Models for Algorithmic Teaching

Frank Balbach

Dissertation

Universität zu Lübeck
Institut für Theoretische Informatik

Aus dem Institut für Theoretische Informatik
der Universität zu Lübeck
Direktor: Prof. Dr. Rüdiger Reischuk

Models for Algorithmic Teaching

Inauguraldissertation

zur Erlangung der Doktorwürde
der Universität zu Lübeck
aus der Technisch-Naturwissenschaftlichen Fakultät

Vorgelegt von

Frank Balbach

Lübeck, November 2006

Vorsitzender des Prüfungsausschusses:	Prof. Dr.-Ing. Achim Schweikard
1. Berichterstatter:	Prof. Dr. math. Rüdiger Reischuk
2. Berichterstatter:	Prof. Dr. rer. nat. Thomas Zeugmann

Tag der mündlichen Prüfung:	23. 02. 2007

Zum Druck genehmigt:	23. 02. 2007

gezeichnet Prof. Dr. rer. nat. Enno Hartmann
Dekan der Technisch-Naturwissenschaftlichen Fakultät der Universität zu Lübeck

Abstract

Learning theory focuses almost entirely on the learner and its efficient realization, but neglects other parts of the learning process, most importantly the teacher, which is merely modeled as a passive data source. In this thesis, however, we study models in which the teacher plays the central, active role and which allow us to investigate teaching algorithms. In practice, teaching algorithms occur in the shape of *intelligent tutoring systems*, computer based interactive systems for teaching students or at least aiding their learning process. But in learning theory so far all teaching models fail to properly describe the intelligent tutoring system scenario. We develop and analyze new teaching models that improve the current ones with respect to this scenario.

The most common teaching model at present is based on the notion of teaching dimension. This dimension specifies the minimum number of argument-value pairs ("examples") needed to describe a given Boolean function ("concept") among a given class of functions. A set of examples that uniquely describes a concept within a class is called a teaching set. One assumption in the teaching dimension model is that all learners are consistent, that is, their hypothesis always matches all known examples. The teaching dimension of a concept then is the minimal number of examples a teacher has to give in order to make all consistent learning algorithms hypothesize that concept. The teaching dimension thus describes the optimal performance of a teacher for a given target concept and therefore in some sense the teachability of the concept. The average teaching dimension over all concepts in a concept class is considered a measure for the teachability of this class, but this value is not known for many natural classes. We show that the classes of monomials and 1-decision lists over n variables both have an average teaching dimension of $O(n)$.

As we demonstrate, a straightforward attempt to analyze intelligent tutoring systems in the teaching dimension model fails, because the model does not capture many real-life aspects of teaching. Teachers cannot benefit from arranging the subject matter suitably; the optimal way of teaching is independent of such crucial properties of the learner as its memory size; and the teacher cannot exploit feedback given by the students. Moreover, the teaching dimension proves to be a counterintuitive measure of teachability. For instance, longer 1-decision lists often have a much smaller teaching dimension than shorter 1-decision lists.

In this thesis we identify two reasons for this lack of realism. First, the underlying model of the student is too simple. Second, the performance of the teacher is measured with respect to the worst student, rather than all students.

We present two approaches to tackle these shortcomings of the teaching dimension model. In the first approach we develop a general, modular framework for algorithmic teaching that allows the specification of various student models and various kinds of students' feedback. Within this framework, a student is modeled as an algorithm that maintains a hypothesis and changes it according to the examples received from the teacher. We then use this general framework to compare concrete student models. Students differ with regard to the way they change the hypothesis and also what examples they memorize and for how long.

In the simplest model we investigate, the students can have a preference for certain hypotheses. We show that if students prefer simple hypotheses over complex ones, 1-decision lists become easier to teach the shorter they are. We also consider students with different hypothesis preferences and show that the teachabilities in all such models can be described by a dimension-like parameter, similar to the teaching dimension. From this it follows that the teacher does not need to pay attention to the order of examples or to the feedback.

Next, we investigate a model of students in which they develop their hypothesis in a more restricted way than in the teaching dimension model. We show that then the teacher must take care to arrange the material in the right order and that the teacher can benefit greatly from receiving feedback. We also show that optimal teachers in this model are harder to find than in the teaching dimension model. The complexity of the corresponding decision problem increases from polynomial time to $\mathcal{NP}$-complete for memoryless learners and from $\mathcal{NP}$-complete to $\mathcal{PSPACE}$-complete for learners with perfect memory.

In the second approach we modify the teaching dimension model so as to analyze the average student instead of the worst student. This is done by introducing a randomized learning algorithm that incorporates all allowed behaviors. We show that to optimize the performance for the average student, the teacher has to pay attention to the order of the material and also to the students' feedback. Furthermore, the performance of a teacher varies with the students' memory size.

Using the theory of Markov decision processes, we characterize optimal teachers for several variants of randomized learners. We then focus on the randomized learner that does not memorize examples and also provides no feedback to the teacher. There is no known algorithm to compute the performance of the optimal teacher in this case; we show this problem to be $\mathcal{NP}$-hard and devise an approximation algorithm for it.

The performance of arbitrary teachers for these learners is hard to calculate, and it is undecidable whether a teacher is successful at all. We identify a class of teachers that can be handled more easily and whose performance can come arbitrarily close to optimal. Finally we show that in general the teacher that chooses the next example greedily does not approximate the optimal performance by a constant factor.

Acknowledgment

I am much obliged to Rüdiger Reischuk for providing excellent working conditions in an inspiring environment and even more for his continuous support and advice throughout the years.

I thank Thomas Zeugmann for introducing me to the learning theory community and also for encouraging me to pursue my own research ideas, yet patiently providing guidance whenever I needed it.

I am grateful to Till Tantau for taking a great interest in my writing, and finishing, this thesis and for providing some vital tips.

I am indebted to all my current and former colleagues at the Institute for Theoretical Computer Science of the University Lübeck. Every single one of them has, in one way or another, contributed to my evolution as computer scientist.

Contents

In addition, however, it is to be wished that the method of human education be mechanical.

(Comenius [25])

Chapter 1

Introduction

1.1 Motivation

All higher life forms on earth learn by imitation, observation, or (the smarter ones) by experimentation. Humans have invented two further mechanisms by which they gain knowledge and skills: teachers and textbooks. Both methods have a long tradition and are opposed to one another in many respects. Teachers are expensive to create and to maintain, books are relatively cheap. Teachers are available only at certain hours at certain places, books can be read anytime anywhere. Teachers can respond to the specific needs of their students, books present their content in the same way for every reader. Teachers can provide their students with hands-on experience in the subject (indispensable, for example, in language teaching), books place the reader in a more passive role.

With the technological advance, a third alternative besides teachers and textbooks becomes imaginable: computers. If nothing else, a computer can present information much in the same way as a book, with the usual additional features like hyperlinks, multimedia, or search capabilities. Beyond this, there are systems that interact with students, monitor their learning progress, and tailor the curriculum to them. Such *intelligent tutoring systems* (ITS) present a middle way between teachers and books. They are more expensive to create than books, but as they are easy to copy, the costs per student can be low. An intelligent tutoring system is available anytime, but requires a computer. It can, ideally, adapt to the student, but not as well as a human teacher can. It can provide hands-on experience in areas like basic arithmetics or learning to type, but not, for example, in athletics.

As a simple illustration, imagine a hypothetical ITS for teaching the school method of adding two numbers (cf. [59]). Such a system would present the student addition problems, such as $127 + 54$, and from the answer it would infer the student's skill and possible misconceptions. For example, the answer 171 suggests that the student has trouble with the carry; the answer 182 indicates trouble with adding certain digits.

The system could then provide the student with specific instruction. The basic idea of intelligent tutoring systems dates back to the 1970s. For the present state of the art we refer the reader to the proceedings of the annual *Intelligent Tutoring Systems* conference [42] as well as to the relevant AAAI website [43].

Apart from some technical details, the task of an intelligent tutoring system is similar to that of a human teacher or a book. The subject matter must be presented in such a way that all students or readers can understand it easily. In a classroom scenario, questions from the audience must be answered helpfully, and the presentation of the information should, in general, be guided by the feedback the audience gives. All teachers, be they human, machines, or books, must be prepared to deal with many different students of which they have little or no prior knowledge. The development of effective intelligent tutoring systems therefore requires knowledge of human learning behavior, or in other words assumptions about how students learn.

Human teachers and authors can draw on results of didactics or psychology. They can use known didactic principles or rules of thumb with a more or less informal image of students in their mind. In contrast, the development and verification of an intelligent tutoring system requires a *formal* model of the underlying teaching process. Our goal in this thesis is to design such a model in a rigorous way.

In the remainder of the introduction we first give an overview over the currently existing formal models for teaching and learning, from which it becomes clear that these models are insufficient for modeling intelligent tutoring systems (Section 1.2). Based thereon we then formulate the requirements for our aspired formal ITS model more precisely (Section 1.3). Afterwards we present our main contributions (Section 1.4) and finally give an outline of our results (Section 1.5).

1.2 Formal Models for Teaching and Learning

Two of the first steps in designing any model are to identify the parts of reality one wants to formalize and to find a suitable way of formalizing them. When Turing modeled a human performing calculations, he decided that writing things down on paper and reading them later are essential, but the color of the ink and the shape of the paper are not. The formalization was an infinite tape with cells, each able to hold a symbol, and a read-write head able to move left or right.

For our teaching model we do not have to start from scratch because teaching and learning are closely connected processes and we can resort to four decades of research in algorithmic learning theory. This theory, although primarily concerned with models for learning instead of teaching, has analyzed the common underlying process and identified its main components. The following abstract description of the process highlights these components (see also Figure 1.1).

The main actors are a *teacher* and a *learner*. The teacher has something in mind he

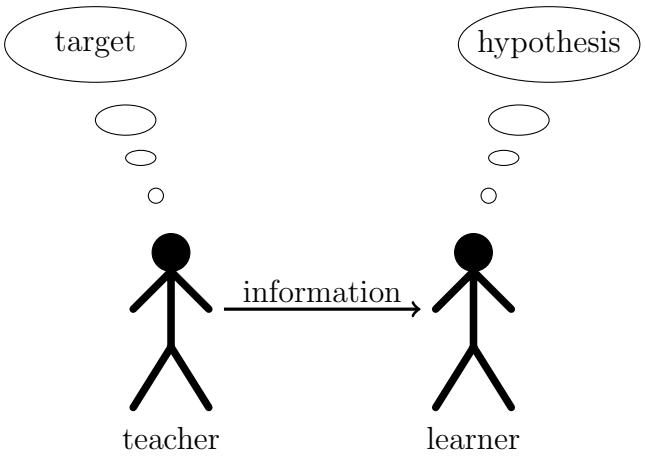

Figure 1.1 *The generic teaching/learning process. The figure visualizes the components of that process, except for the student's prior knowledge and the success criterion.*

wants to teach (the *target*). Note that here "teacher" is a generic term that can refer to any source of information about the target. The learner initially has some incomplete *prior knowledge* about the target. During the teaching process the teacher and the learner perform some sort of *interaction*, which results in *information* about the target flowing from the teacher to the learner. The learner combines this information with the prior knowledge in order to form *hypotheses*. These hypotheses represent the learner's guesses about what the target actually is. Roughly speaking, the process is deemed successful if and when the hypothesis eventually comes close enough to the target. The formal definition is given by a *success criterion*.

This description is general enough to span a wide range of real-world scenarios. For example, in a foreign language course, a teacher has in mind a language, say Spanish. The students may or may not have some prior knowledge about Spanish or a general background on languages, but they certainly have some "feeling" as to how languages work. This would be the prior knowledge. Teacher and learners interact in the usual classroom fashion, during which the teacher communicates Spanish vocabulary, phrases, or grammar rules to the students. The hypotheses represent the students' Spanish competence. As success criterion a certain Spanish language proficiency is required that can be assessed by an examination.

Another scenario that fits the description above is that of an autonomous robot exploring an unknown environment. The environment serves as a teacher. The robot, while moving around, receives information by observing the environment. The target could be, for instance, a complete map of the environment. Prior knowledge could be the ability of the robot to recognize common objects in the environment, like doors, walls, or trees. A hypothesis would then be a map of the region explored so far. Finally, the success is measured by the accuracy and the coverage of the hypothesized map.

The general description of the process also captures the scenario in which an author writes a textbook that is later read by students and the scenario of a scientist doing research. Imagine, for instance, a volcanologist observing volcanos and collecting lava

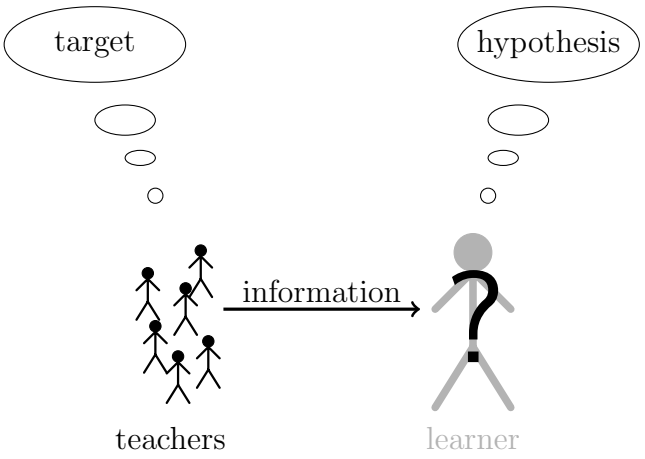

Figure 1.2 *In learning models a learner is sought that works well, regardless of the teacher's behavior.*

in order to learn how to predict eruptions.

The ITS scenario fits in as well. In our hypothetical ITS that teaches basic arithmetics the ITS plays the role of the teacher, the learner is the student using the system, and interaction works by the teacher providing addition problems and the learner supplying answers. Success occurs when the student can perform all additions correctly.

Many models have been developed that formalize the components of the abstract process. The models differ not only in the way they formalize the components but also in their purpose. The original purpose of models in learning theory was, naturally, the investigation of learning algorithms. Common to all these learning models is that their success criterion requires the learner to be successful for *all* teachers defined by the model, even for malicious ones. Such models are therefore suitable for scenarios like the autonomous robot or the volcanologist, who both have to work in whatever environment they are placed or using whatever lava specimen they get. Figure 1.2 is a visualization of this success criterion.

Learning models are primarily used to design learning algorithms, whereas our goal is to design algorithms for teaching. It is nevertheless worthwhile to review these models briefly because they introduce many concepts that we shall meet again later in dedicated teaching models. In particular, even learning models have to give some formal definition of teachers. Mostly, this definition is hidden implicitly in the specification of the information presentation and in the definition of the success criterion and must be made explicit.

The first learning model was introduced by Gold in 1967 [33] and is generally referred to as inductive inference. It is formulated in terms of recursive sets and functions and therefore the most fundamental model. The targets are recursive languages. There are two ways of formalizing the teacher. In one variant the teacher is an infinite series containing, in arbitrary order, every word over the underlying alphabet with a label stating whether or not the word is in the target language. In the other variant the teacher is a series containing only, and all, the words present in the target language.

In any case, the learner is a partial recursive function mapping initial portions of the teacher series to descriptions of languages. The process consists of the teacher giving the series element by element to the learner which in turn builds hypotheses accordingly. There are two fundamentally different definitions of success: *learning in the limit* and *finite learning*. The first variant requires the series of hypotheses to eventually stabilize on a correct description of the target. In contrast, the second success criterion, finite learning, requires the learner to output a single hypothesis that must be a correct description of the target. In both variants the learner must be able to reach a correct hypothesis no matter in which order the teacher presents the information.

The inductive inference model is not suited for finite languages. For instance, learning finite languages in the limit is trivial. Consequently, Boolean functions, which can be seen as characteristic functions of finite languages, have received less attention in the original inductive inference setting. Even for learning simple functions, like monomials, exponentially many argument-value pairs (called *examples*) are needed. In the more recent stochastic finite learning approach [65, 66, 73], however, one is able to study Boolean functions in a stochastic variant of learning in the limit, which yields polynomial teachability results for the class of monomials.

In 1984 Valiant [78] introduced a model that is applicable to finite as well as to infinite languages and focuses on learnability from polynomially many examples. The targets are regarded as 0-1-valued functions, called *concepts*, over arbitrary domains; typical domains are the Boolean domain, all words over an alphabet, or the real numbers. In this so called PAC learning model the teacher presents the data as a sequence of examples of the target concept. The examples are chosen according to a probability distribution that is unknown to the learner. After having received polynomially many examples, the learner has to stop and output a single hypothesis. In PAC learning the learner is required to be successful for every probability distribution according to which the teacher chooses the examples. Because the probability distribution can put very small probabilities on the interesting examples, the learner's hypothesis is allowed to be only approximately correct, and with some small probability the hypothesis may even be arbitrarily bad. The success criterion is thus probabilistic.

The online learning model, introduced by Littlestone [51], combines non-probabilistic learning with the possibility of studying finite concepts. In every round the teacher presents the learner an argument of the target function, but without the value. The learner guesses this value and is charged one mistake whenever this guess is wrong. At the end of every round the learner is told the right value. The learner's quality is measured by the number of mistakes it makes in the worst case over all teachers, that is, over all sequences of unlabeled examples for the target. The online learning model is primarily concerned with the question whether a concept class is learnable with only polynomially many prediction mistakes. It is thus closely related to the PAC model.

Angluin's query model [3, 4] changes the way the learner receives its information. Rather than relying on the teacher selecting examples, the learner can now ask questions, which the teacher answers truthfully. The quality of the learner is measured

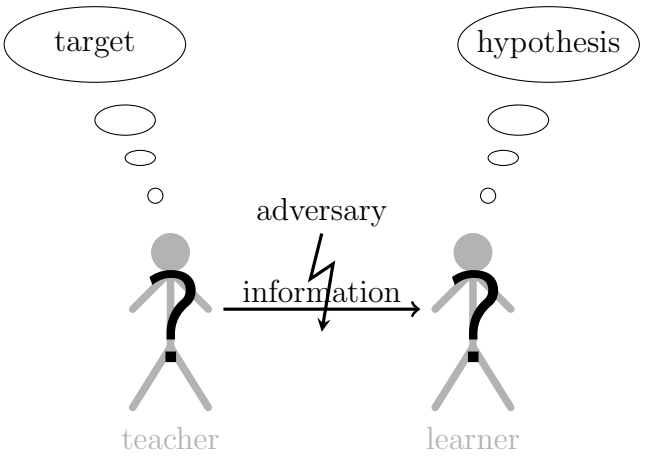

Figure 1.3 *In hybrid models a teacher and a learner are sought that work well together, regardless of the adversary's behavior.*

by the number of questions needed to identify the target. The teacher in the query learning model has the greatest similarity with a real world teacher among all the learning models. But still that teacher remains passive and does not aid the learner of its own accord. When, depending on the model variant, the teacher may choose from several answers, he is assumed to act adversarially. The burden of learning thus remains completely on the part of the learner.

We finish this review of learning models by remarking that all models differ with respect to which targets they deem teachable. There is thus no universal model describing learnability. This contrasts, for example, with the notion of computability, for which all models have turned out to be equivalent.

The learning models in the previous paragraphs provide plenty of formalism for the components of the teaching process, but they are all incapable of modeling the ITS scenario. Their basic assumption is that the teacher may behave badly, which is plausible in some scenarios like the volcanologist, who cannot expect volcanos to give particularly helpful examples, but which is not true in the ITS scenario.

In the ITS scenario the teacher chooses information in a helpful manner. In learning theory several models exist where, in addition to a learner, a teacher selecting helpful examples can be devised. A major problem in this kind of model is the prevention of cheating. For example, in an inductive inference setting, if the teacher wants to teach a recursive language he could outright encode the index of the target language into the example sequence, and the learner could decode it to hypothesize the target without doing "real" learning in any sense. A common remedy is to introduce a kind of adversary that disturbs the information flow in order to prevent coding tricks. Figure 1.3 provides a graphical representation of the basic structure of such models.

Collusion between teacher and learner can be prevented in several ways. Variants of the Gold style learning model, called learning from good examples, have been introduced by Freivalds, Kinber, and Wiehagen [29, 30] and by Lange, Nessel, and Wiehagen [48]. Here, the teacher selects a finite set of words of the target language and

gives it to the learner. The main difference to the original inductive inference models is that the learner is required to find a correct hypothesis not only from this given set of words, but also from all proper supersets thereof. This can be thought of as an adversary adding information to the information given by the teacher. This additional information is truthful, but nevertheless prevents collusion like outright encoding because the learner cannot distinguish words given by the teacher from those given by the adversary. Just like in the original inductive inference model, learning can be successful in the limit or in a finite way.

Similar models have been introduced by Goldman and Mathias [35] and by Mathias [53, 54] for learning concepts. Again, the main difference to the traditional models is an adversary that adds true information to the information given by the teacher.

Yet another model has been devised by Jackson and Tomkins [44]. Here, the teacher and learner interact in a modified prover-verifier session (see Goldwasser, Micali, and Rackoff [37]). The teacher can be designed arbitrarily and may even have complete knowledge about the learner. The learner, on the other hand, may not be designed arbitrarily, but must be designed in such a way that no adversarial teacher can make the learner hypothesize a wrong concept. This prevents the learner from exploiting cheating teachers.

Another way to avoid coding tricks between teacher and learner is to provide them with incompatible hypothesis spaces. Angluin and Kriķis [6, 7] propose a model for learning and teaching partial recursive functions in which the learner's task is to find an index of the target function in a partial recursive enumeration that can be accessed by the learner only like a black box.

The models in which a learner benefits from examples carefully selected by a teacher are closer to the ITS scenario in that they allow the design and analysis of teaching algorithms. But they do not yet fit exactly because they also still require the design of a learner. An intelligent tutoring system, however, is supposed to deal with many different students, not just with a single, specially defined one. The model we are looking for should provide a definition of learners in a symmetric way as the learning models define a notion of teacher. Figure 1.4 shows the basic layout of a teaching model.

In the beginning of the 1990s, several teaching models appeared. Shinohara and Miyano [76] devised a model in which the teacher is required to give information that, when combined with the learners' prior knowledge, uniquely describes the target concept. More precisely, the prior knowledge is modeled as a class of concepts from which the target is drawn. The teacher's task is to give a set of examples such that the target is the only concept in the concept class that matches all examples in the set. The implicit assumption about the learners' behavior is that they are *consistent* and *class-preserving*. Consistent means that they choose only hypotheses that match all received examples; class-preserving means that they choose only hypotheses from the known concept class. The quality measure for the teacher is the number of examples

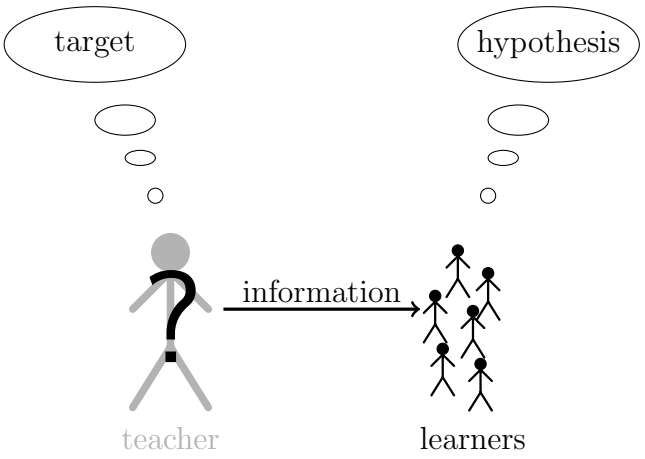

Figure 1.4 *In teaching models a teacher is sought that works well, regardless of the learner's behavior.*

he gives. Shinohara and Miyano call the minimum number of examples necessary for a given target its key size; this value measures the teachability of the target. They show that computing the key size for a given concept is $\mathcal{NP}$-complete, but that it is possible to efficiently find minimum sets of examples for threshold functions.

Independently of the previous teaching model Anthony, Brightwell, Cohen, and Shawe-Taylor [8] introduced a combinatorial parameter, called specification number, for concepts with respect to concept classes. This parameter is identical to the key size. They extensively study the specification number of linearly separable Boolean functions (see also Anthony, Brightwell, and Shawe-Taylor [9]).

The teacher-directed learning model by Goldman, Rivest, and Shapire [36] is based on the online learning model. The teacher's goal is to present examples of the target in such a way that the learners eventually hypothesize the target, making as few mistakes as possible during the teaching process. Again, "all learners" is interpreted as "all consistent and class-preserving learners." Goldman and Kearns [34] show that under the optimal teacher in the teacher-directed model the worst case number of mistakes made by a learner equals the key size, or specification number, of the target concept. They coined the term *teaching dimension*, which has prevailed over the equivalent terms key size and specification number. They also propose to use the maximum teaching dimension over all concepts in a class as a measure for the teachability of that class.

We shall refer to the above three equivalent models as the *teaching dimension model* or *TD model*. The teaching dimensions of many natural concept classes have been determined [34, 8, 46, 75, 11, 50]. As a measure of teachability the teaching dimension has been compared with measures for learnability, such as the Vapnik-Chervonenkis dimension (see [41, 40, 34, 4, 16, 69]). That computing the teaching dimension, or finding a minimum teaching set, for a given concept is $\mathcal{NP}$-complete has been proved by all inventors of the TD model [76, 8, 34]. A hardness result for concepts of a specific class, namely intersections of halfspaces, is given by Servedio [75]. The average teaching dimension over all concepts in a class has been proposed as a better measure for the

teachability of a concept class than the worst case teaching dimension (see [9, 46, 50]).

The approximate testing model by Romanik and Smith [72] and Romanik [70, 71] is similar in spirit to the TD model, but designed specifically for concepts over uncountable domains. It is akin to the PAC learning model in that the learner's hypothesis is allowed to have some error.

Teaching models and learning models can formalize the components of the process in the same way, even though their purposes are different, in fact symmetrical, to each other. Teaching models and learning models can be regarded as lying at opposite sides of a spectrum, with the hybrid models in between (cf. Figures 1.2, 1.3, 1.4). The symmetry between learning and teaching models holds only with respect to the models' structure, not with respect to the amount of attention they have received. This imbalance gives an additional, more theoretical motivation for investigating models for algorithmic teaching.

Both teaching models, the TD model and the approximate testing model, define the learners as all consistent and class-preserving algorithms. This is a natural definition because it is symmetric to the definition of the teachers in the learning models. Roughly speaking, in a learning model a teacher is an entity that has complete knowledge about the target and that makes this knowledge available completely and truthfully. In the Gold style learning, for instance, the teacher has to present all words and only the words in the target language; no information is missing or incorrect. This intuitive characterization of teachers leads to an intuitive characterization of learners: A learner is an entity that incorporates all information it has about the target completely and truthfully into the hypothesis. Incorporating all information means to incorporate the previous knowledge by choosing a class-preserving hypothesis and to incorporate all examples by choosing a consistent hypothesis.

Although the definition of learners is natural, the results given by the TD model often do not match our intuition and are unsatisfactory. A formal version of the following discussion can be found in Section 3.1.

One unrealistic aspect of the model is that the order of examples is irrelevant; the teacher simply gives the examples to the learner all at once. In reality, by contrast, the teacher must present the subject matter in a suitable order. For instance, discussing the recursion theorem before introducing partial recursive functions is not going to work well.

The TD model also assumes that the learners memorize every example and never forget it, which is clearly an unrealistic assumption. Although this assumption could easily be relaxed by allowing the learners to forget an example after, say, m rounds, the model would not become more realistic. A target would be teachable to all these "forgettable" learners if and only if m is at least as large as the teaching dimension of the target. If m is smaller, the learners can always choose a wrong hypothesis without violating the consistency requirement. If m is greater, teaching would nevertheless not

get faster.

In reality, the exact wording of the teacher or the book is hardly ever remembered. Instead, the examples are gradually integrated into the hypothesis. For instance, during a foreign language course, vocabulary and grammar rules have to be learned; but sooner or later the student will use them intuitively without having to recall them whenever saying a sentence. Therefore it seems natural to require that teaching should also be possible if the learner has only a small memory that cannot hold enough examples to describe the target completely.

Another effect that is present in real life teaching, and that should be present in an intelligent tutoring system as well, is that taking into account the learner's feedback when chosing the examples makes the teaching process more efficient. Recall, for instance, the ITS that teaches how to add numbers. Inferring the student's misconceptions from the incorrect answers and accordingly selecting further problems is the main feature of the system. Without getting the student's answers as feedback, adaption to the student would be impossible, and there would not be much tutoring left in the system. Apart from asking questions to assess the student's skills, a teacher can also infer some deficiencies by the questions the student asks. The strongest, most unrealistic, yet easiest formalizable kind of feedback would be the teacher reading the learner's mind. In terms of the model, the teacher has access to the learner's current hypothesis in every round. But in the TD model, even if the teacher is equipped with mind reading capabilities, the worst case learner would still require a number of examples equal to the teaching dimension.

The non-intuitive effects described so far involved the interaction of a teacher and the environment. But in addition to enable us to study concrete teachers, the TD model also permits us to measure the abstract property "teachability" of a target concept and thus to distinguish easily teachable from hard to teach concepts. The teachability of a concept is defined as the teaching dimension.

The teaching dimension for a class is not always a sensible measure for its teachability. For instance, the small and simple class of monomials over n variables has a teaching dimension of 2^n, the same as the huge class of all Boolean concepts. Also the teaching dimension of single concepts can look implausible. For example, Anthony *et al.* [8, 9] show that linearly separable Boolean functions over n variables have a teaching dimension of $(d+1) \cdot 2^{(n-d)}$ where d is the number of relevant variables of the function. Intuitively, however, we would expect that functions with fewer relevant variables are easier to teach. Since 1-decision lists are special cases of linearly separable Boolean functions, this result suggests that 1-decision lists are the easier to teach the longer they are, which seems also implausible.

We conclude that the original TD model is not suitable for the ITS scenario, and also the straightforward modifications allowing feedback and imperfect memory yield no improvement.

Finally, we observe that the teaching dimension model does not work well with

infinite concept classes because in these cases often infinitely many examples are needed to describe the target, which causes the teaching dimension to be infinite, too. It is thus often impossible to compare infinite concepts with respect to their teachability in the teaching dimension model.

1.3 Goal

Our goal is to devise teaching models that remedy the above mentioned flaws. More precisely, we will judge our models by the following aspects:

1. The order in which the teacher presents the information should have an influence on the performance of the teacher.
2. Teaching should get harder when the memory size of the learners decreases, but it should not become impossible for small memory.
3. Teaching should get easier when the learners give feedback to the teacher.
4. Concepts that are more complex should be harder to teach.
5. The teaching model should work for both finite and infinite concepts.

1.4 Main Contributions

As a first step towards our goal we identify two main reasons why the teaching dimension model lacks the five properties described in the previous section. First, the student modeling is unrealistic. A realistic student does not pick an arbitrary hypothesis consistent with all examples seen so far in every round. More realistic assumptions are that a learner chooses hypotheses in a more restricted way and that not all examples ever seen are taken into account. Second, the quality of the teacher is always measured with respect to the worst case student. This neglects the vast majority of students who are not worst case. Often, an action of the teacher helps a reasonable learner, but not a worst case learner. This kind of actions is thus unaccounted for by the teaching performance measure.

To remedy the first reason, in Part I we develop a general framework that contains the TD model as special case. This flexible framework allows us to study various student models that are more realistic than the one used in the TD model. Moreover, it provides a formalization of learner's feedback.

To remedy the second reason, in Part II we modify our framework from Part I to allow an average case analysis of the teaching process. This is done by introducing a single randomized learner that acts as an average case learner.

1.5 Outline of the Results

The analysis of our teaching models parallels the analysis of the TD model in the literature, but with a greater emphasis on the five aspects in Section 1.3. In particular, we study the following questions:

1. Which concepts are teachable?
2. What is the complexity of teaching natural concepts?
3. How does this teachability measure compare with other measures of teachability or learnability?
4. How hard is it to compute the teachability of a concept or to find an optimal teacher?

In Chapter 3 we focus on one of the least well explored points in the TD model, namely the average teaching dimension. We show that the Boolean concept classes of 2-term DNFs as well as of 1-decision lists have a linear average teaching dimension, although their (worst case) teaching dimension is exponential in the number of variables.

In the remainder of Part I we use our framework for two approaches to improving student modeling. In the first approach (Chapter 4), students do not choose *arbitrary* consistent hypotheses any more, but they prefer certain hypotheses over others. As one instantiation of this general idea, in Section 4.2 we consider students that prefer simple hypotheses over complex ones, a property frequently found in learning algorithms and known in learning theory as *Occam's Razor*. We show that according to this teachability measure 1-decision lists are the easier to teach the shorter they are. This contrasts with the counter-intuitive results for 1-decision lists in the TD model. Therefore, this model improves the TD model with respect to Item 4 of Section 1.3. Moreover, it allows to investigate the teachability of infinite concepts, again in contrast to the TD model.

The approach of preferring certain hypotheses is flexible enough to model students that make assumptions about the teacher. In Section 4.3 we introduce learners that prefer hypotheses for which the teacher's examples are optimal. A teacher exploiting this students' assumption can teach more efficiently than in the TD model. If the students assume the teacher behaves in this more efficient way, they change their behavior. The teacher can again adapt to this new behavior and improve the teaching efficiency again. Iterating this results in an infinite series of teachers of monotonically increasing quality. We show that for monomials there is no improvement beyond the second teacher, but that there are infinite concept classes for which every teacher in the series is strictly better than the previous one.

Even if learners prefer certain hypotheses, the teacher still cannot benefit from their giving feedback. This changes when we define the learners such that they memorize an

example only if it is inconsistent with the current hypothesis. This models in some sense the observation that surprising things better stick to memory. We show in Section 4.4 that for those learners there is a quadratic speed-up in teaching time when the teacher receives feedback.

Despite some improvement with respect to feedback and the relationship between complexity and teachability of concepts, the models in Chapter 4 are still essentially batch models, in which the order of examples does not matter and where the teachability can be expressed using a combinatorial parameter of the concept class, in a similar fashion as the teaching dimension in the TD model; no further dependence of the teachability from the memory occurs. The main reason for these shortcomings is that learners are allowed to choose their next hypothesis independently of their current one. It is more realistic to assume that learners change their hypothesis only a little in each round.

In Chapter 5 we introduce a model in which the hypothesis changes are restricted. Now, providing the examples in the right order becomes an important aspect of teaching. Unsuitably ordered examples might lead the learner to assume hypotheses from which it is difficult to reach the target. Moreover, if the learners give no feedback, the teacher must be very careful to give examples such that no learner, regardless of its current hypothesis, is led astray. Carefully defining the hypothesis change restriction can make feedback very advantageous. We show in Section 5.2 that missing feedback can arbitrarily slow down teaching. Due to the hypothesis restriction the behavior of the learner is more complex than in the teaching dimension model. Consequently, it is harder to find good teachers and also to decide whether there is a teacher with a certain performance guarantee at all. We make this intuition precise and show in Section 5.4 that in the new model the question is $\mathcal{PSPACE}$-complete to decide, in contrast to the TD model, in which that question is $\mathcal{NP}$-complete to decide. The corresponding question for memoryless learners is decidable in linear time in the TD model, but is $\mathcal{NP}$-complete in the new model.

In Part II we modify the TD model so as to perform an average case analysis instead of the usual worst case analysis. More precisely, we consider a randomized learner that combines the behavior of the learners in the TD model. In particular, no restrictions as in the Chapters 4 and 5 apply to the learner. This randomized version of the TD model is superior to the traditional TD model in terms of Items 1.–5. of our goal. The performance of a teacher now depends on the order in which it presents the examples. Furthermore, even learners with very small memory can be taught, even though it takes a long time. If the memory grows, the expected duration of teaching decreases smoothly. Teaching success happens sooner when feedback is available to the teacher (see Chapter 6).

We distinguish randomized learners according to their memory size and whether or not they give feedback. In Chapter 7 we study the randomized learners that have infinite memory or give feedback. In Section 7.1 we characterize optimal teachers

for the memoryless learner with feedback and use the characterization to develop an efficient optimal teaching algorithm for the concept class of monomials. In Section 7.2 we show that for learners with infinite memory and with feedback there is always an optimal teacher that gives in each round an example that is inconsistent with the learner's current hypothesis. We use this result to develop a characterization of optimal teachers for these kind of learners. Finally, we show that the optimal performance of a teacher for learners with infinite memory is computable, but hard to even approximate.

Memoryless learners that give no feedback are the hardest to analyze; we dedicate Chapter 8 to the analysis. No algorithm for finding an optimal teacher or for determining the performance of the optimal teacher is known. We first derive characterizations of successful teachers and, using recent results on unobservable stochastic shortest path problems [61, 62], of optimal teachers (see Sections 8.1 and 8.2). Then we prove the problem of finding the optimal teacher's performance to be $\mathcal{NP}$-hard. Moreover, we show that there is an algorithm for approximating this optimal performance value (see Chapter 8.3).

In general it is not possible to compute the performance of a teacher or to decide whether it is successful at all. In Section 8.4 we investigate a kind of teachers that give the same sequence of examples in an endless loop. These cyclic teachers are much easier to handle than teachers in general, but are nevertheless powerful enough to come arbitrarily close to the optimum. We devise a cyclic teacher for the monomials that needs twice the time than our optimal algorithm for learners with feedback (Section 7.1), thus yielding an upper bound for the optimal teaching performance for monomials. Finally, in Section 8.5 we investigate a natural greedy strategy for teaching algorithms. We show that this strategy, although optimal for some simple concept classes, yields no constant factor approximation to the optimal performance in general.

We conclude the thesis with a discussion of the relevance of our models and their relation to traditional teaching and learning models. We also point towards some future research.

Some results in the Chapters 3 and 4 have been presented at the COLT 2005 conference [11]. The contents of Chapter 5, except for Section 5.4, has been presented at the ALT 2005 conference [12]. Most results of the Chapters 6 and 7 have been presented at the COLT 2006 conference [14]. Finally, the results in Chapter 8 have been presented at ALT 2006 [13].

Chapter 2

Preliminaries

2.1 Basic Definitions and Notations

Let $\mathbb{N} = \{0, 1, 2, \dots\}$ be the set of all natural numbers, $\mathbb{N}^+ = \mathbb{N} \setminus \{0\}$ the set of all positive natural numbers, and $\mathbb{R}$ the set of all real numbers. The cardinality of a set S is denoted by $|S|$, and $|S| = \infty$ means S is infinite. We use "$\subseteq$" to denote the *subset* relation and "$\subset$" for the *proper subset* relation. $A \triangle B$ is the *symmetric difference* of the sets A and B. The symbol "$\forall^\infty$" means "for all but finitely many." We set the minimum of the empty set to infinity, that is, $\min \emptyset = \infty$. The *power set* of a set M is written as 2^M. For $a, b \in \mathbb{N}$ with $a \leq b$ the set $\{a, \dots, b\} \subset \mathbb{N}$ is abbreviated as $[a, b]$. For a function $f\colon M \to \mathbb{R}$ over an arbitrary set M and a set $A \subseteq M$ we define

$$\operatorname*{argmin}_{x \in A} f(x) = \{x \in A \mid f(x) = \min\{f(x') \mid x' \in A\}\}.$$

If $\{f(x') \mid x' \in A\}$ has no minimum, then $\operatorname{argmin}_{x \in A} f(x) = \emptyset$. The definition of $\operatorname{argmax}_{x \in A} f(x)$ is analogous.

For any set S we denote by S^* the set of all finite sequences over S and by S^m and $S^{\leq m}$ the set of all sequences of length m and at most length m, respectively. We use bold lowercase letters as identifiers for sequences. Elements forming a sequence are enclosed in angle brackets. The *empty sequence* is denoted by $\langle\,\rangle$, the *length* of a sequence $\boldsymbol{s} \in S^*$ by $|\boldsymbol{s}|$. For the i-th element of a sequence $\boldsymbol{s}$ ($i = 1, \dots, |\boldsymbol{s}|$) we write $\boldsymbol{s}[i]$. We use the symbol $\circ$ for *concatenation* of sequences and, for $m \in \mathbb{N}^+$, the symbol $\circ_m$ for a length-restricted concatenation with a singleton sequence:

$$\langle x_1, \dots, x_k \rangle \circ_m \langle y \rangle = \begin{cases} \langle x_1, \dots, x_k, y \rangle & \text{if } k < m, \\ \langle x_{k-m+2}, \dots, x_k, y \rangle & \text{if } k \geq m. \end{cases}$$

For notational convenience, we set $\circ_\infty$ to be the standard concatenation $\circ$. For a finite sequence $\boldsymbol{s}$ and an $\ell \in \mathbb{N}^+$ we write $\boldsymbol{s}^\ell$ for $\underbrace{\boldsymbol{s} \circ \cdots \circ \boldsymbol{s}}_{\ell \text{ times}}$. For two sequences $\boldsymbol{s}, \boldsymbol{s}'$ we write $\boldsymbol{s} \sqsubseteq \boldsymbol{s}'$ if $\boldsymbol{s}$ is a prefix of $\boldsymbol{s}'$.

We sometimes identify a sequence and the set of all its elements.

As strings are only sequences of characters, we use similar notations for strings as for sequences: For a string σ over an alphabet Σ we denote by $|\sigma|$ its *length* and by Σ^* the set of all finite strings over Σ including the *empty string* Λ; $\sigma[i]$ is the i-th symbol in σ ($i = 1, \ldots, |\sigma|$). In addition, we may omit the symbol $\circ$ in concatenations of strings.

For $b \in \{0,1\}$ we abbreviate $1 - b$ with $\bar{b}$. For $k \in \mathbb{N}$ and $n \in \mathbb{N}^+$, the number $r = k \bmod n$ is defined as the unique number in $[0, n-1]$ with $r \equiv k \pmod{n}$. To save parentheses, we assume that the operator mod has a higher priority than addition. That means "$1 + k \bmod n$" is equivalent to "$1 + (k \bmod n)$."

To describe our teaching models we mostly use standard notions from algorithmic learning theory. We always assume a countable *learning domain* X whose elements we call *instances*. A *concept* is a function $c\colon X \to \{0,1\}$. A *concept class* $\mathcal{C}$ is a set of concepts. A pair $(x, b) \in X \times \{0,1\}$ of an instance and a Boolean *label* is called *example*. It is *positive* if $b = 1$, otherwise *negative*. The set of all examples is denoted by $\mathcal{X} = X \times \{0,1\}$. A concept c is *consistent* with an example (x, b) iff $c(x) = b$ and consistent with a set $S \subseteq \mathcal{X}$ of examples iff $c(x) = b$ for all $(x, b) \in S$. A set of examples is also called a *sample*. We often implicitly identify a concept c with the set $\{x \in X \mid c(x) = 1\}$ and vice versa a set $Y \subseteq X$ with the concept $c\colon c(x) = 1 \Leftrightarrow x \in Y$. For example, the empty concept $\emptyset$ is identified with the constant-0 function over X.

For a sample S, let $\mathcal{C}(S) = \{c \in \mathcal{C} \mid c \text{ is consistent with } S\}$. We denote by $\mathcal{X}(c) = \{(x, c(x)) \mid x \in X\}$ the set of all examples for c.

A sample S is called a *teaching set* [34, 36] (also known as key [76], specifying set [8], discriminant [57], and witness set [47]) for c with respect to $\mathcal{C}$ iff $\mathcal{C}(S) = \{c\}$. The *teaching dimension* of c with respect to $\mathcal{C}$ is defined as the cardinality of the smallest teaching set of c:

$$TD(c, \mathcal{C}) = \min\{|S| \mid \mathcal{C}(S) = \{c\}\}.$$

We simply write $TD(c)$ if the concept class is clear. The teaching dimension of the class $\mathcal{C}$ is defined as the maximum teaching dimension over all concepts: $TD(\mathcal{C}) = \max\{TD(c, \mathcal{C}) \mid c \in \mathcal{C}\}$. If the learning domain X is infinite, there might be no finite teaching set for a given concept. In this case we say the teaching dimension of c is infinite and write $TD(c) = \infty$. Similarly, $TD(\mathcal{C}) = \infty$ if there is a concept with infinite teaching dimension or if the set $\{TD(c) \mid c \in \mathcal{C}\}$ is not bounded from above.

Two concepts differing only with respect to one instance are called *neighbor concepts*. The number of neighbor concepts of c in the class $\mathcal{C}$ is a lower bound for the teaching dimension of c with respect to $\mathcal{C}$ because each neighbor concept must be ruled out by a separate example.

To investigate the computational complexity of algorithms coping with concepts (as input or output) it is necessary to use finite descriptions as *representations* for concepts. Especially for classes over infinite learning domains, defining representations is unavoidable. Formally, representations are strings over a finite alphabet Σ. A *representation function* is a function $\varrho\colon \Sigma^* \times X \to \{0, 1, \uparrow\}$. For a string $\sigma \in \Sigma^*$ and an

	X	1	2	3	4	5	6	7	8
σ									
0		1	1	1	1	1	1	1	1
1		0	1	1	1	1	1	1	1
2		1	0	1	1	1	1	1	1
3		1	1	0	1	1	1	1	1
4		1	1	1	0	1	1	1	1
5		1	1	1	1	0	1	1	1
6		1	1	1	1	1	0	1	1
7		1	1	1	1	1	1	0	1
8		1	1	1	1	1	1	1	0

Figure 2.1 *The concept class $\mathcal{S}_8$ over learning domain $[1,8]$. The concept represented by 0 is the all-concept $[1,8]$, the concepts represented by $i = 1,\ldots,8$ are the co-singletons $[1,8]\setminus\{i\}$.*

instance $x \in X$, $\varrho(\sigma, x)$ specifies whether x belongs to the concept represented by σ or not. The value "$\uparrow$" signals an invalid input, for example, a σ that should not represent any concept. By ϱ_σ we denote the function $\varrho_\sigma\colon X \to \{0,1,\uparrow\}$ with $\varrho_\sigma(x) = \varrho(\sigma, x)$ for all $x \in X$. A string $\sigma \in \Sigma^*$ is a *representation* for a concept $c\colon X \to \{0,1\}$ iff for all $x \in X$, $\varrho_\sigma(x) = c(x)$. The *representation size* $\|c\|$ of a concept with respect to a function ϱ is the length of its shortest representation: $\|c\| = \min\{|\sigma| \mid \varrho_\sigma = c\}$. If c is not represented by any σ, we have $\|c\| = \infty$. The concept class of all concepts represented by ϱ is denoted $\mathcal{C}_\varrho = \{\varrho_\sigma \mid \forall x \in X : \varrho_\sigma(x) \in \{0,1\}\}$ and the set of all representations for concepts in a set $C \subseteq \mathcal{C}_\varrho$ is $\varrho(C) = \{\sigma \in \Sigma^* \mid \varrho_\sigma \in C\}$; we abbreviate $\varrho(\{c\})$ with $\varrho(c)$. Finally, a *represented concept class* is a pair $\langle \varrho, \mathcal{C}\rangle$ of a representation function ϱ and a concept class $\mathcal{C}$ such that $\mathcal{C} \subseteq \mathcal{C}_\varrho$. Note that X and Σ are implicitly given by the function ϱ and thus we do not have to mention them explicitly in the represented class.

For convenience we sometimes identify the representation language Σ^* with the set $\mathbb{N}$. This happens only for concept classes for which the "real" representations are not important. We shall always mention that at the definition of the concept class. Using this interpretation, the minimum $\min S$ and the maximum $\max S$ of a set $S \subseteq \Sigma^*$ are well-defined.

Three learning domains will be used throughout this thesis: The set $[1,n]$ to describe n objects with no particular structure, the set $\{0,1\}^n$ containing all 2^n Boolean vectors of length n, typically interpreted as assignments to n Boolean variables, and the infinite set $\{\mathtt{a},\mathtt{b}\}^*$ of all words over the alphabet $\{\mathtt{a},\mathtt{b}\}$. To denote subsets of $\{0,1\}^n$ we use notations like $1^k\{0,1\}^{n-k}$ for the set $\{1^k y \mid y \in \{0,1\}^{n-k}\}$.

For the purpose of illustrating the various models introduced later, we use a few common natural concept classes. The first and simplest one is $\mathcal{S}_n$ over $[1,n]$. This

class contains all *co-singleton concepts* $c_i = [1,n] \setminus \{i\}$ for $i = 1,\ldots,n$ and the all-concept $c_0 = [1,n]$. There is a straightforward representation function with $\varrho_i = c_i$ for $i = 0,\ldots,n$ (see Figure 2.1). Another concept class over $[1,n]$ is $\mathcal{A}_n = 2^{[1,n]}$, the class of *all* concepts over $[1,n]$.

To describe concept classes over the learning domain $\{0,1\}^n$ of Boolean assignments, we use *Boolean variables* $v_1,\ldots,v_n$. We denote by v_i^0 or $\bar{v}_i$ *negative literals*, and by v_i^1 or v_i *positive literals*. For a literal $w = v_i^\alpha$ we write $\bar{w}$ for $v_i^{1-\alpha}$. An instance $x \in \{0,1\}^n$ is interpreted as an assignment to $v_1,\ldots,v_n$ in the canonical way.

1-decision lists are popular concepts in learning theory [68, 39, 77, 24, 38] and have also been investigated in the context of teaching [34, 35]. A 1-decision list over n Boolean variables is a list

$$D = \langle (w_1,b_1),(w_2,b_2),\ldots,(w_m,b_m),(*,b_{m+1})\rangle$$

of *nodes* (w_j,b_j) consisting of literals $w_j \in \{v_i^\alpha \mid i \in [1,n], \alpha \in \{0,1\}\}$ and *labels* $b_j \in \{0,1\}$. A node (w,b) is called *positive* if $b = 1$, *negative* otherwise. The node $(*,b)$ is called *default node*.

The concept $c_D\colon \{0,1\}^n \to \{0,1\}$ described by D is defined as $c_D(x) = b_j$ for the minimal $j \le m$ such that x satisfies w_j, and $c_D(x) = b_{m+1}$ if no such j exists. Note that a default node is present in every decision list. Two decision lists D, E are called *equivalent* iff $c_D = c_E$. We say a node (w_i,b_i) *absorbs* an instance x if w_i is the first literal in D satisfied by x. We say an instance x *reaches* a node (w_i,b_i) if x is not absorbed by any node before (w_i,b_i). Furthermore, we say an instance x *leaves* a node if it reaches the node, but is not absorbed by it. For example, in the decision list (2.1) below, the instance 01010 is absorbed by node $(\bar{v}_3,0)$; the instance 01101 is absorbed by the default node.

The set of all concepts represented by 1-decision lists over n variables is denoted by $\mathcal{D}_n$. We define the *length* of a decision list as the number of nodes (not counting the default node), $len(D) = m$, and the *length* of a concept $c \in \mathcal{D}_n$ as the length of its shortest decision list: $len(c) = \min\{len(D) \mid c_D = c\}$. We denote by $\mathcal{D}_{len\le\ell}^n \subseteq \mathcal{D}_n$ the set of all 1-decision list concepts of length at most ℓ.

To make 1-decision lists formally fit into our representation framework we need a representation function ϱ_{DL} for $\mathcal{D}_n$. An informal description of ϱ_{DL} will suffice. A node (v_j^α,b) is represented by a string of length $2 + \lceil \log n \rceil$ containing one bit for α and b each, and $\lceil \log n \rceil$ bits for j. The default node is represented by the label alone. We only need the fact that two decision lists have the same representation length if and only if they have the same length. For all our purposes we can regard the length of a 1-decision list as their representation size.

As usual, we assume 1-decision lists to be in reduced form, that is, each variable occurs at most once (either negated or not) and the default node and its predecessor (if any) have different labels. A 1-decision list can be transformed into an equivalent

reduced 1-decision list in linear time [35] and reduced decision lists are of minimal length [77].

Furthermore, we assume that the last two nodes before the default node have the same label. In the special case of decision lists of length 1 there are two equivalent reduced lists: $\langle(w, b), (*, \bar{b})\rangle$ and $\langle(\bar{w}, \bar{b}), (*, b)\rangle$. We then assume that the default label is negative. We call reduced lists with these properties *normal form decision lists* (NFDL). Every reduced 1-decision list either is an NFDL or can be easily transformed into one by inverting the default label and the last node's label and literal. For example, the 1-decision list

$$\langle(v_1, 1), (\bar{v}_3, 0), (v_4, 1), (v_3, 1), (v_5, 0), (v_6, 1), (*, 1)\rangle$$

in reduced form is

$$\langle(v_1, 1), (\bar{v}_3, 0), (v_4, 1), (v_5, 0), (*, 1)\rangle$$

and in normal form

$$\langle(v_1, 1), (\bar{v}_3, 0), (v_4, 1), (\bar{v}_5, 1), (*, 0)\rangle. \tag{2.1}$$

Every NFDL can be divided into *segments* (also called levels [35, 77]), that is, maximum length sequences of consecutive nodes with the same label. For example, the segments in (2.1) are $D_1 = \langle(v_1, 1)\rangle$, $D_2 = \langle(\bar{v}_3, 0)\rangle$, and $D_3 = \langle(v_4, 1), (\bar{v}_5, 1)\rangle$. The segments of the 1-decision list in Figure 4.2 on Page 68 are $D_1 = \langle(v_1, 1)\rangle$, $D_2 = \langle(v_2, 0), (v_3, 0), (v_4, 0)\rangle$, $D_3 = \langle(v_5, 1), (v_6, 1)\rangle$. We write $D = D_1 \circ \cdots \circ D_r \circ \langle(*, b)\rangle$ to indicate the segmentation of D into segments $D_1, D_2, \ldots, D_n$. Note that the default node is determined by the last node in segment D_r. Thus we can denote the segmentation also simply by $D = D_1 \circ \cdots \circ D_r$, provided the length of D is a least one.

Another concept class over $\{0,1\}^n$ is the class of *monomials*. A monomial is a conjunction of Boolean literals, for example, $v_1 \wedge \bar{v}_3 \wedge v_4$. It describes a concept over $\{0,1\}^n$ in the canonical way if the instances in $\{0,1\}^n$ are seen as assignments to n variables. For example, over $\{0,1\}^4$, the monomial $v_1 \wedge \bar{v}_3 \wedge v_4$ describes the concept $\{1001, 1101\} \subseteq \{0,1\}^4$. Note that the empty concept, described by all monomials containing a variable and its negation, is not always considered a monomial. We shall always make clear whether or not we include the empty concept, but always use the same symbol, $\mathcal{M}_n^1$, for the set of all monomials over n variables.

Every monomial, except the contradictory ones, can be represented by a string $M \in \{\mathtt{0}, \mathtt{1}, \mathtt{*}\}^n$, where $M[i] = \mathtt{0}, \mathtt{1}, \mathtt{*}$ specifies whether the variable v_i occurs negated, unnegated, or not at all. Thus, `1*01` represents $v_1 \wedge \bar{v}_3 \wedge v_4$. One gets the set of satisfying assignments by replacing every $*$ with all values from $\{0,1\}$. Sometimes we abuse notation and identify M with the concept it represents. Note that $M_1 \subseteq M_2$ if and only if for all i, $M_1[i] = M_2[i]$ or $M_2[i] = *$. Moreover, $M_1 \subset M_2$ if and only if $M_1 \subseteq M_2$ and there is an i with $M_2[i] = * \neq M_1[i]$. When we include the empty concept into the set of monomials we represent it by the empty string Λ.

The previous paragraph implies a definition of a representation function $\varrho\colon \{0,1,*\}^* \times \{0,1\}^n \to \{0,1,\uparrow\}$ over the alphabet $\{0,1,*\}$, according to which all non-contradictory monomials have a representation size of n and the contradictory monomials one of 0.

A disjunction of at most two monomials is called a *2-term DNF*. For two monomials M_1, M_2 the 2-term DNF $M_1 \vee M_2$ represents the union of the concepts represented by M_1 and M_2. For example, $1{*}01 \vee 00{*}{*}$ represents $\{1001, 1101, 0000, 0001, 0010, 0011\} \subset \{0,1\}^4$. When two 2-term DNFs $M_1 \vee M_2$ and $K_1 \vee K_2$ represent the same concept, we write $M_1 \vee M_2 \equiv K_1 \vee K_2$. We denote by $\mathcal{M}_n^2$ the class of all concepts representable by 2-term DNFs over n variables. We always assume $\emptyset \in \mathcal{M}_n^2$.

When we consider two monomials M_1, M_2 we say that they have a *strong difference at i* iff $M_1[i], M_2[i] \in \{0,1\}$ and $M_1[i] \neq M_2[i]$. They have a *weak difference at i* iff either $M_1[i] = *$ and $M_2[i] \in \{0,1\}$ or $M_2[i] = *$ and $M_1[i] \in \{0,1\}$. Two weak differences, at positions i and j, are said to be *of the same kind* iff $M_q[i] = M_q[j] = *$ for a $q \in \{1,2\}$, that is, iff both $*$ occur in the same monomial; they are called *of different kind* otherwise.

For a string $M \in \{0,1,*\}^n$ we denote by $M[\frac{*}{0}]$ and $M[\frac{*}{1}]$ the string resulting from substituting all $*$'s by 0 and 1, respectively. Strings $s, s' \in \{0,1\}^n$ are called *neighbors* iff they differ in only one position and *i-neighbors* iff they differ only in the i-th position.

2.2 The Formal Teaching Framework

In this section we formally define our teaching framework, which is a generalization and extension of the classical teaching dimension model. It is more general in that it specifies only the shape of the learner, but allows for various concrete definitions of them, whereas in the TD model the learners are fixed to be consistent and class-preserving. The framework is an extension because it adds different kinds of feedback to the TD model, which lacked a notion of feedback altogether. Every kind of feedback requires a different formalization of the teacher and also of the success criterion.

In the following we discuss all components of the teaching process, as introduced on Page 2. First we present a semi-formal description of the framework. Afterwards we go into the details of formalizing the learners, the teachers, and the success criteria. At one point, however, we shall make a rather lengthy and technical digression, comparing two kinds of teachers and finally rejecting one of them. This is not essential for understanding the framework itself, only for understanding one particular design decision. For the reader who is only interested in the bare framework, we list the relevant definitions that constitute the framework at the end of this section.

The *target* is a concept $c^*\colon X \to \{0,1\}$ from a represented concept class $\langle \varrho, \mathcal{C} \rangle$ over a learning domain X.

The *learner* maintains a *hypothesis* from $\varrho(\mathcal{C})$ and has a memory for storing examples from $\mathcal{X}(c^*)$. We model learners as automata whose state consists of hypothesis and

memory and that change their state according to the information received by the teacher. More formally, the state of the learning automaton is a pair $(\boldsymbol{s}, \sigma) \in \mathcal{X}^* \times (\varrho(\mathcal{C}) \cup \{\mathit{init}\})$ consisting of a memory $\boldsymbol{s}$ and a hypothesis σ. Here, *init* is the name of a special hypothesis, which denotes the initial hypothesis of the learner. We decided to introduce a special initial hypothesis in order to avoid to pick an arbitrary hypothesis from $\varrho(\mathcal{C})$ as initial one or to add the initial hypothesis as another parameter on which the teaching process depends. Whenever we give precise definitions of a learner we have to define its behavior for the hypothesis *init*, too. When defining concrete learners, we shall always make sure that this hypothesis is not reached at any round after the initial one.

For generality, the memory is defined as a list of examples rather than as a set. Used as a queue, for instance, a list gives a simple means to implement the age of an example, that is, the number of rounds it has already been memorized. Initially the learner's memory is empty. Thus the initial state of the learner is $(\langle\,\rangle, \mathit{init})$.

The learner is class-preserving because it chooses hypotheses only from $\varrho(\mathcal{C})$. This means the learner has as *prior knowledge* that the target is contained in $\mathcal{C}$.

The teaching proceeds in rounds. The basic form of *interaction* consists in the teacher giving an example from $\mathcal{X}(c^*)$ to the learner in every round. These examples are thus the *information* the learner receives. An optional, additional form of interaction is the learner revealing parts of its state to the teacher. Such a learner will be said to give feedback. The only form of feedback we shall consider is the extreme variant in which the learner reveals the complete state consisting of hypothesis and memory at the beginning of every round. Without receiving feedback, the *teacher* is just an infinite sequence of examples, one for every round of the teaching process. When receiving feedback, the teacher chooses the next example depending on the received feedback.

The process is considered *successful* if and when the hypothesis is correct, that is, from $\varrho(c^*)$. The quality of a given teacher is measured by the number of rounds it needs in order to be successful.

The framework leaves open the precise properties of the learners. In order to complete the definition of learners we have to specify two parts of their behavior:

(MEM) The memory of the learners. How many examples can be memorized? Which examples are memorized and for how long?

(HYP) The way the hypothesis is chosen. Which hypotheses are admissible? Which are preferred?

The form of the teacher depends on the kind of feedback it receives. We thus have to specify yet another aspect.

(FB) The feedback given by the learners and observed by the teacher.

A straightforward attempt to formalize learners is as deterministic automata. And even though we shall later abandon them in favor of non-deterministic automata, we nevertheless discuss deterministic ones to some extent, because teaching them yields a simpler process than teaching non-deterministic learners. Moreover, seeing the problems caused by using deterministic automata gives some explanation as to why we finally opted for the non-deterministic variant.

Definition 2.1 *Let $\langle \varrho, \mathcal{C} \rangle$ be a represented concept class with $\varrho\colon \Sigma^* \times X \to \{0, 1, \uparrow\}$. A* deterministic learner using hypothesis space $\langle \varrho, \mathcal{C} \rangle$ *is a function $L\colon \mathcal{X}^* \times (\varrho(\mathcal{C}) \cup \{\mathit{init}\}) \times \mathcal{X} \to \mathcal{X}^* \times \varrho(\mathcal{C})$ with component functions*

$$L_{mem}\colon \mathcal{X}^* \times (\varrho(\mathcal{C}) \cup \{\mathit{init}\}) \times \mathcal{X} \to \mathcal{X}^*$$

and

$$L_{hyp}\colon \mathcal{X}^* \times (\varrho(\mathcal{C}) \cup \{\mathit{init}\}) \times \mathcal{X} \to \varrho(\mathcal{C}).$$

We write $L = (L_{mem}, L_{hyp})$ for a learner and $L(\boldsymbol{s}, \sigma, z)$ instead of $L((\boldsymbol{s}, \sigma), z)$.

One remark about the term "hypothesis" is in order. Although, strictly speaking, "hypothesis" refers to a representation $\sigma \in \varrho(\mathcal{C}) \cup \{\mathit{init}\}$, we often sloppily say "the hypothesis is c" and mean "the hypothesis is contained in $\varrho(c)$." The phrase "target hypothesis" refers to every hypothesis in $\varrho(c^*)$.

The feedback can be of various forms. Only two forms will be considered in this thesis, although many others would be sensible, too. The two possible implementations for (FB) are the extreme ones, namely: (1) no feedback at all, (2) the complete current state consisting of hypothesis and memory.

Our specifications for the aspect (FB) can be combined with every possible definition of (MEM) and (HYP). We can thus take a model and vary the (FB) aspect in order to investigate the influence of feedback, or vary (MEM) to study different memory sizes.

In the first instance we pose no restrictions on the teachers other than that they use only that information about the learner's state that is specified by (FB). A teacher may remember the whole history of the teaching process (all examples given so far, all feedback received). Throughout the thesis we do not require the teacher to be a computable or even efficiently computable function. Nevertheless, all concrete teacher we present are efficiently computable.

Depending on (FB) the formal description of teachers differs greatly. At first we introduce teachers *without* feedback as they are the simpler ones. They do not depend on the learner's state nor on the history and thus simply output one example in every round.

Definition 2.2 *Let $\langle \varrho, \mathcal{C} \rangle$ with $\varrho\colon \Sigma^* \times X \to \{0, 1, \uparrow\}$ be a represented concept class and $c^* \in \mathcal{C}$ a target. A* teacher without feedback *is a function $T\colon \mathbb{N} \to \mathcal{X}(c^*)$.*

A teacher without feedback and a deterministic learner determine a sequence of learner's states. To denote the state of a deterministic learner $L = (L_{mem}, L_{hyp})$ after t rounds of interaction with a teacher T we use the notation $L(T, t)$ with components $L_{mem}(T, t)$ and $L_{hyp}(T, t)$. Note that after t rounds the teacher has given the examples $T(0), \ldots, T(t-1)$. Thus the next example is $T(t)$. We have

$$\begin{aligned} L(T, 0) &= (\langle\,\rangle, \mathit{init}), \\ L(T, t+1) &= L(L(T, t), T(t)). \end{aligned}$$

Example 2.3 Consider the concept class $\mathcal{S}_n$ and the target $c^* = [1, n]$. Let $L^1 = (L^1_{hyp}, L^1_{mem})$ be a learner that memorizes every example and hypothesizes the consistent hypothesis with least index:

$$\begin{aligned} L^1_{mem}(\boldsymbol{s}, \sigma, z) &= \boldsymbol{s} \circ \langle z \rangle, \\ L^1_{hyp}(\boldsymbol{s}, \sigma, z) &= \min \varrho(\mathcal{C}(L_{mem}(\boldsymbol{s}, \sigma, z))). \end{aligned}$$

A teacher $T_1\colon \mathbb{N} \to \mathcal{X}(c^*)$ with $T_1(i) = (1 + (i \bmod n), 1)$ for all i makes L^1 hypothesize $[1, n]$ after one example, because $L^1_{hyp}(\langle\,\rangle, \mathit{init}, (1, 1)) = 0$.

We define the learner L^2 like L^1 but with max instead of min:

$$\begin{aligned} L^2_{mem}(\boldsymbol{s}, \sigma, z) &= \boldsymbol{s} \circ \langle z \rangle, \\ L^2_{hyp}(\boldsymbol{s}, \sigma, z) &= \max \varrho(\mathcal{C}(L_{mem}(\boldsymbol{s}, \sigma, z))). \end{aligned}$$

The teacher T_2 with $T_2(i) = (n - (i \bmod n), 1)$ makes L^2 hypothesize $[1, n]$ after n rounds. Memory contents and hypothesis of L^2 during the teaching process are as follows:

round t	memory $L^2_{mem}(T, t)$	hypothesis $L^2_{hyp}(T, t)$	example $T(t)$
0	$\langle\rangle$	*init*	$(n, 1)$
1	$\langle(n, 1)\rangle$	$n - 1$	$(n - 1, 1)$
2	$\langle(n, 1), (n - 1, 1)\rangle$	$n - 2$	$(n - 2, 1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n - 1$	$\langle(n, 1), \ldots, (2, 1)\rangle$	1	$(1, 1)$
n	$\langle(n, 1), \ldots, (2, 1), (1, 1)\rangle$	0	$(n, 1)$

■

Rather than teaching only one learner at a time, our ITS scenario requires teaching many learners simultaneously. A straightforward way to incorporate this is to consider a set of deterministic learners and require the teacher to teach them all. As we already

mentioned, we shall, however, use a non-deterministic learner instead, for reasons that become clear when we consider teaching *with* feedback, which we do next.

A teacher receiving in every round the complete state as feedback is harder to model than a feedbackless teacher. In the most general form, such a teacher's behavior can depend on the complete history of the teaching process. Formally, a history is a sequence

$$\langle(\langle\,\rangle, \mathit{init}, z_0), (\boldsymbol{s}_1, \sigma_1, z_1), (\boldsymbol{s}_2, \sigma_2, z_2)\ldots\rangle$$

from the set of all histories

$$\mathit{HIST} = (\mathcal{X}^* \times (\varrho(\mathcal{C}) \cup \{\mathit{init}\}) \times \mathcal{X})^*.$$

The j-th element in a history is a triple consisting of the learner's memory and hypothesis and the example given by the teacher in round j. This would yield a definition for T as a function

$$T\colon \mathbf{(}\mathcal{X}^* \times (\varrho(\mathcal{C}) \cup \{\mathit{init}\})\mathbf{)} \times \mathbf{(}\mathcal{X}^* \times (\varrho(\mathcal{C}) \cup \{\mathit{init}\}) \times \mathcal{X}\mathbf{)}^* \to \mathcal{X}(c^*)$$

where the arguments inside the first pair of bold parentheses describe the current state and the ones in the second pair the history.

Definition 2.4 *Let $\langle\varrho, \mathcal{C}\rangle$ with $\varrho\colon \Sigma^* \times X \to \{0, 1, \uparrow\}$ be a represented concept class and $c^* \in \mathcal{C}$ a target. A* history-aware teacher for c^* *is a function*

$$T\colon (\mathcal{X}^* \times (\varrho(\mathcal{C}) \cup \{\mathit{init}\})) \times (\mathcal{X}^* \times (\varrho(\mathcal{C}) \cup \{\mathit{init}\}) \times \mathcal{X})^* \to \mathcal{X}(c^*).$$

This looks very complicated. It also looks unnecessary because if the learner's behavior does not depend on the history, the behavior of the teacher should also be independent of the history.

For simplicity we would prefer a teacher of the form $T\colon \mathcal{X}^* \times (\varrho(\mathcal{C}) \cup \{\mathit{init}\}) \to \mathcal{X}(c^*)$ that takes into account only the currently observed state of the learner.

Definition 2.5 *Let $\langle\varrho, \mathcal{C}\rangle$ with $\varrho\colon \Sigma^* \times X \to \{0, 1, \uparrow\}$ be a represented concept class and $c^* \in \mathcal{C}$ a target. A* teacher with feedback for c^* *is a function*

$$T\colon \mathcal{X}^* \times (\varrho(\mathcal{C}) \cup \{\mathit{init}\}) \to \mathcal{X}(c^*).$$

Note that history-aware teachers, by definition, also receive feedback and should thus be called "history-aware teachers with feedback." But we shall stick to the shorter term "history-aware."

The next example shows informally that Definition 2.5 is indeed a restriction, because some classes of learners can only be taught by history-aware teachers and not by plain teachers with feedback.

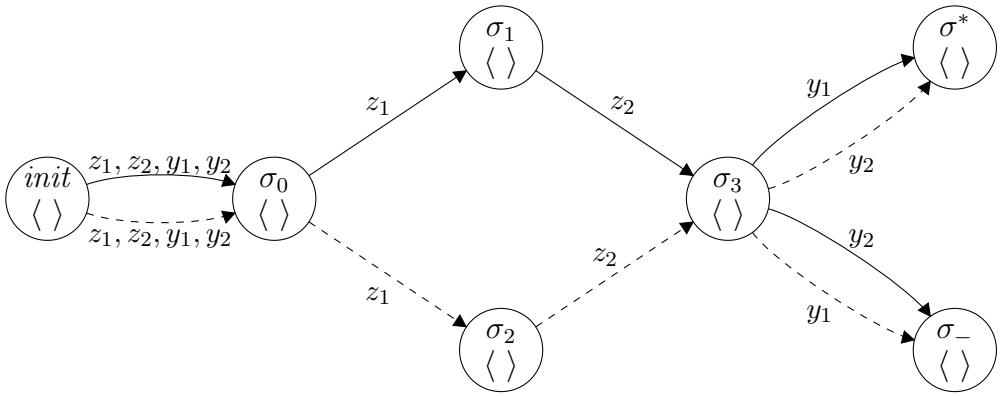

Figure 2.2 *A learning scenario that requires history-aware teachers. The graph describes a concept class with representations $\sigma_0, \sigma_1, \sigma_2, \sigma_3, \sigma^*, \sigma_-$. The nodes are the possible states the two learners L^1 and L^2 can be in. The solid and dashed lines specify the state transitions for L^1 and L^2, respectively. Examples that do not cause a state change are left out. After reaching σ_3, learner L^1 needs example y_1 and L^2 needs y_2 to reach the target σ^*. But a teacher cannot distinguish between L^1 and L^2 based only on the current state. Which learner the teacher deals with can only be seen by looking at the history, namely whether the learner was in σ_1 or σ_2 before.*

Example 2.6 Consider a concept class $\mathcal{C}$ over $\{z_1, z_2, y_1, y_2\}$ whose representation set is $\varrho(\mathcal{C}) = \{\sigma_0, \sigma_1, \sigma_2, \sigma_3, \sigma^*, \sigma_-\}$; σ^* is the target representation and σ_- is a dead end that signals "failure." The concepts themselves are irrelevant for the purpose of this example.

We define two learners L^1, L^2 that do not memorize examples, that is, $L^1_{mem}(\boldsymbol{s}, \sigma, z) = L^2_{mem}(\boldsymbol{s}, \sigma, z) = \langle\,\rangle$ and that choose hypotheses according to the following tables (see also Figure 2.2):

L^1_{hyp}	z_1	z_2	y_1	y_2
$init$	σ_0	σ_0	σ_0	σ_0
σ_0	σ_1	σ_0	σ_0	σ_0
σ_1	σ_1	σ_3	σ_1	σ_1
σ_2	σ_2	σ_2	σ_2	σ_2
σ_3	σ_3	σ_3	σ^*	σ_-
σ^*	σ^*	σ^*	σ^*	σ^*
σ_-	σ_-	σ_-	σ_-	σ_-

L^2_{hyp}	z_1	z_2	y_1	y_2
$init$	σ_0	σ_0	σ_0	σ_0
σ_0	σ_2	σ_0	σ_0	σ_0
σ_1	σ_1	σ_1	σ_1	σ_1
σ_2	σ_2	σ_3	σ_2	σ_2
σ_3	σ_3	σ_3	σ_-	σ^*
σ^*	σ^*	σ^*	σ^*	σ^*
σ_-	σ_-	σ_-	σ_-	σ_-

The learners differ in state $(\langle\,\rangle, \sigma_0)$. When the teacher gives the example z_1, the learner L^1 chooses $(\langle\,\rangle, \sigma_1)$, and L^2 chooses $(\langle\,\rangle, \sigma_2)$. Afterwards both learners receive example z_2, leave their respective state, and enter the state $(\langle\,\rangle, \sigma_3)$. Now L^1 needs example y_1 and L^2 needs example y_2 to reach $(\langle\,\rangle, \sigma^*)$, otherwise the learners enter the failure state $(\langle\,\rangle, \sigma_-)$.

When a teacher receives as feedback the state $(\langle\,\rangle, \sigma_3)$, it cannot distinguish between L^1 and L^2. Thus whatever example the teacher gives next, one of the two learners will fail; for example, if we define $T(\langle\,\rangle, \sigma_3) = y_1$ then L^2 enters the failure state.

But when a history-aware teacher receives as feedback the state $(\langle\,\rangle, \sigma_3)$, it also knows the teaching history and therefore knows whether the learner was in $(\langle\,\rangle, \sigma_1)$ or $(\langle\,\rangle, \sigma_2)$ before. Accordingly the teacher can give y_1 or y_2. Formally,

$$\begin{aligned} T(\quad \langle\,\rangle, \quad \sigma_3, \quad \langle\,(\langle\,\rangle, \mathit{init}),\ (\langle\,\rangle, \sigma_0),\ (\langle\,\rangle, \sigma_1)\,\rangle \quad) &= y_1, \\ T(\quad \langle\,\rangle, \quad \sigma_3, \quad \langle\,(\langle\,\rangle, \mathit{init}),\ (\langle\,\rangle, \sigma_0),\ (\langle\,\rangle, \sigma_2)\,\rangle \quad) &= y_2. \end{aligned}$$

This would lead both learners to the target state. Thus only a history-aware teacher can make both learners reach the target. ■

The reason why both learners in Example 2.6 could be taught by a history-aware teacher is that part of their behavior (in state $(\langle\,\rangle, \sigma_3)$) could be inferred from another part of their behavior (in state $(\langle\,\rangle, \sigma_0)$) that was observed previously.

Since we, for simplicity, do not want to consider history-aware teachers, we have two alternatives. First we could simply ignore history-aware teachers even if they might be superior to plain teachers with feedback. Then we sometimes had to settle for suboptimal teachers. The second alternative, and the one we choose, is to model the learners in such a way that "history awareness effects" do not occur. In other words, in our framework the optimal teacher with feedback and the optimal history-aware teacher are equally good (with respect to the performance measure defined later).

The advantage of history awareness disappears when learners act in each state independently from their actions in the other states. We achieve this by letting the learner choose non-deterministically in every state from a set of allowed follow-up states. In other words, the learner is a *non-deterministic* automaton.

Definition 2.7 *Let $\langle \varrho, \mathcal{C}\rangle$ with $\varrho\colon \Sigma^* \times X \to \{0, 1, \uparrow\}$ be a represented concept class. A* non-deterministic learner using $\langle \varrho, \mathcal{C}\rangle$ *is a function*

$$\mathcal{L}\colon \mathcal{X}^* \times (\varrho(\mathcal{C}) \cup \{\mathit{init}\}) \times \mathcal{X} \to 2^{\mathcal{X}^*} \times 2^{\varrho(\mathcal{C})\cup\{\mathit{init}\}}$$

with component functions

$$\begin{aligned} \mathcal{L}_{\mathit{mem}}\colon \mathcal{X}^* \times (\varrho(\mathcal{C}) \cup \{\mathit{init}\}) \times \mathcal{X} &\to 2^{\mathcal{X}^*}, \\ \mathcal{L}_{\mathit{hyp}}\colon \mathcal{X}^* \times (\varrho(\mathcal{C}) \cup \{\mathit{init}\}) \times \mathcal{X} &\to 2^{\varrho(\mathcal{C})\cup\{\mathit{init}\}}. \end{aligned}$$

Our definition of non-deterministic learner is not as general as it could be. A more general learner could have the codomain $2^{\mathcal{X}^*\times(\varrho(\mathcal{C})\cup\{\mathit{init}\})}$. With our definition, however, we can separate the behavior into the two component functions $\mathcal{L}_{\mathit{mem}}$ and $\mathcal{L}_{\mathit{hyp}}$. This is purely for convenience as it simplifies the definitions of concrete learners later. Nevertheless, all results in this section hold for the more general definition of non-deterministic learner as well.

Using a non-deterministic learner instead of a set of deterministic learners allows us to view the teaching process as a game played by the learner and the teacher on a directed graph. Consider a directed graph whose (possibly infinitely many) nodes are all states a learner can assume, that is, $\mathcal{X}^* \times (\varrho(\mathcal{C}) \cup \{\mathit{init}\})$. An arc labeled with an example $z \in \mathcal{X}$ runs from a state $(\boldsymbol{s}, \sigma)$ to a state $(\boldsymbol{s}', \sigma')$ if and only if $(\boldsymbol{s}', \sigma') \in \mathcal{L}(\boldsymbol{s}, \sigma, z)$ (see Figure 2.2 for an example of such a graph).

The learner starts at the node labeled $(\langle\,\rangle, \mathit{init})$. During the teaching process the learner walks through the graph and the teacher tries to influence its path by giving examples. More precisely, the teacher's moves consist in giving an example $z \in \mathcal{X}$ and the learner's moves consist in choosing the next node by following one of the arcs labeled with z. The teacher's goal is to lead the learner to a target state from $\mathcal{X}^* \times (\varrho(\mathcal{C}) \cup \{\mathit{init}\})$. The learner's goal is to avoid this. The target concept is teachable if and only if there is a winning strategy for the teacher.

Example 2.8 We consider the non-deterministic learner $\mathcal{L}$ that results from "merging" the two deterministic learners in Example 2.6. The definition of $\mathcal{L}$ is based on the following table which describes the possible follow-up hypotheses $f(\sigma, z)$ for all hypotheses σ and examples z.

f	z_1	z_2	y_1	y_2
init	$\{\sigma_0\}$	$\{\sigma_0\}$	$\{\sigma_0\}$	$\{\sigma_0\}$
σ_0	$\{\sigma_1, \sigma_2\}$	$\{\sigma_0\}$	$\{\sigma_0\}$	$\{\sigma_0\}$
σ_1	$\{\sigma_1\}$	$\{\sigma_1, \sigma_3\}$	$\{\sigma_1\}$	$\{\sigma_1\}$
σ_2	$\{\sigma_2\}$	$\{\sigma_2, \sigma_3\}$	$\{\sigma_2\}$	$\{\sigma_2\}$
σ_3	$\{\sigma_3\}$	$\{\sigma_3\}$	$\{\sigma_-, \sigma^*\}$	$\{\sigma_-, \sigma^*\}$
σ^*	$\{\sigma^*\}$	$\{\sigma^*\}$	$\{\sigma^*\}$	$\{\sigma^*\}$
σ_-	$\{\sigma_-\}$	$\{\sigma_-\}$	$\{\sigma_-\}$	$\{\sigma_-\}$

Formally, $\mathcal{L}$ is then defined by $\mathcal{L}(\boldsymbol{s}, \sigma, z) = \{\langle\,\rangle\} \times f(\sigma, z)$. Figure 2.2 above shows the possible state transitions of $\mathcal{L}$. The target is not teachable with feedback to $\mathcal{L}$ because in $(\langle\,\rangle, \sigma_3)$, no matter what example a teacher would give, there is always an arc leading to $(\langle\,\rangle, \sigma_-)$ or not causing a state change at all. ■

Example 2.9 Let us look at a situation in which teaching is possible. Let $\mathcal{L}$ be a learner using hypothesis space $\langle \varrho, \mathcal{S}_n \rangle$. We define for $\sigma \neq \mathit{init}$:

$$\mathcal{L}(\boldsymbol{s}, \sigma, z) = \begin{cases} \{\boldsymbol{s} \circ \langle z \rangle\} \times \varrho(\mathcal{S}_n(\boldsymbol{s} \circ \langle z \rangle)) & \text{if } z = (\sigma, 0) \text{ or } z = (\sigma, 1), \\ \{\boldsymbol{s}\} \times \varrho(\mathcal{S}_n(\boldsymbol{s})) & \text{otherwise;} \end{cases}$$

and for the initial state: $\mathcal{L}(\langle\,\rangle, \mathit{init}, z) = \{\langle z \rangle\} \times \varrho(\mathcal{S}_n(\{z\}))$. This learner maintains a consistent hypothesis, but memorizes the new example z only if it in some sense "matches" the current hypothesis.

Let the target be $c^* = c_0 = [1, n]$. A teacher receiving feedback can always choose a "matching" example and thus make the learner memorize a new example in every round:

$$T(\boldsymbol{s}, \sigma) = \begin{cases} (1,1) & \text{if } \sigma = \mathit{init}, \\ (\sigma, 1) & \text{otherwise.} \end{cases}$$

The learner $\mathcal{L}$ memorizes all examples given by the teacher T. Moreover, if an example $(i, 1)$ is in the memory, the learner cannot hypothesize c_i any more because the concept c_i is inconsistent with the example $(i, 1)$. Therefore every example is given only once and after n rounds the learner knows all examples for c^*. As the hypothesis is always consistent with the memory, the learner hypothesizes the target after n rounds. ■

Sometimes it is helpful to think of a non-deterministic learner as many deterministic learners combined. All these learners act in every state independently from one another and from the other states. In informal explanations we shall often speak of the non-deterministic learner as if there actually were many deterministic learners. The fundamental difference to a set of "real" deterministic learners appears when the same state is reached a second time during the teaching process. A "real" deterministic learner acts the same way as in the first time, the non-deterministic learner may or may not act the same way. The following notation is inspired by the interpretation of $\mathcal{L}$ as a set of learners.

Definition 2.10 *Let $\langle \varrho, \mathcal{C} \rangle$ with $\varrho\colon \Sigma^* \times X \to \{0, 1, \uparrow\}$ be a represented concept class. Let $\mathcal{L}$ be a non-deterministic learner using $\langle \varrho, \mathcal{C} \rangle$. A* deterministic learner for $\mathcal{L}$ *is a deterministic learner L such that*

$$L(\boldsymbol{s}, \sigma, z) \in \mathcal{L}(\boldsymbol{s}, \sigma, z)$$

for all $(\boldsymbol{s}, \sigma, z)$. We write $L \in \mathcal{L}$ to denote that L is a deterministic learner for $\mathcal{L}$.

If all computations of a non-deterministic learner reach a certain state, for example, the target state, then also all deterministic learners reach that state. We shall use the following contrapositive of this fact implicitly in some proofs.

Lemma 2.11 *Let $\langle \varrho, \mathcal{C} \rangle$ with $\varrho\colon \Sigma^* \times X \to \{0, 1, \uparrow\}$ be a represented concept class. Let $\mathcal{L}$ be a non-deterministic learner using $\langle \varrho, \mathcal{C} \rangle$ and let $L \in \mathcal{L}$. Let $\boldsymbol{y} \in \mathcal{X}^*$ be a finite sequence of examples.*

If L, after having received $\boldsymbol{y}$, is not in state $(\boldsymbol{s}', \sigma')$ then not all computations of $\mathcal{L}$ on $\boldsymbol{y}$ end in state $(\boldsymbol{s}', \sigma')$.

In order to prove that non-deterministic learners can be taught equally fast no matter whether or not the teacher takes the history into account, we first have to formalize the

notions of teaching success and the performance measure for both teachers, history-aware and with feedback. We thus postpone the proof until after the formalization (see Lemma 2.18).

A teacher $T\colon \mathcal{X}^* \times (\varrho(\mathcal{C}) \cup \{\mathit{init}\}) \to \mathcal{X}(c^*)$ with feedback and a non-deterministic learner $\mathcal{L}$ determine the possible sequences of states assumed by $\mathcal{L}$ during teaching. For generality we introduce notation that allows the learner to start from any state, not only from the initial state. The set of all possible states of $\mathcal{L}$, starting in state $(\boldsymbol{s}, \sigma)$, after t rounds of interaction with T is denoted by $\mathcal{L}(\boldsymbol{s}, \sigma; T, t)$. Initially we have

$$\mathcal{L}(\boldsymbol{s}, \sigma; T, 0) = \{(\boldsymbol{s}, \sigma)\}.$$

The possible states after $t+1$ rounds are all the follow-up states of all the states after t rounds:

$$\mathcal{L}(\boldsymbol{s}, \sigma; T, t+1) = \bigcup_{\substack{(\boldsymbol{s}', \sigma') \\ \in \mathcal{L}(\boldsymbol{s}, \sigma; T, t)}} \mathcal{L}(\boldsymbol{s}', \sigma', T(\boldsymbol{s}', \sigma')).$$

We denote the set of all hypotheses the learner can assume after t rounds by

$$\mathcal{L}_{hyp}(\boldsymbol{s}, \sigma; T, t) = \{\sigma' \mid \exists \boldsymbol{s}' : (\boldsymbol{s}', \sigma') \in \mathcal{L}(\boldsymbol{s}, \sigma; T, t)\}.$$

When history-aware teachers are involved, the state of the teaching process is not described by the state of the learner alone. Rather we have to add another component, the history. We call the resulting states *history states* $(\boldsymbol{s}, \sigma, h) \in \mathcal{X}^* \times (\varrho(\mathcal{C}) \cup \{\mathit{init}\}) \times \mathit{HIST}$. The initial history is empty, thus teaching starts in history state $(\langle\,\rangle, \mathit{init}, \langle\,\rangle)$. The set of all history states of the teaching process involving T and $\mathcal{L}$ and starting in history state $(\boldsymbol{s}, \sigma, h)$ is denoted by $\mathfrak{L}(\mathcal{L}, \boldsymbol{s}, \sigma, h; T, t)$. Then analogous to non-history-aware teachers (but more complicated) we have:

$$\begin{aligned}
\mathfrak{L}(\mathcal{L}, \boldsymbol{s}, \sigma, h; T, 0) &= \{(\boldsymbol{s}, \sigma, h)\}, \\
\mathfrak{L}(\mathcal{L}, \boldsymbol{s}, \sigma, h; T, t+1) &= \bigcup_{\substack{(\boldsymbol{s}', \sigma', h') \\ \in \mathfrak{L}(\mathcal{L}, \boldsymbol{s}, \sigma, h; T, t)}} \mathcal{L}(\boldsymbol{s}', \sigma', T(\boldsymbol{s}', \sigma', h')) \times \{h \circ \langle \boldsymbol{s}', \sigma', T(\boldsymbol{s}', \sigma', h')\rangle\}.
\end{aligned}$$

We have now defined the notions of learner and teacher and have formalized the teaching process. Given a non-deterministic learner $\mathcal{L}$, the first question is whether there is a teacher that makes $\mathcal{L}$ eventually hypothesize the target. For the sake of precision we now formally define the notions of "teachability" that we consider in this thesis. However, most of the time an intuitive understanding will suffice.

Definition 2.12 *Let $\langle \varrho, \mathcal{C}\rangle$ with $\varrho\colon \Sigma^* \times X \to \{0, 1, \uparrow\}$ be a represented concept class and let $\mathcal{L}$ be a non-deterministic learner using $\langle \varrho, \mathcal{C}\rangle$. Let $c^* \in \mathcal{C}$ be a target concept. We*

say c^* is teachable with feedback to $\mathcal{L}$ *iff there is a teacher* $T\colon \mathcal{X}^* \times (\varrho(\mathcal{C}) \cup \{init\}) \to \mathcal{X}(c^*)$ *and a* $t \in \mathbb{N}$ *such that*

$$\mathcal{L}_{hyp}(\langle\,\rangle, init; T, t) \subseteq \varrho(c^*).$$

Definition 2.13 *Let* $\langle \varrho, \mathcal{C}\rangle$ *with* $\varrho\colon \Sigma^* \times X \to \{0, 1, \uparrow\}$ *be a represented concept class and let* $\mathcal{L}$ *be a non-deterministic learner using* $\langle \varrho, \mathcal{C}\rangle$*. Let* $c^* \in \mathcal{C}$ *be a target concept. We say* c^* is teachable by a history-aware teacher to $\mathcal{L}$ *iff there is a history-aware teacher* T *and a* $t \in \mathbb{N}$ *such that*

$$\mathfrak{L}(\mathcal{L}, \langle\,\rangle, init, \langle\,\rangle; T, t) \subseteq \mathcal{X}^* \times \varrho(c^*) \times HIST.$$

Note that both previous definitions pose no constraints on rounds after round t. The idea here is that the teacher stops teaching immediately after realizing that all learners have reached the target. The state of the learners after round t is thus unimportant.

In general we call a represented concept class *teachable* if all its concepts are teachable. A teacher T satisfying one of the conditions in Definitions 2.12 or 2.13 is called *successful*.

The next step after determining that there is a successful teacher is to find the best one. The performance of the best teacher is a measure for the *teachability* of the target concept, in the same way as the complexity of the most efficient algorithm solving a problem is a measure for that problem's complexity. It is therefore necessary to be able to measure the quality of successful teachers. Roughly speaking, we always measure the teacher's quality by the maximum number of rounds necessary to lead the learner to the target. The maximum is taken over all non-deterministic choices of the learner. We call the minimum number of rounds until one of the above teachability criteria holds the *teaching time*.

Definition 2.14 *Let* $\langle \varrho, \mathcal{C}\rangle$ *with* $\varrho\colon \Sigma^* \times X \to \{0, 1, \uparrow\}$ *be a represented concept class and* $c^* \in \mathcal{C}$ *a target. Let* $\mathcal{L}$ *be a non-deterministic learner using* $\langle \varrho, \mathcal{C}\rangle$ *and let* T *be a teacher with feedback. Let* $(\boldsymbol{s}, \sigma)$ *be a state. The* teaching time of T for c^* to $\mathcal{L}$ starting in $(\boldsymbol{s}, \sigma)$ *is*

$$\tau_T(\boldsymbol{s}, \sigma) = \min\{t \in \mathbb{N} \mid \mathcal{L}_{hyp}(\boldsymbol{s}, \sigma; T, t) \subseteq \varrho(c^*)\}$$

and the optimal teaching time for c^* to $\mathcal{L}$ *is*

$$\tau(\boldsymbol{s}, \sigma) = \min_T \tau_T(\boldsymbol{s}, \sigma)$$

where T *ranges over all teachers* $T\colon \mathcal{X}^* \times (\varrho(\mathcal{C}) \cup \{init\}) \to \mathcal{X}(c^*)$.

The teaching time $\tau_T(\boldsymbol{s}, \sigma)$ is infinite if the teacher cannot force all computations of the non-deterministic learner starting in $(\boldsymbol{s}, \sigma)$ into a target state. In particular, $\tau_T(\langle\,\rangle, init) = \infty$ means that T is not a successful teacher. Analogously, $\tau(\langle\,\rangle, init) = \infty$ means that there is no successful teacher.

The teaching times are defined individually for every state $(\boldsymbol{s}, \sigma)$, but there is a relation between a state's teaching time and that of its follow-up states. For a learner $\mathcal{L}$ and a teacher with feedback T and a state $(\boldsymbol{s}, \sigma)$ with $\tau(\boldsymbol{s}, \sigma) > 0$ we have

$$\begin{aligned}\tau_T(\boldsymbol{s}, \sigma) &= 1 + \min\{t \mid \forall\ (\boldsymbol{s}', \sigma') \in \mathcal{L}(\boldsymbol{s}, \sigma, T(\boldsymbol{s}, \sigma)) : \mathcal{L}_{hyp}(\boldsymbol{s}', \sigma'; T, t) \subseteq \varrho(c^*)\} \\ &= 1 + \max_{\substack{(\boldsymbol{s}', \sigma') \\ \in \mathcal{L}(\boldsymbol{s}, \sigma, T(\boldsymbol{s}, \sigma))}} \min\{t \mid \mathcal{L}_{hyp}(\boldsymbol{s}', \sigma'; T, t) \subseteq \varrho(c^*)\} \\ &= 1 + \max_{\substack{(\boldsymbol{s}', \sigma') \\ \in \mathcal{L}(\boldsymbol{s}, \sigma, T(\boldsymbol{s}, \sigma))}} \tau_T(\boldsymbol{s}', \sigma').\end{aligned} \tag{2.2}$$

A relation like (2.2) holds also for optimal teaching times, as we prove next.

Lemma 2.15 *Let $\langle \varrho, \mathcal{C} \rangle$ with $\varrho\colon \Sigma^* \times X \to \{0, 1, \uparrow\}$ be a represented concept class and $c^* \in \mathcal{C}$ a target. For all states $(\boldsymbol{s}, \sigma) \in \mathcal{X}^* \times (\varrho(\mathcal{C}) \cup \{init\})$ with $\tau(\boldsymbol{s}, \sigma) > 0$:*

$$\tau(\boldsymbol{s}, \sigma) = 1 + \min_{z \in \mathcal{X}} \max_{\substack{(\boldsymbol{s}', \sigma') \\ \in \mathcal{L}(\boldsymbol{s}, \sigma, z)}} \tau(\boldsymbol{s}', \sigma').$$

Proof. We introduce for all states $(\boldsymbol{s}, \sigma)$ the notation $\eta(\boldsymbol{s}, \sigma)$ for the set

$$\eta(\boldsymbol{s}, \sigma) = \operatorname*{argmin}_{z \in \mathcal{X}} \max_{\substack{(\boldsymbol{s}', \sigma') \\ \in \mathcal{L}(\boldsymbol{s}, \sigma, z)}} \tau(\boldsymbol{s}', \sigma') \quad \subseteq\ \mathcal{X}.$$

Intuitively, $\eta(\boldsymbol{s}, \sigma)$ contains the best examples one can give to a learner in state $(\boldsymbol{s}, \sigma)$.

Claim. For all $r \in \mathbb{N}$:

(a) for all $(\boldsymbol{s}, \sigma)$ with $1 \leq \tau(\boldsymbol{s}, \sigma) \leq r + 1$:

$$\tau(\boldsymbol{s}, \sigma) = 1 + \min_{z \in \mathcal{X}} \max_{\substack{(\boldsymbol{s}', \sigma') \\ \in \mathcal{L}(\boldsymbol{s}, \sigma, z)}} \tau(\boldsymbol{s}', \sigma').$$

(b) for all $(\boldsymbol{s}, \sigma)$ with $\tau(\boldsymbol{s}, \sigma) \leq r$: Let $\widetilde{T}$ be a teacher with $\widetilde{T}(\boldsymbol{s}', \sigma') \in \eta(\boldsymbol{s}', \sigma')$ for all $(\boldsymbol{s}', \sigma')$ with $\tau(\boldsymbol{s}', \sigma') < r$ and for $(\boldsymbol{s}', \sigma') = (\boldsymbol{s}, \sigma)$. Then $\tau(\boldsymbol{s}, \sigma) = \tau_{\widetilde{T}}(\boldsymbol{s}, \sigma)$.

Proof. The proof is by induction on r. For the induction basis let $r = 0$. The only states with $\tau(\boldsymbol{s}, \sigma) = 0$ are the target states, for which $\sigma \in \varrho(c^*)$. Since $\tau_T(\boldsymbol{s}, \sigma) = 0$ for every teacher T, Item (b) holds. To show Item (a), let $\tau(\boldsymbol{s}, \sigma) = 1$. Then there is a teacher T such that $\min\{t \mid \mathcal{L}_{hyp}(\boldsymbol{s}, \sigma; T, t) \subseteq \varrho(c^*)\} = \tau_T(\boldsymbol{s}, \sigma) = \tau(\boldsymbol{s}, \sigma) = 1$. Therefore, $\mathcal{L}_{hyp}(\boldsymbol{s}, \sigma; T, 1) \subseteq \varrho(c^*)$ and thus for all $(\boldsymbol{s}', \sigma') \in \mathcal{L}(\boldsymbol{s}, \sigma, T(\boldsymbol{s}, \sigma))$ we have $\tau(\boldsymbol{s}', \sigma') = 0$. It follows that

$$\max_{\substack{(\boldsymbol{s}', \sigma') \\ \in \mathcal{L}(\boldsymbol{s}, \sigma, T(\boldsymbol{s}, \sigma))}} \tau(\boldsymbol{s}', \sigma') = 0$$

and hence

$$\min_{z\in\mathcal{X}} \max_{\substack{(\boldsymbol{s}',\sigma')\\ \in\mathcal{L}(\boldsymbol{s},\sigma,z)}} \tau(\boldsymbol{s}',\sigma') = 0.$$

For the induction step, assume that (a) and (b) hold for some $r \geq 0$. We first show (b) for $r+1$. Let $\tau(\boldsymbol{s},\sigma) = r+1$ and let $\widetilde{T}$ be a teacher with $\widetilde{T}(\boldsymbol{s}',\sigma') \in \eta(\boldsymbol{s}',\sigma')$ for all $(\boldsymbol{s}',\sigma')$ with $\tau(\boldsymbol{s}',\sigma') < r+1$ and for $(\boldsymbol{s}',\sigma') = (\boldsymbol{s},\sigma)$. We have to show $\tau_{\widetilde{T}}(\boldsymbol{s},\sigma) = \tau(\boldsymbol{s},\sigma) = r+1$. Since $\widetilde{T}(\boldsymbol{s},\sigma) \in \eta(\boldsymbol{s},\sigma)$ we know from Item (a) of the induction hypothesis that $r+1 = \tau(\boldsymbol{s},\sigma) = 1 + \max_{(\boldsymbol{s}',\sigma')\in\mathcal{L}(\boldsymbol{s},\sigma,\widetilde{T}(\boldsymbol{s},\sigma))} \tau(\boldsymbol{s}',\sigma')$. Thus all $\tau(\boldsymbol{s}',\sigma')$ in the last expression are at most r and by Item (b) of the induction hypothesis, $\tau(\boldsymbol{s}',\sigma') = \tau_{\widetilde{T}}(\boldsymbol{s}',\sigma')$. This yields

$$r+1 = \tau(\boldsymbol{s},\sigma) = 1 + \max_{\substack{(\boldsymbol{s}',\sigma')\\ \in\mathcal{L}(\boldsymbol{s},\sigma,\widetilde{T}(\boldsymbol{s},\sigma))}} \tau(\boldsymbol{s}',\sigma') = 1 + \max_{\substack{(\boldsymbol{s}',\sigma')\\ \in\mathcal{L}(\boldsymbol{s},\sigma,\widetilde{T}(\boldsymbol{s},\sigma))}} \tau_{\widetilde{T}}(\boldsymbol{s}',\sigma') = \tau_{\widetilde{T}}(\boldsymbol{s},\sigma).$$

Now we show the induction step for (a). Let $\tau(\boldsymbol{s},\sigma) = r+2$. Then there is a teacher T such that

$$\begin{aligned} r+2 = \tau(\boldsymbol{s},\sigma) = \tau_T(\boldsymbol{s},\sigma) &= 1 + \max_{\substack{(\boldsymbol{s}',\sigma')\\ \in\mathcal{L}(\boldsymbol{s},\sigma,T(\boldsymbol{s},\sigma))}} \tau_T(\boldsymbol{s}',\sigma') \\ &\geq 1 + \max_{\substack{(\boldsymbol{s}',\sigma')\\ \in\mathcal{L}(\boldsymbol{s},\sigma,T(\boldsymbol{s},\sigma))}} \tau(\boldsymbol{s}',\sigma') \geq 1 + \min_{z\in\mathcal{X}} \max_{\substack{(\boldsymbol{s}',\sigma')\\ \in\mathcal{L}(\boldsymbol{s},\sigma,z)}} \tau(\boldsymbol{s}',\sigma'). \end{aligned}$$

It remains to show the "$\leq$" part, $\tau(\boldsymbol{s},\sigma) \leq 1 + \min_{z\in\mathcal{X}} \max_{(\boldsymbol{s}',\sigma')\in\mathcal{L}(\boldsymbol{s},\sigma,z)} \tau(\boldsymbol{s}',\sigma')$. For a contradiction, suppose $\tau(\boldsymbol{s},\sigma) > 1 + \min_{z\in\mathcal{X}} \max_{(\boldsymbol{s}',\sigma')\in\mathcal{L}(\boldsymbol{s},\sigma,z)} \tau(\boldsymbol{s}',\sigma')$. Then there is an example $y \in \eta(\boldsymbol{s},\sigma)$ such that for all $(\boldsymbol{s}',\sigma') \in \mathcal{L}(\boldsymbol{s},\sigma,y)$: $1 + \tau(\boldsymbol{s}',\sigma') < \tau(\boldsymbol{s},\sigma) = r+2$, that is, $\tau(\boldsymbol{s}',\sigma') < r+1$.

Let $\widetilde{T}$ be a teacher with $\widetilde{T}(\boldsymbol{s}',\sigma') \in \eta(\boldsymbol{s}',\sigma')$ for all $(\boldsymbol{s}',\sigma')$ with $\tau(\boldsymbol{s}',\sigma') < r+1$ and $\widetilde{T}(\boldsymbol{s},\sigma) = y$. Then

$$\tau_{\widetilde{T}}(\boldsymbol{s},\sigma) = 1 + \max_{\substack{(\boldsymbol{s}',\sigma')\\ \in\mathcal{L}(\boldsymbol{s},\sigma,\widetilde{T}(\boldsymbol{s},\sigma))}} \tau_{\widetilde{T}}(\boldsymbol{s}',\sigma') = 1 + \max_{\substack{(\boldsymbol{s}',\sigma')\\ \in\mathcal{L}(\boldsymbol{s},\sigma,y)}} \tau_{\widetilde{T}}(\boldsymbol{s}',\sigma'). \tag{2.3}$$

Since $\tau(\boldsymbol{s}',\sigma') \leq r$ for all $(\boldsymbol{s}',\sigma') \in \mathcal{L}(\boldsymbol{s},\sigma,y)$, we can apply Item (b) of the induction step and get $\tau_{\widetilde{T}}(\boldsymbol{s}',\sigma') = \tau(\boldsymbol{s}',\sigma') \leq r$. Continuing Equation (2.3) we obtain

$$\tau(\boldsymbol{s},\sigma) \leq \tau_{\widetilde{T}}(\boldsymbol{s},\sigma) = 1 + \max_{\substack{(\boldsymbol{s}',\sigma')\\ \in\mathcal{L}(\boldsymbol{s},\sigma,y)}} \tau_{\widetilde{T}}(\boldsymbol{s}',\sigma') \leq 1 + r,$$

a contradiction to $\tau(\boldsymbol{s},\sigma) = r+2$. Thus the assumption is wrong and the "$\leq$" part proved. □ Claim

The statement of the lemma is just Item (b) of the claim. □

Much in the same way as in Definition 2.14, teaching times can be defined for history-aware teachers, too.

Definition 2.16 *Let $\langle \varrho, \mathcal{C} \rangle$ with $\varrho\colon \Sigma^* \times X \to \{0, 1, \uparrow\}$ be a represented concept class and $c^* \in \mathcal{C}$ a target. Let $\mathcal{L}$ be a non-deterministic learner using $\langle \varrho, \mathcal{C} \rangle$ and let T be a history-aware teacher. Let $(\boldsymbol{s}, \sigma, h)$ be a history state. The* teaching time of T for c^* to $\mathcal{L}$ starting in $(\boldsymbol{s}, \sigma, h)$ *is*

$$\tau_T(\boldsymbol{s}, \sigma, h) = \min\{t \in \mathbb{N} \mid \mathfrak{L}(\mathcal{L}, \boldsymbol{s}, \sigma, h; T, t) \subseteq \mathcal{X}^* \times \varrho(c^*) \times \mathit{HIST}\}.$$

and the optimal teaching time for c^* to $\mathcal{L}$ *is*

$$\tau(\boldsymbol{s}, \sigma, h) = \min_T \tau_T(\boldsymbol{s}, \sigma, h)$$

where T ranges over all history-aware teachers.

The remarks on Page 30 about infinite teaching times apply accordingly for history-aware teachers. The analog of Lemma 2.15 for history-aware teachers is the following.

Lemma 2.17 *Let $\langle \varrho, \mathcal{C} \rangle$ with $\varrho\colon \Sigma^* \times X \to \{0, 1, \uparrow\}$ be a represented concept class and $c^* \in \mathcal{C}$ a target. For all states $(\boldsymbol{s}, \sigma) \in \mathcal{X}^* \times (\varrho(\mathcal{C}) \cup \{\mathit{init}\})$ and histories $h \in \mathit{HIST}$ with $\tau(\boldsymbol{s}, \sigma, h) > 0$:*

$$\tau(\boldsymbol{s}, \sigma, h) = 1 + \min_{z \in \mathcal{X}} \max_{\substack{(\boldsymbol{s}', \sigma') \\ \in \mathcal{L}(\boldsymbol{s}, \sigma, z)}} \tau(\boldsymbol{s}', \sigma', h \circ \langle (\boldsymbol{s}, \sigma, z) \rangle).$$

Proof. Let $(\boldsymbol{s}, \sigma, h)$ be such that $\tau(\boldsymbol{s}, \sigma, h) > 0$.

We start by showing the "$\leq$" part of the statement. Let $y \in \mathcal{X}$ be an example such that

$$\max_{\substack{(\boldsymbol{s}', \sigma') \\ \in \mathcal{L}(\boldsymbol{s}, \sigma, y)}} \tau(\boldsymbol{s}', \sigma', h \circ \langle (\boldsymbol{s}, \sigma, y) \rangle)$$

is minimal. We denote the follow-up history by $h' = h \circ \langle (\boldsymbol{s}, \sigma, y) \rangle$ and all follow-up history states in $\mathcal{L}(\boldsymbol{s}, \sigma, y)$ by $(\boldsymbol{s}_1, \sigma_1, h'), (\boldsymbol{s}_2, \sigma_2, h'), \ldots$ (there are only countably many). For each follow-up history state $(\boldsymbol{s}_i, \sigma_i, h')$ there is a history-aware teacher T_i with $\tau_{T_i}(\boldsymbol{s}_i, \sigma_i, h') = \tau(\boldsymbol{s}_i, \sigma_i, h')$. Every teacher T_i gives in $(\boldsymbol{s}_i, \sigma_i, h')$ the example $y_i = T_i(\boldsymbol{s}_i, \sigma_i, h')$, which results in new histories $h'_i = h' \circ \langle (\boldsymbol{s}_i, \sigma_i, h'_i) \rangle$. Now it suffices to show

$$\tau(\boldsymbol{s}, \sigma, h) \leq 1 + \max_i \tau(\boldsymbol{s}_i, \sigma_i, h').$$

To show this it suffices to specify a history-aware teacher T such that

$$\tau_T(\boldsymbol{s}, \sigma, h) \leq 1 + \max_i \tau(\boldsymbol{s}_i, \sigma_i, h').$$

We claim that the following teacher satisfies this inequality:

$$T(\hat{\boldsymbol{s}},\hat{\sigma},\hat{h}) = \begin{cases} T_i(\hat{\boldsymbol{s}},\hat{\sigma},\hat{h}) & \text{if } \exists i : \hat{h} \sqsupseteq h'_i \quad \text{or} \quad (\hat{\boldsymbol{s}},\hat{\sigma},\hat{h}) = (\boldsymbol{s}_i,\sigma_i,h'), \\ y & \text{if } (\hat{\boldsymbol{s}},\hat{\sigma},\hat{h}) = (\boldsymbol{s},\sigma,h), \\ \text{arbitrary} & \text{otherwise.} \end{cases}$$

Roughly speaking, T is composed of all optimal teachers for all possible follow-up history states of $(\boldsymbol{s},\sigma,h)$. The teacher T is well-defined as there can never by two different i's satisfying the first condition.

To show that T works, let $(\boldsymbol{s}_i,\sigma_i,h') \in \mathcal{L}(\boldsymbol{s},\sigma,y) = \mathcal{L}(\boldsymbol{s},\sigma,T(\boldsymbol{s},\sigma,h))$. Then for all further history states that learner $\mathcal{L}$ reaches under T, T is defined identically to T_i. Therefore $\tau_T(\boldsymbol{s},\sigma,h) = 1 + \max_i \tau_T(\boldsymbol{s}_i,\sigma_i,h') = 1 + \max_i \tau_{T_i}(\boldsymbol{s}_i,\sigma_i,h')$. By definition of T_i we have $\tau_{T_i}(\boldsymbol{s}_i,\sigma_i,h') = \tau(\boldsymbol{s}_i,\sigma_i,h')$ and it follows

$$\tau(\boldsymbol{s},\sigma,h) \le \tau_T(\boldsymbol{s},\sigma,h) = 1 + \max_{\substack{(\boldsymbol{s}',\sigma') \\ \in\mathcal{L}(\boldsymbol{s},\sigma,y)}} \tau(\boldsymbol{s}',\sigma',h') = 1 + \min_{z\in\mathcal{X}} \max_{\substack{(\boldsymbol{s}',\sigma') \\ \in\mathcal{L}(\boldsymbol{s},\sigma,z)}} \tau(\boldsymbol{s}',\sigma',h'),$$

where the last equality is due to our choice of y. This proves the "$\le$" part.

It remains to show "$\ge$". Let T be a teacher such that $\tau_T(\boldsymbol{s},\sigma,h) = \tau(\boldsymbol{s},\sigma,h)$ and let $y = T(\boldsymbol{s},\sigma,h)$. After receiving y the history will be $h' = h \circ \langle(\boldsymbol{s},\sigma,y)\rangle$. Then

$$\begin{aligned} \tau(\boldsymbol{s},\sigma,h) = \tau_T(\boldsymbol{s},\sigma,h) &= 1 + \max_{\substack{(\boldsymbol{s}',\sigma') \\ \in\mathcal{L}(\boldsymbol{s},\sigma,y)}} \tau_T(\boldsymbol{s}',\sigma',h') \\ &\ge 1 + \max_{\substack{(\boldsymbol{s}',\sigma') \\ \in\mathcal{L}(\boldsymbol{s},\sigma,y)}} \tau(\boldsymbol{s}',\sigma',h') \ge 1 + \min_{z\in\mathcal{X}} \max_{\substack{(\boldsymbol{s}',\sigma') \\ \in\mathcal{L}(\boldsymbol{s},\sigma,z)}} \tau(\boldsymbol{s}',\sigma',h'). \end{aligned}$$

□

At last we are ready to show that non-history-aware teachers with feedback can do as well as history-aware teachers.

Lemma 2.18 *Let $\langle\varrho,\mathcal{C}\rangle$ be a represented concept class and let $\mathcal{L}$ be a non-deterministic learner using $\langle\varrho,\mathcal{C}\rangle$. Then a target $c^* \in \mathcal{C}$ is teachable by a history-aware teacher to $\mathcal{L}$ if and only if c^* is teachable by a teacher with feedback. If c^* is teachable, then*

$$\tau(\langle\,\rangle, \mathit{init}) = \tau(\langle\,\rangle, \mathit{init}, \langle\,\rangle).$$

Proof. Note that $\tau(\boldsymbol{s},\sigma,h) \le \tau(\boldsymbol{s},\sigma)$ for all states and histories, because a teacher with feedback is just a special case of a history-aware teacher.

The lemma is implied by the following claim.

Claim. For all $r \in \mathbb{N}$, all $(\boldsymbol{s},\sigma)$, and all $h \in \mathit{HIST}$:

$$\tau(\boldsymbol{s},\sigma) = r \quad \Leftrightarrow \quad \tau(\boldsymbol{s},\sigma,h) = r.$$

Proof. For the induction basis, let $r = 0$. Let $(\boldsymbol{s}, \sigma)$ be a state with $\tau(\boldsymbol{s}, \sigma) = 0$ and let h be an arbitrary history. Then $\sigma \in \varrho(c^*)$ and consequently $\tau(\boldsymbol{s}, \sigma, h) = 0$. On the other hand, if $\tau(\boldsymbol{s}, \sigma, h) = 0$ then also $\sigma \in \varrho(c^*)$ and $\tau(\boldsymbol{s}, \sigma) = 0$.

For the induction step, assume the claim holds for some $r \in \mathbb{N}$.

Let $(\boldsymbol{s}, \sigma)$ be a state with $\tau(\boldsymbol{s}, \sigma) = r + 1$ and let $h \in \mathit{HIST}$. Then by Lemma 2.17,

$$r + 1 = \tau(\boldsymbol{s}, \sigma) \geq \tau(\boldsymbol{s}, \sigma, h) = 1 + \min_{z \in \mathcal{X}} \max_{\substack{(\boldsymbol{s}', \sigma') \\ \in \mathcal{L}(\boldsymbol{s}, \sigma, z)}} \tau(\boldsymbol{s}', \sigma', h \circ \langle (\boldsymbol{s}, \sigma, z) \rangle).$$

Let $y \in \mathcal{X}$ be such that

$$\tau(\boldsymbol{s}, \sigma, h) = 1 + \max_{\substack{(\boldsymbol{s}', \sigma') \\ \in \mathcal{L}(\boldsymbol{s}, \sigma, y)}} \tau(\boldsymbol{s}', \sigma', h \circ \langle (\boldsymbol{s}, \sigma, y) \rangle).$$

Then for all $(\boldsymbol{s}', \sigma') \in \mathcal{L}(\boldsymbol{s}, \sigma, y)$ we have $\tau(\boldsymbol{s}', \sigma', h \circ \langle (\boldsymbol{s}, \sigma, y) \rangle) \leq r$. To all these $(\boldsymbol{s}', \sigma')$ we can apply the induction hypothesis, which leads to $\tau(\boldsymbol{s}', \sigma') \leq r$ for all $(\boldsymbol{s}', \sigma') \in \mathcal{L}(\boldsymbol{s}, \sigma, y)$. For one of these states $(\boldsymbol{s}', \sigma')$ we have $\tau(\boldsymbol{s}', \sigma') = r$ (otherwise by Lemma 2.15, $\tau(\boldsymbol{s}, \sigma) \leq 1 + \max_{(\boldsymbol{s}', \sigma') \in \mathcal{L}(\boldsymbol{s}, \sigma, y)} \tau(\boldsymbol{s}', \sigma') < 1 + r$, a contradiction). Then by the induction hypothesis, for this state also $\tau(\boldsymbol{s}', \sigma', h \circ \langle (\boldsymbol{s}, \sigma, y) \rangle) = r$. It follows

$$\tau(\boldsymbol{s}, \sigma, h) = 1 + \max_{\substack{(\boldsymbol{s}', \sigma') \\ \in \mathcal{L}(\boldsymbol{s}, \sigma, y)}} \tau(\boldsymbol{s}', \sigma', h \circ \langle (\boldsymbol{s}, \sigma, y) \rangle) = 1 + r.$$

This proves $\tau(\boldsymbol{s}, \sigma) = r + 1 \quad \Longrightarrow \quad \tau(\boldsymbol{s}, \sigma, h) = r + 1$.

To show the other direction, let $\tau(\boldsymbol{s}, \sigma, h) = r + 1$. Then by Lemma 2.17 there is a $y \in \mathcal{X}$ such that

$$r + 1 = \tau(\boldsymbol{s}, \sigma, h) = 1 + \max_{\substack{(\boldsymbol{s}', \sigma') \\ \in \mathcal{L}(\boldsymbol{s}, \sigma, y)}} \tau(\boldsymbol{s}', \sigma', h')$$

with $h' = h \circ \langle \boldsymbol{s}, \sigma, y \rangle$. All τ-values in the maximization are at most r and we can apply the induction hypothesis to them. It follows

$$r + 1 = 1 + \max_{\substack{(\boldsymbol{s}', \sigma') \\ \in \mathcal{L}(\boldsymbol{s}, \sigma, y)}} \tau(\boldsymbol{s}', \sigma'). \tag{2.4}$$

We cannot apply Lemma 2.15 to (2.4) because we have not yet shown that $\tau(\boldsymbol{s}, \sigma)$ is finite. But we can also directly show that (2.4) equals $\tau(\boldsymbol{s}, \sigma)$. Let $\widetilde{T}$ be a teacher with $\widetilde{T}(\boldsymbol{s}', \sigma') \in \eta(\boldsymbol{s}', \sigma')$ for all $(\boldsymbol{s}', \sigma') \leq r$ and $\widetilde{T}(\boldsymbol{s}, \sigma) = y$. Then $\mathcal{L}(\boldsymbol{s}, \sigma; \widetilde{T}, 1) = \mathcal{L}(\boldsymbol{s}, \sigma, y)$. Now by the claim in the proof of Lemma 2.15, $\tau_{\widetilde{T}}(\boldsymbol{s}', \sigma') = \tau(\boldsymbol{s}', \sigma')$ for all $(\boldsymbol{s}', \sigma') \in \mathcal{L}(\boldsymbol{s}, \sigma, y)$. We then obtain

$$\tau_{\widetilde{T}}(\boldsymbol{s}, \sigma) = 1 + \max_{\substack{(\boldsymbol{s}', \sigma') \\ \in \mathcal{L}(\boldsymbol{s}, \sigma, y)}} \tau_{\widetilde{T}}(\boldsymbol{s}, \sigma) = 1 + \max_{\substack{(\boldsymbol{s}', \sigma') \\ \in \mathcal{L}(\boldsymbol{s}, \sigma, y)}} \tau(\boldsymbol{s}, \sigma) = r + 1.$$

Therefore $\tau(\boldsymbol{s},\sigma) \leq r+1$. Now if $\tau(\boldsymbol{s},\sigma) < r+1$ then $\tau(\boldsymbol{s},\sigma,h) < r+1$, a contradiction. Hence $\tau(\boldsymbol{s},\sigma,h) = r+1$. □ Claim

□

Thanks to Lemma 2.18 we do not have to consider history-aware teachers any more. We can confine the search for optimal teachers to the teachers with feedback. This concludes the formal framework for teachers with feedback. Now we turn to teaching without feedback again.

There is an intuition of the teaching process without feedback as a game played on a graph, similar to the game for teaching with feedback (see Page 27). The graph is the same in both games. But the new game is a one player game, the one player being the teacher. At the beginning of the game, only the node belonging to the initial state of the learner is marked. The teacher's move then consists in an example $z \in \mathcal{X}(c^*)$. The effect of such a move is that all nodes connected with the already marked nodes by an arc with label z are marked, and the old marks are removed. After round t exactly those nodes are marked whose corresponding states are in $\mathcal{L}(\langle\,\rangle, \mathit{init}; T; t)$. The teacher has won when eventually only target nodes are marked.

In the absence of feedback, a teacher does not necessarily know when to stop. Therefore teaching success can occur either finitely or in the limit. To make the distinction clearer, we interpret the non-deterministic learner as many single learners. *Finite* teaching means that all learners hypothesize the target concept after the teacher has given a finite sequence of examples. Teaching *in the limit* means that the teacher gives an infinite sequence of examples such that every learner eventually hypothesizes the target; but there need not be a certain round in which all learners have reached the target.

Recall that a teacher without feedback is a function $T\colon \mathbb{N} \to \mathcal{X}(c^*)$. To denote the possible states of the learner during the teaching process we use the same notation as for teachers with feedback:

$$\begin{aligned} \mathcal{L}(\boldsymbol{s},\sigma;T,0) &= \{(\boldsymbol{s},\sigma)\},\\ \mathcal{L}(\boldsymbol{s},\sigma;T,t+1) &= \bigcup_{\substack{(\boldsymbol{s}',\sigma')\\ \in\mathcal{L}(\boldsymbol{s},\sigma;T,t)}} \mathcal{L}(\boldsymbol{s}',\sigma',T(t)), \end{aligned}$$

and

$$\mathcal{L}_{hyp}(\boldsymbol{s},\sigma;T,t) = \{\sigma' \mid \exists \boldsymbol{s}' : (\boldsymbol{s}',\sigma') \in \mathcal{L}(\boldsymbol{s},\sigma;T,t)\}.$$

Definition 2.19 *Let $\langle\varrho,\mathcal{C}\rangle$ with $\varrho\colon \Sigma^* \times X \to \{0,1,\uparrow\}$ be a represented concept class and let $\mathcal{L}$ be a non-deterministic learner using $\langle\varrho,\mathcal{C}\rangle$. Let $c^* \in \mathcal{C}$ be a target concept. We say c^** is finitely teachable without feedback to $\mathcal{L}$ *iff there is a teacher $T\colon \mathbb{N} \to \mathcal{X}(c^*)$ and a $t \in \mathbb{N}$ such that*

$$\mathcal{L}_{hyp}(\langle\,\rangle, \mathit{init}; T, t) \subseteq \varrho(c^*).$$

Definition 2.20 *Let $\langle \varrho, \mathcal{C} \rangle$ with $\varrho\colon \Sigma^* \times X \to \{0, 1, \uparrow\}$ be a represented concept class and let $\mathcal{L}$ be a non-deterministic learner using $\langle \varrho, \mathcal{C} \rangle$. Let $c^* \in \mathcal{C}$ be a target concept. We say c^** is teachable in the limit without feedback to $\mathcal{L}$ *iff there is a teacher $T\colon \mathbb{N} \to \mathcal{X}(c^*)$ such that for every series $((\boldsymbol{s}_t, \sigma_t))_{t\in\mathbb{N}}$ with $(\boldsymbol{s}_0, \sigma_0) = (\langle\ \rangle, \mathit{init})$ and $(\boldsymbol{s}_{t+1}, \sigma_{t+1}) \in \mathcal{L}(\boldsymbol{s}_t, \sigma_t, T(t))$:*

$$\forall^\infty t :\ \sigma_t \in \varrho(\sigma).$$

The performance of a given teacher without feedback, again called its *teaching time*, is measured by the number of rounds needed to reach finite teaching success. Consequently, the teaching time may be infinite if the teacher is only successful in the limit. The teachability of a concept is again given by the optimal teaching time over all teachers without feedback.

Definition 2.21 *Let $\langle \varrho, \mathcal{C} \rangle$ with $\varrho\colon \Sigma^* \times X \to \{0, 1, \uparrow\}$ be a represented concept class and $c^* \in \mathcal{C}$ a target. Let $\mathcal{L}$ be a non-deterministic learner using $\langle \varrho, \mathcal{C} \rangle$ and let $T\colon \mathbb{N} \to \mathcal{X}(c^*)$ be a teacher without feedback. Let $(\boldsymbol{s}, \sigma)$ be a state. The* teaching time of T for c^* to $\mathcal{L}$ starting in $(\boldsymbol{s}, \sigma)$ *is*

$$\tau_T(\boldsymbol{s}, \sigma) = \min\{t \in \mathbb{N} \mid \mathcal{L}_{hyp}(\boldsymbol{s}, \sigma; T, t) \subseteq \varrho(c^*)\}$$

and the optimal teaching time for c^* to $\mathcal{L}$ *is*

$$\tau(\boldsymbol{s}, \sigma) = \min_T \tau_T(\boldsymbol{s}, \sigma)$$

where T ranges over all functions $T\colon \mathbb{N} \to \mathcal{X}(c^*)$.

Example 2.22 Again, we want to teach $c^* = [1, n]$ to a learner using $\langle \varrho, \mathcal{S}_n \rangle$ as hypothesis space. This time, the learner can in every round choose to perform either a dumb hypothesis change, in which the hypothesis depends on the size of the memory, or a smart one, in which the hypothesis is consistent with the memory.

$$\mathcal{L}(\boldsymbol{s}, \sigma, z) = \ \{\ \begin{array}{l} (\boldsymbol{s} \circ \langle z \rangle,\ \ |\boldsymbol{s} \circ \langle z \rangle| \bmod (n+1)), \\ (\boldsymbol{s} \circ \langle z \rangle,\ \ \min \varrho(\mathcal{S}_n(\boldsymbol{s} \circ \langle z \rangle))\quad) \end{array}\ \}$$

The target $c^* = [1, n]$ is finitely teachable to $\mathcal{L}$ using the teacher T with $T(i) = (1+i \bmod n, 1)$ for all $i \in \mathbb{N}$. Let $\boldsymbol{s}_{n+1}$ be the memory of the learner in round $n+1$, which is independent of the learner's hypothesis changes. Consider the possible hypotheses of the learner in round $n+1$: the hypothesis is either $|\boldsymbol{s}_{n+1}| \bmod (n+1) = (n+1) \bmod (n+1) = 0$ or $\min \varrho(\mathcal{S}_n(\boldsymbol{s}_{n+1})) = \min \varrho(c_0) = 0$. Thus, $\mathcal{L}$ is in a target state. The optimal teaching time is $\tau(\langle\ \rangle, \mathit{init}) = n + 1$, because it must be greater than n since after only n rounds the learner could hypothesize c_n if it has always chosen the dumb hypothesis change.

The target $c^* = [1, n]$ is not teachable in the limit because in the non-deterministic computation which always chooses $|\boldsymbol{s} \circ \langle z \rangle| \bmod (n+1)$ as next hypothesis a non-target hypothesis is reached infinitely often. ■

The previous example reveals an issue we have neglected so far, namely what happens *after* a learner has reached the target concept. In the variant with feedback, the teacher simply stops teaching when the learner arrives in a target state. In the variant without feedback, however, a learner carries on even after reaching the target and may even assume a different, wrong hypothesis again later.

One way to deal with that problem is to distinguish two interpretations for teaching success: "punctual" teaching success and "lasting", thereby doubling the number of model variants. Another way is to take care that learners do not "unlearn" the target once they have reached it. This can be done most easily by forbidding learners to change their hypothesis after receiving a consistent example. Such learners will never leave the target hypothesis because it is always consistent with any example for the target. This property is called *conservativeness*.

Definition 2.23 *Let $\langle \varrho, \mathcal{C} \rangle$ be a represented concept class. A non-deterministic learner $\mathcal{L} = (\mathcal{L}_{mem}, \mathcal{L}_{hyp})$ using the hypothesis space $\langle \varrho, \mathcal{C} \rangle$ is called* conservative *iff for all states $(\boldsymbol{s}, \sigma)$ and examples z:*

(HYP) *If $z \in \mathcal{X}(\varrho_\sigma)$ then $\mathcal{L}_{hyp}(\boldsymbol{s}, \sigma, z) = \{\sigma\}$.*

For $\sigma = init$ the statement "$z \in \mathcal{X}(\varrho_\sigma)$" is considered false.

In all teaching models presented in this thesis, all learners are defined as conservative.

Example 2.24 Let us consider the situation of Example 2.22, but with the learner modified to be conservative:

$$\mathcal{L}(\boldsymbol{s}, \sigma, z) = \begin{cases} \{(\boldsymbol{s} \circ \langle z \rangle, \sigma)\} & \text{if } z \in \mathcal{X}(\varrho_\sigma), \\ \{ \quad (\boldsymbol{s} \circ \langle z \rangle, \quad |\boldsymbol{s} \circ \langle z \rangle| \bmod (n+1)), & \\ \quad\;\; (\boldsymbol{s} \circ \langle z \rangle, \quad \min \varrho(\mathcal{S}_n(\boldsymbol{s} \circ \langle z \rangle)) \quad) \quad \} & \text{otherwise.} \end{cases}$$

Now $[1, n]$ is teachable in the limit by the teacher T with $T(i) = (1 + i \bmod n, 1)$ for all $i \in \mathbb{N}$, because in round $n + 1$ the hypothesis is $\sigma = 0$ and afterwards all examples are consistent with this hypothesis, that is, "$z \in \mathcal{X}(\varrho_\sigma)$" is true for all upcoming examples z. ■

We summarize this section by giving a list of all relevant definitions that constitute our formal teaching framework.

- The teacher:
  - with feedback: Definition 2.5,
  - without feedback: Definition 2.2.
- The learner: Definitions 2.7 and 2.23.

- The teaching time:
  - with feedback: Definition 2.14,
  - without feedback: Definition 2.21.
- The success criteria:
  - with feedback: Definition 2.12,
  - without feedback, finitely : Definition 2.19,
  - without feedback, in the limit: Definition 2.20.

We finally remark that the learners can not only be modeled as *non-deterministic* automata, but also as *probabilistic* automata, in which case the intermediate hypotheses, the memory, and the teaching times become random variables. A model with this sort of randomized learners is the subject of Part II of this thesis.

Part I

Teaching Models with Non-Deterministic Learner

Chapter 3

Consistent Learners

The central ingredient to all our teaching models is the concrete definition of the learner. In this chapter the only condition we put on the learner is that its hypothesis is always consistent with the memorized examples. There are three reasons why we start with this learner. First, the resulting teaching model is closely related to the teaching dimension model: Regardless of the kind of teachability (finitely, in the limit, or with feedback), the question whether or not a concept can be taught is reducible to calculating the teaching dimension of this concept (see Lemma 3.3). As the teaching dimension model is the only relatively well studied model that fits the ITS scenario, it is a good starting point for the development of our more sophisticated models in later chapters.

Second, this definition of the learner is natural because it is symmetric to the typical definition of a teacher in a learning model. Such a teacher makes all its knowledge about the target available, truthfully and completely. The consistent learner in the teaching model *incorporates* all its knowledge about the target truthfully and completely into its hypothesis.

Third, we argue that a teacher for the consistent learner has a wide range of application because the set of consistent learning algorithms is the largest easy-to-define class of learning algorithms. To make that statement clearer, we now briefly discuss the relationship between "consistent algorithm" and "learning algorithm."

From a purely information theoretical point of view, in which computability is no concern, the set of learning algorithms is closely related to the set of algorithms outputting consistent hypotheses. In inductive inference, learning in the limit and consistent learning in the limit are equivalent if learning strategies need not be computable [60, 45]. Conversely, every class-preserving, conservative, and consistent strategy is successful, provided the target is at all learnable. A special instance of these strategies is the *identification-by-enumeration* strategy. In the PAC learning model, everything that can be learned can be learned using a consistent and class-preserving algorithm if efficency is no concern [22, 27]. Conversely, every PAC algorithm can easily be transformed into one that is consistent with high probability. In the online learning setting, the *standard optimal algorithm* (see Littlestone [51]) also picks only consistent hypotheses.

On the other hand, as soon as computational resource bounds have to be obeyed, learning algorithms may have to be inconsistent in order to be successful. This happens when a consistent hypothesis cannot be found effectively or efficiently (see the articles by Wiehagen and Zeugmann [79, 80] and the references therein for more on the inconsistency phenomenon). In addition, practical algorithms often have some limitations on the size of the example storage. Finally, noisy data might not even admit a consistent hypothesis due to contradictions within the data.

We conclude that even if consistency does not precisely describe the property that makes an algorithm a learning algorithm, there is no more precise description available, and even if, such a description would very likely be much more complicated than the consistency property.

In our discussion above we implicitly assume that the algorithms, when searching for a consistent hypothesis, pay attention to every example they have been given. A common relaxation is consistency only with respect to some, rather than all, given examples. In the extreme variant the learner pays attention only to the last example received. Such *incremental learning algorithms* have been widely studied in learning theory (see [49, 23] and the references therein), but not yet in a teaching scenario, although they are closer to human behavior than algorithms with perfect unlimited memory. For a generalization of incremental learners, we also consider learners that pay attention to the last $m \geq 1$ examples for some fixed m. The normal consistent learners are covered by allowing $m = \infty$.

In Section 3.1 we define the consistent non-deterministic learner formally and show the connections between the resulting teaching model and the teaching dimension.

Most previous research on the teaching dimension focuses on concept classes as a whole, rather than on single concepts. But since the teaching dimension of a class is taken with respect to the worst concept in the class, it is often not suited for judging the teachability of the class. As a more plausible measure for the teachability of a concept class, the *average* teaching dimension over all concepts is sometimes used. In Section 3.2 we analyze the teaching dimensions of single concepts in the classes of 2-term DNFs and 1-decision lists and determine the average teaching dimension of both classes. The latter is $\Theta(n)$ in both cases, whereas the (worst case) teaching dimension is in both cases 2^n.

3.1 Description and Properties of the Model

The consistent learner with memory size m is formally defined as follows.

Definition 3.1 *Let $\langle \varrho, \mathcal{C} \rangle$ be a represented concept class. A non-deterministic learner $\mathcal{L} = (\mathcal{L}_{mem}, \mathcal{L}_{hyp})$ is called* consistent *iff*

(HYP) $\forall \boldsymbol{s} \in \mathcal{X}^* \; \forall \sigma \in \varrho(\mathcal{C}) \cup \{init\} \; \forall z \in \mathcal{X} \; \forall (\boldsymbol{s}', \sigma') \in \mathcal{L}(\boldsymbol{s}, \sigma, z)$:

$$\sigma' \in \begin{cases} \{\sigma\} & \text{if } z \in \mathcal{X}(\varrho_\sigma), \\ \varrho(\mathcal{C}(\boldsymbol{s}')) & \text{otherwise.} \end{cases}$$

Let $m \in \mathbb{N}^+ \cup \{\infty\}$. The learner is said to have memory size m *(or* m-memory*) iff*

(MEM) $\mathcal{L}_{mem}(\boldsymbol{s}, \sigma, z) = \{\boldsymbol{s} \circ_m \langle z \rangle\}$.

Example 3.2 Let $n \in \mathbb{N}$, $n \geq 2$ be arbitrarily fixed. Let us look at teaching $c^* = [1, n] \in \mathcal{S}_n$ to the consistent learner with 1-memory. A natural teacher without feedback would simply teach all examples in an endless loop: $T(i) = (1 + i \bmod n, 1)$ for $i \in \mathbb{N}$.

In round 0 the teacher gives $T(0) = (1, 1)$. Thus, in round 1 the memory is $\boldsymbol{s}_1 = \langle (1, 1) \rangle$ and the possible hypotheses $\mathcal{L}_{hyp}(\langle\,\rangle, init; T, 1) = [0, n] \setminus \{1\}$. Then T gives $T(1) = (2, 1)$ and the memory becomes $\boldsymbol{s}_2 = \langle (2, 1) \rangle$. The possible hypotheses are $\mathcal{L}_{hyp}(\langle\,\rangle, init; T, 1) = [0, n] \setminus \{2\}$ because, for example, from state $(\boldsymbol{s}_1, 2) \in \mathcal{L}(\langle\,\rangle, init; T, 1)$ the learner can switch to state $(\boldsymbol{s}_2, i)$ for all $i \neq 2$. In general, in round $t \geq 1$ the memory is $\boldsymbol{s}_t = \langle (1 + ((t-1) \bmod n), 1) \rangle$ and the possible hypotheses are $[0, n] \setminus \{1 + ((t-1) \bmod n)\}$. In no round is the set of possible hypotheses a subset of $\varrho(c^*) = \{0\}$. Therefore, T is not successful.

Now let $\mathcal{L}$ be the consistent learner with n-memory and let T be as above. The memory in round n is $\boldsymbol{s}_n = \langle (1, 1), \ldots, (n, 1) \rangle$. The hypothesis of $\mathcal{L}$ at round $n - 1$ either is already 0 or must be changed into one consistent with $\boldsymbol{s}_n$. But the only hypothesis consistent with $\boldsymbol{s}_n$ is 0. Thus, $\mathcal{L}(\langle\,\rangle, init; T, n) = \{(\boldsymbol{s}_n, 0)\}$, which means that T is successful and needs at most n rounds. ■

A teacher that is supposed to teach a concept $c^* \in \mathcal{C}$ to a consistent m-memory learner must give the learner a teaching set for c^*. Otherwise the learner can always switch to a hypothesis consistent with the examples but different from c^*. On the other hand every teaching set is suitable because it leaves the learner no other choice than to hypothesize the target. Therefore, the quality of an optimal teacher is given by the size of the smallest teaching set, that is, the teaching dimension of c^* with respect to $\mathcal{C}$. In other words, the question of teachability of a concept in this model is reduced to computing the teaching dimension.

In the following lemma, "teachable" is a placeholder for all three variants: finitely, in the limit, and with feedback.

Lemma 3.3 *Let $\langle \varrho, \mathcal{C} \rangle$ be a represented concept class and let $c^* \in \mathcal{C}$ be a target concept. Let $m \in \mathbb{N}^+ \cup \{\infty\}$. Then c^* is teachable to the consistent learner with memory size m if and only if $TD(c^*, \mathcal{C})$ is finite and $TD(c^*, \mathcal{C}) \leq m$. The optimal teaching time is $TD(c^*, \mathcal{C})$.*

Proof. Let $TD(c^*) \leq m$ and let $\{z_1, \ldots, z_{TD(c^*)}\}$ be a minimum teaching set. Then the following teacher is successful for teaching finitely (and therefore in the limit): $T(i) = z_{i+1}$ for $i = 0, \ldots, TD(c^*) - 1$. The same idea also works for teachers with feedback: $T(\boldsymbol{s}, \sigma) = z_{|\boldsymbol{s}|+1}$. In both situations the learner memorizes all examples $z_1, \ldots, z_{TD(c^*)}$ and in round $TD(c^*)$ outputs a consistent hypothesis. But by the definition of teaching set, there is only one such hypothesis, namely c^*. The teaching time is $TD(c^*)$. This proves the "if" direction.

Let c^* be teachable to the consistent learner $\mathcal{L}$ with memory size $m \in \mathbb{N}^+ \cup \{\infty\}$. Assume $\mathcal{L}$ always chooses a hypothesis that is wrong (but consistent with the memory), as long as such a hypothesis is available.

Case 1: $TD(c^*)$ is infinite.

Then all sets $\mathcal{C}(\boldsymbol{s})$ for finite sequences $\boldsymbol{s}$ contain a concept different from c^*.

We show that $\mathcal{L}$ can always choose a wrong hypothesis. In round 0, the learner $\mathcal{L}$ receives one example and can choose a hypothesis from the set $\varrho(\mathcal{C}(T(0)))$, which contains also a wrong one.

Assume in round t the current hypothesis σ is wrong. The current memory $\boldsymbol{s}$ contains t examples. Let z be the new example given by the teacher. If z is consistent with σ, then the next hypothesis is still σ and wrong. If z is inconsistent with σ then $\mathcal{L}$ can choose the next hypothesis from the set $\varrho(\mathcal{C}(\boldsymbol{s} \circ \langle z \rangle))$, which contains a wrong one. In both cases the hypothesis in round $t + 1$ is wrong. □ Case 1

Case 2: $TD(c^*)$ is finite but greater than m.

Then m is finite, too. Now for all sequences $\boldsymbol{s}$ of length at most $m < TD(c^*)$ the set $\mathcal{C}(\boldsymbol{s})$ contains a concept different from c^*, by the definition of the teaching dimension. A similar argument as in Case 1 shows that $\mathcal{L}$ can in every round choose a wrong hypothesis. □ Case 2

The Cases 1 and 2 contradict the assumption that c^* is teachable. Therefore both cases cannot occur, and thus $TD(c^*)$ is finite and $TD(c^*) \leq m$. This shows the "only if" direction.

To show that the optimal teaching time is $TD(c^*)$, let c^* be teachable. It remains to show that no teacher can be successful in less than $TD(c^*)$ rounds. For $TD(c^*) = 1$ the statement is clear because a learner needs at least one round to reach the target. So assume $TD(c^*) \geq 2$.

In round 1, the learner $\mathcal{L}$ may choose from $\varrho(\mathcal{C}(T(0)))$ and pick a wrong hypothesis (because $TD(c^*) \geq 2$). Now consider round $1 \leq t < TD(c^*) - 1$. Let $(\boldsymbol{s}, \sigma)$ with $\sigma \notin \varrho(c^*)$ be the current state and z the newly arrived example. If z is consistent with σ then due to conservativeness the hypothesis does not change and is still not in $\varrho(c^*)$. If z is inconsistent with σ then $\mathcal{L}$ may choose from $\varrho(\mathcal{C}(\boldsymbol{s} \circ \langle z \rangle))$. But $\boldsymbol{s} \circ \langle z \rangle$ consists

of $t+1 < TD(c^*)$ examples. Hence there is a concept in $\mathcal{C}$ consistent with $\boldsymbol{s} \circ \langle z \rangle$ but different from c^*. Therefore in round $t+1$ the hypothesis is still wrong.

This shows that before round $TD(c^*)$ a teacher cannot be successful. On the other hand, the teachers described in the "if" direction are successful after $TD(c^*)$ rounds. Therefore the optimal teaching time for c^* is $TD(c^*)$. □

For the TD model, the questions from Section 1.5 have already received attention in the literature. The teaching dimension has been calculated for many natural concept classes. We can only cite a few, such as (monotone) monomials, monotone k-term DNFs, k-term μ-DNFs, monotone decision lists and rectangles in $\{0,1,\ldots,n-1\}^d$ [34]; for linearly separable Boolean functions [8, 9]; for threshold functions [76]; and for k-juntas and sparse GF_2 polynomials [50].

The teaching dimension is a measure that is based only on the concept class. It has therefore been compared to other notions of dimensionality and to parameters occurring in the query learning model, in the PAC model, or in the online learning model (see Hegedűs [41, 40], Goldman and Kearns [34], Angluin [4], Ben-David and Eiron [16], and Rivest and Yin [69]).

Deciding whether a given concept in a given class has a teaching dimension of less then a given value is $\mathcal{NP}$-complete [76, 34, 8]. This can be shown by a reduction from the SET-COVER problem. As an optimization problem, SET-COVER has been studied intensively, and it is relatively easy to translate these results to the problem of computing optimal teaching sets. For the formal definition of the MIN-TEACHING-SET problem we assume that $X = [1,k]$ for some $k \in \mathbb{N}^+$ and that $\mathcal{C}$ is represented as a binary $|\mathcal{C}| \times k$ matrix. A concept is represented as a binary string of length k. In the following definitions we use the terminology from Ausiello *et al.* [10].

Definition 3.4 *The problem* MIN-TEACHING-SET *is the optimization problem with*

- *instances of the form* $(\mathcal{C}, c^*)$,
- *feasible solutions* $sol(\mathcal{C}, c^*) = \{\boldsymbol{s} \in X^* \mid \mathcal{C}(\boldsymbol{s}) = \{c^*\}\}$,
- *a measure* μ *with* $\mu((\mathcal{C}, c^*), \boldsymbol{s}) = |\boldsymbol{s}|$ *for all* $\boldsymbol{s} \in sol(\mathcal{C}, c^*)$.

Definition 3.5 *The problem* SET-COVER *is the optimization problem with*

- *instances of the form* $(U, V_1, \ldots, V_k)$ *with a finite set* U *and sets* $V_1, \ldots, V_k \subseteq U$,
- *feasible solutions* $sol(U, V_1, \ldots, V_k) = \{\boldsymbol{s} \in [1,k]^* \mid \bigcup_{i=1}^{|\boldsymbol{s}|} V_{\boldsymbol{s}[i]} = U\}$,
- *a measure* μ *with* $\mu((U, V_1, \ldots, V_k), \boldsymbol{s}) = |\boldsymbol{s}|$ *for all* $\boldsymbol{s} \in sol(U, V_1, \ldots, V_k)$.

The usual reductions from SET-COVER to MIN-TEACHING-SET [76, 34, 8] map an instance $(U, V_1, \ldots, V_k)$ of SET-COVER to the MIN-TEACHING-SET instance $(\{c_1, \ldots, c_{|U|}, c^*\}, c^*)$ in which all concepts are subsets of $[1,k]$: $c_j = \{i \in [1,k] \mid j \in V_i\}$ and $c^* = \emptyset$. Then

the sets $V_{x_1}, \ldots, V_{x_\ell}$ cover U if and only if the examples $(x_1, 0), \ldots, (x_\ell, 0)$ constitute a teaching set for c^*.

On the other hand, every MIN-TEACHING-SET instance $(\{c_1, \ldots, c_n, c^*\}, c^*)$ over a domain $X = [1, k]$ can be mapped to the SET-COVER instance $(U, V_1, \ldots, V_{|X|})$ with $U = [1, n-1]$ and $V_i = \{j \mid c_i(j) \neq c^*(j)\}$. Now the examples $(x_1, c^*(x_1)), \ldots, (x_\ell, c^*(x_\ell))$ are a teaching set for c^* if and only if the sets $V_{x_1}, \ldots, V_{x_\ell}$ cover U.

Thus, instances of both problems can be transferred into each other, with the representation size of the instances changing only slightly. In particular, for two corresponding instances we have $|\mathcal{C}| - 1 = |U|$. This means that the result by Feige [28], who shows non-approximability for SET-COVER within a factor of $(1 - \varepsilon) \ln |U|$, can be formulated as follows.

Theorem 3.6 *The problem of computing a minimal teaching set is $\mathcal{NP}$ hard. It cannot be approximated in polynomial time within a factor of $(1 - \varepsilon) \ln(|\mathcal{C}| - 1)$ for all $\varepsilon > 0$, unless $\mathcal{NP} \subseteq DTime(n^{\log \log n})$.*

In a similar way, other non-approximability results for SET-COVER (for example, Raz and Safra [64]) can be rephrased for MIN-TEACHING-SET.

Lemma 3.3 can be used to judge the model of teaching the consistent learner with regard to the aspects described in Section 1.3. An immediate consequence of the lemma is that the greater the memory size is the more concepts are teachable. For a given teachable concept, however, the teaching time does not improve with growing memory size. Thus memory-size dependence is at least rudimentarily present. In contrast, the feedback is of no use and the order of examples is irrelevant.

The applicability to infinite concepts and classes is limited. Even in rather simple classes, like the class of all finite languages, all concepts have an infinite teaching dimension and are therefore unteachable in this model. One notable exception is the class of pattern languages introduced by Angluin [1]. Every pattern language has a finite teaching dimension with respect to the class of all pattern languages because there are only finitely many pattern languages consistent with any positive example ("finite thickness" [2]). Zeugmann [81] gives the precise values of the teaching dimensions of pattern languages.

For the remainder of this chapter we focus on the question whether the teaching dimension is a plausible measure for teachability. That the teaching dimension does not always capture our intuition about teachability can be seen in Example 3.2. To successfully teach the all-concept in $\mathcal{S}_n$, a teacher has to rule out all co-singleton concepts. This can only be done by providing *all* n examples. The teaching dimension then is the maximum possible, which does not properly reflect the easy teachability one would expect from this class.

There are two main reasons for this implausible result. First, the teaching dimension of the class is determined by the worst case teaching dimension over all concepts.

Easily learnable concepts are not taken into account. Second, the teaching dimension itself represents the worst case over all learners, among which are also "bad" ones. We address remedies for the second reason in the next chapters; in this chapter we consider the average teaching dimension instead of the worst case teaching dimension as a remedy for the first reason.

3.2 The Average Teaching Dimension

Definition 3.7 *Let $\mathcal{C}$ be a concept class. The* average teaching dimension *of $\mathcal{C}$ is defined as $\overline{TD}(\mathcal{C}) = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} TD(c,\mathcal{C})$.*

That the average teaching dimension actually might be more plausible than the normal (worst case) teaching dimension can be illustrated with the help of the class $\mathcal{S}_n$. Its average teaching dimension is

$$\overline{TD}(\mathcal{S}_n) = \frac{TD(c_0) + n \cdot TD(c_1)}{|\mathcal{S}_n|} = \frac{n + n \cdot 1}{n+1} < 2$$

and thus much smaller than the worst case teaching dimension $TD(\mathcal{S}_n) = n$.

Naturally, to calculate the average teaching dimension it is necessary to consider all concepts in a class. Consequently, for classes that are more complex than $\mathcal{S}_n$ the average teaching dimension is often much harder to determine than the (worst case) teaching dimension, and much less results are known.

We have already mentioned that Anthony, Brightwell, and Shawe-Taylor [9] have proved that the class of linearly separable Boolean functions over n variables has an average teaching dimension of $O(n^2)$. Furthermore, Kuhlmann [46] has shown that concept classes with VC dimension 1 have an average teaching dimension of less than 2 and that balls of radius d in $\{0,1\}^n$ have an average teaching dimension of at most $2d$.

Based on our results below, Lee, Servedio, and Wan [50] have shown that the class of DNFs with at most $s \leq 2^{\Theta(n)}$ terms has an average teaching dimension of $O(ns)$. Moreover they show that the average teaching dimension of the class of k-juntas is at most $2^k + o(1)$ and that the average teaching dimension for GF_2 polynomials with $s \leq (1-\varepsilon)\log_2 n$ monomials is at most $ns + 2s$.

A more general result is given by Kushilevitz, Linial, Rabinovich, and Saks [47], who show an upper bound of $O(\sqrt{|\mathcal{C}|})$ for the average teaching dimension of any class $\mathcal{C}$. They also define a family of classes for which the average teaching dimension is $\Omega(\sqrt{|\mathcal{C}|})$.

In this section we show that 2-term DNFs as well as 1-decision lists have an average teaching dimension linear in the number n of variables, although their teaching dimension is 2^n.

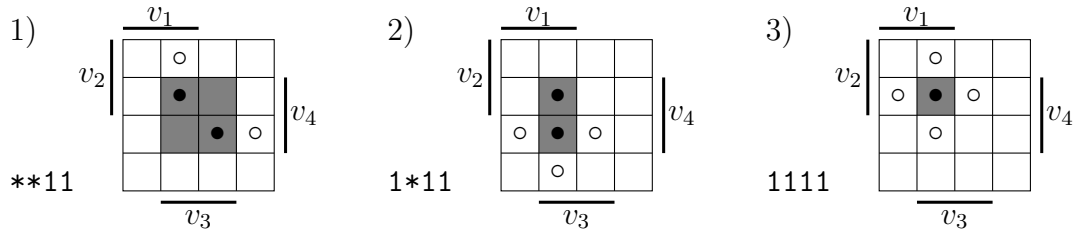

Figure 3.1 *Minimum teaching sets for three monomials over four variables. The monomials are depicted in Karnough diagrams. The positive examples in the teaching set are marked by •, the negative examples by ∘.*

3.2.1 2-Term DNFs

We start by looking at the simpler class of 1-term DNF (monomials). Goldman and Kearns [34] showed that the class of monomials over n variables (without the empty concept) has a teaching dimension of $n+1$. More precisely, a monomial with k literals has a teaching dimension of $\min\{k+2, n+1\}$. An intuition as to the shape of a minimum teaching set for monomials is given in Figure 3.1. The positive examples cause every consistent hypothesis to be at least as "large" as the target (that is, it contains at most the literals of the target monomial); the negative examples take care that the hypothesis does not extend beyond the target (that is, it contains all literals of the target).

In this section, we assume that $\mathcal{M}_n^1$ contains the empty concept. Adding the empty concept to the class of monomials does not change the teaching dimension of the non-empty concepts since their minimum teaching sets always contain a positive example, which rules out the empty concept "for free." The empty concept has a teaching dimension of 2^n since each of the 2^n singleton concepts is contained in $\mathcal{M}_n^1$ and one example can rule out only one such concept. There are 3^n non-empty concepts in $\mathcal{M}_n^1$ and therefore

$$\overline{TD}(\mathcal{M}_n^1) \leq \frac{2^n + 3^n \cdot (n+1)}{3^n + 1} \leq \left(\frac{2}{3}\right)^n + n + 1 \leq n + 2 \ .$$

To calculate the average teaching dimension for 2-term DNFs we identify subsets of $\mathcal{M}_n^2$ for which we can give bounds on the teaching dimension and on the cardinality. The first such subset contains only the empty concept, the second subset contains $\mathcal{M}_n^1 \setminus \{\emptyset\}$, and the third subset contains the remaining concepts from $\mathcal{M}_n^2$.

The empty concept has a teaching dimension of 2^n with respect to $\mathcal{M}_n^2$. Concepts representable as monomial have a teaching dimension with respect to $\mathcal{M}_n^2$ at least as large as the number of negative examples, hence $TD(c) \geq 2^{n-1}$ for $c \in \mathcal{M}_n^1 \setminus \{\{0,1\}^n\}$. The all-concept $\{0,1\}^n \in \mathcal{M}_n^1$ has a teaching dimension of 2 since $\{(0^n, 1), (1^n, 1)\}$ is a minimum teaching set for it. It remains to calculate the teaching dimensions for

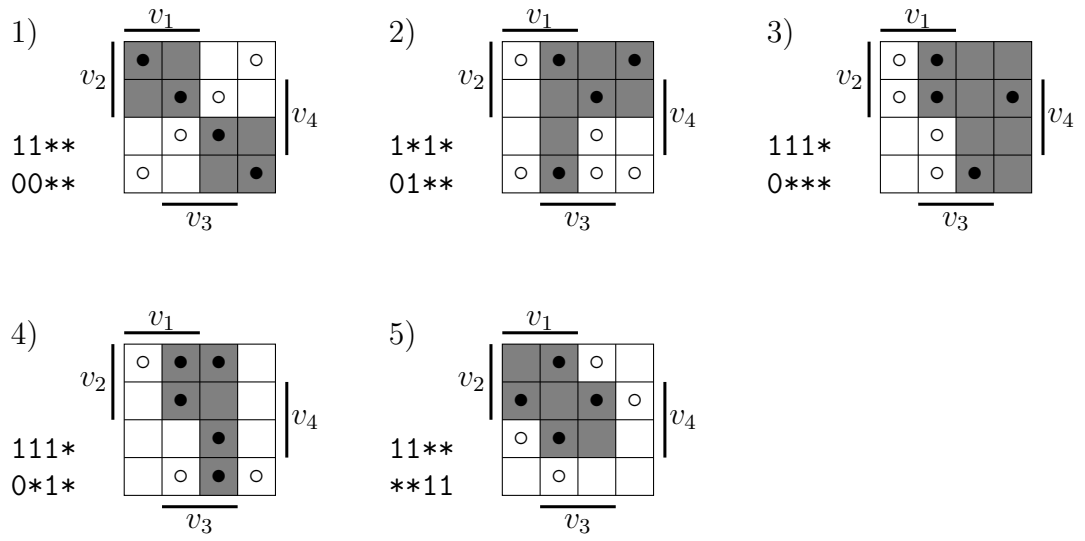

Figure 3.2 *Five 2-term DNFs over four variables corresponding to the five cases distinguished in Lemma 3.8. Positive and negative examples in the teaching set are marked by • and ∘, respectively.*

the concepts in $\mathcal{M}_n^2 \setminus \mathcal{M}_n^1$. These teaching dimensions must be small enough and there must be enough of these concepts in order for the average teaching dimension to become linear in n. The next lemma shows that their teaching dimensions are in $O(n)$.

Lemma 3.8 *For all $c \in \mathcal{M}_n^2 \setminus \mathcal{M}_n^1$, $TD(c, \mathcal{M}_n^2) \leq 2n + 4$.*

Proof. In this proof we use some additional shorthand notation. For $x_1, x_2 \in \{0,1\}^n$ we denote by $x_1 \cup x_2$ the monomial with

$$(x_1 \cup x_2)[i] = \begin{cases} * & \text{if } x_1[i] \neq x_2[i], \\ x_1[i] & \text{if } x_1[i] = x_2[i], \end{cases}$$

for all $i = 1, \ldots, n$. Thus, $x_1 \cup x_2$ is the minimal monomial containing x_1 and x_2. For two monomials $M, P \in \{0, 1, *\}^n$ we write $P[i] \supseteq M[i]$ iff $P[i] = *$ or $P[i] = M[i]$.

Let $c \in \mathcal{M}_n^2 \setminus \mathcal{M}_n^1$ be represented by the 2-term DNF $M_1 \vee M_2$ with monomials $M_1, M_2 \in \{0, 1, *\}^n$.

The basic idea for constructing a teaching set for c is similar to the construction of teaching sets for a single monomial M. We include two complementary positive examples per monomial. They ensure that a monomial P consistent with both examples must at least encompass M. All neighbors of an arbitrary positive example that do not satisfy the monomial are then included into the teaching set as negative examples. They ensure that P cannot be a proper superset of M (see Figure 3.1).

Dealing with 2-term DNFs is more complicated because the two monomials need not be disjoint; moreover, many concepts in $\mathcal{M}_n^2$ can be represented by more than one 2-term DNF. In the following we distinguish five cases according to the number and kind of differences between the monomials M_1 and M_2. Figure 3.2 shows an example for each of the five cases.

Case 1: M_1 and M_2 have at least two strong differences.

Without loss of generality, we assume two strong differences at position 1 and 2: $* \neq M_1[1] \neq M_2[1] \neq *$ and $* \neq M_1[2] \neq M_2[2] \neq *$.

First we define a set $S = S^+ \cup S^-$ of cardinality at most $4+2n$ and then we show that S is a teaching set for $M_1 \vee M_2$. Let $x_1 = M_1[\frac{*}{0}]$, $x_1' = M_1[\frac{*}{1}]$, $x_2 = M_2[\frac{*}{0}]$, $x_2' = M_2[\frac{*}{1}]$. Then $S^+ = \{(x_1, 1), (x_1', 1), (x_2, 1), (x_2', 1)\}$. The set S^- consists of all examples $(x, 0)$ in which x is a neighbor of x_1 or x_2 that neither satisfies M_1 nor M_2. Since each instance has n neighbors, it follows that $|S| \leq 4 + 2n$.

In order to show that S is a teaching set, let $K_1 \vee K_2$ be 2-term DNF consistent with S. We have to show that $K_1 \vee K_2$ is equivalent to $M_1 \vee M_2$. We assume without loss of generality that x_1 satisfies K_1, that is, $x_1 \in K_1$.

Claim 1: $x_1, x_1' \in K_1 \setminus K_2$ and $x_2, x_2' \in K_2 \setminus K_1$.

Proof. First, we show $x_2, x_2' \notin K_1$. Suppose for a contradiction $x_2 \in K_1$. From $x_1, x_2 \in K_1$, $x_1[1] \neq x_2[1]$, and $x_1[2] \neq x_2[2]$ we get that $K_1[1] = K_2[2] = *$. Then K_1 also contains the 1-neighbor x of x_1. On the other hand, $x \notin M_1 \vee M_2$ and hence $(x, 0) \in S^-$. Thus K_1 is satisfied by a negative example, a contradiction. Analogously one can show that $x_2' \notin K_1$. This implies $x_2, x_2' \in K_2$. In a symmetric way one proves $x_1, x_1' \notin K_2$. □ Claim 1

Claim 1 implies $K_1[i] = *$ for all i with $M_1[i] = *$, and it implies $K_1[i] \supseteq M_1[i]$ for all other i. It remains to show that $K_1[i] = M_1[i]$ for all i with $M_1[i] \neq *$ and analogously for K_2 and M_2. Suppose there is an i such that $K_1[i] = * \neq M_1[i]$. Let x be the i-neighbor of x_1. Then $x \in K_1 \setminus M_1$. Additionally $x \notin M_2$ since x certainly differs from M_2 at the first or second position (not necessarily at both, since one of them could be i). Thus $(x, 0) \in S^-$, and since K_1 is consistent with S^-, $x \in K_1$ cannot be true, a contradiction. By the same arguments, one shows that $K_2[i] = M_2[i]$ for all i with $M_2[i] \neq *$. We have now proved $K_1 = M_1$ and $K_2 = M_2$, hence S is a teaching set for c. □ Case 1

Case 2: M_1 and M_2 have one strong difference and two weak differences of different kind.

Without loss of generality, let $M_1 = b_1 * b_3 y_1$ and let $M_2 = \bar{b}_1 b_2 * y_2$ with $y_1, y_2 \in \{0, 1, *\}^{n-3}$ and $b_1, b_2, b_3 \in \{0, 1\}$. Let S^+ contain the four positive examples with instances $x_1 = b_1 \bar{b}_2 b_3 y_1[\frac{*}{0}]$, $x_1' = b_1 b_2 b_3 y_1[\frac{*}{1}]$, $x_2 = \bar{b}_1 b_2 \bar{b}_3 y_2[\frac{*}{0}]$, and $x_2' = \bar{b}_1 b_2 b_3 y_2[\frac{*}{1}]$. Let

S^- contain the negative examples whose instances are again all neighbors of x_1 or x_2 that do not satisfy $M_1 \vee M_2$.

To prove that $S = S^+ \cup S^-$ is a teaching set, let $K_1 \vee K_2$ be a 2-term DNF consistent with S. We assume without loss of generality $x_1 \in K_1$.

Claim 2: $x_1, x_1' \in K_1 \setminus K_2$ and $x_2, x_2' \in K_2 \setminus K_1$.

Proof. First we show $x_2, x_2' \notin K_1$. Suppose for a contradiction that $x_2 \in K_1$. Then $K_1[1] = K_1[2] = K_2[3] = *$. Now let x be the 3-neighbor of x_1. Then $x \in K_1$ (since $K_1[3] = *$), but $x \notin M_1$ (since $x[3] \neq M_1[3]$) and $x \notin M_2$ (since $x[1] \neq M_2[1]$). Thus $(x, 0) \in S^-$, and x cannot satisfy K_1, a contradiction. Therefore $x_2 \notin K_1$.

Now suppose $x_2' \in K_1$. Then $K_1[1] = K_2[2] = *$. Let x be the 1-neighbor of x_1. Then $x \in K_1$, but $x \notin M_1$ and $x \notin M_2$ (since $x[2] = x_1[2] = \overline{M_2[2]}$), hence $(x, 0) \in S^-$, a contradiction. Therefore $x_2' \notin K_1$. It follows that $x_2, x_2' \in K_2$ and analogous arguments show $x_1, x_1' \notin K_2$. □ Claim 2

From Claim 2 it follows $K_1 \supseteq x_1 \cup x_1'$ and $K_2 \supseteq x_2 \cup x_2'$.

Claim 3: $K_1 \not\supset x_1 \cup x_1'$ and $K_2 \not\supset x_2 \cup x_2'$.

Proof. Suppose for a contradiction $K_1 \supset x_1 \cup x_1'$. Then there is an i with $K_1[i] = x[i] \neq (x_1 \cup x_1')[i]$. For this i we have $M_1[i] \neq *$ (otherwise $(x_1 \cup x_1')[i] = *$). Let x be the i-neighbor of x_1. Then $x \in K_1$, but $x \notin M_1$ (since $x[i] \neq M_1[i]$) and $x \notin M_2$. The latter holds because in case of $i = 1$ we have $x[2] = x_1[2] \neq M_2[2]$ and in case $i > 1$ we have $x[1] = x_1[1] \neq M_2[1]$. But then K_1 is satisfied by x and $(x, 0) \in S^-$, a contradiction. Similarly one disproves the assumption $K_2 \supset x_2 \cup x_2'$. □ Claim 3

From Claim 2 and 3 it follows that $K_1 = M_1$ and $K_2 = M_2$. Therefore, S is a teaching set for $M_1 \vee M_2$. □ Case 2

Case 3: M_1 and M_2 have one strong difference, at least two weak differences of the same kind, and no weak differences of different kind.

Without loss of generality, let $M_1 = b_1 b_2 b_3 y_1$ and $M_2 = \bar{b}_1 ** y_2$ with $y_1 \subseteq y_2$. Note that there is a different but equivalent 2-term DNF: $M_1 \vee M_2 \equiv \widehat{M_1} \vee M_2$ with $\widehat{M_1} = *b_2 b_3 y_1$. Let S^+ contain the positive examples with instances $x_1 = b_1 b_2 b_3 y_1[\frac{*}{0}]$, $x_1' = b_1 b_2 b_3 y_1[\frac{*}{1}]$, $x_2 = \bar{b}_1 \bar{b}_2 b_3 y_2[\frac{*}{0}]$, and $x_2' = \bar{b}_1 b_2 \bar{b}_3 y_2[\frac{*}{1}]$. Let S^- contain the negative examples whose instances are again those neighbors of x_1 or x_2 that do not satisfy $M_1 \vee M_2$.

To prove that $S = S^+ \cup S^-$ is a teaching set for $M_1 \vee M_2$, let $K_1 \vee K_2$ be a 2-term DNF consistent with S.

Claim 4: $x_1, x_1' \in K_1 \setminus K_2$ and $x_2, x_2' \in K_2 \setminus K_1$.

Proof: First we show $x_2, x_2' \notin K_1$. Suppose for a contradiction $x_2 \in K_1$. Then $K_1 \supseteq **b_3(y_1[\frac{*}{0}] \cup y_2[\frac{*}{0}])$. Let x be the 2-neighbor of x_1. Then $x \in K_1$, but $x \notin M_1$ and

$x \notin M_2$. Thus $(x,0) \in S^-$, a contradiction. The statement $x_2' \in K_1$ can be disproved analogously using the 3-neighbor of x_1. Therefore $K_2 \supseteq M_2$.

Suppose for a contradiction $x_1' \in K_2$. Then $K_2 \supseteq \texttt{***}(y_1[\frac{*}{1}] \cup y_2[\frac{*}{1}])$. For the 2-neighbor x of x_1 we then have $x \in K_2$, but $(x,0) \in S^-$, a contradiction. □ Claim 4

From Claim 4 it follows $K_1 \supseteq M_1$ and $K_2 \supseteq M_2$. Note that $K_1 = M_1$ need not hold, as $K_1 = \widehat{M_1}$ is also "allowed" since $M_1 \vee M_2 \equiv \widehat{M_1} \vee M_2$.

Claim 5: $K_1 \not\supset \widehat{M_1}$ and $K_2 \not\supset M_2$.

Proof. Suppose for a contradiction $K_1 \supset \widehat{M_1}$. Then there is an i with $K_1[i] = * \neq \widehat{M_1}[i]$ and thus $i > 1$. Moreover, the i-neighbor x of x_1 satisfies K_1. But because x is neither in M_1 (since $x[i] \neq M_1[i]$) nor in M_2 (since $x[1] \neq M_2[1]$), we have $x \in S^-$, a contradiction. The assumption $K_2 \supset M_2$ can be disproved similarly using the i-neighbor of x_2. □ Claim 5

Altogether we have now shown that $M_1 \subseteq K_1 \subseteq \widehat{M_1}$ and $K_2 = M_2$. It follows $M_1 \vee M_2 \equiv K_1 \vee K_2$, and thus S is a teaching set for $M_1 \vee M_2$. □ Case 3

Case 4: M_1 and M_2 have exactly one strong difference and exactly one weak difference.

Without loss of generality, let $M_1 = b_1 b_2 y$ and $M_2 = \bar{b}_1 * y$ with $y \in \{\texttt{0}, \texttt{1}, *\}^{n-2}$. Note that the concept has three equivalent representations. With $\widehat{M_1} = * b_2 y$ and $\widehat{M_2} = \bar{b}_1 \bar{b}_2 y$ we have $M_1 \vee M_2 \equiv M_1 \vee \widehat{M_2} \equiv \widehat{M_1} \vee M_2$.

Let S^+ contain the five positive examples with instances $x_1 = b_1 b_2 y[\frac{*}{0}]$, $x_1' = b_1 b_2 y[\frac{*}{1}]$, $x_2 = \bar{b}_1 \bar{b}_2 y[\frac{*}{0}]$, $x_2' = \bar{b}_1 \bar{b}_2 y[\frac{*}{1}]$, $x_3 = \bar{b}_1 b_2 y[\frac{*}{0}]$. Let S^- contain all negative examples $(x,0)$ such that x is a neighbor of x_1 or x_2 not satisfying $M_1 \vee M_2$. Note that the 1-neighbor of x_1 does satisfy $M_1 \vee M_2$, hence $|S^-| \leq 2n-1$ and therefore $|S^+ \cup S^-| \leq 4 + 2n$.

To show that $S = S^+ \cup S^-$ is a teaching set for $M_1 \vee M_2$, let $K_1 \vee K_2$ be consistent with S. Without loss of generality, we assume $x_1 \in K_1$.

Claim 6: $x_1, x_1' \in K_1 \setminus K_2$ and $x_2, x_2' \in K_2 \setminus K_1$.

Proof: Suppose for a contradiction that $x_2 \in K_1$. Then $K_1 \supseteq \texttt{**}y[\frac{*}{0}]$ and it follows $x_4 \in K_1$, a contradiction. Suppose for a contradiction that $x_2' \in K_1$. Then $K_1 \supseteq \texttt{**}y$, thus K_1 is satisfied by x_4, a contradiction. Therefore we have $x_2, x_2' \notin K_1$. But since x_2 and x_2' satisfy $K_1 \vee K_2$, they must satisfy K_2.

Now suppose $x_1' \in K_2$. Then $x_1', x_2 \in K_2$ and hence $K_2 \supseteq \texttt{**}y$. But then also $x_4 \in K_2$, a contradiction. □ Claim 6

Claim 7: $K_1 \not\supset * b_2 y$ and $K_2 \not\supset \bar{b}_1 * y$.

Proof: Suppose for a contradiction that $K_1 \supset * b_2 y$. Now K_1 cannot equal $\texttt{**}y$ because in this case the 2-neighbor $b_1 \bar{b}_2 y[\frac{*}{0}]$ of x_1, which does not satisfy $M_1 \vee M_2$, would satisfy K_1, a contradiction. But since we suppose $K_1 \supset * b_2 y$, there is an $i \geq 3$ with

$K_1[i] = * \neq y[i-2]$. Let x be the i-neighbor of x_1. Then $x[i] = \overline{y[i-2]} = \overline{M_1[i]} = \overline{M_2[i]}$, hence $(x, 0) \in S^-$. But since $K_1[i] = *$, we have $x \in K_1$, a contradiction.

Similarly one shows $K_2 \not\supset \bar{b}_1 * y$. □ Claim 7

It follows from the Claims 6 and 7 that $\widehat{M_1} = *b_2y \supseteq K_1 \supseteq (x_1 \cup x_1') = M_1$ and $M_2 = \bar{b}_1 * y \supseteq K_2 \supseteq (x_2 \cup x_2') = \widehat{M_2}$. Thus for both K_1 and K_2 there are two possibilities. The combination $K_1 = M_1$ and $K_2 = \widehat{M_2}$ is not consistent with S because it is not satisfied by x_3. The other three combinations are, as we have already mentioned above, equivalent to $M_1 \vee M_2$. We conclude $K_1 \vee K_2 \equiv M_1 \vee M_2$. □ Case 4

So far all cases with at least one strong difference have been covered. The cases without strong difference still remain. Since the target concept $M_1 \vee M_2$ represents no concept in $\mathcal{M}_n^1$, we have neither $M_1 \subseteq M_2$ nor $M_2 \subseteq M_1$. Therefore we only need to consider situations with at least two weak differences of different kind. Some of these cases have already been covered. Case 4 treats the case of exactly two differences of different kind (and otherwise identical terms). Case 3 treats the case of exactly two differences of different kind plus exactly one more weak difference. Thus, only the following case remains.

Case 5: M_1 and M_2 have at least two disjoint pairs of weak differences of different kind.

Without loss of generality, let $M_1 = b_1b_2{*}{*}y_1$ and $M_2 = {*}{*}b_3b_4y_2$.

Let the set S^+ contain the positive examples with the instances $x_1 = b_1b_2\bar{b}_3b_4y_1[\frac{*}{0}]$, $x_1' = b_1b_2b_3\bar{b}_4y_1[\frac{*}{1}]$, $x_2 = \bar{b}_1b_2b_3b_4y_2[\frac{*}{0}]$, and $x_2' = b_1\bar{b}_2b_3b_4y_2[\frac{*}{1}]$. Let S^- contain the negative examples whose instances are all neighbors of x_1 or x_2 that do not satisfy $M_1 \vee M_2$.

Let $K_1 \vee K_2$ be consistent with $S = S^+ \cup S^-$, and let without loss of generality $x_1 \in K_1$.

Claim 8: $x_1, x_1' \in K_1 \setminus K_2$ and $x_2, x_2' \in K_2 \setminus K_1$.

Proof. Suppose for a contradiction $x_2 \in K_1$. Then $K_1 \supseteq {*}b_2{*}b_4(y_1[\frac{*}{0}] \cup y_2[\frac{*}{1}])$. Using the 1-neighbor $\bar{b}_1b_2\bar{b}_3b_4y_1[\frac{*}{0}]$ of x_1, we can get a contradiction. Analogously we disprove $x_2' \in K_1$ using the 2-neighbor of x_1. Therefore $x_2, x_2' \notin K_1$.

A symmetrical reasoning shows $x_1' \notin K_2$. □ Claim 8

Claim 8 shows $K_1 \supseteq (x_1 \cup x_1') = M_1$ and $K_2 \supseteq (x_2 \cup x_2') = M_2$.

Claim 9: $K_1 \not\supset M_1$ and $K_2 \not\supset M_2$.

Proof. Suppose for a contradiction $K_1 \supset M_1$. Then there is an i with $K_1[i] = * \neq \overline{M_1[i]}$. Let x be the i-neighbor of x_1. Then $x \in K_1$, but $x \notin M_1$ since $x[i] = \overline{x_1[i]} = \overline{M_1[i]}$. Because of $M_1[i] \neq *$ the index i cannot be 3, hence x differs from M_2 on the third position, that is, $x \notin M_2$. Therefore $(x, 0) \in S^-$, a contradiction. Analogously we disprove $K_2 \supset M_2$. □ Claim 9

We conclude that $K_1 \vee K_2 \equiv M_1 \vee M_2$ and that S is a teaching set. □ Case 5

□

Lemma 3.8 presents a complete distinction of cases for the concepts in the class $\mathcal{M}_n^2 \setminus \mathcal{M}_n^1$ and in each case the teaching dimension is bounded by $2n+4$. We therefore get the following corollary.

Corollary 3.9 *$TD(\mathcal{M}_n^2 \setminus \mathcal{M}_n^1) \leq 2n+4$.*

To complete the calculation of the average teaching dimension of $\mathcal{M}_n^2$ we have to count the concepts in $\mathcal{M}_n^2 \setminus \mathcal{M}_n^1$. The next lemma provides bounds for the number of these concepts.

Lemma 3.10 *$\frac{1}{3} \cdot 9^n \leq |\mathcal{M}_n^2 \setminus \mathcal{M}_n^1| \leq \frac{2}{3} \cdot 9^n$ for all $n \geq 10$.*

Proof. All 2-term DNFs of the form considered in Case 1 of Lemma 3.8 represent pairwise different concepts (modulo permutation of the monomials). Each such 2-term DNF can be described by the number k of strong differences ($2 \leq k \leq n$), their kind (two possibilities: `0/1` or `1/0`), and the kind of the positions without strong differences ($n-k$ positions with seven possibilities each: `0/0`, `1/1`, `0/*`, `1/*`, `*/*`, `*/0`, `*/1`). The number of concepts represented by such 2-term DNFs is thus

$$\frac{1}{2} \cdot \sum_{k=2}^{n} \binom{n}{k} \cdot 2^k \cdot 7^{n-k} = \frac{1}{2} \cdot (9^n - 7^n - 2n \cdot 7^{n-1}),$$

which is greater than $\frac{1}{3} \cdot 9^n$ for $n \geq 10$ and therefore proves the lower bound.

There are $(3^n+1)^2$ syntactically different 2-term DNFs of which the 3^n+1 ones with two identical monomials do not represent true 2-term DNFs. The remaining $3^n(3^n+1)$ 2-term DNFs represent $\frac{1}{2} \cdot 3^n(3^n+1) = \frac{1}{2}(9^n+3^n)$ concepts. This number is therefore an upper bound for $|\mathcal{M}_n^2 \setminus \mathcal{M}_n^1|$. □

The above proof actually shows the number of true 2-term DNF concepts to be asymptotically equal to $\frac{1}{2} \cdot 9^n$. We are now ready to calculate the average teaching dimension of $\mathcal{M}_n^2$.

Theorem 3.11 *$\overline{TD}(\mathcal{M}_n^2) \leq 4n+10$ for all $n \geq 10$.*

Proof. The teaching dimension of each of the 3^n+1 concepts in $\mathcal{M}_n^1$ can be upper bounded by 2^n and that of the concepts in $\mathcal{M}_n^2 \setminus \mathcal{M}_n^1$ by $2n+4$. Therefore by Lemma 3.10 for all $n \geq 10$

$$\overline{TD}(\mathcal{M}_n^2) \leq \frac{(3^n+1) \cdot 2^n + \frac{2}{3} \cdot 9^n \cdot (4+2n)}{(3^n+1) + \frac{1}{3} \cdot 9^n} \leq \frac{9^n + 9^n \cdot (4+2n)}{\frac{1}{2} \cdot 9^n} = 4n+10 \ .$$

□

3.2.2 1-Decision Lists

The class of 1-decision lists has a teaching dimension of 2^n (cf. [44]). We use a result from Anthony, Brightwell, and Shawe-Taylor [9], which states the teaching dimension of linearly separable functions in dependence of the number of relevant variables. Since 1-decision lists are linearly separable and their number of relevant variables equals their length, we get the following lemma.

Lemma 3.12 *Let $n \in \mathbb{N}$ and $c \in \mathcal{D}_n$. Then $TD(c, \mathcal{D}_n) \leq (len(c)+1) \cdot 2^{n-len(c)}$.*

For concepts attaining this bound, see Lemma 4.20.

The teaching dimension of the concepts grows roughly exponentially as their length decreases. However, as we show in this section, the number of concepts of a certain length grows faster, thus leading to a small average teaching dimension. We denote the number of length-m concepts in $\mathcal{D}_n$ by A_m^n.

In order to determine A_m^n it suffices to count the number of inequivalent NFDLs of length m. To do so, we derive a criterion for the equivalence of two NFDLs.

One half of the criterion can be seen easily: If consecutive nodes with the same label are permuted, the represented concept remains the same. However, the converse is not true, even when only reduced decision lists are considered. For example, the reduced decision lists $\langle (v_1, 0), (v_2, 1), (*, 0)\rangle$ and $\langle (\bar{v}_2, 0), (\bar{v}_1, 1), (*, 0)\rangle$ are equivalent, but cannot be transformed into one another by permuting consecutive nodes with the same label. On the other hand, if only NFDLs are considered, a converse does hold.

Definition 3.13 *Two segments $G = \langle (w_1^{\alpha_1}, b), \ldots, (w_\ell^{\alpha_\ell}, b)\rangle$ and $H = \langle (y_1^{\beta_1}, b'), \ldots, (y_r^{\beta_\ell}, b')\rangle$ are called* similar *(denoted $G \sim H$) iff they contain the same nodes, that is,*

$$\{(w_1^{\alpha_1}, b), \ldots, (w_\ell^{\alpha_\ell}, b)\} = \{(y_1^{\beta_1}, b'), \ldots, (y_r^{\beta_\ell}, b')\} \ .$$

The next lemma presents the equivalence criterion for NFDLs.

Lemma 3.14 *Two decision lists D and E in normal form are equivalent, that is, $c_D = c_E$, if and only if they have the same sequence of node labels and for their segmentations $D = D_1 \circ \cdots \circ D_r$ and $E = E_1 \circ \cdots \circ E_r$ it holds $D_i \sim E_i$ for all i.*

Proof. For the "if" part, let $D = D_1 \circ \cdots \circ D_r$ and $E = E_1 \circ \cdots \circ E_r$ be two NFDLs with equal label sequences and equivalent segments. Let $1 \leq i \leq r$ and let $x \in \{0,1\}^n$ be an instance.

Assume x is absorbed by a node (w, b) in segment D_i. Then it is not absorbed in $D_1 \circ \cdots \circ D_{i-1}$ and neither in $E_1 \circ \cdots \circ E_{i-1}$. Instead it is absorbed by the node (w, b) which also occurs in segment E_i. Since x was arbitrary, it follows $c_D(x) = c_E(x)$ for all $x \in \{0,1\}^n$.

For the "only if" part, let D and E be two equivalent NFDLs.

Case 1: $len(D) = 0$.

Then D consists only of a default node $(*, b)$. This is obviously a unique NFDL representation. Therefore D and E are the same 1-decision list. □ Case 1

Case 2: $len(D) = 1$.

Then $D = \langle (w, 1), (*, 0) \rangle$ for some literal w (recall the special definition of NFDL for length-1 lists). Again, this is a unique NFDL representation and D and E are identical. □ Case 2

Case 3: $len(D) \geq 2$.

Let $b_1, \ldots, b_m$ be the label sequence of D (with default label b_{m+1}). The default node and the node before both absorb 2^{n-m} instances. Since one of the nodes is labeled positive, both nodes together classify 2^{n-m} instances as positive. The number of instances positively classified by the other nodes in D is $\sum_{i=1}^{m-1} b_i 2^{n-i}$.

Since E is equivalent to D, it must have the same number of positive instances and therefore the same label sequence $b_1, \ldots, b_{m-1}$. Moreover, the definition of NFDL requires $b_m = b_{m-1}$ and therefore the last node in D and in E are both labeled b_m. Overall, D and E have the same label sequence. Hence both lists are divided into the same number of segments, $D = D_1 \circ \cdots \circ D_r$ and $E = E_1 \circ \cdots \circ E_r$, and corresponding segments are of equal length and have the same label.

It remains to prove $D_i \sim E_i$ for all $i \leq r$. We show:

Claim: If the instances reaching D_i are the same as those reaching E_i, then $E_i \sim D_i$ and the instances leaving D_i are the same as those leaving E_i.

Proof. Let $Z \subseteq \{0,1\}^n$ be the set of instances reaching both D_i and E_i. Let $D_i = \langle (w_1^{\alpha_1}, b), \ldots, (w_s^{\alpha_s}, b) \rangle$ and $E_i = \langle (y_1^{\beta_1}, b), \ldots, (y_s^{\beta_s}, b) \rangle$ with $b \in \{0,1\}$. Let $\langle (y^{\beta}, \bar{b}) \rangle$ be the first node after E_i in E (or $(*, \bar{b})$ if D_i is the last segment).

Suppose that there is a literal $w_k^{\alpha_k}$ in D_i that is missing in E_i. Since none of the variables $w_1, \ldots, w_s, y_1, \ldots, y_s, y$ appears in any of the first $i-1$ segments (since D and E are NFDLs), for every truth assignment to these variables there must be an instance in Z. Let $x \in Z$ be an assignment satisfying y^{β} (or the default node if D_i is the last segment) and $w_k^{\alpha_k}$, but none of the $y_j^{\beta_j}$. This x is then classified as b by D and as $\bar{b}$ by E, a contradiction. It follows that every literal in D_i is also contained in E_i and vice versa by an analogous argument. Therefore both segments are permutations of each other, that is, $D_i \sim E_i$. This means that D_i and E_i absorb the same instances, hence the same instances leave D_i and E_i. □ Claim

Since all instances in $\{0,1\}^n$ reach D_1 and E_1, the claim can be used to prove by induction that $D_i \sim E_i$ for all i. □ Case 3

□

Now that we can use Lemma 3.14 to recognize inequivalent NFDLs, we can analyze A_m^n more closely.

Lemma 3.15 *For all n and $2 \leq m \leq n$, $A_m^n \geq 2(n-m+1) \cdot A_{m-1}^n$.*

Proof. Let $n \geq 2$.

There are $2n$ NFDLs of length 1, namely $\langle (v_i^\alpha, 1), (*, 0) \rangle$ for $i \in \{1, \ldots, n\}$ and $\alpha \in \{0, 1\}$. Hence $A_1^n = 2n$.

We now consider the case $m = 2$. A normal form decision list of length 2 has the shape $\langle (w_1, 1), (w_2, 1)(*, 0) \rangle$ or $\langle (w_1, 0), (w_2, 0)(*, 1) \rangle$. There are $2 \cdot \binom{n}{2}$ ways to choose two different literals over n variables. Every selection of two literals yields one decision list of each shape. All these lists are mutually inequivalent because they have pairwise non-similar segments. Therefore $A_2^n = 2 \cdot 2 \cdot \binom{n}{2} = 2(n-1) \cdot A_1^n$.

Now let $m > 2$. We show that each of the A_{m-1}^n pairwise inequivalent NFDLs of length $m-1$ can be extended to an NFDL of length m in $2(n-m+1)$ ways such that all extended lists are mutually inequivalent.

Let D be an NFDL of length $m-1$ and let b_1 be the label of the first node. The list D contains exactly $m-1$ different variables, hence there are $n-m+1$ variables left. By prepending each of these variables (negated or not) as nodes with label $\bar{b}_1$ we get $2(n-m+1)$ new NFDLs of length m. The prepended node certainly forms a segment of its own, because its label is different from that of the second node.

In this way we get $2(n-m+1) \cdot A_{m-1}^n$ NFDLs of length m. These are all mutually inequivalent since they either differ in their first segment, or if their first nodes are equal they are extension of two already inequivalent NFDLs. □

We need the lemma in the following form in which the number of NFDLs of a certain length is related to the total number $\sum_{m'=1}^n A_{m'}^n$ of NFDLs.

Corollary 3.16 *For $n \geq 2$ and $1 \leq m \leq n$: $\sum_{m'=1}^n A_{m'}^n \geq 2^{n-m} \cdot (n-m)! \cdot A_m^n$.*

Theorem 3.17 *The average teaching dimension of $\mathcal{D}_n$ is linear in n, $\overline{TD}(\mathcal{D}_n) \leq O(n)$.*

Proof. We first prove the statement for the concept class of NFDLs of length at least 1. Then we argue that the inclusion of the missing concepts, $\emptyset$ and $\{0,1\}^n$, does not matter.

Using Lemma 3.12 we bound the average teaching dimension from above by

$$\frac{\sum_{m=1}^n (m+1) 2^{n-m} \cdot A_m^n}{\sum_{m=1}^n A_m^n} = \sum_{m=1}^n (m+1) 2^{n-m} \cdot \frac{A_m^n}{\sum_{m'=1}^n A_{m'}^n} \; .$$

Now we apply Corollary 3.16 to the fraction and get a new upper bound of

$$\sum_{m=1}^n (m+1) 2^{n-m} \cdot \frac{1}{2^{n-m} \cdot (n-m)!} = \sum_{m=1}^n \frac{m+1}{(n-m)!} = \sum_{m=0}^{n-1} \frac{n+1-m}{m!} \; .$$

To see that this value grows linearly in n, we divide by n and obtain

$$\frac{1}{n}\sum_{m=0}^{n-1}\frac{n+1-m}{m!} \leq \frac{n+1}{n}\sum_{m=0}^{n-1}\frac{1}{m!}$$

which converges to Euler's number as $n \to \infty$.

Since $|\mathcal{D}_n|$ grows faster than 2^n, the two missing concepts amount only to a fraction of less than 2^{1-n} of all concepts. Their teaching dimension of 2^n increases the average teaching dimension therefore by less than 2. □

3.3 Discussion

The average teaching dimension is a more plausible measure for the teachability of concept classes as a whole than the (worst case) teaching dimension. To support this claim we showed that 2-term DNFs and 1-decision lists have an average teaching dimension only *linear* in the number of variables. This seems more plausible than their *exponential* worst case teaching dimension.

The teaching dimension is not only used to measure the teachability for classes, but also for the teachability of *individual* concepts inside a class. Of course these individual teaching dimensions can be implausible, too. For example, Lemma 3.12 suggests that longer decision lists are easier to teach than shorter ones, which is clearly not realistic. But the average teaching dimension does not affect the teachabilities of individual concepts inside a class.

In the next chapter we therefore tackle the question of individual teaching dimensions and investigate ways to judge the teachability of single concepts more plausibly.

Chapter 4

Learners With Restricted Admissible Hypothesis Spaces

The learners discussed in the last chapter are not all particularly clever. They only satisfy some minimum requirements for learners: "Remember what you are taught" and "think of a consistent hypothesis." Typically, there are many consistent hypotheses available and most of them are just unreasonable. The TD model measures the difficulty of teaching the most unreasonable learners. Teaching those, however, is often hard, which is one reason for the teaching dimension sometimes being implausibly high.

In this chapter we consider learners that choose their hypotheses among other (and typically smaller) spaces than the space of all consistent hypotheses. For every sample S of memorized examples we specify a space $\mathfrak{H}(S) \subseteq \mathcal{C}$ of *admissible hypotheses* from which the learner must choose. This includes the TD model as a special case in which $\mathfrak{H}(S) = \mathcal{C}(S)$ for all samples S.

The generalization of Definition 3.1 to arbitrary admissible hypothesis spaces is simple, but the generalization of Lemma 3.3 is straightforward, too. Therefore, the only improvement in realism we can expect from the new model is an improvement with respect to the plausibility of the induced teachability measure. Thus, our goal is to define the admissible hypothesis spaces such as to simulate smarter learners.

In this chapter we consider two natural ways to define the admissible hypothesis spaces. In the first one, only consistent hypotheses of least complexity are allowed. This remedies unrealistic relations between the complexity and the teachability of concepts. The optimal teaching time for 1-decision lists is now upper bounded by their length, whereas the teaching dimension may decrease for lists of increasing length (compare Lemma 3.12 and Lemma 4.20). A similar result holds for monomials (see Section 4.2). However, this teachability measure depends on the way the complexity of concepts is measured, rather than only on the concept class itself.

The second variant defines the admissible hypothesis space so as to simulate learners that assume their teacher to be optimal. Theses learners disregard hypotheses for which the received examples would not be optimal. This yields a teachability measure that depends only on the concept class again, but whose values are much harder to calculate. For monomials the optimal teaching time equals their teaching dimension,

except that the empty concept can be taught in $n+2$ rounds, instead of 2^n rounds in the TD model. For certain 1-decision lists teaching can take as long as $\sqrt{n} \cdot 2^{n/2}$ rounds (see Section 4.3).

In the final Section 4.4 we introduce a new specification of (MEM), called *selective memory*. This can be combined with every definition of admissible hypothesis spaces. It yields our first teaching model in which the teacher benefits from feedback. The optimal teaching time without feedback is always quadratic in the optimal teaching time with feedback.

4.1 Description and Properties of the Model

The definition of the learners is based on the notion of *admissible hypothesis space*. Given a sample S, the set $\mathfrak{H}(S) \subseteq \mathcal{C}$ specifies among which hypotheses a learner knowing the examples in S may choose. Since we want a specification for all samples $S \subseteq \mathcal{X}$, we regard $\mathfrak{H}$ as a function $\mathfrak{H}\colon 2^{\mathcal{X}} \to 2^{\mathcal{C}}$. We call $\mathfrak{H}$ a *hypothesis restriction*. The following is a generalization of Definition 3.1 in which $\mathcal{C}(\cdot)$ is replaced by $\mathfrak{H}(\cdot)$.

Definition 4.1 *Let $\langle \varrho, \mathcal{C} \rangle$ be a represented concept class. Let $\mathfrak{H}\colon 2^{\mathcal{X}} \to 2^{\mathcal{C}}$ be a hypothesis restriction. A non-deterministic learner $\mathcal{L} = (\mathcal{L}_{mem}, \mathcal{L}_{hyp})$ using hypothesis space $\langle \varrho, \mathcal{C} \rangle$ is called* $\mathfrak{H}$-restricted *iff*

(HYP) $\forall \boldsymbol{s} \in \mathcal{X}^* \; \forall \sigma \in \varrho(\mathcal{C}) \cup \{\mathit{init}\} \; \forall z \in \mathcal{X} \; \forall (\boldsymbol{s}', \sigma') \in \mathcal{L}(\boldsymbol{s}, \sigma, z)$:

$$\sigma' \in \begin{cases} \{\sigma\} & \text{if } z \in \mathcal{X}(\varrho_\sigma), \\ \varrho(\mathfrak{H}(\boldsymbol{s}')) & \text{otherwise.} \end{cases}$$

Let $m \in \mathbb{N}^+ \cup \{\infty\}$. The learner is said to have memory size m *iff*

(MEM) $\mathcal{L}_{mem}(\boldsymbol{s}, \sigma, z) = \{\boldsymbol{s} \circ_m \langle z \rangle\}$.

Example 4.2 Let us again consider teaching concepts from the class $\mathcal{S}_n$. We set $\mathfrak{H}(S) = \mathcal{C}(S \cap ([1,n] \times \{1\}))$ for all $S \subseteq \mathcal{X}$, that is, the learner considers only positive examples when choosing the hypothesis. Now for every target $c^* \in \mathcal{S}_n$ the teacher has to give all positive examples since the most discriminative negative examples are without effect. The optimal teaching time is thus n for the all-concept and $n-1$ for the other concepts. ■

Every hypothesis restriction $\mathfrak{H}$ induces a dimensionality notion similar to the teaching dimension. As mentioned above, the teaching dimension is the special case in which $\mathfrak{H}(S) = \mathcal{C}(S)$ for all S.

Definition 4.3 *Let $\mathcal{C}$ be a concept class, and let $\mathfrak{H}\colon 2^{\mathcal{X}} \to 2^{\mathcal{C}}$ be a hypothesis restriction. We define the* $\mathfrak{H}$-dimension of c with respect to $\mathcal{C}$ *as*

$$\mathfrak{H}D(c,\mathcal{C}) = \min\{|S| \mid S \subseteq \mathcal{X} \text{ and } \mathfrak{H}(S) = \{c\}\} \ .$$

The characterization Lemma 3.3 can be generalized to arbitrary $\mathfrak{H}$-restricted learners:

Lemma 4.4 *Let $\langle \varrho, \mathcal{C}\rangle$ be a represented concept class and let $c^* \in \mathcal{C}$ be a target concept. Let $\mathfrak{H}\colon 2^{\mathcal{X}} \to 2^{\mathcal{C}}$ be a hypothesis restriction, and let $m \in \mathbb{N}^+ \cup \{\infty\}$. Then c^* is teachable to the $\mathfrak{H}$-restricted learner with memory size m if and only if $\mathfrak{H}D(c^*,\mathcal{C})$ is finite and $\mathfrak{H}D(c^*,\mathcal{C}) \leq m$. The optimal teaching time is $\mathfrak{H}D(c^*,\mathcal{C})$.*

Lemma 4.4 has the same consequences for $\mathfrak{H}$-restricted learners as Lemma 3.3 for non-restricted learners (see Page 45): feedback is of no use, there is a weak form of memory-size dependence, and the order of examples is irrelevant. These consequences are independent of the hypothesis restriction. But still we can try to define the hypothesis restriction $\mathfrak{H}$ so as to get a more plausible teachability measure. Indeed we show in the following sections that for suitable hypothesis restrictions the teaching times for concepts can become more meaningful.

Before we define our first concrete hypothesis restriction we present some general considerations about the properties of such a restriction. After all, Definition 4.1 allows arbitrary functions $\mathfrak{H}$, but not all choices are natural. A natural hypothesis restriction should allow the learner to choose among consistent hypotheses only (just not necessarily from all of them). Another natural constraint is that the set of *all* examples for a concept leaves no choice to the learner but to hypothesize this very concept. These conditions on $\mathfrak{H}$ can be expressed as

(H 1) For all $S \subseteq \mathcal{X}$: $\mathfrak{H}(S) \subseteq \mathcal{C}(S)$.

(H 2) For all $c \in \mathcal{C}$: $\mathfrak{H}(\mathcal{X}(c)) = \{c\}$.

As the only operation on the learner's memory is to add an example to it, the relation of $\mathfrak{H}(S)$ to $\mathfrak{H}(S \cup \{z\})$, or more generally of $\mathfrak{H}(S)$ and $\mathfrak{H}(S')$ to $\mathfrak{H}(S \cup S')$, is particularly important. The most straightforward condition is that the samples S and S' combine their power to eliminate hypotheses:

(H 3) For all $S, S' \subseteq \mathcal{X}$: $\mathfrak{H}(S \cup S') = \mathfrak{H}(S) \cap \mathfrak{H}(S')$.

This condition is satisfied by $\mathfrak{H}(S) = \mathcal{C}(S)$. In fact, (H 3) gives us, together with (H 1) and (H 2), a topological characterization for $\mathcal{C}$, regarded as a function $\mathcal{C}\colon S \mapsto \mathcal{C}(S)$, and thus indirectly for the teaching dimension.

Fact 4.5 *A hypothesis restriction $\mathfrak{H}\colon 2^{\mathcal{X}} \to 2^{\mathcal{C}}$ satisfies* (H 1), (H 2), (H 3) *if and only if for all $S \subseteq \mathcal{X}$: $\mathfrak{H}(S) = \mathcal{C}(S)$.*

Proof. The "if" direction is easily checked using the definition of $\mathcal{C}(S)$.

For the "only if" direction, let $\mathfrak{H}$ satisfy (H 1), (H 2), and (H 3). Since $\mathfrak{H}(S) \subseteq \mathcal{C}(S)$ holds by (H 1), it remains to show $\mathcal{C}(S) \subseteq \mathfrak{H}(S)$ for all $S \subseteq \mathcal{X}$. Let $S \subseteq \mathcal{X}$ be an arbitrary sample and let $c \in \mathcal{C}(S)$. Then

$$\{c\} = \mathfrak{H}(\mathcal{X}(c)) = \mathfrak{H}(S) \cap \mathfrak{H}(\mathcal{X}(c) \setminus S),$$

where the first equality holds by (H 1) and the second one by (H 3). It follows that $c \in \mathfrak{H}(S)$, which shows the "only if" direction. □

In order to get $\mathfrak{H}$-dimensions different from the teaching dimension, we have to relax one of the conditions (H 1), (H 2), or (H 3). Relaxing (H 1) would mean to allow also inconsistent hypotheses. But as we do not want to allow *all* inconsistent hypotheses, we would have to define the admissible inconsistent hypotheses; and there seems to be no natural and universally agreed way to do this. Therefore, we leave (H 1) untouched. Relaxing (H 2) would allow the learner, even if it knew everything about the target, to hypothesize a non-target concept. This also seems undesirable.

It thus remains to consider relaxations of (H 3). Two obvious variants are to replace the "$=$" by "$\subseteq$" or by "$\supseteq$". The former variant means that the unified sample $S \cup S'$ can be more powerful than the sum of its parts S and S'. The latter variant means that two samples might be unable to fully combine their power. Indeed, there are natural hypothesis restrictions satisfying either of these relaxations. In Section 4.2 we consider a natural hypothesis restriction $\mathfrak{H}$ satisfying $\mathfrak{H}(S \cup S') \supseteq \mathfrak{H}(S) \cap \mathfrak{H}(S')$ instead of (H 3). A natural choice of $\mathfrak{H}$ satisfying $\mathfrak{H}(S \cup S') \subseteq \mathfrak{H}(S) \cap \mathfrak{H}(S')$ instead of (H 3) is discussed in Section 4.3.

4.2 Learners Preferring Simple Hypotheses

Imagine a learner being taught a 2-term DNF and having received some examples for which there is a consistent monomial. Then, as discussed before, there are also exponentially many consistent 2-term DNFs and the learner could hypothesize any one of them until they are all eliminated. This elimination requires exponentially many further examples. A more reasonable learner, however, would rather hypothesize that consistent monomial because it is the simplest consistent hypothesis available.

Under the name *Occam's Razor*, the strategy of choosing the simplest consistent hypothesis is a common principle in learning theory, in particular in the PAC setting (see, for example, Blumer *et. al.* [21] and Natarajan [56]). The similar *Minimum Description Length Principle* by Rissanen [67] has been used to develop learning algorithms (see, for example, [31, 58]).

In this section we study learners that output hypotheses that are not only consistent, but also of minimum complexity. We call such a learner *complexity based*. A natural

choice for the complexity measure is the representation size $\|c\|$ of a concept c in a represented concept class. The admissible hypothesis spaces are defined as

$$\begin{aligned}\mathfrak{H}_{compl}(S) &= \{h \in \mathcal{C}(S) \mid \|h\| = \min\{\|h'\| \mid h' \in \mathcal{C}(S)\}\} \\ &= \operatorname*{argmin}_{h \in \mathcal{C}(S)} \|h\| \ .\end{aligned}$$

This definition gives rise to the $\mathfrak{H}_{compl}$-dimension which we call the *complexity teaching dimension* and abbreviate by *CTD* instead of the canonical name $\mathfrak{H}_{compl}D$. A *complexity teaching set* (or *CT-set* for short) for c is a sample S with $\mathfrak{H}_{compl}(S) = \{c\}$ and $CTD(c, \mathcal{C})$ is defined as $CTD(c, \mathcal{C}) = \min\{|S| \mid \mathfrak{H}_{compl}(S) = \{c\}\}$.

In order to teach a complexity based learner a concept c, a teacher has to provide enough examples to rule out all concepts with complexity less or equal to $\|c\|$; concepts with higher complexity need not be ruled out. The complexity teaching dimension can therefore be calculated as the teaching dimension with respect to the subclasses of concepts with bounded complexity:

Lemma 4.6 *Let $\langle \varrho, \mathcal{C} \rangle$ be a represented concept class and let $\mathcal{C}_k = \{c \in \mathcal{C} \mid \|c\| \leq k\}$. Then for all $c \in \mathcal{C}$: $CTD(c, \mathcal{C}) = TD(c, \mathcal{C}_{\|c\|})$.*

It follows from the previous lemma that the complexity teaching dimension of a concept $c \in \mathcal{C}$ with maximum complexity equals its teaching dimension with respect to $\mathcal{C}$. A more interesting consequence is that $CTD(c)$ is always finite, even for concepts over an infinite domain, because $\mathcal{C}_{\|c\|}$ is always finite. In this respect *TD* and *CTD* differ.

Corollary 4.7 *Let $\langle \varrho, \mathcal{C} \rangle$ be a represented concept class. Then we have $CTD(c, \mathcal{C}) < \infty$ for all $c \in \mathcal{C}$.*

Note that the complexity teaching dimension $CTD(\mathcal{C})$ of a class can still be infinite.

Before we go on with some examples of the *CT*-dimension, we remark that $\mathfrak{H}_{compl}$ satisfies not only (H 1) and (H 2), but also the following relaxation of (H 3) in which "=" is replaced by "$\supseteq$".

Fact 4.8 *Let $\langle \varrho, \mathcal{C} \rangle$ be a represented concept class. Then $\mathfrak{H}_{compl}(S \cup S') \supseteq \mathfrak{H}_{compl}(S) \cap \mathfrak{H}_{compl}(S')$ for all $S, S' \subseteq \mathcal{X}$.*

Proof. Let $S, S' \subseteq \mathcal{X}$, and let $\ell = \min\{\|h\| \mid h \in \mathcal{C}(S)\}$, $\ell' = \min\{\|h\| \mid h \in \mathcal{C}(S')\}$, and $k = \min\{\|h\| \mid h \in \mathcal{C}(S \cup S')\}$. These values can be infinite, but certainly $\ell, \ell' \leq k$. We distinguish four cases which can all be solved easily.

Case 1: $k = \infty$.

Then $\mathcal{C}(S) \cap \mathcal{C}(S') = \emptyset$ and therefore $\mathfrak{H}_{compl}(S) \cap \mathfrak{H}_{compl}(S') \subseteq \mathcal{C}(S) \cap \mathcal{C}(S') = \emptyset \subseteq \mathfrak{H}_{compl}(S \cup S')$. □ Case 1

Case 2: $k < \infty$ and $\ell \neq \ell'$.

Then $\mathcal{C}(S) \cap \mathcal{C}(S') = \emptyset$ and thus $\mathfrak{H}_{compl}(S) \cap \mathfrak{H}_{compl}(S') = \emptyset \subseteq \mathfrak{H}_{compl}(S \cup S')$. □ Case 2

Case 3: $k < \infty$ and $\ell = \ell' = k$.

Then $\mathfrak{H}_{compl}(S \cup S') = \{h \in \mathcal{C}(S) \cap \mathcal{C}(S') \mid \|h\| = k\} = \{h \in \mathcal{C}(S) \mid \|h\| = \ell\} \cap \{h \in \mathcal{C}(S') \mid \|h\| = \ell'\} = \mathfrak{H}_{compl}(S) \cap \mathfrak{H}_{compl}(S')$. □ Case 3

Case 4: $k < \infty$ and $\ell = \ell' \neq k$.

Then $\mathfrak{H}_{compl}(S) \cap \mathfrak{H}_{compl}(S') = \{h \in \mathcal{C}(S) \mid \|h\| = \ell\} \cap \{h \in \mathcal{C}(S') \mid \|h\| = \ell'\} = \{h \in \mathcal{C}(S) \cap \mathcal{C}(S') \mid \|h\| = \ell\} = \{h \in \mathcal{C}(S \cup S') \mid \|h\| = \ell\} = \emptyset$. The last equality holds because $k > \ell$. □ Case 4

□

4.2.1 2-Term DNFs

The complexity of a concept in $\mathcal{M}_n^2$ is essentially the minimal number of terms necessary to represent the concept. Formally,

$$\|c\| = \begin{cases} 0 & \text{if } c = \emptyset, \\ n & \text{if } c \in \mathcal{M}_n^1 \setminus \{\emptyset\}, \\ 2n & \text{if } c \in \mathcal{M}_n^2 \setminus \mathcal{M}_n^1. \end{cases}$$

We calculate the CT-dimensions for all concepts in $\mathcal{M}_n^2$ using Lemma 4.6. The empty concept has a CT-dimension of 1. The CT-dimension of non-empty concepts representable as monomial is at most $n + 1$. To complete the results for $\mathcal{M}_n^2$, we still have to determine the CT-dimension for the concepts in $\mathcal{M}_n^2 \setminus \mathcal{M}_n^1$. Since they have maximum complexity, their CT-dimension equals their teaching dimension.

Theorem 4.9 $CTD(\mathcal{M}_n^2) \leq 2n + 4$.

Proof. Let $c \in \mathcal{M}_n^2$. Lemma 3.8 and the discussion above show that

- if $\|c\| = 0$ then $TD(c) = 1$ with respect to $\{c \in \mathcal{M}_n^2 \mid \|c\| = 0\}$,
- if $\|c\| = n$ then $TD(c) \leq n + 1$ with respect to $\{c \in \mathcal{M}_n^2 \mid \|c\| \leq n\}$,
- if $\|c\| = 2n$ then $TD(c) \leq 2n + 4$ with respect to $\{c \in \mathcal{M}_n^2 \mid \|c\| \leq 2n\}$.

From Lemma 4.6 it follows that the CT-dimensions are 0, at most $n + 1$, and at most $2n + 4$ for c with $\|c\| = 0$, n, and $2n$, respectively. □

The last theorem illustrates that teaching concepts to the complexity based learner can be much faster than teaching them to the plain consistent learner.

v_1	v_2	v_3	v_4
0	0	0	0
0	0	1	1
0	1	0	1
1	0	0	1
1	1	1	0

v_1	v_2	v_3	v_4	v_5
0	0	0	0	0
0	0	0	1	1
0	0	1	0	1
0	1	0	0	1
0	1	1	1	0
1	0	0	0	1
1	0	1	1	0
1	1	0	1	0
1	1	1	0	0
1	1	1	1	1

Figure 4.1 *On the left, a* $(4,2)$*-universal set of strings. On the right, a* $(5,3)$*-universal set. Both sets are of minimum size since* $f(4,2) = 5$ *and* $f(5,3) = 10$*.*

The CT-dimension also gives a more plausible relation between teachability and complexity of a 2-term DNF than the teaching dimension does. For all $c, c' \in \mathcal{M}_n^2 \setminus \{\{0,1\}^n\}$ we have $\|c\| < \|c'\| \Longrightarrow TD(c) > TD(c')$, that is, the teaching dimension decreases as the concepts become more complex.

In contrast there are many concepts $c, c' \in \mathcal{M}_n^2$ with $\|c\| < \|c'\|$ and $CTD(c) < CTD(c')$, that is, the CT-dimensions and the complexities have the same relation.

4.2.2 1-Decision Lists

In this section we show that the complexity teaching dimension for 1-decision lists is linear in the number of variables. This will be a corollary of a more general theorem (Theorem 4.12) which we prove first and which we will also use again later (see Fact 5.10).

Definition 4.10 ([74, 15]) *Let* $0 \leq k \leq m$*. A set* B *of truth assignments to variables* $v_1, \ldots, v_m$ *is said to be* (m,k)-universal *iff for every set* $\{w_1, \ldots, w_k\}$ *of literals over* $v_1, \ldots, v_m$ *there is an assignment in* B *that satisfies all literals* $w_1, \ldots, w_k$*. The minimum size of a* (m,k)*-universal set is denoted by* $f(m,k)$*. We set* $f(m,0) = 1$*.*

An equivalent definition of $f(m,k)$ is the minimum number of binary strings of length m such that every binary subsequence of length k occurs in one of the strings. This makes the definition $f(m,0)$ reasonable because for the empty subsequence to occur there must be at least one string. See Figure 4.1 for some examples of (m,k)-universal sets.

Fact 4.11 1. $f(m,m) = 2^m$,

$\langle (v_1, 1) \quad (v_2, 0) \quad (v_3, 0) \quad (v_4, 0) \quad (v_5, 1) \quad (v_6, 1) \quad (*, 0) \rangle$

v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	label
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	1
0	0	0	0	1	0	0	0	0	0	1
0	0	0	1	1	1	0	0	0	0	0
0	0	1	0	1	1	0	0	0	0	0
0	1	0	0	1	1	0	0	0	0	0
1	1	1	1	1	1	0	0	0	0	1

Figure 4.2 *A 1-decision list of length 6 over 10 variables and a teaching set for that list with respect to the class of all 1-decision lists of length 6. Every row represents an example consisting of an instance from* $\{0,1\}^{10}$ *(interpreted as an assignment to the variables* $v_1, \ldots, v_{10}$*) and a label. Each node (including the default) absorbs exactly one example of the teaching set.*

2. $f(m,k) \leq k \cdot 2^k \cdot \log m \cdot \ln 2$.

Proof.

1. The subsequences of length m are just the binary strings of length m. And the minimum set that contains all strings of length m has cardinality 2^m.

2. This has been proved by Becker and Simon [15]. □

Theorem 4.12 *Let* $0 \leq \lambda \leq \ell \leq n \geq 1$ *and let* $c \in \mathcal{D}_n$ *with* $len(c) = \lambda$. *Then* $TD(c, \mathcal{D}^n_{len \leq \ell}) \leq (\lambda + 1) \cdot f(n - \lambda, \ell - \lambda)$.

Proof. Let c be represented by the NFDL

$$D = \langle (v_1, b_1), \ldots, (v_\lambda, b_\lambda), (*, \bar{b}_\lambda) \rangle$$

where we assume without loss of generality that the literals in the nodes are the variables $v_1, \ldots, v_\lambda$. Let the segmentation of D be $D_1 \circ \cdots \circ D_r \circ \langle (*, \bar{b}_\lambda) \rangle$ and let the label of the nodes in segment D_j be B_j. We call the variables not in D, that is, $v_{\lambda+1}, \ldots, v_n$, *irrelevant*.

Now we define a sample S whose cardinality is upper bounded by $(\lambda+1) \cdot f(n-\lambda, \ell-\lambda)$. Then we show that S is a teaching set for c with respect to $\mathcal{D}^n_{len \leq \ell}$ (see Figures 4.2 and 4.3 for two such samples).

The sample S contains exactly one example for each node in the segments $D_1, \ldots, D_{r-1}$ and $f(n - \lambda, \ell - \lambda)$ examples for each node in the segment D_r and for the default

$\langle(v_1, 1) \quad (v_2, 0) \quad (v_3, 0) \quad (v_4, 0) \quad (v_5, 1) \quad (v_6, 1) \quad (*, 0)\rangle$

v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	label
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	1	0
0	0	0	0	0	0	0	1	0	1	0
0	0	0	0	0	0	1	0	0	1	0
0	0	0	0	0	0	1	1	1	0	0
0	0	0	0	0	1	0	0	0	0	1
0	0	0	0	0	1	0	0	1	1	1
0	0	0	0	0	1	0	1	0	1	1
0	0	0	0	0	1	1	0	0	1	1
0	0	0	0	0	1	1	1	1	0	1
0	0	0	0	1	0	0	0	0	0	1
0	0	0	0	1	0	0	0	1	1	1
0	0	0	0	1	0	0	1	0	1	1
0	0	0	0	1	0	1	0	0	1	1
0	0	0	0	1	0	1	1	1	0	1
0	0	0	1	1	1	0	0	0	0	0
0	0	1	0	1	1	0	0	0	0	0
0	1	0	0	1	1	0	0	0	0	0
1	1	1	1	1	1	0	0	0	0	1

Figure 4.3 *A 1-decision list of length 6 over 10 variables. The examples constitute a teaching set with respect to all decision lists of length 8 over 10 variables. For the default node and each node in the last segment, there are five examples in the teaching set. The relevant variables $v_1, \ldots, v_6$ are assigned the same values as in Figure 4.2; the irrelevant variables v_7, v_8, v_9, v_{10} are assigned values such that all assignments for all sets of two variables occur (cf. the $(4, 2)$-universal set in Figure 4.1). For every other node the teaching set contains only one example, as in Figure 4.2.*

node. More precisely, for a node (v_i, B_j) in a segment D_j with $j < r$, the sample S contains the example (x, B_j) where the instance x (interpreted as an assignment to the variables $v_1, \ldots, v_n$) satisfies only the node's literal v_i and all literals in the segments $D_{j+1}, \ldots, D_r$; no other literal in D_j is satisfied by x and neither are the literals in $D_1, \ldots, D_{j-1}$. Which irrelevant variables x satisfies is arbitrary.

For a node (v_i, B_r) in segment D_r, the sample S contains examples $(x_1, B_r), \ldots, (x_{f(n-\lambda,\ell-\lambda)}, B_r)$ with the following property: All instances $x_1, \ldots, x_{f(n-\lambda,\ell-\lambda)}$ are equivalent with respect to the relevant variables: they only satisfy v_i. With respect to the $n-\lambda$ irrelevant variables $v_{\lambda+1}, \ldots, v_n$ the instances form an $(n-\lambda, \ell-\lambda)$-universal set. This means that for every possible assignment to every set of $\ell-\lambda$ irrelevant variables there is an instance in $\{x_1, \ldots, x_{f(n-\lambda,\ell-\lambda)}\}$ that coincides with this assignment on these $\ell-\lambda$ variables. For the default node $(*, \bar{B}_r)$ we include $f(n-\lambda, \ell-\lambda)$ examples in the same way as for the nodes in D_r: The instances satisfy none of the relevant variables, and with respect to the irrelevant variables the instances form an $(n-\lambda, \ell-\lambda)$-universal set.

We call a sample as defined above for an NFDL a *canonical sample* for this NFDL. Examples that are defined for a certain node are said to *belong* to that node. Clearly, to each node belong at most $f(n-\lambda, \ell-\lambda)$ examples and there are $\lambda+1$ nodes (including the default). This shows the claim about the cardinality of S.

Now we have to show that S is a teaching set for c with respect to $\mathcal{D}^n_{len\leq\ell}$. To this end, let E be an NFDL of length at most ℓ that is consistent with S. We have to show that E is equivalent to D. The segmentation of E is denoted $E_1 \circ \cdots \circ E_s$. In the following we show how S determines the nodes in E and that E and D are equal up to permutation of nodes within a segment.

We treat the special cases $\lambda = 0$ and $\lambda = 1$ first. The general proof is simplified if we do not need to pay attention to these cases.

If $\lambda = 0$ then D consists of a default node only, namely $(*, \bar{B}_r)$. If moreover $\ell = \lambda = 0$ then also E only consists of a default node, which must be the same as the default node of D. If however $\ell > \lambda = 0$ then S contains for every sequence of $\ell - \lambda = \ell$ literals over $v_1, \ldots, v_n$ an example satisfying all literals. Now suppose for a contradiction that E consists of more than $\lambda = 0$ nodes. Then there is a node (w, B_r) labeled B_r (if all nodes were labeled $\bar{B}_r$, the list E would not be in normal form). Consequently, the negated literals in the nodes before (w, B_r) and the literal w are a sequence of at most ℓ literals such that every example satisfying theses literals is absorbed by (w, B_r). But by definition, S contains an example that satisfies these at most ℓ literals and is labeled $\bar{B}_r$. This example is then classified as B_r by the node (w, B_r), a contradiction. It follows that E consists of a default node only.

The case $\lambda = 1$ is similar to $\lambda = 0$. Without loss of generality, let $D = \langle (v_1, 1), (*, 0) \rangle$. If in addition $\ell = \lambda = 1$ then E is of length 1 (it cannot contain a default node only, because there are differently labeled examples in S). Trying all possibilities, shows that the only length-1 decision lists consistent with S are $E = D$ and $E = \langle (\bar{v}_1, 0), (*, 1) \rangle$. But the latter is not in normal form, hence $E = D$.

If $\ell > \lambda = 1$ then the first node of E can be either $(v_1, 1)$ or $(\bar{v}_1, 0)$. In any case the node absorbs half the examples in S and the other half is an $(n-\lambda, \ell-\lambda)$-universal set with respect to the variables $v_2, \ldots, v_n$. Now an analogous reasoning as in the case $\lambda = 0$ shows that no sequence of $\ell - 1$ literals over $v_2, \ldots, v_n$ can come next. But E cannot

be longer than ℓ and therefore no node at all comes next. The list E thus ends with a default node and since it is in normal form, it can only be $E = \langle (v_1, 1), (*, 0) \rangle = D$.

For the general proof, we assume that $\lambda \geq 2$ and consequently that the last segment D_r of D contains at least two nodes.

Claim: (1) For all $j = 1, \ldots, r$: The set S_j of examples from S that reach the first node of segment E_j is a canonical teaching set for the NFDL $D_j \circ \cdots \circ D_r \circ \langle (*, \bar{b}_\lambda) \rangle$. (2) $E_1 \supseteq D_1, \ldots, E_{j-1} \supseteq D_{j-1}$.

Proof. The proof is by induction on j. As induction basis, let $j = 1$. Then $S_j = S_1$ is the original sample S which is a canonical sample for $D_1 \circ \cdots \circ D_r \circ \langle (*, \bar{b}_\lambda) \rangle$ by definition. Claim (2) holds trivially for $j = 1$.

Now assume the claim for some $j \in \{1, \ldots, s-1\}$. We consider segments D_j and E_j.

Let B be the label common to all nodes in D_j and let (w, b) be the first node in E_j. By the induction hypothesis, S_j is a canonical sample for $D_j \circ \cdots \circ D_r \circ \langle (*, \bar{b}_\lambda) \rangle$. The nodes in segment E_j must not contradict the examples in S_j. Therefore S_j rules out several possibilities for w and b:

1. w occurs in D_j and $b = \bar{B}$. Then $(w, \bar{b})$ is in D_j and the example belonging to this node contradicts (w, b) as first node in E_j.

2. $\bar{w}$ occurs in D_k with $k \in \{1, \ldots, r\}$. Then the example belonging to the default node satisfies w. Since there are at least two nodes in D_r, one example belonging to D_r satisfies w (if $k < r$ then all examples belonging to D_r satisfy w). But these two examples are labeled differently, hence one of them would be classified incorrectly by (w, b).

3. w occurs in D_k with $k \geq j+1$. Every example belonging to a node in D_j satisfies w and has label B. In D_{j+1} there is a node whose example satisfies w as well (if $k \neq 2$ all examples do) and has label $\bar{B}$. The node (w, b) is thus inconsistent with one of these two examples.

4. w does not occur in D and $w = \bar{v}$ for a variable v. Then the examples belonging to the default node and the node before satisfy w but are labeled differently.

In the case $\lambda < \ell$, that is, $f(n - \lambda, \ell - \lambda) > 1$, a stronger form of Item 4 applies:

5. w does not occur in D. Then one of the $f(n - \lambda, \ell - \lambda)$ examples belonging to the default node and one example belonging to the node before satisfy v but are labeled differently. The node (w, b) would thus be inconsistent with one of these variables.

We conclude that either (w, b) is a node from D_j or w is a variable not occurring in D (an irrelevant variable). In the latter case, we can repeat the above argumentation for the second and the following nodes in E_j until one of them is also in D_j. Only then does the list E constructed so far absorb one example from S_j, namely the example belonging to (w, b).

The above arguments can be repeated until all nodes in D_j, and hence all examples in S_j belonging to them, are used up. At this point E_j consists only of nodes from D_j and irrelevant nodes. Since only examples from nodes in D_j are absorbed by E_j, the remaining examples form a canonical sample for $D_{j+1} \circ \cdots \circ D_r \circ \langle (*, \bar{b}_\lambda) \rangle$. □ Claim

The claim says that after segment E_{r-1} all examples not yet absorbed (that is, the sample S_r) constitute a canonical teaching set for $D_r \circ \langle (*, \bar{b}_\lambda) \rangle$.

If we reason as above for the last segment of D, we would eventually reach a situation in which only one example is left. But then the argumentation in Item 2 would fail. Until this happens, however, Items 1–4 (or 1–3 and 5) apply. Now consider the situation in which all but one node of D_r, say (w, B_r), have been put into E_r.

Case 1: $\lambda = \ell$.

Then there are two examples from S not yet absorbed, namely that of (w, B_r) and that of the default node. Moreover, E so far contains at least $\ell - 1$ nodes (all nodes from D except (w, B_r)). On the other hand, E is of length at most ℓ. So at most one node may be added and at least one node has to be added because the two remaining examples have different labels. Now there are two ways for the segment E_r to end and be consistent with both examples: First, $(w, B_r), (*, \bar{B}_r)$ and second $(\bar{w}, \bar{B}_r), (*, B_r)$. Both endings, however, lead to equivalent 1-decision lists because every instance $x \in \{0, 1\}^n$ reaching the last node is classified as B if it satisfies w and as $\bar{B}$ otherwise. Therefore, we can assume the first ending, which means that also the last node in E_r is from D_r. □ Case 1

Case 2: $\lambda < \ell$.

Then there are the examples for (w, B_r) left and those for the default node. Moreover no irrelevant variable has been introduced into E so far (due to Item 5). There are two possible next nodes for E_r. First (w, B_r), second $(\bar{w}, \bar{B}_r)$. In both cases the remaining examples are a $(n - \lambda, \ell - \lambda)$-universal set with respect to the irrelevant variables. This prevents any sequence of $\ell - \lambda$ nodes with irrelevant variables to be appended. It does not prevent the appending of more then $\ell - \lambda$ irrelevant nodes, but if this happens the resulting list E would be longer than ℓ, a contradiction. Therefore, no irrelevant nodes are appended, and since the relevant variables are all used up, the list E ends at this point with a default node. This means that of the two possible nodes, only (w, B_r) is really possible because in normal form decision lists, like E, the two nodes before the default node have the same label. □ Case 2

We have shown that D and E differ only by permutations of nodes within corresponding segments, which means $c_D = c_E$. □

We can draw conclusions from the previous theorem by setting ℓ and λ to certain values. For example, if we set $\ell = n$ then we get with Fact 4.11:

$$TD(c, \mathcal{D}^n_{len\leq\ell}) = TD(c, \mathcal{D}_n) \leq (len(c)+1) \cdot f(n - len(c), n - len(c)) \\ = (len(c)+1) \cdot 2^{n-len(c)}.$$

This is just the bound from Lemma 3.12.

Corollary 4.13 *$CTD(\mathcal{D}_n) \leq n+1$ and for all $c \in \mathcal{D}_n$: $CTD(c) \leq len(c)+1$.*

Proof. Let $c \in \mathcal{D}_n$. Then $CTD(c, \mathcal{D}_n) = TD(c, \mathcal{D}^n_{len\leq len(c)})$. From Theorem 4.12 it follows $CTD(c, \mathcal{D}_n) \leq (len(c)+1) \cdot f(n - len(c), 0) = len(c)+1$. □

Teaching 1-decision lists to the complexity based learner requires much fewer examples than teaching them to the plain consistent learner (compare Lemma 3.12 and Corollary 4.13). Moreover, Corollary 4.13 suggests that the difficulty of teaching grows with the complexity of the concepts. Again, just as for monomials, the CT-dimension yields more intuitive results than the teaching dimension.

Unlike the teaching dimension the CT-dimension is not a combinatorial parameter of the concept class $\mathcal{C}$ alone, but depends on the complexity measure $\|\cdot\|$, that is, ultimately on the representation function ϱ. Although there are often natural choices for ϱ, for example, the size of a minimal standard representation for concepts in $\mathcal{C}$, it would be nice to be able to measure teachability without somewhat arbitrary additional assumptions.

4.3 Learners Assuming Optimal Teachers

In this section we investigate another reasonable behavior for learners. This will yield a teachability measure that depends only on the concept class.

Consider the class $\mathcal{S}_n$ and assume the teacher has given a positive example. Although this example eliminates only one hypothesis, it strongly suggests that the target concept is *not* a co-singleton one, since an optimal teacher would have used the unique negative example in this case. Thus, a learner who believes that the teacher is optimal would immediately hypothesize the all-concept, which would therefore be teachable in only one round instead of n rounds.

More generally speaking, an optimal teacher does not give superfluous examples. In other words, in each round the set S of examples given by an optimal teacher so far can be extended to a minimum teaching set for the target. All hypotheses with the property that every minimum teaching set is not a superset of S could be ignored by

the learners. But this requires the learners to know all minimum teaching sets of all concepts, which seems quite demanding.

It is less demanding to require that the learners know only the teaching dimensions of all concepts. This knowledge allows the learners to ignore all hypotheses whose teaching dimensions are smaller than $|S|$. After all, an optimal teacher would not give $|S|$ examples if the target had a teaching dimension of less than $|S|$.

Definition 4.14 *Let $\mathcal{C}$ be a concept class over X, and let $S \subseteq \mathcal{X}$ be a sample. We define*

$$\mathfrak{H}_{opt}(S) = \{h \in \mathcal{C}(S) \mid TD(h, \mathcal{C}) \geq |S|\} .$$

A set S with $\mathfrak{H}_{opt}(S) = \{c\}$ is called an OT-set for c. The optimal teacher teaching dimension *for a concept $c \in \mathcal{C}$ with respect to $\mathcal{C}$ is then defined as the size of the smallest OT-set for c:*

$$OTD(c, \mathcal{C}) = \min\{|S| \mid \mathfrak{H}_{opt}(S) = \{c\}\},$$

and for the class $\mathcal{C}$ as $OTD(\mathcal{C}) = \max\{OTD(c, \mathcal{C}) \mid c \in \mathcal{C}\}$.

The hypothesis restriction $\mathfrak{H}_{opt}$ is easily seen to satisfy (H 1) and (H 2). Moreover, it satisfies the following relaxation of (H 3).

Fact 4.15 *$\mathfrak{H}_{opt}(S \cup S') \subseteq \mathfrak{H}_{opt}(S) \cap \mathfrak{H}_{opt}(S')$ for all $S, S' \subseteq \mathcal{X}$.*

Proof. Let $S, S' \subseteq \mathcal{X}$. Then

$$\begin{aligned}
\mathfrak{H}_{opt}(S \cup S') &= \{h \in \mathcal{C}(S \cup S') \mid TD(h) \geq |S \cup S'|\} \\
&= \{h \in \mathcal{C}(S) \mid TD(h) \geq |S \cup S'|\} \cap \{h \in \mathcal{C}(S') \mid TD(h) \geq |S' \cup S'|\} \\
&\subseteq \{h \in \mathcal{C}(S) \mid TD(h) \geq |S|\} \cap \{h \in \mathcal{C}(S') \mid TD(h) \geq |S'|\} \\
&= \mathfrak{H}_{opt}(S) \cap \mathfrak{H}_{opt}(S') .
\end{aligned}$$

□

Intuitively, the job of an OT-set for a concept c is to eliminate every concept $c' \neq c$ either by proving that c' is inconsistent with the examples or by being larger than $TD(c')$. We refer to these two ways as to *elimination by inconsistency* and *elimination by size*, respectively. In contrast, to become a teaching set for c, a sample must eliminate all concepts by inconsistency alone. It is therefore clear that every teaching set is an OT-set and that $OTD(c, \mathcal{C}) \leq TD(c, \mathcal{C})$ for all concepts $c \in \mathcal{C}$ and all classes $\mathcal{C}$.

The class $\mathcal{S}_n$ shows that there can be big differences between TD and OTD. A set of two positive examples eliminates all co-singleton concepts by size as their teaching dimension is 1. Thus, $OTD([1, n], \mathcal{S}_n) = 2 < n = TD([1, n], \mathcal{S}_n)$.

The OT-dimension for a class $\mathcal{C}$ can be calculated using the teaching dimension of certain subclasses $\mathcal{C}_{\geq i} := \{c \in \mathcal{C} \mid TD(c) \geq i\}$ for $i = 1, \ldots, |X|$.

Lemma 4.16 *For all concept classes $\mathcal{C}$, $OTD(\mathcal{C}) = \min\{j \mid j \geq TD(\mathcal{C}_{\geq j})\}$.*

Proof. Set $J = \min\{j \mid j \geq TD(\mathcal{C}_{\geq j})\}$. We first show $OTD(c) \leq J$ for all $c \in \mathcal{C}$. For c with $TD(c) \leq J$ we have $OTD(c) \leq J$ since every teaching set is an OT-set. Now let c be such that $TD(c) > J$. Then $c \in \mathcal{C}_{\geq J}$ and $TD(\mathcal{C}_{\geq J}) \leq J$. Thus there is a set S of size at most J that uniquely describes c with respect to $\mathcal{C}_{\geq J}$. This S eliminates by size all c with $TD(c) < J$, hence it is an OT-set for c. It follows that $OTD(c) \leq J$.

Now we show that there is a concept c with $OTD(c) \geq J$. From the definition of J it follows that $J - 1 < TD(\mathcal{C}_{\geq J-1})$. Hence there is a $c \in \mathcal{C}_{\geq J-1}$ whose teaching dimension with respect to $\mathcal{C}_{\geq J-1}$ is at least J. Let S be an example set of size $J - 1$. Then S eliminates no concept in $\mathcal{C}_{\geq J-1}$ by size and it is also too small to uniquely describe c with respect to $\mathcal{C}_{\geq J-1}$. Thus S is no OT-set for c and $OTD(c) \geq J$. □

Corollary 4.17 *Let $\mathcal{C}$ be a concept class and let g_i, G_i be such that $g_i \leq TD(\mathcal{C}_{\geq i}) \leq G_i$ for all i. Then* $\max\{j \mid j < g_j\} \leq OTD(\mathcal{C}) \leq \min\{j \mid j \geq G_j\}$.

4.3.1 Teaching Monomials and 1-Decision Lists

For concept classes that have a single concept with a much higher teaching dimension than the other concepts, the OT-dimension for that concept drops down to one plus the second largest teaching dimension. This happens for the monomials as well.

Theorem 4.18 *Let $n \geq 2$. Then for all $c \in \mathcal{M}_n^1 \setminus \{\emptyset\}$, $OTD(c, \mathcal{M}_n^1) = TD(c, \mathcal{M}_n^1) = \min\{k + 2, n + 1\}$, but $OTD(\emptyset, \mathcal{M}_n^1) = n + 2$.*

Proof. For symmetry reasons it suffices to consider the monomials of the form $c = 1^k *^{n-k}$ for $k \in \{0, \ldots, n\}$ and the monomial $\emptyset$.

Case 1: $k = 0$.

Then $TD(c) = 2$. A sample S with $|S| < 2$ cannot eliminate any monomials by size since the teaching dimension of every monomial is at least two. The sample S cannot eliminate enough concepts by inconsistency either, otherwise $TD(c)$ would be $|S| < 2$. Therefore, there is no OT-set for c of size less than two. □ Case 1

Case 2: $1 \leq k \leq n - 1$.

Then $TD(c) = k + 2$. Let S be a sample with $|S| = k + 1$. We show that S is no OT-set for c.

Case 2.1: S contains no positive example.

Then S does not eliminate $\emptyset$ by inconsistency and neither by size, as $TD(\emptyset) = 2^n$. Therefore, S is no OT-set for c.

Case 2.2: S contains exactly one positive example, say $(x, 1)$.

Then the *monomial* whose representation is x has a teaching dimension of $n+1 > |S|$ and is consistent with S. Thus S eliminates that monomial neither by size nor by inconsistency. Therefore, S is no OT-set for c.

Case 2.3: S contains at least two positive examples.

Then S contains at most $k-1$ negative examples and there is an $i \in \{1,\ldots,k\}$ such that for all $\mu \in \{0,1\}^{n-k}$ the negative example $(1^{i-1}01^{k-i}\mu, 0) \in \mathcal{X}(c)$ is not in S. The monomial $c' = \mathtt{1}^{i-1}\mathtt{*}\mathtt{1}^{k-i}\mathtt{*}^{n-k}$ has a teaching dimension of $k+1$ and is thus not eliminated by S by size. It is also not eliminated by inconsistency as it is consistent with S: First, $c' \supset c$, hence c' is consistent with all positive examples in S. Second, all instances in $c' \setminus c$ are of the form $1^{i-1}01^{k-i}\mu$ with $\mu \in \{0,1\}^{n-k}$, but no such instance occurs in a negative example in S (by choice of i); therefore all negative examples in S are also negative examples for c'. Since S does not eliminate c', it is no OT-set for c. □ Case 2

Case 3: $k = n$.

Then $(1^n, 1)$ is the only positive example for c and $TD(c) = n+1$. Let S be a sample with $|S| = n$. We show that S is no OT-set for c.

Case 3.1: S contains no positive example.

Then S does not eliminate $\emptyset$ by inconsistency and neither by size, as $TD(\emptyset) = 2^n$.

Case 3.2: S contains the positive example.

Then S contains $n-1$ negative examples and there is an $i \in \{1,\ldots,n\}$ such that the negative example $(1^{i-1}01^{n-i}, 0)$ is not in S. But then the monomial $c' = \mathtt{1}^{i-1}\mathtt{*}\mathtt{1}^{n-i}$ is not eliminated by S: First, $TD(c') = n+1$ as c' has $n-1$ literals; thus no elimination by size occurs. Second, c' is consistent with S because $c' = \{1^n, 1^{i-1}01^{n-i}\}$ and by choice of i the example $(1^{i-1}01^{n-i}, 0)$ is not in S. □ Case 3

Case 4: $c = \emptyset$.

Every sample containing $n+2$ negative examples is an OT-set for $\emptyset$ because all other concepts have a teaching dimension of at most $n+1$ and are thus eliminated by size. On the other hand, let S contain only $n+1$ negative examples. Then there is an instance $x \in \{0,1,\}^n$ that does not occur in any example in S (for $n \geq 2$ we have $2^n > n+1$). The monomial with representation x has a teaching dimension of $n+1$, is consistent with S, and thus not eliminated by S. This shows $OTD(\emptyset) = n+2$. □ Case 4

□

For the class of monomials, the OT-dimension, the CT-dimension, and the average teaching dimension are all linear in n. This appears to be more reasonable than the teaching dimension, which is exponential in n.

The class of 1-decision lists has quite a different distribution of teaching dimensions than the class of monomials. Instead of a single high-dimensional concept, $\mathcal{D}_n$ contains

many concepts with high, medium, and low teaching dimensions. Nevertheless, the average teaching dimension and the CT-dimension are only linear in n. The OT-dimension, however, is much larger, as we want to show next.

To analyze the OT-dimension of 1-decision lists we need some notations, in which we omit the subscript n for readability. We define

$$\mathcal{D}_{len\leq i} = \{c \in \mathcal{D}_n \mid len(c) \leq i\},$$
$$\mathcal{D}_{\geq j} = \{c \in \mathcal{D}_n \mid TD(c) \geq j\}.$$

By $\mathcal{D}^+$ we denote the set of all concepts representable by NFDLs whose only positive node is the default node, for example, $\langle (v_4, 0), (\bar{v}_2, 0), (v_1, 0), (v_3, 0), (*, 1)\rangle$. Finally, we set as shortcut

$$\begin{aligned}\ell_j &= \max\{\ell \mid (\ell+1)\cdot 2^{n-\ell} \geq j\} \\ &= \min\{\ell \mid (\ell+1)\cdot 2^{n-\ell} < j\} - 1\end{aligned}$$

for $j = 0, \ldots, 2^n$. In other words, ℓ_j is the minimal length such that all longer 1-decision lists have a teaching dimension of less than j (compare Lemma 3.12).

To apply Corollary 4.17 we derive bounds for the teaching dimension of $\mathcal{D}_{\geq i}$. We start with an upper bound.

Lemma 4.19 *$TD(\mathcal{D}_{\geq j}) \leq \ell_j \cdot 2^{\ell_j} \cdot \log n \cdot \ln 2$.*

Proof. First we prove $\mathcal{D}_{\geq j} \subseteq \mathcal{D}_{len\leq \ell_j}$. Let $c \in \mathcal{D}_{\geq j}$ and suppose that $len(c) \geq \ell_j + 1$. Then $TD(c) \leq (\ell_j + 2)\cdot 2^{n-(\ell_j+1)}$ and $TD(c) \geq j$, hence $j \leq (\ell_j + 2)\cdot 2^{n-(\ell_j+1)}$, a contradiction to the definition of ℓ_j.

To prove the lemma, it suffices to show $TD(\mathcal{D}_{len\leq \ell}) \leq \ell \cdot 2^\ell \cdot \log n \ln 2$ for all ℓ, in particular for $\ell = \ell_j$. Let $c \in \mathcal{D}_{len\leq \ell}$ with $len(c) = \lambda$. Then it follows from Theorem 4.12 and Fact 4.11 that

$$TD(c, \mathcal{D}_{len\leq \lambda}) \leq (\lambda+1)\cdot f(n-\lambda, \ell-\lambda) \leq (\lambda+1)\cdot(\ell-\lambda)\cdot 2^{\ell-\lambda}\cdot \log(n-\lambda)\cdot \ln 2.$$

The expression on the right is upper bounded by $\ell \cdot 2^\ell \cdot \log n \cdot \ln 2$ because $(\lambda+1)\cdot 2^{\ell-\lambda}$ is upper bounded by 2^ℓ for $0 \leq \lambda \leq \ell$. Putting it all together and setting $\ell = \ell_j$, we get

$$TD(c, \mathcal{D}_{len\leq \lambda}) \leq \ell_j \cdot 2^{\ell_j} \cdot \log n \cdot \ln 2$$

for all $c \in \mathcal{D}_{len\leq \ell_j}$. It follows that $TD(\mathcal{D}_{\geq j}) \leq \ell_j \cdot 2^{\ell_j} \cdot \log n \cdot \ln 2$. □

The next lemma is needed for our lower bound on $TD(\mathcal{D}_{\geq i})$ and moreover presents concepts whose teaching dimensions attain the upper bound stated in Lemma 3.12.

Lemma 4.20 *Let $c \in \mathcal{D}^+$. Then $TD(c, \mathcal{D}_n) = (len(c)+1)\cdot 2^{n-len(c)}$.*

Proof. Because of Lemma 3.12 we only need to show $TD(c) \geq (len(c)+1) \cdot 2^{n-len(c)}$. Without loss of generality, let $c \in \mathcal{D}^+$ be represented using the variables $v_1, \ldots, v_\ell$ only, that is, by $\langle (v_1, 0), \ldots, (v_{\ell-1}, 0), (v_\ell, 0), (*, 1)\rangle$.

We prove that there are at least $(\ell+1) \cdot 2^{n-\ell}$ neighbor concepts of c in the class of *all* 1-decision lists. Of these neighbor concepts $\ell \cdot 2^{n-\ell}$ are represented by lists of the form

$$\langle (v_{i_1}, 0), \ldots, (v_{i_{\ell-1}}, 0), (\bar{v}_{i_\ell}, 1), (v_{\ell+1}^{\alpha_{\ell+1}}, 0), \ldots, (v_n^{\alpha_n}, 0), (*, 1)\rangle$$

where v_{i_ℓ} ranges over all ℓ variables $v_1, \ldots, v_\ell$ and $\{i_1, \ldots, i_\ell\} = \{1, \ldots, \ell\}$. Each of the $2^{n-\ell}$ combinations of the α's yields a 1-decision list representing a concept containing exactly one instance more than c. There are $\ell \cdot 2^{n-\ell}$ such lists, which represent just as many pairwise different concepts.

There are another $2^{n-\ell}$ neighbor concepts of c represented by decision lists of the form

$$\langle (v_1, 0), \ldots, (v_{\ell-1}, 0), (v_\ell, 0), (v_{\ell+1}^{\alpha_{\ell+1}}, 1), \ldots, (v_n^{\alpha_n}, 1), (*, 0)\rangle.$$

All these lists represent pairwise different concepts each of which contains exactly one instance less than c.

Overall there are $(\ell+1) \cdot 2^{n-\ell}$ neighbor concepts of c. □

Lemma 4.21 *$TD(\mathcal{D}_{\geq j}) \geq 2^{\ell_j - 1}$ for all $j = 1, \ldots, 2^n$.*

Proof. All concepts in $c \in \mathcal{D}^+$ with $len(c) \leq \ell_j - 1$ are also in $\mathcal{D}_{\geq j}$. This follows from Lemma 4.20 because $TD(c) \geq (len(c)+1) \cdot 2^{n-len(c)} = \ell_j \cdot 2^{n-\ell_j+1} \geq (\ell_j+1) \cdot 2^{n-\ell_j} \geq j$. The last inequality holds by definition of ℓ_j.

In particular, all concepts representable by decision lists of the form

$$\langle (v_1^{\alpha_1}, 0), \ldots, (v_{\ell_j-2}^{\alpha_{\ell_j-2}}, 0), (v_{\ell_j-1}^{\alpha_{\ell_j-1}}, 0), (*, 1)\rangle$$

are in $\mathcal{D}_{\leq j}$. Since all α's can be either 0 or 1, there are 2^{ℓ_j-1} such concepts. Moreover, all these concepts, regarded as subsets of $X = \{0,1\}^n$, are mutually disjoint.

A teaching set for $\emptyset \in \mathcal{D}_{\geq j}$ contains only negative examples and must rule out all these 2^{ℓ_j-1} concepts. But as these concepts are mutually disjoint, every example can rule out at most one of them. Therefore a teaching set for $\emptyset$ contains at least 2^{ℓ_j-1} examples. This number is then a lower bound for $TD(\mathcal{D}_{\geq j})$. □

Now we have all the bounds for $TD(\mathcal{D}_{\geq j})$ needed in order to bound $OTD(\mathcal{D}_n)$.

Theorem 4.22 $\Omega(\sqrt{n} \cdot 2^{n/2}) \leq OTD(\mathcal{D}_n) \leq O(n \cdot \sqrt{\log n} \cdot 2^{n/2})$.

Proof. For easier calculation we consider teaching dimensions of the form $j = (\lambda+1) \cdot 2^{n-\lambda}$ for $\lambda = 0, \ldots, n$. Thus we have $\ell_j = \lambda$.

Lower bound: From Corollary 4.17 we know that a j with $j < TD(\mathcal{D}_{\geq j})$ is a lower bound for $OTD(\mathcal{D}_n)$. From Lemma 4.21 we know that $TD(\mathcal{D}_{\geq j}) \geq 2^{\ell_j - 1}$. Hence we are seeking a j with $j < 2^{\ell_j - 1}$.

Claim 1: $j < 2^{\ell_j - 1}$ holds for $\lambda > \frac{1}{2}(n + 1 + \log(n+1))$.

Proof. Let $\lambda > \frac{1}{2}(n + 1 + \log(n+1))$. Using $\log(n+1) \geq \log(\lambda + 1)$ it follows that $\lambda > \frac{1}{2}(n + 1 + \log(\lambda + 1))$ and $\log(\lambda + 1) + n - \lambda < \lambda - 1$. This is equivalent to

$$(\lambda + 1) \cdot 2^{n-\lambda} < 2^{\lambda - 1}.$$

Therefore $j = (\lambda + 1) \cdot 2^{n-\lambda} < 2^{\lambda-1} = 2^{\ell_j - 1}$. □ Claim 1

By Claim 1 we can conclude that $j < 2^{\ell_j - 1}$ holds for $\lambda = \frac{1}{2}(n + 2 + \log(n+1))$. Calculating j using this λ yields a lower bound for $OTD(\mathcal{D}_n)$:

$$\begin{aligned} j = (\lambda + 1) \cdot 2^{n-\lambda} &\geq \tfrac{1}{2}(n + 4 + \log(n+1)) \cdot 2^{n/2 - (\log(n+1) - 2)/2} \\ &\geq \frac{\frac{1}{2}(n+4) \cdot 2^{n/2}}{\sqrt{\frac{1}{4}(n+1)}} \geq \Omega(\sqrt{n} \cdot 2^{n/2}). \end{aligned}$$

Upper bound: We are seeking a j with $j \geq TD(\mathcal{D}_{\geq j})$ to apply Corollary 4.17.

Claim 2: $j \geq TD(\mathcal{D}_{\geq j})$ holds for $\lambda \leq \frac{1}{2}(n - \log\log n)$.

Proof. $\lambda \leq \frac{1}{2}(n - \log\log n)$ is equivalent to $2^{n-2\lambda} \geq \log n$ which in turn is equivalent to $(\lambda+1) \cdot 2^{n-\lambda} \geq (\lambda+1) \cdot 2^{\lambda} \cdot \log n$ where the right hand side is greater than $TD(\mathcal{D}_{\geq j})$ by Lemma 4.19 and the left hand side equals j, hence $j \geq TD(\mathcal{D}_{\geq j})$. □ Claim 2

Calculating j for $\lambda = \frac{1}{2}(n - \log\log n)$ yields

$$\begin{aligned} j = (\tfrac{1}{2}(n - \log\log n) + 1) \cdot 2^{n - (n - \log\log n)/2} &= (\tfrac{1}{2}(n - \log\log n) + 1) \cdot 2^{(n + \log\log n)/2} \\ &\leq O(n \cdot \sqrt{\log n} \cdot 2^{n/2}) \end{aligned}$$

as an upper bound for $OTD(\mathcal{D}_n)$. □

The teachability of 1-decision lists is judged quite differently by our various measures. While the average teaching dimension and the CT-dimension are linear, the OT-dimension is exponential in n. Whether linear or exponential is more "realistic" is not clear. On the one hand, $OTD(\mathcal{D}_n)$ should be greater than linear because the class of 1-decision lists is bigger and more complex than the class of monomials and should therefore be harder to teach. On the other hand, exponential OT-dimensions should be reserved for the most complex classes. For example, the class of all Boolean functions over n variables has an OT-dimension of 2^n.

More results seem necessary for deciding whether the OT-dimension matches our intuition about teachability. In particular, knowing OT-dimensions for other classes, as well as for individual concepts, would be helpful.

4.3.2 Iterated Optimal Teacher Teaching Dimensions

The optimal teacher teaching dimension is defined in terms of TD (see Definition 4.14). Using TD to define $\mathfrak{H}_{opt}$ reflects the assumption of the learners that the teacher will not give more examples than necessary, that is, no more than $TD(c^*)$. But for a teacher knowing this assumption of the learner, less than $TD(c^*)$ examples would suffice for teaching c^*; more precisely, $OTD(c^*)$ examples are enough. If the learners in turn are aware of this fact they might then assume that the teacher does not give more than $OTD(c^*)$ examples (instead of no more than $TD(c^*)$ as before).

The number of examples needed to teach those learners is measured by another dimension, which we get after substituting TD by OTD in Definition 4.14. This substitution can occur iteratively, leading to an infinite number of *iterated optimal teacher teaching dimensions*:

Definition 4.23 *Let $\mathcal{C}$ be a concept class and let $c \in \mathcal{C}$. We define*

$$OTD^0(c, \mathcal{C}) = TD(c, \mathcal{C})$$

and for $j \geq 1$:

$$\mathfrak{H}^j_{opt}(S) = \{h \in \mathcal{C}(S) \mid OTD^{j-1}(h, \mathcal{C}) \geq |S|\},$$

$$OTD^j(c, \mathcal{C}) = \min\{|S| \mid \mathfrak{H}^j_{opt}(S) = \{c\}\}.$$

*A sample S with $\mathfrak{H}^j_{opt}(S) = \{c\}$ is called an OT^j-*set for c. *As usual we may write $OTD^j(c)$ for $OTD(c, \mathcal{C})$, and we define $OTD^j(\mathcal{C}) = \max\{OTD^j(c, \mathcal{C}) \mid c \in \mathcal{C}\}$.*

We already know the first two iterated OT-dimensions: OTD^0 is the teaching dimension, and OTD^1 is the plain OT-dimension.

Intuitively speaking, OTD^2 measures the number of examples needed for learners that know that the teacher knows that they assume the teacher to be optimal. OTD^3 measures the number of examples needed for learners that know that the teacher knows that they know that the teacher knows that they assume the teacher to be optimal, and so on.

Calculating $OTD^j(c)$ for a fixed c and growing j results in a monotonically decreasing sequence of values.

Lemma 4.24 *Let $\mathcal{C}$ be a concept class. Then for all $c \in \mathcal{C}$ and all $j \in \mathbb{N}$, $OTD^j(c) \geq OTD^{j+1}(c)$.*

Proof. We use induction on j. It follows immediately from the definitions that for all $c \in \mathcal{C}$, $OTD^0(c) \geq OTD^1(c)$. Assume the statement for some j. From the induction hypothesis $\forall c$: $OTD^{j-1}(c) \geq OTD^j(c)$ we conclude

$$\mathfrak{H}^{j}_{opt}(S) = \{c \mid c \in \mathfrak{H}_{opt}(S) \wedge OTD^{j-1}(c) \geq |S|\}$$
$$\supseteq \{c \mid c \in \mathfrak{H}_{opt}(S) \wedge OTD^{j}(c) \geq |S|\} = \mathfrak{H}^{j+1}_{opt}(S) .$$

Therefore every OT^j-set for a given concept c is also an OT^{j+1}-set for this concept, hence $OTD^{j+1}(c) \leq OTD^j(c)$. □

As a simple example, consider the class $\mathcal{S}_n$. Here, for all $j \geq 1$, $OTD^j([1,n]) = 2$ and $OTD^j(c) = 1$ for $c \neq [1,n]$. For this class the iterated OT-dimension is of little interest. A similar result holds for the monomials.

Fact 4.25 *For all $j \geq 1$ and all $c \in \mathcal{M}^1_n$, $OTD^j(c) = OTD^1(c)$.*

Proof. It suffices to show $OTD^2(c) = OTD^1(c)$ for all $c \in \mathcal{M}^1_n$. Since $TD(c) = OTD(c)$ for all $c \neq \emptyset$, we can use almost the same reasoning as in the proof of Theorem 4.18, with TD substituted by OTD. Only Case 1 uses the concept $\emptyset$, whose TD and OTD differ. But a set S of size $k+1$ with no positive example is not an OT^2-set for a monomial with k variables because it is consistent with $\emptyset$ and smaller than $OTD^1(\emptyset) = n+2$. □

Since the series $(OTD^j(c))_{j=0}^{\infty}$ is monotonicalls decreasing and lower bounded by 0, it converges. And if there are only finitely many concepts in a class, there is a j such that for all concepts c the limit $\lim_{i\to\infty} OTD^i(c) = OTD^j(c)$. Intuitively speaking, at this point the teacher and the learners cannot benefit any more from knowing the other's behavior. How many iteration it takes until this *fixed point* is reached is a natural question. We have already seen that $\mathcal{S}_n$ and $\mathcal{M}^1_n$ reach this fixed point after one iteration only. In contrast to these two classes, we next show that the fixed point can occur arbitrarily late.

Theorem 4.26 *For all $k \geq 1$ there is a class $\mathcal{C}$ over a learning domain X such that $\min\{j \mid \forall c \in \mathcal{C} : OTD^j(c) = OTD^{j+1}(c)\} \geq k$.*

Proof. The idea of the proof is to construct a class $\mathcal{C}$ containing (among others) $k+1$ special concepts $c_0, \ldots, c_k$. In the first iteration, that is, from OTD^0 to OTD^1 only the dimension of c_0 is reduced. This reduction makes c_1 easier to teach because c_0 can be ruled out by less examples. Consequently in the second iteration, the dimension of c_1 is reduced (and none else). This causes the dimension of c_2 to be reduced in the third iteration, and so on until finally in the transition from OTD^k to OTD^{k+1} the dimension of c_k changes.

See Figure 4.4 for the class $\mathcal{C}$ for $k = 2$. For arbitrary k the class $\mathcal{C}$ is constructed as follows. The learning domain X is the union of $k+1$ disjoint sets $X_0, \ldots, X_k$ with $|X_0| = 3$ and $|X_j| = j+2$ for all $j \geq 1$. The concept class $\mathcal{C}$ is a union of $k+1$ disjoint classes $\mathcal{C}_0, \ldots, \mathcal{C}_k$ over X. $\mathcal{C}_0$ contains the empty concept $c_0 = \emptyset$ and the singleton concepts for all instances in X_0. Let c_1 be an arbitrary such singleton concept. All $c \in \mathcal{C}_1 \cup \cdots \cup \mathcal{C}_k$ coincide with c_1 on X_0, that is, $\forall c \in \mathcal{C}_1 \cup \cdots \cup \mathcal{C}_k : c \cap X_0 = c_1 \cap X_0$.

	X_0			X_1			X_2				OTD^0	OTD^1	OTD^2	OTD^3	
c_0	0	0	0	0	0	0	0	0	0	0	**3**	**2**	**2**	**2**	
c_1	1	0	0	0	0	0	0	0	0	0	**4**	**4**	**3**	**3**	
	0	1	0	0	0	0	0	0	0	0	1	1	1	1	$\mathcal{C}_0$
	0	0	1	0	0	0	0	0	0	0	1	1	1	1	
c_2	1	0	0	1	0	0	0	0	0	0	**7**	**5**	**5**	**4**	
	1	0	0	0	1	0	0	0	0	0	3	3	3	3	
	1	0	0	0	0	1	0	0	0	0	3	3	3	3	
	1	0	0	1	0	1	0	0	0	0	3	3	3	3	$\mathcal{C}_1$
	1	0	0	0	1	1	0	0	0	0	3	3	3	3	
	1	0	0	1	1	0	0	0	0	0	3	3	3	3	
	1	0	0	1	1	1	0	0	0	0	3	3	3	3	
c_3	1	0	0	1	0	0	1	0	0	0	9	6	6	6	
	1	0	0	1	0	0	0	1	0	0	4	4	4	4	
	1	0	0	1	0	0	0	0	1	0	4	4	4	4	
	1	0	0	1	0	0	0	0	0	1	4	4	4	4	
	1	0	0	1	0	0	1	0	0	1	4	4	4	4	
	1	0	0	1	0	0	0	1	1	0	4	4	4	4	
	1	0	0	1	0	0	0	1	0	1	4	4	4	4	
	1	0	0	1	0	0	1	0	1	0	4	4	4	4	$\mathcal{C}_2$
	1	0	0	1	0	0	1	1	0	0	4	4	4	4	
	1	0	0	1	0	0	0	0	1	1	4	4	4	4	
	1	0	0	1	0	0	1	1	1	0	4	4	4	4	
	1	0	0	1	0	0	1	1	0	1	4	4	4	4	
	1	0	0	1	0	0	1	0	1	1	4	4	4	4	
	1	0	0	1	0	0	0	1	1	1	4	4	4	4	
	1	0	0	1	0	0	1	1	1	1	4	4	4	4	

Figure 4.4 *The concept class $\mathcal{C}_0 \cup \mathcal{C}_1 \cup \mathcal{C}_2$ over $X_0 \cup X_1 \cup X_2$ is constructed in the proof of Theorem 4.26 for $k = 2$. The dimensions of concept c_2 stabilize only at OTD^3.*

As a shortcut we define $q_j = |X_j|$ for all $j \leq k$.

The concepts in $\mathcal{C}_1 \cup \cdots \cup \mathcal{C}_k$ on $X_1 \cup \cdots \cup X_k$ as well as the special concepts $c_2, \ldots, c_{k+1}$ are defined inductively as follows. For $j = 1, \ldots, k$ the concepts in $\mathcal{C}_j$ are identical to c_j on $X_0, \ldots, X_{j-1}$ and for all $2^{q_j} - 1$ non-empty subsets $Z \subseteq X_j$ there is a concept in $\mathcal{C}_j$ containing exactly the elements in Z. Then there are $|X_j|$ neighbors of c_j in $\mathcal{C}_j$ and we call an arbitrary one of them c_{j+1}.

To prove the theorem we make use of the following claims:

Claim 1: $\forall j \ \forall \ell \geq 1 \ \forall c \in \mathcal{C}_\ell \cup \{c_\ell\} : OTD^j(c) \geq q_\ell$.

Proof. Let $\ell \geq 1$. The proof is by induction on j. Let $j = 0$ and $c \in \mathcal{C}_\ell \cup \{c_\ell\}$. Then c has q_ℓ neighbor concepts in $\mathcal{C}_\ell \cup \{c_\ell\}$ (all with respect to X_ℓ). Since each neighbor concept must be eliminated by a separate example, we have $OTD^0(c) \geq q_\ell$.

Now assume that the claim holds for some j. Again c has q_ℓ neighbor concepts in $\mathcal{C}_\ell \cup \{c_\ell\}$, all having an OT^j-dimension of at least q_ℓ. An example set of size less than q_ℓ does not eliminate any such neighbor concept by size and is also unable to eliminate all neighbor concepts by inconsistency. Hence $OTD^{j+1}(c) \geq q_\ell$. □ Claim 1

Claim 2: $\forall j\ \forall \ell\ \forall c \in \mathcal{C}_\ell \setminus \{c_\ell, c_{\ell+1}\}\colon OTD^j(c) = OTD^0(c)$.

Proof. The claim trivially holds for $\ell = 0$, since $OTD^0(c) = 1$ for all $c \in \mathcal{C}_0 \setminus \{c_0, c_1\}$. Let $\ell \geq 1$ and let $c \in \mathcal{C}_\ell$ with $c \neq c_{\ell+1}$. By Lemma 4.24 we know that $OTD^j(c) \leq OTD^0(c) = q_\ell$. On the other hand, Claim 1 yields $OTD^j(c) \geq q_\ell$, and therefore we have $OTD^j(c) = q_\ell$. □ Claim 2

Claim 3: $\forall j\ \forall \ell \geq 1$: If $\forall c \in \mathcal{C}_0 \cup \cdots \cup \mathcal{C}_{\ell-1} \setminus \{c_\ell\} : OTD^j(c) < q_\ell$ then $OTD^{j+1}(c_\ell) = q_\ell$.

Proof. Let $\ell \geq 1$ and j such that $\forall c \in \mathcal{C}_0 \cup \cdots \cup \mathcal{C}_{\ell-1} \setminus \{c_\ell\} : OTD^j(c) < q_\ell$. Now every example set of size q_ℓ excludes all concepts in $\mathcal{C}_{\ell-1} \setminus \{c_\ell\}$ by size. Moreover, all $c \notin \mathcal{C}_{\ell-1}$ are inconsistent with c_ℓ on X_ℓ. Thus, $\{(x, c_\ell(x)) \mid x \in X_\ell\}$ is an OT^{j+1}-set for c_ℓ of size $|X_\ell| = q_\ell$. □ Claim 3

Claim 4: $\forall j\ \forall \ell \in \{1, \ldots, k\}$: If $OTD^j(c_{\ell-1}) \geq q_\ell$ then $OTD^{j+1}(c_\ell) > q_\ell$.

Proof. Let $\ell \geq 1$ and j with $OTD^j(c_{\ell-1}) \geq q_\ell$. First we have that $c_{\ell-1}$ is a neighbor concept of c_ℓ and that $OTD^j(c_{\ell-1}) \geq q_\ell$ according to the assumption. Second, c_ℓ has q_ℓ neighbor concepts c' in $\mathcal{C}_\ell$, all of them having $OTD^j(c') \geq q_\ell$ (by Claim 1).

Then c_ℓ has $q_\ell + 1$ neighbor concepts c' with $OTD^j(c') \geq q_\ell$. Therefore it follows that $OTD^{j+1}(c_\ell) > q_\ell$. □ Claim 4

The next claim shows that with each iteration of the OT-dimension the dimension of one of the special concepts $c_1, \ldots, c_k$ decreases.

Claim 5: For $j = 1, \ldots, k$: $OTD^j(c_j) > OTD^{j+1}(c_j) = q_j$.

Proof. By counting the neighbor concepts one can see that $OTD^0(c_0) = 3$ and $OTD^1(c_0) = 2$. It is also easy to check that $OTD^0(c_1) = OTD^1(c_1) = 4$ and $OTD^2(c_1) = 3 = q_1$. This proves the claim for $j = 1$.

We proceed by induction on j. Let the claim be true for $1, \ldots, j < k$. We show that it holds for $j + 1$. For the OT^j-dimensions of the concepts we have: $OTD^j(c) \geq q_j$ for $c \in \mathcal{C}_j \cup \cdots \cup \mathcal{C}_k$ (by Claim 1); $OTD^j(c_0), \ldots, OTD^j(c_{j-1}) \leq q_j - 1$ (by the induction hypothesis); for the other $c \in \mathcal{C}_0 \cup \cdots \cup \mathcal{C}_{j-1}$: $OTD^j(c) \leq q_{j-1}$ (by Claim 2).

Applying Claim 3 with $\ell = j + 1$ yields $OTD^{j+1}(c_{j+1}) = q_{j+1}$. Applying Claim 4 with $\ell = j + 1$ yields $OTD^{j+2}(c_{j+1}) > q_{j+1}$. This proves the claim for $j + 1$. □ Claim 5

From Claim 5 it follows $OTD^k(c_k) > OTD^{k+1}(c_k)$, which proves the theorem. □

If we continue the construction described in the previous proof, we get an infinite concept class such that the iterated OT-dimensions have no fixed point.

Corollary 4.27 *There is an infinite concept class $\mathcal{C}$ such that for all $j \in \mathbb{N}$ there is a $c \in \mathcal{C}$ with $OTD^j(c, \mathcal{C}) < OTD^{j+1}(c, \mathcal{C})$.*

When it exists, the fixed point of the iterated OT-dimensions appears to be a natural teachability measure whose values are typically less than those of the original OT-dimension. It would be interesting to know whether the fixed point values of the 1-decision lists are subexponential. However, calculating even $OTD^2(\mathcal{D}_n)$ looks daunting because to do so we need to know $OTD(c, \mathcal{D}_n)$ for all $c \in \mathcal{D}_n$, and it was already difficult to derive only a bound for one OTD-value in Section 4.3.1.

The next section is less technical. It describes a teaching model in which we, for the first time, observe feedback effects.

4.4 Learners with Selective Memory

So far, the teacher always had full knowledge about the learner's memory content, even if the learner gave no feedback. The memory of a learner with memory size m simply contained the last m examples (or all examples if $m = \infty$) in the order they were given. This enabled a feedbackless teacher to follow the simple strategy of teaching an $\mathfrak{H}$-set of the target. Such a strategy is optimal for teachers without feedback and cannot be improved when the teacher receives feedback (see Lemma 4.4).

In this section we let the memory behavior depend on the current hypothesis. As the latter is typically not known to a feedbackless teacher, this measure should make feedback more valuable. More precisely, now the learners only memorize an example if it is inconsistent with the current hypothesis at the round it is taught; consistent examples are ignored. We call such a memory *selective*.

Selective memory bears some resemblance with the human phenomenon that surprising or previously unknown observations better stick in memory than those that are already well-known. Granted, this is a very coarse resemblance, but within our general framework this is probably the closest we get to a formalization of this phenomenon.

Definition 4.28 *Let $\langle \varrho, \mathcal{C} \rangle$ be a represented concept class and $m \in \mathbb{N}^+ \cup \{\infty\}$. A non-deterministic learner $\mathcal{L} = (\mathcal{L}_{mem}, \mathcal{L}_{hyp})$ is said to have* selective memory of size m *iff*

(MEM) $\mathcal{L}_{mem}(\langle\,\rangle, init, z) = \{\langle z \rangle\}$,

$$\mathcal{L}_{mem}(\boldsymbol{s}, \sigma, z) = \begin{cases} \{\boldsymbol{s} \circ_m \langle z \rangle\} & \text{if } z \notin \mathcal{X}(\varrho_\sigma), \\ \{\boldsymbol{s}\} & \text{otherwise.} \end{cases}$$

The previous definition can be combined with the $\mathfrak{H}$-restriction (see Definition 4.1). This yields $\mathfrak{H}$-restricted learners with selective memory of size m. For the rest of the section we study only these learners.

When receiving feedback, the teacher can always give an example that is inconsistent with the current hypothesis and that will thus be added to the memory. Moreover, no matter whether or not feedback is present, the learner must be able to memorize an $\mathfrak{H}$-set. Therefore, teaching the selective memory learner with feedback is just as easy as teaching the non-selective memory learner (compare with Lemmas 3.3 and 4.4).

In the situation without feedback, the learner still must be able to memorize an $\mathfrak{H}$-set for c^*. But nevertheless, teaching c^* may take longer than $\mathfrak{H}D(c^*)$ rounds. The following lemma is similar to the Lemmas 3.3 and 4.4 for non-selective memory, except that it does not state the optimal teaching time for teaching without feedback.

Lemma 4.29 *Let $\langle \varrho, \mathcal{C} \rangle$ be a represented concept class and $c^* \in \mathcal{C}$ be a target. Let $\mathfrak{H}\colon 2^{\mathcal{X}} \to 2^{\mathcal{C}}$ and $m \in \mathbb{N}^+ \cup \{\infty\}$.*

For teaching with or without feedback, the target c^* is teachable to the $\mathfrak{H}$-restricted learner with selective m-memory if and only if $\mathfrak{H}D(c^*, \mathcal{C})$ is finite and at most m.

For teaching with feedback, the optimal teaching time is $\mathfrak{H}D(c^*, \mathcal{C})$.

Proof. Suppose $\mathfrak{H}D(c^*, \mathcal{C})$ is finite and at most m. Then there is a sample $S = \{z_1, \ldots, z_k\}$ with $k = \mathfrak{H}D(c^*, \mathcal{C}) \leq m$ and $\mathfrak{H}(S) = \{c^*\}$. First, consider teaching with feedback. We define a teacher T_1 by

$$T_1(\boldsymbol{s}, \sigma) = \begin{cases} z_1 & \text{if } \sigma = \mathit{init}, \\ z_j \text{ with } j = \min\{i \mid z_i \notin \mathcal{X}(\varrho_\sigma)\} & \text{if } \exists i : \ z_i \notin \mathcal{X}(\varrho_\sigma), \\ z_1 & \text{otherwise.} \end{cases}$$

In other words, T_1 teaches an inconsistent example from S as long as the learner has not reached the target. In this way, one example from S is given in each round and in particular memorized by the learner. Therefore after k rounds the learner knows an $\mathfrak{H}$-set for c^* and has to hypothesize c^* due to the $\mathfrak{H}$-restriction.

Now consider teaching without feedback. Let T_2 be the teacher that first teaches z_1 and then $z_2, \ldots, z_k$ in an endless loop: $T_2(0) = z_1$, $T_2(i) = z_{1+i \bmod (k-1)}$ for all $i \geq 1$. The first example, z_1, automatically memorized, and during each iteration through $z_2, \ldots, z_k$ the learner, unless it already hypothesizes c^*, encounters an example that is inconsistent with the current hypothesis. This example is then memorized according to the definition of selective memory. Therefore after at most $k-1$ iterations the learner knows the $\mathfrak{H}$-set $\{z_1, \ldots, z_k\}$. Then, since the learner is $\mathfrak{H}$-restricted, it hypothesizes c^*. This shows that c^* can be taught finitely.

Now suppose c^* is teachable. Then by the same arguments as in the non-selective memory situation, $\mathcal{L}$ eventually memorizes an $\mathfrak{H}$-set for c^*. Therefore the memory of $\mathcal{L}$ must at least have size $\mathfrak{H}D(c^*, \mathcal{C})$.

Moreover $\mathfrak{H}D(c^*, \mathcal{C})$ is finite. Memorizing an $\mathfrak{H}$-set can only occur after at least $\mathfrak{H}D(c^*, \mathcal{C})$ rounds. This number of rounds is achieved by the teacher T_1 with feedback defined above. Therefore the optimal teaching time for teaching with feedback is $\mathfrak{H}D(c^*, \mathcal{C})$. □

In the case without feedback, it can happen that the teacher gives an example that the learner cannot memorize. Teaching without feedback is thus harder than teaching with feedback. But, just as in the non-selective memory models, there is no difference in teaching time between teaching finitely and in the limit. To see this, consider teacher T_2 in the proof of Lemma 4.29. This teacher works for all targets that are teachable without feedback.

The previous lemma only characterizes teachability without feedback, it does not give the optimal teaching time. We denote the optimal teaching times for the $\mathfrak{H}$-restricted ∞-memory learner by *Sel*-$\mathfrak{H}D(c^*, \mathcal{C})$. So far, Lemma 4.29 provides us only with a trivial lower bound of $\mathfrak{H}D(c^*)$ and with an quadratic upper bound for *Sel*-$\mathfrak{H}D(c^*, \mathcal{C})$:

Corollary 4.30 *Let $\langle \varrho, \mathcal{C} \rangle$ be a represented concept class, $\mathfrak{H}\colon 2^X \to 2^{\mathcal{C}}$ a hypothesis restriction, and $c^* \in \mathcal{C}$ a target concept with $\mathfrak{H}D(c^*, \mathcal{C}) < \infty$. Then Sel-$\mathfrak{H}D(c^*, \mathcal{C}) \leq (\mathfrak{H}D(c^*, \mathcal{C}) - 1)^2 + 1$.*

Proof. The teacher that works like T_2 in the proof of Lemma 4.29 has the claimed teaching time. □

Next we give a lower bound for *Sel*-$\mathfrak{H}D$, which matches the upper bound up to a factor of at most two. This shows that lack of feedback *always* handicaps teaching learners with selective memory.

Theorem 4.31 *Let $\langle \varrho, \mathcal{C} \rangle$ be a represented concept class, $\mathfrak{H}\colon 2^X \to 2^{\mathcal{C}}$ a hypothesis restriction, and $c^* \in \mathcal{C}$ a target concept with $\mathfrak{H}D(c^*\mathcal{C}) < \infty$. Then Sel-$\mathfrak{H}D(c^*, \mathcal{C}) \geq 1 + \frac{1}{2} \cdot \mathfrak{H}D(c^*, \mathcal{C}) \cdot (\mathfrak{H}D(c^*, \mathcal{C}) - 1)$.*

Proof. In this proof we abbreviate $\mathfrak{H}D(c^*, \mathcal{C})$ with k. Let $T\colon \mathbb{N} \to \mathcal{X}(c^*)$ be a feedbackless teacher for c^* and let $\mathcal{L}$ be the non-deterministic learner with infinite selective memory. We construct a computation of $\mathcal{L}$ that assumes the hypothesis c^* only after at least $\frac{1}{2} \cdot k \cdot (k-1) + 1$ rounds. The idea behind that computation is to always assume a hypothesis that maximizes the number of rounds until T gives the next inconsistent example. A computation of $\mathcal{L}$ is represented by a series $((\boldsymbol{s}_i, \sigma_i))_{i \in \mathbb{N}}$ of states such that $(\boldsymbol{s}_0, \sigma_0) = (\langle\,\rangle, \mathit{init})$ and

$$(\boldsymbol{s}_{i+1}, \sigma_{i+1}) \in \mathcal{L}(\boldsymbol{s}_i, \sigma_i, T(i))$$

for all $i \in \mathbb{N}$. The behavior of the memory is already specified by the definition of selective memory. We define the hypotheses by $\sigma_{i+1} = \sigma_i$ if $T(i) \in \mathcal{X}(\varrho_{\sigma_i})$; if

$T(i) \notin \mathcal{X}(\varrho_{\sigma_i})$ we define

$$\sigma_{i+1} \in \operatorname*{argmax}_{h \in \mathfrak{H}(\boldsymbol{s}_i \circ \langle T(i) \rangle)} \min\{t \mid t > i, \text{ and } \boldsymbol{s} \circ \langle T(i+1), \ldots, T(t-1) \rangle \text{ is consistent with } h, \text{ and } T(t) \text{ is not consistent with } h\}.$$

Since in the latter case $\boldsymbol{s}_i \circ \langle T(i) \rangle = \boldsymbol{s}_{i+1}$ and σ_{i+1} is chosen from the set $\mathfrak{H}(\boldsymbol{s}_i \circ \langle T(i) \rangle)$, the hypothesis σ_{i+1} satisfies the conditions of the $\mathfrak{H}$-restricted learner $\mathcal{L}$.

By $t_0, t_1, \ldots$ we denote the rounds in which the memory grows, that is, $j+1 = |\boldsymbol{s}_{t_j}| > |\boldsymbol{s}_{t_j-1}| = j$. We have $t_0 = 1$ since the first example is automatically memorized. The values t_j also describe exactly the rounds before which an inconsistent example has arrived, that is, $T(t_j - 1) \notin \mathcal{X}(\varrho_{\sigma_{t_j-1}})$, and thus they describe the rounds in which the hypothesis might change. By definition, σ_i can be a target hypothesis only when $\boldsymbol{s}_i$ contains a teaching set. Therefore the series t_j is defined at least until $j = k-1$.

Claim: $t_{j+1} - t_j \geq k - (j+1)$ for all $j = 0, \ldots, k-2$.

Proof. Let $0 \leq j \leq k-2$. Then $|\boldsymbol{s}_{t_j}| = j+1 < k$, and the hypothesis σ_{t_j} is chosen to maximize the number of rounds during which examples consistent with σ_{t_j} arrive. There is a hypothesis such that this number is at least $k-(j+1)$. Otherwise all hypotheses in $\mathfrak{H}(\boldsymbol{s}_{t_j} \circ \langle T(t_j+1) \rangle)$ were inconsistent with one of the following $k-(j+1)-1$ examples and we could build an $\mathfrak{H}$-set for c^* using these $k-(j+1)-1$ examples plus the examples in $\boldsymbol{s}_{t_j}$. This would yield a sample of size $k-(j+1)-1+j+1 = k-1$, a contradiction to $k = \mathfrak{H}D(c^*)$. It follows that the next inconsistent example can arrive only after at least $k-(j+1)$ rounds. □ Claim

Using the claim we conclude that the number of rounds until the hypothesis of our computation reaches the target is at least

$$t_{k-1} \geq 1 + \sum_{j=0}^{k-2} k - (j+1) = 1 + \sum_{j=1}^{k-1} j = 1 + \frac{1}{2} \cdot k \cdot (k-1).$$

□

Selective memory is the first model in which the absence of feedback actually increases the optimal teaching time. But this feedback effect is roughly the same for all concepts and classes. Consequently, the model does not help to distinguish classes for which feedback is more useful from those for which it is less useful.

4.5 Discussion

Our goal in this section was to improve the TD model with respect to plausibility. However, we have not provided a clear definition of what plausibility actually means.

Is the empty concept in the class of monomials easy to teach because it is so simple? Is it hard to teach because there are so many similar looking concepts from which it must be distinguished?

The CT-dimension was designed with the implicit understanding that simpler concepts are easier to teach. It is not surprising that it resulted in a model that shows just that. If we had assumed, for instance, that larger concepts are easier to teach and had defined $\mathfrak{H}$ accordingly, we had gotten a model in which larger concepts were indeed easier to teach. Thus the $\mathfrak{H}$-approach allows us to encode our interpretation of plausibility into the teaching model. The use of this is, however, rather limited because then we have to know what targets are easy to teach and what are hard before we devise the model. But the hardness of teaching is just what we want to measure.

On the other hand, we are not forced to use hypothesis restrictions this way. The OT-dimension shows that we can use $\mathfrak{H}$ to encode plausible learning behaviors into the model, rather than plausible teachabilities. But the one problems persists, that we cannot judge whether the model yields realistic teaching times. After all there is no real-life data on teaching Boolean functions to humans.

As a consequence in the remainder of this thesis we shall focus less on plausible teaching times. Rather we concentrate on the other aspects described in Section 1.3: the influence of memory size, feedback, and the order of examples.

Chapter 5

Learners with Restricted Hypothesis Changes

In the $\mathfrak{H}$-restricted model the learner can choose its follow-up hypothesis independently of its current one. The current hypothesis therefore reveals no information about the next hypothesis. As a consequence, we have seen that feedback is of no use for the teacher in this model. In contrast, it is a common feature of real-world learners (both humans and machines) to change the hypothesis only a little in each step. But then a hypothesis does give some information about the next hypothesis, which implies that feedback is often helpful in the real world.

In this chapter we introduce a model that allows to formalize the idea of small hypothesis changes. But since there is no universal definition of a small hypothesis change, we define the model general enough to allow for any restriction of hypothesis changes. At the core of the model, a relation over all hypotheses, called *neighborhood relation*, specifies which hypothesis transitions are allowed.

Consider, for example, teaching the target $[1, n] \in \mathcal{S}_n$ in the teaching dimension model. This takes n rounds because the learner, even after receiving $n - 1$ positive examples, may still assume a co-singleton hypothesis. If we now modify the rules and forbid transitions between co-singleton concepts (see Figure 5.1), then teaching with feedback is much faster than without: The teacher first gives one positive example and then observes the resulting hypothesis. If it is a co-singleton concept, the teacher can identify the "missing" instance and give it as positive example. Now, according to our hypothesis change restriction the learner can switch to the target $[1, n]$, but not to any co-singleton concept. Therefore teaching takes only two rounds. On the other hand, if the teacher gets no feedback, still all n examples must be given.

In the new model there is one subtle point concerning the consistency of the hypothesis with the memorized examples. Whereas in the $\mathfrak{H}$-restricted model the learner always outputs a consistent hypothesis, in the new model all admissible follow-up hypotheses might be inconsistent. We thus have to relax the consistency requirement. Now we demand that the learner chooses only among those admissible hypotheses that have least error with respect to the memorized examples. In Section 5.1 we give a formal definition of the learner in the model.

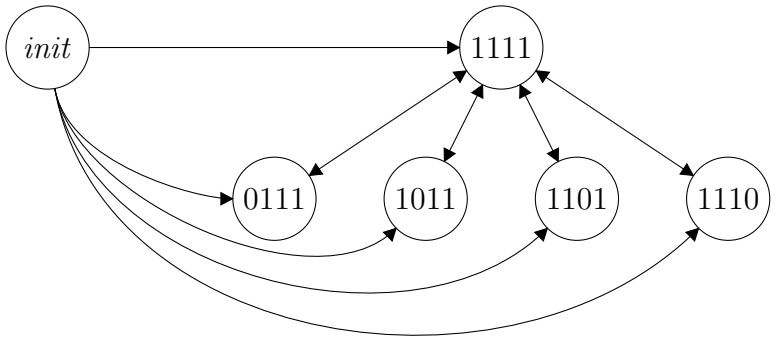

Figure 5.1 *A neighborhood relation for the concept class $\mathcal{S}_4$. The nodes represent the hypotheses, the arcs describe the allowed hypothesis changes. From the initial hypothesis all other hypotheses are accessible. Within the other hypotheses, transitions are allowed only between concepts that differ in exactly one instance.*

The neighborhood relation adds a lot of freedom, and different relations can cause different teachability results for the same concept class. We demonstrate this in Section 5.2 using the class of all finite languages. For a certain neighborhood relation, concepts can be taught much faster with feedback than without; for another neighborhood relation, concepts cannot be taught finitely unless feedback is available; and using yet another neighborhood relation, concepts cannot be taught in the limit unless feedback is available.

For the natural concept classes of monomials and 1-decision lists, together with natural neighborhood relations, feedback effects do not seem to occur. However, we show in Section 5.3 that the order of examples is crucial and that for monomials teaching becomes faster if the learners have a bigger memory.

Finding an optimal teacher or at least computing the optimal teaching time for a given concept class and neighborhood relation is a fundamental task. From a computational complexity perspective, the difficulty of solving this task varies. For learners with limited memory an optimal teacher with feedback can be found in polynomial time by a dynamic programming approach using Lemma 2.15. Finding the optimal teaching time for teaching without feedback is $\mathcal{NP}$-hard. The same is true for teaching without feedback to learners with infinite memory. Teaching learners with infinite memory with feedback is even $\mathcal{PSPACE}$-hard. The details are given in Section 5.4.

5.1 Description of the Model

A neighborhood relation over all hypotheses is formally defined as follows.

Definition 5.1 *Let $\langle \varrho, \mathcal{C} \rangle$ with $\varrho\colon \Sigma^* \times X \to \{0, 1, \uparrow\}$ be a represented concept class. A* neighborhood relation $\rhd$ *is a relation $\rhd \subseteq (\varrho(\mathcal{C}) \cup \{init\}) \times (\varrho(\mathcal{C}) \cup \{init\})$.*

A neighborhood relation specifies the allowed hypothesis changes: "$\sigma \rhd \zeta$" means that a learner currently hypothesizing σ may assume ζ next; "$\sigma \not\rhd \zeta$" means that ζ is no allowed follow-up hypothesis. The learner's behavior with respect to memory is not affected by the neighborhood relation. It can thus be combined with m-memory as well as with selective memory. We confine ourselves to the former.

To formally define our relaxed consistency requirement, we have to introduce some notation for the number of errors a sequence of examples has with respect to some hypothesis. We define for $\sigma \in \varrho(\mathcal{C})$ and $\boldsymbol{s} = \langle (x_1, b_1), \ldots, (x_\ell, b_\ell) \rangle \in \mathcal{X}^*$:

$$err(\sigma, \boldsymbol{s}) = \sum_{i=1}^{\ell} \varrho_\sigma(x_i) \oplus b_i$$

where $\oplus$ is the XOR-function. The definition of *err* is such that multiple occurrences of the same example in $\boldsymbol{s}$ count multiple times. We define the hypothesis *init* to have infinitely many errors: $err(init, \boldsymbol{s}) = \infty$. This ensures that *init* is left in the first round and never reached again.

Similar to the $\mathfrak{H}$-restricted model, the model here is really a family of models with the neighborhood relation as parameter.

Definition 5.2 *Let $\langle \varrho, \mathcal{C} \rangle$ with $\varrho\colon \Sigma^* \times X \to \{0, 1, \uparrow\}$ be a represented concept class and let $\rhd$ be a neighborhood relation. A non-deterministic learner $\mathcal{L} = (\mathcal{L}_{mem}, \mathcal{L}_{hyp})$ is called* $\rhd$-restricted using hypothesis space $\langle \varrho, \mathcal{C} \rangle$ *iff it satisfies:*

(HYP) $\forall \boldsymbol{s} \in \mathcal{X}^* \; \forall \sigma \in \varrho(\mathcal{C}) \cup \{init\} \; \forall z \in \mathcal{X} \; \forall (\boldsymbol{s}', \sigma') \in \mathcal{L}(\boldsymbol{s}, \sigma, z)$*:*

$$\sigma' \in \begin{cases} \{\sigma\} & \text{if } z \in \mathcal{X}(\varrho_\sigma) \text{ or } \min\limits_{\zeta\colon \sigma \rhd \zeta} err(\zeta, \boldsymbol{s}') \geq err(\sigma, \boldsymbol{s}'), \\ \operatorname{argmin}\limits_{\zeta\colon \sigma \rhd \zeta} err(\zeta, \boldsymbol{s}') & \text{otherwise.} \end{cases}$$

Let $m \in \mathbb{N}^+ \cup \{\infty\}$. The learner $\mathcal{L}$ is said to have memory size m *iff*

(MEM) $\forall \boldsymbol{s} \in \mathcal{X}^* \; \forall \sigma \in \Sigma^* \; \forall z \in \mathcal{X} : \mathcal{L}_{mem}(\boldsymbol{s}, \sigma, z) = \{\boldsymbol{s} \circ_m \langle z \rangle\}$.

Note that the $\rhd$-restricted learner is conservative by definition.

5.2 Teaching the Class of All Finite Languages

In this section we apply the $\rhd$-restricted learner model to the class $\mathcal{C}_{fin}$ of all finite languages over the alphabet $\{\mathtt{a}, \mathtt{b}\}$. Using different $\rhd$-restrictions we demonstrate various effects.

As representation alphabet we use $\Sigma = \{\mathtt{a}, \mathtt{b}, \mathtt{\#}\}$. A string $\sigma \in \{\mathtt{a}, \mathtt{b}, \mathtt{\#}\}^*$ represents the set of all words separated by `#`. For example, $\sigma =$ `ab##baab#bb#ab` represents the concept $\{\mathtt{ab}, \Lambda, \mathtt{baab}, \mathtt{bb}\}$. Therefore the representation size of a concept

$c = \{w_1, \ldots, w_k\} \subseteq \{\mathtt{a}, \mathtt{b}\}^*$ is $\|c\| = k - 1 + \sum_{i=1}^{k} |w_i|$. We denote the representation function just defined by ϱ_{fin}.

Small changes among finite sets consist in the addition or removal of a word. Formally we define the neighborhood relation $\rhd_{fin}$ by

$$\sigma \rhd_{fin} \zeta \quad :\Leftrightarrow \quad |\varrho_\sigma \triangle \varrho_\zeta| \leq 1$$

for $\sigma \neq \mathit{init}$. For the initial hypothesis we allow only a transition to the empty language, which is represented by the empty string Λ, that is, $\mathit{init} \rhd_{fin} \zeta :\Leftrightarrow \zeta = \Lambda$.

The following results were originally obtained for a model in which all learners start with the hypothesis Λ. For that reason, teachers often start by giving an arbitrary negative example which leads all learners from *init* to Λ where the "real teaching" begins. In this section all learners have infinite memory.

Fact 5.3 *The represented concept class $\langle \varrho_{fin}, \mathcal{C}_{fin} \rangle$ is finitely teachable to the $\rhd_{fin}$-restricted learner.*

Proof. For a target language $\{w_1, \ldots, w_k\} \in \mathcal{C}_{fin}$ a teacher first gives an arbitrary negative example and then presents all positive examples $(w_1, 1), \ldots, (w_k, 1)$. In the first round, the learner switches to Λ and then in every round the learner may either add or remove a word from the hypothesis. But there is only one possible way to stay consistent with the examples, namely by adding them to the hypothesis. Therefore, after $k + 1$ rounds the $\rhd_{fin}$-restricted learner has arrived at a target hypothesis. □

For the $\rhd_{fin}$-restricted learner feedback is of no use. But a modification of $\rhd_{fin}$ makes feedback very valuable. We define

$$\begin{aligned} \sigma \rhd'_{fin} \zeta \quad :\Leftrightarrow \quad & \big(\varrho_\zeta = \varrho_\sigma \cup \{w_1, w_2\} \text{ or } \varrho_\zeta = \varrho_\sigma \setminus \{w_1\}\big) \\ & \text{and } \|\varrho_\zeta\| \leq 2\|\varrho_\sigma\|. \end{aligned}$$

In other words, the learner may either add two words or remove one word from the hypothesis. In both cases the size of the hypothesis may at most double in each round. In the special case $\sigma = \mathit{init}$ we allow every singleton concept as neighbor: $\mathit{init} \rhd'_{fin} \zeta$ for all ζ with $|\varrho_\zeta| = 1$. For the $\rhd'_{fin}$-restricted learner there is a big difference in teaching time between teaching with and without feedback.

Fact 5.4 *The represented concept class $\langle \varrho_{fin}, \mathcal{C}_{fin} \rangle$ is teachable to the $\rhd'_{fin}$-restricted learner with feedback such that for all $c \in \mathcal{C}$ the teaching time is in $O(|c|) \leq O(\|c\|)$.*

Proof. The $\rhd'_{fin}$-restricted learner may either add two words to the hypothesis or remove one. As a consequence, whenever the $\rhd'_{fin}$-restricted learner receives a positive example, it can add it to the hypothesis and "invent" another word and add it to the hypothesis as well. Due to the size restriction there are always only finitely many words that can be invented.

Let $c^* = \{w_1, \ldots, w_k\}$ be a target concept. A teacher with feedback first teaches a negative example and then all words w_i as positive examples. After w_k the hypothesis contains all words from c^* plus $\ell \leq k$ invented words $u_1, \ldots, u_\ell$. From the feedback, the teacher gets to know these words and teaches them as negative examples $(u_1, 0), \ldots, (u_\ell, 0)$. Since at most one word can be removed per round, the learner has to remove the negative example it is taught and thus arrives at the correct hypothesis after ℓ rounds. Overall, teaching takes at most $2k + 1 = 2|c^*| + 1$ rounds. □

The teaching time for a teacher without feedback is exponentially larger.

Fact 5.5 *The represented concept class $\langle \varrho_{fin}, \mathcal{C}_{fin} \rangle$ is finitely teachable to the $\rhd'_{fin}$-restricted learner. Every such teacher needs $\Omega(2^{\|c\|})$ rounds for some $c \in \mathcal{C}$ and there is no upper bound for the teaching time that depends only on $|c|$.*

Proof. A successful teacher can be defined as follows. Let $c^* \in \mathcal{C}_{fin}$. First of all, the teacher gives all words of length at most $2\|c^*\|$ that are not in c^* as negative examples. Afterwards, all words in c^*, starting with a longest one, are taught as positive examples. The hypothesis Λ is consistent with all negative examples, hence no hypothesis change happens after the transition from *init* to Λ in the first round. While the positive examples are taught, the learner cannot include any words outside of c^* into the hypothesis since all these words either have been ruled out by the negative examples or are too long to be included. Also, since a longest word is taught first, the hypothesis growth limitation cannot be violated by positive examples included later. Therefore, the $\rhd'_{fin}$-restricted learner must reach the target hypothesis after the positive examples are taught.

For the lower bound, let T be a teacher that teaches $\mathcal{C}_{fin}$ finitely to the $\rhd'_{fin}$-restricted learner $\mathcal{L}$. Let $c^* = \{\mathtt{a}^k, \mathtt{b}\}$ be the target concept of size $k + 2$ for an arbitrary $k > 3$. Let M be the teaching time of T for the target c^*.

Both words, $\mathtt{a}^k$ and $\mathtt{b}$, must occur as positive examples, otherwise the deterministic learner $L_0 \in \mathcal{L}$ that never "invents" a word could not be taught. Moreover, $\mathtt{a}^k$ must occur before $\mathtt{b}$, since otherwise L_0 would at some point have $\{\mathtt{b}\}$ as hypothesis. But because of the growth restriction, $\{\mathtt{b}\}$ cannot be changed to $\{\mathtt{a}^k, \mathtt{b}\}$ later, thus L_0 cannot learn c^*. Let $T(j_1) = (\mathtt{a}^k, 1)$ be the first occurrence of $\mathtt{a}^k$ and let $T(j_2) = (\mathtt{b}, 1)$ be the first occurrence of $\mathtt{b}$ in the example sequence.

It suffices to show that $Z(0), \ldots, T(M)$ contains all words of length at most $k - 3$. This implies $M \geq 2^{k-2} - 1 = \Omega(2^{\|c^*\|})$. Assume that there was a word $\hat{w} \notin c^*$ with $|\hat{w}| \leq k - 3$ that is not taught. We define a $\rhd'_{fin}$-restricted learner $L \in \mathcal{L}$ that does not arrive at c^* during teaching. On $T(j_1)$, L switches to hypothesis $\mathtt{a}^\mathtt{k}$ and does not change it until $T(j_2)$ arrives. Then L chooses the hypothesis $\mathtt{a}^k\mathtt{\#b\#}\hat{w}$, which is incorrect, but consistent with the examples so far. The length restriction is obeyed, since $\|\mathtt{a}^k\| = k \leq 2k \leq \|\mathtt{a}^k\mathtt{\#b\#}\hat{w}\|$. From then on, L will never change the hypothesis, since the only inconsistent example, $(\hat{w}, 0)$, is never taught, according to the assumption.

As k can be chosen arbitrarily large, there is no bound on the number of rounds needed that depends on the cardinality $|c^*|$ only. □

If we remove the size restriction from $\rhd'_{fin}$, we get $\rhd''_{fin}$. Using this neighborhood relation, we observe a difference between finite teaching and teaching in the limit. This is the first time we witness such an effect in any of our models.

Fact 5.6 *The represented concept class $\langle \varrho_{fin}, \mathcal{C}_{fin} \rangle$ is not finitely teachable to the $\rhd''_{fin}$-restricted learner, but it is teachable with feedback as well as in the limit.*

Proof. Suppose that there is teacher that finitely teaches $\{\mathtt{a}, \mathtt{b}\} \in \mathcal{C}_{fin}$ to the $\rhd''_{fin}$-restricted learner with a teaching time of M. If the learner, when the second positive example arrives, "invents" a word not occurring in the finitely many examples $T(0), \ldots, T(M)$ given by the teacher, then it does not arrive at a correct hypothesis, a contradiction.

Next, we describe a teacher that teaches $\mathcal{C}_{fin}$ with feedback. For a target $c^* \in \mathcal{C}_{fin}$, the teacher first gives an arbitrary negative example and then all positive examples. This may lead to at most $|c|$ superfluous words in the hypothesis of the $\rhd''_{fin}$-restricted learner. The teacher observes these words and gives them as negative examples, thus forcing the learner to remove the excessive words and to reach a target hypothesis.

A teacher for teaching $\mathcal{C}_{fin}$ in the limit without feedback first teaches an arbitrary negative example and then all positive examples. Again, the $\rhd''_{fin}$-restricted learner's hypothesis may contain finitely many excessive words. By teaching all infinitely many words not in the target concept as negative examples, the superfluous words can be removed in the limit. □

Finally we define $\rhd'''_{fin}$. It differs from $\rhd''_{fin}$ in that a word may only be removed from the hypothesis if neither its predecessor nor its successor (we order the words in $\{\mathtt{a}, \mathtt{b}\}^*$ first by length and then lexicographically) is contained in the hypothesis. For instance, $\mathtt{aa\#aab} \rhd'''_{fin} \mathtt{aa}$, but $\mathtt{aa\#ab} \not\rhd'''_{fin} \mathtt{aa}$.

Fact 5.7 *The represented concept class $\langle \varrho_{fin}, \mathcal{C}_{fin} \rangle$ is not teachable to the $\rhd'''_{fin}$-restricted learner in the limit, but it is teachable with feedback.*

Proof. Let $c^* = \{w_1, w_2, w_3\} \in \mathcal{C}_{fin}$ be a target concept. Suppose that there is a teacher $T\colon \mathbb{N} \to \mathcal{X}(c^*)$ that teaches c^* to the $\rhd'''_{fin}$-restricted learner in the limit. All three examples $(w_1, 1), (w_2, 1), (w_3, 1)$ must occur in the example sequence, otherwise the learner that does not "invent" words could not be taught. Let $T(j_i) = (w_i, 1)$ be the first occurrence of w_i for $i = 1, 2, 3$. Without loss of generality, we assume that $j_1 < j_2 < j_3$.

Now we construct a $\rhd'''_{fin}$-restricted deterministic learner L that fails under teacher T. After $T(j_1)$ the hypothesis of L is w_1 (this is consistent because prior to round j_1 only negative examples can occur). When taught $T(j_2)$, the learner L adds w_2 to the

hypothesis, as well as a word $u_1 \notin c^*$ such that (1) neither u_1 nor its successor u_2 occurs in $T(0), \ldots, T(j_3)$, and (2) $u_2 \notin c^*$. When taught $T(j_3)$, the learner L adds w_3 and u_2 to the hypothesis. Adding u_2 is possible, because it has not yet occurred as negative example. At this point the hypothesis of L contains the words u_1 and u_2, neither of which can be deleted any more due to the definition of $\rhd'''_{fin}$. Thus, L cannot end up with a correct hypothesis, a contradiction.

Teaching $\mathcal{C}_{fin}$ to the $\rhd'''_{fin}$-restricted learner with feedback can be done as follows. Let $c^* \in \mathcal{C}_{fin}$ be the target concept. The teacher first teaches all negative examples that are predecessors or successors of a word in c^*. Then all positive examples are taught and as soon as the teacher discovers that a learner has introduced a wrong word $u \notin c^*$ into the hypothesis, the negative example $(u, 0)$ is given immediately. The word u cannot be predecessor or successor of any other word in the hypothesis and is thus deleted from the hypothesis. After at most $(2 + 1 + 1) \cdot |c| = O(\|c\|)$ examples the $\rhd'''_{fin}$-restricted learner has reached the target. □

5.3 Teaching Finite Natural Concept Classes

After considering somewhat artificial situations, we have seen that the $\rhd$-restricted model allows for various feedback and finite vs. in the limit effects and is sensitive to the order of examples. Now we investigate whether these effects can be observed for more natural concept classes and $\rhd$-restrictions, too.

5.3.1 Monomials

We consider the class of all monomials over n variables excluding the contradictory ones. Then all representations are strings from $\{0, 1, *\}^n$. We define the neighborhood relation

$$\sigma \rhd_{Mon} \zeta \quad :\Leftrightarrow \quad \text{there is exactly one } i \text{ with } \sigma[i] \neq \zeta[i].$$

For the initial hypothesis we set $\mathit{init} \rhd_{Mon} *^n$ and $\mathit{init} \not\rhd_{Mon} \zeta$ for all $\zeta \neq *^n$. For example, $\mathtt{1*01} \rhd \mathtt{**01}$, $\mathtt{000*} \rhd \mathtt{100*}$, $\mathtt{1010} \not\rhd \mathtt{**10}$.

Fact 5.8 *Let $1 \leq k \leq n$. A monomial $c^* \in \mathcal{M}^1_n$ with k literals is finitely teachable without feedback to the $\rhd_{Mon}$-restricted learner with memory size $m \geq 2$. The optimal teaching time is at most $k + \lceil \frac{k}{m-1} \rceil$. For $k = 0$ the optimal teaching time is 1.*

Proof. For $k = 0$ the teacher only gives one arbitrary example and the learner must switch from *init* to $*^n$.

Let $k \geq 1$ and let without loss of generality $\sigma^* = \mathtt{1}^k\mathtt{*}^{n-k}$ be the target. We denote by $z^+ = (1^k 0^{n-k}, 1)$ a positive example and, for $j = 1, \ldots, k$, by $z_j^- = (1^{j-1}01^{k-j}0^{n-k}, 0)$ negative examples for the target. The teacher T first gives the example z^+. Then it

	Hypothesis	Memory	Example
0	*init*	$\langle\,\rangle$	(11100, 1)
1	`*****`	$\langle(11100,1)\rangle$	(01100, 0)
2	`1****`	$\langle(11100,1),$ $(01100,0)\rangle$	(11100, 1)
3	`1****`	$\langle(01100,0),$ $(11100,1)\rangle$	(10100, 0)
4	`11***`	$\langle(11100,1),$ $(10100,0)\rangle$	(11100, 1)
5	`11***`	$\langle(10100,0),$ $(11100,1)\rangle$	(11000, 0)
6	`111**`	$\langle(11100,1),$ $(11000,0)\rangle$	— — —

	Hypothesis	Memory	Example
0	*init*	$\langle\,\rangle$	(11100, 1)
1	`*****`	$\langle(11100,1)\rangle$	(01100, 0)
2	`1****`	$\langle(11100,1),$ $(01100,0)\rangle$	(10100, 0)
3	`11***`	$\langle(11100,1),$ $(01100,0),$ $(10100,0)\rangle$	(11100, 1)
4	`11***`	$\langle(01100,0),$ $(10100,0),$ $(11100,1)\rangle$	(11000, 0)
5	`111**`	$\langle(10100,0),$ $(11100,1),$ $(11000,0)\rangle$	— — —

Figure 5.2 *Teaching the monomial* `111**` *finitely to the* $\rhd_{Mon}$*-restricted learner with memory size 2 (left) and 3 (right). Note how the positive example* $(11100,1)$ *is taught less often to the learner with bigger memory.*

basically goes on with $z_1^-, \ldots, z_k^-$, but whenever the learner is about to forget z^+, that is, every m-th round, the teacher provides z^+ again (see Figure 5.2). Formally, for all $i < k + \lceil \frac{k}{m-1} \rceil$:

$$T(i) = \begin{cases} z^+ & \text{if } i \equiv 0 \pmod m, \\ z_j^- \text{ with } j = i - \lceil i/m \rceil & \text{otherwise.} \end{cases}$$

Let $\mathcal{L}$ be the $\rhd_{Mon}$-restricted learner with memory size m, and let its hypothesis after the i-th round be h_i. Then $h_0 = \mathit{init}$ and $h_1 = \mathtt{*}^n$. Next, the teacher T gives z_1^- and $\mathcal{L}$ has to find a hypothesis consistent with z^+ and z_1^- differing from $\mathtt{*}^n$ in only one position. The only such hypothesis is $h_1 = \mathtt{1*}^{n-1}$ because a "`0`" among the first k positions would contradict z^+, a "`1`" in a position other than the first one would contradict z_1^-. A "`0`" within the last $n-k$ positions would also contradict z_1^-, and a "`1`" within the last $n-k$ positions would contradict z^+. Therefore the only consistent hypothesis in the neighborhood of $\mathtt{*}^n$ has a "`1`" in the first position.

Similarly one shows that the examples $z_2^-, \ldots, z_{m-1}^-$ enforce the hypotheses $h_2 = \mathtt{11*}^{n-2}, \ldots, h_{m-1} = \mathtt{1}^{m-1}\mathtt{*}^{n-m+1}$. Then the teacher gives z^+ again. This ensures that for the next $m-1$ negative examples the memory contains z^+ and the above reasoning can be applied when $z_m^-, z_{m+1}^- \ldots$ are given. This shows that when finally z_k^- is taught, the learner must reach the target hypothesis.

Overall, there are k negative examples given plus $\lceil \frac{k}{m-1} \rceil$ positive examples. □

We conjecture that the teacher in the last fact is optimal. This would give us a natural situation in which the teaching time decreases with growing memory; although the effect is not very strong. It certainly gives us a natural situation in which the order of examples matters.

The monomials cannot be taught if the memory gets smaller than 2, even if the teacher receives feedback.

Fact 5.9 *Let $n \geq 2$ and $1 \leq k \leq n$. A monomial $c^* \in \mathcal{M}_n^1$ with k literals is not teachable with feedback to the $\rhd_{Mon}$-restricted learner with memory size 1.*

Proof. We describe a deterministic learner L that cannot be forced to reach the target hypothesis $\sigma^* = \mathtt{1}^k\mathtt{*}^{n-k}$. It suffices to show that when L assumes a hypothesis σ with $\sigma \rhd_{Mon} \sigma^*$ it never switches to σ^*.

Let L assume a hypothesis σ with $\sigma \rhd_{Mon} \sigma^*$, and let $z = (x, b)$ be the newly received example. As a learner with 1-memory, the behavior of L does not depend on the memory, only on σ and z. Moreover we assume that z is inconsistent with the current hypothesis σ (otherwise L would not change the hypothesis anyway). We denote the follow-up hypothesis of L by σ'.

Case 1: σ and σ^* differ weakly at one of the first k positions.

Without loss of generality, let $\sigma = \mathtt{*1}^{k-1}\mathtt{*}^{n-k}$. Then z must be a negative example, and x must start with 01^{k-1}. We define $\sigma' = \mathtt{01}^{k-1}\mathtt{*}^{n-k}$. □ Case 1

Case 2: σ and σ^* differ strongly at one of the first k positions.

Without loss of generality, let $\sigma = \mathtt{01}^{k-1}\mathtt{*}^{n-k}$.

Case 2.1: z is a negative example.

Then x must start with 01^{k-1}. If $k \geq 2$ we set $\sigma' = \mathtt{001}^{k-2}\mathtt{*}^{n-k}$. If $k = 1$ then $n - k \geq 1$ and we set $\sigma' = \mathtt{01}^{k-1}\overline{x[2]}\mathtt{*}^{n-k-1}$.

Case 2.2: z is a positive example.

Then x starts with 1^k. We set $\sigma' = \mathtt{*1}^{k-1}\mathtt{*}^{n-k}$. □ Case 2

Case 3: σ and σ^* differ at one of the last $n - k$ positions.

Without loss of generality, let $\sigma = \mathtt{1}^k\mathtt{*}^{n-k-1}0$. Then z must be positive, and x must end with 1. We set $\sigma' = \mathtt{1}^k\mathtt{*}^{n-k-1}1$. □ Case 3

In all cases we have $\sigma \rhd_{Mon} \sigma'$, and σ' is consistent with the example z. But all σ' are different from σ^*. Therefore, the learner L never reaches σ^*. □

5.3.2 1-Decision Lists

There are several natural ways to define a neighborhood relation over the 1-decision lists. In the following we allow the insertion, removal, or replacement of exactly one node in the list at an arbitrary position. The default node may not be removed. For the hypothesis *init* the neighbors are the two decision lists containing only a default node. We denote the neighborhood relation so defined by $\rhd_{DL}$.

Fact 5.10 *The class $\mathcal{D}_n$ of 1-decision lists over n variables can be finitely taught to the $\rhd_{DL}$-restricted learner with $(n+1)$-memory. The optimal teaching time for a 1-decision list concept $c \in \mathcal{D}_n$ is $len(c) + 1$.*

Proof. The idea of the proof is that a teacher for 1-decision lists can use a canonical sample, as defined in the proof of Theorem 4.12 (for $\ell = \lambda$), but must take care of the correct order. Teaching is successful if the examples are given from back to front, that is, starting with the example belonging to the default node and ending with the example belonging to the first node. In the course of teaching, the learner is forced to reconstruct the target NFDL node by node, beginning with the default. This takes $len(c) + 1$ rounds, which is no more than the memory size $n + 1$, hence no example is ever forgotten.

Let $c^* \in \mathcal{D}_n$ be a target concept. Without loss of generality, we assume that c is represented by an NFDL D^* containing only the variables $v_1, \ldots, v_\ell$ in that order:

$$D^* = \langle (v_1, b_1), (v_2, b_2), \ldots, (v_\ell, b_\ell), (*, \bar{b}_\ell) \rangle \ .$$

The example belonging to node (v_i, b_i) is denoted by $z_i = (x_i, b_i)$; the default node's example by $z_{\ell+1} = (x_{\ell+1}, \bar{b}_\ell)$. Thus, the examples are taught in the order $z_{\ell+1}, \ldots, z_1$.

Claim: For all $i = \ell+1, \ldots, 1$: After the teacher has given example z_i, the hypothesis is equivalent to the decision list

$$D = \langle (v_i, b_i), \ldots, (v_\ell, b_\ell), (*, \bar{b}_\ell) \rangle.$$

Proof. We proof the claim by induction on i starting with $i = \ell + 1$. After $z_{\ell+1}$ the hypothesis is $\langle (*, \bar{b}_\ell) \rangle$ and the claim holds.

Now assume that the claim holds for some $i > 1$. We show it for $i - 1$. Before receiving z_i the hypothesis is of length $\ell + 1 - i$ and the learner memorizes $z_{\ell+1}, \ldots, z_i$. After receiving z_{i-1} the learner memorizes $z_{\ell+1}, \ldots, z_{i-1}$. We know from the proof of Theorem 4.12 that there is only one concept in $\mathcal{D}_n$ of length $\ell + 2 - i$ that is consistent with these examples. This concept is represented by the 1-decision list

$$D' = \langle (v_{i-1}, b_{i-1}), (v_i, b_i), \ldots, (v_\ell, b_\ell), (*, \bar{b}_\ell) \rangle.$$

Moreover, no shorter 1-decision list is consistent with the examples $z_{\ell+1}, \ldots, z_{i-1}$.

In order to maintain a consistent hypothesis, the learner has to switch from the current hypothesis, which is equivalent to D, to a hypothesis that is equivalent to D'. This is possible by inserting the node (v_{i-1}, b_{i-1}) into the first segment of D. Therefore all learners switch to a hypothesis equivalent to D'. This proves the claim for $i-1$. □ Claim

For $i = 1$ the claim implies that after all examples have been given, the hypothesis is equivalent to the target D^*. The claim also shows that D^* can be taught within $\ell + 1$ rounds. On the other hand, $\ell + 1$ rounds is also a lower bound for the teaching time since in each round the hypothesized 1-decision list can only grow by one node. Therefore, $\ell + 1 = len(D^*) + 1$ is the optimal teaching time. □

Fact 5.11 *The class of 1-decision lists over n variables cannot be taught to the $\rhd_{DL}$-restricted learner with $(n-1)$-memory, not even with feedback.*

Proof. We call a 1-decision list *positive* iff all nodes except the default node and its predecessor consist of a positive literal and a positive label. To simplify notation we denote the representation function for 1-decision lists by ϱ.

To show that the class cannot be taught to the learner with $(n-1)$-memory, it suffices to define a $\rhd_{DL}$-restricted deterministic learner that cannot reach the target concept $c^* = \{0,1\}^n \setminus \{0^n\} \in \mathcal{D}_n$. This target concept is represented by the positive NFDL $\langle (v_1, 1), \ldots, (v_n, 1), (*, 0) \rangle$.

Let L be a deterministic $\rhd_{DL}$-restricted learner with $(n-1)$-memory that, whenever possible, hypothesizes a positive decision list of minimum length consistent with all memorized examples. The behavior of L in case there is no such hypothesis available is irrelevant for this proof.

Let T be a teacher with feedback and let $(\boldsymbol{s}_t)_{t\in\mathbb{N}}$, $(\sigma_t)_{t\in\mathbb{N}}$, and $(z_t)_{t\in\mathbb{N}}$ be the series of memory contents, hypotheses, and examples resulting from T teaching c^* to L. The instance in example z_t is denoted x_t for all $t \in \mathbb{N}$.

The first example, $z_0 = T(\langle\ \rangle, init)$, can either be positive or negative. We first assume that it is negative, that is, $z_0 = (0^n, 0)$ and consider the positive case after the next claim.

Claim: For all $t \geq 1$:

(1) σ_t is a positive decision list,

(2) $\varrho_{\sigma_t} \subset c^*$,

(3) σ_t is consistent with $\boldsymbol{s}_t$.

Proof. Since the first example is negative, the learner switches from *init* to $\sigma_1 = \langle (*, 0) \rangle$, and $\boldsymbol{s}_t = \langle z_0 \rangle$. Therefore $\varrho_{\sigma_t} = \emptyset$ and (1), (2), (3) hold for the induction basis.

Now assume (1), (2), (3) hold for some $t \geq 1$. Then $\boldsymbol{s}_{t+1} = \boldsymbol{s}_t \circ_{n-1} z_i$. We distinguish three cases depending on z_t.

Case 1: z_t is consistent with σ_t.

Then, by definition of $\rhd_{DL}$-restricted learner, $\sigma_{t+1} = \sigma_t$ and (1), (2), (3) for $t+1$ follow immediately from the induction hypothesis. □ Case 1

Case 2: z_t is inconsistent with σ_t and $len(\sigma_t) \leq n-2$.

Whatever σ_{t+1} will be, its length is less then n, hence (2) holds for $t+1$. Now we show that there is a positive and consistent hypothesis in the neighborhood of σ_t.

The example z_t is a positive example, otherwise it would be consistent with σ_t (due to Item (2)). Then there must be an $i \in \{1, \ldots, n\}$ such that $x_t[i] = 1$. Moreover, the node $(v_i, 1)$ is not contained in σ_t. We define σ' by prepending the node $(v_i, 1)$ to σ_t. This hypothesis is in the neighborhood of σ_t and positive. It is also consistent with all examples in $\boldsymbol{s}_{t+1}$ as we shall show next.

All instances classified as positive by σ_t are also classified as positive by σ', hence $\varrho_{\sigma'} \supseteq \varrho_{\sigma_t}$. Therefore all positive examples in $\boldsymbol{s}_t$ are consistent with σ'. The only negative example, $(0^n, 0)$, is consistent with σ' as well. The only positive example in $\boldsymbol{s}_{t+1}$ not already in $\boldsymbol{s}_t$ is z_t. This example is consistent with σ' since σ' contains the node $(v_i, 1)$ and $x_t[i] = 1$.

As σ' is a positive 1-decision list of length less than n, the concept $\varrho_{\sigma'}$ is a proper subset of c^*, which is represented by a positive 1-decision list of length n.

We have now shown that there is a hypothesis satisfying (1), (2), (3) and reachable from σ_t. Therefore the learner L will switch to such a hypothesis, for which (1), (2), (3) hold. □ Case 2

Case 3: z_t is inconsistent with σ_t and $len(\sigma_t) = n-1$.

Then σ_t contains all nodes with positive label except for one, say $(v_k, 1)$. Then $x_t = 0^{k-1}10^{n-k}$ since z_t is a positive example classified as negative by σ_t. There are at most $n-1$ examples in $\boldsymbol{s}_{t+1}$, hence there is a $j \in \{1, \ldots, n\}$ such that $0^{j-1}10^{n-j}$ is *not* contained in $\boldsymbol{s}_{t+1}$. Since z_t *is* contained in $\boldsymbol{s}_{t+1}$, we have $j \neq k$. It follows that the node $(v_j, 1)$ is in σ_t. Let σ' be the 1-decision list resulting from replacing $(v_j, 1)$ by $(v_k, 1)$. This is a positive decision list in the neighborhood of σ_t and $\varrho_{\sigma'} \subset c^*$.

Now we show that σ' is consistent with $\boldsymbol{s}_{t+1}$. Node $(v_j, 1)$ absorbs z_t and classifies it correctly. The instance 0^n is also classified correctly (as negative) by σ'.

It remains to prove consistency for the other positive examples $z = (x, 1)$ in $\boldsymbol{s}_t$. If $x[j] = 1$ then there is another $j' \neq j$ with $x[j'] = 1$ since $0^{j-1}10^{n-j}$ is not contained in $\boldsymbol{s}_{t+1}$. The node $(v_{j'}, 1)$ is in σ' because it was in σ_t and has not been replaced. Hence z is classified as positive by σ'. If $x[j] = 0$ then there is another $j' \neq j$ with $x[j'] = 1$ since z is a positive example. Again $(v_{j'}, 1)$ is in σ', hence σ' classifies z positive.

We have now shown that there is a hypothesis satisfying (1), (2), (3) and reachable from σ_t. Therefore the learner L will switch to such a hypothesis, for which $(1), (2), (3)$ hold. □ Case 3

□ Claim

The claim shows that L never hypothesizes c^*.

It remains to consider the case of a positive first example z_0. Then the learner assumes the hypothesis $\sigma_1 = \langle(*, 1)\rangle$. As long as the teacher continues giving positive examples, the hypothesis does not change due to the learner L being conservative. When the first negative example has arrived, the memory contains a positive and a negative example and thus the hypothesis changes to $\langle(\bar{v}, 0), (*, 1)\rangle$ for some variable v, because removal of the default node is forbidden and replacing the default node would result in a decision list that misclassifies the positive examples; a node $(v, 0)$ with a variable v would not absorb the negative example $(0^n, 0)$. This new hypothesis is equivalent to the positive 1-decision list $\langle(v, 1), (*, 0)\rangle$. This 1-decision list satisfies the conditions (1), (2), (3) in the claim above. Using the same arguments as in the claim, one can show that L does not reach the target c^*. □

We have shown that for monomials and 1-decision lists there is an m^* such that for $m < m^*$ they are unteachable even with feedback and for $m \geq m^*$ they are teachable even without feedback. We know of no finite, natural concept class and hypothesis restriction where teaching to restricted learners without feedback is more difficult than with feedback.

In general, a teacher without feedback knows the learner's hypothesis in the beginning, but can quickly lose track of it during teaching. However, the teachers presented in the Facts 5.8 and 5.10 for monomials and 1-decision lists do not have this problem. The examples they give force the learner to perform certain hypothesis changes without being able to choose from several alternative hypotheses. Our results about monomials and 1-decision lists suggest that natural concept classes, together with natural $\rhd$-restrictions, are susceptible to this kind of "enforcement strategy."

5.4 Computing the Optimal Teaching Time

In this section we investigate the computational complexity of finding the optimal teaching time for $\rhd$-restricted learners. We define several decision problems according to various combinations of feedback and memory size.

Before we begin we have to give some thoughts to the representation of the teaching problem instances. Depending on whether the set of valid representations $\varrho(\mathcal{C})$ is finite or infinite (the latter can happen also for finite concept classes) we have to specify the problem in different ways. If $\varrho(\mathcal{C})$ is finite then $(\varrho(\mathcal{C}) \cup \{init\}, \rhd)$ is a finite directed

graph whose nodes are all representations and whose arcs specify the allowed transitions between the representations.

Definition 5.12 *Let $\langle \varrho, \mathcal{C} \rangle$ be a represented concept class and $\rhd$ a neighborhood relation. The* $\rhd$-restriction graph *is then the directed graph $(\varrho(\mathcal{C}) \cup \{init\}, \rhd)$ where each node $\sigma \in \varrho(\mathcal{C})$ is labeled with the function ϱ_σ.*

If X and $\varrho(\mathcal{C})$ are finite then the $\rhd$-restriction graph is finite and all node labels $\varrho_\sigma \colon X \to \{0,1\}$ can be represented finitely. As an example, Figure 5.1 shows the graph for the concept class $\mathcal{S}_n$ with $\rhd$-restriction defined so as to allow only transitions between concepts with Hamming distance 1.

A finitely representable restriction graph allows a finite description of a teaching problem. More precisely, we use a representation of size $(|\varrho(\mathcal{C})|+1)^2$ bits for the graph plus $|\varrho(\mathcal{C})| \cdot |X|$ bits for the labels. We assume this representation for the definition of our teaching problems. As usual, we distinguish problems of teaching with and without feedback and with varying memory sizes.

Definition 5.13 *Let $m \in \mathbb{N}^+ \cup \{\infty\}$. The decision problem* m-FB-TEACHINGTIME *is defined as follows:*

Instance: *A restriction graph $(\varrho(\mathcal{C}) \cup \{init\}, \rhd)$, a target representation σ^*, a number $\ell \in \mathbb{N}$.*

Question: *Is ϱ_{σ^*} teachable with feedback to the $\rhd$-restricted learner with memory size m in at most ℓ rounds?*

Definition 5.14 *The decision problem* FB-TEACHINGTIME *is defined as follows:*

Instance*: A restriction graph $(\varrho(\mathcal{C}), \rhd)$, a target representation σ^*, a number $m \in \mathbb{N}^+$, a number $\ell \in \mathbb{N}$.*

Question*: Is ϱ_{σ^*} finitely teachable with feedback to the $\rhd$-restricted learner with memory size m in at most ℓ rounds?*

Definition 5.15 *Let $m \in \mathbb{N}^+ \cup \{\infty\}$. The decision problem* m-NOFB-TEACHINGTIME *is defined as follows:*

Instance: *A restriction graph $(\varrho(\mathcal{C}) \cup \{init\}, \rhd)$, a target representation σ^*, a number $\ell \in \mathbb{N}$.*

Question: *Is ϱ_{σ^*} finitely teachable without feedback to the $\rhd$-restricted learner with memory size m in at most ℓ rounds?*

As we will see later, it makes sense to consider the problems in which the number ℓ of rounds is represented unary.

Definition 5.16 *Let $m \in \mathbb{N}^+ \cup \{\infty\}$. The decision problem* UNARY-m-FB-TEACHING-TIME *is defined as follows:*

Instance: *A restriction graph $(\varrho(\mathcal{C}) \cup \{init\}, \rhd)$, a target representation σ^*, the string 1^ℓ for an $\ell \in \mathbb{N}$.*

Question: *Is ϱ_{σ^*} teachable with feedback to the $\rhd$-restricted learner with memory size m in at most ℓ rounds?*

Definition 5.17 *Let $m \in \mathbb{N}^+ \cup \{\infty\}$. The problem* UNARY-$m$-NOFB-TEACHINGTIME *is defined as follows:*

Instance: *A restriction graph $(\varrho(\mathcal{C}) \cup \{init\}, \rhd)$, a target representation σ^*, the string 1^ℓ for an $\ell \in \mathbb{N}$.*

Question: *Is ϱ_{σ^*} finitely teachable without feedback to the $\rhd$-restricted learner with memory size m in at most ℓ rounds?*

The unary and non-unary versions differ with respect to their computational complexity only if the optimal teaching time can be superpolynomially large in the representation of the graph. It is open whether this can occur.

In the $\mathfrak{H}$-restricted model (and therefore also in the TD model) the optimal teaching time cannot be greater than the representation of the teaching problem (that, is the class $\mathcal{C}$) because it is bounded from above by $|X|$. Thus, in the $\mathfrak{H}$-restricted model the unary and non-unary versions are equivalent.

Theorem 5.18 *For every $m \geq 1, m \neq \infty$ the problem* m-FB-TEACHINGTIME *can be solved in time $(|X|^m \cdot |\varrho(\mathcal{C})|)^{O(1)}$:*

Proof. The learner can be in any state from the set $State = \mathcal{X}^m \times (\varrho(\mathcal{C}) \cup \{init\})$. The idea of the algorithm is to compute the τ-values for all states using the formula in Lemma 2.15. This lemma also implies that all τ-values are upper bounded by $|State|$ because for every occurring τ-value, say r, there must be a state with τ-value $r-1$.

For a convenient description of the algorithm, let $\mathcal{L}$ be the m-memory learner with $\rhd$-restriction.

Input: Restriction graph $(\varrho(\mathcal{C}) \cup \{init\}, \rhd)$; a number ℓ.

1 **if** $\ell > |State|$ **then reject**

2 **for all** $(\boldsymbol{s}, \sigma) \in State : u(\boldsymbol{s}, \sigma) := \uparrow$

3 **for all** $(\boldsymbol{s}, \sigma) \in State$ with $\varrho_\sigma = \varrho_{\sigma^*} : u(\boldsymbol{s}, \sigma) := 0$

4 **for** $i = 1$ **to** ℓ :

5 **for all** $(\boldsymbol{s}, \sigma) \in State$ with $u(\boldsymbol{s}, \sigma) = \uparrow$:

6 **if** $\exists z \in \mathcal{X}\ \forall (\boldsymbol{s}', \sigma') \in \mathcal{L}(\boldsymbol{s}, \sigma, z) : u(\boldsymbol{s}, \sigma') \neq \uparrow$ **then** $u(\boldsymbol{s}, \sigma) := i$

7 **if** $u(\langle\,\rangle, init) \leq \ell$ **then accept**

8 **else reject**

To show correctness, it suffices to show that at the end of every iteration $i \leq \ell$ of the outmost for-loop, $\tau(\boldsymbol{s}, \sigma) = i \Leftrightarrow u(\boldsymbol{s}, \sigma)$. This is clearly true for $i = 0$ (that is, before iteration $i = 1$). Assume the equivalence holds for all iterations up to an $i < \ell$. For

the one implication, let $\tau(\boldsymbol{s},\sigma) = i+1$. Then $u(\boldsymbol{s},\sigma) = \uparrow$ at the beginning of iteration $i+1$, and by the formula in Lemma 2.15 there is an example z such that all states in $\mathcal{L}(\boldsymbol{s},\sigma,z)$ have an τ-value of at most i. By the induction hypothesis the corresponding u-values have already been set to something different than "$\uparrow$". Thus the condition in line 6 is satisfied and $u(\boldsymbol{s},\sigma)$ is set to $i+1$. This value will never be changed because the for-loop in line 5 will never consider $(\boldsymbol{s},\sigma)$ again.

For the other implication, let $u(\boldsymbol{s},\sigma) = i+1$ at the end of iteration $i+1$. Then the u-value has been "$\uparrow$" at the beginning of the iteration, and by the induction hypothesis $\tau(\boldsymbol{s},\sigma) \not\leq i$. Moreover, since the condition in line 6 has been satisfied, there is an example z such that all states in $\mathcal{L}(\boldsymbol{s},\sigma,z)$ have an u-value of at most i. By the induction hypothesis, their τ-values are at most i, too. Then applying the formula in Lemma 2.15 yields $\tau(\boldsymbol{s},\sigma) = i+1$.

The runtime is $O(|\mathit{State}|)$ for the first three lines plus $O(\ell \cdot |\mathit{State}| \cdot |X| \cdot |\mathit{State}|) \leq O(|\mathit{State}| \cdot |\mathit{State}| \cdot |X| \cdot |\mathit{State}|)$ for the rest. This proves the claimed runtime, since $|\mathit{State}| \leq O(|X|^m \cdot |\varrho(\mathcal{C})|)$. □

If the memory size m is part of the input, the above theorem shows that the teachability problem is fixed-parameter tractable with parameter m. We do not expect that there is an efficient algorithm for the problem when m is part of the input, since we show later (Corollary 5.22) that this problem is $\mathcal{PSPACE}$-hard.

For our $\mathcal{PSPACE}$-hardness proofs we use the `SEQUENTIAL-TRUTH-ASSIGNMENT` problem, which is known to be $\mathcal{PSPACE}$-complete (see Stockmeyer and Meyer [55] and also Garey and Johnson [32]). We use the variant in which the formula is in 3-CNF.

Definition 5.19 *The following problem is called* `SEQUENTIAL-TRUTH-ASSIGNMENT`*:*
Instance: *Set $V = \{v_1, \ldots, v_n\}$ of variables, clauses $K_1, \ldots, K_r$ with three literals.*
Question: *Is the formula $\exists v_1\ \forall v_2\ \exists v_3 \ldots : K_1 \wedge \ldots \wedge K_r$ true?*

We can assume that the number n of variables is even by adding a dummy variable if necessary.

A sequential truth assignment can be seen as a two player game. Player 1 assigns a truth value to variable v_1, then Player 2 assigns a truth value to v_2 and so on until after n rounds all variables have been assigned a value. Player 1 wins if this variable assignment satisfies the formula $K_1 \wedge \ldots \wedge K_r$. The problem `SEQUENTIAL-TRUTH-ASSIGNMENT` can then be interpreted as the question whether Player 1 has a winning strategy for this game. In the following proofs we exploit that the teaching process can be seen as a two player game, too (see Page 27).

Theorem 5.20 *The problem ∞-`FB-TEACHINGTIME` is $\mathcal{PSPACE}$-hard.*

Proof. We prove the $\mathcal{PSPACE}$-hardness via a polynomial time reduction from the `SEQUENTIAL-TRUTH-ASSIGNMENT` problem.

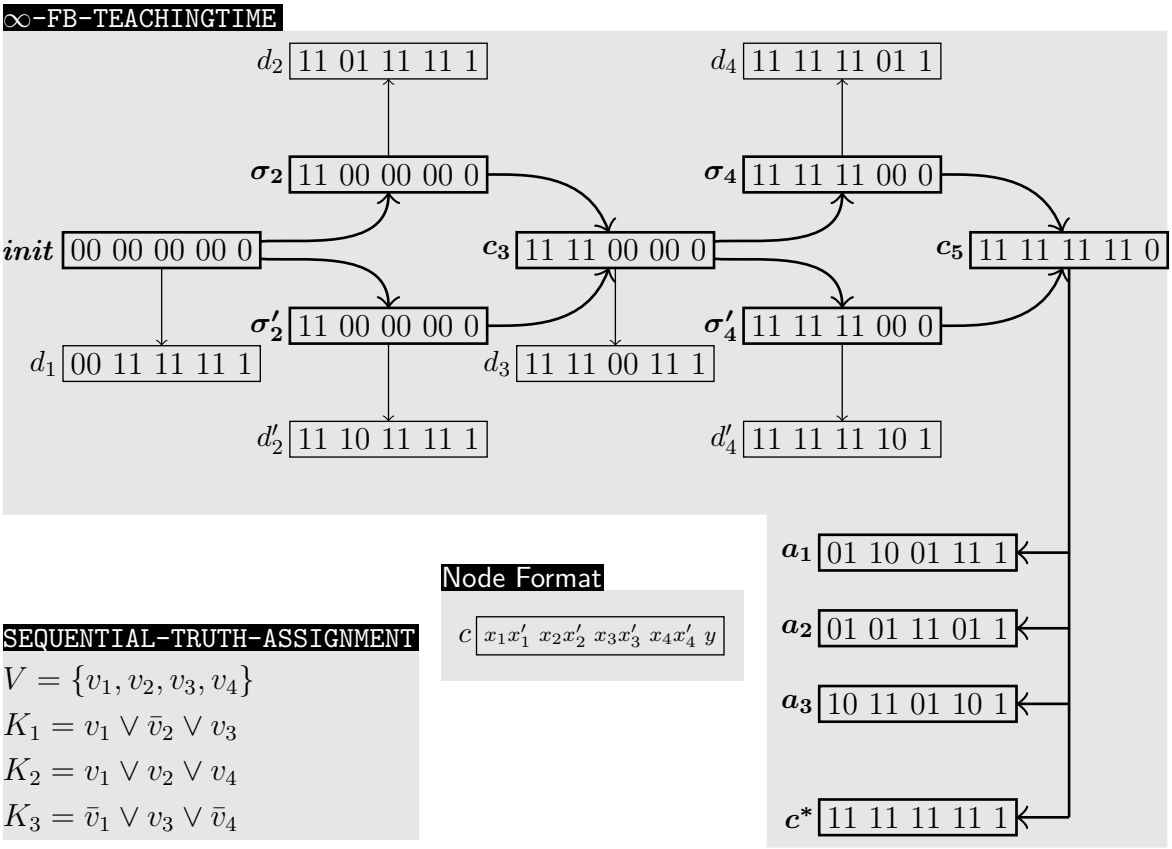

Figure 5.3 *Example for the reduction from the* SEQUENTIAL-TRUTH-ASSIGNMENT *problem to* ∞*-FB-TEACHINGTIME used in the proof of Theorem 5.20. The instance in the lower left corner is mapped to the restriction graph displayed. Each node is labeled with a concept over* $\{x_1, x'_1, \ldots, x_4, x'_4, y\}$*. The bold nodes and arcs describe paths that a learner is supposed to take. The non-bold arcs lead to dead ends in the graph; a successful teacher must prevent the learners from going there. The nodes between init and* c_5 *force the teacher to give examples that can be interpreted as a truth assignment to the variables in* V*. The hypotheses* a_1, a_2, a_3 *correspond to the clauses* $K_1, K_2.K_3$*.*

Let $V = \{v_1, \ldots, v_n\}$ be a set of Boolean variables and $\{K_1, \ldots, K_r\}$ a set of clauses over V with each clause containing three literals: $K_i = v_{i_1}^{\alpha_{i_1}} \vee v_{i_2}^{\alpha_{i_2}} \vee v_{i_3}^{\alpha_{i_3}}$ for $i = 1, \ldots, r$. We assume that n is even. The idea of the reduction is to build a restriction graph that consists of two parts. The first part ensures that a successful teacher gives a sequence of examples that can be interpreted as alternating assignments of truth values to the variables in V. The other part checks whether this truth assignment satisfies all clauses.

Now we describe the instance of ∞-FB-TEACHINGTIME in detail for the sake of precision. The construction can most likely be understood much better from the example shown in Figure 5.3. Let $X = \{x_1, x'_1, \ldots, x_n, x'_n\} \cup \{y\}$ be the learning domain. The

concept class $\mathcal{C}$ contains the following concepts: $c_j = \{x_1, x'_1, \ldots, x_{j-1}, x'_{j-1}\}$ for all $j = 2, \ldots, n+1$ and the target $c^* = X$. For odd numbers $j = 1, 3, 5, \ldots, n-1$ there are concepts $d_j = X \setminus \{x_j, x'_j\}$. For even numbers $j = 2, 4, 6, \ldots, n$ there are concepts $d_j = X \setminus \{x_j\}$ and $d'_j = X \setminus \{x'_j\}$. Finally, there is a concept a_i for every clause K_i: $a_i = X \setminus (\{x_j \mid v_j \text{ is in } K_i\} \cup \{x'_j \mid \bar{v}_j \text{ is in } K_i\})$. For example, if $K_i = v_1 \vee \bar{v}_2 \vee v_3$ then $a_i = X \setminus \{x_1, x'_2, x_3\}$. Note that "officially" there is no concept called c_1. According to the above definition, we have $c_1 = \emptyset$, which means that c_1 behaves just like the hypothesis *init*. So, whenever we refer to "c_1" for convenience, we actually mean "*init*".

The nodes in the restriction graph are representations, but we do not need to introduce an identifier for all representations. It suffices to say that there is one representation for every concept defined, but for every concept c_j with j even there are *two* representations, denoted σ_j and σ'_j. To the other concepts we need not refer with their representations, as they are unique. The arcs of the restriction graph are as follows. There is an arc from *init* to d_1, to σ_2, and to σ'_2. There is an arc from each concept c_i with $i > 1$ odd to d_i, and for i even: from σ_i to d_i and from σ'_i to d'_i. There is also an arc from c_i to c_{i+1} for $i = 3, 5, \ldots, n-1$ and from σ_i to c_{i+1} and from σ'_i to c_{i+1} for $i = 2, 4, \ldots, n-2$. Finally there is an arc from c_{n+1} to each node a_i $(i = 1, \ldots, r)$ and to c^*. To complete the ∞-FB-TEACHINGTIME instance we set $\ell = n + 1$.

There are $3n + 2 + r$ representations. Each represented concept can be written using $|X| = n + 1$ bits. The total size of the restriction graph and all concepts is thus polynomial in the total size of the clauses. Moreover the restriction graph can be computed straightforwardly given the clauses. Therefore the reduction described above is polynomial time.

Intuitively, the teaching scenario built by the reduction works as follows. In an even round i the learner hypothesizes c_{i+1} and the teacher can either give the example $(x_{i+1}, 1)$ or $(x'_{i+1}, 1)$. This choice corresponds to an assignment to the variable v_{i+1} by Player 1. In the following odd round $i+1$ the learner can choose to either go to σ_{i+2} or to σ'_{i+2}. This choice cannot be influenced by the teacher because both representations belong to the same concept. If the learner chooses σ_{i+2}, the teacher is forced to give the example $(x_{i+2}, 1)$. If the learner chooses σ'_{i+2}, the teacher is forced to give $(x'_{i+2}, 1)$. This corresponds to an assignment to the variable v_{i+2} by Player 2.

In round n, for each $i \in \{1, \ldots, n\}$ either $(x_i, 1)$ or $(x'_i, 1)$ has been given. This corresponds to an assignment to all variables in V. In the final step the teacher gives the example $(y, 1)$, which enforces on the learner a transition from c_{n+1} to c^* if and only if the variable assignment satisfies all clauses $K_1, \ldots, K_r$. If one clause K_j is unsatisfied then the learner could as well switch to the dead-end hypothesis a_j, and the teacher would fail.

Now we show more formally that the reduction maps positive instances to positive instances and negative ones to negative ones.

Assume that the SEQUENTIAL-TRUTH-ASSIGNMENT instance is positive, that is, the formula $\exists v_1\ \forall v_2 \ldots \exists v_{n-1}\ \forall v_n : K_1 \wedge K_2 \wedge \ldots \wedge K_r$ is true. As shortcut we set $F := K_1 \wedge \ldots \wedge K_r$, and for a partial variable assignment $\beta\colon V \to \{\mathit{true}, \mathit{false}, \mathit{undefined}\}$ we let F_β be the formula that results by substituting all literals in F with *true* or *false* as specified by β. For a partial assignment β and a value $b \in \{\mathit{true}, \mathit{false}\}$ we write $\beta \cup \{v_i \mapsto b\}$ to denote the assignment that assigns b to v_i and otherwise works as β.

For non-empty memory $\boldsymbol{s} \in \mathcal{X}(c^*)^*$ we define the partial assignment $\beta_{\boldsymbol{s}}\colon V \to \{\mathit{true}, \mathit{false}, \mathit{undefined}\}$ encoded in $\boldsymbol{s}$ by

$$\beta_{\boldsymbol{s}}(v_i) = \begin{cases} \mathit{true} & \text{if } (x_i, 1) \in \boldsymbol{s}, \\ \mathit{false} & \text{if } (x_i, 1) \notin \boldsymbol{s} \text{ and } (x_i', 1) \in \boldsymbol{s}, \\ \mathit{undefined} & \text{otherwise.} \end{cases}$$

If both $(x_i, 1)$ and $(x_i', 1)$ are present in $\boldsymbol{s}$ then v_i is, by definition of $\beta_{\boldsymbol{s}}$, assigned *true*. However, this case will not occur.

Now the idea for the successful teacher is to follow the winning strategy for Player 1 in the rounds $0, 2, \ldots, n-2$. This strategy ensures that there is always a suitable example, no matter what examples have to be given in the rounds $1, 3, \ldots, n-1$. More formally, we define the following teacher T. At first

$$T(\langle\,\rangle, \mathit{init}) = \begin{cases} (x_1, 1) & \text{if } \forall v_2 \exists v_3 \ldots \forall v_n : F_{\{v_1 \mapsto \mathit{true}\}} \text{ is true,} \\ (x_1', 1) & \text{otherwise.} \end{cases}$$

Note that the "otherwise" case is equivalent to "$\forall v_2 \exists v_3 \ldots \forall v_n : F_{\{v_1 \mapsto \mathit{false}\}}$ is true." Thus the teacher starts according to the first move of the winning strategy of Player 1. In general, for $i = 3, 5, \ldots, n-1$,

$$T(\boldsymbol{s}, c_i) = \begin{cases} (x_i, 1) & \text{if } \forall v_{i+1} \exists v_{i+2} \ldots \forall v_n : F_{\beta_{\boldsymbol{s}} \cup \{v_i \mapsto \mathit{true}\}} \text{ is true,} \\ (x_i', 1) & \text{otherwise.} \end{cases} \tag{5.1}$$

Furthermore, for $i = 2, 4, \ldots, n$,

$$\begin{aligned} T(\boldsymbol{s}, \sigma_i) &= (x_i, 1), \\ T(\boldsymbol{s}, \sigma_i') &= (x_i', 1). \end{aligned} \tag{5.2}$$

And finally for the hypothesis c_{n+1}:

$$T(\boldsymbol{s}, c_{n+1}) = (y, 1).$$

Now we show that T teaches c^* within $n+1$ rounds to the non-deterministic $\rhd$-restricted learner $\mathcal{L}$ with ∞-memory, where $\rhd$ is defined by the restriction graph described above. The proof is based on the following claim.

Claim 1: At the even rounds $t = 0, 2, 4, \ldots, n$ of the teaching process:

1. $\mathcal{L}$ hypothesizes c_{t+1},
2. the memory of $\mathcal{L}$ contains t examples, namely for all $i = 1, \ldots, t$ exactly one example from $\{(x_i, 1), (x'_i, 1)\}$,
3. the formula $\exists v_{t+1} \ldots \forall v_n : F_{\beta_{\boldsymbol{s}}}$ is true.

Proof. We proceed by induction on the number of rounds.

As induction basis we prove the claim for $t = 0$. Item 1 holds because *init* has the same behavior as c_1 had. Item 2 holds because initially the memory is empty. Item 3 holds because $\exists v_1 \ldots \forall v_n : F$ is true by assumption.

Now, let us assume the claim for some round $t \in \{0, 2, 4, \ldots, n-2\}$. We show it for $t+2$. Let $\boldsymbol{s}$ be the memory of $\mathcal{L}$ at round t, and let $b \in \{\mathit{true}, \mathit{false}\}$ be such that $\forall v_{t+2} \exists v_{t+3} \ldots \forall v_n : F_{\beta_{\boldsymbol{s}} \cup \{v_{t+1} \mapsto b\}}$ is true. Such a b exists due to the induction hypothesis. Now the definition (5.1) of T implies that the example $(x_{t+1}, 1)$ is taught if $b = \mathit{true}$ and example $(x'_{t+1}, 1)$ if $b = \mathit{false}$. Then the memory of $\mathcal{L}$ in round $t+1$ is $\boldsymbol{s}' = \boldsymbol{s} \circ \langle T(\boldsymbol{s}, c_{t+1}) \rangle$. Thus the assignment $\beta_{\boldsymbol{s}'}$, which corresponds to the new memory of $\mathcal{L}$, equals the assignment $\beta' = \beta \cup \{v_{t+1} \mapsto b\}$, which makes $\forall v_{t+2} \exists v_{t+3} \ldots \forall v_n : F_{\beta'}$ true.

The example given by T causes the current hypothesis c_{t+1} and the hypothesis d_{t+1} to have exactly one error, namely with respect to the instance x_{t+1}. The remaining neighborhood hypotheses, σ_{t+2} and σ'_{t+2}, have no errors (here we need Item 2 of the induction hypothesis). The learner's next hypothesis is therefore one of these two. If it is σ_{t+1} then T gives the example $(x_{t+2}, 1)$ next; if it is σ'_{t+1} then T gives the example $(x'_{t+2}, 1)$, according to (5.2). The new memory $\boldsymbol{s}'' = \boldsymbol{s}' \circ \langle T(\boldsymbol{s}', \sigma^{(\prime)}_{t+2}) \rangle$ corresponds to an assignment $\beta'' = \beta_{\boldsymbol{s}''}$ that is like β' but assigns to v_{t+2} either *true* or *false*. In any case this assignment β'' makes $\exists v_{t+3} \ldots \forall v_n : F_{\beta''}$ true. This proves Item 3 of the claim.

Item 2 is also satisfied because T has first given an example from $\{(x_{t+1}, 1), (x'_{t+1}, 1)\}$ and then one from $\{(x_{t+2}, 1), (x'_{t+2}, 1)\}$. As for Item 1, giving $(x_{t+2}, 1)$ to a learner in σ_{t+2} causes σ_{t+2} and d_{t+2} to have one error (with respect to x_{t+2}) whereas c_{t+3} has no errors. Hence, the learner switches to c_{t+3}. Symmetrically, a learner in σ'_{t+2} switches to c_{t+3} after receiving $(x'_{t+2}, 1)$. Therefore the hypothesis of $\mathcal{L}$ at the next round, $t+2$, is c_{t+3}. This shows Item 1 and finishes the induction step. □ Claim 1

By Claim 1 it follows that at round n the learner hypothesizes c_{n+1} and that its memory corresponds to an assignment that makes the formula $F = K_1 \wedge \ldots \wedge K_r$ true. In this round, the teacher gives the example $(y, 1)$ causing c_{n+1} to have one error whereas the target c^* has no errors. It remains to show that all hypotheses $a_1, \ldots, a_r$ have at least one error with respect to the memory of $\mathcal{L}$.

Let $i \in \{1, \ldots, r\}$, and let the memory of the learner at round n be $\boldsymbol{s}$. Then the memory at round $n+1$ is $\boldsymbol{s} \circ \langle (y, 1) \rangle$. Since $\beta_{\boldsymbol{s}}$ satisfies $K_i = v_{i_1}^{\alpha_{i_1}} \vee v_{i_2}^{\alpha_{i_2}} \vee v_{i_3}^{\alpha_{i_3}}$, it satisfies at least one literal, say $v_{i_1}^{\alpha_{i_1}}$. If this is a positive literal ($\alpha_{i_1} = 1$) then $\beta_{\boldsymbol{s}}$ satisfies v_{i_1}

and thus the example $(x_{i_1}, 1)$ is in $\boldsymbol{s}$. On the other hand, by definition of a_i, $x_{i_1} \notin a_i$ and therefore a_i has one error with respect to $\boldsymbol{s}$. If the literal is negative, the argument works symmetrically. As i was arbitrary, we have shown that the memory of the learner causes all hypotheses $a_1, \dots a_r$ to have an error. Thus the learner has no choice but to switch to the zero-error hypothesis c^*. This means, the teacher T finishes teaching successfully after $n+1$ rounds.

For the converse, we now assume that there is a teacher T successfully teaching c^* to $\mathcal{L}$ after $n+1$ rounds. We have to show that $\exists v_1 \forall v_2 \dots \exists v_{n-1} \forall v_n : F$ is true.

Since the distance from the initial hypothesis to the target is $n+1$, the learner, when taught by T, must switch its hypothesis in every round. In other words, in round $t = 0, \dots, n$ the learner must hypothesize c_{t+1} (if t even) or σ_{t+1} or σ'_{t+1} (if t odd). Therefore, in round t the learner memorizes exactly one example from $\{(x_i, 1), (x'_i, 1)\}$ for all $i = 1, \dots, t$. In particular, in round n the memory corresponds to a complete variable assignment $V \to \{\mathit{true}, \mathit{false}\}$.

Claim 2: For all even rounds $t \in \{n, n-2, \dots, 2, 0\}$: the learner $\mathcal{L}$ memorizes $\boldsymbol{s}$ such that the formula $\exists v_{t+1} \dots \forall v_n : F_{\beta_{\boldsymbol{s}}}$ is true.

Proof. We begin the inductive proof at $t = n$. At this round the learner hypothesizes c_{n+1}. As a successful teacher, T has to give the example $(y, 1)$ now. With the same reasoning as in the proof of Claim 1, the learner switches to c^* if and only if its memory corresponds to an assignment that satisfies $F = K_1 \wedge \dots \wedge K_r$. Thus the claim holds for $t = n$.

Now assume the claim for some even $t \in \{n, n-2, \dots, 2\}$. We show it for round $t-2$. Let $\mathcal{L}$ memorize $\boldsymbol{s}$. In round $t-2$, $\mathcal{L}$ hypothesizes c_{t-1}. The teacher must give an example from $\{(x_{t-1}, 1), (x'_{t-1}, 1)\}$, otherwise $\mathcal{L}$ would either not change the hypothesis or switch to the dead-end d_{t-1}. This adds one example to $\mathcal{L}$'s memory and thereby assigns a value $b_{t-1} \in \{\mathit{true}, \mathit{false}\}$ to the variable v_{t-1}. Now one computation of the non-deterministic learner $\mathcal{L}$ switches to σ_t, another switches to σ'_t. In the first case, the teacher must give the example $(x_t, 1)$ and in the second case the example $(x'_t, 1)$. In both cases $\mathcal{L}$ moves on to c_{t+1}, but there are two possible memories: $\boldsymbol{s}_2 = \boldsymbol{s} \circ \langle T(\boldsymbol{s}, c_{t-1}), (x_t, 1)\rangle$ and $\boldsymbol{s}'_2 = \boldsymbol{s} \circ \langle T(\boldsymbol{s}, c_{t-1}), (x'_t, 1)\rangle$. The respective assignments are $\beta_2 = \beta_{\boldsymbol{s}} \cup \{v_{t-1} \mapsto b_{t-1}, v_t \mapsto \mathit{true}\}$ and $\beta'_2 = \beta_{\boldsymbol{s}} \cup \{v_{t-1} \mapsto b_{t-1}, v_t \mapsto \mathit{false}\}$. By the induction hypothesis, the learner's memory at round t is such that $\exists v_{t+1} \dots \forall v_n : F_{\beta_2}$ and $\exists v_{t+1} \dots \forall v_n : F_{\beta'_2}$ are true. In other words, there is an assignment to v_{t-1}, namely b_{t-1}, such that for all assignments to v_t the formula $\exists v_{t+1} \dots \forall v_n : F_{\beta_{\boldsymbol{s}}}$ is true. But this means $\exists v_{t-1}\ \forall v_t\ \exists v_{t+1} \dots \forall v_n : F_{\beta_{\boldsymbol{s}}}$ is true, which proves the induction step. □ Claim 2

For round $t = 0$, Claim 2 says that the learner's memory $\boldsymbol{s}$ is such that $\exists v_1 \dots \forall v_n : F_{\beta_{\boldsymbol{s}}}$ is true. But as the initial memory is empty, the assignment $\beta_{\boldsymbol{s}}$ actually does not assign a value to any variable. This means $\exists v_1 \dots \forall v_n : F$ is true. This completes the proof of the reduction and shows that ∞-FB-TEACHINGTIME is $\mathcal{PSPACE}$-hard. □

The ∞-FB-TEACHINGTIME problem can be solved by the algorithm we have presented for m-FB-TEACHINGTIME by setting $m = \ell$. Therefore ∞-FB-TEACHINGTIME can be solved in double exponential time. For the unary version of the problem we can specify the computational complexity more precisely. In the next theorem we show it to be $\mathcal{PSPACE}$-complete. This also implies that ∞-FB-TEACHINGTIME is solvable in exponential space, a slight improvement over double exponential time.

Theorem 5.21 *The problem* UNARY-∞-FB-TEACHINGTIME *is $\mathcal{PSPACE}$-complete.*

Proof. The $\mathcal{PSPACE}$-hardness can be shown via essentially the same reduction as in the proof of Theorem 5.20. The reduction is still polynomial time even if the number $\ell = n + 1$ has to be encoded unary.

The size of an instance of UNARY-∞-FB-TEACHINGTIME is in $\Theta(|\varrho(\mathcal{C})|^2 \cdot |X| + \ell)$, where the first summand is for the restriction graph and the second for the unary encoding of the number of rounds. To show UNARY-∞-FB-TEACHINGTIME $\in \mathcal{PSPACE}$ we present a non-deterministic algorithm that uses polynomial space. The idea of the algorithm is to guess a teacher and to check whether it is successful after ℓ rounds. Since such a teacher has to be defined for $O(|X|^\ell \cdot |\varrho(\mathcal{C})|)$ states, storing a teacher completely would exceed the space bound. To circumvent that problem, our algorithm only teaches one learner at a time for ℓ rounds and runs through all learners in a backtracking manner. Teaching only one learner for ℓ rounds requires to store only $O(\ell)$ examples, which is within the polynomial space bound. More formally, the algorithm is as follows.

Input: Restriction graph $(\varrho(\mathcal{C}) \cup \{\mathit{init}\}, \rhd)$, target $c^* \in \mathcal{C}$, string 1^ℓ.

1 **if** *teachable*$(\langle\,\rangle, \mathit{init}) = 1$ **then accept**

2 **else reject**

The function *teachable*() takes as arguments a list $\boldsymbol{s}$ of examples and a hypothesis σ:

function *teachable*$(\boldsymbol{s}, \sigma)$:

1 **if** $|\boldsymbol{s}| = \ell$ **and** $\sigma \in \varrho(c^*)$ **then return** 1

2 **if** $|\boldsymbol{s}| = \ell$ **and** $\sigma \notin \varrho(c^*)$ **then return** 0

3 **guess** $z \in \mathcal{X}(c^*)$

4 $\mathit{hyps} := \mathcal{L}_{\mathit{hyp}}(\boldsymbol{s}, \sigma, z)$

5 **if** $\forall \zeta \in \mathit{hyps} : \mathit{teachable}(\boldsymbol{s} \circ \langle z \rangle, \zeta) = 1$ **then return** 1

6 **else return** 0

The function *teachable*() calls itself recursively. At any given time the call stack contains at most ℓ instantiations of the function because the length of $\boldsymbol{s}$ increases by one with every recursive call and the recursion ends as soon as $|\boldsymbol{s}| = \ell$ (see lines 1 and

2). Every instantiation of *teachable*() stores $\boldsymbol{s}, \sigma, z$, and *hyps*. The space requirements for σ and z are $O(\log|\varrho(\mathcal{C})|)$ and $O(\log|X|)$; the list $\boldsymbol{s}$ contains at most ℓ examples, which overall requires $O(\ell{\cdot}\log|X|)$ bits; the set *hyps* contains at most $|\varrho(\mathcal{C})|$ hypotheses, which amounts to $O(|\varrho(\mathcal{C})|\cdot\log|\varrho(\mathcal{C})|)$ bits. Storing all this for at most ℓ instantiations requires space polynomial in the input size because ℓ is represented unary in the input.

For the correctness we have to show that there is a teacher successful after ℓ rounds if and only if there is an accepting computation of the algorithm. First, suppose there is a teacher $T\colon \mathcal{X}^*\times(\varrho(\mathcal{C})\cup\{\mathit{init}\})\to\mathcal{X}(c^*)$ such that the non-deterministic $\rhd$-restricted learner, when taught by T, hypothesizes c^* in round ℓ. Now consider the computation of the algorithm in which the guesses always conform to the teacher, that is, $z=T(\boldsymbol{s},\sigma)$ in all calls of *teachable*$(\boldsymbol{s},\sigma)$. One can show by induction over the calls of *teachable*() that *teachable*$(\boldsymbol{s},\sigma)$ is called with parameters $\boldsymbol{s},\sigma$ if and only if the state $(\boldsymbol{s},\sigma)$ can be reached by a learner under teacher T. The main reason for this is that in line 5, the function *teachable*() is called for all possible follow-up hypotheses ζ and the follow-up memory contents $\boldsymbol{s}\circ\langle z\rangle$.

By assumption about T, as soon as $|\boldsymbol{s}|=\ell$ all learners hypothesize c^*. Therefore, on the ℓ-th level of the recursion *teachable*() always returns 1. Then on the $(\ell-1)$-st level the condition in line 5 is satisfied and 1 is returned. By induction on the recursion level one can show that *teachable*$(\langle\,\rangle,\mathit{init})$ also returns 1. This means that we have an accepting computation of the algorithm.

Now assume that the algorithm has an accepting computation. This computation yields guesses for "many" states $(\boldsymbol{s},\sigma)$. However, there can be several different guesses for the same state since *teachable*() could have been called several times with the same parameters. Thus we cannot extract a unique teacher $T\colon \mathcal{X}^*\times(\varrho(\mathcal{C})\cup\{\mathit{init}\})\to\mathcal{X}(c^*)$ from the guess.

But we can extract a history-aware teacher from the guesses. Two calls of the function *teachable*$(\boldsymbol{s},\sigma)$ with the same parameters differ in the call stack, that is, in the sequence of recursive calls that have led to *teachable*$(\boldsymbol{s},\sigma)$. A call stack is described by a sequence $(\boldsymbol{s}_0,\sigma_0),(\boldsymbol{s}_1,\sigma_1),\ldots,(\boldsymbol{s}_t,\sigma_t)$ with $t\in\{0,\ldots,\ell\}$, $(\boldsymbol{s}_0,\sigma_0)=(\langle\,\rangle,\mathit{init})$, and $(\boldsymbol{s}_t,\sigma_t)=(\boldsymbol{s},\sigma)$. The example z that is guessed in the call *teachable*$(\boldsymbol{s}_i,\sigma_i)$ for $i\in\{0,\ldots,t-1\}$ can be reconstructed from the call stack because it is just the last element of $\boldsymbol{s}_{i+1}$. We denote this example by z_i. We define a history-aware teacher

$$T\colon(\mathcal{X}^*\times(\varrho(\mathcal{C})\cup\{\mathit{init}\}))\times(\mathcal{X}^*\times(\varrho(\mathcal{C})\cup\{\mathit{init}\})\times\mathcal{X})^*\to\mathcal{X}(c^*)$$

as follows. Let *teachable*$(\boldsymbol{s},\sigma)$ be a call occurring in the accepting computation and let its call stack be $(\boldsymbol{s}_0,\sigma_0),(\boldsymbol{s}_1,\sigma_1),\ldots,(\boldsymbol{s}_t,\sigma_t)$ with $t\in\{0,\ldots,\ell\}$, $(\boldsymbol{s}_0,\sigma_0)=(\langle\,\rangle,\mathit{init})$, and $(\boldsymbol{s}_t,\sigma_t)=(\boldsymbol{s},\sigma)$. The corresponding history is then

$$\mathit{hist}=\langle\,(\boldsymbol{s}_0,\sigma_0,z_0),(\boldsymbol{s}_1,\sigma_1,z_1),\ldots,(\boldsymbol{s}_{t-1},\sigma_{t-1},z_{t-1})\,\rangle$$

and we define $T(\boldsymbol{s},\sigma,\mathit{hist})=z_{t+1}$.

By induction on the recursive calls one can show that a call to *teachable*$(\boldsymbol{s}, \sigma)$ with a certain call stack occurs if and only if there is a learner that reaches $(\boldsymbol{s}, \sigma)$ with a history that corresponds to that call stack. Again the reason is that the recursive calls in line 5 simulate exactly the possible behavior of all learners.

As we consider an accepting computation of the algorithm, all calls to *teachable*$(\boldsymbol{s}, \sigma)$ with $|\boldsymbol{s}| = \ell$ return 1. Therefore all learners must hypothesize c^* as soon as their memory contains ℓ examples, that is, in round ℓ. Consequently, T is a teacher that is successful after ℓ rounds.

Finally according to Lemma 2.18, there is also a teacher without history that teaches within the same number of rounds as T. □

Corollary 5.22 *The problem* FB-TEACHINGTIME *is fixed-parameter tractable (with parameter m) and $\mathcal{PSPACE}$-hard.*

Proof. The fixed-parameter tractability follows from Theorem 5.18.

The problem is $\mathcal{PSPACE}$-hard because we can reduce ∞-FB-TEACHINGTIME to it. Given an instance for the latter problem involving some ℓ, we construct an instance for FB-TEACHINGTIME by setting $m = \ell$ and leaving the rest unchanged. Now the result follows because during the first ℓ rounds, an ℓ-memory learner and an ∞-memory learner behave the same. □

In the teaching dimension model it is easy to decide whether a given target can be taught to a learner with 1-memory. This is equivalent to deciding whether the teaching dimension of the target concept is one, which can be done by checking every singleton sample for already being a teaching set. In the $\rhd$-restricted model the analog question in the presence of feedback is just the 1-FB-TEACHINGTIME problem and thus remains easy to decide. In contrast, the same problem in the absence of feedback is $\mathcal{NP}$-hard.

Theorem 5.23 *The problem* 1-NOFB-TEACHINGTIME *is $\mathcal{NP}$-hard and in $\mathcal{PSPACE}$.*

Proof. For the $\mathcal{NP}$-hardness we present a reduction from 3-SAT. Let $K_1 \wedge K_2 \wedge \ldots \wedge K_r$ be a formula in 3-CNF over variables $v_1, \ldots, v_n$. The literals in K_i are denoted $v_{i_j}^{\alpha_{i_j}}$ for $j = 1, 2, 3$. Then the instance of 1-NOFB-TEACHINGTIME looks as follows (see also Figure 5.4). The learning domain is $X = \{x_1, x'_1, \ldots, x_n, x'_n\} \cup \{y\}$. There are $r + 1$ representations $\sigma_0, \ldots, \sigma_r$ for the concept $\emptyset$. All other concepts have only one representation. These concepts are

- $c_i = \{x_i\}$, $c'_i = \{x'_i\}$ for $i = 1, \ldots, n$;
- $a_i = \{x_k \mid v_k \text{ is a literal in } K_i\} \cup \{x'_k \mid \bar{v}_k \text{ is a literal in } K_i\}$ for $i = 1, \ldots, r$;
- $d_i = \{x_1, x'_1, \ldots, x_{i-1}, x'_{i-1}, x_{i+1}, x'_{i+1}, \ldots, x_n, x'_n\}$ for $i = 0, \ldots, n$;
- the target concept $c^* = X$.

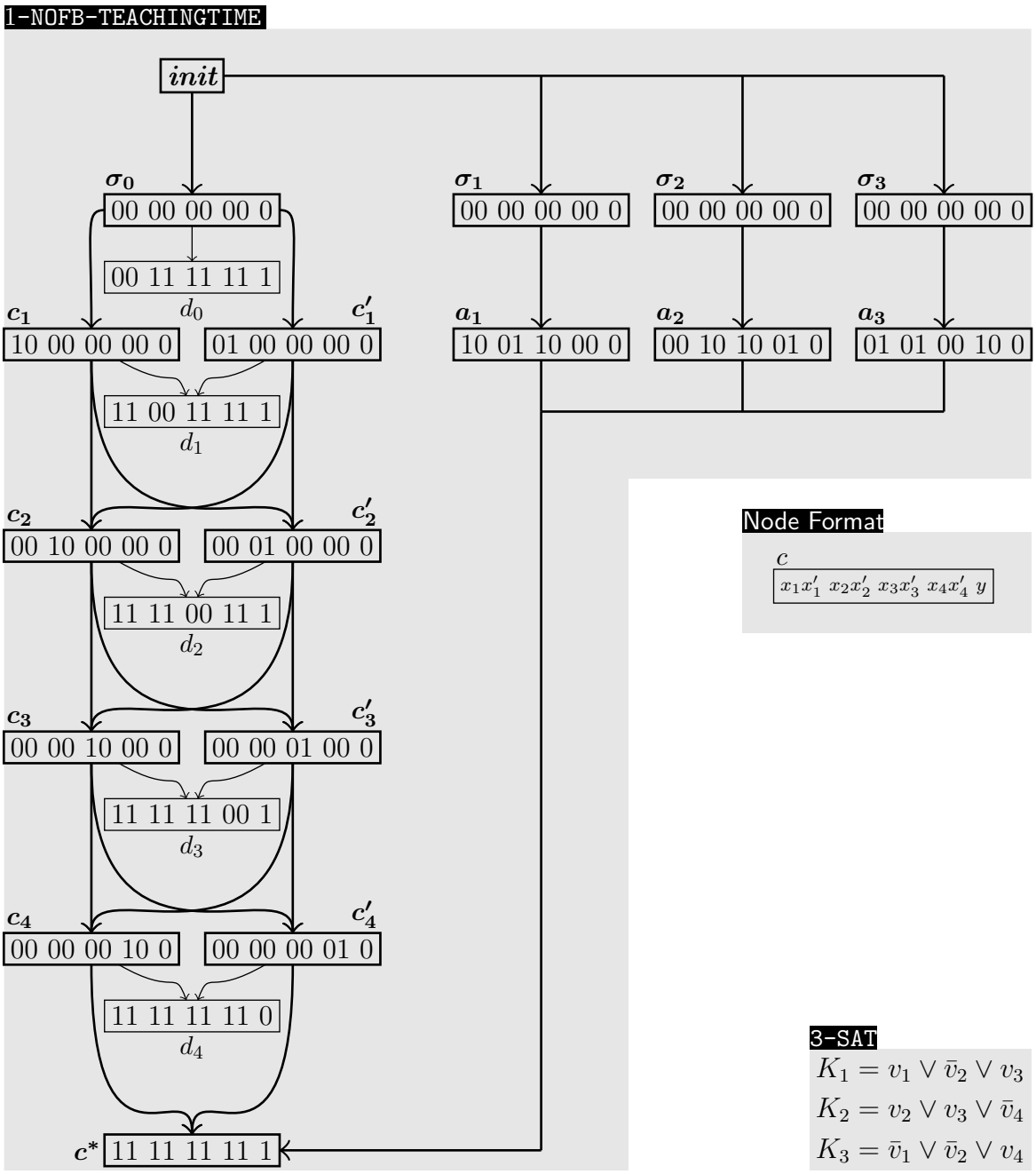

Figure 5.4 *The reduction from* 3-SAT *to* 1-NOFB-TEACHINGTIME *applied to the formula* $(v_1 \vee \bar{v}_2 \vee v_3) \wedge (v_2 \vee v_3 \vee \bar{v}_4) \wedge (\bar{v}_1 \vee \bar{v}_2 \vee v_4)$ *yields this* $\triangleright$*-restriction graph. The nodes contain concepts over* $X = \{x_1, x_1', x_2, x_2', x_3, x_3', x_4, x_4', y\}$*. See the proof of Theorem 5.23 for details.*

There are arcs from *init* to $\sigma_0, \ldots, \sigma_r$ and from σ_i to a_i for all $i \in \{1, \ldots, r\}$. Moreover we have arcs from c_i and c_i' to d_i for $i = 1, \ldots, n$ and from σ_0 to d_0. Finally the concepts a_i, c_n, and c_n' have an arc to the target c^*. The number ℓ of rounds is set to $n + 2$.

The concepts and arcs can be computed easily from the clauses $K_1, \ldots, K_r$. There are $3n + 2r + 4$ nodes in the restriction graph, each of which is labeled with $|X| = n + 1$ bits. The instance of 1-NOFB-TEACHINGTIME can thus be computed in polynomial time.

Let $\mathcal{L}$ be the non-deterministic $\rhd$-restricted learner with 1-memory, where $\rhd$ is defined by the restriction graph described above.

Intuitively, the 1-NOFB-TEACHINGTIME instance produced by the reduction works as follows. The examples $(x_i, 1), (x_i', 1)$ correspond to the assignments *true* or *false* to the variable v_i. After the first example, one computation of the non-deterministic learner move to each hypothesis $\sigma_0, \ldots, \sigma_r$. The computation moving to σ_0 forces the teacher to give an example from $\{(x_1, 1), (x_1', 1)\}$ in the next round, then one from $\{(x_2, 1), (x_2', 1)\}$ and so on. This eventually defines a complete assignment to all variables $v_1, \ldots, v_n$. The computations of $\mathcal{L}$ in σ_i for $i = 1, \ldots, r$ can only switch to a_i if they get an example that "matches" a literal in K_i. If that never happens, the computation remains in σ_i and does not reach the target, which causes the teacher to fail. If, on the other hand, all computations of $\mathcal{L}$ in all σ_i's switch to a_i they will then reach the target when the teacher finally gives the example $(y, 1)$.

Now we prove the above intuition more formally. Let $K_1 \wedge \ldots \wedge K_r$ be satisfied by an assignment $\beta\colon \{v_1, \ldots, v_n\} \to \{\mathit{true}, \mathit{false}\}$. We define a feedbackless teacher $T\colon \mathbb{N} \to \mathcal{X}(c^*)$ by $T(0) = (x_1, 1)$, $T(n + 1) = (y, 1)$, and

$$T(i) = \begin{cases} (x_1, 1) & \text{if } \beta(v_i) = \mathit{true}, \\ (x_1', 1) & \text{if } \beta(v_i) = \mathit{false} \end{cases}$$

for $i = 1, \ldots, n$. The values $T(i)$ for $i > n + 1$ do not matter. We have to show that T is successful and needs at most $n + 2$ rounds.

Claim 1: At round $n + 2$ the learner $\mathcal{L}$ hypothesizes c^*.

Proof. We have to distinguish two kinds of computations of $\mathcal{L}$, depending on the behavior after the first example $T(0)$. We show that the computations of both kinds reach c^* at the latest in round $n+2$. Recall that $\mathcal{L}$ has 1-memory and thus the follow-up hypothesis is determined only by the current hypothesis and the example just received.

Case 1: $\mathcal{L}$ hypothesizes σ_0 in round 1.

Then $T(1)$ causes one error among the instances x_1, x_1'. The neighbor hypothesis d_0 has the same error. The same is true for one of the hypotheses c_1, c_1'. The other hypothesis, however, has no error and thus $\mathcal{L}$ moves to this one. Similarly, receiving $T(2)$ causes an error at the new hypothesis and the only error-free neighbor is c_2 or

c'_2 depending on whether $T(2) = (x_2, 1)$ or $T(2) = (x'_2, 1)$. Applying this argument n times, one can see that $\mathcal{L}$ hypothesizes c_n or c'_n at round $n+1$. The next example then is $(y, 1)$ and the only consistent neighbor is the target. Therefore, at round $n+2$, the learner $\mathcal{L}$ hypothesizes c^*.

Case 2: $\mathcal{L}$ hypothesizes σ_i in round 1 for some $i \in \{1, \ldots, r\}$.

Since β satisfies K_i there is a minimum $i_j \in \{1, \ldots, n\}$ such that β satisfies the literal $v_{i_j}^{\alpha_{i_j}}$ in K_i. Until round $i_j - 1$ all examples given by the teacher cause one error at σ_i and all neighbor hypotheses have the same error. Thus no hypothesis change takes place. When the example $T(i_j)$ arrives, it causes an error, but now the neighbor a_i is error-free. Consequently, $\mathcal{L}$ moves to a_i. The hypothesis a_i has only one neighbor, the target. A hypothesis change to the target is triggered at the latest by the example $T(n+1) = (y, 1)$ and therefore $\mathcal{L}$ is in c^* at round $n+2$. □ Claim 1

For the converse, assume that T is a teacher successful within $n+2$ rounds. We have to show that $K_1 \wedge \ldots \wedge K_r$ is satisfiable. The first step is to show that in rounds $t = 1, \ldots, n$ the teacher must give an example from $\{(x_t, 1), (x'_t, 1)\}$.

After example $T(0)$ one computation of $\mathcal{L}$ is in σ_0 since all hypotheses $\sigma_0, \ldots, \sigma_r$ are equivalent and the *init* hypothesis must be left. To reach the target within the next $n+1$ rounds, $\mathcal{L}$ must hypothesize c_t or c'_t at round $t+1$ for all $t = 1, \ldots, n$. From c_t or c'_t the neighbor c_{t+1} is only entered on example $(x_{t+1}, 1)$; all other examples are either consistent with c_t or c'_t, or with the dead end d_t. But an example consistent with d_t would cause $\mathcal{L}$ to switch to d_t and teaching would fail. Similarly, c'_{t+1} is only entered on example $(x'_{t+1}, 1)$. Therefore the teacher T must give an example from $\{(x_{t+1}, 1), (x'_{t+1}, 1)\}$ at round $t = 1, \ldots, n$. The examples $T(1), \ldots, T(n)$ then correspond to the following assignment β to all variables $v_1, \ldots, v_n$:

$$\beta(v_i) = \begin{cases} \mathit{true} & \text{if } T(i) = (x_1, 1), \\ \mathit{false} & \text{if } T(i) = (x'_1, 1). \end{cases}$$

We say the example $(x_i, 1)$ *matches* the literal v_i and the example $(x'_i, 1)$ *matches* $\bar{v}_i$.

It remains to show that the assignment β satisfies K_i for all $i = 1, \ldots, r$. Let $i \in \{1, \ldots, r\}$. After $T(0)$ there is a computation of $\mathcal{L}$ hypothesizing σ_i. This computation will eventually reach the target (by assumption about T). The only path is via the hypothesis a_i. By construction of a_i a transition to it from σ_i is only possible on an example that matches a literal in K_i. Therefore at least one example that matches one of the literals of K_i will be given by the teacher. That example makes β satisfy this literal and thus the whole clause. Since i was arbitrary we have just shown that β satisfies all clauses.

To show 1-NOFB-TEACHINGTIME $\in \mathcal{PSPACE}$ we present a non-deterministic algorithm that uses polynomial space. The idea of the algorithm is to guess ℓ examples and to keep an account of all hypotheses the learner could be in after every guess.

Input: Restriction graph $(\varrho(\mathcal{C}) \cup \{init\}, \rhd)$, target $c^* \in \mathcal{C}$, number $\ell \in \mathbb{N}$.

1 $hyps := \{init\}$

2 **for** $t = 0$ **to** $\ell - 1$:

3 $\quad newhyps := \emptyset$

4 $\quad$ **guess** $z \in \mathcal{X}(c^*)$

5 $\quad$ **for all** $\sigma \in hyps$:

6 $\quad\quad newhyps := newhyps \cup \mathcal{L}(\langle\,\rangle, \sigma, z)$

7 $\quad hyps := newhyps$

8 **if** $hyps \subseteq \varrho(c^*)$ **then accept**

9 $\quad$ **else reject**

Throughout the execution of the algorithm, both *hyps* and *newhyps* can contain at most $|\varrho(\mathcal{C})|$ elements, each of which can be written in $\log|\varrho(\mathcal{C})|$ bits. The space requirement of the algorithm is thus polynomial.

Whenever the algorithm accepts the input, there is a sequence of ℓ guessed examples such that all computations of $\mathcal{L}$, after having receiving all these examples, hypothesize c^*. Thus there is a teacher successful after ℓ rounds.

On the other hand, whenever there is a teacher successful after ℓ rounds there is also an accepting computation of the algorithm. Guessing $z = T(t)$ in all rounds $t = 0, \ldots, \ell - 1$ is such a computation. □

For the unary variant we can again determine the computational complexity more precisely.

Theorem 5.24 *The problem* UNARY-1-NOFB-TEACHINGTIME *is $\mathcal{NP}$-complete.*

Proof. The $\mathcal{NP}$-hardness can be shown using the same reduction as in the proof of Theorem 5.23, except that $\ell = n + 2$ is encoded unary.

To show containment in $\mathcal{NP}$ we use the non-deterministic algorithm from the proof of Theorem 5.23. This algorithm runs for ℓ iterations of the outer loop. The inner loop has $|hyps| \leq |\varrho(\mathcal{C})|$ iterations and there are at most $|\varrho(\mathcal{C})|$ hypotheses to add to *newhyps*. Thus the inner loop takes time $O(|\varrho(\mathcal{C})|^2)$. So overall the running time is $O(\ell \cdot |\varrho(\mathcal{C})|^2)$, which is polynomial in the input size since ℓ is represented unary. □

For the complexity of ∞-NOFB-TEACHINGTIME we confine ourselves to a simple observation.

Fact 5.25 *The problem ∞-*NOFB-TEACHINGTIME *is $\mathcal{NP}$-hard.*

Proof. The problem of finding an optimal teacher without feedback for $\rhd$-restricted learners is $\mathcal{NP}$-hard, since it is a generalization of finding an optimal teaching set, namely if $\rhd = (\Sigma^* \cup \{init\}) \times (\Sigma^* \cup \{init\})$ (see [76, 34, 8]). □

Broadly speaking, the results in this section indicate that computing the optimal teaching time in the $\rhd$-restricted model is more difficult than in the TD model. Only in the case of teaching 1-memory learners with feedback, both problems are provably solvable in polynomial time. In the case of teaching ∞-memory learners with feedback we have $\mathcal{NP}$-completeness vs. $\mathcal{PSPACE}$-completeness. For teaching 1-memory learners without feedback we have containment in $\mathcal{P}$ vs. $\mathcal{NP}$-completeness.

The decision problems we studied in this section are decision versions of the optimization problem of finding the minimum teaching time over all successful teachers. Another important question, which we have neglected so far, is the teachability problem. This is the problem whether or not a successful teacher exists at all. In the $\mathfrak{H}$-restricted model this teachability problem can be decided by comparing the $\mathfrak{H}$-dimension of the target with the memory size of the learner (see Lemmas 3.3 and 4.29).

In the $\rhd$-restricted model, the teachability can sometimes be decided by algorithms we introduced in this section. In the case of teaching 1-memory learners with feedback Theorem 5.18 also yields a decision algorithm for teachability. In the case of teaching 1-memory learners without feedback a similar algorithm can be devised. The hypotheses have to be replaced by "meta hypotheses" in $2^{\varrho(\mathcal{C})}$. Each meta hypothesis describes the set of possible hypotheses the learner can assume. Target meta hypotheses are then all subsets of $\varrho(c^*)$. The running time of the algorithm would then be polynomial in the number of meta states $2^{\varrho(\mathcal{C})} \cdot |X|$, that is, exponential in the representation size.

Learners with infinite memory can assume infinitely many states. The algorithm in Theorem 5.21 cannot be used to decide the teachability in this situation. Moreover, the $\mathcal{PSPACE}$-hardness result for ∞-`FB-TEACHINGTIME` does not imply $\mathcal{PSPACE}$-hardness of the teachability problem. The reason is that for the ∞-`FB-TEACHINGTIME` instances generated by the reduction (see, for example, Figure 5.3) the teachability of the target can be decided easily: The target is always teachable. A teacher can make the learner reach c_{n+1} and then, by giving all examples $(x_1, 1), (x'_1, 1), \ldots, (x_n, 1), (x'_n, 1), (y, 1)$, force it into the target state.

We thus pose as an open problem to determine the computational complexity of the teachability problem for the ∞-memory learner (with or without feedback).

5.5 Discussion

The $\rhd$-restricted model offers the possibility for all kinds of realistic effects. Regarding feedback, we observed that it can be useless, helpful, or even indispensable for teaching. In addition, natural infinite concept classes can be taught in this model, as well as finite concept classes. The order of examples is also crucial, and the learner's memory size

influences the teaching time and also whether or not a concept is teachable at all. These effects can be achieved thanks to the great freedom in defining the relation $\rhd$.

On the other hand, that freedom can also be a burden because it allows us to define the neighborhood relation in unnatural ways. We have used, or rather abused, the freedom for defining the relation $\rhd$ in our complexity results in Section 5.4. For example, having dead-end hypotheses available is very helpful, but not exactly realistic. It would be interesting to impose some natural restrictions on the neighborhood relation so as to prevent too artificial and abusive definitions. However, it is not quite clear how such natural restrictions look like. One idea for an additional condition is that the $\rhd$-restriction graph should consist of only one strongly connected component, which would at least prevent dead-end hypotheses. Another way to prevent such hypotheses is to require the neighborhood relation to be symmetric. An entirely different approach is to stipulate some relation between the syntactic closeness and the semantic closeness of hypotheses. For example, one could require that hypotheses that are close in the $\rhd$-restriction graph are also semantically close, that is, equal with respect to most instances, or vice versa.

To obtain meaningful results in the $\rhd$-restricted model, the neighborhood relation has to be defined in a natural way. This definition has to be done individually for every natural concept class we want to investigate. Even though it is impossible to give a general "recipe" how to do that, our definitions of $\rhd_{Mon}$ and $\rhd_{DL}$ for monomials and 1-decision lists suggest a usable, general approach, which is based on the representation function. Using this function one can define two hypotheses as neighbors of each other if the two representations differ only a little. Since representations are strings, the difference between two of them can naturally be measured by the Hamming distance or the edit distance. For instance, our relation $\rhd_{Mon}$ is defined via the Hamming distance, and $\rhd_{DL}$ via the edit distance; in both cases two hypotheses are neighbors of each other if and only if their distance is 1. This approach to defining $\rhd$ can be used for all represented concept classes, but still leaves some details to be specified, such as the maximal allowed distance.

Besides adding further restrictions on the neighborhood relation $\rhd$ there are other natural ways the learner can be modified. One idea is to combine the $\rhd$-restriction with selective memory (see Definition 4.28). Another idea is to remove the conservativeness requirement and to allow the learner to switch to a hypothesis with less errors than the current one, even if the newly received example is consistent. But also these variations on the learner do not absolve us from defining the neighborhood relation appropriately. Therefore, in our opinion, future work should concentrate on the definition of $\rhd$ rather than on new model variants.

Part II

Teaching Models with Randomized Learner

Chapter 6

The Randomized Framework

6.1 Introduction

In Part I we were able to enhance the basic teaching dimension model to get feedback and memory effects. These enhancements, however, mostly came at the price of arbitrary, hard-to-justify additional assumptions, such as hypothesis restrictions $\mathfrak{H}$ or neighborhood relations $\rhd$. In contrast, Part II is devoted to a new, randomized teaching model, which uses no additional assumptions, but which does give us memory and feedback effects. It is introduced in this chapter.

The new model remedies a fundamental flaw in the teaching dimension model, namely that it measures teachability always with respect to the worst learner. The teaching dimension model lacks feedback effects because there is always a bad learner that chooses follow-up hypotheses independently from the current one. The teaching dimension model lacks memory size effects because there is always a bad learner that avoids the target hypothesis if its memory is smaller than the target's teaching dimension. Until now we always tried to *prohibit* such bad behavior, for example, by forcing the follow-up hypothesis to depend on the current one or by disallowing implausible hypotheses. In our new approach we accept that there are good learners and bad learners and again *allow* all of them, as long as they are consistent and conservative.

Rather than changing the learners that are taught, we change the measurement of how well they are taught. Instead of measuring the number of rounds needed by the worst learner, we measure the number of rounds needed by the average learner. At first, however, it may seem unclear what such an average learner is supposed to be. Of course it would be useless to pick out one of the deterministic learners and call it "average." Rather, we have to devise a new learner that somehow combines all learners into one. Perhaps the most straightforward way to do so is to take a randomized learner for the task. Whenever there are several alternatives from which the non-deterministic learner in the basic model can choose, the randomized learner would choose each alternative with the same probability. In this way all possible behaviors, good ones and bad ones, are incorporated with equal rights. The performance of a teacher interacting with such a randomized learner is then measured by the *expected* teaching time.

In more formal terms, the new approach causes only small changes to our teaching framework described in Section 2.2. Essentially, the non-deterministic automaton is replaced by a probabilistic automaton and the teaching time becomes a random variable. Also the notions of teaching success have to be reinterpreted a bit in light of the randomization. But in any case, no additional assumptions like $\mathfrak{H}$ or $\rhd$ are introduced. The formalities can be found in the following Section 6.2.

That the randomized model actually works, can be demonstrated already with quite simple concept classes such as $\mathcal{S}_n$. After a few rounds of teaching the target $[1, n] \in \mathcal{S}_n$, with some probability the current hypothesis of the randomized learner is the target. The remaining probability mass is distributed among the co-singleton hypotheses. Now every example for $[1, n]$ is consistent with the current hypothesis with some probability. Therefore a feedbackless teacher cannot cause a hypothesis change with probability 1 in every round. The teacher with feedback, however, can do so and thus achieves teaching success faster. Likewise, the more examples the learner memorizes, the higher the probability that a hypothesis change leads to the target because there are fewer consistent non-target concepts. Thus the teaching speed increases with growing memory size. In Section 6.3 we give more details and demonstrate how varying the order of examples and the frequency of examples influences the teaching time of monomials.

From a more general point of view, the randomized teaching process can be regarded as a special case of *Markov decision processes*. These probabilistic processes have been subject of research for decades and we shall benefit from some results of their theory in subsequent chapters. Section 6.4 provides a short introduction into the necessary terminology.

6.2 Formal Description

The basic scenario is the same as in the non-deterministic case (see Section 2.2). The teaching process is divided into rounds. In each round the teacher gives the learner an example of a target concept. The learner memorizes this example and computes a new hypothesis based on the last hypothesis and the memorized examples. The target and the hypotheses are taken from a concept class known to both teacher and learner.

The learner is again modeled as an automaton, but with randomized state transitions. Its behavior can be described using specifications for (HYP) and (MEM) just as in the non-deterministic models. Whenever the specifications allow a set of alternatives, the learner is supposed to pick one alternative uniformly at random. We assume that there is always a finite number of alternatives, since otherwise the uniform distribution would not be defined. The teacher's goal consists in making the learner hypothesize the target as quickly as possible. But now we do not measure the worst case time until this happens, rather we measure the expected teaching time of the learner.

As the analysis of randomized learners is much more difficult, we investigate only one specification of (HYP), namely conservative and consistent. We also limit our

attention to non-selective memory. The learner is thus a randomized version of the non-deterministic learner from Definition 3.1. We simplify even further and do not distinguish between representations and concepts any more. We thus use concepts in places where before we only used representations; for example, hypotheses are now from $\mathcal{C}$ rather than from $\varrho(\mathcal{C})$. One can think of this as having an implicit 1-1-mapping from representations to concepts. This measure helps reduce notation and will not cause problems, because throughout Part II we consider only finite classes.

As before, we stipulate a special initial hypothesis called *init*, with which every example is inconsistent. Moreover, the initial memory is empty. The randomized learners are randomized automata whose every state consists of a list $\boldsymbol{s} \in \mathcal{X}^*$ of memorized examples and a hypothesis $h \in \mathcal{C} \cup \{\mathit{init}\}$. The state space is thus $\mathcal{X}^* \times (\mathcal{C} \cup \{\mathit{init}\})$. The learners we consider differ only with respect to their memory size.

Definition 6.1 *Let $\mathcal{C}$ be a concept class over X and let $m \in \mathbb{N}^+ \cup \{\infty\}$. The* randomized m-memory learner using hypothesis space $\mathcal{C}$ *is the randomized automaton that performs state changes according to the following randomized algorithm. It is denoted by $\mathcal{L}_{m,\mathcal{C}}$ (or by $\mathcal{L}_m$ if $\mathcal{C}$ is clear or unimportant).*

Current state: *Memory $\boldsymbol{s} \in \mathcal{X}^*$, hypothesis $h \in \mathcal{C} \cup \{\mathit{init}\}$.*

Input: *Example $z \in \mathcal{X}$.*

Follow-up state: *Memory $\boldsymbol{s}' \in \mathcal{X}^*$, hypothesis $h' \in \mathcal{C}$.*

1 $\boldsymbol{s}' := \boldsymbol{s} \circ_m \langle z \rangle$;

2 **if** $z \notin \mathcal{X}(h)$ **then** *choose h' uniformly at random from $\mathcal{C}(\boldsymbol{s}')$;*

3 **else** $h' := h$;

Definition 6.1 implicitly defines the probabilities $p((\boldsymbol{s}, h), z, (\boldsymbol{s}', h'))$ of a state change from $(\boldsymbol{s}, h)$ to $(\boldsymbol{s}', h')$ on input $z \in \mathcal{X}$:

$$p((\boldsymbol{s}, h), z, (\boldsymbol{s}', h')) = \begin{cases} 1 & \text{if } z \in \mathcal{X}(h) \wedge \boldsymbol{s}' = \boldsymbol{s} \circ_m \langle z \rangle \wedge h' = h, \\ 1/|\mathcal{C}(\boldsymbol{s}')| & \text{if } z \notin \mathcal{X}(h) \wedge \boldsymbol{s}' = \boldsymbol{s} \circ_m \langle z \rangle \wedge h' \in \mathcal{C}(\boldsymbol{s}'), \\ 0 & \text{otherwise.} \end{cases}$$

The teacher needs to be reformalized only slightly. Wherever necessary we replace representations with the represented concept. That means that "$\varrho(\mathcal{C})$" is replaced by "$\mathcal{C}$." In the *presence of feedback* a teacher is thus a function

$$T\colon \mathcal{X}^* \times (\mathcal{C} \cup \{\mathit{init}\}) \to \mathcal{X}(c^*).$$

In the *absence of feedback* the teacher is the same as in the non-deterministic model, namely a function

$$T\colon \mathbb{N} \to \mathcal{X}(c^*).$$

A learner $\mathcal{L}_{m,\mathcal{C}}$ and a teacher determine a teaching process. The state of the process in a round $t \in \mathbb{N}$ is described by the probability distribution over the learner's state space that specifies for each state the probability of the learner being in this state in round t. We denote this probability distribution by

$$\delta_T^{(t)} \colon \mathcal{X}^* \times (\mathcal{C} \cup \{\mathit{init}\}) \to [0,1].$$

The initial distribution is $\delta^{(\mathit{init})}$ with $\delta^{(\mathit{init})}(\langle\,\rangle, \mathit{init}) = 1$, because initially the learner hypothesizes *init* and has an empty memory.

Let us first consider the teaching process involving a teacher T *without feedback* and the learner $\mathcal{L}_{m,\mathcal{C}}$. The probability distributions evolve as follows. Let $\delta_T^{(t)}$ be the distribution in round t and let $z = T(t)$ be the example given in round t. Then for every state $(\boldsymbol{s}, h)$ the definition of $\mathcal{L}_{m,\mathcal{C}}$ implies a distribution over the follow-up states. The distribution $\delta_T^{(t+1)}$ for the next round is then the weighted sum over all the distributions for the single states of the learner. Formally, we have $\delta_T^{(0)} = \delta^{(\mathit{init})}$, and for all $t \geq 0$,

$$\delta_T^{(t+1)}(\boldsymbol{s}', h') = \sum_{\substack{(\boldsymbol{s},h) \\ \in \mathcal{X}^* \times (\mathcal{C} \cup \{\mathit{init}\})}} \delta_T^{(t)}(\boldsymbol{s}, h) \cdot p((\boldsymbol{s}, h), T(t), (\boldsymbol{s}', h')). \tag{6.1}$$

Let us now consider the teaching process involving a teacher T *with feedback* and the learner $\mathcal{L}_{m,\mathcal{C}}$. In this situation we have $\delta_T^{(0)} = \delta^{(\mathit{init})}$, and for all $t \geq 0$,

$$\delta_T^{(t+1)}(\boldsymbol{s}', h') = \sum_{\substack{(\boldsymbol{s},h) \\ \in \mathcal{X}^* \times (\mathcal{C} \cup \{\mathit{init}\})}} \delta_T^{(t)}(\boldsymbol{s}, h) \cdot p((\boldsymbol{s}, h), T(\boldsymbol{s}, h), (\boldsymbol{s}', h')), \tag{6.2}$$

which is similar to Equation (6.1) except that $T(t)$ is replaced by $T(\boldsymbol{s}, h)$.

Since we are mostly interested in the probability for certain hypotheses, as opposed to the memory, we define as shortcut:

$$\delta_T^{(t)}(c) = \sum_{\boldsymbol{s} \in \mathcal{X}^*} \delta_T^{(t)}(\boldsymbol{s}, c).$$

We distinguish two teaching success variants: finite and in the limit. Finite teaching success means that after finitely many rounds the probability of having reached the target is 1. Teaching in the limit means that the probability of reaching the target converges to 1.

Definition 6.2 *Let $\mathcal{C}$ be a concept class, $c^* \in \mathcal{C}$ be a target, and $m \in \mathbb{N}^+ \cup \{\infty\}$. Let T be a teacher and $\left(\delta_T^{(t)}\right)_{t \in \mathbb{N}}$ be the series of probability distributions over states of $\mathcal{L}_{m,\mathcal{C}}$. The* success probability *of T is then*

$$\lim_{t \to \infty} \delta_T^{(t)}(c^*).$$

A teacher is successful *iff its success probability equals 1. A successful teacher is called* finitely successful *iff there is a t such that $\delta^{(t)}(c^*) = 1$, otherwise it is called* successful in the limit. *For a successful teacher we define the* expected teaching time *as*

$$\mathbb{E}[T, \mathcal{L}_{m,\mathcal{C}}, c^*] = \sum_{t \geq 1} t \cdot \left(\delta_T^{(t)}(c^*) - \delta_T^{(t-1)}(c^*) \right).$$

The expected teaching time need not be finite, even if the teacher is successful. The limit $\lim_{t\to\infty} \delta_T^{(t)}(c^*)$ exists for every teacher because $\delta_T^{(t)}(c^*)$ is monotonically increasing due to the learner being conservative. The teachability of a concept is then measured by the minimal expected teaching time over all teachers.

Definition 6.3 *Let $\mathcal{C}$ be a concept class, $c^* \in \mathcal{C}$ and $m \in \mathbb{N}^+ \cup \{init\}$. The* optimal teaching time for teaching c^* with feedback to $\mathcal{L}_{m,\mathcal{C}}$ *is defined as*

$$E_m^+(c^*, \mathcal{C}) = \inf_T \mathbb{E}[T, \mathcal{L}_{m,\mathcal{C}}, c^*]$$

where T ranges over all teachers $T\colon \mathcal{X}^* \times (\mathcal{C} \cup \{init\}) \to \mathcal{X}(c^*)$. The optimal teaching time for teaching c^* without feedback to $\mathcal{L}_{m,\mathcal{C}}$ *is defined as*

$$E_m^-(c^*, \mathcal{C}) = \inf_T \mathbb{E}[T, \mathcal{L}_{m,\mathcal{C}}, c^*]$$

where T ranges over all teachers $T\colon \mathbb{N} \to \mathcal{X}(c^*)$. For a class $\mathcal{C}$ we set $E_m^-(\mathcal{C}) = \max\{E_m^-(c, \mathcal{C}) \mid c \in \mathcal{C}\}$ and $E_m^+(\mathcal{C}) = \max\{E_m^+(c, \mathcal{C}) \mid c \in \mathcal{C}\}$.

If the concept class is clear, we may write $E_m^+(c)$ instead of $E_m^+(c, \mathcal{C})$ and $E_m^-(c)$ for $E_m^-(c, \mathcal{C})$.

For finite concept classes $\mathcal{C}$, the infimum in the definition of E_m^+ can be replaced by the minimum since in this case there are only finitely many teachers with feedback.

We conclude this section with some simple facts about the notions of teachability and the teaching times just defined.

Fact 6.4 *Let $\mathcal{C}$ be a concept class and let $m \in \mathbb{N}^+$. A concept c^* is teachable to $\mathcal{L}_{\mathcal{C},m}$ finitely (with or without feedback) if and only if $TD(c^*, \mathcal{C}) \leq m$.*

Fact 6.5 *For all $\mathcal{C}$, all $m \in \mathbb{N}^+ \cup \{\infty\}$, all $c^* \in \mathcal{C}$, and all $\alpha \in \{+, -\}$:*

1. $E_m^+(c^*, \mathcal{C}) \leq E_m^-(c^*, \mathcal{C})$,
2. $E_\infty^\alpha(c^*, \mathcal{C}) \leq E_{m+1}^\alpha(c^*, \mathcal{C}) \leq E_m^\alpha(c^*, \mathcal{C})$.

Proper inequality holds for the concepts in the class $\mathcal{A}_n$ of all concepts over $[1, n]$.

6.3 The Influence of Feedback, Memory Size, and the Order of Examples

In this section we calculate the teaching times for the concept $[1, n] \in \mathcal{S}_n$ for varying memory size and feedback. Rather than using a provably optimal teacher, we use a "reasonable" teacher whose optimality for the special case $m = 1$ we show later. These results will nevertheless illustrate that feedback and memory size have a somewhat realistic influence on the duration of teaching.

Let us first consider a teacher *with* feedback and the learner $\mathcal{L}_1$ using hypothesis space $\mathcal{S}_n$. Our "reasonable" teacher now gives an example inconsistent with the current hypothesis in every round until $\mathcal{L}_1$ reaches the target $[1, n]$. The probability of reaching the target is $1/n$ in each round. Therefore the expected number of rounds until the target is reached is n. This teacher is optimal because it is basically the only one. Giving an example consistent with the current hypothesis would not change the learner's state and would therefore be useless. Thus $E_1^+([1, n], \mathcal{S}_n) = n$.

Our teacher can easily be generalized to $1 < m \leq n$. A small problem for calculating the expected teaching time is that the learner's memory needs some rounds to fill. More precisely, in the first round the probability of reaching the target is $1/n$, in the second round $1/(n-1)$ and in round $i \leq m$ it is $1/(n-i)$. Beginning with round m the probability remains constant $1/(n-m)$. Thus on entering the m-th round, the expected number of remaining rounds is $n-m$. Putting it all together and simplifying the expression, we get the following formula for the expected number of rounds until $\mathcal{L}_m$ $(1 \leq m \leq n)$ reaches $[1, n]$: $\frac{m(m-1)}{2n} + n - m + 1$. For memory sizes greater then n the teaching time improves no further. Thus we obtain the teaching time for $m = \infty$ by setting $m = n$. Again, also for $m > 1$ this is essentially the only teacher and its teaching time is therefore optimal:

$$E_m^+([1, n], \mathcal{S}_n) = \frac{m(m-1)}{2n} + n - m + 1 \ .$$

Teaching is more difficult without feedback. In this situation the teacher can merely guess examples hoping that they are inconsistent with the current hypothesis. Recall that only in this case a hypothesis change is triggered and the learner gets the chance to reach the target. In particular, giving one example a second time within an interval of m rounds will certainly not trigger a hypothesis change. Therefore, it seems like a good strategy to put a maximum length interval between two occurrences of the same example. This is achieved by the teacher T^- which gives all n examples for $[1, n]$ in the canonical order in an infinite loop, that is, $T^-(i) = (1 + i \bmod n, 1)$ for all $i \in \mathbb{N}$. The analysis of T^- is a little more complicated and we summarize it in the following fact.

Fact 6.6 *Let T^- be a teacher for $c^* = [1, n] \in \mathcal{S}_n$ with $T^-(i) = (1 + i \bmod n, 1)$ for all*

$i \in \mathbb{N}$. *Then*

$$\mathbb{E}[T^-, \mathcal{L}_{1,\mathcal{S}_n}, c^*] = 1 + \frac{n(n-1)}{2}$$

and for $m \in \{2, \ldots, n\}$*:*

$$\mathbb{E}[T^-, \mathcal{L}_{m,\mathcal{S}_n}, c^*] = \frac{1}{n} + \sum_{i=2}^{m} \frac{i}{(n-i+1)(n-i+2)} + \frac{(n-m)(1+(n-m)^2+2n)}{2(n-m+1)} \ .$$

Proof. We start with the case $m = 1$. We denote the expectation we seek by F. First we derive some properties of F that in turn allow us to derive a formula for it. For a learner $\mathcal{L}_1$ with 1-memory the actual memory contents can be neglected since it does not influence the state change probability. The state is thus represented by the hypothesis alone.

Assume that the learner hypothesizes c_i $(1 \leq i \leq n)$ and the teacher T^- presents example $(i, 1)$ next. The resulting probability distribution is the same as after the regular first example: all hypotheses except c_i have a probability of $1/n$ and the teacher presents $(i, 1)$ only after all other examples have been presented. Therefore, the expected number of rounds to reach c_0 for this learner is also F.

Now suppose the learner assumes hypothesis c_i and the teacher presents example $(i-1, 1)$ next. This will not change the learners hypothesis because c_i is consistent with $(i-1, 1)$. Only in the next round, in which T^- gives example $(i, 1)$, the hypothesis changes. Moreover this hypothesis change is the same as before; it only happens one round later. The expectation for this learner is therefore $F + 1$.

In general, if it takes $\ell \in \{0, \ldots, n-2\}$ rounds before the next inconsistent example arrives, the expectation is $F + \ell$. Of course, starting in the target concept c_0 has an expectation of 0.

Now let us go back to our "real" teaching process in which the learner starts in *init*. After the first example, the learner assumes hypothesis c_1 with probability 0 and all other hypotheses $c_0, c_2, \ldots, c_n$ with probability $1/n$. This means that with probability $1/n$ the learner is in a state, namely c_2, in which the next example triggers a hypothesis change. More generally, the learner is with probability $1/n$ in a state in which after $\ell = 0, \ldots, n-2$ examples a hypothesis change is triggered (the states are $c_2, c_3, \ldots, c_n$). The expected number F of rounds is thus composed of $n-1$ individual expectations, each of which is to be weighted by $1/n$. This yields

$$F = 1 + \sum_{\ell=0}^{n-2} \frac{1}{n} \cdot (F + \ell) \tag{6.3}$$

which is a linear equation with one variable and the solution

$$F = 1 + \frac{n(n-1)}{2} \ .$$

Now we turn to the case $m > 1$. This is more complicated because it takes m rounds until the memory is filled. We consider this initial phase later and first focus on the situation in which the memory already contains m examples.

The arguments are similar as above for the case $m = 1$. Assume that a learner hypothesizes c_i and memorizes the last m examples $\langle (i-m,1),\ldots,(i-1,1)\rangle$. This means that example $(i,1)$ comes next. We denote the expected number of rounds for this learner to reach the target by F'. After the example $(i,1)$ is given, the learner is in each hypothesis consistent with the new memory $\langle (i-m+1,1),\ldots,(i,1)\rangle$ with a probability of $1/(n-m+1)$.

More generally, assume that the learner hypothesizes $c_{i+\ell}$ for some $\ell > 0$ and the memory is the same as before. Then it takes $1+\ell$ rounds until a hypothesis change happens. The situation reached afterwards is essentially the same as in the special case $\ell = 0$ just discussed: the learner is in each hypothesis consistent with the new memory with a probability of $1/(n-m+1)$. Of course, the consistent hypotheses are different for different ℓ. The point, however, is that there is always a hypothesis that is inconsistent with the next example, one that is consistent with the next but inconsistent with the example after the next, and so on. In effect the expectation for a learner starting in $c_{i+\ell}$ and receiving example $(i,1)$ next is $F' + \ell$.

Our observations above allow us to state a formula for F' similar as in the case $m = 1$:

$$F' = 1 + \sum_{\ell=0}^{n-m-1} \frac{1}{n-m+1}(F' + \ell) \ . \tag{6.4}$$

Note that for $m = 1$ we get Equation (6.3). The solution of Equation (6.4) is

$$F' = 1 + \frac{(n-m)(n-m+1)}{2} \ . \tag{6.5}$$

Now, if n is large compared to m the initial m rounds can be neglected and Equation (6.5) gives a good approximation for the true expectation. Now we derive the exact values.

The probability of reaching the target in the first round is $1/n$. Otherwise after $i \le m$ rounds the learner knows the examples $(1,1),\ldots,(i,1)$ and hypothesizes $c_0, c_{i+1},\ldots,c_n$ with equal probability, namely $p_i = 1/(n-i+1)$. It follows that the probability of reaching the target c_0 in round $i > 1$ is $p_i - p_{i-1} = \frac{1}{(n-i+1)(n-i+2)}$. Calculating the expectation would therefore begin like this:

$$\frac{1}{n} \cdot 1 + \sum_{i=2}^{m} \frac{1}{(n-i+1)(n-i+2)} \cdot i \ . \tag{6.6}$$

Computing the "in-the-target-probabilities" p_i for $i \ge m$ is more difficult, but we can use the values for F' instead.

After m examples have been given, the learner memorizes $(1,1),\ldots,(m,1)$ and hypothesizes $c_0, c_{m+1},\ldots,c_n$ with probability $1/(n-m+1)$ each. Then for all $\ell \in$

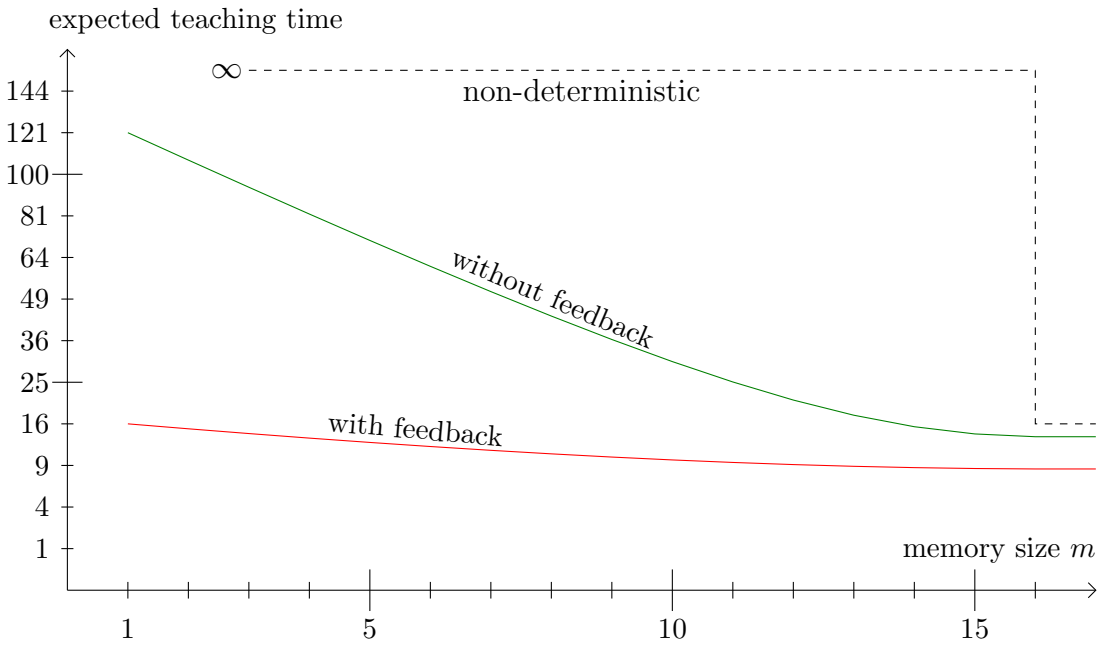

Figure 6.1 *Teaching the concept* $[1, 16] \in \mathcal{S}_{16}$ *to the randomized learners* $\mathcal{L}_m$ *with and without feedback. The y-axis is "square-rootish." The values for* $m = 1$ *and those for "with feedback" are the optimal teaching times. The values for* $1 < m \leq 16$ *without feedback are based on a reasonable, supposedly optimal, teacher. In contrast, teaching is impossible in the TD model unless the memory size is at least 16.*

$\{0, \ldots, n - m - 1\}$ there is a hypothesis, namely $c_{m+\ell}$, that is inconsistent only with the example given ℓ rounds later. Thus, the expected number of rounds to reach the target from this probability distribution is

$$\sum_{\ell=0}^{n-m-1} \frac{1}{n-m+1} \cdot (F' + \ell) \; .$$

But this probability distribution is reached after m rounds. Thus the expectations have to be considered higher by m. Therefore we have to add to Equation (6.6) the expression:

$$\sum_{\ell=0}^{n-m-1} \frac{1}{n-m+1} \cdot (F' + \ell + m) = \frac{(n-m)(1+(n-m)^2+2n)}{2(n-m+1)},$$

which will yield the sought expectation, as claimed. □

The teaching times under teacher T^- in the previous fact are optimal in the case $m = 1$, as we shall show in Fact 8.8. We conjecture that T^- is optimal for $1 < m \leq n$, too.

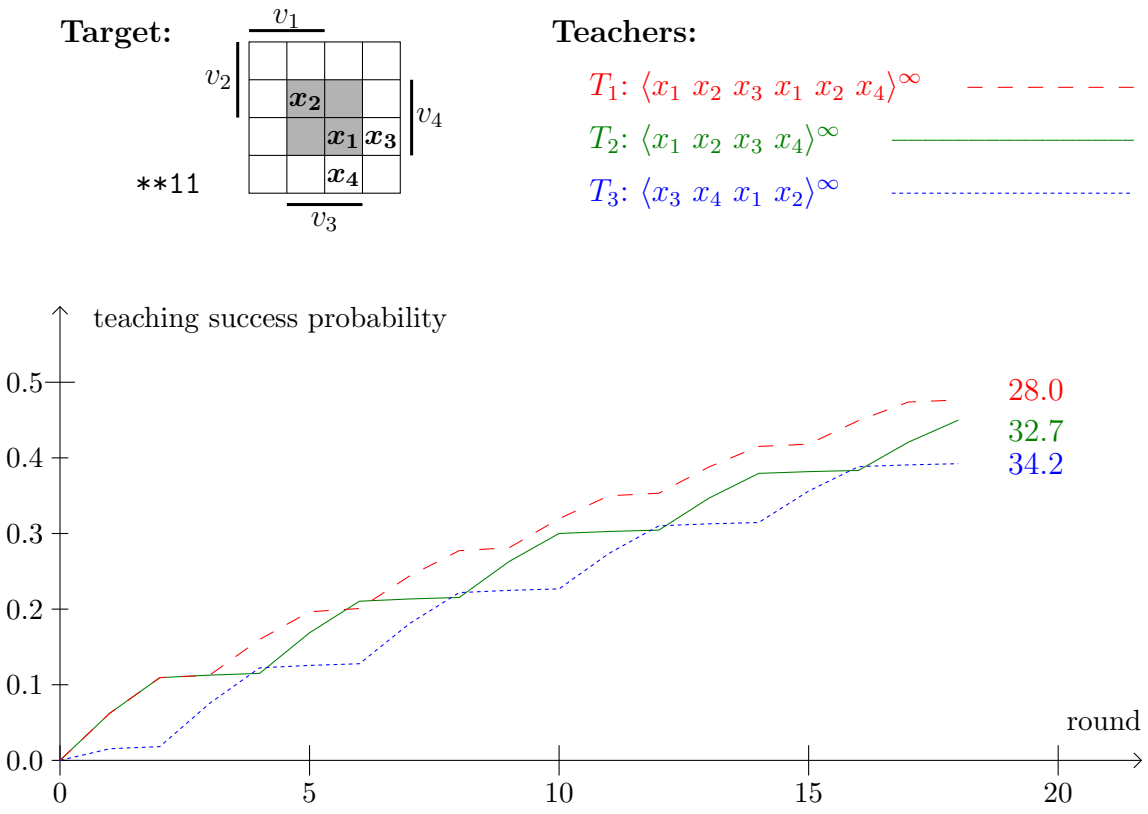

Figure 6.2 *The learner $\mathcal{L}_1$ using $\mathcal{M}_4^1$ is taught the monomial* `**11` *by three teachers T_1, T_2, T_3. The y-axis shows the probability of the learner for being in the target state in dependence of the round of the teaching process. All teachers use the same examples, but give them in different orders. This leads to different success probability curves and to different expected teaching times (the numbers on the right end of the curves).*

As an illustration, all teaching times for $n = 16$ and $m = 1, \ldots, 16$ are shown in Figure 6.1. Clearly, teaching becomes faster with growing m. Moreover the teaching speed increases continuously with m and not abruptly as in the $\mathfrak{H}$-restricted models. In particular, teaching is possible even with the smallest memory size ($m = 1$).

In addition, regardless of the memory size, the teaching times clearly improve when feedback is present.

To illustrate the influence of the order of examples in the randomized teaching model, we have calculated a numerical example. Figure 6.2 shows three teachers teaching the monomial $v_3 \wedge v_4$ without feedback to the learner $\mathcal{L}_1$ using hypothesis space $\mathcal{M}_4^1$. All teachers use the same four examples from a minimum teaching set. Every teacher, however, arranges these examples into a different sequence and teaches this sequence in an infinite loop.

We refrain from including all the numerical calculations of the teaching success probabilities in the curves of Figure 6.2. The expected teaching times will be proved in Fact 8.18.

6.4 Markov Decision Processes

All our randomized teaching model variants can be regarded as special cases of so called *Markov decision processes* (MDP). In this section we introduce some basic terminology.

Under a fixed teacher, the behavior of a randomized learner with feedback can be described as a Markov chain. In each state of the learner the teacher gives an example and the learner switches to another state with probabilities as described by the function p (see Page 123). One can then, for example, compute the expected number of rounds until a target state is reached.

In our teaching model, however, the teacher is not fixed. Rather we want to find a teacher that minimizes the expected teaching time, or at least we want to find that minimal teaching time. A generalization of Markov chains to this situation are *Markov decision processes* (MDP). These processes have been subject of research for several decades. For an extensive treatment see the books by Derman [26], by Puterman [63], and by Bertsekas [18, 17].

An MDP is a probabilistic system whose state transitions can be influenced during the process by actions which incur costs. Formally, an MDP consists of a finite set *State* of states, an initial state $s_0 \in \mathit{State}$, a finite set *Action* of actions, a function $\mathit{cost}\colon \mathit{State} \times \mathit{Action} \to \mathbb{R}$, and a function $\mathit{prob}\colon \mathit{State} \times \mathit{Action} \times \mathit{State} \to [0,1]$. The value $\mathit{cost}(s,a)$ specifies the cost incurred if action a is performed in state s. The value $\mathit{prob}(s,a,s')$ specifies the probability for the MDP to change from state s to s' under action a. A *policy* $\pi\colon \mathit{State} \to \mathit{Action}$ assigns an action to every state and thus induces a Markov chain over the state space *State*.

A special case of Markov decision processes, which is still more general than our teaching scenario, are *stochastic shortest path problems* (SSPP). In an SSPP there is a set $\mathit{State}_* \subset \mathit{State}$ of target states. Once a target state has been reached, it cannot be left and all actions in a target state incur no costs. In an SSPP the costs are then interpreted as lengths and a minimum expected cost policy corresponds to a tour with minimum expected length from the initial state to any of the target states.

The basic analogy between SSPPs and our teaching model is as follows. The set *State* contains all states of the learner; State_* contains all states in which the learner hypothesizes the target; *Action* contains all examples for the target; *cost* is set to 1, except for the target states, which incur no costs; policies correspond to teachers. The function *prob* is identical to the function p defined on Page 123. The teaching time of a teacher corresponds to the expected length of the path from the initial state to a target state under the policy corresponding to that teacher. The optimal teaching time corresponds to the minimal expected path length over all policies.

A policy $\pi\colon \mathit{State} \to \mathit{Action}$ defines a Markov chain over State and for all $s \in \mathit{State}$ an expected time $H_\pi(s)$ for reaching the target c^* from s. These expectations, called *hitting times*, satisfy the following linear equations for all $s \in \mathit{State}$:

$$H_\pi(s) = \mathit{cost}(s, \pi(s)) + \sum_{s' \in \mathit{State}} \mathit{prob}(s, \pi(s), s') \cdot H_\pi(s'). \tag{6.7}$$

For a given policy π it is therefore possible, by solving a linear equation system of size $|\mathit{State}|$, to calculate the hitting times.

Under certain assumptions, optimal policies and their expectations for stochastic shortest path problems can be characterized (see Bertsekas and Tsitsiklis [19], Bertsekas [17, Chapter 2], and Puterman [63, Chapter 7]). The first such assumption is that all costs, except in the target state, are positive. This assumption is satisfied trivially in our teaching model, as all these costs are 1. The second assumption requires the existence of a so called *proper* policy. A sufficient condition for properness is that in every state the action is chosen such that there is a positive probability of reaching the target state in the next round. A straightforward teacher that corresponds to a proper policy is a teacher that gives for every state an example inconsistent with the hypothesis. Such an example triggers a hypothesis change that leads to the target with positive probability. All our randomized teaching scenarios are thus proper in the sense of MDP theory.

Now we state the optimality condition in terms of SSPPs. Interpretations in terms of the teaching model are given in Sections 7.1 and 7.2 for learners with 1-memory and ∞-memory, respectively.

Lemma 6.7 *All hitting times $H(s)$ simultaneously assume their minimal values if and only if for all $s \in \mathit{State}$:*

$$H(s) = \min_{a \in \mathit{Action}} \left(\mathit{cost}(s, a) + \sum_{s' \in \mathit{State}} \mathit{prob}(s, a, s') \cdot H(s') \right).$$

A policy π has minimal hitting times for all states if and only if for all states $s \in \mathit{State}$:

$$\pi(s) \in \operatorname*{argmin}_{a \in \mathit{Action}} \left(\mathit{cost}(s, a) + \sum_{s' \in \mathit{State}} \mathit{prob}(s, a, s') \cdot H(s') \right).$$

The hitting time for a state $s \in \mathit{State}_*$ is $H(s) = 0$.

A policy $\pi\colon \mathit{State} \to \mathit{Action}$ corresponds to a teacher that receives feedback and that can thus choose an action depending on the current state of the learner. If the teacher receives no feedback, the results about SSPPs, including Lemma 6.7, do not apply. The corresponding notion to this teaching scenario is that of an *unobservable stochastic shortest path problem* (USSPP). Only recently Patek [61, 62] has analyzed such problems and derived an optimality characterization for them, analogous to Lemma 6.7. It requires some additional notations, which we introduce in Chapter 8 where we analyze teaching the learner $\mathcal{L}_1$ without feedback.

6.5 Discussion

Already the simple examples in this chapter have shown that the randomized model is sensitive to feedback, memory size, and the order of examples. This sensitivity is also qualitatively correct, that is, teaching becomes faster with growing memory or with feedback. Whether or not it is also quantitatively correct is hard to say. In any case one could add $\mathfrak{H}$-restrictions or $\rhd$-restrictions to influence the teaching times in order to make them more natural. But there is also another way in which the teaching times can be influenced and that is not available in the non-deterministic models, namely changing the uniform distribution into something more "plausible," like a distribution that favors simple concepts over complex ones.

Changing the probability distribution is also necessary if we want to consider infinite concept classes, because in this situation the learner can often choose between countably infinitely many follow-up hypotheses, for which the uniform distribution is undefined.

Chapter 7

Learners with Feedback or Infinite Memory

7.1 Memoryless Learners with Feedback

We call teachers with 1-memory *memoryless* because their memory contents does not influence the probability for state transitions. Teaching memoryless learners with feedback presents the simplest situation. The teacher faces no uncertainty about the current state of the learner and there are only few states. In this section we apply the general Lemma 6.7 to the special case of teaching the memoryless randomized learner $\mathcal{L}_1$ with feedback and thus derive a characterization of optimal teachers (Lemma 7.1). We then use this criterion to develop an optimal teacher for the monomials (Fact 7.2). Afterwards we treat more teaching specific questions and show that the greedy teacher is not optimal in general (Fact 7.4). Finally we compare the teachability measure E_1^+ with other popular measures of teachability and learnability (Fact 7.5).

When $\mathcal{L}_1$ receives an example z, the new memory $\boldsymbol{s}'$ will contain only this example, $\boldsymbol{s}' = \langle z \rangle$, and the follow-up hypothesis is chosen from $\mathcal{C}(\langle z \rangle)$. Thus the behavior of $\mathcal{L}_1$ in a state $(\boldsymbol{s}, h)$ does not depend on $\boldsymbol{s}$, and in effect the memory is not part of the state. Therefore the state can be described by the hypothesis alone. More precisely, the learner $\mathcal{L}_1$ looks as follows (cf. Definition 6.1):

Current state: Hypothesis $h \in \mathcal{C} \cup \{\mathit{init}\}$.

Input: Example $z \in \mathcal{X}$.

Follow-up state: Hypothesis $h' \in \mathcal{C}$.

1 **if** $z \notin \mathcal{X}(h)$ **then** choose h' uniformly at random from $\mathcal{C}(z)$;

2 **else** $h' := h$;

A teacher for teaching c^* to $\mathcal{L}_1$ with feedback is then a function $T\colon \mathcal{C} \cup \{\mathit{init}\} \to \mathcal{X}(c^*)$.

A teaching process with feedback involving $\mathcal{L}_1$ can be modeled as a stochastic shortest path problem with $\mathit{State} = \mathcal{C} \cup \{\mathit{init}\}$, $\mathit{State}_* = \{c^*\}$, $\mathit{Action} = \mathcal{X}(c^*)$, $\mathit{cost}(h, z) = 1$

for $h \neq c^*$ and $cost(c^*, z) = 0$ for all $z \in \mathcal{X}(c^*)$. Furthermore,

$$prob(h, z, h') = \begin{cases} 1/|\mathcal{C}(z)| & \text{if } z \in \mathcal{X}(h') \setminus \mathcal{X}(h), \\ 0 & \text{otherwise,} \end{cases}$$

and $prob(c^*, z, c^*) = 1$ for all $z \in \mathcal{X}(c^*)$. The initial state is *init*.

From the characterization of optimal policies (Lemma 6.7) we can derive a characterization of optimal teachers and of the minimum teaching time. Note that if $\mathcal{L}_1$ is in state h, an example $z \in \mathcal{X}(h)$ does not change its state and is therefore useless. An optimal teacher refrains from teaching such examples.

Lemma 7.1 *Let $\mathcal{C}$ be a finite concept class and $c^* \in \mathcal{C}$ be a target. Let $H\colon \mathcal{C} \cup \{init\} \to \mathbb{R}$ be such that for all $h \in \mathcal{C} \cup \{init\} \setminus \{c^*\}$,*

$$H(h) = \min_{\substack{z \in \mathcal{X}(c^*) \\ z \notin \mathcal{X}(h)}} \left(1 + \frac{1}{|\mathcal{C}(z)|} \sum_{h' \in \mathcal{C}(z)} H(h') \right) \tag{7.1}$$

and $H(c^*) = 0$. A teacher $T\colon \mathcal{C} \cup \{init\} \to \mathcal{X}(c^*)$ is optimal for teaching c^* to $\mathcal{L}_1$ with feedback if and only if for all $h \in \mathcal{C} \cup \{init\} \setminus \{c^*\}$,

$$T(h) \in \operatorname*{argmin}_{\substack{z \in \mathcal{X}(c^*) \\ z \notin \mathcal{X}(h)}} \left(1 + \frac{1}{|\mathcal{C}(z)|} \sum_{h' \in \mathcal{C}(z)} H(h') \right). \tag{7.2}$$

The minimum teaching time for teaching c^* to $\mathcal{L}_1$ with feedback is $H(init)$.

The characterization in Lemma 7.1 can be used to prove the optimality of teachers and the optimal teaching time for concepts. We show this for the class of monomials without the concept $\emptyset$.

Fact 7.2 *Let $n \geq 2$, and let $\mathcal{M}_n^1$ not contain $\emptyset$. Then the concept $\mathtt{1}^k\mathtt{*}^{n-k}$ has an optimal teaching time of*

$$E_1^+(\mathtt{1}^k\mathtt{*}^{n-k}, \mathcal{M}_n^1) = \frac{(3^n - 2^n)(2^n + 2^k) - 2^{n+k-1} + 2^{n+1} - 3^n}{3^n - 2^n + 2^{k-1}}$$

for all $k \in \{1, \ldots, n\}$. The optimal teaching time for the all-concept is

$$E_1^+(\mathtt{*}^n, \mathcal{M}_n^1) = 2^n \ .$$

Proof. Let T be the teacher defined in Figure 7.1. We begin with the simpler case $k = 0$, that is, $c^* = \mathtt{*}^n$, and claim that H with $H(h) = 2^n$ for all $h \neq c^*$ is optimal. In this case every teacher that always gives an arbitrary positive example until the learner

Input: Target $c^* \in \{\mathtt{0}, \mathtt{1}, *\}^n$, hypothesis $h \in \{\mathtt{0}, \mathtt{1}, *\}^n$.

Output: Example $(x, b) \in \mathcal{X}(c^*)$.

1 **if** $h \supset c^*$ **then output** $(x, 0)$ with

$$x[i] = \begin{cases} 1 - c^*[i] & \text{if } i = \min\{j \mid h[j] = * \neq c^*[j]\}, \\ c^*[i] & \text{if } i \neq \min\{j \mid h[j] = * \neq c^*[j]\} \text{ and } c^*[i] \neq *, \\ 0 & \text{otherwise.} \end{cases}$$

2 **else output** $(x, 1)$ with arbitrary $x \in c^* \setminus h$.

Figure 7.1 *Optimal teacher with feedback for the concept class of monomials (without the concept $\emptyset$) and the learner $\mathcal{L}_1$. When the hypothesis encompasses the target, the teacher gives a negative examples that maximizes the probability that the learner reaches, in the next round, a hypothesis that does* not *encompass the target.*

is in the target state is optimal. Every such example leads to one of 2^n hypotheses with equal probability of 2^{-n}. Therefore we have for all $z \in \mathcal{X}(c^*)$ and for all $h \neq c^*$:

$$1 + \frac{1}{|\mathcal{C}(z)|} \sum_{h' \in \mathcal{C}(z)} H(h') = 1 + 2^{-n} \cdot (2^n - 1)2^n = 2^n = H(h) \ .$$

The expectations H thus satisfy Condition (7.1) in Lemma 7.1. The teacher T satisfies Condition (7.2) in Lemma 7.1.

Now let the target concept c^* be represented by $\mathtt{1}^k *^{n-k}$ with $k \geq 1$. The behavior of the teacher is based on a partition of all hypotheses into two groups. Within a group, all hypotheses are assigned the example in the same way and have the same expected teaching time. The first group contains all hypotheses h with $h \supset c^*$; these hypotheses are called $\supset$-hypotheses. The other group contains the remaining hypotheses (including *init*), called the $\not\supset$-hypotheses.

Now we define the expectations $H \colon \mathcal{C} \cup \{init\} \to \mathbb{R}$ by

$$H(h) = H_{\supset} := \frac{(3^n - 2^n)(2^n + 2^k) - 2^{n+k-1}}{3^n - 2^n + 2^{k-1}}$$

for all $\supset$-hypotheses h and

$$H(h) = H_{\not\supset} := \frac{(3^n - 2^n)(2^n + 2^k) - 2^{n+k-1} + 2^{n+1} - 3^n}{3^n - 2^n + 2^{k-1}}$$

for all $\not\supset$-hypotheses h. Note that for $n \geq 2$ we have $H_{\not\subset} < H_{\subset}$.

We have to prove that H and T satisfy the Conditions (7.1) and (7.2) in Lemma 7.1. To this end, we shall make use of two claims.

Claim 1: Let $(x, 0)$ be consistent with c^*, that is, $x \notin c^*$. Then

(a) there are $3^n - 2^n$ hypotheses consistent with $(x, 0)$;

(b) x is of the form $y\{0,1\}^{n-k}$ with y containing $\ell \geq 1$ zeros;

(c) the number of $\supset$-hypotheses consistent with $(x, 0)$ is exactly $2^k - 2^{k-\ell} - 1$.

Proof. Without loss of generality, let $x \in 0^\ell 1^{k-\ell}\{0,1\}^{n-k}$ for some $\ell \geq 0$.
(a) There are 2^n concepts containing x, hence there are $3^n - 2^n$ concepts consistent with $(x, 0)$.
(b) If $\ell = 0$, then x would be of the form $1^k\{0,1\}^{n-k}$ and therefore in c^*, a contradiction.
(c) A concept $d \in \mathcal{M}_n^1$ encompasses c^* if and only if d is of the form $\{\mathtt{1}, *\}^k *^{n-k}$. In addition, such a concept d is consistent with $(x, 0)$ (that is, $x \notin d$) if and only if d is of the form $y\{\mathtt{1}, *\}^{k-\ell} *^{n-k}$ with $y \in \{\mathtt{1}, *\}^\ell$ containing at least one "$\mathtt{1}$". There are exactly $(2^\ell - 1) \cdot 2^{k-\ell} = 2^k - 2^{k-\ell}$ concepts satisfying the latter condition. Since c^* does not count as $\supset$-hypothesis the sought number is one less, as claimed. □ Claim 1

Claim 2: For every positive example there are 2^n consistent hypotheses of which $2^k - 1$ are $\supset$-hypotheses.

Proof. For every instance $x \in 1^k\{0,1\}^{n-k}$ there are exactly 2^n concepts containing x. The concepts that result from substituting $\mathtt{1}$'s by $*$ in c^* are the only concepts containing x and c^*. There are $2^k - 1$ such concepts (c^* itself is not a $\supset$-hypothesis). □ Claim 2

Now we prove that H satisfies Condition (7.1).

Case 1: $h \supset c^*$.

Then only negative examples $z = (x, 0)$ are inconsistent with h. Without loss of generality, let $x \in 0^\ell 1^{k-\ell}\{0,1\}^{n-k}$ with $1 \leq \ell \leq k$ (the case $\ell = 0$ is impossible by Claim 1 (b)). Then

$$1 + \frac{1}{|\mathcal{C}(z)|} \sum_{h' \in \mathcal{C}(z)} H(h') = 1 + \frac{2^k - 2^{k-\ell} - 1}{3^n - 2^n} \cdot H_{\supset} + \frac{3^n - 2^n - 2^k + 2^{k-\ell}}{3^n - 2^n} \cdot H_{\not\supset}.$$

The sum of the coefficients of $H_{\supset}$ and $H_{\not\supset}$ is $(3^n - 2^n - 1)/(3^n - 2^n)$ and thus independent of ℓ. The right hand side of the equation becomes minimal for $\ell = 1$, since the coefficient of $H_{\supset}$ becomes minimal for $\ell = 1$ and moreover $H_{\supset} > H_{\not\supset}$. After plugging in $H_{\supset}$ and $H_{\not\supset}$, a tedious calculation shows that this minimal value of the right hand side equals $H_{\supset}$. Thus in Case 1 the Condition (7.1) holds. □ Case 1

Case 2: $h \not\supset c^*$.

If z is a positive example then by Claim 2,

$$1 + \frac{1}{|\mathcal{C}(z)|} \sum_{h' \in \mathcal{C}(z)} H(h') = 1 + \frac{2^k - 1}{2^n} \cdot H_{\supset} + \frac{2^n - 2^k}{2^n} \cdot H_{\not\supset} = H_{\not\supset},$$

where the last equality is again due to a tedious calculation. If z is a negative example then

$$1 + \frac{1}{|\mathcal{C}(z)|} \sum_{h' \in \mathcal{C}(z)} H(h') = 1 + \frac{2^k - 2^{k-\ell} - 1}{3^n - 2^n} \cdot H_{\supset} + \frac{3^n - 2^n - 2^k + 2^{k-\ell}}{3^n - 2^n} \cdot H_{\not\supset}.$$

Again this expression is minimized for $\ell = 1$ and its minimal value is $H_{\supset}$. The value for a positive example is thus smaller. This means that the minimal value of $1 + \frac{1}{|\mathcal{C}(z)|} \sum_{h' \in \mathcal{C}(z)} H(h')$ is $H_{\not\supset}$, and the Condition (7.1) holds also for $\not\supset$-hypothesis. □ Case 2

In the two previous cases we have identified examples z that minimize the value $1 + \frac{1}{|\mathcal{C}(z)|} \sum_{h' \in \mathcal{C}(z)} H(h')$. The teacher in Figure 7.1 always teaches such examples. Therefore, that teacher satisfies Condition (7.2) in Lemma 7.1 and is thus optimal. □

The teacher from Figure 7.1 can be computed in linear time. It outputs a positive example whenever possible (that is, when $h \not\supset c^*$). Since there are 2^n hypotheses consistent with a positive example and $3^n - 2^n$ consistent with a negative one, this means that T follows a greedy strategy minimizing the number of consistent hypotheses for the learner to choose from, and thereby maximizing the probability for the learner to reach the target c^* in the next round.

Definition 7.3 *Let $\mathcal{C}$ be a class over X and $c^* \in \mathcal{C}$. A teacher $T\colon \mathcal{C} \cup \{init\} \to \mathcal{X}$ for c^* is called* greedy *iff for all $h \in \mathcal{C}$: $T(h) \in \operatorname{argmin}_{\substack{z \in \mathcal{X}(c^*) \\ z \notin \mathcal{C}(h)}} |\mathcal{C}(z)|$.*

The notion of greedy teacher cannot be generalized to arbitrary stochastic shortest path problems, because in general the target state cannot be reached from all states under all actions. The question of how good a greedy teacher can be is thus a teaching-specific question, which cannot readily be answered by MDP theory. Such a greedy strategy seems sensible in general and is provably optimal in the case of monomials. However, there are classes where no greedy teacher is optimal.

Fact 7.4 *There is a class $\mathcal{C}$ and target c^* such that no greedy teacher is optimal.*

Proof. Figure 7.2 displays such a concept class $\mathcal{C}$ and target c^*. The teacher T^* with teaching times H^* is an optimal teacher and T^g with H^g is the only greedy teacher. The values H^* and H^g are obtained by solving the system of linear equations (6.7). That T^* is indeed optimal can be checked using Lemma 7.1. □

h	x_1	x_2	x_3	x_4	x_5	$T^*(h)$	$H^*(h)$	$T^g(h)$	$H^g(h)$
init	–	–	–	–	–	$(x_1, 1)$	$176/35 \approx 5.0285$	$(x_1, 1)$	$2536/504 \approx 5.0317$
c^*	1	1	1	1	1	–	0	–	0
c_1	0	0	0	0	1	$(x_1, 1)$	176/35	$(x_1, 1)$	2536/504
c_2	0	0	0	1	1	$(x_1, 1)$	176/35	$(x_1, 1)$	2536/504
c_3	0	0	1	0	1	$(x_1, 1)$	176/35	$(x_1, 1)$	2536/504
c_4	0	0	1	1	1	$(x_1, 1)$	176/35	$(x_1, 1)$	2536/504
c_5	0	1	0	1	1	$(x_1, 1)$	176/35	$(x_1, 1)$	2536/504
c_6	0	1	1	0	1	$(x_1, 1)$	176/35	$(x_1, 1)$	2536/504
c_7	0	1	1	1	1	$(x_1, 1)$	176/35	$(x_1, 1)$	2536/504
c_8	1	0	0	1	0	$(x_2, 1)$	186/35	$(x_2, 1)$	2680/504
c_9	1	1	1	0	0	$(x_5, 1)$	189/35	$(x_4, 1)$	2725/504
c_{10}	1	1	1	1	0	$(x_5, 1)$	189/35	$(x_5, 1)$	2723/504

Figure 7.2 *The concept class $\mathcal{C} = \{c^*, c_1, \ldots, c_{10}\}$ over $X = \{x_1, \ldots, x_5\}$ and two teachers, T^* and T^g, teaching c^* to $\mathcal{L}_1$ with feedback. Teacher T^* is optimal and T^g is the only greedy teacher. The optimal teaching time is 176/35, but that of T^g is 2536/504 and thus suboptimal.*

In contrast to the $\mathfrak{H}$-dimension and the teaching times in the $\rhd$-restricted models, the expected teaching times $E_1^+(c, \mathcal{C})$ do not depend on additional assumptions. It is therefore possible to compare E_1^+ with other dimensions that occur in learning theory. In particular, the comparison of E_1^+ with the number MQ of membership queries (see Angluin [3]) is interesting because MQ and E_1^+ are both lower bounded by the teaching dimension.

Fact 7.5

1. *For all $\mathcal{C}$ and $c \in \mathcal{C}$: $E_1^+(c, \mathcal{C}) \geq TD(c, \mathcal{C})$.*
2. *There is no function of TD upper bounding E_1^+.*
3. *There is no function of E_1^+ upper bounding MQ.*
4. *There is a concept class $\mathcal{C}$ with $E_1^+(\mathcal{C}) > MQ(\mathcal{C})$.*
5. *For all concept classes $\mathcal{C}$, $E_1^+(\mathcal{C}) \leq 2^{MQ(\mathcal{C})}$.*

Proof.

1. For every example $z \in \mathcal{X}(c)$ there are at least $TD(c, \mathcal{C})$ consistent hypotheses. Consequently, in every round the probability of reaching the target is at most $1/TD(c, \mathcal{C})$. The expected number of rounds is therefore at least $TD(c, \mathcal{C})$.

2. Let $\mathcal{C}_n = \{c \subseteq [1,n] \mid |c| = 2\}$. Then $TD(\mathcal{C}_n) = 2$, but $E_1^+(\mathcal{C}_n) = n-1$ because the optimal teacher gives positive examples all the time and there are $n-1$ hypotheses consistent with such an example.

3. Let $\mathcal{C}_n = \{c \subseteq [1,n] \mid |c| = 1\}$. Then $E_1^+(c, \mathcal{C}_n) = 1$ for all $c \in \mathcal{C}_n$, but $MQ(\mathcal{C}_n) = n-1$.

4. $MQ(\mathcal{A}_n) = n$ and $E_1^+(\mathcal{A}_n) = 2^{n-1}$.

5. It is known (see, for example, Angluin [5]) that $\log|\mathcal{C}| \leq MQ(\mathcal{C})$ for all classes $\mathcal{C}$. Also, $E_1^+(\mathcal{C}) \leq |\mathcal{C}|$ because in every step the learner cannot choose from more than $|\mathcal{C}|$ hypotheses. Combining both inequalities yields the fact. □

Roughly speaking, teaching $\mathcal{L}_1$ can take arbitrarily longer than teaching in the teaching dimension model, but is still incomparable with membership query learning.

7.2 Learners with Infinite Memory and Feedback

Learners with infinite memory can have infinitely many states $(\boldsymbol{s}, h) \in \mathcal{X}^* \times \{\mathcal{C} \cup \{\mathit{init}\}\}$. But the behavior of the learner $\mathcal{L}_\infty$ is not affected by having the same example in the memory multiple times. This makes it pointless to teach the same example twice. Thus it suffices to consider the finitely many memories of length at most $|X|$. The number of states we have to consider is therefore only finite. From the SSPP optimality criterion Lemma 6.7 we can then immediately derive a characterization of optimal teachers for $\mathcal{L}_\infty$. This characterization is more complicated than in the $\mathcal{L}_1$ case (Lemma 7.1) and difficult to write in closed form. Our first task is thus to simplify this criterion (Lemma 7.8) by proving that an optimal teacher always gives examples that are inconsistent with the current hypothesis (see Lemma 7.7).

The optimality criterion also yields an algorithm, called *backward induction*, for computing the optimal teaching time. The runtime of this algorithm is, however, not polynomial in the representation size of the concept class. A straightforward idea to improve the backward induction is to consider only the first $TD(c^*)$ rounds of the teaching process, since there is always a teacher successful after that many rounds. But, as we show in Fact 7.9, this modified algorithm does not always yield the optimal teaching time. Indeed, that an efficient algorithm for computing E_∞^+ is unlikely to exist is shown afterwards in Theorem 7.11. This theorem is based on a general lemma (Lemma 7.10) that relates E_∞^+ to the teaching dimension.

It is not difficult to formally describe the SSPP corresponding to teaching $c^* \in \mathcal{C}$ to $\mathcal{L}_\infty$ with feedback. We stipulate that in every state $(\boldsymbol{s}, h)$ only examples $z \notin \boldsymbol{s}$ can be given. As we mentioned above, an optimal teacher would not teach other examples

anyway. This allows us to consider only memories in which no example occurs twice. The set of states is thus

$$\mathit{State} = \{(\boldsymbol{s}, h) \in \mathcal{X}(c^*)^{\leq |X|} \times (\mathcal{C} \cup \{\mathit{init}\}) \mid h \in \mathcal{C}(\boldsymbol{s}) \text{ and } (i \neq j \Rightarrow \boldsymbol{s}[i] \neq \boldsymbol{s}[j])\}.$$

The initial state is $(\langle\,\rangle, \mathit{init})$, and the set of target states is $\mathit{State}_* = \{(\boldsymbol{s}, h) \in \mathit{State} \mid h = c^*\}$. For a state $(\boldsymbol{s}, h)$ and an example $z \notin \boldsymbol{s}$ the transition probabilities are

$$\begin{aligned}&\mathit{prob}((\boldsymbol{s}, h), z, (\boldsymbol{s}', h')) \\ &\qquad = p((\boldsymbol{s}, h), z, (\boldsymbol{s}', h')) = \begin{cases} 1 & \text{if } z \in \mathcal{X}(h) \wedge \boldsymbol{s}' = \boldsymbol{s} \circ \langle z \rangle \wedge h' = h, \\ 1/|\mathcal{C}(\boldsymbol{s}')| & \text{if } z \notin \mathcal{X}(h) \wedge \boldsymbol{s}' = \boldsymbol{s} \circ \langle z \rangle \wedge h' \in \mathcal{C}(\boldsymbol{s}'), \\ 0 & \text{otherwise.} \end{cases}\end{aligned}$$

As usual, the costs are 1 for each example, except for examples given in the target state, which are costless:

$$\mathit{cost}((\boldsymbol{s}, h), z) = \begin{cases} 1 & \text{if } h \neq c^*, \\ 0 & \text{if } h = c^*. \end{cases}$$

Plugging the above into Lemma 6.7 yields an optimality characterization that is hard to write concisely. This is so because for p we have to distinguish three cases. In comparison, Lemma 7.1 looks rather simple because we could confine the actions to those examples that are inconsistent with the current hypothesis. This was possible because a consistent example would not change the state of $\mathcal{L}_1$. Giving a consistent example to $\mathcal{L}_\infty$, however, does change the learner's state. Below we show that it nevertheless suffices to consider teachers that always give inconsistent examples. This not only yields a simpler characterization, it is also interesting in its own right. As an example, some consequences for learners with selective memory are discussed at the end of this section.

Our first step towards this result is to prove that the order of the examples in an infinite memory is not important. Therefore, we can regard the memory as a set rather than a sequence. As a by-product, this reduces the number of states we have to consider.

Fact 7.6 *Let H satisfy the condition in Lemma 6.7 for an SSPP corresponding to teaching $c^* \in \mathcal{C}$ to $\mathcal{L}_\infty$ with feedback. Let $(\boldsymbol{s}, h), (\tilde{\boldsymbol{s}}, h) \in \mathit{State}$ be two states with $\boldsymbol{s}, \tilde{\boldsymbol{s}} \in \mathcal{X}^k$ and $\{\boldsymbol{s}[1], \ldots, \boldsymbol{s}[k]\} = \{\tilde{\boldsymbol{s}}[1], \ldots, \tilde{\boldsymbol{s}}[k]\}$. Then $H(\boldsymbol{s}, h) = H(\tilde{\boldsymbol{s}}, h)$.*

Proof. The proof is by induction on the length k of $\boldsymbol{s}$ and $\tilde{\boldsymbol{s}}$. We start at the maximal possible length, $k = |X|$. Let $(\boldsymbol{s}, h), (\tilde{\boldsymbol{s}}, h) \in \mathit{State}$ with $|\boldsymbol{s}| = |X|$. Then both $\boldsymbol{s}$ and $\tilde{\boldsymbol{s}}$ contain all examples in $\mathcal{X}(c^*)$ and thus $h = c^*$. Therefore $H(\boldsymbol{s}, h) = H(\tilde{\boldsymbol{s}}, h) = 0$.

Now assume that the statement holds for all memories of length $k > 0$. We show the fact for memories of length $k-1$. Let $\boldsymbol{s}, \tilde{\boldsymbol{s}} \in \mathcal{X}(c^*)^{k-1}$ be memories with identical range and h a hypothesis such that $(\boldsymbol{s}, h), (\tilde{\boldsymbol{s}}, h) \in \mathit{State}$. If $h = c^*$ then both H-values are again 0. We therefore assume that $h \neq c^*$. Since $\boldsymbol{s}$ and $\tilde{\boldsymbol{s}}$ have the same range, we have for all h' and all z by the definition of p:

$$p((\boldsymbol{s}, h), z, (\boldsymbol{s} \circ \langle z \rangle, h')) = p((\tilde{\boldsymbol{s}}, h), z, (\tilde{\boldsymbol{s}} \circ \langle z \rangle, h')).$$

With this and Lemma 6.7 we obtain:

$$\begin{aligned}
H(\boldsymbol{s}, h) &= \min_{\substack{z \in \mathcal{X}(c^*) \\ z \notin \boldsymbol{s}}} \left(1 + \sum_{(\boldsymbol{s}', h') \in \mathit{State}} p((\boldsymbol{s}, h), z, (\boldsymbol{s}', h')) \cdot H(\boldsymbol{s}', h') \right) \\
&= \min_{\substack{z \in \mathcal{X}(c^*) \\ z \notin \boldsymbol{s}}} \left(1 + \sum_{(\boldsymbol{s} \circ \langle z \rangle, h') \in \mathit{State}} p((\boldsymbol{s}, h), z, (\boldsymbol{s} \circ \langle z \rangle, h')) \cdot H(\boldsymbol{s} \circ \langle z \rangle, h') \right) \\
&= \min_{\substack{z \in \mathcal{X}(c^*) \\ z \notin \boldsymbol{s}}} \left(1 + \sum_{(\tilde{\boldsymbol{s}} \circ \langle z \rangle, h') \in \mathit{State}} p((\tilde{\boldsymbol{s}}, h), z, (\tilde{\boldsymbol{s}} \circ \langle z \rangle, h')) \cdot H(\tilde{\boldsymbol{s}} \circ \langle z \rangle, h') \right) \\
&= H(\tilde{\boldsymbol{s}}, h).
\end{aligned}$$

The third equality holds because $H(\boldsymbol{s} \circ \langle z \rangle, h') = H(\tilde{\boldsymbol{s}} \circ \langle z \rangle, h')$ by the induction hypothesis. □

Lemma 7.7 *Let $\mathcal{C}$ be a class and c^* be a target. Then there is an optimal teacher T' for teaching c^* to $\mathcal{L}_\infty$ with feedback that never gives an example consistent with the current hypothesis, that is,*

$$T'(\boldsymbol{s}, h) \notin \mathcal{X}(h)$$

for all $(\boldsymbol{s}, h) \in \mathit{State} \setminus \mathit{State}_*$.

Proof. Let H satisfy the first condition in Lemma 6.7. Let T be a teacher that gives a consistent example $z_1 = T(\boldsymbol{s}, h) \in \mathcal{X}(h)$ when the learner is in a state $(\boldsymbol{s}, h) \in \mathit{State} \setminus \mathit{State}_*$. We assume that $\boldsymbol{s}$ is of maximal length with this property. That means that in the follow-up state $(\boldsymbol{s} \circ \langle z_1 \rangle, h)$ the teacher T gives an inconsistent example $z_2 = T(\boldsymbol{s} \circ \langle z_1 \rangle, h) \notin \mathcal{X}(h)$.

We show that this teacher does not satisfy the second condition in Lemma 6.7 for state $(\boldsymbol{s}, h)$. That means we show

$$z_1 \notin \operatorname*{argmin}_{\substack{z \in \mathcal{X}(c^*) \\ z \notin \boldsymbol{s}}} \left(\mathit{cost}((\boldsymbol{s}, h), z) + \sum_{c \in \mathcal{C}(\boldsymbol{s} \circ \langle z \rangle)} \mathit{prob}((\boldsymbol{s}, h), z, (\boldsymbol{s} \circ \langle z \rangle, c)) \cdot H(\boldsymbol{s} \circ \langle z \rangle, c) \right).$$

As a shortcut we denote the expression in the large parentheses by Y_z. In general, the value of Y_z is

$$Y_z = 1 + \frac{1}{|\mathcal{C}(\boldsymbol{s} \circ \langle z \rangle)|} \cdot \sum_{c \in \mathcal{C}(\boldsymbol{s} \circ \langle z \rangle)} H(\boldsymbol{s} \circ \langle z \rangle, c).$$

for examples $z \notin \mathcal{X}(h)$ and

$$Y_z = 1 + H(\boldsymbol{s} \circ \langle z \rangle, h)$$

for examples $z \in \mathcal{X}(h)$. The value of Y_z with z set to z_1 is

$$Y_{z_1} = 1 + H(\boldsymbol{s} \circ \langle z_1 \rangle, h) = 2 + q \cdot \sum_{c \in \mathcal{C}(\boldsymbol{s} \circ \langle z_1, z_2 \rangle)} H(\boldsymbol{s} \circ \langle z_1, z_2 \rangle, c) \tag{7.3}$$

with $q = 1/|\mathcal{C}(\boldsymbol{s} \circ \langle z_1, z_2 \rangle)|$.

Now we show that (7.3) is not the minimal value of Y_z over all examples z by showing that $Y_{z_2} < Y_{z_1}$. At first we have

$$Y_{z_2} = 1 + q' \cdot \sum_{c \in \mathcal{C}(\boldsymbol{s} \circ \langle z_2 \rangle)} H(\boldsymbol{s} \circ \langle z_2 \rangle, c) \tag{7.4}$$

with $q' = 1/|\mathcal{C}(\boldsymbol{s} \circ \langle z_2 \rangle)|$. The values $H(\boldsymbol{s} \circ \langle z_2 \rangle, c)$ in the summation are

$$\begin{aligned} &H(\boldsymbol{s} \circ \langle z_2 \rangle, c) = \\ &= \min_{\substack{z \in \mathcal{X}(c^*) \\ z \notin \boldsymbol{s} \circ \langle z_2 \rangle}} \left(1 + \sum_{c' \in \mathcal{C}(\boldsymbol{s} \circ \langle z_2, z \rangle)} prob((\boldsymbol{s} \circ \langle z_2 \rangle, c), z, (\boldsymbol{s} \circ \langle z_2, z \rangle, c')) \cdot H(\boldsymbol{s} \circ \langle z_2, z \rangle, c') \right) \\ &< 1 + \sum_{c' \in \mathcal{C}(\boldsymbol{s} \circ \langle z_2, z_1 \rangle)} prob((\boldsymbol{s} \circ \langle z_2 \rangle, c), z_1, (\boldsymbol{s} \circ \langle z_2, z_1 \rangle, c')) \cdot H(\boldsymbol{s} \circ \langle z_2, z_1 \rangle, c') \end{aligned}$$

where the upper bound results from setting z to z_1. When substituting the upper bounds just derived for the H-values in (7.4), we have to distinguish between hypotheses c for which z_1 is consistent and those for which z_1 is inconsistent and triggers a hypothesis change. We get as upper bound for (7.4):

$$\begin{aligned} Y_{z_2} < 1 + q' \Bigg(&\sum_{\substack{c \in \mathcal{C}(\boldsymbol{s} \circ \langle z_2 \rangle) \\ c \notin \mathcal{C}(\boldsymbol{s} \circ \langle z_2, z_1 \rangle) \\ c \neq c^*}} \left(1 + q'' \sum_{c' \in \mathcal{C}(\boldsymbol{s} \circ \langle z_2, z_1 \rangle)} H(\boldsymbol{s} \circ \langle z_2, z_1 \rangle, c') \right) \\ &+ \sum_{\substack{c \in \mathcal{C}(\boldsymbol{s} \circ \langle z_2, z_1 \rangle) \\ c \neq c^*}} (1 + H(\boldsymbol{s} \circ \langle z_2, z_1 \rangle, c)) \Bigg) \end{aligned}$$

with $q'' = 1/|\mathcal{C}(\boldsymbol{s} \circ \langle z_2, z_1 \rangle)| = q$. Now all occurring H-values have $\boldsymbol{s} \circ \langle z_2, z_1 \rangle$ as first argument. Removing the first summation yields

$$1 + q' \left(|\mathcal{C}(\boldsymbol{s} \circ \langle z_2 \rangle) \setminus \mathcal{C}(\boldsymbol{s} \circ \langle z_2, z_1 \rangle) \setminus \{c^*\}| \cdot \left(1 + q'' \sum_{c' \in \mathcal{C}(\boldsymbol{s} \circ \langle z_2, z_1 \rangle)} H(\boldsymbol{s} \circ \langle z_2, z_1 \rangle, c') \right) + \sum_{\substack{c \in \mathcal{C}(\boldsymbol{s} \circ \langle z_2, z_1 \rangle) \\ c \neq c^*}} (1 + H(\boldsymbol{s} \circ \langle z_2, z_1 \rangle, c)) \right).$$

We set $r = |\mathcal{C}(\boldsymbol{s} \circ \langle z_2 \rangle) \setminus \mathcal{C}(\boldsymbol{s} \circ \langle z_2, z_1 \rangle) \setminus \{c^*\}|$ and $r' = |\mathcal{C}(\boldsymbol{s} \circ \langle z_2, z_1 \rangle) \setminus \{c^*\}|$. After multiplying out we obtain

$$Y_{z_2} < 1 + q'r' + q'r'q'' \sum_{c' \in \mathcal{C}(\boldsymbol{s} \circ \langle z_2, z_1 \rangle)} H(\boldsymbol{s} \circ \langle z_2, z_1 \rangle, c') + q'r + q'r \sum_{c' \in \mathcal{C}(\boldsymbol{s} \circ \langle z_2, z_1 \rangle)} H(\boldsymbol{s} \circ \langle z_2, z_1 \rangle, c')$$

and after sorting the terms,

$$Y_{z_2} < 1 + q'r' + q'r + (q'rq'' + q') \sum_{c' \in \mathcal{C}(\boldsymbol{s} \circ \langle z_2, z_1 \rangle)} H(\boldsymbol{s} \circ \langle z_2, z_1 \rangle, c'). \tag{7.5}$$

The first three terms can be upper bounded by $1 + q'r' + q'r = 1 + q'(r + r') = 2 - q' < 2$. The coefficient $q'rq'' + q'$ of the summation can be upper bounded as follows:

$$\begin{aligned} q'rq'' + q' &= q'' \cdot \frac{|\mathcal{C}(\boldsymbol{s} \circ \langle z_2 \rangle) \setminus \mathcal{C}(\boldsymbol{s} \circ \langle z_2, z_1 \rangle) \setminus \{c^*\}|}{|\mathcal{C}(\boldsymbol{s} \circ \langle z_2 \rangle)|} + q' \\ &= q'' \cdot \left(1 - \frac{|\mathcal{C}(\boldsymbol{s} \circ \langle z_2, z_1 \rangle)| + 1}{|\mathcal{C}(\boldsymbol{s} \circ \langle z_2 \rangle)|} \right) + q' \\ &< q'' \cdot \left(1 - \frac{|\mathcal{C}(\boldsymbol{s} \circ \langle z_2, z_1 \rangle)|}{|\mathcal{C}(\boldsymbol{s} \circ \langle z_2 \rangle)|} \right) + q' \\ &= q'' - \frac{1}{|\mathcal{C}(\boldsymbol{s} \circ \langle z_2 \rangle)|} + q' \\ &= q''. \end{aligned}$$

Applying these upper bounds to (7.5), it follows

$$\begin{aligned} Y_{z_2} &< 2 + q'' \cdot \sum_{c' \in \mathcal{C}(\boldsymbol{s} \circ \langle z_2, z_1 \rangle)} H(\boldsymbol{s} \circ \langle z_2, z_1 \rangle, c') \\ &= 2 + q \cdot \sum_{c' \in \mathcal{C}(\boldsymbol{s} \circ \langle z_2, z_1 \rangle)} H(\boldsymbol{s} \circ \langle z_1, z_2 \rangle, c') \\ &= Y_{z_1} \end{aligned}$$

where the first equality holds because $q'' = q$ and $H(\boldsymbol{s} \circ \langle z_2, z_1 \rangle, c') = H(\boldsymbol{s} \circ \langle z_1, z_2 \rangle, c')$ by Fact 7.6. Therefore Y_{z_2} is strictly less than Y_{z_1}, which means that the example z_1 is not in the set $\operatorname{argmin}_{z \in \mathcal{X}(c^*), z \notin \boldsymbol{s}} Y_z$. This shows that the teacher T does not satisfy the second condition in Lemma 6.7. □

In every state $(\boldsymbol{s}, h) \in \mathit{State}$ we have $h \in \mathcal{C}(\boldsymbol{s})$, and therefore the condition $z \notin \mathcal{X}(h)$ implies $z \notin \boldsymbol{s}$. Because of this we have omitted the condition $z \notin \boldsymbol{s}$ in the following optimality characterization.

Lemma 7.8 *Let $\mathcal{C}$ be a finite concept class and $c^* \in \mathcal{C}$ be a target. Let $H \colon \mathcal{X}^{\leq |X|} \times (\mathcal{C} \cup \{\mathit{init}\}) \to \mathbb{R}$ be such that for all $(\boldsymbol{s}, h) \in \mathit{State} \setminus \mathit{State}_*$:*

$$H(\boldsymbol{s}, h) = \min_{\substack{z \in \mathcal{X}(c^*) \\ z \notin \mathcal{X}(h)}} \left(1 + \frac{1}{|\mathcal{C}(\boldsymbol{s} \circ \langle z \rangle)|} \sum_{h' \in \mathcal{C}(\boldsymbol{s} \circ \langle z \rangle))} H(\boldsymbol{s} \circ \langle z \rangle, h') \right) \tag{7.6}$$

and for all $(\boldsymbol{s}, h) \in \mathit{State}_*$: $H(\boldsymbol{s}, h) = 0$. A teacher $T \colon \mathcal{X}^* \times (\mathcal{C} \cup \{\mathit{init}\}) \to \mathcal{X}(c^*)$ is optimal for teaching c^* to $\mathcal{L}_\infty$ with feedback if and only if for all $(\boldsymbol{s}, h) \in \mathit{State} \setminus \mathit{State}_*$:

$$T(\boldsymbol{s}, h) \in \operatorname*{argmin}_{\substack{z \in \mathcal{X}(c^*) \\ z \notin \mathcal{X}(h)}} \left(1 + \frac{1}{|\mathcal{C}(\boldsymbol{s} \circ \langle z \rangle)|} \sum_{h' \in \mathcal{C}(\boldsymbol{s} \circ \langle z \rangle))} H(\boldsymbol{s} \circ \langle z \rangle, h') \right). \tag{7.7}$$

The minimum teaching time for teaching c^* to $\mathcal{L}_\infty$ with feedback is $H(\langle\,\rangle, \mathit{init})$.

Lemma 7.8 is virtually the same as Lemma 7.1, only with a larger set of states. For a given state $(\boldsymbol{s}, h)$ the sum in the minimization ranges over the hitting times of all possible follow-up states. The main difference to the condition in Lemma 7.1 is that in (7.6) and (7.7) there are no cyclic dependencies within the H-values. Intuitively, the reason for this is that a learner with infinite memory cannot reach the same state twice during a teaching process, because in each round the memory grows.

The lack of cyclic dependencies yields a straightforward inductive algorithm for computing all optimal hitting times. A state $(\boldsymbol{s}, h)$ with $|\boldsymbol{s}| = |X|$ has a hitting time of $H(\boldsymbol{s}, h) = 0$. The optimal hitting times for states with smaller memories can be computed using formula (7.6) in Lemma 7.8, until finally the states with empty memory are reached. This algorithm is called *backward induction* and runs in time polynomial in the representation size of the MDP, but not polynomial in the representation size of the teaching problem, that is, in the matrix representation of $\mathcal{C}$.

A tempting idea for improvement is based on the observation that a teacher that gives a minimum teaching set is always successful after $\mathit{TD}(c^*)$ rounds. Not every such teacher is optimal, but one could conjecture that there is at least one optimal teacher among them. This is, however, not always the case.

Fact 7.9 *There is a concept class $\mathcal{C}$ and a concept $c^* \in \mathcal{C}$ such that all teachers teaching c^* with feedback to the learner $\mathcal{L}_\infty$ finitely within $\mathit{TD}(c^*)$ rounds are suboptimal.*

Proof. The concept class $\mathcal{C}$ and the concept c^* are defined by Figure 7.3. The teaching dimension of c^* is three and the unique minimum teaching set is $S = \{(x_1, 1), (x_2, 1), (x_3, 1)\}$. We first prove that every teacher that teaches finitely within three rounds needs expected 2.6 rounds.

Let T be such a teacher. In order to be successful after three rounds, T must teach the examples in S in some order and hence start with $(x_1, 1), (x_2, 1)$, or $(x_3, 1)$. For symmetry reasons we can assume, without loss of generality, that T starts with $(x_1, 1)$. The probability for the learner to reach c^* in the first step is $1/9$. The remaining probability mass of $8/9$ is equally distributed between the possibility of reaching a hypothesis containing x_2, but not x_3, and the possibility of reaching a hypothesis containing x_3, but not x_2. As both cases are symmetric, we assume the first one without loss of generality. The teacher then goes on teaching $(x_3, 1)$, which is the only example in S inconsistent with the current hypothesis. Then the probability of reaching c^* is $1/5$ since there are five hypotheses consistent with $(x_1, 1)$ and $(x_3, 1)$, namely $c^*, c_2, c_5, c_8, c_{11}$. With probability $4/5$ the learner reaches a hypothesis not containing x_2. Finally the teacher gives $(x_2, 1)$, which leads to c^* with certainty. Altogether the expected number of rounds is

$$\frac{1}{9} \cdot 1 + \frac{8}{9} \cdot \frac{1}{5} \cdot 2 + \frac{8}{9} \cdot \frac{4}{5} \cdot 3 = 2.6 \ .$$

On the other hand, let T' be the teacher starting with $(x_4, 1)$ and then teaching in each round an inconsistent example from S. In the first round, the probability for an immediate transition to c^* is $1/4$. In the second, third, and fourth round the probabilities are $1/3$, $1/2$, and 1, respectively, since each example rules out exactly one hypothesis. Thus the expected number of rounds for T' is

$$\frac{1}{4} \cdot 1 + \frac{3}{4} \cdot \frac{1}{3} \cdot 2 + \frac{3}{4} \cdot \frac{2}{3} \cdot \frac{1}{2} \cdot 3 + \frac{3}{4} \cdot \frac{2}{3} \cdot \frac{1}{2} \cdot 4 = 2.5 \ .$$

Therefore the teacher T is not optimal. □

We are going to show that most likely there is no polynomial time algorithm to even approximate $E^+_\infty(c^*, \mathcal{C})$ up to a constant factor. This result relies on the following lemma, which shows that $E^+_\infty(c^*, \mathcal{C})$ differs from $TD(c^*, \mathcal{C})$ by at most a factor of two.

Lemma 7.10 *Let $\mathcal{C}$ be a class and let $c^* \in \mathcal{C}$ be a target. For all $m \in [1, TD(c^*, \mathcal{C})]$,*

$$E^-_m(c^*, \mathcal{C}) \geq E^+_m(c^*, \mathcal{C}) \geq \frac{m(m-1)}{2\,TD(c^*, \mathcal{C})} + TD(c^*) + 1 - m,$$

and for all $m > TD(c^*, \mathcal{C})$ and for $m = \infty$,

$$TD(c^*, \mathcal{C}) \geq E^-_m(c^*, \mathcal{C}) \geq E^+_m(c^*, \mathcal{C}) \geq \frac{TD(c^*, \mathcal{C})}{2}.$$

	x_1	x_2	x_3	x_4	x_5	x_6
init	–	–	–	–	–	–
c^*	1	1	1	1	1	1
c_1	1	1	0	1	1	1
c_2	1	0	1	1	1	1
c_3	0	1	1	1	1	1
c_4	1	1	0	0	1	0
c_5	1	0	1	0	1	0
c_6	0	1	1	0	1	0
c_7	1	1	0	0	0	1
c_8	1	0	1	0	0	1
c_9	0	1	1	0	0	1
c_{10}	1	1	0	0	0	0
c_{11}	1	0	1	0	0	0
c_{12}	0	1	1	0	0	0

Figure 7.3 *Concept class and target for which the optimal $\mathcal{L}_\infty$-teacher with feedback has not finished after $TD(c^*) = 3$ rounds. The optimal teacher starts with $(x_4, 1)$ and finishes after 4 rounds (see Fact 7.9).*

Proof. Let $k = TD(c^*)$ and $m \in [1, TD(c^*)]$. It suffices to show the statement for E_m^+. The proof is based on the following observation.

Claim: For i examples $z_0, \ldots, z_{i-1} \in \mathcal{X}(c^*)$: $\quad |\mathcal{C}(\{z_0, \ldots, z_{i-1}\})| \geq k + 1 - i$.

Proof. Assume $|\mathcal{C}(\{z_0, \ldots, z_{i-1}\})| \leq k - i$. Then c^* can be specified with $k - i - 1$ examples with respect to $\mathcal{C}(\{z_0, \ldots, z_{i-1}\})$ (each example rules out at least one concept). Thus, c^* can be uniquely specified with $z_0, \ldots, z_{i-1}$ plus $k-i-1$ other examples, which amounts to $k - 1$ examples. This contradicts $TD(c^*) = k$. □ Claim

Using the claim we upper bound the probabilities for reaching the target in round $i = 0, \ldots, m-2$. At the end of round i the learner knows $i+1$ examples and therefore can choose between at least $k - i$ consistent hypotheses (see the claim). Thus, the probability for entering c^* in round $i + 1$ is at most $p_i = 1/(k - i)$. Beginning with round $m - 1$, the learner knows m examples at the end of the round and has in each following round $i \geq m - 1$ a probability of at most $p_i = p_{m-1} = 1/(k + 1 - m)$ of reaching c^*.

No teaching process can be faster than one with the probabilities p_i described above. The expected teaching time of such a process is

$$\sum_{i=0}^{m-2} (i+1) \cdot p_i \cdot \prod_{j=0}^{i-1} (1 - p_j) \;+\; \sum_{i=m-1}^{\infty} (i+1) \cdot p_i \cdot \prod_{j=0}^{i-1} (1 - p_j) \;. \tag{7.8}$$

We first calculate the second sum in (7.8). Since $\prod_{j=0}^{m-2}(1-p_j) = \frac{k-m+1}{k}$ the product $\prod_{j=0}^{i-1}(1-p_j)$ in the right sum equals $\frac{k-m+1}{k} \cdot (1-p_{m-1})^{i-m+1}$ and the whole sum can be written as

$$\begin{aligned}
&\sum_{i=m-1}^{\infty} (i+1) \cdot p_{m-1} \cdot \tfrac{k-m+1}{k} \cdot (1-p_{m-1})^{i-m+1} \\
&= \tfrac{k-m+1}{k} \cdot \sum_{i=0}^{\infty} (m+i) \cdot p_{m-1} \cdot (1-p_{m-1})^{i} \\
&= \tfrac{k-m+1}{k} \cdot \left(m-1+\sum_{i=0}^{\infty} (i+1) \cdot p_{m-1} \cdot (1-p_{m-1})^{i} \right) .
\end{aligned}$$

The sum appearing in the last line is the expectation of the first success in a Bernoulli experiment with probability p_{m-1} and thus equals $1/p_{m-1} = k-m+1$. For the second sum in (7.8) we therefore get

$$\tfrac{k-m+1}{k} \cdot (m-1+k-m+1) = k-m+1 \; .$$

Calculating the first sum in (7.8) yields

$$\sum_{i=0}^{m-2} (i+1) \cdot \tfrac{1}{k-i} \cdot \prod_{j=0}^{i-1} \tfrac{k-j-1}{k-j} = \sum_{i=0}^{m-2} (i+1) \cdot \tfrac{1}{k-i} \cdot \tfrac{k-i}{k} = \frac{m(m-1)}{2k} \; .$$

Putting it together we obtain $\frac{m(m-1)}{2k} + k + 1 - m$ as the value of (7.8).

For $m > TD(c^*)$ the teaching process described above takes at most $TD(c^*)$ rounds. The lower bound is therefore the same as for $m = TD(c^*)$. Moreover, a teacher giving the examples of a minimum teaching set is successful after $TD(c^*)$ rounds, from which it follows that $TD(c^*) \geq E_m^-(c^*) \geq E_m^+(c^*)$. □

Setting $m = \infty$ we conclude from the previous lemma that

$$TD(c^*, \mathcal{C}) \geq E_\infty^+(c^*, \mathcal{C}) \geq \frac{TD(c^*, \mathcal{C})}{2}, \tag{7.9}$$

which means that every algorithm computing $E_\infty^+(c^*, \mathcal{C})$ also computes a factor 2 approximation of the teaching dimension. Theorem 3.6 shows that the teaching set problem is a hard approximation problem, and this suggests that a similar result holds for E_∞^+ as well.

The problem corresponding to `MIN-TEACHING-SET` in the TD model would be the problem of finding an optimal teacher for $\mathcal{L}_\infty$ with feedback. Such a teacher has a representation size of $O(|X|^{|X|! \cdot |\mathcal{C}|})$ and is thus not polynomial in the representation size $|\mathcal{C}| \cdot |X|$ of the teaching problem. Finding such an optimal teacher is therefore not

an $\mathcal{NPO}$ optimization problem (see Ausiello *et al.* [10]), and results about the hardness of SET-COVER cannot be immediately transferred. But a closer look at the proofs of these hardness results shows that also approximating this "set cover number" is hard in some sense. We demonstrate such a reasoning in the next theorem, which shows that E^+_∞ is hard to approximate within a factor of $\frac{1}{2}(1-\varepsilon)\ln(|\mathcal{C}|-1)$ for any $\varepsilon > 0$.

Theorem 7.11 *If there is a polynomial time algorithm computing for all finite classes $\mathcal{C}$ and concepts $c \in \mathcal{C}$ rational number $A(c,\mathcal{C})$ such that*

$$\frac{E^+_\infty(c,\mathcal{C})}{\frac{1}{2}\sqrt{(1-\varepsilon)\ln(|\mathcal{C}|-1)}} \quad\leq\quad A(c,\mathcal{C}) \quad\leq\quad \tfrac{1}{2}\sqrt{(1-\varepsilon)\ln(|\mathcal{C}|-1)}\cdot E^+_\infty(c,\mathcal{C})$$

for some $\varepsilon > 0$, then $\mathcal{NP} \subseteq DTime(n^{O(\log\log n)})$.

Proof. We suppose for a contradiction that there is such a polynomial time algorithm. Using (7.9) we get

$$\frac{TD(c,\mathcal{C})}{\sqrt{(1-\varepsilon)\ln(|\mathcal{C}|-1)}} \quad\leq\quad A(c,\mathcal{C}) \quad\leq\quad \tfrac{1}{2}\sqrt{(1-\varepsilon)\ln(|\mathcal{C}|-1)}\cdot TD(c,\mathcal{C}).$$

Using the correspondence between SET-COVER and MIN-TEACHING-SET instances (see Page 48) we conclude that there is also a polynomial time algorithm A' computing for every SET-COVER instance $(U, V_1, \ldots, V_k)$ a value $A'(U, V_1, \ldots, V_k)$ with

$$\frac{SC(U,V_1,\ldots,V_k)}{\sqrt{(1-\varepsilon)\ln|U|}} \quad\leq\quad A'(U,V_1,\ldots,V_k) \quad\leq\quad \tfrac{1}{2}\sqrt{(1-\varepsilon)\ln|U|}\cdot SC(U,V_1,\ldots,V_k)$$

where $SC(U, V_1, \ldots, V_k)$ is the minimal number of sets in $V_1, \ldots, V_k$ needed to cover U.

Now let Π be an $\mathcal{NP}$ decision problem and let Π^+ and Π^- be the set of positive and negative instances, respectively. Feige [28, Theorem 4.4] shows that it is possible to map every instance $\pi \in \Pi$ in time $n^{O(\log\log n)}$ to a SET-COVER instance $(U, V_1, \ldots, V_k)$ such that for an easily computable value Q:

$$\begin{aligned} \pi \in \Pi^+ &\Leftrightarrow SC(U,V_1,\ldots,V_k) \leq Q, \\ \pi \in \Pi^- &\Leftrightarrow SC(U,V_1,\ldots,V_k) > (1-\varepsilon)\ln|U|\cdot Q. \end{aligned} \tag{7.10}$$

By checking the condition "$A'(U, V_1, \ldots, V_k) < \sqrt{(1-\varepsilon)\ln|U|}\cdot Q$" one can decide whether π is a positive or negative instance for Π: Assume the condition holds; then $SC(U, V_1, \ldots, V_k) < (1-\varepsilon)\ln|U|\cdot Q$ and by (7.10) we have $\pi \in \Pi^+$. Now let $\pi \in \Pi^+$; then by (7.10) $SC(U, V_1, \ldots, V_k) \leq Q$ and $A'(U, V_1, \ldots, V_k) \leq \frac{1}{2}\sqrt{(1-\varepsilon)\ln|U|}\cdot SC(U, V_1, \ldots, V_k) \leq \frac{1}{2}\sqrt{(1-\varepsilon)\ln|U|}\cdot Q < \sqrt{(1-\varepsilon)\ln|U|}\cdot Q$.

It follows that it can be decided in time $n^{O(\log\log n)}$ whether any given π is a positive instance for the arbitrarily chosen $\mathcal{NP}$ problem Π. This means that $\mathcal{NP} \subseteq DTime(n^{O(\log\log n)})$. □

Although Lemma 7.10 is responsible for a negative result about the approximability of E^+_∞, we can also draw a positive conclusion from it: if the TD-value is known, there is often no need to compute the E^+_∞-value. For example, the teachabilities of concepts or classes for infinite memory learners with feedback can be compared by comparing the teaching dimensions of these concepts or classes.

We finally mention some implications of Lemma 7.7 to teaching learners with infinite *selective* memory (see Definition 4.28). An optimal teacher for infinite selective memory learners with feedback always teaches an example inconsistent with the current hypothesis; a consistent example would not change the learner's state. Such an optimal teacher can be applied to the non-selective memory learner, which then behaves exactly like the selective memory learner. On the other hand, an optimal teacher that satisfies Lemma 7.7 makes the selective memory learner behave like the non-selective memory learner. This means that the optimal teaching times for learners with infinite selective memory and infinite non-selective memory are equal. The same is true for learners with memory size one. However, it is open whether it holds for m-memory learners with $1 < m < \infty$, because the proof of Fact 7.6 does not generalize to $m < \infty$.

7.3 Learners with Infinite Memory and without Feedback

We limit our discussion to the hardness of approximating E^-_∞. From Lemma 7.10 we know that $TD(c^*, \mathcal{C}) \geq E^-_\infty(c^*, \mathcal{C}) \geq TD(c^*, \mathcal{C})/2$. We can thus prove an analog to Theorem 7.11.

Theorem 7.12 *If there is a polynomial time algorithm computing for all $c^*, \mathcal{C}$ a rational number $A(c^*, \mathcal{C})$ such that*

$$\frac{E^-_\infty(c, \mathcal{C})}{\frac{1}{2}\sqrt{(1-\varepsilon)\ln(|\mathcal{C}|-1)}} \quad \leq \quad A(c^*, \mathcal{C}) \quad \leq \quad \tfrac{1}{2}\sqrt{(1-\varepsilon)\ln(|\mathcal{C}|-1)} \cdot E^-_\infty(c^*, \mathcal{C})$$

for some $\varepsilon > 0$, then $\mathcal{NP} \subseteq DTime(n^{O(\log\log n)})$.

Another consequence of Lemma 7.10 is that the optimal teaching times E^+_∞ and E^-_∞ of teaching with and without feedback differ by a factor of at most two. This means that feedback is not that much of a help when teaching randomized learners with infinite memory.

Chapter 8

Memoryless Learners without Feedback

8.1 A Characterization of Successful Teachers

For convenience we restate some notation introduced in Section 6.2 for the special case of teaching the learner $\mathcal{L}_1$ without feedback. Again, as in Section 7.1, the state of the learner is only the hypothesis. Given a teacher $T\colon \mathbb{N} \to \mathcal{X}(c^*)$ the series of probability distributions during the teaching process is $\delta_T^{(t)}\colon \mathcal{C} \cup \{init\} \to [0,1]$ with $\delta^{(0)} = \delta^{(init)}$ and

$$\delta_T^{(t+1)}(h) = \begin{cases} \delta_T^{(t)}(h) + \dfrac{1}{|\mathcal{C}(T(t))|} \cdot \displaystyle\sum_{c \notin \mathcal{C}(T(t))} \delta^{(t)}(c) & \text{if } h \in \mathcal{C}(T(t)), \\ 0 & \text{if } h \notin \mathcal{C}(T(t)). \end{cases}$$

Intuitively speaking, the following happens to the probability distribution during one round of teaching. At the beginning of round t, let $\delta^{(t)}$ be the probability distribution over all hypotheses. Then an example $z = T(t) \in \mathcal{X}(c^*)$ for the target is given. This example *activates* the probability mass of all hypotheses it is inconsistent with, that is, $\sum_{c \notin \mathcal{C}(z)} \delta^{(t)}(c)$. This probability mass is then equally distributed among all hypotheses that are consistent with z. That means, every hypothesis $h \in \mathcal{C}(z)$ receives a share of $\frac{1}{|\mathcal{C}(z)|} \cdot \sum_{c \notin \mathcal{C}(z)} \delta^{(t)}(c)$ of additional probability. The probability of all other hypotheses is set to zero.

For illustration, Figures 8.1 and 8.2 show two attempts to teach $c_0 = [1,4] \in \mathcal{S}_4$. The teacher in Figure 8.2 is not successful because it fails to activate the probability mass of hypothesis c_4. We observe that activating all hypotheses at least "once in a while" is necessary to be a successful teacher. The easiest way to achieve this is to iterate through a teaching set forever. The teacher in Figure 8.1 does just this and is successful. More general, we have the following fact.

Fact 8.1 *Let $\mathcal{C}$ be a finite concept class and $c^* \in \mathcal{C}$. Let $S = \{z_1, \ldots, z_k\}$ be a teaching set for c^* with respect to $\mathcal{C}$. The teacher $T\colon \mathbb{N} \to \mathcal{X}(c^*)$ with $T(i) = z_{1+(i \bmod k)}$ for all $i \in \mathbb{N}$ teaches c^* successfully to $\mathcal{L}_1$ without feedback. The expected teaching time is at most $|S| \cdot |\mathcal{C}|$.*

t	0	1	2	3	4	5	6	$\cdots$	31	32
$\delta_T^{(t)}(init)$	*1.000*	0.000	0.000	0.000	0.000	0.000	0.000	$\cdots$	0.000	0.000
$\delta_T^{(t)}(c_0)$	0.000	0.250	0.312	0.391	0.488	0.548	0.607	$\cdots$	0.988	0.990
$\delta_T^{(t)}(c_1)$	0.000	0.000	0.062	0.141	*0.238*	0.000	0.059	$\cdots$	0.004	*0.005*
$\delta_T^{(t)}(c_2)$	0.000	*0.250*	0.000	0.078	0.176	*0.235*	0.000	$\cdots$	0.002	0.003
$\delta_T^{(t)}(c_3)$	0.000	0.250	*0.312*	0.000	0.098	0.157	*0.216*	$\cdots$	0.000	0.002
$\delta_T^{(t)}(c_4)$	0.000	0.250	0.312	*0.391*	0.000	0.060	0.118	$\cdots$	*0.006*	0.000
$T(t)$	$(1,1)$	$(2,1)$	$(3,1)$	$(4,1)$	$(1,1)$	$(2,1)$	$(3,1)$	$\cdots$	$(4,1)$	$(1,1)$

Figure 8.1 *Probability distributions over the hypotheses in $\mathcal{S}_4 \cup \{init\}$ during a teaching process. The teacher T gives the examples $(1,1),(2,1),(3,1),(4,1)$ for the target concept c_0 in an endless loop. The slanted numbers show the probability mass activated in each round by the given example. During the teaching process the probability mass aggregates at the target hypothesis.*

t	0	1	2	3	4	5	6	$\cdots$	10	11
$\delta_T^{(t)}(init)$	*1.000*	0.000	0.000	0.000	0.000	0.000	0.000	$\cdots$	0.000	0.000
$\delta_T^{(t)}(c_0)$	0.000	0.250	0.312	0.391	0.426	0.454	0.470	$\cdots$	0.495	0.497
$\delta_T^{(t)}(c_1)$	0.000	0.000	0.062	*0.141*	0.000	0.028	0.044	$\cdots$	0.000	0.002
$\delta_T^{(t)}(c_2)$	0.000	*0.250*	0.000	0.078	*0.113*	0.000	0.016	$\cdots$	*0.007*	0.000
$\delta_T^{(t)}(c_3)$	0.000	0.250	*0.312*	0.000	0.035	*0.063*	0.000	$\cdots$	0.003	*0.005*
$\delta_T^{(t)}(c_4)$	0.000	0.250	0.312	0.391	0.426	0.454	0.470	$\cdots$	0.495	0.497
$T(t)$	$(1,1)$	$(2,1)$	$(3,1)$	$(1,1)$	$(2,1)$	$(3,1)$	$(1,1)$	$\cdots$	$(2,1)$	$(3,1)$

Figure 8.2 *Probability distributions over the hypotheses in $\mathcal{S}_4 \cup \{init\}$ during a teaching process. The teacher T gives the examples $(1,1),(2,1),(3,1)$ for the target concept c_0 in an endless loop. The slanted numbers show the probability mass activated in each round by the given example. The probability mass belonging to c_4 is never activated. Consequently, the probability mass aggregates not only at the target c^* but also at c_4. This teacher is therefore not successful.*

Proof. Unless the learner hypothesizes c^*, there is at least one example z_i inconsistent with the current hypothesis. Therefore during every iteration through $z_1, \ldots, z_k$ at least one hypothesis change occurs. This change leads to c^* with probability at least $1/|\mathcal{C}|$. The expected number of rounds to reach the target is therefore at most $k \cdot |\mathcal{C}|$. □

The upper bound in the previous fact can be improved because typically much less than $|\mathcal{C}|$ concepts are consistent with any given memory contents and thus the probability of reaching the target is often much greater than $1/|\mathcal{C}|$. A more important consequence of Fact 8.1 is that for all concepts c and all finite concept classes $\mathcal{C}$ the

value $E_1^-(c, \mathcal{C})$ is finite.

We have seen that presenting a teaching set infinitely often always yields a teacher with a finite teaching time. And in some sense conversely, a teacher *must* present teaching sets infinitely often in order to be successful. This is in fact a characterization of successful teachers.

Fact 8.2 *Let $\mathcal{C}$ be a finite concept class and $c^* \in \mathcal{C}$ with $TD(c^*, \mathcal{C}) > 1$. A teacher $T\colon \mathbb{N} \to \mathcal{X}(c^*)$ is a successful teacher for c^* to $\mathcal{L}_1$ if and only if for all $t \in \mathbb{N}$ there is a $t' \geq t$ such that the set $\{T(i) \mid t \leq i \leq t'\}$ is a teaching set for c^* with respect to $\mathcal{C}$.*

Proof. We begin with the "only if" direction. Suppose there was a minimal t such that for all $t' \geq t$, $\{T(i) \mid t \leq i \leq t'\}$ is no teaching set for c^*. Then the set $S = \{T(i) \mid i \geq t\}$ is also no teaching set.

Since $TD(c^*) > 1$ there are always at least two concepts consistent with any example given by the teacher. Therefore, at every round there are at least two hypotheses with positive probability. In particular after round t not the whole probability mass is concentrated in the target. Let $h \neq c^*$ be a hypothesis with positive probability after round t. Then there must eventually come an example that is inconsistent with h, say $T(t')$ for some $t' > t$. In round $t' + 1$ all hypotheses in $\mathcal{C}(T(t'))$ have a positive probability. In particular, since $T(t') \in S$, all hypotheses in $\mathcal{C}(S)$ have a positive probability. Among these hypotheses there is an $h' \neq c^*$ because S is no teaching set for c^*. This h' is consistent with S and therefore after round t the teacher gives no example inconsistent with h'. Consequently the probability for h' will remain positive forever and the probability for the target c^* does not converge to 1.

For the "if" direction assume that for all $t \in \mathbb{N}$ there is a $t' \geq t$ such that the set $\{T(i) \mid t \leq i \leq t'\}$ is a teaching set for c^* with respect to $\mathcal{C}$. Recall that for all $t \in \mathbb{N}$ and $h \in \mathcal{C} \cup \{init\}$ the value $\delta_T^{(t)}(h)$ is the probability for being in state h at round t, that is, immediately before example $T(t)$ is given.

By assumption there are rounds $0 = t_0 < t_1 < t_2 < \dots$ such that for all j the set $\{T(i) \mid t_j \leq i < t_{j+1}\}$ is a teaching set for c^*. We show that $\lim_{j\to\infty}(1 - \delta_T^{(t_j)}(c^*)) = 0$. As abbreviation for the "non-target probability" we set $\alpha_j := 1 - \delta_T^{(t_j)}(c^*)$.

At round t_j the average probability of a non-target hypothesis is least $\alpha_j/|\mathcal{C}|$, and hence there is a hypothesis h whose probability is at least $\alpha_j/|\mathcal{C}|$. Before round t_{j+1} an example is given that is inconsistent with h. At this point the probability for the target increases by at least $\alpha_j/|\mathcal{C}|^2$ because the probability mass of h is distributed among at most $|\mathcal{C}|$ hypotheses (including the target). That means that the non-target probability decreases by at least $\alpha_j/|\mathcal{C}|^2$. In other words, $\alpha_{j+1} \leq \alpha_j - \alpha_j/|\mathcal{C}|^2$. It follows $\alpha_{j+1} \leq (1 - 1/|\mathcal{C}|^2) \cdot \alpha_j$ and therefore $(\alpha_j)_{j\in\mathbb{N}}$ converges to 0 for $j \to \infty$. This shows that $\lim_{j\to\infty} \delta_T^{(t_j)}(c^*) = 1$, and thus T is successful. □

Successful teachers are good and well, but of course we would prefer an optimal teacher. In addition, the optimal teaching time describes the teachability that is as-

signed to each concept. In the next two sections we shall see that characterizing optimal teachers or computing optimal teaching times is much more difficult than characterizing teachers that are merely successful.

8.2 A Characterization of Optimal Teachers

In all situations discussed in Chapter 7 the optimal teaching time could, in principle, be computed by computing the teaching times of finitely many "reasonable" teachers. A simpler method was to use characterizations of optimal teachers and their hitting times. In contrast, in the situation of memoryless learners without feedback there are uncountably many teachers. Therefore the optimal teaching time cannot be computed, not even in principle, by evaluating all teachers. It remains the approach via a characterization of optimal teachers.

Patek [61, 62] presents an optimality criterion for policies in the setting of *unobservable* stochastic shortest path problems (USSPP). The idea of this criterion is to use the probability distributions δ over the learner's states as new states. This results in an *observable* Markov decision process that, however, has uncountably many states.

In the same way as SSPPs are more general than our randomized teaching models with feedback, USSPPs are more general than our randomized models without feedback. However, we face two problems when trying to adapt Patek's result to our teaching model. First, the criterion holds only for USSPPs satisfying certain assumptions. The most general such assumption presented by Patek is called "Assumption C." Second, policies (that is, teachers) have to be defined over so called *information states*, which are slightly different from the probability distributions δ. Our plan to get the optimal teacher characterization in this section now is as follows:

1. Explain information states and teachers over information states and introduce other necessary notation.

2. Translate Assumption C into our teaching terminology and show that it holds for teaching $\mathcal{L}_1$ without feedback.

3. Translate Patek's characterization of optimal policies and teaching times into our teaching terminology.

Teachers Over Information States

An *information state* is a probability distribution $\gamma\colon (\mathcal{C} \cup \{\mathit{init}\}) \setminus \{c^*\} \to [0,1]$ over all learner's states except the target state. Formally, for a distribution $\delta\colon \mathcal{C} \cup \{\mathit{init}\} \to [0,1]$ the information state is γ with

$$\gamma(h) = \frac{\delta(h)}{1 - \delta(c^*)}$$

for all $h \in (\mathcal{C} \cup \{init\}) \setminus \{c^*\}$. We abbreviate $(\mathcal{C} \cup \{init\}) \setminus \{c^*\}$ with $\widehat{\mathcal{C}}$. In other words, γ describes the conditional probabilities of being in the states $h \neq c^*$ provided that the target has not yet been reached. The set of all information states is denoted by

$$\Gamma = \{\gamma \colon \widehat{\mathcal{C}} \to [0,1] \mid \sum_{c \in \widehat{\mathcal{C}}} \gamma(c) = 1\} \cup \{\mathbf{0}\},$$

where $\mathbf{0}$ is the constant zero function over $\widehat{\mathcal{C}}$. We use $\widehat{\mathcal{C}}(z)$ to denote the set $\mathcal{C}(z) \cap \widehat{\mathcal{C}} = \mathcal{C}(z) \setminus \{c^*\}$ for examples z. Moreover, we write "$d \notin \widehat{\mathcal{C}}(z)$" as shorthand for "$d \in \widehat{\mathcal{C}} \setminus \widehat{\mathcal{C}}(z)$." The special information state $\mathbf{0}$, in which all probabilities are zero, means that the target has been reached with probability 1. This can only happen when a complete teaching set fits into the learner's memory, that is, if $TD(c^*) = 1$. Other than in this special case there is no target state. The initial information state is $\gamma^{(init)}$ with $\gamma^{(init)}(init) = 1$.

The behavior of the learner $\mathcal{L}_1$ can also be described in terms of information states, rather than in terms of the probability distribution δ (compare with Page 153). When $\mathcal{L}_1$ in information state γ receives the example $z \in \mathcal{X}(c^*)$ it switches to an information state $f(\gamma, z)$. The function $f \colon \Gamma \times \mathcal{X}(c^*) \to \Gamma$ maps γ and z to an information state $\hat{\gamma}$ with

$$\hat{\gamma}(c) = \begin{cases} \dfrac{\gamma(c) + \frac{1}{|\widehat{\mathcal{C}}(z)|} \sum_{d \notin \widehat{\mathcal{C}}(z)} \gamma(d)}{1 - \frac{1}{|\widehat{\mathcal{C}}(z)|} \sum_{d \notin \widehat{\mathcal{C}}(z)} \gamma(d)} & \text{if } c \in \widehat{\mathcal{C}}(z), \\ 0 & \text{if } c \notin \widehat{\mathcal{C}}(z). \end{cases} \tag{8.1}$$

Crucial for the optimality criterion is the interpretation of teachers as functions $\widetilde{T} \colon \Gamma \to \mathcal{X}(c^*)$ over information states. Such a teacher $\widetilde{T}$, when applied to an initial state $\gamma^{(0)} \in \Gamma$ and subsequently to all emerging states, yields a teacher $T \colon \mathbb{N} \to \mathcal{X}(c^*)$. Formally, $T(0) = \widetilde{T}(\gamma^{(0)})$, $\gamma^{(i+1)} = f(\gamma^{(i)}, T(i))$, and $T(i+1) = \widetilde{T}(\gamma^{(i+1)})$. We call T the *sequential teacher for* $\widetilde{T}$ *starting in* $\gamma^{(0)}$. Most important are the sequential teachers starting in $\gamma^{(init)}$. For generality, Patek considers series of policies and thus we have to consider series $\widetilde{T} = (\widetilde{T}_t)_{t \in \mathbb{N}}$ of teachers. Such a series is called *stationary* if all teachers in it are identical. We identify a stationary teacher series with the unique teacher it contains. A sequential teacher for a teacher series can be defined similarly as above for a single teacher.

Given a teacher series $(\widetilde{T}_t)_{t \in \mathbb{N}}$, the expected time for the learner $\mathcal{L}_1$ to reach the target when starting in $\gamma \in \Gamma$ is denoted by $G_{\widetilde{T}}(\gamma)$. This yields a function $G_{\widetilde{T}} \colon \Gamma \to \mathbb{R}$. We denote by $\mathcal{G}$ the set of all functions $G \colon \Gamma \to \mathbb{R}$ and by $\mathcal{G}_B$ the subset of $\mathcal{G}$ containing all bounded functions. Intuitively, $\mathcal{G}$ contains all functions that map each information state γ to the expected number of rounds for a learner starting in γ.

We also need two *dynamic programming operators* D and $D_{\widetilde{T}}$ mapping functions $G \colon \Gamma \to \mathbb{R}$ to functions with the same domain and codomain. We denote the application of D or $D_{\widetilde{T}}$ to a function G by DG or $D_{\widetilde{T}}G$; the value of the resulting function for an

argument $\gamma \in \Gamma$ is denoted by $[DG](\gamma)$ and $[D_{\widetilde{T}}G](\gamma)$. The operators are now defined as follows:

$$[D_{\widetilde{T}}G](\gamma) = 1 + G(f(\gamma, \widetilde{T}(\gamma))) \cdot \sum_{c,d \in \widehat{\mathcal{C}}} \gamma(c) \cdot p(c, \widetilde{T}(\gamma), d) \ ,$$

$$[DG](\gamma) = \min_{z \in \mathcal{X}(c^*)} \Bigg(1 + G(f(\gamma, z)) \cdot \sum_{c,d \in \widehat{\mathcal{C}}} \gamma(c) \cdot p(c, z, d) \Bigg).$$

The value $G(\gamma)$ can be thought of as expectation of the number of rounds to reach the target when the teaching process starts in information state γ. Roughly speaking, $[D_{\widetilde{T}}G](\gamma)$ is the expectation for the learner starting in γ under teacher $\widetilde{T}$, assuming that for all other states the expectations are given by G.

We denote the composition $D_{\widetilde{T}_0} \circ (D_{\widetilde{T}_1} \circ (\cdots \circ (D_{\widetilde{T}_{t-1}} \circ D_{\widetilde{T}_t}) \ldots))$ of finitely many operators by $D_{\widetilde{T}_0} D_{\widetilde{T}_1} \cdots D_{\widetilde{T}_t}$. The expected number of rounds for reaching c^* when $\mathcal{L}_1$ starts in some $\gamma^{(0)} \in \Gamma$ and is taught by the teacher series $(\widetilde{T}_t)_{t \in \mathbb{N}}$ is given by

$$G_{\widetilde{T}}(\gamma^{(0)}) = \liminf_{t \to \infty} [D_{\widetilde{T}_0} D_{\widetilde{T}_1} \cdots D_{\widetilde{T}_t} \mathbf{0}](\gamma^{(0)}).$$

An information state teacher series $(\widetilde{T}_t)_{t \in \mathbb{N}}$ is called *optimal* if and only if

$$G_{\widetilde{T}}(\gamma) = \inf_{\widetilde{U}} G_{\widetilde{U}}(\gamma)$$

for all $\gamma \in \Gamma$, where $\widetilde{U}$ ranges over all teacher series $(\widetilde{U}_t)_{t \in \mathbb{N}}$. This notion of optimality for information state teachers is compatible with our notion of optimality for sequential teachers.

Lemma 8.3 *Let $(\widetilde{T}_t)_{t \in \mathbb{N}}$ be an optimal series of information state teachers and let T be the sequential teacher for $\widetilde{T}$ starting in $\gamma^{(init)}$. Then T is an optimal sequential teacher and its teaching time is $G_{\widetilde{T}}(\gamma^{(init)})$.*

Proof. The behavior of the learner $\mathcal{L}_1$ under teacher series $\widetilde{T}$ is the same as under teacher T, from which it follows that the teaching time of T is $G_{\widetilde{T}}(\gamma^{(init)})$.

It remains to show that T has the minimal teaching time of all sequential teachers. Suppose for a contradiction that there is a sequential teacher U whose teaching time E_U is less than that of T. Let $(\widetilde{U}_t)_{t \in \mathbb{N}}$ be the information state teacher series with $\widetilde{U}_t(\gamma) = U(t)$ for all $t \in \mathbb{N}$ and all $\gamma \in \Gamma$. Then this information state teacher series has an expectation $G_{\widetilde{U}}(\gamma^{(init)}) = E_U$, which by assumption about U is less than $G_{\widetilde{T}}(\gamma^{(init)})$. But this is a contradiction because $\widetilde{T}$ is optimal. □

Assumption C

The statement of Lemma 8.5 below is Patek's Assumption C expressed in our terminology. Its proof shows that Assumption C holds in our teaching setting, regardless of the concept class $\mathcal{C}$ and the target $c^* \in \mathcal{C}$. Intuitively it demands that there is a successful stationary teacher series and that every non-successful series of teachers has an infinite teaching time. To show that there is a successful stationary teacher series we can always use a greedy teacher. A teacher is called greedy if in every round it maximizes the probability for the learner to reach the target in the next round.

Definition 8.4 *A teacher $\widetilde{T}$ for $c^* \in \mathcal{C}$ is called* greedy *iff for all $\gamma \in \Gamma$*

$$\widetilde{T}(\gamma) \in \operatorname*{argmax}_{z \in \mathcal{X}(c^*)} \sum_{c \in \widehat{\mathcal{C}}} \gamma(c) \cdot p(c, z, c^*) \ .$$

To formulate Assumption C we need one last notation. We denote by $\Pr^k(\gamma, \widetilde{T})$ the probability of reaching c^* within k rounds when the learner starting in $\gamma \in \Gamma$ is taught by teacher $\widetilde{T}_t$ in round t. More formally, let $\gamma = \gamma^{(0)} \in \Gamma$ be an information state and $\widetilde{T} = (\widetilde{T}_t)_{t \in \mathbb{N}}$ a series of teachers. Let $\delta^{(0)} \colon \mathcal{C} \cup \{init\} \to [0,1]$ be the probability distribution with

$$\delta^{(0)}(h) = \begin{cases} 0 & \text{if } h = c^*, \\ \gamma(h) & \text{if } h \neq c^*. \end{cases}$$

The δ- and γ-distributions during the teaching process under the teacher series $\widetilde{T}$ are then inductively defined as $\gamma^{(t+1)} = f(\gamma^{(t)}, \widetilde{T}_t(\gamma^{(t)}))$ and as

$$\delta^{(t+1)}(h) = \begin{cases} \delta^{(t)}(h) + \frac{1}{|\mathcal{C}(\widetilde{T}_t(\gamma^{(t)}))|} \cdot \sum\limits_{c \notin \mathcal{C}(\widetilde{T}_t(\gamma^{(t)}))} \delta^{(t)}(c) & \text{if } h \in \mathcal{C}(\widetilde{T}_t(\gamma^{(t)})), \\ 0 & \text{otherwise} \end{cases}$$

for all $h \in \mathcal{C} \cup \{init\}$. Then for all $k \in \mathbb{N}$ we define $\Pr^k(\gamma, \widetilde{T}) = \delta^{(k)}(c^*)$. Note that with δ and γ defined as above, we have $\gamma^{(t)}(h) = \delta^{(t)}(h)/(1 - \delta^{(t)}(c^*))$ for all $h \neq c^*$.

Lemma 8.5 *Let $\mathcal{C}$ be a concept class over X and $c^* \in \mathcal{C}$ be a target. Then:*

1. *There is a stationary teacher series $(\widetilde{T}_t)_{t \in \mathbb{N}}$ with $\lim\limits_{k \to \infty} \Pr^k(\gamma, \widetilde{T}) = 1$ for all $\gamma \in \Gamma$.*

2. *Every teacher series $(\widetilde{T}_t)_{t \in \mathbb{N}}$ that does not satisfy Condition 1 is such that a subsequence of*

$$\Big([D_{\widetilde{T}_0} D_{\widetilde{T}_1} \cdots D_{\widetilde{T}_t} \mathbf{0}](\gamma) \Big)_{t=0}^{\infty}$$

tends to infinity for some $\gamma \in \Gamma$.

Proof. 1. We show that the greedy teacher $\widetilde{T}$ satisfies this condition. Let $\gamma = \gamma^{(0)} \in \Gamma$ and let $\delta^{(t)}$ and $\gamma^{(t)}$ be defined as above for the definition of $\Pr^k(\cdot,\cdot)$.

In each round t, the teacher $\widetilde{T}$ picks an example z maximizing $\sum_{c\in\widehat{\mathcal{C}}} \gamma^{(t)}(c)\cdot p(c,z,c^*)$, which is the same as maximizing $\sum_{c\in\mathcal{C}\cup\{init\}} \delta^{(t)}(c)\cdot p(c,z,c^*)$.

In round t a probability mass of $1-\delta^{(t)}(c^*)$ is distributed among the non-target concepts. Therefore there is a concept $c' \neq c^*$ with $\delta^{(t)}(c') \geq (1-\delta^{(t)}(c^*))/|\mathcal{C}|$. Let z' be an example inconsistent with c'. Then $p(c',z',c^*) \geq 1/|\mathcal{C}|$ and therefore

$$\sum_{c\in\mathcal{C}\cup\{init\}} \delta^{(t)}(c)\cdot p(c,z',c^*) \geq \delta^{(t)}(c')\cdot p(c',z',c^*) \geq \frac{1-\delta^{(t)}(c^*)}{|\mathcal{C}|^2}.$$

Since $\widetilde{T}$ chooses an example that maximizes this sum, we also have for $z = \widetilde{T}(\gamma^{(t)})$ that

$$\sum_{c\in\mathcal{C}\cup\{init\}} \delta^{(t)}(c)\cdot p(c,z,c^*) \geq \frac{1-\delta^{(t)}(c^*)}{|\mathcal{C}|^2}.$$

Because $p(c,z,c^*) = 1/|\mathcal{C}(z)|$ for $c \notin \mathcal{C}(z)$ this sum also equals $\delta^{(t+1)}(c^*) - \delta^{(t)}(c^*)$ and therefore

$$1-\delta^{(t+1)}(c^*) \leq \left(1-\tfrac{1}{|\mathcal{C}|^2}\right)\cdot(1-\delta^{(t)}(c^*)).$$

Hence, $1-\delta^{(t)}(c^*) \to 0$ as $t\to\infty$ and the probability $\delta^{(t)}(c^*)$ tends to one. Since $\Pr^k(\gamma,\widetilde{T}) = \delta^{(k)}(c^*)$, this means that $\lim_{k\to\infty}\Pr^k(\gamma,\widetilde{T}) = 1$.

2. Let $\widetilde{T} = (\widetilde{T}_t)_t$ be a series that does not satisfy Condition 1. We show that the sequence $([D_{\widetilde{T}_0} D_{\widetilde{T}_1}\cdots D_{\widetilde{T}_t}\mathbf{0}](\gamma^{(init)}))_{t=0}^{\infty}$ tends to infinity.

Starting in the state $\gamma^{(0)} = \gamma^{(init)}$, the teacher series $(\widetilde{T}_t)_{t\in\mathbb{N}}$ generates a sequence of examples and states: $z^{(t)} = \widetilde{T}_t(\gamma^{(t)})$ and $\gamma^{(t+1)} = f(\gamma^{(t)}, z^{(t)})$. Let $T\colon \mathbb{N}\to\mathcal{X}(c^*)$ be the sequential teacher with $T(i) = z^{(i)}$ for all i. By assumption, $\Pr^t(\gamma^{(init)},\widetilde{T})$ does not converge to 1 for $t\to\infty$. According to the definition of $\Pr^t(\cdot,\cdot)$ that means that T is no successful teacher. Its expected teaching time is thus infinite. Therefore, also the expected teaching time for $(\widetilde{T}_t)_{t\in\mathbb{N}}$ is infinite. But this teaching time is also given by

$$\liminf_{t\to\infty}[D_{\widetilde{T}_0} D_{\widetilde{T}_1}\cdots D_{\widetilde{T}_t}\mathbf{0}](\gamma^{(init)})$$

which is therefore infinite, too. □

The Characterization

We can now state a first version of the optimality criterion, which is directly taken from Patek [61, 62].

Lemma 8.6 ([61, 62]) *Let $\mathcal{C}$ be a concept class and $c^* \in \mathcal{C}$ a target.*

1. *The operator D has a unique fixed point in $\mathcal{G}_B$, denoted G^*, that is, $DG^* = G^*$.*
2. *For every $G \in \mathcal{G}_B$ we have $D^k G \xrightarrow{pw} G^*$, where $\xrightarrow{pw}$ denotes pointwise convergence.*
3. *A (stationary) teacher $\widetilde{T}\colon \Gamma \to \mathcal{X}(c^*)$ is optimal for teaching c^* to $\mathcal{L}_1$ without feedback if and only if $D_{\widetilde{T}} G^* = G^*$.*

This criterion, directly adapted from Patek's criterion, requires us to find a $G\colon \Gamma \to \mathbb{R}$ and to define a teacher $\widetilde{T}\colon \Gamma \to \mathcal{X}(c^*)$. However, most of the states in Γ cannot be reached from the initial state $\gamma^{(init)}$ of the learner and it seems unnecessary to specify behavior of $\widetilde{T}$ for the unreachable states, too. As a matter of fact, it suffices to define G and $\widetilde{T}$ for the *reachable states* in Γ, which we denote by

$$\Gamma_0 = \{\gamma \in \Gamma \mid \exists t\ \exists z_0, \ldots, z_t\colon \gamma = f(\ldots f(f(\gamma^{(init)}, z_0), z_1) \ldots, z_t)\}.$$

We can now state the final version of our optimality criterion.

Corollary 8.7 *Let $\mathcal{C}$ be a concept class and $c^* \in \mathcal{C}$ a target. A teacher $\widetilde{T}\colon \Gamma_0 \to \mathcal{X}(c^*)$ is optimal if and only if there is a $G\colon \Gamma_0 \to \mathbb{R}$ such that $DG = G$ and $D_{\widetilde{T}} G = G$, where the operator D has to be restricted suitably to operate on functions $G\colon \Gamma_0 \to \mathbb{R}$.*

Proof. For the "if" part, let $\widetilde{T}$ be a teacher with expectations $G\colon \Gamma_0 \to \mathbb{R}$. Now, for all $\gamma \in \Gamma_0$:

$$\begin{aligned} G(\gamma) = [D_{\widetilde{T}} G](\gamma) &= 1 + \sum_{c,d \in \widehat{\mathcal{C}}} p(c, \widetilde{T}(\gamma), d) \cdot G(\gamma, f(\gamma, \widetilde{T}(\gamma))) \\ &= \min_{z \in \mathcal{X}(c^*)} \left[1 + \sum_{c,d \in \widehat{\mathcal{C}}} \gamma_c \cdot p(c, z, d) \cdot G(f(\gamma, z)) \right] = [DG](\gamma) \end{aligned}$$

in short: $D_{\widetilde{T}} G = DG = G$. To see this, suppose that the third equality would not hold. Then the teacher could be improved by setting $\widetilde{T}(\gamma)$ to the z minimizing the term in square brackets. The first equality holds because the $G(\gamma)$ values are the expectations of $\widetilde{T}$; the second and last equality are definitions of the respective operators.

For the "only if" part, let $G\colon \Gamma_0 \to \mathbb{R}$, and let $\widetilde{T}\colon \Gamma_0 \to \mathcal{X}(c^*)$ be a teacher such that $D_{\widetilde{T}} G = G = DG$. Let $G^* \in \mathcal{G}_B$ be the unique fixed point of D from Lemma 8.6, Item 1. We shall show that G^* is an extension of G to the domain $\Gamma \supset \Gamma_0$, from which we conclude that G describes the optimal expectations for information states in Γ_0.

Define G' as extension of G by G^*:

$$G'(\gamma) = \begin{cases} G(\gamma) & \text{if } \gamma \in \Gamma_0, \\ G^*(\gamma) & \text{if } \gamma \in \Gamma \setminus \Gamma_0. \end{cases}$$

As $G, G^* \in \mathcal{G}_B$, also $G' \in \mathcal{G}_B$ and by Lemma 8.6, Item 2, the sequence $D^k G'$ converges pointwise (that is, γ-wise) to G^*.

We then use that $[DG'](\gamma) = G'(\gamma) = G(\gamma)$ for all $\gamma \in \Gamma_0$. To see this consider

$$[DG'](\gamma) = \min_{z \in \mathcal{X}(c^*)} \left[1 + \sum_{c,d \in \widehat{\mathcal{C}}} \gamma_c \cdot p(c,z,d) \cdot G'(f(\gamma,z)) \right] \tag{8.2}$$

together with the fact that $f(\gamma, z) \in \Gamma_0$ for all z and thus $G'(\gamma, f(\gamma,z)) = G(\gamma, f(\gamma,z))$. Therefore, we can substitute $G()$ for $G'()$ in the right hand side of Equation (8.2), hence $[DG'](\gamma) = [DG](\gamma) = G(\gamma)$.

It follows that we also have for all $k \geq 1$ and $\gamma \in \Gamma_0$, $[D^k G'](\gamma) = G(\gamma)$ and therefore $G(\gamma) = \lim_{k\to\infty}[D^k G'](\gamma) = G^*(\gamma)$. But then $G'(\gamma) = G^*(\gamma)$ for all $\gamma \in \Gamma$. □

One advantage of using Γ_0 instead of Γ is that we have to consider only one state with $\gamma(\mathit{init}) > 0$, namely the initial state $\gamma^{(\mathit{init})}$. For illustration we apply Corollary 8.7 to the class $\mathcal{S}_n$.

Fact 8.8 *Let $c^* = [1,n] \in \mathcal{S}_n$ be the target concept. Then the teacher $T\colon \mathbb{N} \to \mathcal{X}(c^*)$ with $T(i) = (1 + (i \mod n), 1)$ is an optimal teacher for $[1,n]$ to the learner $\mathcal{L}_1$. Its teaching time is $1 + \frac{1}{2}(n-1)n$.*

Proof. The proof proceeds in several steps. First we define a teacher $\widetilde{T}\colon \Gamma_0 \to \mathcal{X}(c^*)$ and a function $G\colon \Gamma_0 \to \mathbb{R}$. Then we show that $DG = G$ and $D_{\widetilde{T}} = D$, from which we conclude that $\widetilde{T}$ is optimal. Finally we show that $\widetilde{T}$, when applied to $\gamma^{(\mathit{init})}$, generates the same example sequence as T.

For a $\gamma \in \Gamma$ and $i \in [1,n]$ we set as shortcut $\gamma_i := \gamma(c)$ for $c = [1,n] \setminus \{i\}$. A positive example $(x,1)$ is inconsistent only with the concept $[1,n] \setminus \{x\}$. Teaching $(x,1)$ in a state $\gamma \neq \gamma^{(\mathit{init})}$ results in a state $f(\gamma,(x,1)) = \hat{\gamma}$ with $\hat{\gamma}_i = \frac{\gamma_i + \gamma_x/n}{1-\gamma_x/n}$ for $i \neq x$, and $\hat{\gamma}_x = 0$. For $\gamma = \gamma^{(\mathit{init})}$ we have $\hat{\gamma}_i = 1/(n-1)$ for all $i \neq x$ and $\hat{\gamma}_x = 0$.

We define $\widetilde{T}$ to be a greedy teacher. If there are several equally "greedy" examples, $\widetilde{T}$ picks the one with least instance. As every example is inconsistent with exactly one concept, $\widetilde{T}$ greedily picks an example that is inconsistent with a most probable hypothesis.

For defining G, let $\gamma \in \Gamma_0 \setminus \{\gamma^{(\mathit{init})}\}$ and assume without loss of generality that $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_n$. Let $F = \frac{(n-1)n}{2}$. Then we define

$$G(\gamma) = F + \sum\nolimits_{i=1}^{n} \gamma_i \cdot i \qquad \text{and} \qquad G(\gamma^{(\mathit{init})}) = F + 1.$$

Next we show $DG = G$. Let $\gamma \in \Gamma_0 \setminus \{\gamma^{(\mathit{init})}\}$ again with $\gamma_1 \geq \cdots \geq \gamma_n$. We have to show that $[DG](\gamma) = G(\gamma)$, in other words that $1 + \min_{(x,1)\in\mathcal{X}} G(f(\gamma,(x,1))) \cdot \sum_{c,d\in\widehat{\mathcal{C}}} p(c,(x,1),d) = G(\gamma)$. Since $\sum_{c,d\in\widehat{\mathcal{C}}} p(c,(x,1),d) = 1 - \gamma_x/n$ this means that

$$1 + \min_{(x,1)\in\mathcal{X}} G(f(\gamma,(x,1))) \cdot (1 - \gamma_x/n) = G(\gamma). \tag{8.3}$$

Let $z = (x, 1) \in \mathcal{X}$ and $\hat{\gamma} = f(\gamma, z)$. Then $\hat{\gamma}_1 \geq \cdots \geq \hat{\gamma}_{z-1} \geq \hat{\gamma}_{z+1} \geq \cdots \geq \hat{\gamma}_n \geq \hat{\gamma}_z = 0$. The expression to be minimized is

$$\begin{aligned}\left(1 - \frac{\gamma_x}{n}\right) \cdot G(\hat{\gamma}) &= \left(1 - \frac{\gamma_x}{n}\right) \cdot \left(F + \sum_{i \leq x-1} i \cdot \hat{\gamma}_i + \sum_{i \geq x+1} (i-1) \cdot \hat{\gamma}_i\right) \\ &= \left(1 - \frac{\gamma_x}{n}\right) \cdot \left(F + \sum_{i \leq x-1} i \cdot \frac{\gamma_i + \gamma_x/n}{1 - \gamma_x/n} + \sum_{i \geq x+1} (i-1) \cdot \frac{\gamma_i + \gamma_x/n}{1 - \gamma_x/n}\right) \\ &= F + \sum_{i=1}^{n} i\gamma_i - \left(x \cdot \gamma_x + \sum_{i \geq x+1} \gamma_i\right). \qquad (*)\end{aligned}$$

From $\gamma_1 \geq \cdots \geq \gamma_n$, it follows $1 \cdot \gamma_1 + \sum_{i \geq 2} \gamma_i \geq 2 \cdot \gamma_2 + \sum_{i \geq 3} \gamma_i \geq \cdots \geq n \cdot \gamma_n$. This means that the expression $(*)$ is minimal for $x = 1$, or $\gamma_x = \gamma_1$. Setting $x = 1$ yields $\min_{(x,1) \in \mathcal{X}} G(f(\gamma, (x, 1))) \cdot (1 - \gamma_x/n) = F - 1 + \sum_{i=1}^{n} i\gamma_i = G(\gamma) - 1$ and thus Equation (8.3) is satisfied.

It remains to show $[DG](\gamma^{(init)}) = G(\gamma^{(init)})$. For all examples $(x, 1) \in \mathcal{X}$ we have

$$\begin{aligned}[DG](\gamma^{(init)}) &= 1 + (1 - \tfrac{1}{n}) \cdot G(f(\gamma^{(init)}, (x, 1))) \\ &= 1 + (1 - \frac{1}{n}) \cdot \left(F + \sum_{i=1}^{n-1} i \frac{1/n}{1 - 1/n}\right) = 1 + \frac{n-1}{n} \cdot \left(F + \frac{1}{n-1} \cdot \frac{n(n-1)}{2}\right) \\ &= 1 + \frac{n(n-1)}{2} = F + 1 = G(\gamma^{(init)}).\end{aligned}$$

It follows that $[DG](\gamma) = G(\gamma)$ for all $\gamma \in \Gamma_0$. Moreover, the teacher $\widetilde{T}$ always picks the example $(x, 1)$ minimizing the term in Equation (8.3), thus $D_{\widetilde{T}} G = G$ and $\widetilde{T}$ is optimal according to Corollary 8.7.

The teacher $\widetilde{T}$, when started in $\gamma^{(init)}$, generates the same sequence of examples as the sequential teacher T. By the definition of $\widetilde{T}$ we have $\widetilde{T}(\gamma^{(init)}) = (1, 1)$ and for $\gamma \neq \gamma^{(init)}$ with $\gamma_1 \geq \cdots \geq \gamma_n$ (without loss of generality) $\widetilde{T}$ chooses example $(1, 1)$ and the next information state is $\hat{\gamma}$ with $\hat{\gamma}_2 \geq \cdots \geq \hat{\gamma}_n \geq \hat{\gamma}_1 = 0$. Therefore, $\widetilde{T}$ chooses $(2, 1)$ as next example and so on. □

The previous proof shows how Corollary 8.7 is to be applied in order to prove a given sequential teacher optimal. First, a information state teacher has to be defined which is then shown to be optimal by applying Corollary 8.7 directly. Afterwards one has to show that the information state teacher, when applied initially to $\gamma^{(init)}$, yields the same example sequence as the sequential teacher. Finding a suitable information state teacher to perform such a proof seems to be a difficult task. In Fact 8.8 we could exploit the fact that the greedy teacher was suitable. But in general, greedy teachers do not have this property, as we shall show in Section 8.5.

8.3 Computing the Optimal Teaching Time

From the last section we know that for all targets c^* and all concept classes $\mathcal{C}$ there is always an optimal sequential teacher, even though we do not know how to find one effectively. The method we derived for checking whether a given sequential teacher is optimal requires us to find a suitable information state teacher first. And we do not know a general method for doing that either. In this section we investigate the problem of finding the optimal teaching time, rather than finding a teacher that achieves this optimum. More precisely, we study the complexity of the following decision problem.

Definition 8.9 *We call the following problem* `OPT-TEACHINGTIME`.
Input: *Concept class $\mathcal{C}$, concept $c^* \in \mathcal{C}$, rational number F.*
Question: *Is $E_1^-(c^*, \mathcal{C}) \leq F$?*

In the more general setting of USSPPs the analogous problem is undecidable (see Madani, Hanks, and Condon [52] and Blondel and Canterini [20]). This can be seen as evidence for the undecidability of `OPT-TEACHINGTIME`. On the other hand, USSPPs differ from our model in some aspects related to computational complexity. For example, deciding whether there is a teacher with at least a given success probability is easy (because there is always one), whereas the analogous problem for USSPPs is undecidable [52, 20].

Although the decidability of `OPT-TEACHINGTIME` is open, we can at least show that it is $\mathcal{NP}$-hard. So even if there is an algorithm, it is presumably inefficient. The $\mathcal{NP}$-hardness proof for `OPT-TEACHINGTIME` is lengthy, so we proceed in several steps.

1. We give a polynomial time reduction from the `EXACT-3-COVERING` (`X3C`) problem. The instances of `OPT-TEACHINGTIME` produced by the reduction are such that all examples for the target are inconsistent with exactly three concepts (see Figure 8.3).

2. We consider concept classes and targets with the property that all examples for the target are inconsistent with exactly three concepts and for all sets of three concepts there is an example that is inconsistent with these three concepts. We show that these classes have an optimal teaching time of $1 + \frac{3}{2}n(n-1)$. We also show that all optimal teachers for these classes are greedy.

3. We show that for a class resulting from the reduction there is a teacher with teaching time $1 + \frac{3}{2}n(n-1)$ if and only if the class was built for a positive `X3C` instance.

Lemma 8.10 *Algorithm 8.1 computes* `OPT-TEACHINGTIME` *instances from instances of* `X3C` *in polynomial time.*

X3C

$B = \{1, 2, 3, 4, 5, 6\}$,
$A_1 = \{2, 4, 5\}$,
$A_2 = \{1, 3, 5\}$,
$A_3 = \{1, 3, 6\}$

OPT-TEACHINGTIME

	x_1	x_2	x_3	y_1	y_2	y_3	y_4	y_5	y_6
c^*	1	1	1	1	1	1	1	1	1
c_1	1	0	0	0	1	1	1	1	1
c_2	0	1	1	1	0	1	1	1	1
c_3	1	0	0	1	1	0	1	1	1
c_4	0	1	1	1	1	1	0	1	1
c_5	0	0	1	1	1	1	1	0	1
c_6	1	1	0	1	1	1	1	1	0

Figure 8.3 *Illustration of the reduction from* X3C *to* OPT-TEACHINGTIME*. Every example is inconsistent with exactly three concepts; $y_1, \ldots, y_6$ are "dummy" instances making all rows unique. The examples $(x_1, 1), (x_3, 1)$ have the X3C property.*

Input: Number $n \in \mathbb{N}$, sets $A_1, \ldots, A_m \subseteq [1, 3n]$ with $|A_i| = 3$.

1 $X := \{x_1, \ldots, x_m\} \cup \{y_1, \ldots, y_{3n}\}$

2 $c_j := \{x_i \mid j \notin A_i\} \cup \{y_i \mid i \neq j\}$ for $j = 1, \ldots, 3n$

3 $c^* := X$

4 $\mathcal{C} := \{c^*, c_1, \ldots, c_{3n}\}$

5 **output** $\langle \mathcal{C}, c^*, 1 + \frac{3}{2}n(n-1) \rangle$

Algorithm 8.1. *Algorithm computing* OPT-TEACHINGTIME *instances from* X3C *instances in polynomial time. See Figure 8.3 for an example.*

Proof. The computation of $\mathcal{C}$ is straightforward and $\mathcal{C}$ is represented by a $(3n + 1) \times (n + m)$ matrix which is polynomial in the input size. The value $1 + \frac{3}{2}n(n-1)$ is polynomially computable and representable, too. □

We call a concept class $\mathcal{C}$ resulting as output of Algorithm 8.1 a *positive* or *negative X3C class* depending on whether the X3C instance was positive or negative.

An X3C class is positive if and only if there are examples $z_1, \ldots, z_n \in \mathcal{X}(c^*)$ such that the sets $\mathcal{C} \setminus \mathcal{C}(z_j)$ for $j = 1, \ldots, n$ are pairwise disjoint and $\bigcup_j (\mathcal{C} \setminus \mathcal{C}(z_j)) = \mathcal{C} \setminus \{c^*\}$.

If $A_1, \ldots, A_m$ consists of all $m = \binom{3n}{3}$ subsets of B we call the class a *full X3C class*. Every full X3C class is a positive X3C class.

Of all X3C classes, the full X3C classes are easiest to analyze because of their intrinsic symmetries. Moreover, the optimal teachers are just the greedy teachers, which simplifies the application of our optimality criterion. In general, for X3C classes a greedy teacher need not be optimal.

The proof of the following lemma bears some similarity with the proof of Fact 8.8. Indeed, co-singleton classes and X3C classes are both special cases of classes in which all examples for the target are inconsistent with a fixed number of concepts.

Lemma 8.11 *Let $n \in \mathbb{N}$, $n \geq 2$. Let $\mathcal{C}$ be a full X3C class containing $3n+1$ concepts, and let c^* be the concept containing all instances. Then a teacher $T\colon \Gamma_0 \to \mathcal{X}(c^*)$ is optimal if and only if $\widetilde{T}$ is greedy. The expected teaching time, when starting in $\gamma^{(init)}$, is $1 + \frac{3}{2}n(n-1)$.*

Proof. We first prove that the function $G\colon \Gamma_0 \to \mathbb{R}$ defined below states the optimal teaching times by showing $DG = G$. Then we show that the only teachers having this expectation are the greedy ones.

To ease notation we identify each concept in $\mathcal{C} \setminus \{c^*\}$ with a number in $[1, 3n]$. A state $\gamma \in \Gamma_0$ is then a function $[1, 3n] \cup \{init\} \to [0, 1]$ with values denoted by $\gamma_1, \ldots, \gamma_{3n}, \gamma_{init}$. As a shortcut we shall also use γ_z to denote $\sum_{i \notin \mathcal{X}(z)} \gamma_i$ for an example $z \in \mathcal{X}(c^*)$.

Let $F = \frac{3}{2}n(n-1)$, and define G for $\gamma \in \Gamma_0 \setminus \{\gamma^{(init)}\}$ with $\gamma_1 \geq \gamma_2 \geq \gamma_{3n} \geq \gamma_{init} = 0$ by

$$G(\gamma) = F + \sum_{i=1}^{3n} \gamma_i \cdot \lceil \tfrac{i}{3} \rceil$$

and $G(\gamma^{(init)}) = F + 1$. Intuitively, each γ_i is multiplied with a factor depending on the rank of i in the sorted list of probabilities.

We distinguish two kinds of examples: useful examples and dummy examples. An example z is called *useful* iff $|\mathcal{C} \setminus \mathcal{C}(z)| = 3$. The other examples z, for which $|\mathcal{C} \setminus \mathcal{C}(z)| = 1$, are called *dummy* examples (see Figure 8.3). The function f for the follow-up information states takes two forms depending on whether the example is a useful one or a dummy. For the former we have $f(\gamma, z) = \hat{\gamma} \in \Gamma_0$ with

$$\hat{\gamma}_i = \begin{cases} (\gamma_i + \gamma_z/(3n-2)) \;/\; (1 - \gamma_z/(3n-2)) & \text{if } i \in \mathcal{C}(z), \\ 0 & \text{otherwise.} \end{cases}$$

A dummy example is only inconsistent with one concept, thus $3n-2$ is replaced by $3n$:

$$\hat{\gamma}_i = \begin{cases} (\gamma_i + \gamma_z/(3n)) \;/\; (1 - \gamma_z/(3n)) & \text{if } i \in \mathcal{C}(z), \\ 0 & \text{otherwise.} \end{cases}$$

We are now going to show $[DG](\gamma) = G(\gamma)$ for $\gamma \in \Gamma_0 \setminus \{\gamma^{(init)}\}$. Without loss of generality, we assume that $\gamma_1 \geq \cdots \geq \gamma_{3n}$. We are required to determine the example(s) z minimizing

$$G(f(\gamma, z)) \cdot \sum_{c,d \in \mathcal{C} \setminus \{c^*\}} p(c, z, d).$$

We shall first determine the useful examples z minimizing this term and then show that dummy examples z cannot make that value smaller. For useful examples z we

have $\sum_{c,d\in\widehat{C}} p(c,z,d) = 1-\gamma_z/(3n-2)$. We denote the three hypotheses with which z is inconsistent by $u<v<w\in[1,3n]$. Then for $\hat{\gamma}=f(\gamma,z)$ it follows

$$\hat{\gamma}_1 \geq \cdots \geq \hat{\gamma}_{u-1} \geq \hat{\gamma}_{u+1} \geq \cdots \geq \hat{\gamma}_{v-1} \geq \hat{\gamma}_{v+1} \geq \cdots \geq \hat{\gamma}_{w-1} \geq \hat{\gamma}_{w+1} \geq \ldots \hat{\gamma}_{3n} > \hat{\gamma}_u = \hat{\gamma}_v = \hat{\gamma}_w = 0.$$

In other words, u, v, and w are put at the end of the "probability list" whereas the indices between u and v rise one rank, those between v and w rise two ranks, and those between w and $3n$ three ranks.

Now, setting $g := \gamma_z/(3n-2)$, we get

$$\begin{aligned} G(\hat{\gamma}) &= F + \sum_{i=1}^{u-1} \hat{\gamma}_i \lceil \tfrac{i}{3} \rceil + \sum_{i=u+1}^{v-1} \hat{\gamma}_i \lceil \tfrac{i-1}{3} \rceil + \sum_{i=v+1}^{w-1} \hat{\gamma}_i \lceil \tfrac{i-2}{3} \rceil + \sum_{i=w+1}^{3n} \hat{\gamma}_i \lceil \tfrac{i-3}{3} \rceil \\ &= F + \sum_{i=1}^{u-1} \tfrac{g+\gamma_i}{1-g} \lceil \tfrac{i}{3} \rceil + \sum_{i=u+1}^{v-1} \tfrac{g+\gamma_i}{1-g} \lceil \tfrac{i-1}{3} \rceil + \sum_{i=v+1}^{w-1} \tfrac{g+\gamma_i}{1-g} \lceil \tfrac{i-2}{3} \rceil + \sum_{i=w+1}^{3n} \tfrac{g+\gamma_i}{1-g} \lceil \tfrac{i-3}{3} \rceil, \end{aligned}$$

and the term to minimize is $(1-g)G(\hat{\gamma})$, that is,

$$(1-g)F + \sum_{i=1}^{u-1} (g+\gamma_i) \lceil \tfrac{i}{3} \rceil + \sum_{i=u+1}^{v-1} (g+\gamma_i) \lceil \tfrac{i-1}{3} \rceil + \sum_{i=v+1}^{w-1} (g+\gamma_i) \lceil \tfrac{i-2}{3} \rceil + \sum_{i=w+1}^{3n} (g+\gamma_i) \lceil \tfrac{i-3}{3} \rceil.$$

Multiplying out $(g+\gamma_i)$ we get a sum of terms containing g, namely

$$(1-g)F + \sum_{i=1}^{u-1} g \lceil \tfrac{i}{3} \rceil + \sum_{i=u+1}^{v-1} g \lceil \tfrac{i-1}{3} \rceil + \sum_{i=v+1}^{w-1} g \lceil \tfrac{i-2}{3} \rceil + \sum_{i=w+1}^{3n} g \lceil \tfrac{i-3}{3} \rceil, \tag{8.4}$$

and of terms containing no g, namely

$$\sum_{i=1}^{u-1} \gamma_i \lceil \tfrac{i}{3} \rceil + \sum_{i=u+1}^{v-1} \gamma_i \lceil \tfrac{i-1}{3} \rceil + \sum_{i=v+1}^{w-1} \gamma_i \lceil \tfrac{i-2}{3} \rceil + \sum_{i=w+1}^{3n} \gamma_i \lceil \tfrac{i-3}{3} \rceil. \tag{8.5}$$

Claim 1: The term (8.4) has a value of F.

Proof. By shifting the indices in the last three summations, we get for (8.4):

$$\begin{aligned} &(1-g)F + \sum_{i=1}^{u-1} g \lceil \tfrac{i}{3} \rceil + \sum_{i=u}^{v-2} g \lceil \tfrac{i}{3} \rceil + \sum_{i=v-1}^{w-3} g \lceil \tfrac{i}{3} \rceil + \sum_{i=w-2}^{3n-3} g \lceil \tfrac{i}{3} \rceil = (1-g)F + g \cdot \sum_{i=1}^{3n-3} \lceil \tfrac{i}{3} \rceil \\ &= (1-g)F + g\tfrac{3}{2}n(n-1) = (1-g)F + gF = F\ . \end{aligned}$$

□ Claim

For a useful example z the value γ_z is maximal if and only if u, v, w satisfy the following three *greedy conditions*:

$$(G1) \quad u = \max_{i \in [1,3n]} \gamma_i \ , \qquad (G2) \quad v = \max_{\substack{i \in [1,3n] \\ i \neq u}} \gamma_i \ , \qquad (G3) \quad w = \max_{\substack{i \in [1,3n] \\ i \neq u,v}} \gamma_i \ .$$

The term (8.5) can be rewritten as

$$\sum_{i=1}^{3n} \gamma_i \lceil \tfrac{i}{3} \rceil - \left(\sum_{i \in I_{u,v,w}} \gamma_i \lceil \tfrac{i}{3} \rceil + \gamma_u \lceil \tfrac{u}{3} \rceil + \gamma_v \lceil \tfrac{v}{3} \rceil + \gamma_w \lceil \tfrac{w}{3} \rceil \right)$$

where

$$I_{u,v,w} = \{ i \in [1,3n] \mid \begin{array}{l} (u < i < v \wedge \lceil \frac{i-1}{3} \rceil < \lceil \frac{i}{3} \rceil) \\ \text{or } (v < i < w \wedge \lceil \frac{i-2}{3} \rceil < \lceil \frac{i}{3} \rceil) \\ \text{or } (w < i \leq 3n) \} \end{array}$$

is the set of all hypotheses that are multiplied with a different factor in γ than in $\hat{\gamma}$. We denote the term in parentheses by $K_{u,v,w}$. This value has some monotonicity properties, as shown in the next claim.

Claim 2:

(a) For all $1 \leq u < v < w < 3n$: $K_{u,v,w} \geq K_{u,v,w+1}$ and proper inequality holds if and only if $\gamma_w > \gamma_{w+1}$.

(b) For all $1 \leq u < v < 3n-1$: $K_{u,v,v+1} \geq K_{u,v+1,v+2}$ and proper inequality holds if and only if $\gamma_v > \gamma_{v+2}$.

(c) For all $1 \leq u < 3n-2$: $K_{u,u+1,u+2} \geq K_{u+1,u+2,u+3}$ and proper inequality holds if and only if $\gamma_u > \gamma_{u+3}$.

Proof. (a) We distinguish two cases:

Case 1: $w \bmod 3 = 1$ or $w \bmod 3 = 2$.

Then $I_{u,v,w} = I_{u,v,w+1}$ and $K_{u,v,w} - K_{u,v,w+1} = \lceil \frac{w}{3} \rceil \gamma_w - \lceil \frac{w+1}{3} \rceil \gamma_{w+1} = \lceil \frac{w}{3} \rceil (\gamma_w - \gamma_{w+1}) \geq 0$, which proves Part (a) for Case 1.

Case 2: $w \bmod 3 = 3$.

Then $I_{u,v,w} = I_{u,v,w+1} \cup \{w+1\}$ and $K_{u,v,w} - K_{u,v,w+1} = \gamma_{w+1} + \lceil \frac{w}{3} \rceil \gamma_w - \lceil \frac{w+1}{3} \rceil \gamma_{w+1} = (\lceil \frac{w}{3} \rceil + 1)\gamma_w - (\lceil \frac{w}{3} \rceil + 1)\gamma_{w+1} = \lceil \frac{w}{3} \rceil (\gamma_w + \gamma_{w+1}) \geq 0$, which proves Part (a) for Case 2.

(b) We distinguish two cases:

Case 1: $v \bmod 3 = 1$.

Then $I_{u,v,v+1} = I_{u,v+1,v+2}$ and $K_{u,v,v+1} - K_{u,v+1,v+2} = \lceil \frac{v}{3} \rceil \gamma_v + \lceil \frac{v+1}{3} \rceil \gamma_{v+1} - \lceil \frac{v+1}{3} \rceil \gamma_{v+1} - \lceil \frac{v+2}{3} \rceil \gamma_{v+2} = \lceil \frac{v}{3} \rceil (\gamma_v - \gamma_{v+2}) \geq 0$, which proves Part (b) for Case 1.

Case 2: $v \bmod 3 = 2$ or $v \bmod 3 = 3$.

Then $I_{u,v,v+1} = I_{u,v+1,v+2} \cup \{v+2\}$ and $K_{u,v,v+1} - K_{u,v+1,v+2} = \gamma_{v+2} + \lceil \frac{v}{3} \rceil \gamma_v + \lceil \frac{v+1}{3} \rceil \gamma_{v+1} - \lceil \frac{v+1}{3} \rceil \gamma_{v+1} - \lceil \frac{v+2}{3} \rceil \gamma_{v+2} = \lceil \frac{v}{3} \rceil (\gamma_v - \gamma_{v+2}) \geq 0$, which proves Part (b) for Case 2.

(c) We have $I_{u,u+1,u+2} = I_{u+1,u+2,u+3} \cup \{u+3\}$. Therefore $K_{u,u+1,u+2} - K_{u+1,u+2,u+3} = \gamma_{u+3} + \lceil \frac{u}{3} \rceil \gamma_u + \lceil \frac{u+1}{3} \rceil \gamma_{u+1} + \lceil \frac{u+2}{3} \rceil \gamma_{u+2} - \lceil \frac{u+1}{3} \rceil \gamma_{u+1} - \lceil \frac{u+2}{3} \rceil \gamma_{u+2} - \lceil \frac{u+3}{3} \rceil \gamma_{u+3} = \lceil \frac{u}{3} \rceil (\gamma_u - \gamma_{u+3})$, which proves Part (c). □ Claim 2

It is now possible to characterize the u, v, w that maximize $K_{u,v,w}$. The maximum value of $K_{u,v,w}$ is $K_{1,2,3}$, which can be seen as follows. Claim 2(c) yields $K_{u,v,w} \leq K_{u,v,w-1} \leq \cdots \leq K_{u,v,v+1}$, then Claim 2(b) leads to $K_{u,v,v+1} \leq K_{u,v-1,v} \leq \cdots \leq K_{u,u+1,u+2}$, and finally Claim 2(a) shows $K_{u,u+1,u+2} \leq K_{u-1,u,u+1} \leq \cdots \leq K_{1,2,3}$. Thus, $K_{u,v,w} \leq K_{1,2,3}$ for all $u < v < w$.

Claim 3: $K_{u,v,w}$ is maximal if and only if $u < v < w$ satisfy the greedy conditions.

Proof. For the "if" direction, assume $u < v < w$ satisfying (G1), (G2), and (G3). We distinguish four cases.

Case 1: $\gamma_1 > \gamma_2 > \gamma_3$.

Then $u = 1, v = 2, w \geq 3$ and $\gamma_3 = \cdots = \gamma_w$. Applying Claim 2(a) yields $K_{u,v,w} = K_{1,2,w} = K_{1,2,w-1} = \cdots = K_{1,2,3}$.

Case 2: $\gamma_1 = \gamma_2 > \gamma_3$.

Works exactly like Case 1.

Case 3: $\gamma_1 > \gamma_2 = \gamma_3$.

Then $u = 1, v \geq 2, w > v$ and $\gamma_2 = \cdots = \gamma_v = \cdots = \gamma_w$. Applying Claim 2(a) yields $K_{u,v,w} = K_{1,v,w} = K_{1,v,v+1}$. Then using Claim 2(b) we get $K_{1,v,v+1} = K_{1,v-1,v} = \cdots = K_{1,2,3}$.

Case 4: $\gamma_1 = \gamma_2 = \gamma_3$.

Then $\gamma_1 = \cdots = \gamma_w$ and from Claim 2(a) we get $K_{u,v,w} = K_{u,v,v+1}$. Then Claim 2(b) yields $K_{u,v,v+1} = K_{u,v-1,v} = \cdots = K_{u,u+1,u+2}$ and finally Claim 2(a): $K_{u,u+1,u+2} = K_{u-1,u,u+1} = \cdots = K_{1,2,3}$.

For the "only if" direction, let $K_{u,v,w} = K_{1,2,3}$ and suppose for a contradiction that not all greedy conditions hold.

Case 1: $\neg(G1)$.

Then $\gamma_u < \max_i \gamma_i$ and there is an $s < u$ with $\gamma_s = \max_i \gamma_i$, hence $\gamma_s > \gamma_u$. Then there is also a t with $s \leq t < u$ such that $\gamma_s \geq \gamma_t > \gamma_{t+1} \geq \gamma_u$. Using Claim 2(a) we conclude $K_{u,v,w} \leq K_{u,u+1,u+2} \leq K_{t+1,t+2,t+3} < K_{t,t+1,t+2}$ (the strict inequality holds because $\gamma_t > \gamma_{t+3}$). Therefore $K_{u,v,w}$ is not maximal, a contradiction.

Case 2: $(G1) \wedge \neg(G2)$.

Then $\gamma_u = \max_i \gamma_i$, but $\gamma_v < \max_{i \neq u} \gamma_i$ and there is an s with $u < s < v$ such that $\gamma_s > \gamma_v$. Then there is a t with $s \leq t < v$ such that $\gamma_s \geq \gamma_t > \gamma_{t+1} \geq \gamma_v$. Applying Claim 2(a) yields $K_{u,v,w} \leq K_{u,v,v+1}$ and from Claim 2(b) it follows $K_{u,v,v+1} \leq K_{u,t+1,t+2} < K_{u,t,t+1}$ (strict inequality because $\gamma_t > \gamma_{t+2}$). Thus, $K_{u,v,w}$ is not maximal, a contradiction.

Case 3: $(G1) \wedge (G2) \wedge \neg(G3)$.

Then $\gamma_w < \max_{i \neq u,v}$ and there is an s with $v < s < w$ such that $\gamma_s > \gamma_w$ and a t with $s \leq t < w$ such that $\gamma_s \geq \gamma_t > \gamma_{t+1} \geq \gamma_w$ It follows by Claim 2(a) that $K_{u,v,w} \leq K_{u,v,t+1} < K_{u,v,t}$. Thus $K_{u,v,w}$ is not maximal, a contradiction. □ Claim 3

Claim 3 implies that an optimal teacher has to choose an example z maximizing γ_z. On the other hand, every such example minimizes $(1-g)G(\hat{\gamma})$. This minimal value is

$$(1-g)G(\hat{\gamma}) = F - 1 + \sum_{i=1}^{3n} \lceil \tfrac{i}{3} \rceil \gamma_i = G(\gamma) - 1.$$

We still have to show that no dummy example reaches this minimal value. Let z be an example inconsistent with only one concept, say $u \in [1, 3n]$. Then for $\hat{\gamma} = f(\gamma, z)$:

$$\hat{\gamma}_1 \geq \cdots \geq \hat{\gamma}_{u-1} \geq \hat{\gamma}_{u+1} \geq \cdots \geq \hat{\gamma}_{3n} > \hat{\gamma}_u = 0$$

and $\sum_{c,d \in \widehat{\mathcal{C}}} p(c,z,d) = 1 - \gamma_z/(3n) =: g$. Therefore

$$\begin{aligned} & G(f(\gamma,z)) \cdot \sum_{c,d \in \widehat{\mathcal{C}}} p(c,z,d) \\ &= G(\hat{\gamma}) \cdot (1-g) \\ &= (1-g)\left(F + \sum_{i=1}^{u-1} \hat{\gamma}_i \cdot \lceil \tfrac{i}{3} \rceil + \sum_{i=u+1}^{3n} \hat{\gamma}_i \cdot \lceil \tfrac{i-1}{3} \rceil \right) \end{aligned}$$

$$
\begin{aligned}
&= (1-g)F + \sum_{i=1}^{u-1}(\gamma_i+g)\lceil\tfrac{i}{3}\rceil + \sum_{i=u+1}^{3n}(\gamma_i+g)(\lceil\tfrac{i}{3}\rceil-1)\\
&= (1-g)F + \sum_{i=1}^{u-1}\gamma_i\lceil\tfrac{i}{3}\rceil + \sum_{i=u+1}^{3n}\gamma_i(\lceil\tfrac{i}{3}\rceil-1) + g\sum_{i\neq u}\lceil\tfrac{i}{3}\rceil\\
&= (1-g)F + \sum_{i=1}^{3n}\lceil\tfrac{i}{3}\rceil\gamma_i - \left(\lceil\tfrac{u}{3}\rceil\gamma_u + \sum_{\substack{i>u\\ i \bmod 3=1}}\gamma_i\right) + g\cdot\left(\tfrac{3}{2}n(n+1)-n\right)\\
&= F + \sum_{i=1}^{3n}\lceil\tfrac{i}{3}\rceil\gamma_i - \left(\lceil\tfrac{u}{3}\rceil\gamma_u + \sum_{\substack{i>u\\ i \bmod 3=1}}\gamma_i\right) + g\cdot\left(\tfrac{3}{2}n(n+1)-F-n\right)\\
&= F + \sum_{i=1}^{3n}\lceil\tfrac{i}{3}\rceil\gamma_i - \left(\lceil\tfrac{u}{3}\rceil\gamma_u + \sum_{\substack{i>u\\ i \bmod 3=1}}\gamma_i\right) + g\cdot 2n\ .
\end{aligned}
$$

The expression in parentheses is at most $u\gamma_u + \sum_{i=u+1}^{3n}\gamma_i \leq \sum_{i=1}^{3n}\gamma_i \leq 1$. The expression $g\cdot 2n = \frac{2}{3}\gamma_z$ is positive. Therefore for all dummy examples z:

$$
G(f(\gamma,z))\cdot\sum_{c,d\in\widehat{\mathcal{C}}}p(c,z,d) > F - 1 + \sum_{i=1}^{3n}\lceil\tfrac{i}{3}\rceil\gamma_i
$$

where the right hand side is the minimum attained by useful examples, as proven above. We conclude that in a state $\gamma \neq \gamma^{(init)}$ an optimal teacher chooses the example greedily and that every greedy choice is optimal.

To finish the proof we have to show a similar result for the initial state $\gamma = \gamma^{(init)}$. Let us first consider useful examples. Without loss of generality, we assume z with $\mathcal{C}\setminus\mathcal{C}(z) = \{1,2,3\}$. The following state $\hat{\gamma}$ satisfies

$$
g := \tfrac{1}{3n-2} = \hat{\gamma}_4 = \cdots = \hat{\gamma}_{3n} > \hat{\gamma}_1 = \hat{\gamma}_2 = \hat{\gamma}_3 = 0\ .
$$

It follows with $\sum_{c,d\in\widehat{\mathcal{C}}}p(c,z,d) = (1-g)$ and $\gamma_i = 0$ for $i\in[1,3n]$:

$$
\begin{aligned}
G(f(\gamma,z))\cdot\sum_{c,d\in\widehat{\mathcal{C}}}p(c,z,d) &= G(\hat{\gamma})\cdot(1-g)\\
&= (1-g)F + \sum_{i=4}^{3n}(\gamma_i+g)\lceil\tfrac{i-3}{3}\rceil\\
&= (1-g)F + g\cdot\sum_{i=4}^{3n}\lceil\tfrac{i-3}{3}\rceil
\end{aligned}
$$

$$= \quad (1-g)F + g \cdot \tfrac{3}{2}n(n-1) \;=\; (1-g)F + gF \;=\; F \;.$$

For a dummy example z we assume that $\mathcal{C}\backslash\mathcal{C}(z) = \{1\}$. Then we have for $\hat{\gamma} = f(\gamma, z)$:

$$g := \tfrac{1}{3n} = \hat{\gamma}_2 = \cdots = \hat{\gamma}_{3n} > \hat{\gamma}_1 = 0,$$

from which it follows that

$$\begin{aligned} G(f(\gamma,z)) \cdot \sum_{c,d \in \widehat{\mathcal{C}}} p(c,z,d) &= G(\hat{\gamma}) \cdot (1-g) \\ &= (1-g)F + \sum_{i=2}^{3n} (\gamma_i + g)\lceil \tfrac{i-1}{3} \rceil \\ &= (1-g)F + g \cdot \sum_{i=2}^{3n} \lceil \tfrac{i-1}{3} \rceil \\ &= (1-g)F + g \cdot \underbrace{(\tfrac{3}{2}n(n+1) - n)}_{>F} > (1-g)F + gF = F. \end{aligned}$$

The minimum value F is therefore attained only if a useful example is given and among these examples exactly the greedy choices are optimal. □

The next lemma describes the optimal teachers as sequential teachers instead of information state teachers.

Lemma 8.12 *Let $n \in \mathbb{N}$ with $n \geq 2$. Let $\mathcal{C}$ be a full X3C class containing $3n+1$ concepts, and let c^* be the concept containing all instances. A teacher $T\colon \mathbb{N} \to \mathcal{X}(c^*)$ is optimal if and only if $T(t) = z_{1+t \bmod n}$ for all t, where the examples z_j are chosen such that $\mathcal{C} \setminus \mathcal{C}(z_j)$ for $j = 1, \ldots, n$ are pairwise disjoint and $\bigcup_j (\mathcal{C} \setminus \mathcal{C}(z_j)) = \mathcal{C} \setminus \{c^*\}$.*

Proof. For the "only if" part, let $T\colon \mathbb{N} \to \mathcal{X}(c^*)$ be an optimal teacher for c^*, that is, with an expected teaching time of $1 + \frac{3}{2}n(n-1)$. We denote the information states reached by learner $\mathcal{L}_1$ during teaching by $\gamma^{(0)}, \gamma^{(1)}, \ldots$

From Lemma 8.11 we know that the only teachers with an expectation of $1+\frac{3}{2}n(n-1)$ are the greedy teachers. Thus T picks examples greedily in all states $\gamma^{(t)}$. In state $\gamma^{(0)}$ all examples are equally greedy and we assume without loss of generality that $T(0) = z_1$ with $\mathcal{C} \setminus \mathcal{C}(z_1) = \{c_1, c_2, c_3\}$. For the probabilities in $\gamma^{(1)}$ it follows

$$\gamma_4^{(1)} = \cdots = \gamma_{3n}^{(1)} > \gamma_1^{(1)} = \gamma_2^{(1)} = \gamma_3^{(1)}.$$

In state $\gamma^{(1)}$ every example z_2 with $\mathcal{C} \setminus \mathcal{C}(z_2) \subseteq \{c_4, \ldots, c_{3n}\}$ is a greedy choice. Without loss of generality, we assume that $\mathcal{C} \setminus \mathcal{C}(z_1) = \{c_4, c_5, c_6\}$. Then state $\gamma^{(2)}$ satisfies

$$\gamma_7^{(2)} = \cdots = \gamma_{3n}^{(2)} > \gamma_1^{(2)} = \gamma_2^{(2)} = \gamma_3^{(2)} > \gamma_4^{(2)} = \gamma_5^{(2)} = \gamma_6^{(2)}.$$

The next example of T then satisfies $\mathcal{C} \setminus \mathcal{C}(z_3) \subseteq \{7, \ldots, 3n\}$ and so on until the learner reaches the state $\gamma^{(n-1)}$ with

$$\gamma_{3n-2}^{(n-1)} = \cdots = \gamma_{3n}^{(n-1)} > \gamma_1^{(n-1)} = \gamma_2^{(n-1)} = \gamma_3^{(n-1)} > \cdots > \gamma_{3n-5}^{(n-1)} = \gamma_{3n-4}^{(n-1)} = \gamma_{3n-3}^{(n-1)}.$$

Therefore $T(n-1) = z_n$ with $\mathcal{C} \setminus \mathcal{C}(z_n) = \{c_{3n-2}, c_{3n-1}, c_{3n}\}$ yielding $\gamma^{(n)}$ with

$$\gamma_1^{(n)} = \gamma_2^{(n)} = \gamma_3^{(n)} > \gamma_4^{(n)} = \gamma_5^{(n)} = \gamma_6^{(n)} > \cdots > \gamma_{3n-2}^{(n)} = \gamma_{3n-1}^{(n)} = \gamma_{3n}^{(n)}. \tag{8.6}$$

At this point, the teacher T must choose $z_1, z_2, z_3, \ldots, z_n$ again. This results in a state $\gamma^{(2n)}$, which satisfies (8.6) with (n) replaced by $(2n)$. It follows that the teacher T chooses the sequence $z_1, \ldots, z_n$ in an endless loop.

In addition, the sets $\mathcal{C} \setminus \mathcal{C}(z_j)$ for $j = 1, \ldots, n$ are mutually disjoint and their union is $\mathcal{C} \setminus \{c^*\}$, as required.

For the "if" part, note that we have just shown the existence of an optimal teacher iterating through n examples with the property stated in the lemma. Every other teacher with this property can be mapped to that one by renaming instances or concepts. Thus, every teacher with this property is optimal. □

Until now we have characterized the optimal teachers for *full* X3C classes only. We can use this result to state the optimal teaching time for positive X3C classes.

Lemma 8.13 *Let $\mathcal{C}$ be an X3C class. Then $E_1^-(c^*, \mathcal{C}) = 1 + \frac{3}{2}n(n-1)$ if and only if $\mathcal{C}$ is a positive X3C class.*

Proof. For the "if" direction, let without loss of generality $z_1, \ldots, z_n \in \mathcal{X}(c^*)$ be such that $\mathcal{C} \setminus \mathcal{C}(z_j)$ for $j = 1, \ldots, n$ are pairwise disjoint and $\bigcup_j (\mathcal{C} \setminus \mathcal{C}(z_j)) = \mathcal{C} \setminus \{c^*\}$.

The teacher T with $T(t) = z_{1+t \bmod n}$ for all $t \in \mathbb{N}$ has an expected teaching time of $1 + \frac{3}{2}n(n-1)$. This follows similar to Lemma 8.12. If there was a better teacher, this teacher would also have a smaller expectation when used for teaching the target concept in a full X3C class, a contradiction to Lemma 8.12.

For the "only if" direction, assume that $E_1^-(c^*, \mathcal{C}) = 1 + \frac{3}{2}n(n-1)$, and suppose for a contradiction that $\mathcal{C}$ is a negative X3C class. Then there is a teacher T for c^* with expectation $1+\frac{3}{2}n(n-1)$, but not iterating through a set of examples $z_1, \ldots, z_n \in \mathcal{X}(c^*)$ with the "positive X3C class property" (because negative X3C classes have no such examples). The teacher T would then have the same expectation with respect to a full X3C class, too. Therefore, T would be an optimal teacher for the full X3C class, a contradiction to Lemma 8.12. □

We can now combine the previous four lemmas into our main result.

Theorem 8.14 *The problem* `OPT-TEACHINGTIME` *is $\mathcal{NP}$-hard.*

Input: Concept class $\mathcal{C}$, concept $c^* \in \mathcal{C}$, rational number $\varepsilon > 0$.

$J := TD(c^*) \cdot |\mathcal{C}|$

for $\ell = 1, 2, \ldots$:

  for all $\alpha \in \mathcal{X}(c)^\ell$:

    // denote by $\delta^{(i)}(c)$ $(i = 1, \ldots, \ell)$ the probability of being in hypothesis c

    // at round i when taught α.

    $b(\alpha) := \sum_{i=1}^{\ell} i \cdot (\delta^{(i)}(c^*) - \delta^{(i-1)}(c^*)) + (\ell + 1)(1 - \delta^{(\ell)}(c^*))$

    $B(\alpha) := \sum_{i=1}^{\ell} i \cdot (\delta^{(i)}(c^*) - \delta^{(i-1)}(c^*)) + (\ell + J)(1 - \delta^{(\ell)}(c^*))$

  $b_\ell := \min\{b(\alpha) \mid |\alpha| = \ell\}$

  if $\exists \alpha \in \mathcal{X}(c)^\ell : B(\alpha) - b_\ell < \varepsilon$ **then output** $B(\alpha)$.

Algorithm 8.2. *Algorithm computing an ε-approximation of $E_1^-(c^*, \mathcal{C})$.*

Proof. Let $\langle B, A_1, \ldots, A_m \rangle$ with $|B| = 3n$ be an instance of `X3C`, and let the instance of `OPT-TEACHINGTIME` resulting from the polynomial time reduction Algorithm 8.1 be $\langle \mathcal{C}, c^*, 1 + \frac{3}{2}n(n-1) \rangle$.

Now, $\langle B, A_1, \ldots, A_m \rangle$ is a positive instance of `X3C` if and only if $\mathcal{C}$ is a positive X3C class (by definition). The latter holds if and only if $E_1^-(c^*, \mathcal{C}) = 1 + \frac{3}{2}n(n-1)$ (by Lemma 8.13). This in turn holds if and only if $\langle \mathcal{C}, c^*, 1 + \frac{3}{2}n(n-1) \rangle$ is a positive `OPT-TEACHINGTIME` instance. □

Although it is open whether the value $E_\infty^-(c^*, \mathcal{C})$ can be computed for given c^* and $\mathcal{C}$, it is at least possible to effectively approximate that value with arbitrary precision.

Fact 8.15 *There is an algorithm with:*
Input: *Concept class $\mathcal{C}$, concept $c^* \in \mathcal{C}$, precision $\varepsilon > 0$.*
Output: *$F \in \mathbb{R}$ with $|F - E_1^-(c^*)| < \varepsilon$.*

Proof. The idea of Algorithm 8.2 is to approximate the expectations for growing finite sequences of examples. This is done until for one such sequence the probability of not being in the target state at the end of the sequence is very small.

We have introduced the value J into Algorithm 8.2 as a crude upper bound for $E_1^-(c^*)$. The values $b(\alpha)$ and $B(\alpha)$ are a lower and an upper bound for the expected teaching time of a teacher starting with example sequence α. The values $\delta^{(i)}(c)$ can be calculated according to the state transition function (8.1). Its values are always rational numbers, which can be calculated and stored exactly.

We have to show that (1) the algorithm always terminates and (2) the output is an ε-approximation for $E_1^-(c^*)$.

Claim 1: For all $\alpha \in \mathcal{X}(c^*)^* : E_1^-(c^*) \leq B(\alpha)$.

Proof. A teacher, after giving the examples in α, can be continued such that it takes at most expected J further examples to make the learner hypothesize the target. Thus, $B(\alpha)$ is an upper bound for the expected teaching time of all teachers starting with α. This teaching time cannot be less than the optimal one, $E_1^-(c^*)$. □ Claim 1

Claim 2: For all $\ell \geq 1 : b_\ell \leq E_1^-(c^*)$.

Proof. Let $\alpha \in \mathcal{X}(c^*)^\ell$. If the teacher has already finished teaching after presenting α (which can only happen if $TD(c^*) = 1$) then $b(\alpha)$ equals its expected teaching time, namely 1. If the teacher has not yet finished teaching after giving the examples in α, it needs at least one additional round. Thus, $b(\alpha)$ lower bounds the expected teaching time for all teachers starting with α. Consequently, b_ℓ is a lower bound for the expected teaching time for all teachers (since every teacher starts with some sequence $\alpha \in \mathcal{X}(c^*)^\ell$). □ Claim 2

Claim 3: $\lim_{\ell\to\infty} b_\ell = E_1^-(c^*)$.

Proof. Let $\zeta > 0$ and $\ell_0 = (J \cdot E_1^-(c^*))/\zeta$. We show that for all $\ell \in \mathbb{N}$ with $\ell \geq \ell_0$, $|E_1^-(c^*) - b_\ell| < \zeta$. Let $\ell \geq \ell_0$. Then $\ell \geq (J \cdot E_1^-(c^*))/\zeta$.

Let $\alpha \in \mathcal{X}(c)^\ell$ with $b(\alpha) = b_\ell$. Then $b(\alpha) \leq E_1^-(c^*)$ and therefore $(\ell + 1) \cdot (1 - \delta^{(\ell)}(c^*)) \leq E_1^-(c^*)$. It follows $1 - \delta^{(\ell)}(c^*) \leq E_1^-(c^*)/(\ell + 1)$.

For $B(\alpha)$ we have

$$B(\alpha) = b(\alpha) + (1 - \delta^{(\ell)}(c^*)) \cdot (J - 1) \leq b(\alpha) + \frac{E_1^-(c)}{\ell + 1} \cdot (J - 1) < b(\alpha) + \frac{E_1^-(c)}{\ell} \cdot J.$$

Substituting ℓ yields

$$B(\alpha) < b(\alpha) + \zeta.$$

On the other hand, $E_1^-(c^*) \leq B(\alpha)$ and therefore $E_1^-(c^*) < b(\alpha) + \zeta$, hence $E_1^-(c^*) - b(\alpha) < \zeta$. Since ζ was arbitrary, the claim follows. □ Claim 3

To prove the termination of the algorithm we have to show that there is an α such that $B(\alpha) - b_\ell < \varepsilon$. Let $T\colon \mathbb{N} \to \mathcal{X}(c^*)$ be an optimal teacher and denote by $T[0\!:\!\ell]$ the sequence $\langle T(0), \ldots, T(\ell)\rangle$. Then $\lim_{\ell\to\infty} B(T[0\!:\!\ell]) = E_1^-(c^*)$. Together with Claim 3 it follows that there is an ℓ such that $B(T[0\!:\!\ell]) - E_1^-(c^*) < \varepsilon/2$ and $E_1^-(c^*) - b_\ell < \varepsilon/2$. That means $B(T[0\!:\!\ell]) - b_\ell < \varepsilon$ and the algorithm terminates at the latest for this ℓ.

When the algorithm terminates, it outputs $B(\alpha)$ for an α with $B(\alpha) - b_\ell < \varepsilon$. This is an ε-approximation for $E_1^-(c^*)$ since $b_\ell \leq E_1^-(c^*) \leq B(\alpha)$ (see Claims 1 and 2). □

Since the optimal teacher is at least hard and probably impossible to determine, it is natural to study teaching heuristics instead. We devote the next two sections to this kind of study.

8.4 Cyclic Teachers

A natural heuristic, which we already mentioned in Section 6.3, is providing the same sequence of examples in an endless loop. We call such teachers *cyclic teachers*. Each such teacher can be identified with the example sequence it iterates over.

Cyclic teachers are easier to handle than teachers in general. It is simple to decide whether a given cyclic teacher is successful (see Fact 8.16). Furthermore, their teaching times can be calculated relatively easily (see Lemma 8.17). At the same time, cyclic teachers are powerful enough to provide arbitrarily close approximations to an optimal teacher (see Corollary 8.21).

Fact 8.16 *It can be decided efficiently whether a given cyclic teacher is successful.*

Proof. Let $\langle z_1, \ldots, z_\ell \rangle$ be the example sequence the teacher iterates over. From Fact 8.2 it follows that the teacher is successful if and only if $\{z_1, \ldots, z_\ell\}$ is a teaching set. Checking whether or not a set of $k = |\{z_1, \ldots, z_\ell\}|$ examples is a teaching set, can be done by checking each of the $|\mathcal{C}|$ concepts for consistency with all k examples. The examples form a teaching set if and only if the only consistent concept is the target. All consistency checks can be done in time $O(|\mathcal{C}| \cdot k \cdot |X|)$, that is, in polynomial time in the representation size of the concept class and the example sequence. □

Lemma 8.17 *The expected teaching time of a cyclic teacher can be computed from the sequence of examples that the teacher iterates over.*

Proof. Let $\mathcal{C}$ be a concept class and let $c^* \in \mathcal{C}$. Let T be a cyclic teacher iterating through $z_0, \ldots, z_{\ell-1}$. If $\{z_0, \ldots, z_{\ell-1}\}$ is not a teaching set for c^* then the expectation of T is infinite. We therefore assume $\{z_0, \ldots, z_{\ell-1}\}$ to be a teaching set.

Teaching will be successful no matter at which of the examples z_i the loop starts. We denote by F_i $(0 \leq i < \ell)$ the expected teaching time for the teacher $T_i \colon T_i(t) = z_{i+t \bmod \ell}$ starting with example z_i. For $h \in \mathcal{C}$ we denote by $F_i(h)$ the expectation for teacher T_i when the learner's initial state is h.

For notational convenience throughout this proof all subscripts of T, z, and F are to be taken modulo ℓ.

We can now state a linear equation for F_i involving all F_j with $j \neq i$. Consider the teacher T_i and the probability distribution δ over hypotheses after the first example, z_i, has been given. The learner assumes all hypotheses $h \in \mathcal{C}(z_i)$ with equal probability $\delta_h = 1/|\mathcal{C}(z_i)|$ and all other hypotheses $h \notin \mathcal{C}(z_i)$ with probability $\delta_h = 0$.

The expectation F_i is one plus the weighted sum of the expectations of teacher T_{i+1} starting in the states h, that is,

$$F_i = 1 + \sum_{h \in \mathcal{C} \setminus \{c^*\}} \delta_h \cdot F_{i+1}(h).$$

Now we determine $F_{i+1}(h)$. Consider a learner in state $h \neq c^*$ and a teacher giving $z_{i+1}, z_{i+2}, \ldots$. The learner will change its state only when the first example inconsistent with h arrives (such an example exists since the z_i's form a teaching set for c^*). Let z_{i+k} be this example. Beginning with z_{i+k}, teaching proceeds as if teacher T_{i+k} had started from the *init* state. Therefore $F_{i+1}(h) = (k-1) + F_{i+k}$.

If we denote for $i = 0, \ldots, \ell-1$ and for $k = 1, \ldots, \ell$,

$$\mathcal{C}_{i,k} = \{h \in \mathcal{C} \setminus \{c^*\} \mid h \in \mathcal{C}(z_i), h \in \mathcal{C}(z_{i+1}), \ldots, h \in \mathcal{C}(z_{i+k-1}), h \notin \mathcal{C}(z_{i+k})\},$$

then we get the following linear equation for F_i:

$$F_i = 1 + \sum_{1 \leq k \leq \ell} \frac{|\mathcal{C}_{i,k}|}{|\mathcal{C}(z_i)|} \cdot ((k-1) + F_{i+k}) \ .$$

In this manner we get ℓ linear equations in the variables $F_0, \ldots, F_{\ell-1}$. Denoting the solution vector by $\boldsymbol{F}$ we get a linear equation system of the form $(\mathbf{1} - C) \cdot \boldsymbol{F} = K$, where $\mathbf{1}$ is the $\ell \times \ell$ unit matrix, C is a matrix composed of entries of the form $\frac{|\mathcal{C}_{i,k}|}{|\mathcal{C}(z_i)|}$ (and zeros). Therefore C is substochastic and $\mathbf{1} - C$ is invertible. Hence the values $F_0, \ldots, F_{\ell-1}$ are uniquely determined by the linear equations derived above. □

The previous lemma allows us to calculate the expectations we had claimed in Figure 6.2.

Fact 8.18 *Let $c^* \in \mathcal{M}_4^1$ be the monomial over 4 variables represented by* `**11`*. Let $z_1 = (0011, 1)$, $z_2 = (1111, 1)$, $z_3 = (0001, 0)$, and $z_4 = (0010, 0)$ be examples for c^*.*

1. *The cyclic teacher $\langle z_1, z_2, z_3, z_4 \rangle$ teaches c^* to the learner $\mathcal{L}_1$ without feedback in expected $580357/17732 \approx 32.7294$ rounds.*
2. *The cyclic teacher $\langle z_3, z_4, z_1, z_2 \rangle$ teaches c^* to the learner $\mathcal{L}_1$ without feedback in expected $607109/17732 \approx 34.2380$ rounds.*
3. *The cyclic teacher $\langle z_1, z_2, z_3, z_1, z_2, z_4 \rangle$ teaches c^* to the learner $\mathcal{L}_1$ without feedback in expected $26315/938 \approx 28.0544$ rounds.*

Proof. 1. We denote the expectation of the cyclic teacher $\langle z_1, z_2, z_3, z_4 \rangle$ by F_0, the expectation of $\langle z_2, z_3, z_4, z_1 \rangle$ by F_1, that of $\langle z_3, z_4, z_1, z_2 \rangle$ by F_2, and that of $\langle z_4, z_1, z_2, z_3 \rangle$ by F_3.

On receiving z_1 the learner can choose between 16 consistent hypotheses. Of these, 4 are consistent with z_2 as well (namely `**11`, `**1*`, `***1`, `****`); the other 12 hypotheses are inconsistent. Of the 4 consistent ones, 2 are consistent even with z_3 (namely `**11`, `**1*`). Of these, just one is inconsistent with z_4 (namely `**1*`). For F_0 we obtain the equation

$$F_0 = 1 + \frac{1}{16} \cdot (12 \cdot F_1 + 2 \cdot (F_2 + 1) + 1 \cdot (F_3 + 2)).$$

If the learner receives z_2 as first example it may choose between 16 consistent hypotheses again. Of these, only 2 are inconsistent with z_3 (namely `***1`, `****`). Of the 14 consistent ones, only `**1*` is inconsistent with the next example z_4. Finally, z_1 is inconsistent with 12 of the 13 remaining hypotheses (all but the target). Therefore we get

$$F_1 = 1 + \frac{1}{16} \cdot (2 \cdot F_2 + 1 \cdot (F_3 + 1) + 12 \cdot (F_0 + 2)).$$

If the learner receives z_3 first, it can choose between 65 consistent hypotheses. The example z_4 is inconsistent with 12 of them and of the remaining 53 hypotheses, 49 are inconsistent with z_1. Finally, z_2 is inconsistent with 3 of the last 4 hypotheses (all but the target). This yields

$$F_2 = 1 + \frac{1}{65} \cdot (12 \cdot F_3 + 49 \cdot (F_0 + 1) + 3 \cdot (F_1 + 2)).$$

If the learner receives z_4 first, it can choose between 65 consistent hypotheses. Only 8 of them are consistent with the next example z_1 (namely `0011`, `00*1`, `0*11`, `0**1`, `*011`, `*0*1`, `**11`, `***1`). Of these, 6 are inconsistent with z_2 (all but `**11`, `***1`). The hypothesis `***1` is then inconsistent with z_3. It follows

$$F_3 = 1 + \frac{1}{65} \cdot (57 \cdot F_0 + 6 \cdot (F_1 + 1) + 1 \cdot (F_2 + 2)).$$

The solution of the equation system is

$$F_0 = \frac{580357}{17732}, \qquad F_1 = \frac{593656}{17732}, \qquad F_2 = \frac{607109}{17732}, \qquad F_3 = \frac{592982}{17732}.$$

The value F_0 is the sought expectation.

2. This teacher is the same as in Item 1 shifted by two examples. Thus we have already derived the necessary equations. The sought expectation is the value F_2 above.

3. We spare the reader the tedious derivations of the equations and just present them:

$$\begin{aligned}
F_0 &= 1 + \frac{1}{16} \cdot (12 \cdot F_1 + 2 \cdot (F_2 + 1) + 0 \cdot (F_3 + 2) + 0 \cdot (F_4 + 3) + 1 \cdot (F_5 + 4)),\\
F_1 &= 1 + \frac{1}{16} \cdot (2 \cdot F_2 + 12 \cdot (F_3 + 1) + 0 \cdot (F_4 + 2) + 1 \cdot (F_5 + 3) + 0 \cdot (F_0 + 4)),\\
F_2 &= 1 + \frac{1}{65} \cdot (57 \cdot F_3 + 6 \cdot (F_4 + 1) + 1 \cdot (F_5 + 2) + 0 \cdot (F_0 + 3) + 0 \cdot (F_1 + 4)),\\
F_3 &= 1 + \frac{1}{16} \cdot (12 \cdot F_4 + 2 \cdot (F_5 + 1) + 0 \cdot (F_0 + 2) + 0 \cdot (F_1 + 3) + 1 \cdot (F_2 + 4)),\\
F_4 &= 1 + \frac{1}{16} \cdot (2 \cdot F_5 + 12 \cdot (F_0 + 1) + 0 \cdot (F_1 + 2) + 1 \cdot (F_2 + 3) + 0 \cdot (F_3 + 4)),
\end{aligned}$$

	`111 **`		`1111 *`
z_0	(111 00, 1)	z_0	(1111 0, 1)
z_1	(011 00, 0)	z_1	(0111 0, 0)
z_2	(111 11, 1)	z_2	(1111 1, 1)
z_3	(101 11, 0)	z_3	(1011 1, 0)
z_4	(111 00, 1)	z_4	(1111 0, 1)
z_5	(110 00, 0)	z_5	(1101 0, 0)
z_6	(111 11, 1)	z_6	(1111 1, 1)
z_7	(011 11, 0)	z_7	(1110 1, 0)
z_8	(111 00, 1)		
z_9	(101 00, 0)		
z_{10}	(111 11, 1)		
z_{11}	(110 11, 0)		

Figure 8.4 *The cyclic teacher for monomials as in Fact 8.19. The left column states the examples for the target* `111**`*, the right column for* `1111*`*.*

$$F_5 \;=\; 1 + \frac{1}{65} \cdot (57 \cdot F_0 + 6 \cdot (F_1 + 1) + 1 \cdot (F_2 + 2) + 0 \cdot (F_3 + 3) + 0 \cdot (F_4 + 4)).$$

The solutions are:

$$F_0 = \frac{26315}{938}, \quad F_1 = \frac{53233}{1876}, \quad F_2 = \frac{13501}{469}, \quad F_3 = \frac{26315}{938}, \quad F_4 = \frac{53233}{1876}, \quad F_5 = \frac{13501}{469},$$

and the sought expectation is the value F_0. □

Lemma 8.17 also makes it possible to compute upper bounds for the optimal teaching time of concepts. We only have to devise a successful cyclic teacher and calculate its teaching time. In the next fact we illustrate just that by calculating an upper bound for the teaching time of monomials. This will tell us how much slower teaching monomials without feedback is than with feedback (compare Fact 7.2).

Fact 8.19 *Let $k \geq 3$ and $n \geq k$. Let $c^* \in \mathcal{M}_n^1$ be a monomial with k variables. The optimal teaching time $E_1^-(c^*, \mathcal{M}_n^1)$ is then upper bounded by*

$$\frac{-2 + 2^{k+1} + 7 \cdot 2^n - 2^{n+k+2} + 2^{k+2} \cdot 3^n + 2^{n+2} \cdot 3^n - 2 \cdot 3^{n+1} - 4^{n+1} - 2k \cdot (2^k - 2^{n+1} + 2 \cdot 3^n)}{2 \cdot 3^n - 2^{n+1} + 2^k}.$$

Proof. This expectation is achieved by the following cyclic teacher T (see also Figure 8.4). Without loss of generality, let the target concept be $\mathtt{1}^k\mathtt{*}^{n-k}$. The teacher provides alternately positive and negative examples. The positive examples alternate between the two complementary examples $(1^k0^{n-k}, 1)$ and $(1^k1^{n-k}, 1)$. The first k characters of the instances in the negative examples iterate through 01^{k-1}, 101^{k-2},...,$1^{k-1}0$;

the last $n-k$ characters equal the last $n-k$ characters of the immediately preceding positive example. The number m of examples in this sequence is $2k$ if k is even, and $4k$ if k is odd. We denote the examples by $z_0, \ldots, z_{m-1}$.

We have to find the linear equations for the expected teaching times of the teachers $T_0, T_1, \ldots, T_{m-1}$ starting the cycle at $z_0, z_1, \ldots, z_{m-1}$. Normally this would yield m equations with m variables. Here, however, the examples are chosen in such a way that all teachers starting with a positive example $(z_0, z_2, \ldots)$ have the same expectation. Likewise, all teachers starting with a negative example $(z_1, z_3, \ldots)$ have the same expectation.

The reason for many expectations to be equal is the presence of symmetries within the instances. If we permute the first k bits in all instances in the same way we get a new example sequence. This new sequence leads to the same expectation as the old one because the target concept is insensitive to permutations of the first k characters. Similarly if we flip the $n-k$ last bits in all instances we get a new example sequence. But this new sequence leads to the same expectation as the old one. Now consider the teacher T_2 starting at z_2. By flipping the last $n-k$ bits and cyclically left-shifting the first k bits in each instance we get a new teacher T_2' with the same expectation. But the teacher T_2' is exactly the teacher that starts at z_0. Similar arguments apply for the teachers starting at the other positive examples. Furthermore, analogous arguments show that also the teachers starting at negative examples all have the same expectation. We denote the expectation of the teachers starting at positive examples by F_+ and the expectation of the teachers starting at negative examples by F_-.

We first derive an equation for F_+. Consider the cyclic teacher starting with z_0. After receiving z_0 the learner is in a state in which all hypotheses from $\mathcal{C}(z_0)$ have the probability $1/|\mathcal{C}(z_0)| = 2^{-n}$. For every following round we have to determine how many of the initially 2^n hypotheses are consistent with all examples up to that round:

- Consistent with z_0: the 2^n hypotheses $\{\mathtt{1}, *\}^k\{\mathtt{0}, *\}^{n-k}$,
- consistent with z_0, z_1: the 2^{n-1} hypotheses $\mathtt{1}\{\mathtt{1}, *\}^{k-1}\{\mathtt{0}, *\}^{n-k}$,
- consistent with z_0, z_1, z_2: the 2^{k-1} hypotheses $\mathtt{1}\{\mathtt{1}, *\}^{k-1}*^{n-k}$.

Among z_0, z_1, z_2 there are two complementary positive examples. This limits the set of consistent hypotheses in the remaining rounds to subsets of $\{\mathtt{1}, *\}^k*^{n-k}$. Moreover the upcoming positive examples $z_4, z_6, \ldots$ will not be inconsistent with any hypothesis, only negative examples will. Each negative example introduces one "$\mathtt{1}$":

- Consistent with $z_0, \ldots, z_3$: the 2^{k-2} hypotheses $\mathtt{11}\{\mathtt{1}, *\}^{k-2}*^{n-k}$,
- consistent with $z_0, \ldots, z_4$: the 2^{k-2} hypotheses $\mathtt{11}\{\mathtt{1}, *\}^{k-2}*^{n-k}$,
- consistent with $z_0, \ldots, z_5$: the 2^{k-3} hypotheses $\mathtt{111}\{\mathtt{1}, *\}^{k-3}*^{n-k}$.

In general,

- for $j=3,\ldots,k$: consistent with $z_0,\ldots,z_{2j-1}$: the 2^{k-j} hypotheses $\mathtt{1}^j\{\mathtt{1},*\}^{k-j}*^{n-k}$,
- for $j=3,\ldots,k$: consistent with $z_0,\ldots,z_{2j}$: the 2^{k-j} hypotheses $\mathtt{1}^j\{\mathtt{1},*\}^{k-j}*^{n-k}$.

The only hypothesis consistent with $z_0,\ldots,z_{2k-1}$ is the target concept $\mathtt{1}^k*^{n-k}$ with which no example can be inconsistent. The equation for F_+ can now be stated:

$$\begin{aligned}F_+ &= 1+2^{-n}\cdot\Big(2^{n-1}\cdot F_- + (2^{n-1}-2^{k-1})\cdot(F_++1) + 2^{k-2}\cdot(F_-+2)\\ &\qquad + 2^{k-3}\cdot(F_-+4)+\cdots+2^1\cdot(F_-+(2k-4))+2^0\cdot(F_-+(2k-2))\Big)\\ &= 1+2^{-n}\Big(2^{n-1}\cdot F_- + (2^{n-1}-2^{k-1})\cdot(F_++1)+\sum_{i=2}^{k}2^{k-i}\cdot(F_-+2i-2)\Big).\end{aligned}$$

For the sum we have

$$\sum_{i=2}^{k}2^{k-i}\cdot(F_-+2i-2) = (2^{k-1}-1)\cdot F_- + 2^{k+1}-2k-2,$$

from which follows the equation for F_+:

$$F_+ = 1+2^{-n}\cdot\Big((2^{n-1}+2^{k-1}-1)\cdot F_- + (2^{n-1}-2^{k-1})\cdot(F_++1)+2^{k+1}-2k-2\Big). \quad (8.7)$$

In a similar manner we now derive the equation for F_-. Consider the teacher that starts with the negative example z_1. There are 3^n-2^n examples consistent with z_1.

- Consistent with z_1: all hypotheses except $\{\mathtt{0},*\}\{\mathtt{1},*\}^{k-1}\{\mathtt{0},*\}^{n-k}$ (3^n-2^n hypotheses),
- consistent with z_1,z_2: the hypotheses $\{\mathtt{1},*\}^n$ except $*\{\mathtt{1},*\}^{k-1}*^{n-k}$ (2^n-2^{k-1} hypotheses),
- consistent with z_1,z_2,z_3: the hypotheses $\{\mathtt{1},*\}\mathtt{1}\{\mathtt{1},*\}^{n-2}$ except $*\mathtt{1}\{\mathtt{1},*\}^{k-2}*^{n-k}$ ($2^{n-1}-2^{k-2}$ hypotheses),
- consistent with z_1,z_2,z_3,z_4: the 2^{k-2} hypotheses $\mathtt{11}\{\mathtt{1},*\}^{k-2}*^{n-k}$.

Among the examples z_1,z_2,z_3,z_4 there are two complementary positive examples (z_2 and z_4) which limit the hypotheses to subsets of $\{\mathtt{1},*\}^k*^{n-k}$. Positive examples will not make the set of consistent hypotheses smaller. Negative examples introduce one "1" into the first k characters.

- Consistent with $z_1,\ldots,z_5$: the 2^{k-3} hypotheses $\mathtt{111}\{\mathtt{1},*\}^{k-3}*^{n-k}$,

- consistent with $z_1, \ldots, z_6$: the 2^{k-3} hypotheses $\mathtt{111}\{\mathtt{1}, *\}^{k-3}*^{n-k}$.

More general,

- for $j = 3, \ldots, k$: consistent with $z_0, \ldots, z_{2j-1}$: the 2^{k-j} hypotheses $\mathtt{1}^j\{\mathtt{1}, *\}^{k-j}*^{n-k}$,
- for $j = 3, \ldots, k$: consistent with $z_0, \ldots, z_{2j}$: the 2^{k-j} hypotheses $\mathtt{1}^j\{\mathtt{1}, *\}^{k-j}*^{n-k}$.

After example z_{2k-1} only the target hypothesis is left. We obtain the following equation:

$$\begin{aligned}
F_- &= 1 + \frac{1}{3^n - 2^n} \cdot \Big((3^n - 2^{n+1} + 2^{k-1}) \cdot F_+ \\
&\quad + (2^{n-1} - 2^{k-2}) \cdot (F_- + 1) + (2^{n-1} - 2^{k-1}) \cdot (F_+ + 2) \\
&\quad + 2^{k-3} \cdot (F_- + 3) + \cdots + 2^1 \cdot (F_- + (2k-5)) + 2^0 \cdot (F_- + (2k-3)) \Big) \\
&= 1 + \frac{1}{3^n - 2^n} \cdot \Big((3^n - 2^{n+1} + 2^{k-1}) \cdot F_+ + (2^{n-1} - 2^{k-2}) \cdot (F_- + 1) \\
&\quad + (2^{n-1} - 2^{k-1}) \cdot (F_+ + 2) + \sum_{i=3}^{k} 2^{k-i} \cdot (F_- + 2i - 3) \Big) \\
&= 1 + \frac{1}{3^n - 2^n} \cdot \Big((3^n - 2^{n+1} + 2^{k-1}) \cdot F_+ + (2^{n-1} - 2^{k-2}) \cdot (F_- + 1) \\
&\quad + (2^{n-1} - 2^{k-1}) \cdot (F_+ + 2) + (2^{k-2} - 1) \cdot F_- + 5 \cdot 2^{k-2} - 2k - 1 \Big).
\end{aligned} \tag{8.8}$$

Solving the equations (8.7) and (8.8) for F_+ yields the desired expectation, and thus Fact 8.19 is shown. □

Applying the teacher from the previous proof to the target in Figure 6.2 yields an expectation of $2148/67 \approx 32.0597$.

We can now compare the teachability of monomials without feedback and the teachability with feedback (see Fact 7.2). Roughly speaking, teaching without feedback takes at most twice as long as teaching with feedback. Ideally we would like to show that the quotient of the bound from Fact 8.19 and $E_1^+(c^*, \mathcal{M}_n^1)$ is upper bounded by two, which is however not true: For $k = n = 13$ this quotient is $17309416200/8650679311 \approx 2.00093$. This is the highest value in the range $1 \leq k \leq n \leq 1000$. What we can show in fact is that the limit of the quotient is 2.

Corollary 8.20 *Let $n \geq k \geq 3$ and denote the upper bound from Fact 8.19 by $F(k, n)$. Let for all $k \leq n$, $c_k \in \mathcal{M}_n^1$ be a monomial with k literals. Then for all k,*

$$\lim_{n \to \infty} \frac{F(k, n)}{E_1^+(c_k, \mathcal{M}_n^1)} = 2.$$

Moreover,

$$\lim_{n\to\infty} \frac{F(n,n)}{E_1^+(c_n, \mathcal{M}_n^1)} = 2.$$

Proof. Dividing $F(k,n)$ by $E_1^+(c_k, \mathcal{M}_n^1)$ and simplifying yields

$$\frac{2 - 2^{1+k} - 7 \cdot 2^n + 2^{2+k+n} - 2^{2+k} \cdot 3^n - 2^{2+n} \cdot 3^n + 2 \cdot 3^{1+n} + 4^{1+n} + 2(2^k - 2^{1+n} + 2 \cdot 3^n)k}{-2 \cdot 3^n(-1 + 2^k + 2^n) + 2^n(-4 + 3 \cdot 2^k + 2^{1+n})}.$$

For constant k and growing n the dominating term in the numerator is $-2^{2+n} \cdot 3^n$ and in the denominator $-2 \cdot 3^n \cdot 2^n$, the quotient of which is 2.

Dividing $F(n,n)$ by $E_1^+(c_n, \mathcal{M}_n^1)$ and simplifying yields

$$\frac{-2 + 9 \cdot 2^n - 2^{3+2n} + 2 \cdot 3^n(-3 + 2^{2+n} - 2n) + 2^{1+n}n}{2^{2+n} - 2 \cdot 3^n + 2^{2+n} \cdot 3^n - 5 \cdot 4^n}.$$

The dominating terms are $2 \cdot 3^n \cdot 2^{2+n}$ and $2^{2+n} \cdot 3^n$. Their quotient is 2. □

Cyclic teachers are not only good for obtaining upper bounds of the optimal teaching time; sometimes, they are even optimal (see Lemma 8.12 and Fact 8.8). But sometimes, such as for monomials, we do not know whether there is an optimal, cyclic teacher. In any case, cyclic teachers can always be used to *approximate* the optimal one.

Corollary 8.21 *Let $\mathcal{C}$ be a finite concept class, $c^* \in \mathcal{C}$, and $\varepsilon > 0$. Then there is a cyclic teacher T with $|E_1^-(c^*, \mathcal{C}) - \mathbb{E}[T, \mathcal{L}_{1,\mathcal{C}}, c^*]| < \varepsilon$.*

Proof. Let α be the example sequence constructed during the execution of Algorithm 8.2 (see Page 174) and output in line 7. The proof of Fact 8.15 shows that any teacher starting with α yields an ε-approximation of the optimal teaching time. In particular this is true for the cyclic teacher iterating over α. □

Algorithm 8.2 is not an efficient means to find good cyclic teachers. Moreover, finding the optimal cyclic teacher is $\mathcal{NP}$-hard. This follows from our results about X3C classes in Section 8.3.

Corollary 8.22 *The following problem is $\mathcal{NP}$-hard.*

`OPT-CYCLIC-TEACHINGTIME`

Instance: *Concept class $\mathcal{C}$, concept $c^* \in \mathcal{C}$, rational number F.*

Question: *Is there a cyclic teacher with expected teaching time of at most F?*

Proof. We reduce `X3C` to `OPT-CYCLIC-TEACHINGTIME` with the same reduction function as in Theorem 8.14, that is, with Algorithm 8.1. An `X3C` instance $\langle B, A_1, \ldots, A_m \rangle$ is transformed into an `OPT-CYCLIC-TEACHINGTIME` instance $\langle \mathcal{C}, c^*, 1 + \frac{3}{2}n(n-1) \rangle$.

The tuple $\langle B, A_1, \ldots, A_m \rangle$ is a positive instance of `X3C` if and only if $\mathcal{C}$ is a positive X3C class (by definition). The latter holds if and only if the optimal cyclic teacher

	x_1	x_2	x_3	y_1	$\dots$	$\dots$	$\dots$	y_n
c^*	1	1	1	1	$\dots$	$\dots$	$\dots$	1
c_1	0	1	1	1	$\dots$	$\dots$	$\dots$	1
c_2	1	0	1	1	$\dots$	$\dots$	$\dots$	1
c_3	1	1	0	1	$\dots$	$\dots$	$\dots$	1
c_4	0	0	1	0	1	$\dots$	1	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	1	0	1	$\dots$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	1	$\ddots$		$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\ddots$	1
c_{3+n}	0	0	1	1	1	$\dots$	1	0

$\Big\}\ n$

Figure 8.5 *Family of classes for which the greedy teacher is much worse than the optimal one. The greedy teacher for c^* has a teaching time of $\Theta(\log n)$ whereas the optimal teaching time is in $O(1)$ (see Theorem 8.23).*

for $c^* \in \mathcal{C}$ has a teaching time of $1 + \frac{3}{2}n(n-1)$ (by Lemma 8.13 because the optimal teacher is cyclic). This in turn holds if and only if $\langle \mathcal{C}, c^*, 1 + \frac{3}{2}n(n-1) \rangle$ is a positive `OPT-CYCLIC-TEACHINGTIME` instance. □

8.5 Greedy Teachers

We know from the proof of Lemma 8.5 that a greedy teacher is always successful. Moreover, a greedy teacher allows a direct application of the optimality criterion Corollary 8.7. Thus we were able to prove the optimality of some greedy teachers (see Fact 8.8 and Lemma 8.11). However, greedy teachers are not always optimal. In fact, they can be arbitrarily far off the optimal teacher as they, in general, do not approximate the optimal teacher by a constant factor.

Theorem 8.23 *For every $d > 1$ there is a class $\mathcal{C}$ and a target c^* such that for all greedy teachers T, $\mathbb{E}[T, \mathcal{L}_{1,\mathcal{C}}, c^*] > d \cdot E_1^-(c^*, \mathcal{C})$.*

Proof. Figure 8.5 describes a family of classes $\mathcal{C}_n$. For $n \geq 1$ the class $\mathcal{C}_n$ contains $n+4$ concepts over the learning domain $X = \{x_1, x_2, x_3\} \cup \{y_1, \dots, y_n\}$. The target concept is $c^* = X$. For $i = 1, 2, 3$ there are concepts $c_i = X \setminus \{x_i\}$, and for $i = 1, \dots, n$ there are concepts $c_{3+i} = \{x_3\} \cup \{y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n\}$. The instances $y_1, \dots, y_n$ are merely dummy instances which allow us to have n concepts $c_4, \dots, c_{3+n}$ that are equal with respect to x_1, x_2, x_3.

A greedy teacher would never chose any of the examples $(y_1, 1), \dots, (y_n, 1)$. Such an example would be inferior to $(x_1, 1)$ and to $(x_2, 1)$ since both examples activate more

i	0	1	2	3	4	5	6	7	8
$\delta_*^{(i)}$	0	1/3	4/9	13/27	40/81	41/81	54/81	176/243	542/729
$\delta_1^{(i)}$	0	0	1/9	0	1/81	2/81	0	14/243	0
$\delta_2^{(i)}$	0	1/3	0	1/27	0	1/81	14/81	0	14/729
$\delta_3^{(i)}$	0	1/3	4/9	13/27	40/81	0	13/81	53/243	173/729
$\delta_4^{(i)}$	0	0	0	0	0	1/81	0	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\delta_{40}^{(i)}$	0	0	0	0	0	1/81	0	0	0
$\delta_{init}^{(i)}$	1	0	0	0	0	0	0	0	0
$T(i)$	$(x_1, 1)$	$(x_2, 1)$	$(x_1, 1)$	$(x_2, 1)$	$(x_3, 1)$	$(x_1, 1)$	$(x_2, 1)$	$(x_1, 1)$	$(x_2, 1)$

Figure 8.6 *The teaching process of a greedy teacher T for the concept c^* in the class $\mathcal{C}_{37}$ which is described in Figure 8.5. Each column shows the probability distribution $\delta^{(i)}$ over all hypotheses as well as the example $T(i)$ taught in round i. The teacher iterates through x_1, x_2, x_1, x_2, x_3 and corresponds to $\ell = 2$ in Claim 2 in the proof of Theorem 8.23.*

probability mass and distribute it among less hypotheses. Hence we need not consider $(y_1, 1), \ldots, (y_n, 1)$ in the following.

The proof is somewhat lengthy and technical, although the main ideas are simple enough. The main steps are as follows.

1. The cyclic teacher iterating over $\langle (x_1, 1), (x_2, 1), (x_3, 1) \rangle$ has a teaching time of less than five, regardless of n. Therefore $E_1^-(c^*, \mathcal{C}_n) < 5$ for all $n \in \mathbb{N}$.

2. For every $\ell \geq 2$, setting $n = \frac{3}{2}(3^{2\ell-1} - 3) + 1$ makes the greedy teacher for $c^* \in \mathcal{C}_n$ be a cyclic teacher that alternates between $(x_1, 1)$ and $(x_2, 1)$ for 2ℓ rounds and only then gives $(x_3, 1)$. Intuitively speaking, when n grows, the example $(x_3, 1)$ becomes less attractive for the greedy teacher and is given less often.

3. Cyclic teachers as in Step 2 have a teaching time of $\Theta(\ell)$. The longer it takes until $(x_3, 1)$ is taught, the more probability mass accumulates in hypothesis c_3. The remaining probability mass activated alternately by $(x_1, 1)$ and $(x_2, 1)$ is then very small. Consequently the target probability increases slowly, which leads to a high expected teaching time.

   On the other hand, giving $(x_3, 1)$ might increase the target probability even less in the very next round. But in the round after the next round, $(x_1, 1), (x_2, 1)$ are much more effective.

4. For n as above, the greedy teacher is cyclic with $\ell = \log(7 + 2n)/\log(9)$ and according to step 3 its teaching time is in $\Theta(\log n)$ and can thus be larger than $E_1^-(c^*, \mathcal{C}_n)$ by any factor d.

Now we prove four claims that correspond to the four steps just outlined.

Claim 1: The cyclic teacher T_{123} iterating over $\langle (x_1,1),(x_2,1),(x_3,1)\rangle$ has a teaching time of $\mathbb{E}[T_{123},\mathcal{L}_{1,\mathcal{C}_n},c^*] = (16+5n)/(4+n)$.

Proof. We apply the idea of the proof of Lemma 8.17. The expectations when the teacher starts at $(x_1,1),(x_2,1),(x_3,1)$ are denoted by F_1,F_2,F_3, respectively. Then we have

$$\begin{aligned} F_1 &= 1+\frac{1}{3}F_2+\frac{1}{3}(F_3+1),\\ F_2 &= 1+\frac{1}{3}F_3+\frac{1}{3}(F_1+1),\\ F_3 &= 1+\frac{n+1}{n+3}F_2+\frac{1}{n+3}(F_3+1). \end{aligned}$$

Solving this system of three linear equations yields $F_1 = (16+5n)/(4+n) < 5$. □ Claim 1

Claim 2: Let $\ell \geq 2$ and set $n = \frac{3}{2}(3^{2\ell-1}-3)+1$. Then the greedy teacher for $c^* \in \mathcal{C}_n$ is a cyclic teacher iterating over $\langle (x_1,1),(x_2,1)\rangle^\ell \circ \langle (x_3,1)\rangle$ or over $\langle (x_2,1),(x_1,1)\rangle^\ell \circ \langle (x_3,1)\rangle$.

Proof. Let T be a greedy teacher. As usual, we denote by $\delta^{(j)}$ the learner's state in round $j \in \mathbb{N}$. For readability, we abbreviate $\delta^{(j)}(c_i)$ by $\delta_i^{(j)}$ and $\delta^{(j)}(c^*)$ by $\delta_*^{(j)}$ and $\delta^{(j)}(init)$ by $\delta_{init}^{(j)}$.

By definition, $\delta_{init}^{(0)} = 1$. Now $T(0)$ can either be $(x_1,1)$ or $(x_2,1)$. Both examples are equally greedy. Without loss of generality, we assume that $T(0) = (x_1,1)$. Then we get $\delta_*^{(1)} = \delta_2^{(1)} = \delta_3^{(1)} = 1/3$. Figure 8.6 displays the teaching process for the first rounds in the case $\ell = 2$ and $n = 37$.

We want to show that T is cyclic with a period of $2\ell+1$. For technical reasons, we let the first cycle start at round 1 (rather than 0) and show that from then on T iterates over $\langle (x_2,1),(x_1,1)\rangle^{\ell-1} \circ \langle (x_3,1),(x_1,1)\rangle$. Generally, we denote by $r_k = k(2\ell+1)$ the last round of the k-th cycle for all $k \geq 0$.

In order to show the "cyclicity" of T, we shall show

Claim 2.1: For all $k \geq 0$:

$$\delta_1^{(r_{k+1}+1)} = 0, \tag{8.9}$$

$$\delta_2^{(r_{k+1}+1)} = \frac{3^{-1-2\ell}((-9+4\cdot 9^\ell+81^\ell)\cdot \delta_2^{(r_k+1)} + 2(9^\ell+81^\ell)\cdot \delta_3^{(r_k+1)})}{2(-1+9^\ell)}, \tag{8.10}$$

$$\delta_3^{(r_{k+1}+1)} = \frac{3^{-1-2\ell}((9-2\cdot 9^\ell+81^\ell)\cdot \delta_2^{(r_k+1)} + 2(-5\cdot 9^\ell+81^\ell)\cdot \delta_3^{(r_k+1)})}{2(-1+9^\ell)}, \tag{8.11}$$

$$\delta_j^{(r_{k+1}+1)} = 0 \qquad \text{for } j = 4, \ldots, n+3, \tag{8.12}$$

and

$$\begin{aligned} T(r_k + i) &= (x_{1+i \bmod 2}, 1) \qquad \text{for } i = 1, \ldots, 2\ell - 1, \\ T(r_k + 2\ell) &= (x_3, 1), \\ T(r_{k+1}) = T(r_k + 2\ell + 1) &= (x_1, 1). \end{aligned}$$

Proof. We prove the claim by induction on k. First assume that $k = 0$ and thus $r_k = 0$. In round 1, the teacher T will choose the example $(x_2, 1)$: $T(1) = (x_2, 1)$. The resulting state is $\delta^{(2)}$ with $\delta_2^{(2)} = 0$, $\delta_1^{(2)} = 1/9$, $\delta_3^{(2)} = 1/3 + 1/9 = 4/9$. Now T may either choose $(x_1, 1)$ or $(x_3, 1)$. It chooses $(x_1, 1)$ if $\delta_1^{(2)}/3 > \delta_3^{(2)}/(n+3)$ and chooses $(x_3, 1)$ if $\delta_1^{(2)}/3 < \delta_3^{(2)}/(n+3)$. Assume for a moment that $(x_1, 1)$ is chosen. Then we get $\delta_1^{(3)} = 0$, $\delta_2^{(3)} = 1/27$, $\delta_3^{(3)} = 4/9 + 1/27 = 13/27$. Then either $(x_1, 1)$ or $(x_3, 1)$ is chosen next, depending on the relation of $\delta_1^{(3)}/3$ to $\delta_3^{(3)}/(n+3)$. In general $(x_3, 1)$ is chosen only in that round $i \in \{1, 2, \ldots\}$ in which the inequality $\delta_{1+i \bmod 2}^{(i)}/3 < \delta_3^{(i)}/(n+3)$ holds for the first time. Until then the examples $(x_2, 1)$ and $(x_1, 1)$ are chosen alternately.

As long as either $(x_1, 1)$ or $(x_2, 1)$ are chosen, the probabilities $\delta_1^{(i)}$, $\delta_2^{(i)}$, $\delta_3^{(i)}$ develop as follows: $\delta_{1+(i+1) \bmod 2}^{(i+1)} = \delta_{1+i \bmod 2}^{(i)}/3$, $\delta_{1+(i+1) \bmod 2}^{(i+1)} = 0$, and $\delta_3^{(i+1)} = \delta_3^{(i)} + \delta_{1+i \bmod 2}^{(i)}/3$. For all these i either $\delta_1^{(i)}$ or $\delta_2^{(i)}$ is greater than zero. We abbreviate this unique positive value by $\delta_{12}^{(i)}$. Resolving the recurrence, we get for $i = 1, 2, \ldots$:

$$\delta_{12}^{(i)} = \delta_2^{(1)}/3^{i-1}, \tag{8.13}$$

$$\delta_3^{(i)} = \delta_3^{(1)} + \sum_{\nu=1}^{i-1} \delta_2^{(1)}/3^{\nu} = \delta_3^{(1)} + \delta_2^{(1)} \cdot \left(\frac{1}{2} - \frac{1}{2 \cdot 3^{i-1}} \right). \tag{8.14}$$

These equations trivially hold for round $i = 1$ and in addition only for rounds i when in the preceding round

$$\frac{\delta_{12}^{(i-1)}}{3} - \frac{\delta_3^{(i-1)}}{(n+3)} > 0, \tag{$*$}$$

because only then $(x_1, 1)$ or $(x_2, 1)$ is chosen. Now we show that Condition $(*)$ is satisfied for $i = 2, \ldots, 2\ell - 1$. Afterwards we show that it does not hold for $i = 2\ell$.

The proof is by induction on i. For the induction basis let $i = 2$. Then $\delta_{12}^{(1)}/3 = \delta_2^{(1)}/3 = 1/9$ and $\delta_3^{(1)}/(n+3) = (1/3)/(n+3) < 1/9$ since $n + 3 > 3$. Thus $(*)$ holds.

Now assume $(*)$ for some $i \in \{2, \ldots, 2\ell - 2\}$. Then we have $\delta_{12}^{(i+1)}/3 = \delta_2^{(1)}/3^{i+1} = 1/3^{i+2}$ and $\delta_3^{(i+1)}/(n+3) = 1/3 + 1/3 \cdot \left(\frac{1}{2} - \frac{1}{2 \cdot 3^i}\right)$. Subtracting the second term from the first and plugging in n, yields

$$\frac{\delta_{12}^{(i+1)}}{3} - \frac{\delta_3^{(i+1)}}{(n+3)} = \frac{3^{-2-i} \cdot (2 - 3^{2+i} + 3^{2\ell})}{3^{2\ell} - 1} \tag{8.15}$$

which is positive because $i \leq 2\ell - 2$. Therefore $(*)$ holds for $i+1$.

For $i = 2\ell - 1$ we can check $(*)$ using Equation (8.15) again. We obtain

$$\frac{\delta_{12}^{(i+1)}}{3} - \frac{\delta_3^{(i+1)}}{(n+3)} = \frac{3^{-1-2\ell} \cdot (2 - 3^{1+2\ell} + 3^{2\ell})}{3^{2\ell} - 1} = \frac{3^{-1-2\ell} \cdot (2 - 2 \cdot 3^{2\ell})}{3^{2\ell} - 1}$$

which is negative. This means $(*)$ is not satisfied for $i = 2\ell - 1$.

So far we have shown that $(x_1, 1)$ and $(x_2, 1)$ are chosen alternately until round $2\ell - 1$, that is, $T(i) = (x_{1+i \bmod 2}, 1)$ for $i = 1, \ldots, 2\ell - 1$. The state at round 2ℓ is $\delta_1^{(2\ell)} = 1/3^{2\ell}$, $\delta_2^{(2\ell)} = 0$, $\delta_3^{(2\ell)} = 1/3 + 1/3 \cdot (\frac{1}{2} - \frac{1}{2 \cdot 3^{2\ell-1}})$. Then $(x_3, 1)$ is the greedy choice: $T(2\ell) = (x_3, 1)$. The resulting state is $\delta_1^{(2\ell+1)} = \delta_1^{(2\ell)} + \delta_3^{(2\ell)}/(n+3)$, $\delta_2^{(2\ell+1)} = \delta_4^{(2\ell+1)} = \cdots = \delta_{n+3}^{(2\ell+1)} = \delta_3^{(2\ell)}/(n+3)$, $\delta_3^{(2\ell+1)} = 0$.

It remains to show that $T(2\ell + 1) = (x_1, 1)$ and the statement about $\delta^{(r_{k+1}+1)} = \delta^{(2\ell+2)}$. In state $\delta^{(2\ell+1)}$ the example $(x_3, 1)$ is certainly not the greedy choice because $\delta_3^{(2\ell+1)} = 0$. From the two remaining examples, $(x_1, 1)$ activates more probability mass and is thus the greedy choice. Therefore $T(2\ell + 1) = (x_1, 1)$. For the state $\delta^{(2\ell+2)}$ it follows:

$$\begin{aligned}
\delta_1^{(2\ell+2)} &= 0, \\
\delta_2^{(2\ell+2)} &= \delta_2^{(2\ell+1)} + (\delta_1^{(2\ell+1)} + n \cdot \delta_4^{(2\ell+1)})/3 \\
&= \frac{3^{-1-2\ell}((-9 + 4 \cdot 9^\ell + 81^\ell) \cdot \delta_2^{(1)} + 2(9^\ell + 81^\ell) \cdot \delta_3^{(1)})}{2(-1 + 9^\ell)}, \\
\delta_3^{(2\ell+2)} &= (\delta_1^{(2\ell+1)} + n \cdot \delta_4^{(2\ell+1)})/3 \\
&= \frac{3^{-1-2\ell}((9 - 2 \cdot 9^\ell + 81^\ell) \cdot \delta_2^{(1)} + 2(-5 \cdot 9^\ell + 81^\ell) \cdot \delta_3^{(1)})}{2(-1 + 9^\ell)}, \\
\delta_4^{(2\ell+2)} = \cdots = \delta_{n+3}^{(2\ell+2)} &= 0.
\end{aligned}$$

We have now proven the induction basis for Claim 2.1. For the induction step, assume that the claim holds for some $k \geq 0$. We show it for $k+1$. In principle, the proof for $k+1$ works exactly as the induction basis proof. However, instead of starting in a state $\delta^{(1)}$ with known probabilities ($\delta_1^{(1)} = 0$, $\delta_2^{(1)} = \delta_3^{(1)} = 1/3$), we start in a state $\delta^{(r_{k+1}+1)}$ about which we only know Equations (8.9)–(8.12). The whole argument thus becomes more complicated.

By the induction hypothesis we know that $T(r_{k+1}) = T(r_k + 2\ell + 1) = (x_1, 1)$. This corresponds to $T(0) = (x_1, 1)$ in the induction base. Beginning with round $r_{k+1} + 1$ the teacher gives the examples $(x_2, 1)$, $(x_1, 1)$ alternately until $(x_3, 1)$ would be more greedy. At what round this happens can be checked via a condition similar to $(*)$. Again we abbreviate the unique positive value of δ_1 and δ_2 by δ_{12}. Then we get the following equations for $i = 1, 2, \ldots$ (cf. (8.13), (8.14)):

$$\delta_{12}^{(r_{k+1}+i)} = \delta_2^{(r_{k+1}+1)}/3^{i-1}, \tag{8.16}$$

$$\begin{aligned}\delta_3^{(r_{k+1}+i)} &= \delta_3^{(r_{k+1}+1)} + \sum_{\nu=1}^{i-1} \delta_2^{(r_{k+1}+1)}/3^\nu \\ &= \delta_3^{(r_{k+1}+1)} + \delta_2^{(r_{k+1}+1)} \cdot \left(\frac{1}{2} - \frac{1}{2 \cdot 3^{i-1}}\right). \end{aligned} \tag{8.17}$$

These equations hold trivially for round $r_{k+1} + i = r_{k+1} + 1$ and in addition only for rounds i when in the preceding round

$$\frac{\delta_{12}^{(r_{k+1}+i-1)}}{3} - \frac{\delta_3^{(r_{k+1}+i-1)}}{n+3} > 0, \tag{$**$}$$

because only then $(x_1, 1)$ or $(x_2, 1)$ is chosen. Now we show that condition $(**)$ is satisfied for $i = 2, \ldots, 2\ell - 1$. Afterwards we show that it does not hold for $i = 2\ell$.

Again we use induction on i. For $i = 2$ the left hand side of condition $(**)$ becomes (using Equations (8.10) and (8.11)):

$$\frac{3^{-3-2\ell}}{2(-1+3^{2\ell})^2} \cdot \Big((-99 - 7 \cdot 3^{1+4\ell} - 3^{2\ell} + 3^{6\ell}) \cdot \delta_2^{(r_k+1)} + 2(-8 \cdot 3^{1+4\ell} + 83 \cdot 3^{2\ell} + 3^{6\ell}) \cdot \delta_3^{(r_k+1)} \Big).$$

In this expression, the coefficients of $\delta_2^{(r_k+1)}$ and $\delta_3^{(r_k+1)}$ are positive for $\ell \geq 2$ because the summand $3^{6\ell}$ dominates. Therefore the whole expression is positive, which means that $(**)$ holds for $i = 2$.

For the induction step, assume that $(**)$ is satisfied for some $i \in \{2, \ldots, 2\ell - 2\}$. Then using Equations (8.16) and (8.17) for i, in addition to (8.10) and (8.11), we get as left hand side of condition $(**)$ for $i + 1$:

$$\frac{3^{-2-2\ell-i}}{2(-1+3^{2\ell})^2} \cdot \Big((-18 + 2 \cdot 3^{1+4\ell} - 3^{3+i} - 3^{2+4\ell+i} - 3^{2\ell} + 3^{6\ell}) \cdot \delta_2^{(r_k+1)} + 2(3^{1+4\ell} + 3^{3+2\ell+i} - 3^{2+4\ell+i} + 2 \cdot 3^{2\ell} + 3^{6\ell}) \cdot \delta_3^{(r_k+1)} \Big).$$

Again we show that the coefficients of $\delta_2^{(r_k+1)}$ and $\delta_3^{(r_k+1)}$ are positive. The coefficient of $\delta_2^{(r_k+1)}$ is

$$\begin{aligned} -18 + 2 \cdot 3^{1+4\ell} - 3^{3+i} - 3^{2+4\ell+i} - 3^{2\ell} + 3^{6\ell} &> -18 + 2 \cdot 3^{1+4\ell} - 3^{2\ell+1} - 3^6 - 3^{2\ell} + 3^{6\ell} \\ &= -18 + 2 \cdot 3^{1+4\ell} - 4 \cdot 3^{2\ell} \\ &> -18 + 2 \cdot 3^{1+4\ell} - 2 \cdot 3^{1+2\ell} \end{aligned}$$

where the first inequality holds because $i \leq 2\ell - 2$. Thus, the coefficient of $\delta_2^{(r_k+1)}$ is positive for $\ell \geq 2$. The coefficient of $\delta_3^{(r_k+1)}$ is (ignoring the factor 2):

$$3^{1+4\ell} + 3^{3+2\ell+i} - 3^{2+4\ell+i} + 2 \cdot 3^{2\ell} + 3^{6\ell} \;>\; 3^{1+4\ell} + 3^{3+2\ell+i} - 3^{6\ell} + 2 \cdot 3^{2\ell} + 3^{6\ell}$$

$$\begin{aligned} &= 3^{1+4\ell} + 3^{3+2\ell+i} + 2 \cdot 3^{2\ell} \\ &> 0. \end{aligned}$$

The first inequality is again due to $i \leq 2\ell - 2$. This shows that $(**)$ is true for $i+1$, too.

Condition $(**)$ is not satisfied any more when $i = 2\ell$. In this case the left hand side of $(**)$ is

$$\frac{3^{-1-4\ell}}{2(-1+3^{2\ell})^2} \cdot \Big((-18 + 2 \cdot 3^{1+4\ell} - 3^{1+6\ell} - 10 \cdot 3^{2\ell} + 3^{6\ell}) \cdot \delta_2^{(r_k+1)} + 2(4 \cdot 3^{1+4\ell} - 3^{1+6\ell} + 2 \cdot 3^{2\ell} + 3^{6\ell}) \cdot \delta_3^{(r_k+1)} \Big).$$

Similar as above, the coefficients of $\delta_2^{(r_k+1)}$ and $\delta_3^{(r_k+1)}$ can be seen to be negative, as the summand $-3^{1+6\ell}$ dominates.

We have now shown that the greedy teacher gives example $(x_2, 1), (x_1, 1)$ alternately until round $r_{k+1} + 2\ell$, that is, $T(r_{k+1} + i) = (x_{1+i \bmod 2}, 1)$ for $i = 1, \ldots, 2\ell - 1$. In state $\delta^{(r_{k+1}+2\ell)}$ condition $(**)$ is not true and thus $T(r_{k+1} + 2\ell) = (x_3, 1)$. For the resulting state we get $\delta_1^{(r_{k+1}+2\ell+1)} = \delta_1^{(r_{k+1}+2\ell)} + \delta_3^{(r_{k+1}+2\ell)}/(n+3)$, $\delta_2^{(r_{k+1}+2\ell+1)} = \delta_4^{(r_{k+1}+2\ell)} = \cdots = \delta_{n+3}^{(r_{k+1}+2\ell+1)} = \delta_3^{(r_{k+1}+2\ell)}/(n+3)$, $\delta_3^{(r_{k+1}+2\ell+1)} = 0$. In this state, the greedy teacher must chose the example $(x_1, 1)$, hence $T(r_{k+1} + 2\ell + 1) = (x_1, 1)$. For the state $\delta^{(r_{k+1}+2\ell+2)} = \delta^{(r_{k+2}+1)}$ it follows:

$$\begin{aligned} \delta_1^{(r_{k+2}+1)} &= 0, \\ \delta_2^{(r_{k+2}+1)} &= \delta_2^{(r_{k+1}+2\ell+1)} + (\delta_1^{(r_{k+1}+2\ell+1)} + n \cdot \delta_4^{(r_{k+1}+2\ell+1)})/3 \\ &= \frac{3^{-1-2\ell}((-9 + 4 \cdot 9^\ell + 81^\ell) \cdot \delta_2^{(r_{k+1}+1)} + 2(9^\ell + 81^\ell) \cdot \delta_3^{(r_{k+1}+1)})}{2(-1+9^\ell)}, \\ \delta_3^{(r_{k+2}+1)} &= (\delta_1^{(r_{k+1}+2\ell+1)} + n \cdot \delta_4^{(r_{k+1}+2\ell+1)})/3 \\ &= \frac{3^{-1-2\ell}((9 - 2 \cdot 9^\ell + 81^\ell) \cdot \delta_2^{(1)} + 2(-5 \cdot 9^\ell + 81^\ell) \cdot \delta_3^{(r_{k+1}+1)})}{2(-1+9^\ell)}, \\ \delta_4^{(r_{k+2}+1)} &= \ldots = \delta_{n+3}^{(r_{k+2}+1)} = 0. \end{aligned}$$

This finishes the induction step and the proof of Claim 2.1. □ Claim 2.1

Now we know that the greedy teacher is a cyclic teacher and iterates through the example sequence $\langle (x_1, 1), (x_2, 1) \rangle^\ell \circ \langle (x_3, 1) \rangle$ of length $2\ell + 1$. □ Claim 2

Claim 3: For all $\ell \geq 1$, the cyclic teacher with example sequence $\langle (x_1, 1), (x_2,) \rangle^\ell \circ \langle (x_3, 1) \rangle$ has an expected teaching time of

$$\frac{-n + 9^\ell(16 + 5n) + 4\ell(1 + 3^{2\ell+1} + 9^\ell n)}{2(-1+9^\ell)(4+n)}.$$

Proof. We use the idea of the proof of Lemma 8.17. We denote the $2\ell+1$ examples by $z_0,\ldots,z_{2\ell}$ and by F_i the expected number of rounds when teaching would start with example z_i $(i=0,\ldots,2\ell)$. We then get the following $2\ell+1$ equations:

$$\begin{aligned}
F_0 &= 1+\frac{1}{3}\cdot F_1+\frac{1}{3}\cdot(F_{2\ell}+2\ell-1), \qquad (8.18)\\
F_1 &= 1+\frac{1}{3}\cdot F_2+\frac{1}{3}\cdot(F_{2\ell}+2\ell-2),\\
F_2 &= 1+\frac{1}{3}\cdot F_3+\frac{1}{3}\cdot(F_{2\ell}+2\ell-3),\\
&\vdots \qquad\qquad \vdots\\
F_{2\ell-2} &= 1+\frac{1}{3}\cdot F_{2\ell-1}+\frac{1}{3}\cdot(F_{2\ell}+1),\\
F_{2\ell-1} &= 1+\frac{1}{3}\cdot F_{2\ell}+\frac{1}{3}\cdot(F_0+1),\\
F_{2\ell} &= 1+\frac{n+1}{n+3}\cdot F_0+\frac{1}{n+3}\cdot(F_1+1). \qquad (8.19)
\end{aligned}$$

Now we plug F_1 into the first equation, then F_2 into the result and so on until $F_{2\ell-1}$. Then we arrive at an equation containing only F_0 and $F_{2\ell}$:

$$\begin{aligned}
F_0 &= 1+\frac{1}{3}+\left(\frac{1}{3}\right)^2 F_2+\left(\frac{1}{3}\right)^2\cdot(F_{2\ell}+2\ell-2)+\frac{1}{3}\cdot(F_{2\ell}+2\ell-1)\\
&= 1+\frac{1}{3}+\left(\frac{1}{3}\right)^2+\left(\frac{1}{3}\right)^3 F_3+\left(\frac{1}{3}\right)^3\cdot(F_{2\ell}+2\ell-3)+\left(\frac{1}{3}\right)^2\cdot(F_{2\ell}+2\ell-2)\\
&\quad+\frac{1}{3}\cdot(F_{2\ell}+2\ell-1)+\frac{1}{3}\cdot(F_{2\ell}+2\ell-1)\\
&= \ldots\\
&= \sum_{i=0}^{2\ell-2}\left(\frac{1}{3}\right)^i+\left(\frac{1}{3}\right)^{2\ell-1}\cdot F_{2\ell-1}+\sum_{i=1}^{2\ell-1}\left(\frac{1}{3}\right)^i(F_{2\ell}+2\ell-i)\\
&= \sum_{i=0}^{2\ell-1}\left(\frac{1}{3}\right)^i+\sum_{i=1}^{2\ell}\left(\frac{1}{3}\right)^i(F_{2\ell}+2\ell-i)+\left(\frac{1}{3}\right)^{2\ell}\cdot(F_0+1).
\end{aligned}$$

Together with (8.18) and (8.19) we thus have a system of three linear equations over $F_0, F_1, F_{2\ell}$. Solving this system yields

$$\begin{aligned}
F_0 &= \frac{-n+9^\ell(16+5n)+4\ell(1+3^{1+2\ell}+9^\ell n)}{2(-1+9^\ell)(4+n)},\\
F_1 &= \frac{2(1+\ell)(5+3^{1+2\ell}+n+9^\ell n)}{(-1+9^\ell)(4+n)},
\end{aligned}$$

$$F_{2\ell} = \frac{-4 + 8\ell - 3n + 9^\ell(20 + 7n + 4\ell(2+n))}{2(-1+9^\ell)(4+n)}.$$

The value F_0 is the sought expectation. □ Claim 3

Claim 4: The greedy teacher has a teaching time of

$$\mathbb{E}[T, \mathcal{L}_{1,\mathcal{C}_n}, c^*] = \frac{(56 + n(33+5n))\log(3) + (22 + n(13+2n))\log(7+2n)}{(3+n)(4+n)\log(9)}\ .$$

Proof. From $n = \frac{3}{2}(3^{2\ell-1} - 3) + 1$ it follows that $\ell = \log(7+2n)/\log 9$ and plugging ℓ into the expression from Claim 3 yields the proof. □ Claim 4

The numerator of $\mathbb{E}[T, \mathcal{L}_{1,\mathcal{C}_n}, c^*]$ is in $\Theta(n^2 \log n)$ and the denominator in $\Theta(n^2)$. This places the expectation itself in $\Theta(\log n)$. Since the optimal value is at most five, this shows that the greedy teacher provides only a factor $\log n$ approximation for the concept $c^* \in \mathcal{C}_n$. □

Since the optimal teachers in a positive X3C class are all greedy, the problem of finding optimal teaching times for greedy teachers is $\mathcal{NP}$-hard, too. This follows in the same way as Corollary 8.22 from the results about X3C classes in Section 8.3.

Corollary 8.24 *The following problem is $\mathcal{NP}$-hard.*
`OPT-GREEDY-TEACHINGTIME`
Instance: *Concept class $\mathcal{C}$, concept $c^* \in \mathcal{C}$, rational number F.*
Question: *Is there a greedy teacher with expected teaching time of at most F?*

This concludes our investigation of teaching the randomized learner $\mathcal{L}_1$ without feedback. That $\mathcal{L}_1$ is memoryless seems to make it a less realistic model for real-world students. But we expect that most results in this section hold *mutatis mutandis* for the case of larger, but finite, memory, provided the teaching dimension of the target is greater than the memory size. Obtaining such results, however, appears to be even more complicated than in the memoryless case considered here. After all, even in this case many questions remain open. We already mentioned the question of the complexity of `OPT-TEACHINGTIME`. Another open problem is to determine an upper bound of the approximation quality of greedy teachers. Furthermore, the role of cyclic teachers should be explored further. For example, experiments indicate that greedy teachers become cyclic teachers after a finite prefix of examples. For all we know, this limited class of teachers could always contain an optimal teacher.

The limits of our knowledge about this teaching model are probably best demonstrated by the fact that we do not even know an optimal teacher for the class of monomials over 4 variables.

Further research regarding randomized learners and their relation to the non-deterministic learners of Part I is discussed in the following, final chapter.

Chapter 9

Conclusion

In this chapter we summarize our results, discuss some relations to learning theory, and sketch ideas for further work.

9.1 Summary

Our goal was to create formal models for teaching that resemble the scenario of an intelligent tutoring system more closely than traditional models, especially the teaching dimension model. To this end, we pursued two ideas. First, we developed a new framework of non-deterministic learners, within which we devised new, more realistic models of the students. Second, we analyzed the performance of the teacher in a novel way, namely by replacing the worst case learner with a randomized average case learner.

Both approaches improve the realism of the teaching dimension model in that they make the teaching performance of a teacher depend on the order in which the teacher presents the examples, on the amount of feedback given by the learners, and on the memory size of the learners. The two approaches achieve this in different ways and to a different extent. To highlight the differences we compare both approaches.

The influence of varying memory size is rather weak in the non-deterministic framework. Most often a target is teachable if and only if the learner's memory has at least a certain size. The Lemmas 4.4 and 4.29 demonstrate this for the hypothesis space restriction model. In the hypothesis change restriction model this effect occurs for 1-decision lists (see Fact 5.10), but not for the monomials with a certain neighborhood relation, as shown in Fact 5.8. When this effect occurs it is caused by the teacher having to force all computations of the non-deterministic learner into the target state. Typically, some transitions on this way can only be enforced by providing sufficiently many examples to the learner. If the memory size is too small, there is a learner that can avoid this transition and thus avoid reaching the target. By contrast, in the randomized framework the performance of a teacher varies smoothly with the memory size of the learner (see Figure 6.1). This is because the more examples the learner memorizes the higher the probability of reaching the target.

The influence of feedback can be rather strong in the hypothesis change restriction model (see Section 5.2). The reason is that giving a certain example to learners in different states can have very different effects on them. In the randomized model the influence is generally much weaker because a certain example can have only two

possible effects, namely triggering a hypothesis change or not. Moreover, triggering a hypothesis change always has the same consequences, regardless of the learner's current state. That feedback effects tend to be small in the randomized model is shown by Lemma 7.10, which states that in the case of infinite memory the absence of feedback at most doubles the expected teaching time.

For the order of examples similar things can be said as for feedback. The order can be crucial in the hypothesis change restriction model because one wrong example at the wrong time can lead the learner into complete failure. In the randomized model arranging the examples differently often causes the teaching time to change only a little (see Figure 6.2).

In the non-deterministic model one wrong example can lead a learner into a dead end, which makes it impossible for the teacher to succeed for all learners. The impact of unsuited examples can thus be fatal, which is not realistic. In the randomized model teaching success can be achieved with probability one, regardless of the situation the learner is in; but the influence of bad examples or a bad order of examples is rather unrealistically small. In reality it seems that on the one hand teaching should be possible regardless of the learner's situation, as in the randomized model, and on the other hand unsuited examples or a badly laid out curriculum should cause a much higher teaching time than a well laid out curriculum. Our results strongly suggest that this kind of behavior can only be achieved if we combine both approaches.

A combined model would allow the learner to switch from every state to every other state with positive probability. The probability distribution would not be uniformly, but biased towards the neighborhood hypotheses in the underlying non-deterministic model. Note that here the randomization has two tasks. First it smoothes out the worst case behavior (and the best case behavior) of the students; second it smoothes out the distinction between admissible and non-admissible hypotheses, preventing dead end situations for the learner.

Our analysis of both single approaches themselves indicates that such a combined model would indeed combine the advantages of the models and would allow to model more complex phenomena occurring in real life. On the other hand, our results also indicate that the resulting model would be much harder to analyze. For instance, the optimality criteria (Lemmas 7.1 and 7.8, and Corollary 8.7) would become even more complicated if we allow the probability distributions to be non-uniform and to be different in every state of the learner. Even in our original randomized model in Chapter 8 we were unable to find an optimal teacher for the simple class of monomials. Thus, unsurprisingly, greater realism comes at the price of greater complexity.

9.2 Relations between Teaching and Learning

Comparing concepts and classes with respect to their teachability and learnability suggests itself, but is not as straightforward as it may look. A fundamental difference

between teachability and learnability is that learnability is typically measured for concept classes, whereas teachability is measured for single concepts. The latter claim holds even though the teaching dimension and average teaching dimension are defined for classes, too. But these dimension values are not defined by the model itself, rather they are calculated afterwards as an "add-on measure." The pure teaching models always measure performance for a certain single target concept. The learnability of a certain concept, on the other hand, cannot be defined as the performance of an optimal learner, because for every fixed target concept there is a learner that guesses this target right from the beginning. Learning makes sense only when a whole class of possible concepts is involved. Therefore, comparisons of learnability measures, such as the VC-dimension or the number of membership queries, with teachability measures are somewhat hard to interpret.

In addition to these intrinsic obstacles, our teachability measures are hard to compare with classical learnability measures because the former often do not depend on the concept class alone, but also on additional assumptions such as a complexity measure or a neighborhood relation over the concepts.

Teaching is connected with another fundamental question in learning theory, namely the question of the information content of a sample. Since good teachers are supposed to present informative samples, it seems natural to measure the information content of a sample by the performance of a teacher using this sample. In particular, teaching sets can be considered to have a high information content. We did not focus on this connection and much more research is necessary to explore it, but nevertheless our results give two suggestions. First, the information content depends on the receiver of the sample, in this case the learner. We have seen in the case of 1-decision lists that replacing plain consistent learners with complexity based learners radically changes the size and the form of the teaching sets (or $\mathfrak{H}$-sets, respectively). Second, our results discourage the use of the randomized models for determining the information content. This is because the randomized learner is sensitive to the order in which the examples are given (see Figure 6.2). Moreover, the optimal teaching times of teachers in the randomized models are very similar to the optimal teaching times of teachers in the corresponding non-deterministic model (see Lemma 7.10). Thus the randomized models would not yield very different measures for the information content, but they would be much more difficult to use.

9.3 Extensions of Our Models and Further Work

As we already mentioned above, a combination of hypothesis change restrictions with randomization most likely yields a more realistic teaching model at the price of higher complexity. But there are more ways in which the models can be varied either to make them match the intelligent tutoring setting more closely or for more theoretical reasons, similarly to variations occurring in learning models. Sometimes such variations

require bigger structural changes in the framework than only adjusting the probability distribution of follow-up hypotheses. In the following we will just mention some ideas.

A classical variation is to require the teacher to present positive examples only. This is intended for studying language acquisition because languages are learned primarily by sentences or phrases that are correct and thus belong to the language to be learned. In another variant the teacher could be allowed to give wrong examples, that is, examples that are not consistent with the target concept. This is similar to a human teacher who gives a rough intuition or oversimplified explanation that is, strictly speaking, wrong, but nevertheless helps the students understand the subject matter.

Besides modifications of the teacher, also the interaction between learner and teacher could resemble reality more closely. After all, intelligent tutoring systems typically use other kinds of feedback than mind reading, which is the only kind of feedback in our models. Information about the current hypothesis can be gathered by asking the student questions or by "watching" the student solve exercises proposed by the system. To some extent the student might even be able to ask the system questions. These would also reveal something about the student's current hypothesis. These ways of interacting can be formalized rather easily. For example, the teacher could choose to give an instance, without label, to the learner that answers with a label according to the current hypothesis. This would partially reveal the hypothesis to the teacher and as such be a middle course between no feedback and complete feedback.

The largest room for improvement still lies in the student model. The memory could be modeled more realistically by separating it into a short term memory and a long term memory. In addition, its behavior could be randomized similar to the hypothesis changes. The hypothesis changes are probably the most difficult part of the learner to get "right." Determining the probabilities of hypothesis changes in a human requires a deep psychological analysis of the human learning behavior, which theoretical computer science, obviously, cannot provide. It can, however, provide a framework for the analysis of any realistic student model a psychologist may come up with.

All these modifications to the randomized model can be regarded as special cases of (unobserved) stochastic shortest path problems, albeit with a larger or more complex state space than our models in Part II had.

All our models and also all extensions suggested above do not require the teacher to pay attention to the history of the teaching process. But an intelligent tutoring system, just as a human teacher, should consider the student's past behavior when deciding how to teach the student in the future. This would result in an interesting intertwining of teaching and learning. The teacher would not only teach the target, but also learn the learner's behavior; the learner would not only learn the target, but also in a sense teach its own behavior to the teacher. When we defined our models, we took special measures to avoid what we called history awareness. Defining the learners as non-deterministic or randomized automata effectively makes it impossible to predict

their future behavior from their past. A remedy could be to define a set of deterministic learners all of which must be taught. The learners in such a set would have to have individual behaviors that are worthwhile to be learned by the teacher. Devising a natural set of such teachers seems difficult. Moreover, the average case analysis via a single randomized teacher would not work any more. A remedy could be to perform an average case analysis by averaging over the teacher's performance over all learners.

Discussing all these possible extensions to our basic framework may give the impression that our teaching models are far too simplistic for humans. While this is true, we reply that these models are only the first steps and certainly considerably more realistic than all previous models. It could also be mentioned that in other fields the models for human behavior are even more simple, yet they are worthwhile to study and produce interesting results. For example, in the game theoretic models of auctions, a bidder is fully characterized by a single number representing the highest bid the bidder would make for the auctioned object. Even with this simple model of the bidder, meaningful comparisons of various kinds of auctions are possible.

Regarding models for intelligent tutoring systems, there is a large gap between what theoretical computer science provides and what is a psychologically realistic and desirable model. As it is certainly impossible to bridge this gap between teaching theory and practice in one big step, continuous efforts are necessary from both sides. With this thesis we believe to have done one step from the theoretical side of the gap.

Bibliography

[1] D. Angluin. Finding patterns common to a set of strings. *J. of Comput. Syst. Sci.*, 21(1):46–62, 1980.

[2] D. Angluin. Inductive inference of formal languages from positive data. *Inform. Control*, 45(2):117–135, May 1980.

[3] D. Angluin. Queries and concept learning. *Machine Learning*, 2(4):319–342, Apr. 1988.

[4] D. Angluin. Queries revisited. In *Algorithmic Learning Theory, 12th International Conference, ALT 2001, Washington, DC, USA, November 25–28, 2001, Proceedings*, volume 2225 of *Lecture Notes in Artificial Intelligence*, pages 12–31. Springer, 2001.

[5] D. Angluin. Queries revisited. *Theoret. Comput. Sci.*, 313(2):175–194, 2004. Special issue for ALT 2001.

[6] D. Angluin and M. Krikis. Teachers, learners and black boxes. In *Proc. 10th Annu. Conf. on Comput. Learning Theory*, pages 285–297. ACM Press, New York, NY, 1997.

[7] D. Angluin and M. Kriķis. Learning from different teachers. *Machine Learning*, 51(2):137–163, 2003.

[8] M. Anthony, G. Brightwell, D. Cohen, and J. Shawe-Taylor. On exact specification by examples. In *Proc. 5th Annual ACM Workshop on Comput. Learning Theory*, pages 311–318. ACM Press, New York, NY, 1992.

[9] M. Anthony, G. Brightwell, and J. Shawe-Taylor. On specifying boolean functions by labelled examples. *Discrete Applied Mathematics*, 61(1):1–25, 1995.

[10] G. Ausiello, P. Crescenzi, G. Gambosi, V. Kann, A. Marchetti-Spaccamela, and M. Protasi. *Complexity and Approximation. Combinatorial optimization problems and their approximability properties.* Springer, 1999.

[11] F. J. Balbach. Teaching classes with high teaching dimension using few examples. In *Learning Theory: 18th Annual Conference on Learning Theory, COLT 2005, Bertinoro, Italy, June 27-30, 2005, Proceedings*, volume 3559 of *Lecture Notes in Artificial Intelligence*, pages 668–683. Springer, 2005.

[12] F. J. Balbach and T. Zeugmann. Teaching learners with restricted mind changes. In *Algorithmic Learning Theory, 16th International Conference, ALT 2005, Singapore, October 2005, Proceedings*, volume 3734 of *Lecture Notes in Artificial Intelligence*, pages 474–489. Springer, Oct. 2005.

[13] F. J. Balbach and T. Zeugmann. Teaching memoryless randomized learners without feedback. In *Algorithmic Learning Theory, 17th International Conference, ALT 2006, Barcelona, Spain, October 2006, Proceedings*, volume 4264 of *Lecture Notes in Artificial Intelligence*, pages 93–108. Springer, Oct. 2006.

[14] F. J. Balbach and T. Zeugmann. Teaching randomized learners. In *Learning Theory: 19th Annual Conference on Learning Theory, COLT 2006, Pittsburgh, PA, USA, June 2006, Proceedings*, volume 4005 of *Lecture Notes in Artificial Intelligence*, pages 229–243. Springer, Berlin, 2006.

[15] B. Becker and H.-U. Simon. How robust is the n-cube? *Information and Computation*, 77(2):162–178, 1988.

[16] S. Ben-David and N. Eiron. Self-directed learning and its relation to the VC-dimension and to teacher-directed learning. *Machine Learning*, 33(1):87–104, 1998.

[17] D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 2. Athena Scientific, second edition, 2001.

[18] D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 1. Athena Scientific, third edition, 2005.

[19] D. P. Bertsekas and J. N. Tsitsiklis. An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16:580–595, Oct 1991.

[20] V. D. Blondel and V. Canterini. Undecidable problems for probabilistic automata of fixed dimension. *Theory of Computing Systems*, 36(3):231–245, 2003.

[21] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Occam's razor. *Inform. Proc. Lett.*, 24:377–380, Apr. 1987.

[22] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. ACM*, 36(4):929–965, 1989.

[23] J. Case, S. Jain, S. Lange, and T. Zeugmann. Incremental concept learning for bounded data mining. *Inform. Comput.*, 152(1):74–110, 1999.

[24] J. Castro and J. L. Balcázar. Simple PAC learning of simple decision lists. In *Algorithmic Learning Theory, 6th International Workshop, ALT '95, Fukuoka, Japan, October 18-20, 1995, Proceedings*, volume 997 of *Lecture Notes in Artificial Intelligence*, pages 239–248. Springer, 1995.

[25] I. A. Comenius. Machina didactica. In O. Chlup, editor, *Opera Omnia Didactica*. Czechoslovak Academy of Science, Prague, 1957. German translation: `www.didactools.de/comenius/machdidk.htm`.

[26] C. Derman. *Finite State Markovian Decision Processes*, volume 67 of *Mathematics in Science and Engineering*. Academic Press, 1970.

[27] A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant. A general lower bound on the number of examples needed for learning. *Inform. Comput.*, 82(3):247–261, 1989.

[28] U. Feige. A threshold of ln n for approximating set cover. *Journal of the ACM*, 45(4):634–652, 1998.

[29] R. Freivalds, E. B. Kinber, and R. Wiehagen. Inductive inference from good examples. In *Analogical and Inductive Inference, International Workshop AII '89, Reinhardsbrunn Castle, GDR, October 1989, Proceedings*, volume 397 of *Lecture Notes in Artificial Intelligence*, pages 1–17. Springer-Verlag, 1989.

[30] R. Freivalds, E. B. Kinber, and R. Wiehagen. On the power of inductive inference from good examples. *Theoret. Comput. Sci.*, 110(1):131–144, 1993.

[31] Q. Gao and M. Li. An application of minimum description length principle to online recognition of handprinted alphanumerals. In *Proc. 11th International Joint Conference on Artificial Intelligence*, pages 843–848. Morgan Kaufmann, 1989.

[32] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1979.

[33] E. M. Gold. Language identification in the limit. *Inform. Control*, 10(5):447–474, 1967.

[34] S. A. Goldman and M. J. Kearns. On the complexity of teaching. *J. of Comput. Syst. Sci.*, 50(1):20–31, 1995.

[35] S. A. Goldman and H. D. Mathias. Teaching a smarter learner. *J. of Comput. Syst. Sci.*, 52(2):255–267, 1996.

[36] S. A. Goldman, R. L. Rivest, and R. E. Schapire. Learning binary relations and total orders. *SIAM J. Comput.*, 22(5):1006–1034, Oct. 1993.

[37] S. Goldwasser, S. Micali, and C. Rackoff. The knowledge complexity of interactive proofs. In *Proceedings of the 17th ACM Symposium on Theory of Computing*, pages 291–304, 1985.

[38] D. Guijarro, V. Lavin, and V. Raghavan. Monotone term decision lists. *Theoret. Comput. Sci.*, 259(1-2):549–575, 2001.

[39] T. Hancock, T. Jiang, M. Li, and J. Tromp. Lower bound on learning decision lists and trees. *Inform. Comput.*, 126(2):114–122, 1996.

[40] T. Hegedűs. Combinatorial results on the complexity of teaching and learning. In *Proc. Mathematical Foundations of Computer Science 1994, 19th International Symposium*, volume 841 of *Lecture Notes in Computer Science*, pages 393–402. Springer-Verlag, 1994.

[41] T. Hegedűs. Generalized teaching dimensions and the query complexity of learning. In *Proc. 8th Annu. Conf. on Comput. Learning Theory*, pages 108–117. ACM Press, New York, NY, 1995.

[42] M. Ikeda, K. Ashlay, and T.-W. Chan, editors. *Intelligent Tutoring Systems, 8th International Conference, ITS 2006 Jhongli, Taiwan, June 26-30, 2006 Proceedings*, volume 4053 of *Lecture Notes in Computer Science*. Springer, 2006.

[43] Intelligent tutoring systems, American Association for Artificial Intelligence, 2006. `www.aaai.org/AITopics/html/tutor.html`.

[44] J. Jackson and A. Tomkins. A computational model of teaching. In *Proc. 5th Annual ACM Workshop on Comput. Learning Theory*, pages 319–326. ACM Press, New York, NY, 1992.

[45] S. Jain, D. Osherson, J. S. Royer, and A. Sharma. *Systems that Learn: An Introduction to Learning Theory, second edition*. MIT Press, Cambridge, Massachusetts, 1999.

[46] C. Kuhlmann. On teaching and learning intersection-closed concept classes. In *Computational Learning Theory, 4th European Conference, EuroCOLT '99, Nordkirchen, Germany, March 29-31, 1999, Proceedings*, volume 1572 of *Lecture Notes in Artificial Intelligence*, pages 168–182. Springer, 1999.

[47] E. Kushilevitz, N. Linial, Y. Rabinovich, and M. Saks. Witness sets for families of binary vectors. *J. Comb. Theory Ser. A*, 73(2):376–380, 1996.

[48] S. Lange, J. Nessel, and R. Wiehagen. Learning recursive languages from good examples. *Annals of Mathematics and Artificial Intelligence*, 23(1/2):27–52, 1998.

[49] S. Lange and T. Zeugmann. Incremental learning from positive data. *J. of Comput. Syst. Sci.*, 53(1):88–103, 1996.

[50] H. Lee, R. A. Servedio, and A. Wan. DNF are teachable in the average case. In *Learning Theory: 19th Annual Conference on Learning Theory, COLT 2006, Pittsburgh, PA, USA, June 2006, Proceedings*, volume 4005 of *Lecture Notes in Artificial Intelligence*, pages 214–228. Springer, Berlin, 2006.

[51] N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, 1988.

[52] O. Madani, S. Hanks, and A. Condon. On the undecidability of probabilistic planning and infinite-horizon partially observable markov decision problems. In *Proceedings AAAI 1999*, pages 541–548, 1999.

[53] D. H. Mathias. DNF - if you can't learn 'em, teach 'em: an interactive model of teaching. In *Proc. 8th Annu. Conf. on Comput. Learning Theory*, pages 222–229. ACM Press, New York, NY, 1995.

[54] H. D. Mathias. A model of interactive teaching. *J. of Comput. Syst. Sci.*, 54(3):487–501, 1997.

[55] A. R. Meyer and L. J. Stockmeyer. Word problems requiring exponential time. In *Proceedings of the 5th ACM Symposium on the Theory of Computing*, pages 1–9, 1973.

[56] B. Natarajan. On learning functions from noise-free and noisy samples via occam's razor. *SIAM J. Comput.*, 29(3):712–727, 1999.

[57] B. K. Natarajan. *Machine Learning: A Theoretical Approach*. Morgan Kaufmann, San Mateo, CA, 1991.

[58] A. L. Oliveira and A. Sangiovanni-Vincentelli. Using the minimum description length principle to infer reduced ordered decision graphs. *Machine Learning*, 25:23–50, 1996.

[59] J. Ong and S. Ramachandran. Intelligent tutoring systems: The what and the how. Online, Feb 2000. `http://www.learningcircuits.org/2000/feb2000/ong.htm`.

[60] D. N. Osherson, M. Stob, and S. Weinstein. *Systems that Learn: An Introduction to Learning Theory for Cognitive and Computer Scientists*. MIT Press, Cambridge, Massachusetts, 1986.

[61] S. D. Patek. On partially observed stochastic shortest path problems. In *Proceedings of the 40-th IEEE Conference on Decision and Control (CDC 2001)*, pages 5050–5055, 2001.

[62] S. D. Patek. Partially observed stochastic shortest path problems with approximate solution by neuro-dynamic programming. Technical Report SIE-030002, Systems and Information Engineering, Univ. of Virginia, 2002. to appear in IEEE Transactions on Systems, Man, and Cybernetics.

[63] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 1994.

[64] R. Raz and S. Safra. A sub-constant error-probability low-degree test, and a sub-constant error-probability PCP characterization of NP. In *Proceedings of the 29th ACM Symposium on Theory of Computing*, pages 475–484, 1997.

[65] R. Reischuk and T. Zeugmann. A complete and tight average-case analysis of learning monomials. In *STACS 99, 16th Annual Symposium on Theoretical Aspects of Computer Science, Trier, Germany, March 1999, Proceedings*, volume 1563 of *Lecture Notes in Computer Science*, pages 414–423. Springer, 1999.

[66] R. Reischuk and T. Zeugmann. An average-case optimal one-variable pattern language learner. *J. Comput. Syst. Sci.*, 60(2):302–335, 2000.

[67] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[68] R. L. Rivest. Learning decision lists. *Machine Learning*, 2:229–246, 1987.

[69] R. L. Rivest and Y. L. Yin. Being taught can be faster than asking questions. In *Proc. 8th Annu. Conf. on Comput. Learning Theory*, pages 144–151. ACM Press, New York, NY, 1995.

[70] K. Romanik. Approximate testing and learnability. In *Proc. 5th Annual ACM Workshop on Comput. Learning Theory*, pages 327–332. ACM Press, New York, NY, 1992.

[71] K. Romanik. Approximate testing and its relationship to learning. *Theoret. Comput. Sci.*, 188(1-2):175–194, 1997.

[72] K. Romanik and C. Smith. Testing geometric objects. In J. Urrutia, editor, *Proceedings of the Second Canadian Conference in Computational Geometry*, pages 14–19, 1990.

[73] P. Rossmanith and T. Zeugmann. Stochastic finite learning of the pattern languages. *Machine Learning*, 44(1/2):67–91, 2001.

[74] G. Seroussi and N. H. Bshouty. Vector sets for exhaustive testing of logic circuits. *IEEE Transactions on Information Theory*, 34(3):513–522, 1988.

[75] R. A. Servedio. On the limits of efficient teachability. *Information Processing Letters*, 79(6):267–272, 2001.

[76] A. Shinohara and S. Miyano. Teachability in computational learning. *New Generation Computing*, 8(4):337–348, 1991.

[77] H. U. Simon. Learning decision lists and trees with equivalence-queries. In *Computational Learning Theory, Second European Conference, EuroCOLT '95, Barcelona, Spain, March 1995, Proceedings*, number 904 in Lecture Notes in Artificial Intelligence, pages 322–336. Springer, 1995.

[78] L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, Nov. 1984.

[79] R. Wiehagen and T. Zeugmann. Ignoring data may be the only way to learn efficiently. *J. of Experimental and Theoret. Artif. Intell.*, 6(1):131–144, 1994.

[80] R. Wiehagen and T. Zeugmann. Learning and consistency. In *Algorithmic Learning for Knowledge-Based Systems*, volume 961 of *Lecture Notes in Artificial Intelligence*, pages 1–24. Springer, 1995.

[81] T. Zeugmann. Lange and Wiehagen's pattern language learning algorithm: An average-case analysis with respect to its total learning time. *Annals of Mathematics and Artificial Intelligence*, 23:117–145, 1998.

Curriculum Vitae

Personal Data

Name:	Frank Jürgen Balbach
Date of birth:	11 September 1976
Place of birth:	Ludwigshafen am Rhein, Germany

Education and Work Experience

08/1983 – 07/1987	Grundschule Neuhofen (primary school)
09/1987 – 06/1996	Gymnasium Schifferstadt (secondary school)
07/1996 – 08/1997	Military service (Wehrdienst) in Bexbach and Speyer
10/1997 – 11/2002	Study of Computer Science (Informatik) with minor subject Mathematics (Mathematik) at the University Kaiserslautern
11/2002	Diploma in Computer Science (Diplom-Informatiker)
12/2002 – 02/2003	Research associate at the Research Group Algorithmic Learning at the University Kaiserslautern
04/2003 – 02/2007	Research associate at the Institute for Theoretical Computer Science at the University Lübeck
02/2007	Ph. D. in Computer Science (Dr. rer. nat.)