Modeling Human Memory for Adaptive Educational Systems

Master’s Thesis

Bc. Pavel Dedík

Brno, Fall 2015
Replace this page with a copy of the official signed thesis assignment and the copy of the Statement of an Author.
Declaration

Hereby I declare that this paper is my original authorial work, which I have worked out by my own. All sources, references and literature used or excerpted during elaboration of this work are properly cited and listed in complete reference to the due source.

Bc. Pavel Dedík

Advisor: doc. Mgr. Radek Pelánek, Ph.D.
Acknowledgement

Firstly, I would like to express my gratitude to my advisor, Radek Pelánek, for his patience, motivation and helpful guidance. My thanks also goes to members of the Adaptive Learning group for their comments and remarks. I am also grateful to my brother for his stylistic revision, and my friends who provided some feedback.
Abstract

Adaptive educational systems enable students to practice different domains of educational content adaptively and make learning personalized. In this thesis, we study and design extensions of student models suitable for applications in adaptive systems while focusing on key aspects of human memory and learning. We evaluate the proposed models using data from the adaptive system for practice of geography facts available at outlinemaps.org, and compare currently used models with ours. Finally, we analyze parameter values of the models and discuss the results.
Keywords

Memory, Learning, Forgetting, Knowledge, Adaptive Educational System, Machine Learning, Student Modeling, Performance Factor Analysis
1 Introduction

Memory could be defined as the biological storage for the information acquired from past experience with the environment. Human memory has been scientifically studied for at least 100 years [1] and today we know that there are many types of memory, all cooperating in the process of memorization so that they could be used to govern subsequent behavior of each individual. Memory is not a single unitary system such as the heart, but a collections of systems where one system is responsible for the encoding of new information while another system helps with its storage over a long period of time. The systems responsible for retention of information in the human brain are also very closely related to learning and forgetting, both of which in turn depend to some extent on memory. That is true because our memory provides the means and structure to link new knowledge faster by association and inference. Hence, human memory is a complex system and the whole underlying process is still broadly studied by researchers who try to understand the complex cognitive processes by exploring the different memory systems of animals (e.g. rats) or patients who suffer amnesia as a consequence of brain damage [2].

The exploration of human memory and learning has applications particularly in education, where our goal is to increase the amount of material students learn in one study session. Before the invention of computers and the growth of the Internet, the ways to test and evaluate new methods of educating students involved classrooms with usually only a limited number of participants. In the last 20 years, students and educational institutions started to use and develop new e-learning systems as a complimentary tool for education. Adaptive educational systems (sometimes referred as adaptive practice systems) are systems that provide online environment for practicing different domains of educational content adaptively.

In adaptive educational systems, our effort is to create sufficiently accurate representation of students in order to make the system personalized, increase students’ motivation and the speed of learning. A part of all adaptive systems are mathematical models which are constructed in order to model learning of individual students and adapt the system to students’ abilities, behavior and knowledge of the subject. These models and their evaluation is very often based on machine learning techniques and is closely related to statistics. The ability to model the adaptive behavior also requires research connected to cognitive psychology.
1. Introduction

One example of a free web-based adaptive educational system is the project called Outline Maps\(^1\) developed at the Faculty of Informatics at Masaryk University. This project helps students practice all kinds of geography facts \(^3\), including world countries, cities, rivers, lakes, mountains, islands, Czech regions, and many others. The practice procedure of geography in the system involves rehearsal of contextual information about a place on a map, i.e. the location, shape or neighbors of a country. The test of students’ knowledge is done by presenting questions requiring the identification of correct association between a name of a place and its position on an outline map (see Figure 1.1).

![Figure 1.1: The screenshot corresponds to a multiple-choice question (with 5 distractors) requiring the student to identify the name of the highlighted country on an outline map of Europe.](image)

Since the system is adaptive, it examines knowledge and skills of students adaptively based on previous answers—by the selection of optimal repeat frequency, the type of question, its difficulty etc. In this context, the practice of geography is somewhat different from ordinary memorization of vocabulary pairs as there are bigger variations in the prior knowledge of individual students.

---

1. Available at [outlinemaps.org](http://outlinemaps.org)

---

2
1. Introduction

1.1 Objectives

The first objective of the thesis is to study the related research to modeling human memory and forgetting mainly in the context of adaptive educational systems. The first objective also involves the study of the relevant machine learning models and the determination of which one is the best candidate considering all its aspects.

Our primary objective is to design, evaluate and analyze the models suitable for adaptive practice systems while accounting for key aspects of human memory and forgetting. The findings should provide insights into how modeling the effect of forgetting improves model’s performance. We are particularly interested with student modeling for application in the adaptive system Outline Maps.

1.2 Outline

The second chapter covers the background of our thesis, we summarize the most relevant attributes of human memory and forgetting that are related to our research, we also describe characteristics of adaptive educational systems as well as mathematical models used in student modeling. In the third chapter, we propose models and their modifications that focus on timing information (e.g. the ages of student’s past trials and response times), we also describe the methods commonly used in machine learning for parameter estimation and metrics for the quantification of model’s performance. The fourth chapter includes evaluation of the proposed models and further analysis of their parameters. Finally, the last chapter concludes the results of our experiments and outlines suggestions for possible future research.
2 Background

In this chapter we present an overview of the related research with the topic of our thesis. Firstly, we describe the key aspects of human memory, how it can acquire new information and knowledge, the process of retention of information, and the retrieval of information from memory. Secondly, we give an overview of techniques suitable for educational systems with the focus on adaptive learning of facts.

2.1 Human Memory

The Greek philosopher and mathematician Plato compared human mind to an aviary in which each bird represented a memory [4]. Today, we have a much better analogy on the human biological storage—the hardware responsible for the storage and organization of data in modern computers. Memory gives humans and other animals the ability to remember, adapt from previous experiences and govern subsequent behavior.

We recognize at least three types of memory as categorized by its capacity and the duration of retained memories:

**Sensory memory** Retains sensory information (e.g. auditory and visual inputs from the environment) for less than one second.

**Short-term memory** Also known as working memory retains information for less than one minute. Short-term memory is very vulnerable to interruption or interference.

**Long-term memory** Has much more capacity than the short-term memory. Retains information potentially for decades, even lifetime.

The gradual change from short-term memory to long-term memory is called *memory consolidation* and is attributed to the *hippocampal system* located in the temporal lobe of the human brain. The hippocampal system is essential for the formation of associations between various elements of specific events and experiences [2], it was also suggested that it is especially relevant in the formation of memories involving places or locations in the environment [5]. Even though consolidation occurs in the hippocampus, the long-term memories are transferred somewhere else—possibly to the *neocortical regions* of the brain. This theory is supported by the fact that patients with damaged hippocampus show deficit for memories obtained shortly before their lesion whereas very remote memories appear to be intact [2].
2. Background

2.1.1 Declarative and Procedural Memory

Similarly as we classified the various types of memory by its duration and capacity, we can also differentiate memory by the type of information it retains. In 1980, N. J. Cohen and Squire introduced the term declarative memory [7]. Declarative memory retains the conscious memory, this involves mainly events and facts. Another type of memory is called the procedural memory and is the type of memory we are not consciously aware of and which makes humans capable of the acquisition of both the motor and cognitive skills. For example, riding a bike or swimming are tasks that are learned and collected as procedural memories [4].

The declarative memory can be further classified into the episodic memory and the semantic memory. The episodic memories are memories that store information associated with a specific event whereas the semantic memories include the knowledge of the meanings of words, factual information and encyclopedic memories [7]. In the human brain, semantic memories seem to be organized hierarchically, i.e. if we acquire the information that Husky is a dog, our brain automatically infers that it has fur, four legs, two ears, it must eat to stay alive, and so on [2]. This representation of semantic relation between concepts is called a semantic network. An example of a semantic network is illustrated on Figure 2.1.

![Figure 2.1: A semantic network from the domain of animals [6].](image)

2.2 The Science of Learning

Throughout the last one hundred years, researchers have been trying to figure out the complex cognitive processes in the human brain responsible for our ability to learn [8]. So far, we have explored human memory from its structural, organizational and functional point of view. In this chapter, we look at memory and learning from the perspective particularly relevant in education and educational systems. Our objective is to understand how to present study material in ways to help people learn, i.e. increase their motivation, the speed of learning, and reduce the need for repetition caused by forgetting.
In a book called *Applying the Science of Learning* by Richard E. Mayer, a professor of psychology at the University of California, the science of learning is organized into the following three components:

- Science of Learning
- Science of Instruction
- Science of Assessment

All of these components are relevant for our research. However, the science of learning is the most related component since it concerns human memory and learning. The science of instruction is about the manipulation of the student’s environment in order to foster learning, it’s the scientific study of how to help people learn. In educational systems, this is usually done by incorporating *instructional policies*, the strategies or a set of instructions that help students maintain engagement and increase the amount gained knowledge in one session. Instructional policies guide students during learning by, for instance, the selection of appropriate questions or the number of options in multiple-choice tests. Finally, the science of assessment seeks to determine what people know, which is important so that we are able to quantify the effectiveness of different instructional methods [8].

### 2.2.1 Self-regulated Study

An important aspect of learning and self-regulated study is the ability to choose the most optimal methods of studying. It is sometimes hard to correctly determine how long or what to study and when to consider the material learned. Often, people stop studying when they do not fully know the material. For example, the effect of forgetting (i.e. the loss of information) can be reduced more dramatically by the correct type of repetition. Repetition can be *massed* or *spaced*, in a massed presentation, an item is typically revised in a short interval many times over with no intervening delay. In contrast, a spaced presentation usually consists of revisions performed across longer periods of time with delayed presentations [9, 10].

Interesting fact about self-regulated study is that in situations where it is allowed to study without time limit or some different higher pressure situation, people often choose to study the most difficult items. Under pressure, people give high priority to relatively easy items. However, in both situations, people do not study items that they think they have already learned [11].
2. Background

2.3 Student Modeling and Memory

Understanding the process of learning is very helpful in educational systems, where we wish to improve student’s representation. This is useful when we need to adapt our system to the needs and knowledge of individual students practicing a particular domain (e.g. the knowledge of animals, Japanese vocabulary, geography, etc.). The construction of a quantitative representation, called a student model, is known as student modeling [12].

The process of learning is very closely related to the topic of our thesis as it leads to the creation of new memory. The study of learning and memory can be divided even further when we realize that we need some way to observe what students know by experiments. Thus, we cover the relevant research to student modeling and memory in three parts:

- Learning
- Memory
- Performance

Although this distinction is not very important, it can help the reader better understand the underlying problem discussed in our thesis and the connection between computer science and cognitive psychology.

2.3.1 Learning

Richard E. Mayer formulated learning as a change in what the student knows caused by the student’s experience [8]. In more psychological terms, learning is the process of encoding, modifying and reinforcing information [13]. For example, the encoding of the location of Portugal can be seen as learning the shape of the country, its neighbor Spain, the surrounding ocean, or even the fact that Portugal was once a part of the Roman Empire, and so on.

A learning curve is the rate of student’s progress in gaining new skill or knowledge by experience in the environment (e.g. by participating in a discussion, reading a book or riding a bike). Generally, the speed of learning is proportional to the amount learned and amount to yet learn. This principal can be written mathematically as follows:

\[
\frac{dK}{dt} = a \cdot (K_{\text{max}} - K) \quad (2.1)
\]

The coefficient \(a\) in the Equation 2.1 represents the rate of learning, \(K\) is the student’s knowledge and \(K_{\text{max}}\) the maximum knowledge possible.
Several functions modeling the learning curve were proposed in the past, namely the power law, exponential and hyperbolic function. Particularly the power law of learning has been suggested by a lot of researches (e.g. Newell and Rosenbloom [14]) and it and in many cases it provides the best fit [15, 16].

![Learning curve](image)

Figure 2.2: Learning curve (left) and forgetting curve (right).

2.3.2 Memory

Memory is the biological storage that retains encoded information, for example the location of Portugal. The long-term memory decay is called forgetting and similarly as learning, the speed of forgetting is inversely proportional to the volume of remaining material.

\[
\frac{dK}{dt} = -bK \tag{2.2}
\]

In Equation 2.2, the parameter \( b \) represents the rate of forgetting and \( K \) is the student’s knowledge of the material. Solving the differential equation yields a negative exponential. Most researchers believe that forgetting respects the power law [4, 17], although in some cases it is argued that exponential function with time scaled to \( \sqrt{t} \) describes the phenomenon better [18].

In chapter 2.2.1, we explained that a spaced presentation leads to better and more stable long-term memory. This phenomenon is called the spacing effect and is illustrated on Figure 2.3. We should also mention that our concern is only the factual knowledge which is stored as declarative memory, i.e. the conscious knowledge (knowing what) such as the world’s countries or the English vocabulary.
2. Background

Figure 2.3: Example of the spacing effect. The student $s_1$ attended three sessions at times $t_0$, $t_1$ and $t_2$, and practiced 50 items on each, the student $s_2$ attended one session at the time $t_0$ where they practiced 150 items.

2.3.3 Performance

The performance of students helps us determine what they know. Performance can be estimated from the speed and precision of recall, for example by a multiple-choice test where the correctness of answers and response time is measured [13]. Performance can be seen as an instrument that describes the student’s knowledge, the understanding of which is important because it helps us guide the instructional policy.

One of the most useful metric of student’s performance is the memory activation (a measurement of the ability to retrieve an item from memory). In our case, an item is any factual information that can be learned. Memory activation can be measured indirectly by observing the following attributes of students when an item is practiced:

**Probability of recall** Probability of the student recalling the practiced item, this can be measured as the fraction of the number of successful recollections and the number of all presentations [13].

**Latency of recall** Latency of the student when retrieving the practiced item from memory. Latency of recall can be measured by observing the response times of students [13].

10
2. Background

**Savings in relearning** The number of required revisions of the practiced item in order to fully regain its knowledge [13, 4].

We can further distinguish the following levels of learning as measured by memory activation (illustrated in Figure 2.4):

![Figure 2.4](image)

**Figure 2.4:** Levels of learning [13]. The line indicates a simplified example of the progress of student’s knowledge. Study time takes place between the times $t_0$ and $t_1$, this is followed by the post-study time.

**Familiarity** The student has feeling they knew the item in the past but cannot remember anymore.

**Recognition** The student recognized the item when presented multiple-choice options but could not remember otherwise.

**Recall** The student is able to recall the item with some effort. Note that in cognitive science we distinguish *free recall* and *cued recall*. Free recall is the ability to remember an item without any help (e.g. recalling the name of a country), in the case of cued recall, we are given an information which can help us remember (e.g. first letter of the country we are to remember).

**Automaticity** The student recalls the item instantaneously when presented. Note that the level of automaticity can be measured by the latency of recall [13].
2. Background

2.4 Relevant Student Models

In this section, we discuss the models relevant to our work. There are two main components that concern us and we can treat them independently—the first is the estimation of probability that a student knows an item before they answer the very first question about an item, the second is the estimation of the current knowledge of a student which combines the prior knowledge and the knowledge the student acquired as they practiced [3].

2.4.1 Logistic Function

We begin with the equation of the sigmoid logistic function, which is used for prediction of correct performance in all of the following models.

\[ \sigma(m) = \frac{1}{1 + e^{-m}} \]  \hspace{1cm} (2.3)

Equation 2.3 shows the sigmoid function \( \sigma : \mathbb{R} \rightarrow \mathbb{R} \) with argument \( m \) representing memory activation. If \( m = 0 \), the probability of correct performance is 0.5.

In the case of multiple-choice questions, we also want to take into account the probability that a student guesses the correct answer. This can be done by introducing the argument \( r \) representing the ratio of the number of correct answers and all possible answers, for instance \( r = 1/6 \) if there are exactly 6 options while only 1 option is correct. The modified version of the sigmoid function is formulated in Equation 2.4.

\[ \sigma(m, r) = r + (1 - r) \frac{1}{1 + e^{-m}} \]  \hspace{1cm} (2.4)

Note that \( \sigma(m, 0) = \sigma(m) \). The shape of such a logistic function is shown in Figure 2.5.

2.4.2 Elo System

Elo rating system and its extension Glicko are very popular mathematical models [19] used in competitor-versus-competitor games such as chess where the goal is to rate players with scores according to the number of opponents they defeated or were defeated by while also honoring each player’s skill [20]. Recent research has shown that Elo system is also suited for student modeling and it was previously very successfully adopted for the estimation of prior knowledge of students [21].
In the adopted version of Elo system, we have the student’s skill $\theta_s$ and the difficulty of an item $d_i$. Equations 2.6 and 2.7 represent an update of the student’s skill and the item’s difficulty after one answered question. The parameter $R$ is 0 or 1 depending on the correctness of the student’s answer. The probability that a student $s$ with a given skill $\theta_s$ will answer correctly on the presented question of difficulty $d_i$ is estimated by the logistic function in Equation 2.5.

$$P(R = 1 | s, i) = \sigma(\theta_s - d_i) \quad (2.5)$$

$$\theta_s \leftarrow \theta_s + K(R - P(R = 1 | s, i)) \quad (2.6)$$

$$d_i \leftarrow d_i - K(R - P(R = 1 | s, i)) \quad (2.7)$$

The constant parameter $K$ affects the change in the estimates $\theta_s$ and $d_i$. Higher value means faster change after few questions, in contrast lower value makes the change slower. This is a problem in cases where the number of answers varies over time. It has been demonstrated that the use of an uncertainty function $\frac{\alpha}{1 + \beta n}$, which considers the number of answers of students in the system, makes the predictions more stable and increases accuracy [20].
2. Background

2.4.3 Performance Factor Analysis

Performance Factor Analysis (PFA) is a student modeling approach based on the Learning Factor Analysis (LFA) [22, 23]. Memory activation in the PFA model can be seen as the sum of the item’s difficulty and all student’s answers with estimated weight for successful and unsuccessful answers. The standard PFA equation is formulated with the incorporation of knowledge components (KCs), which may include skills, concepts, facts, or any fragment of a domain-specific information that should be used to accomplish a task [24]. In our work, we use PFA and its extensions for the estimation of current knowledge.

\[ m(i, s, f) = \sum_{j \in KCs} \beta_j + \gamma_j s_{i,j} + \delta_j f_{i,j} \]  

(2.8)

\[ P(m) = \sigma(m) \]  

(2.9)

In Equation 2.8, where \( m \) is a function of the student’s knowledge of item \( i \), the parameter \( \beta_j \) is the difficulty of the knowledge component \( j \). The counts of current successes and failures of the practiced item \( i \) and the knowledge component \( j \) are represented by \( s_{i,j} \) (number of successes) and \( f_{i,j} \) (number of failures), where \( \gamma_j \) and \( \delta_j \) give a weight to each success and failure. An item is retrieved only if its memory activation \( m \) is above certain value, e.g. if \( m = 0 \), we can calculate from Equation 2.11 that the probability of correct performance is 0.5.

The Elo/Extended Version

The standard PFA model is defined in terms of knowledge components which is not always needed. The main disadvantage, however, is the inability to consider the order of answers. Another problem with the standard PFA model is that it does not take into account the probability of guessing [3]. Both of these issues are solved by the adjustments made in Equation 2.10.

\[ m \leftarrow \begin{cases} 
  m + \gamma \cdot (1 - P(m)), & \text{if the answer was correct} \\
  m + \delta \cdot P(m), & \text{otherwise}
\end{cases} \]  

(2.10)

\[ P(m) = \sigma \left( m, \frac{1}{n} \right) \]  

(2.11)

The initial value of \( m \) can be estimated from the Elo model which was discussed in chapter 2.4.2, i.e. \( m = \theta_s - d_i \). The variable \( n \) in Equation 2.11
2. Background

represents the number of options in a multiple-choice question. The probability of guess is thus $\frac{1}{n}$. Note that this is a simplification as it does not consider the possibility that a student answers correctly only by ruling out some portion of the other displayed options (in which case they do not know the correct answer even though there is higher chance of guess) [25].

To get a better idea of how this model works, suppose that a new student $s$ decides to practice autonomous communities of Spain. Before the first question about any item is presented, the memory activations of all items in the system are estimated from the Elo model, i.e. $m = \theta_s - d_i$ for each student $s$ and item $i$. Now, the system can select an appropriate question for the student, this is often based on the probability of correct performance $P(m)$. After the student $s$ answers the presented question, the memory activation $m$ of the autonomous community $i$ is locally updated for the student $s$ depending on correctness of their answer. In the next repetition of the same item, the updated memory activation of the student is used again for the estimation of correct performance and updated subsequently. This process is repeated for each answered item.

The Decay Factor

Another variation of the original PFA model was proposed by Yue Gong [26]. The idea of the model is based on the fact that we expect the student who answered in four presentations two times correctly in the last two presentations to perform better than the student who answered correctly in the first two presentations. The extended model introduces a decay factor $\xi$ that changes the behavior of the parameters $s_{i,j}$ and $f_{i,j}$ by penalizing the older answers. The modifications are shown in Equations 2.12 and 2.13.

$$s_{i,j} = \sum_{k=1}^{n} y_k \cdot \xi^{n-k}$$ (2.12)

$$f_{i,j} = \sum_{k=1}^{n} |y_k - 1| \cdot \xi^{n-k}$$ (2.13)

The variable $y_k$ represents the correctness of the $k$-th question. For example, suppose that a student was presented with a question in five trials. Further, suppose that the student answered incorrectly in the first three trials and answered correctly in the last two. Now for the given value of $\xi = 0.8$, we can calculate the values of $s_{i,j}$ and $f_{i,j}$. Equation 2.14 demonstrates the result of such scenario where we see that the weight of all successes surpasses the weight of all failures.

15
2. Background

\[ s_{i,j} = 0 \cdot 0.8^4 + 0 \cdot 0.8^3 + 0 \cdot 0.8^2 + 1 \cdot 0.8^1 + 1 \cdot 0.8^0 \]
\[ = 1.8 \]
\[ f_{i,j} = 1 \cdot 0.8^4 + 1 \cdot 0.8^3 + 1 \cdot 0.8^2 + 0 \cdot 0.8^1 + 0 \cdot 0.8^0 \]
\[ = 1.56 \] (2.14)

The problem of this model is that it cannot be straightforwardly adjusted so that it includes the probability of guessing, particularly in cases where the practices were presented in the form of multiple-choice questions with varied number of options.

2.4.4 Models of Forgetting

So far, all the presented models did not account for the effect of forgetting. Although the use of a decay factor can be seen as a way to penalize past answers, it does not consider intervals of time between repetitions of an item. In the ACT-R (Adaptive Character of Thought–Rational) modeling system [27], the memory activation \( m \) of a student is a function of prior practices. This function does not account for correctness of prior answers or their difficulty, it considers only counts and times of past repetitions.

\[ m(t) = \ln \sum_{i=1}^{n} t_i^{-d} \] (2.15)

The memory activation function used in ACT-R model is shown in Equation 2.15, the argument \( t \) is a vector of seconds that passed since each of the \( n \) repetitions were performed by a student. The parameter \( d \) represents memory decay (the speed of forgetting). Note that the equation is just a simplification of the reality and does not take into account many very important aspects of forgetting, e.g. the mentioned spacing effect.

An item will be retrieved only if its memory activation is above certain threshold. In ACT-R model, the probability of recall is given by the value of the threshold \( \tau \) and the measure of noise \( s \):

\[ P(m) = \sigma \left( \frac{\tau - m}{s} \right) \] (2.16)

Philip I. Pavlik and John R. Anderson [28] developed an extended version of the equation in which the decay is a function of the activation at the time the item was presented. The Equations 2.17 and 2.18 demonstrate the replacement of the parameter \( d \) with a recurrent function \( d_i \) that depends on the past timing distances between presentations.
2. Background

\[ d_i = c e^{m_i} - 1 + a \]  \hspace{1cm} (2.17)

\[ m(t) = \ln \sum_{i=1}^{n} t_i^{-d_i} \] \hspace{1cm} (2.18)

The parameter \( c \) affects the impact of the spacing effect while the parameter \( a \) affects the slope of the decay. Since \( m_0 = -\infty \), the value of \( d_i \) is always equal to \( a \) at the first student’s practice of an item. Instinctively, when \( c = 0 \), the result of the equation is equivalent with the Equation 2.15. Because the computation is recursive and a bit complex, we present the pseudo-code of the algorithm computing the memory activation function (see Algorithm 1). Note that the time complexity of the function is \( O \left( n(n-1)^2 \right) \).

Algorithm 1 The function MemoryActivation : \( \mathbb{N}^n \rightarrow \mathbb{R}^n \) takes the vector \( t \) in descending order, e.g. \([56800, 56400, 3600, 60, 0]\) (the last zero is the current practice). The result of the computation is a vector \( m \) of student’s memory activations during each practice.

1: function MemoryActivation(t)
3: \( n \leftarrow \text{dimension}(t) \)
4: \( m_0 \leftarrow -\infty \)
5: for \( i \leftarrow 2 \) to \( n \) do
6: \( s \leftarrow 0 \)
7: for \( j \leftarrow 1 \) to \( i - 1 \) do
8: \( d_j \leftarrow c e^{m_j} - 1 + a \)
9: \( s \leftarrow s + (t_j - t_i)^{-d_j} \)
10: \( m_i \leftarrow \log(s) \)

The purpose of this adjustment is to address the spacing effect by extending the ACT-R’s activation equation. The experiments done by P. I. Pavlik and J. Anderson showed better fits and less variations in parameters. However, the experiments were performed in controlled environment on students who had no prior knowledge of the material they were practicing (Japanese vocabulary) [28]. In domains where the students differ in their age, education and the prior knowledge of the practiced material, the parameters might be very unstable or even impossible to estimate.


3 Models and Methods

The models presented in chapter 2.4 seem to work well either in the context of adaptive systems where we are not concerned with timing information or have been tested in controlled environment where the students had no prior knowledge of the presented material. In this chapter, we elaborate the methods we used to further improve the performance of models by taking into account the timing information of students’ answers, i.e. the response times or the age of previous presentations. We are interested primarily in domains where the prior knowledge varies widely between students.

3.1 Models Based on Timing Information

In this section we discuss several models which aim at modeling students’ memory with the usage of timing information of answers.

3.1.1 The Extended Model

The extended PFA model summarized in chapter 2.4.3 can be further enhanced when we utilize the timing information by locally changing the memory activation in prediction. The difference in seconds between the times of the current question and the last answer is passed to a time effect function (we use this term as it was used in a related work [29]). The time effect function increases or decreases the probability of recall or the memory activation depending on the age of the previous trial. The updated equation with a time effect function is shown in Equation 3.1.

\[ P(m) = \sigma(m + f(t)) \] (3.1)

The advantage of this model is that it is easy to employ Elo model for estimation of prior knowledge as well as to account for the probability of guessing.

3.1.2 The Alternative Model

Another way of dealing with timing distances between trials is by changing the core idea of the model with a decay factor \( \xi \) presented in chapter 2.4.3. The model takes into account the order of questions by penalizing past trials with the decay factor, yet does not consider the ages of student’s previous attempts. This problem can be resolved by replacing \( \xi^{n-k} \) with two time effect functions for correct and incorrect performances.
3. Models and Methods

\[ s_{i,j} = \sum_{k=1}^{n-1} y_k \cdot f_{succ}(t_k) \]  
\[ f_{i,j} = \sum_{k=1}^{n-1} |y_k - 1| \cdot f_{fail}(t_k) \]

Equations 3.2 and 3.3 show the incorporation of time effect functions \( f_{succ} \) and \( f_{fail} \) in the model (\( f_{succ} \) for successful attempts and \( f_{fail} \) for failed attempts). Each \( t_k \) equals the number of seconds that passed between the current and the \( k \)-th practice. The weight of successes and failures is thus dependent on the ages of the prior practices.

In this model, we also ignore the parameters \( \gamma \) and \( \delta \) as the weight of each success and failure depends on parameters of the chosen time effect functions.

The updated memory activation function \( m \) is shown in Equation 3.4.

\[ m(i,s,f) = \sum_{j \in KCs} \beta_j + s_{i,j} + f_{i,j} \]

As we discussed in the chapter 2.4.3, the problem arises with multiple-choice questions. Another complication is the choice of some good time effect functions that fit the data well.

3.1.3 Response Times

The response time of student may indicate student’s level of learning. If the student answers quickly, it often implies knowledge at the level of automaticity. We address this phenomenon by increasing the memory activation \( m \) after each answered item.

\[ m \leftarrow \begin{cases} 
    m + \gamma \cdot (1 - P(m)) + r_{succ}(t_d), & \text{if the answer was correct} \\
    m + \delta \cdot P(m) + r_{fail}(t_d), & \text{otherwise}
\end{cases} \]

Equation 3.5 shows the adjusted update rule with unary functions \( r_{succ} \) for correct answer and \( r_{fail} \) for incorrect answer—this distinction is necessary since it has been shown that the response time and the probability of correct performance depends on the performance in previous trials [30]. The function argument \( t_d \) represents student’s delay, i.e. the difference between the time an item was presented and the time the item was answered.
3.2 Quantifying the Quality of Predictions

One way to quantify the quality of model predictions is to use metric functions. A good choice of a metric function is important for accurate evaluation of performance of student models. In our case, we are not interested strictly in the correctness of student’s answers, the goal is to precisely assess their knowledge. The information of student’s knowledge is crucial if we do not want to demotivate the student by too difficult or simple questions. For this purpose, a good choice is the Root Mean Squared Error metric (see Equation 3.6) or the less popular Log Likelihood (see Equation 3.7) [25].

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - y_i)^2}{n}} \quad (3.6)
\]

\[
LL = \sum_{i=1}^{n} y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \quad (3.7)
\]

The RMSE metric represents “error”, its the square root of the sum of all squared differences between predicted values \( p_i \) and true values \( y_i \) divided by the number of samples in a data set. RMSE yields a number between 0 and 1, where 0 represents perfect predictions. LL metric is the logarithm of likelihood, the LL value is a negative number dependent on the size of the data set (the bigger the dataset, the lower the LL value). Higher values of LL represent the better predictions.

In domains such as information retrieval or pattern recognition, other types of metrics are commonly used which are based on qualitative understanding of errors [25] instead of their probabilistic value. An advantage of this type of metrics is that it can be easily used in multi-classification tasks. These metrics, however, depend on a chosen threshold, e.g. in case the threshold is set to 0.5, the predictions 0.51 and 0.98 are classified as positive while predictions 0.49 and 0.04 are classified as negative.

<table>
<thead>
<tr>
<th>Predicted Outcome</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Observed Value</strong></td>
<td>Positive</td>
<td>True Positive (TP)</td>
</tr>
<tr>
<td></td>
<td>Negative</td>
<td>False Positive (FP)</td>
</tr>
</tbody>
</table>

Table 3.1: Confusion matrix.
3. Models and Methods

The Table 3.1 shows the confusion matrix of binary classifier (i.e. the mis-labeling of classes). Once the confusion of classes is calculated, we can derive for example precision (see Equation 3.8) and accuracy (see Equation 3.9).

\[
Precision = \frac{TP}{TP + FP} \quad (3.8)
\]

\[
Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (3.9)
\]

Precision and accuracy are not very fit in our case as we are more interested in probabilistic understanding of errors. There is, however, one commonly used metric based on marking predictions as true/false positives/negatives called Area Under the ROC Curve (AUC). This metric measures the performance across all possible thresholds. The result is a number between 0.5 and 1, where 1 represents perfect predictions of the classifier. Note that if all predictions are divided by 2, AUC stays the same, this property is sometimes criticized—hence it should be used with caution [25].

3.3 Parameter Estimation

The goal of the fitting procedure is to find the optimal parameters that perform well on new samples. In our case, this involves maximizing the predictive ability of our student models by discovering the most fitted parameters. In this section, we describe the principal algorithms that are suitable for parameter estimation in our thesis.

3.3.1 Grid Search

This technique is used for parameter optimization, it exhaustively generates a grid of all combinations of selected parameter values where each point in the grid is a function value of a chosen objective metric (e.g. RMSE). The fittest combination of parameters is represented by the best score of the metric (e.g. has lowest error). Figure 3.1 illustrates a result of the grid search algorithm performed on PFA model with two parameters $\gamma$ and $\delta$.

The disadvantage of this optimization method is the high computational complexity especially in cases when the model contains a lot of parameters. Also, there is no simple way to plot or visualize a high-dimensional grid.

3.3.2 Hill Climbing

This method starts from a chosen position in the parameter space and evaluates the score of an objective function (metric function in our case). The
Figure 3.1: Result of the grid search performed on the PFA model. The figure shows a grid of model’s performance evaluated as RMSE (left) and AUC (right) for each combination of parameters.

The model is then trained using the values from the enclosing neighborhood of the parameter space, afterwards the best combination of parameters is selected and the process repeated. This continues until there is no other parameter value in the neighborhood with a better score of the objective function, i.e. in the case of discrete values, we found the local minimum, in the case of continuous values, we are as close to the local minimum as possible under the preset conditions [31].

3.3.3 Gradient Descent

Gradient descent is optimization algorithm very similar to hill climbing, the difference is that gradient descent has indication which way to “climb”. The goal of gradient descent is to find the best parameters for a function called hypothesis [32], this is done by finding the lowest value of the cost function $J(w)$ (or the objective function). The cost function represents the error of a chosen hypothesis $h_w(x_i) = w^T x_i = \sum_{j=0}^{n} w_j x_{i,j}$, where $w_j$ is the value of the $j$-th parameter and $x_{i,j}$ the $i$-th value of the $j$-th feature (e.g. the skill of the student $i$).

$$J(w) = \frac{1}{2m} \sum_{i=1}^{m} (h_w(x_i) - y_i)^2$$  (3.10)

The definition of the cost function is formalized in Equation 3.10, it is the sum of all squared errors of hypotheses $h_w$ over all examples from data set of size $m$ divided by $2m$. Our goal is to minimize the value of $J(w)$ which
can be done efficiently by the estimation of the gradient using the following update rule:

\[ w_j \leftarrow w_j - \alpha \frac{\partial}{\partial w_j} J(w) \quad \text{for all } j \quad (3.11) \]

The partial derivatives help indicate the surface of the cost function, it gives us the information which direction to take in order to reach the closest local minimum. The value of \( \alpha \) is the size of one step also called the learning rate, if the value is too big the algorithm might not be able to ever reach a local minimum, on the other hand if the value is too small, it is less efficient and the computation takes longer. Notice that the main difference between gradient descent and hill climbing is that the latter does not have any indication of which direction to take next and how big one step in that direction should be [31].

\[ \text{Figure 3.2: 3D visualization showing first 7 iterations of gradient descent applied to an example with two parameters.} \]

\[ \text{3.3.4 Approximating the Gradient} \]

In this chapter, we discuss possible approximations of the gradient descent to optimize parameters of the PFA model and its variations. The most ob-
vious and easiest way to implement approximate gradient could average the difference between correctness of student’s answer $y_i$ and model’s predicted probability $p_i$ that the item $i$ is answered correctly. The batch update rule would be defined as follows:

$$w_j \leftarrow w_j - \alpha \frac{1}{n} \sum_{i=0}^{n} (p_i - y_i) \text{ for all } j$$ (3.12)

This method has several problems. First, it is easy to demonstrate that it does not work well even in cases where the number of parameters is very low. The parameters get stuck in local minimum as can be observed in Figure 3.3, e.g. different values of the step size $\alpha$ lead to different values of the parameters $\gamma$ and $\delta$. Another disadvantage of this technique is that the whole process of parameter optimization has to run periodically, which is of course not very desirable in online learning.

Pelánek in his work [29] figured out a bit more complicated way of parameter estimation which combines some aspects of gradient descent and is suitable in online learning. The update of parameters happens after arrival of every new data point (i.e. student’s answer). This method is very practical particularly in our case where we need to find the most optimal and best calibrated time effect function.

Algorithm 2 formally describes an update of parameters in PFA model with two parameters $\gamma$ and $\delta$. The update is performed whenever a question
3. Models and Methods

is answered by a student. To understand the idea of this method, we can examine four situations:

- Answer A was correct and \( P(m) > 0.5 \) – memory activation \( m \) is low and should be increased (e.g. by increasing \( \gamma \) or \( \delta \))

- A was correct and \( P(m) \leq 0.5 \) – memory activation \( m \) is very low and should be greatly increased (e.g. by increasing \( \gamma \) or \( \delta \))

- A was incorrect and \( P(m) \leq 0.5 \) – memory activation \( m \) is high and should be decreased (e.g. by decreasing \( \gamma \) or \( \delta \))

- A was incorrect and \( P(m) > 0.5 \) – memory activation \( m \) is very high and should be greatly decreased (e.g. by decreasing \( \gamma \) or \( \delta \))

The process of updating the parameters is repeated for all answers, i.e. it tries to approximate the optimal combination of parameter values by minimizing discrepancy in the above situations.

The process of change of parameters in PFA model using this method is visualized in Figure 3.4. We can also notice that with each iteration of the algorithm, the parameters are updated while error in predictions as measured by RMSE generally decreases.

**Figure 3.4:** Iterations of the algorithm based on gradient descent for parameter estimation in PFA model. The figure on left shows the progress of \( \gamma \) and \( \delta \) with each new answer. The figure on right indicates the improvement of predictive ability of the model by measuring RMSE each time 500 places are answered by students.
Algorithm 2 The algorithm demonstrates update of parameters after one trial of the item $I$. The function expects the practiced item $I$ and current values of global parameters $\gamma$ and $\delta$. Output is a tuple of updated $\gamma$ and $\delta$. Note that $w_\gamma$ and $w_\delta$ (initially equal to 0) help indicate how much to change activation $m$ of the item $I$ given the previous values of $P(m)$ and student’s answers.

1: function UpdateParameters($I, \gamma, \delta$)
2: \hspace{1em} $m \leftarrow$ Activation($I$) \Comment{Returns memory activation of the item $I$.}
3: \hspace{1em} $A \leftarrow$ LastAnswer($I$) \Comment{Returns the last practice of the item $I$.}
4: \hspace{1em} if IsCorrect($A$) then \Comment{Checks correctness of the last answer.}
5: \hspace{2em} $s \leftarrow 1 - P(m)$
6: else
7: \hspace{2em} $s \leftarrow 0 - P(m)$
8: \hspace{1em} $w_\gamma, w_\delta \leftarrow$ GetWeights($I$) \Comment{Returns $w_\gamma, w_\delta$ local for the item $I$.}
9: \hspace{1em} $\gamma \leftarrow \gamma + \alpha \cdot s \cdot w_\gamma$ \Comment{The parameter $\alpha$ is the learning rate.}
10: \hspace{1em} $\delta \leftarrow \delta + \alpha \cdot s \cdot w_\delta$
11: \hspace{1em} if IsCorrect($A$) then
12: \hspace{2em} $w_\gamma \leftarrow w_\gamma + s$
13: \hspace{1em} else
14: \hspace{2em} $w_\delta \leftarrow w_\delta + s$
15: \hspace{1em} SetWeights($I, w_\gamma, w_\delta$) \Comment{Updates $w_\gamma, w_\delta$ locally for the item $I$.}
16: \hspace{1em} return $\gamma, \delta$
3. Models and Methods

3.4 Time Effect Functions

The goal of a time effect function is to locally increase student’s memory activation $m$ of the practiced item while taking into account the time a previous trial of the item was performed. A time effect function thus increases memory activation based on the age of the last answer. For our purpose, we have chosen the following functions with parameter $t$ representing the time of the last presentation and two parameters $a$ and $c$:

- $f_{\text{log}}(t) = a - c \cdot \log(t)$
- $f_{\text{exp}}(t) = a \cdot e^{-c\sqrt{t}}$
- $f_{\text{pow}}(t) = a/t^c$

These functions were chosen as simple candidates with only a few parameters, other candidates we experimented with were just similar variations and did not improve the performance of our models in any significant way. The shape of each of these time effect functions with logarithmically scaled $x$-axis is depicted in Figure 3.5.

![Figure 3.5: Candidates of time effect functions we inspected and evaluated in our analysis. Note that the $x$-axis is log scaled.](image-url)
3. Models and Methods

3.4.1 The Staircase Function

If we want to approximate the exact shape of a time effect function, we can define a staircase function suggested by Pelánek [29] with fixed intervals $i$. In each interval $(i_j, i_{j+1}]$, we preserve a learned value $a_j$ which represents an increase in memory activation. The formal definition of the staircase function $f_{\text{staircase}}$ is formalized in Equation 3.13.

$$f_{\text{staircase}}(t) = \begin{cases} 
    a_1, & \text{if } i_0 < t \leq i_1 \\
    a_2, & \text{if } i_1 < t \leq i_2 \\
    \vdots \\
    a_n, & \text{if } i_{n-1} < t \leq i_n 
\end{cases} \quad (3.13)$$

Applying simple linear algebra, we can further modify the staircase function so that the memory activation between two points is a linear function, this makes a better approximation of the learned values. The definition of the adjusted function is described in Equations 3.14 and 3.15, where $T$ is a set of ages of all answers in seconds.

$$\hat{i}_0 = i_0$$
$$\hat{i}_j = \text{mean} \{ t \in T \mid i_{j-1} < t \leq i_j \} \quad (3.14)$$

$$f_{\text{staircase}}(t) = \begin{cases} 
    a_1, & \text{if } \hat{i}_0 < t \leq \hat{i}_1 \\
    (t - \hat{i}_1) \frac{a_2-a_1}{i_2-i_1} + a_1, & \text{if } \hat{i}_1 < t \leq \hat{i}_2 \\
    \vdots \\
    (t - \hat{i}_{n-1}) \frac{a_n-a_{n-1}}{i_n-i_{n-1}} + a_{n-1}, & \text{if } \hat{i}_{n-1} < t \leq \hat{i}_n 
\end{cases} \quad (3.15)$$

Note that $\hat{i}_0$ is in our case always equal to 0 and $\hat{i}_n$ to infinity (in which case the memory activation in the interval is equal to $a_n$).
4 Evaluation and Analysis

In the first part of this chapter, we briefly characterize the nature of the data set used for evaluation of the models summarized in chapters 2.4 and 3.1 and discuss some specifics of the implementation. In the second part, we report on evaluation of models predictions, parameter fitting procedure, stability of parameters, and extensions of the models. In the last part, we summarize results of further analysis of the models.

4.1 Data Set

The experiments were performed using data generated by real students from the adaptive system Outline Maps (outlinemaps.org)—a web-based educational tool for learning geography. The system enables to practice sets of different types of geography facts including countries, cities, rivers, lakes, countries’ regions, provinces or autonomous communities, mountains and islands. Each type contains tens to hundreds distinct places.

Since the educational system is adaptive, the selection of questions is based on student’s knowledge instead of being statically programmed by an expert. The system has target probability 75%, i.e. the questions which are expected to be too hard or too easy for the student are penalized by a scoring function. The scoring function is responsible for the selection of question to be practiced [33]. The adaptability of the system should be also taken into consideration as it might have affected the results of our analysis.

The data set is available online [34] and consists of more than 10 million answers from tens of thousands of unique users, mostly Czech students. In our experiments, we filter the data set so that it contains only answers of students who practiced at least 50 items, furthermore, the places that were answered by less than 100 students were also removed. Note that we usually run our experiments only on a portion of the whole data set.

Finally, the data set contains answers of very different students with distinct prior knowledge, some students are still in school and practice geography in classes, other students already finished school and want to reinstate their forgotten knowledge (this is further analyzed in chapter 4.7). There are also huge differences between continents, regions and other types of places, all of which may vary in the difficulty of learning, some may be harder for the majority of Czech students to retain and are thus forgotten faster (e.g. some small African country that most students never heard of even in a context of another country).
4. Evaluation and Analysis

4.2 Toolchain

The models were implemented in Python programming language. Experiments were performed in the Jupyter Notebook interactive environment\(^1\). Here is a list of the used libraries and modules:

- SciPy, NumPy, Pandas – numerical calculations and data analysis.
- Matplotlib, Seaborn, NetworkX – data visualization.
- Scikit-Learn – miscellaneous machine learning helpers.

4.3 Baseline

As a baseline, we use the standard PFA model and two of its extensions. We use acronyms for all evaluated models, here is a list of the baseline models as their acronyms occur in our experiments:

- PFA – The original performance factor analysis which we described in chapter 2.4.3. Note that our implementation doesn’t consider multiple knowledge components.
- PFA/E – A version of the original PFA model with some aspects of Elo model which we discussed in chapter 2.4.3.
- PFA/G – Another version of the original PFA model with a decay factor, the characteristics of the model were outlined in chapter 2.4.3.

Note that for the estimation of prior knowledge, we use the Elo model briefly described in chapter 2.4.2. In all our models, the initial memory activation \(m\) is estimated as \(\theta_s - d_i\) as we explained in chapter 2.4.3.

4.4 Response Time

The response time of a student may indicate student’s knowledge of an item. If the student answered quickly, almost automatically, it is very likely they either know the place very well or are purely guessing, depending on the correctness of their answer. On the other hand, when the response is longer, the student is probably familiar with the item and might eventually recall the correct answer. The relationship between response times and students’

\(^1\) Jupyter Notebook is an open sourced web application for interactive computing, see https://jupyter.org/
knowledge was discussed in a related work [30], here we perform experiments with models that reflect this phenomenon.

Figure 4.1 illustrates the relationship between students’ response times and correctness of answers. If the student answers suspiciously fast (response time is lower than 800 milliseconds), it usually means they are guessing. The data set suggests that the highest probability of correct performance is between 1500 and 2000 milliseconds of response time.

![Graph showing response times and correctness.](image)

**Figure 4.1:** Response times of students. Figure (a) shows relationship between response times and correctness of answers. Each line represents items which students answered either correctly or incorrectly in the previous trial. The data set was divided into 40 bins where each point is an average of correctness of answers from the corresponding bin. The number of items in each bin is depicted in figure (b) (note that the y-axis is logarithmically scaled).

We evaluate one model focused on improving performance by taking into account the response times of students:

- **PFA/E/RT** – An extended version of the PFA/E model which alters student’s knowledge by respecting past response times. We described the model in chapter 3.1.3.

Note that some places cover wider area on the map than others, i.e. a question asking a student to locate Russia has generally lower response time than a question requiring to locate Andorra since usually students have to zoom using the mouse wheel before they can respond.
4. Evaluation and Analysis

4.4.1 Parameters

The extended PFA model has 2 parameters when we use Elo model for the estimation of prior knowledge. Thus we need to estimate $\gamma$, $\delta$ and the response time functions $r_{\text{succ}}$ and $r_{\text{fail}}$. We learned the shape of both functions from the data as indicated in Figure 4.1a, however, we also use a parameter $b$ that controls the influence of both functions. The parameter $b$ was estimated using the hill climbing algorithm.

4.5 Memory Decay

In this chapter we describe parameter optimization and calibration techniques of the models focused on memory decay and forgetting. We focus on the following extensions of models:

- **PFA/E/T** – An extended version of the PFA/E model which adapts the idea of a time effect function. We described the model in chapter 3.1.1. In our analysis we examine several time effect functions formalized in chapter 3.4 including the staircase function.

- **PFA/G/T** – Our version of the PFA/G model which uses a time effect function instead of a decay factor. The differences were stated in chapter 3.1.2.

4.5.1 Parameters

Standard PFA model has 3 global parameters when we consider only one knowledge component—the difficulty $\beta$ and the weight of each success ($\gamma$) and failure ($\delta$). In cases where we use Elo model for prior knowledge estimation, we can replace the global parameter $\beta$ with $\theta_s - d_i$ (i.e. the difference between skill of the student $s$ and difficulty of the item $i$) which leaves 2 parameters we need to estimate in the model (disregarding the parameters of a time effect function).

However, a time effect function has to be selected as well. The list of functions we used in our experiments was presented in chapter 3.4. Because all these functions have only the parameters $a$ and $c$, in the PFA/E/T model, we are required to estimate $\gamma$, $\delta$ and parameters of a chosen time effect function. In the PFA/G/T model, we need to estimate parameters of two time effect functions, $f_{\text{succ}}$ and $f_{\text{fail}}$. We were able to estimated the parameters of both models using hill climbing algorithm detailed in chapter 3.3.2.
In our experiments related to the staircase function formalized in chapter 3.4.1, we divided the vector \( \mathbf{i} \) producing 10 intervals with values between 0 seconds, 60 seconds, 90 seconds, 150 seconds, 5 minutes, 10 minutes, 30 minutes, 3 hours, 24 hours, 5 days, and more than 5 days. We picked the values so that the intervals containing the ages of last trials are easy to read and also contain sufficient amount of answers. Thus, the number of parameters to estimate is 12 including \( \gamma \) and \( \delta \). We learned the estimates from the data set using the adjusted gradient descent described in chapter 3.3.3. The fitted staircase function is illustrated in Figure 4.2.

The figure shows that the parameters are rather stable (with \( \gamma = 1.814 \pm 0.276 \) and \( \delta = 0.827 \pm 0.095 \)) and also that the staircase function appears to be a good choice as a time effect function. The learned values also suggest that none of the simple time effect functions’ shapes fit the staircase function, although the logarithmic and power functions might be close approximations—we analyze this in the following chapters.

![Figure 4.2: Time effect function as learned from the data with standard deviations. Each point represents average of 10 independent data sets.](image-url)
4. Evaluation and Analysis

4.5.2 Calibration

Calibration can be used to detect the deviation of model’s predictions from the averaged observed frequency of correctness of answers. This is very useful in our case where we need to find out how a model performs with each time effect function over timing distances between concrete practices of individual students. Note that a metric function (e.g. RMSE) is not a good choice in this case due to the fact that it considers all deviations simply as errors in predictions, i.e. it doesn’t show whether the model overestimates or underestimates the observed regularity of correct answers. The metric we used in order to calibrate the models is shown in Equation 4.1.

\[
Correctness - Prediction = \frac{1}{n} \sum_{i=0}^{n} (y_i - p_i) \quad (4.1)
\]

We compared the calibrations of models that do not take into account ages of previous practices with our models and visualized the results in Figure 4.3a. Our conclusion is that these models are not very well calibrated and analysis also suggests that the decay factor introduced in the PFA/G model does not solve this issue at all. We also analyzed several previously mentioned time effect functions which we used in the PFA/E/T and PFA/G/T models. Figure 4.3b indicates that this method leads to much better calibration of models.

Figure 4.3: Visualization of calibrations for (a) the baseline models and (b) their extended versions with selected time effect functions. We learned the parameters of the staircase function from data as was described in chapter 3.4.1, both power functions were estimated using the hill climbing method.
4. Evaluation and Analysis

Table 4.1: Parameters of models used in calibration analysis. Note that the PFA/G/T model requires the estimation of parameters of both time effect functions $f_{\text{succ}}$ and $f_{\text{fail}}$.

<table>
<thead>
<tr>
<th>Time Effect Function</th>
<th>$f_{\text{log}}$</th>
<th>$f_{\text{exp}}$</th>
<th>$f_{\text{pow}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$a$</td>
<td>$c$</td>
<td>$a$</td>
</tr>
<tr>
<td>PFA/E/T</td>
<td>1.802</td>
<td>0.119</td>
<td>1.642</td>
</tr>
<tr>
<td>PFA/G/T</td>
<td>$f_{\text{succ}}$</td>
<td>1.637</td>
<td>0.118</td>
</tr>
<tr>
<td></td>
<td>$f_{\text{fail}}$</td>
<td>0.94</td>
<td>0.107</td>
</tr>
</tbody>
</table>

The Figure 4.3a also indicates that the models with no time effect function are least calibrated especially in cases where the time from previous attempt is bigger.

Figure 4.4 shows calibration of all three time effect functions used in our experiments. The parameters of both models are summarized in Table 4.1. The analysis also suggests that the exponential function is not a very good choice as a time effect function for both models while the most calibrated models can be achieved with the power function.

Figure 4.4: Calibration analysis for (a) the PFA/E/T model and (b) the PFA/G/T model with three time effect functions used in our experiments.
4.6 Evaluation

We evaluated the models on several different data sets, including answers of all students, African and European countries, USA states, Czech rivers, world mountains and lakes. Parameters of all PFA-based models were estimated by running either the hill climbing method or the adjusted gradient descent. In all our experiments we used Elo model for the estimation of prior knowledge, where we estimated the parameters $\alpha$ and $\beta$ using grid search algorithm. On account of the quantification of model’s performance, we use several previously mentioned metrics, namely RMSE, LL, AUC and Accuracy.

4.6.1 All Answers

We compare the performance of all evaluated variations of examined models on a data set containing 150 thousand answers. The results are summarized in Table 4.2 where we report on the estimated parameters and the scores of each model as quantified by several metric functions. The best metric scores overall are marked bold. In Table 4.3 we summarize the standard deviations of the estimated parameters from 10 independent data sets.

It is clear that the models based on PFA/E perform best, even though the improvement is not huge. The performance of the PFA/E/T model with all compared time effect functions is very similar, however, the staircase offers the best performance with RMSE 3.402. Both the PFA/E model and the PFA/G model outperform our PFA/G/T model which uses a time effect function for penalizing old trials, nevertheless the difference is very small and the PFA/G/T model still shows significant improvement over the standard PFA model. Compared with the PFA/E model, there is also a small increase in performance of the PFA/E/RT model. Another interesting and perhaps unexpected result is the estimated value of the parameter $\xi$ (the decay factor) in the PFA/G model, i.e. $\xi = 0.425$, which penalizes the past trials to such an extent that if three new trials were performed subsequently, the weight of the old trial would have been multiplied by $0.425^3 = 0.077$. Also notice that the estimated value of $\delta$ is very close to 0.

The improvement of models which account for ages of past trials may be more visible when we evaluate the performance on items where the last trial was practiced at more distant time. Table 4.4 shows performance of all models evaluated only on answers where the last attempt is older than 6 hours. The results show that the PFA/G/T model outperforms the PFA/G model in cases where the last trials are older (e.g. older than 6 hours).
Table 4.2: Performance of all variations of models focused on timing information of students’ answers. The upper part of the table contains estimated parameters of each model. The lower part contains metric scores of each model.

<table>
<thead>
<tr>
<th>Model</th>
<th>Time Effect Fun.</th>
<th>Parameters</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\gamma$</td>
<td>$\delta$</td>
<td>$\xi$</td>
<td>$b$</td>
</tr>
<tr>
<td>PFA</td>
<td></td>
<td>1.022</td>
<td>−0.081</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PFA/E</td>
<td></td>
<td>2.614</td>
<td>−0.642</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PFA/G</td>
<td></td>
<td>1.798</td>
<td>0.091</td>
<td>0.425</td>
<td></td>
</tr>
<tr>
<td>PFA/E/RT</td>
<td></td>
<td>1.453</td>
<td>−1.356</td>
<td>1.886</td>
<td></td>
</tr>
<tr>
<td>PFA/E/T</td>
<td></td>
<td>$f_{\text{pow}}$</td>
<td>2.004</td>
<td>−0.713</td>
<td>2.931</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$f_{\text{log}}$</td>
<td>1.906</td>
<td>−0.806</td>
<td>1.789</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$f_{\text{exp}}$</td>
<td>2.006</td>
<td>−0.757</td>
<td>1.005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$f_{\text{staircase}}$</td>
<td>1.814</td>
<td>−0.827</td>
<td></td>
</tr>
<tr>
<td>PFA/G/T</td>
<td></td>
<td>$f_{\text{pow}}$</td>
<td>3.138</td>
<td>0.198</td>
<td>5.068</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$f_{\text{log}}$</td>
<td>1.669</td>
<td>0.102</td>
<td>0.914</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$f_{\text{exp}}$</td>
<td>0.96</td>
<td>0.002</td>
<td>0.508</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Time Effect Fun.</th>
<th>RMSE</th>
<th>AUC</th>
<th>Acc</th>
<th>LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFA</td>
<td></td>
<td>0.3651</td>
<td>0.753</td>
<td>0.822</td>
<td>−75 215</td>
</tr>
<tr>
<td>PFA/E</td>
<td></td>
<td>0.3456</td>
<td>0.8</td>
<td>0.836</td>
<td>−55 503</td>
</tr>
<tr>
<td>PFA/G</td>
<td></td>
<td>0.35</td>
<td>0.782</td>
<td>0.836</td>
<td>−58 420</td>
</tr>
<tr>
<td>PFA/E/RT</td>
<td></td>
<td>0.3449</td>
<td>0.799</td>
<td>0.836</td>
<td>−55 436</td>
</tr>
<tr>
<td>PFA/E/T</td>
<td></td>
<td>$f_{\text{pow}}$</td>
<td>0.3406</td>
<td>0.806</td>
<td><strong>0.841</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$f_{\text{log}}$</td>
<td>0.3407</td>
<td>0.806</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$f_{\text{exp}}$</td>
<td>0.341</td>
<td><strong>0.808</strong></td>
<td><strong>0.841</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$f_{\text{staircase}}$</td>
<td><strong>0.3402</strong></td>
<td><strong>0.807</strong></td>
<td><strong>0.841</strong></td>
</tr>
<tr>
<td>PFA/G/T</td>
<td></td>
<td>$f_{\text{pow}}$</td>
<td>0.3493</td>
<td>0.783</td>
<td>0.834</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$f_{\text{log}}$</td>
<td>0.3516</td>
<td>0.775</td>
<td>0.834</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$f_{\text{exp}}$</td>
<td>0.3508</td>
<td>0.775</td>
<td>0.834</td>
</tr>
</tbody>
</table>

$^1$ The parameters of the staircase function are depicted in Figure 4.2.
4. Evaluation and Analysis

Table 4.3: Standard deviations of estimated parameters.

<table>
<thead>
<tr>
<th>Model</th>
<th>Time Effect Fun.</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>γ</td>
<td>δ</td>
</tr>
<tr>
<td>PFA</td>
<td>0.163</td>
<td>0.02</td>
</tr>
<tr>
<td>PFA/E</td>
<td>0.099</td>
<td>0.059</td>
</tr>
<tr>
<td>PFA/G</td>
<td>0.259</td>
<td>0.080</td>
</tr>
<tr>
<td>PFA/E/RT</td>
<td>0.18</td>
<td>0.111</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Time Effect Fun.</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>γ</td>
<td>δ</td>
</tr>
<tr>
<td>PFA/E/T</td>
<td>f_{pow}</td>
<td>0.157</td>
</tr>
<tr>
<td></td>
<td>f_{log}</td>
<td>0.129</td>
</tr>
<tr>
<td></td>
<td>f_{exp}</td>
<td>0.142</td>
</tr>
<tr>
<td></td>
<td>f_{staircase}</td>
<td>0.276</td>
</tr>
<tr>
<td>PFA/G/T</td>
<td>f_{pow}</td>
<td>0.348</td>
</tr>
<tr>
<td></td>
<td>f_{log}</td>
<td>0.029</td>
</tr>
<tr>
<td></td>
<td>f_{exp}</td>
<td>0.104</td>
</tr>
</tbody>
</table>

1 Standard deviations of the staircase function are depicted in Figure 4.2.

Table 4.4: Performance of models when we consider only answers where the last trial is older than 6 hours.

<table>
<thead>
<tr>
<th>Model</th>
<th>Time Effect Fun.</th>
<th>RMSE</th>
<th>AUC</th>
<th>Acc</th>
<th>LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFA</td>
<td>0.3665</td>
<td>0.727</td>
<td>0.831</td>
<td>-16120</td>
<td></td>
</tr>
<tr>
<td>PFA/E</td>
<td>0.3491</td>
<td>0.752</td>
<td>0.84</td>
<td>-10575</td>
<td></td>
</tr>
<tr>
<td>PFA/G</td>
<td>0.357</td>
<td>0.754</td>
<td>0.84</td>
<td>-11783</td>
<td></td>
</tr>
<tr>
<td>PFA/E/RT</td>
<td>0.3487</td>
<td>0.768</td>
<td>0.84</td>
<td>-11308</td>
<td></td>
</tr>
<tr>
<td>PFA/E/T</td>
<td>f_{pow}</td>
<td>0.3417</td>
<td>0.777</td>
<td>0.842</td>
<td>-10056</td>
</tr>
<tr>
<td></td>
<td>f_{log}</td>
<td>0.342</td>
<td>0.779</td>
<td>0.84</td>
<td>-10061</td>
</tr>
<tr>
<td></td>
<td>f_{exp}</td>
<td>0.3433</td>
<td>0.776</td>
<td>0.84</td>
<td>-10142</td>
</tr>
<tr>
<td></td>
<td>f_{staircase}</td>
<td>0.3418</td>
<td>0.778</td>
<td>0.842</td>
<td>-10045</td>
</tr>
<tr>
<td>PFA/G/T</td>
<td>f_{pow}</td>
<td>0.3516</td>
<td>0.755</td>
<td>0.84</td>
<td>-10677</td>
</tr>
<tr>
<td></td>
<td>f_{log}</td>
<td>0.3524</td>
<td>0.771</td>
<td>0.841</td>
<td>-10924</td>
</tr>
<tr>
<td></td>
<td>f_{exp}</td>
<td>0.3521</td>
<td>0.75</td>
<td>0.841</td>
<td>-10939</td>
</tr>
</tbody>
</table>

40
4. Evaluation and Analysis

4.6.2 Countries and USA States

Most answers in our data set are from the citizens of the Czech Republic, i.e. Europeans, who generally have much better knowledge of the countries in Europe than for example the countries in Africa. Very important aspect of learning the locations of individual places is their context. It is easier to encode and retain the locations that are neighbors with an already known place.

![Graph showing learned values of the staircase function for African countries, European countries and USA states.]

**Figure 4.5:** Learned values of the staircase function for African countries, European countries and USA states.

In Figure 4.5, we observe the differences between learned values of the staircase function when the data set is separated in three by the type of place—African countries, European countries and USA states. The learned values suggest that the rate of forgetting (or more precisely the change in memory activation) is very similar in all examined types of places when the timing distance between current and the last practice is lower than one day, however, the African countries and USA states are forgotten faster after longer periods of time.

An overview of the estimated parameters of models and their scores is included in Table 4.5. Note that for both model that consider ages of trials we report only on the results of the best time effect function. The parameters of the staircase function are pictured in Figure 4.5.
### 4. Evaluation and Analysis

<table>
<thead>
<tr>
<th>Model</th>
<th>( \gamma/a_s )</th>
<th>( \delta/c_s )</th>
<th>( \xi/a_f )</th>
<th>( b/c_f )</th>
<th>RMSE</th>
<th>LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFA</td>
<td>1.172</td>
<td>-0.169</td>
<td></td>
<td></td>
<td>0.3686</td>
<td>-49.706</td>
</tr>
<tr>
<td>PFA/E</td>
<td>2.134</td>
<td>-0.24</td>
<td></td>
<td></td>
<td>0.3483</td>
<td>-38.531</td>
</tr>
<tr>
<td>PFA/G</td>
<td>1.705</td>
<td>0.191</td>
<td>0.507</td>
<td></td>
<td>0.3486</td>
<td>-39.958</td>
</tr>
<tr>
<td>PFA/E/RT</td>
<td>1.009</td>
<td>-1.236</td>
<td>2.037</td>
<td></td>
<td>0.3465</td>
<td>-38.457</td>
</tr>
<tr>
<td>PFA/E/T ( f_{\text{staircase}} )</td>
<td>1.889</td>
<td>-0.777</td>
<td></td>
<td></td>
<td>0.3425</td>
<td>-37.286</td>
</tr>
<tr>
<td>PFA/G/T ( f_{\text{pow}} )</td>
<td>3.287</td>
<td>0.191</td>
<td>3.7</td>
<td>0.476</td>
<td>0.3524</td>
<td>-43.725</td>
</tr>
</tbody>
</table>

- **(a) European Countries**

<table>
<thead>
<tr>
<th>Model</th>
<th>( \gamma/a_s )</th>
<th>( \delta/c_s )</th>
<th>( \xi/a_f )</th>
<th>( b/c_f )</th>
<th>RMSE</th>
<th>LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFA</td>
<td>0.837</td>
<td>-0.164</td>
<td></td>
<td></td>
<td>0.3842</td>
<td>-33.364</td>
</tr>
<tr>
<td>PFA/E</td>
<td>2.235</td>
<td>-0.684</td>
<td></td>
<td></td>
<td>0.3608</td>
<td>-26.866</td>
</tr>
<tr>
<td>PFA/G</td>
<td>1.619</td>
<td>0.021</td>
<td>0.433</td>
<td></td>
<td>0.3651</td>
<td>-28.828</td>
</tr>
<tr>
<td>PFA/E/RT</td>
<td>1.042</td>
<td>-0.943</td>
<td>2.017</td>
<td></td>
<td>0.3597</td>
<td>-26.512</td>
</tr>
<tr>
<td>PFA/E/T ( f_{\text{staircase}} )</td>
<td>2.121</td>
<td>-0.913</td>
<td></td>
<td></td>
<td>0.356</td>
<td>-26.144</td>
</tr>
<tr>
<td>PFA/G/T ( f_{\text{pow}} )</td>
<td>3.197</td>
<td>0.227</td>
<td>4.588</td>
<td>0.745</td>
<td>0.3675</td>
<td>-43.725</td>
</tr>
</tbody>
</table>

- **(b) African Countries**

<table>
<thead>
<tr>
<th>Model</th>
<th>( \gamma/a_s )</th>
<th>( \delta/c_s )</th>
<th>( \xi/a_f )</th>
<th>( b/c_f )</th>
<th>RMSE</th>
<th>LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFA</td>
<td>1.068</td>
<td>-0.14</td>
<td></td>
<td></td>
<td>0.3917</td>
<td>-35.440</td>
</tr>
<tr>
<td>PFA/E</td>
<td>2.816</td>
<td>-0.568</td>
<td></td>
<td></td>
<td>0.3643</td>
<td>-25.677</td>
</tr>
<tr>
<td>PFA/G</td>
<td>1.809</td>
<td>0.084</td>
<td>0.443</td>
<td></td>
<td>0.3683</td>
<td>-27.083</td>
</tr>
<tr>
<td>PFA/E/RT</td>
<td>1.251</td>
<td>-1.02</td>
<td>1.663</td>
<td></td>
<td>0.3704</td>
<td>-28.577</td>
</tr>
<tr>
<td>PFA/E/T ( f_{\text{staircase}} )</td>
<td>2.144</td>
<td>-0.842</td>
<td></td>
<td></td>
<td>0.3588</td>
<td>-24.864</td>
</tr>
<tr>
<td>PFA/G/T ( f_{\text{pow}} )</td>
<td>3.285</td>
<td>0.205</td>
<td>4.982</td>
<td>0.623</td>
<td>0.3698</td>
<td>-28.368</td>
</tr>
</tbody>
</table>

- **(c) USA States**

1 Parameters of staircase functions are depicted in Figure 4.5.
4. Evaluation and Analysis

4.6.3 Mountains, Rivers and Lakes

Another interesting question is how the number of facts affects difficulty of learning. Intuitively, it’s easier to learn all facts from domains of small sizes. We divided the data set into three, one containing only the answers to questions about either 21 lakes, 71 mountains or 84 rivers. The result of learned parameters of $\gamma$, $\delta$ and the staircase function is depicted in Figure 4.6.

![Figure 4.6: Learned values of the staircase function from data sets containing only rivers, lakes and mountains.](image)

Estimated parameters and scores of models for each data set are summarized in Table 4.7. Again, we report only on performance of the best time effect functions.

4.7 Further Analysis

In this section we perform further analysis of our models and their parameters. Although the interpretation is debatable, we believe that it shows useful results and opens some interesting questions.

4.7.1 Parameters vs Types of Places

The parameters $\gamma$ and $\delta$ in the PFA model affect penalty for each successful or failed answer. Higher values of $\gamma$ cause each correct answer increase the memory activation more, this usually means that the material is easier to
4. Evaluation and Analysis

Table 4.7: Mountains, rivers and lakes.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\gamma/a_s$</th>
<th>$\delta/c_s$</th>
<th>$\xi/a_f$</th>
<th>$b/c_f$</th>
<th>RMSE</th>
<th>LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFA</td>
<td>1.517</td>
<td>−0.195</td>
<td>0.4092</td>
<td>−73694</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PFA/E</td>
<td>2.609</td>
<td>−0.153</td>
<td>0.38</td>
<td>−47170</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PFA/G</td>
<td>1.86</td>
<td>0.023</td>
<td>0.3845</td>
<td>−50530</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PFA/E/RT</td>
<td>1.399</td>
<td>−0.95</td>
<td>1.997</td>
<td>0.3827</td>
<td>−50209</td>
<td></td>
</tr>
<tr>
<td>PFA/E/T</td>
<td>$f_{\text{staircase}}$ 1</td>
<td>2.108</td>
<td>−0.709</td>
<td>0.3743</td>
<td>−45824</td>
<td></td>
</tr>
<tr>
<td>PFA/G/T</td>
<td>$f_{\text{pow}}$</td>
<td>3.203</td>
<td>0.188</td>
<td>4.905</td>
<td>0.523</td>
<td>0.3935</td>
</tr>
</tbody>
</table>

(a) Mountains

<table>
<thead>
<tr>
<th>Model</th>
<th>$\gamma/a_s$</th>
<th>$\delta/c_s$</th>
<th>$\xi/a_f$</th>
<th>$b/c_f$</th>
<th>RMSE</th>
<th>LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFA</td>
<td>1.545</td>
<td>−0.21</td>
<td>0.3768</td>
<td>−66675</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PFA/E</td>
<td>2.92</td>
<td>−0.759</td>
<td>0.352</td>
<td>−43512</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PFA/G</td>
<td>2.024</td>
<td>0.066</td>
<td>0.3569</td>
<td>−47487</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PFA/E/RT</td>
<td>1.299</td>
<td>−0.951</td>
<td>1.916</td>
<td>0.3522</td>
<td>−44757</td>
<td></td>
</tr>
<tr>
<td>PFA/E/T</td>
<td>$f_{\text{staircase}}$ 1</td>
<td>2.554</td>
<td>−0.766</td>
<td>0.3476</td>
<td>−32161</td>
<td></td>
</tr>
<tr>
<td>PFA/G/T</td>
<td>$f_{\text{pow}}$</td>
<td>3.196</td>
<td>0.21</td>
<td>4.705</td>
<td>0.695</td>
<td>0.359</td>
</tr>
</tbody>
</table>

(b) Rivers

<table>
<thead>
<tr>
<th>Model</th>
<th>$\gamma/a_s$</th>
<th>$\delta/c_s$</th>
<th>$\xi/a_f$</th>
<th>$b/c_f$</th>
<th>RMSE</th>
<th>LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFA</td>
<td>2.178</td>
<td>−0.288</td>
<td>0.4102</td>
<td>−5795</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PFA/E</td>
<td>2.984</td>
<td>−0.057</td>
<td>0.3778</td>
<td>−4073</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PFA/G</td>
<td>1.846</td>
<td>0.096</td>
<td>0.38</td>
<td>−4263</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PFA/E/RT</td>
<td>1.871</td>
<td>−0.85</td>
<td>2.179</td>
<td>0.3757</td>
<td>−4070</td>
<td></td>
</tr>
<tr>
<td>PFA/E/T</td>
<td>$f_{\text{staircase}}$ 1</td>
<td>2.699</td>
<td>−0.618</td>
<td>0.3681</td>
<td>−3900</td>
<td></td>
</tr>
<tr>
<td>PFA/G/T</td>
<td>$f_{\text{exp}}$</td>
<td>1.2</td>
<td>0.002</td>
<td>1.496</td>
<td>0.012</td>
<td>0.3815</td>
</tr>
</tbody>
</table>

(c) Lakes

1 Parameters of staircase functions are depicted in Figure 4.5.
learn. On the other hand, the parameter $\delta$ influences the weight of each incorrect answer. In order to observe the relationship between the type of place and the parameters $\gamma$ and $\delta$, we divided the data set with answers of students into 10 bins each containing one type of places.

The size and the number inside each circle represents counts of unique places of the type.

Figure 4.7: Estimated parameters of the PFA/E model vs the type of place. The learned parameters are depicted in Figure 4.7. Although the interpretation is debatable, we expect the easier places to be located in the upper right corner whereas the hardest places to be located in the lower left corner. Probably one of the easiest questions students practice concern Czech lakes (note that the values of $\gamma$ and $\delta$ seem to correlate with the number of places).

4.7.2 At School vs At Home

An interesting question is how much students’ motivation affects the speed of learning or the rate of forgetting. Figure 4.8 shows the learned parameters of a staircase function on two different data sets. The first data set labeled as
4. Evaluation and Analysis

“At School” contained only the answers from groups of 10 or more students with the same IP address (the assumption is that these students practiced geography in classrooms). The second data set labeled as “At Home” contained all the other answers of students.

![Graph](image)

**Figure 4.8:** Comparison of students who practice geography at school (i.e. in classrooms) and students who practice at home (i.e. on their own initiative).

Again, the interpretation is open to discussion. One theory could suggest the following explanation—the learned values of the staircase function and the parameters $\gamma$, $\delta$ indicate that students who practice geography at school learn slower and forget the learned material faster then the students who practice geography at home and are more willing to study. Another explanation could propose that the difference in the rate of forgetting between both types of students is caused by teaching methods, e.g. due to more material being learned in shorter periods of time. Since the system we analyze is adaptive and selects questions based on student’s knowledge, the shape of the staircase function and the learned parameters could be caused by combination of both.
5 Discussion

In this chapter, we review results of our experiments and discuss applications of the analyzed models in other systems. We also give tips on how to perhaps further improve performance of our models and which aspects of human memory could be considered in future research.

5.1 Results

In chapter 4, we analyzed several models based on timing information (taking into account either ages of past trials of items or past response times of students). Most promising results in the context of learning geography facts shows the PFA/E/T model with a staircase time effect function, although the power function proves to be almost as equally a good choice. Main advantage of the staircase function is that the parameters are easy to learn for the purpose of online learning, i.e. there is no need to periodically run parameter optimization procedure since the parameters are slightly adjusted with each new answer.

The PFA/G/T model suffers from the inability to account for multiple-choice questions. It might be interesting to analyze the models on data from adaptive systems where questions with multiple-choice options are not used—this is not possible using the data set described in section 4.1 since the system adaptively selects number of options in multiple choice tests, i.e. filtering out all multiple-choice questions creates data set with skewed difficulties and abilities of students.

The PFA/E/RT model that considers past response times of students shows in most cases only a very small improvement. However, very interesting results provides the PFA/G model with values of the decay factor lower then 0.5, this may indicate that the past trials of items do not actually influence model’s performance in the context of learning geography facts as much as we thought. When we compare the PFA/G model and the PFA/G/T model, it seems that penalizing old trials based on order of answers leads in most cases to a better performance than the penalty of trials based on timing information.

5.2 Applications and Recommendations

Currently, models used in the adaptive system Outline Maps do not take into account ages of past trials. We have demonstrated that the timing
5. Discussion

formation improves model’s performance and thus, the extensions of presented models should be considered in production. As we already discussed in previous section, the best choice is the PFA/E/T model with a time effect function fitted by gradient descent described in chapter 3.3.4.

Another application of our models could be considered in the adaptive system for practicing medical anatomy (practiceanatomy.com) targeted mostly at medical students. The system allows the practice of various organ systems, body parts, bones, muscles, etc. The evaluation of our models on data from students who practice anatomy could also open interesting questions about difficulty of items and the effect of forgetting where most students are collage students.

5.3 Possible Extensions and Alternative Models

The examined models approximate the effect of human memory decay with a time effect function which ignores some very important aspects of memory and forgetting. For example, in our models the spacing effect is completely disregarded. One extension of the PFA/G/T model could reuse some properties of the memory activation equation 2.18 described in chapter 2.4.4. Next, an extension of the PFA/G/T model which considers multiple-choice questions might also be possible and could significantly improve its performance.

Another extension of the models could consider the fact that students mix up the places that are, for instance, geographically very close, have similar names, or seem somehow similar for entirely different reasons. Students typically know the approximate geographical location of the Nordic countries, however, they might struggle when determining exact location of each country, often they confuse Norway with Sweden as can be seen in Figure 5.1. If a students chooses Finland when the highlighted place is actually Norway, it usually does not mean that they have no knowledge of the country’s location. In such cases it might be more accurate to slightly increase student’s memory activation of the place.

A completely different approach on student modeling is the usage of Recurrent Neural Networks (RNNs), the family of RNN model have important advantage over models analyzed in our thesis—they do not require explicit encoding of the essential aspects of human memory and can capture more complex patterns of learning. The usage of RNNs in education, especially the Long Short Term Memory (LSTM) model, has been already explored and shows promising results [35]. It might be interesting to analyze the RNN model in the context of the system for learning geography and also compare the performance with PFA models.
Figure 5.1: Confusion network of European countries. The edges between countries represent confusion of students. Dark blue edge means that the two connected countries are confused by students quite often while light blue indicates that the countries are still sometimes confused but at least less often.
6 Conclusion

In the first part of the thesis, we summarized the key aspects of human memory—learning and forgetting. We discussed applications and properties of adaptive educational systems, basics of student modeling and the most relevant models convenient in the context of our thesis. We described the Elo model successfully used for estimation of prior knowledge of students and mainly the Performance Factor Analysis (PFA) used for estimation of the current knowledge students acquire while interacting with an adaptive system. We also studied some extensions of the PFA model and the models used in the ACT-R modeling system.

The primary objective of our thesis was to explore relevant models and design their extensions which take into account the key aspects of human memory (mainly forgetting). We described several models based on previous extensions and introduced a time effect function which penalizes the age of student’s past attempts. Next, we summarized machine learning techniques used for parameter estimation and quantification of model’s performance. The evaluation of models was performed using data from the adaptive practice system Outline Maps. Finally, we analyzed students’ response times for the possibility that it may indicate level of student’s knowledge, and the values of model’s parameters depending on the domain of the used data set—the type of place, or the purpose of practice.

Our experiments demonstrated that both student models which take into account ages of past trials results in a better performance. Even though the knowledge of a country leads to a faster response time, our model showed minor improvements when the duration of past response times was considered. The analysis of the model with a decay factor shows that penalizing old trials based on order of answers leads in most cases to a better performance than the penalty of trials based on timing information.

We suggested the usage of the extended PFA model with a staircase function in production, this method is computationally very efficient and does not require periodic estimations of parameters. Other models do not perform so well especially in environments where the test of student’s knowledge involves answer to a multiple-choice question. Future research may focus on analysis of models in other adaptive practice systems (e.g. practiceanatomy.com), where more complex patterns of human memory (e.g. the spacing effect) might be more apparent.
Bibliography


BIBLIOGRAPHY


A Links

The source code with our experiments, implementations of several student models, optimization algorithms and other miscellaneous helpers is available on GitHub: https://github.com/paveldedik/thesis.