TCS-TR-A-**05-5**

# TCS Technical Report

# Teaching Learners that can only Perform Restricted Mind Changes

**by**

Frank J. Balbach and Thomas Zeugmann

**Division of Computer Science**

**Report Series A**

July 18, 2005

Hokkaido University
Graduate School of
Information Science and Technology

Email: thomas@ist.hokudai.ac.jp

Phone: +81-011-706-7684
Fax: +81-011-706-7684

# Teaching Learners that can only Perform Restricted Mind Changes

Frank J. Balbach
*Institut für Theoretische Informatik*
*Universität zu Lübeck*
*Ratzeburger Allee 160*
*23538 Lübeck, Germany*
balbach@tcs.uni-luebeck.de

Thomas Zeugmann
*Division of Computer Science*
*Hokkaido University*
*N-14, W-9*
*Sapporo 060-0814, Japan*
thomas@ist.hokudai.ac.jp

**Abstract**

Within algorithmic learning theory teaching has been studied in various ways. In a common variant the teacher has to teach all learners that are restricted to output only consistent hypotheses. The complexity of teaching is then measured by the maximum number of mistakes a consistent learner can make until successful learning. This is equivalent to the so-called teaching dimension. However, many interesting concept classes have an exponential teaching dimension and it is only meaningful to consider the teachability of finite concept classes.

A refined approach of teaching is proposed by introducing a neighborhood relation over all possible hypotheses. The learners are then restricted to choose a new hypothesis from the neighborhood of their current one. Teachers are either required to teach finitely or in the limit. Moreover, the variant that the teacher receives the current hypothesis of the learner as *feedback* is considered.

The new models are compared to existing ones and to one another in dependence of the neighborhood relations given. In particular, it is shown that feedback can be very helpful. Moreover, within the new model one can also study the teachability of infinite concept classes with potentially infinite concepts such as languages. Finally, it is shown that in our model teachability and learnability can be rather different.

## 1. Introduction

Teaching has been modeled and investigated in various ways within algorithmic learning theory. Already in Angluin's query model [1, 2] the oracles have some characteristics of teachers. However, they remain completely passive. In order to study teachers in a more active role, several models have been developed, each of which follows one of two basically different approaches.

In the first approach, the goal is to find a teacher *and* a learner such that a given learning task can be carried out by them. For the inductive inference framework, Freivalds Kinber, and Wiehagen [7] and Jain, Lange, and Nessel [13] developed a model in which a rather implicit teacher provides the learning strategy with good examples. Jackson and Tomkins [12] as well as Goldman and Mathias [9, 14] defined models of teacher/learner pairs where teachers and learners are constructed explicitly. In all these models, some kind of adversary disturbing the teaching process is necessary to avoid collusion between the teacher and the learner. Angluin and Kriķis' [3, 4] model prevents collusion by giving incompatible hypothesis spaces to teacher and learner. This makes simple encoding of the target impossible.

In the second approach, a teacher has to be found that teaches *all* learners. This prevents collusion, since teaching happens the same way for all learners and cannot be tailored to a specific one. Goldman, Rivest, and Shapire [10] and Goldman and Kearns [8] substitute the adversarial teacher in the online learning model by a helpful one selecting good examples. They investigate how many mistakes a consistent learner can make in the worst case. In Shinohara and Miyano's [16] model the teacher produces a set of examples for the target concept such that it is the only consistent one in the concept class. The size of this set is the same as the worst case number of mistakes in the online model. This number is termed the *teaching dimension* of the target. Because of this similarity we will from now on refer to both models as the *teaching dimension (TD-)model*.

One difficulty of teaching in the TD-model results from the teacher not knowing anything about the learners besides them being consistent. In reality a teacher can benefit a lot from knowing the learners' behavior or their current hypotheses. It is therefore natural to ask how teaching can be improved if the teacher may observe the learners' hypotheses after each example.

After translating this question into the TD-model, one sees that there is no gain in sample size at all. The current hypothesis of a consistent learner reveals nothing about its following hypothesis. Even if the teacher knew the hypothesis and provided a special example in response, he can only be sure that the learner's next hypothesis will be consistent. But this was already known to the teacher.

In this paper we extend the TD-model by a neighborhood relation over all hypotheses and by the requirement that all learners may only switch to a hypothesis in the neighborhood of their current one. We then compare basically two variants: In the first, the teacher receives the learner's hypothesis after every example taught. In the second, the teacher has no feedback available. It turns out that in the extended model the existence of feedback can really make a difference. Some concept classes can be taught much faster with feedback than without and some cannot be taught unless feedback is available to the teacher.

As a side effect the model can be used to study the teachability of infinite classes with potentially infinite concepts, e.g., languages. In the class containing all finite languages, for example, all concepts have an infinite teaching dimension and are thus

unteachable in the TD-model. With appropriate neighborhood relations this class can be taught, as we shall show in Section 3.

## 2. Preliminaries

A *concept* $c$ is a subset of an *instance space* $X$ and a *concept class* is a set of concepts over $X$. We consider two instance spaces: $\{0,1\}^n$ for Boolean functions and $\Sigma^*$ for languages over a finite and non-empty alphabet $\Sigma$. By $\mathcal{X} = X \times \{0,1\}$ we denote the set of *examples* over $X$. An example $(x,b)$ is either *positive*, if $b = 1$, or *negative*, if $b = 0$. A concept $c$ is *consistent* with $(x,b)$ iff $x \in c \Leftrightarrow b = 1$.

Let $R$ be a set of strings. $R$ *represents* $\mathcal{C}$ iff there is a function $\gamma\colon R \times X \to \{0,1\}$ such that $\mathcal{C} = \{\mathcal{C}_r \mid r \in R\}$, where $\mathcal{C}_r = \{x \mid \gamma(r,x) = 1\}$. The length of $r$ is denoted by $|r|$ and $\text{size}(c) := \min\{|r| \mid \mathcal{C}_r = c\}$ for every $c \in \mathcal{C}$. For any set $S$, we denote by $\text{card}(S)$ its cardinality and by $S^*$ the set of all finite tuples over $S$. We use the symbols $\circ$ for concatenation of tuples and $\triangle$ for the symmetric difference of two sets. Let $c$ be a concept and let $\vec{x} \in \mathcal{X}^*$ be a list of examples, then $err(\vec{x}, c)$ is the set of all examples in $\vec{x}$ that are inconsistent with $c$.

A *teaching set* for a concept $c$ with respect to $\mathcal{C}$ is a set $S$ of examples such that $c$ is the only concept in $\mathcal{C}$ consistent with $S$. The *teaching dimension* $TD(c)$ is the size of the smallest teaching set for $c$, the teaching dimension of $\mathcal{C}$ is $TD(\mathcal{C}) = \max\{TD(c) \mid c \in \mathcal{C}\}$.

For studying feedback, the learners in our model have to evolve over time. We adopt the online learning model and divide the teaching process into rounds. In each round the teacher provides an example to the learner who then computes a hypothesis from $R$. At the end of the round the teacher observes this hypothesis.

Thus, we describe a teacher by a function $T\colon R \times R^* \to \mathcal{X}$ receiving a concept's representation and a sequence of previously observed hypotheses as input and outputting an example.

A learner can be described by a function $L\colon \mathcal{X}^* \to R$ receiving a sequence of examples as input and outputting a hypothesis. Let $\nu \subseteq R \times R$ be a relation over $R$. Then $L$ is called *restricted* to $\nu$ iff $\forall \vec{x} \in \mathcal{X}^*\ \forall z \in \mathcal{X}\ [(L(\vec{x}), L(\vec{x} \circ z)) \in \nu]$, that is $\nu$ defines the admissible mind changes of $L$. Now, $(R,\nu)$ is a directed graph and we define the neighborhood of $r \in R$ as $Nb(r) := \{s \in R \mid (r,s) \in \nu\} \cup \{r\}$ and denote by $dist(r,s)$ the length of a shortest path from $r$ to $s$.

In the TD-model, the learner is required to always output a consistent hypothesis. Since in the restricted model all admissible hypotheses might be inconsistent, we have to modify this demand. We require that $L$ chooses only among the admissible hypotheses with least error with respect to the known examples. Moreover, we require a form of *conservativeness*: $L$ may only change its hypothesis if the new one has a smaller error. This ensures that $L$ will not change its mind after reaching a correct hypothesis. On the other hand, we also *require* $L$ to search for a better hypothesis if

it receives an inconsistent example. Otherwise, $L$ could stay at the initial hypothesis forever and teaching were impossible.

DEFINITION 1. *Let $R$ be a representation language for a concept class $\mathcal{C}$ and let $\nu \subseteq R \times R$ be a relation over $R$ and $h_0 \in R$ a starting hypothesis. A* $\nu$-learner *is a function $L\colon \mathcal{X}^* \to R$ such that $L(\emptyset) = h_0$ and for all $\vec{x} \in \mathcal{X}^*$ and for all $z \in \mathcal{X}$:*

(1) $(L(\vec{x}), L(\vec{x} \circ z)) \in \nu$,

(2) *if $L(\vec{x}) \neq L(\vec{x} \circ z)$ then $z$ is inconsistent with $\mathcal{C}_{L(\vec{x})}$,*

(3) *if $z$ is inconsistent with $\mathcal{C}_{L(\vec{x})}$ then*
$L(\vec{x} \circ z) \in \arg\min_{s \in Nb(L(\vec{x}))} \operatorname{card}(err(\vec{x} \circ z, \mathcal{C}_s))$.

We briefly remark that one can think of many plausible variants of the above definition. For instance, the learner could be allowed to change its mind on a consistent example if its hypothesis is inconsistent with an example received earlier. In this paper, however, all learners follow Definition 1.

The teaching process for a concept $c = \mathcal{C}_r$ is fully described by a teacher $T$ and a learner $L$ together with an initial hypothesis $h_0$. Such a process will result in a series $(h_i)_{i \in \mathbb{N}}$ of hypotheses and a series $(z_i)_{i \in \mathbb{N}}$ of examples: $h_{i+1} = L(z_0, \ldots, z_i)$ and $z_i = T(r, (h_0, \ldots, h_i))$.

DEFINITION 2. *Let $\mathcal{C}$ be a concept class with representation $R$ and let $\nu \subseteq R \times R$. We call* $\mathcal{C}$ teachable to $\nu$-learners in the limit with feedback *iff there is a teacher $T$ such that for all representations $r \in R$ and all $\nu$-learners $L$ the series $(h_i)_{i \in \mathbb{N}}$ of hypotheses converges to an $h$ with $\mathcal{C}_h = \mathcal{C}_r$.*

*The* teaching time of $T$ on $r$ *is the maximum $i$ such that there is a $\nu$-learner $L$ that reaches a representation of $\mathcal{C}_r$ at round $i$ for the first time.*

Note that an infinite teaching time does not imply unteachability of a concept.

For studying the influence of feedback, we also have to define teaching without feedback. In this situation the teacher is modeled as a function $T\colon R \times \mathbb{N} \to \mathcal{X}$, where the second argument specifies the round. The series of hypotheses is then given by $h_{i+1} = L(T(r, 0), \ldots, T(r, i))$. With this notation the definition of teaching *in the limit without feedback* is literally the same as Definition 2.

In the situation with feedback the teacher can stop teaching as soon as the learner has reached the goal. If there is no feedback, the teacher may or may not know when to stop. A teacher stopping after finitely many examples and still ensuring the learning success is said to teach *finitely without feedback*. More formally we consider $T\colon R \times \mathbb{N} \to \mathcal{X} \cup \{\perp\}$ where $\perp$ means "teaching has stopped."

With feedback we do not need to distinguish teaching finitely from teaching in the limit and we shall call this kind of teaching simply *teaching with feedback*.

DEFINITION 3. *Let $\mathcal{C}$ be a concept class with representation $R$ and let $\nu \subseteq R \times R$. We call $\mathcal{C}$* finitely teachable to $\nu$-learners without feedback *iff there is a teacher $T$ such that for all representations $r \in R$ and all $\nu$-learners $L$ the hypothesis $h_j$ with $j = \min\{i \mid T(r,i) = \perp\}$ satisfies $\mathcal{C}_{h_j} = \mathcal{C}_r$.*

If we set $\nu = R \times R$ in Definition 3 then we have essentially the teacher-directed learning model from Goldman, Rivest and Schapire [10] which has no restriction on hypothesis changes. The following theorem justifies the use of *arbitrary* $\nu$'s for studying the impact of feedback on the teaching process.

THEOREM 1. *Let $\mathcal{C}$ be a concept class with representations $R$ and let $\nu = R \times R$. Then the following statements are equivalent:*

(1) *$\mathcal{C}$ is finitely teachable to $\nu$-learners without feedback,*

(2) *$\mathcal{C}$ is teachable in the limit to $\nu$-learners without feedback,*

(3) *$\mathcal{C}$ is teachable to $\nu$-learners with feedback.*

*Furthermore in all three cases the same teacher can be used to obtain minimum teaching time which for all $c \in \mathcal{C}$ equals $TD(c)$ with respect to $\mathcal{C}$.*

*Proof.* The implication 1. $\Rightarrow$ 2. $\Rightarrow$ 3. is clear from the definitions.

It remains to show 3. $\Rightarrow$ 1. Let $\mathcal{C}$ be teachable to $\nu$-learners with feedback for $\nu = R \times R$ and let $c \in \mathcal{C}$. We first prove that for all $c \in \mathcal{C}$, $TD(c) < \infty$. Suppose there is a $c^* \in \mathcal{C}$ with $TD(c^*) = \infty$. Then there is a $\nu$-learner $L$ always assuming a consistent hypothesis not representing $c^*$. This is possible because there is no finite set of examples specifying $c^*$ and because every hypothesis can be reached from every other. Obviously $L$ cannot be taught $c^*$ in the limit, not even with feedback; a contradiction.

Now, since all teaching dimensions are finite, we can define a teacher $T$ that outputs for each $c \in \mathcal{C}$ a teaching sequence and stops. Teacher $T$ does not need any feedback. Obviously, teacher $T$ teaches $\mathcal{C}$ to all $\nu$-learners finitely and without feedback, because at the end of teaching there is only one consistent hypothesis left which is certainly reachable.

In order to see that teacher $T$ has optimal teaching time for $c \in \mathcal{C}$ with respect to all three teaching models, we consider a $\nu$-learner $L$ that always outputs a consistent hypothesis not representing concept $c$, unless $c$ is the only consistent concept in which case $L$ outputs a representation for $c$. It is easy to see that concept $c$ cannot be taught to learner $L$ with less than $TD(c)$ examples, no matter whether or not feedback is allowed. ∎

Note that Theorem 1 relies on the fact that neither the teacher nor the learners nor the function $\gamma$ are required to be recursive. Adding these requirements leads to new questions which we shall investigate in a forthcoming paper.

## 3. Comparison of the Teaching Models

In this section we will apply the new framework to the class $\mathcal{C}_{fin}$ of all finite languages over an alphabet $\Sigma$. This class cannot be taught in the TD-model. By using different $\nu$-restrictions we demonstrate various effects.

We fix any total ordering on all strings over $\Sigma$ and use as representation language $R$ the set of all comma-separated ordered lists of strings over $\Sigma$, i.e., $r = w_1, \ldots, w_m \in R$ represents the language $\{w_1, \ldots, w_m\}$. To simplify proofs later, we set $|r| := \sum_{i=1}^{m} |w_i|$, i.e., without counting the commas. We define the allowed transitions from $r$ to $s$ by $((r, s) \in \nu)$ iff $\mathrm{card}(\mathcal{C}_r \triangle \mathcal{C}_s) \leq 1$. The initial hypothesis is the empty string $\varepsilon$ representing the empty concept. Now we have:

FACT 2. *$\mathcal{C}_{fin}$ is finitely teachable to $\nu$-learners without feedback.*

*Proof.* For a finite language with representation $w_1, \ldots, w_m$ a teacher simply provides all positive examples $(w_1, 1), \ldots, (w_m, 1)$. In every round, the learners may either add or remove a string from their hypothesis. Starting at the empty language, there is only one possibility to stay consistent with the examples, namely by adding them to the hypothesis. Therefore, after $m$ rounds all $\nu$-learners have arrived at the target hypothesis. ∎

Feedback can be utilized when the restriction is modified. We define $(r, s) \in \nu'$ iff $\mathcal{C}_s = \mathcal{C}_r \cup \{w_1, w_2\}$ for some $w_1, w_2 \in \Sigma^*$ or $\mathcal{C}_s = \mathcal{C}_r \setminus \{w_1\}$. In both cases, we require that the size of the hypotheses may at most double each round: $|s| \leq 2|r|$. In the special case $r = \varepsilon$ we allow every singleton concept as neighbor: $(\varepsilon, s) \in \nu$ for all $s$ with $\mathrm{card}(\mathcal{C}_s) = 1$. For $\nu'$-learners there is a big difference in teaching time between teaching with and without feedback.

FACT 3. *$\mathcal{C}_{fin}$ is teachable to $\nu'$-learners with feedback such that for all $c \in \mathcal{C}$ the number of examples is $O(\mathrm{card}(c)) \leq O(\mathrm{size}(c))$.*

*Proof.* All $\nu'$-learners may either add two strings to their hypothesis or remove one. As a consequence, whenever a $\nu'$-learner receives a positive example, he can add it to the hypothesis and "invent" another string and add it to the hypothesis as well. Due to the size restriction there are always only finitely many strings that can be invented.

Let $c^* = \{w_1, \ldots, w_m\}$ be a target concept. A teacher with feedback first teaches all strings $w_i$ as positive examples. After $w_m$ the hypothesis of each learner contains $c^*$ plus at most $m$ invented strings $u_1, \ldots, u_\ell$. From the feedback, the teacher gets to know these strings and can teach them as negative examples. Since at most one string can be removed per round, the learners have to remove the negative example they are taught and thus arrive at the correct hypothesis after $\ell$ rounds. Alltogether teaching takes at most $2m = 2\mathrm{card}(c^*)$ rounds. ∎

FACT 4. *$\mathcal{C}_{fin}$ is finitely teachable to $\nu'$-learners without feedback. Every such teacher needs $\Omega(2^{\mathrm{size}(c)})$ examples for some $c \in \mathcal{C}$ and there is no upper bound for the number of examples that depends only on* $\mathrm{card}(c)$.

*Proof.* A suitable teacher is defined as follows. Let $c \in \mathcal{C}_{fin}$. First of all, the teacher gives all strings of length at most 2size($c$) that are not in $c$ as negative examples. Afterwards, all strings in $c$, starting with a longest one, are taught as positive examples. The initial hypothesis is consistent with all negative examples, hence no hypothesis change happens. During the positive examples, the learners cannot include any strings outside of $c$ into their hypotheses, since all these strings either have been ruled out by the negative examples or are too long to be included. Also, since a longest string is taught first, the hypothesis growth limitation cannot be violated by positive examples included later. Hence, all $\nu'$-learners 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 without feedback to all $\nu'$-learners. Let $\mathtt{a}$ and $\mathtt{b}$ be symbols from the alphabet and $c^* = \{\mathtt{a}^m, \mathtt{b}\}$ a concept of size $m+1$ for an arbitrary $m > 2$. Let $z_0, \ldots, z_M$ be all examples taught by $T$ on concept $c^*$. Let $L$ be a $\nu'$-learner.

Clearly both strings, $\mathtt{a}^m$ and $\mathtt{b}$, must occur as positive examples, otherwise the $\nu'$-learner $L_0$ that never "invents" a string could not be taught. Moreover, $\mathtt{a}^m$ 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 by changed to $\mathtt{a}^m, \mathtt{b}$ later, thus $L_0$ cannot learn $c^*$. Let $z_{j_1} = (\mathtt{a}^m, 1)$ be the the first occurrence of $\mathtt{a}^m$ and let $z_{j_2} = (\mathtt{b}, 1)$ be the first occurrence of $\mathtt{b}$.

It suffices to show that $z_1, \ldots, z_M$ contains all strings of length at most $m-1$. This implies $M \geq 2^m - 1 = \Omega(2^{\text{size}(c^*)})$. Assume there were a string $\hat{w} \notin c^*$ with $|\hat{w}| \leq m-1$ which is not taught. We give a $\nu'$-learner $L$ that does not arrive at $c^*$ during teaching. On $z_{j_1}$, $L$ switches to hypothesis $h_{j_1+1} = \mathtt{a}^m$ and does not change it until $z_{j_2}$ arrives. Then $L$ chooses the hypothesis $h_{j_2+1} = \mathtt{a}^m, \mathtt{b}, \hat{w}$ which is incorrect, but consistent with the examples so far. The length restriction is obeyed, since size$(\mathtt{a}^m) = m = |\mathtt{b}| + |\hat{w}|$. 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 $m$ can be choosen arbitrarily large, there is no bound on the number of examples needed that depends on card($c^*$) only. ∎

If we remove the size restriction from $\nu'$ we yield $\nu''$.

FACT 5. *$\mathcal{C}_{fin}$ is not finitely teachable to $\nu''$-learners without feedback, but it is finitely teachable with feedback as well as in the limit without feedback.*

*Proof.* Suppose there is a teacher that finitely teaches $\mathcal{C}_{fin}$ to $\nu''$-learners without feedback. Let $c = \{w_1, w_2\} \in \mathcal{C}_{fin}$. Then a learner that, when the second positive example arrives, "invents" a word not occurring in the examples does not arrive at a correct hypothesis, a contradiction.

Next, we describe a teacher $T$ which teaches $\mathcal{C}_{fin}$ finitely with feedback. On $c \in \mathcal{C}_{fin}$, $T$ first gives all positive examples. This may lead to at most card($c$) superfluous strings in the hypothesis of a $\nu''$-learner. $T$ observes these strings and gives them as negative

examples, thus forcing all learners to remove the excessive strings and to reach the correct hypothesis.

A teacher for teaching $\mathcal{C}_{fin}$ in the limit without feedback, first teaches all positive examples. Again, a $\nu''$-learner's hypothesis may contain finitely many excessive strings. By teaching all strings outside the target concept, the superfluous strings can be removed in the limit. ∎

Finally we define $\nu'''$. It differs from $\nu''$ in that a string may only be removed from the hypothesis if neither its predecessor nor its successor (wrt. the fixed ordering on $\Sigma^*$) is contained in the hypothesis.

FACT 6. *$\mathcal{C}_{fin}$ is not teachable to $\nu'''$-learners in the limit without feedback, but it is finitely teachable with feedback.*

*Proof.* Suppose there is a teacher $T$ which teaches $\mathcal{C}_{fin}$ to $\nu'''$-learners in the limit without feedback. Let $c^* = \{w_1, w_2, w_3\} \in \mathcal{C}_{fin}$ and let $(z_i)_{i\in\mathbb{N}}$ be the sequence of examples taught by $T$ on $c^*$. All three strings must occur in the example sequence, otherwise the learner that does not "make up" strings could not be taught. Let $z_{j_i} = (w_i, 1)$ be first occurrence of $w_i$ for $i = 1, 2, 3$. Without loss of generality, we assume $j_1 < j_2 < j_3$.

We now construct a $\nu'''$-learner $L$ which fails on the above example sequence. After $z_{j_1}$, $L$'s hypothesis is $w_1$. When taught $z_{j_2}$, $L$ adds $w_2$ to the hypothesis, as well as a string $u_1 \notin c^*$ such that (1) neither $u_1$ nor its successor $u_2$ occurs in $z_1, \ldots, z_{j_3}$, and (2) $u_2 \notin c^*$. When taught $z_{j_3}$, $L$ 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 $L$'s hypothesis contains the strings $u_1$ and $u_2$ neither of which can be deleted any more. Thus, $L$ cannot end up with a correct hypothesis (because of the definition of $\nu'''$), a contradiction.

Teaching $\mathcal{C}_{fin}$ to $\nu'''$-learners finitely 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 string in $c^*$. Then all positive examples are taught and as soon as the teacher discovers that a learner has introduced a wrong string $u$ into the hypothesis, the negative example $(u, 0)$ is given. The string $u$ cannot be predecessor or successor of any other string in the hypothesis and is thus deleted from the hypothesis. After at most $(2+1+1)\cdot\mathrm{card}(c) = O(\mathrm{size}(c))$ examples all $\nu'''$-learners have reached the target. ∎

If we denote by $TFIN$, $TFB$, $TLIM$ the set of all $(\mathcal{C}, R, \nu, h_0)$ such that $\mathcal{C}$ is finitely teachable without feedback, with feedback or in the limit, respectively, we have just proved the following theorem.

THEOREM 7. *$TFIN \subset TLIM \subset TFB$.*

The teaching times in our model can hardly be compared to the teaching dimension, since the latter depends only on $\mathcal{C}$, whereas different choices of $\nu$ can lead to different teaching times for the same $\mathcal{C}$.

## 4. Finding Teachers

The problem of finding an optimal teacher (with or without feedback) for $\nu$-learners is NP-hard, since it is a generalization of finding an optimal teaching set, namely if $\nu = R \times R$ (see [16, 8, 5]).

Concept classes over finite instance spaces can always be taught in the TD-model. Given $\nu$-learners, however, the first question is whether teaching is possible at all. We shall show that this is difficult to decide in general.

The next theorem assumes that $\mathcal{C}$ and $\nu$ over an instance space $X$ and representation language $R$ are represented as a 0-1-valued matrix with $\mathrm{card}(R)$ rows and $\mathrm{card}(X) + \mathrm{card}(R)$ columns. Each row describes the represented concept in the first $\mathrm{card}(X)$ bits, and its neighborhood in the last $\mathrm{card}(R)$ bits (cf. Fig 1).

| | $x_1$ | $\bar{x}_1$ | $x_2$ | $\bar{x}_2$ | $x_3$ | $\bar{x}_3$ | $x_4$ | $\bar{x}_4$ | $w$ | $y_1$ | $y'_1$ | $y_2$ | $y'_2$ | $y_3$ | $y'_3$ | $y_4$ | $y'_4$ | $r_0$ | $r_1$ | $r_2$ | $r_3$ | $s_0$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s^*$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $r_0$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| $r_1$ | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $r_2$ | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $r_3$ | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $s_0$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| $s_1$ | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| $s_2$ | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| $s_3$ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| $s_4$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| $s^*$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Figure 1: Example for the reduction in Theorem 8 for the formula $F = (v_1 \vee \bar{v}_2 \vee v_3) \wedge (v_2 \vee v_4 \vee v_1) \wedge (\bar{v}_1 \vee v_3 \vee \bar{v}_4)$. The left part of the matrix defines $\mathcal{C}$, the right one $\nu$.

THEOREM 8. *For all notions of teaching, the following problem is NP-hard:*

Instance: $\mathcal{C}, R, \nu$, *and a concept* $c^*$ *as 0-1-vector of length* $\mathrm{card}(X)$.

Question: *Can* $c^*$ *be taught to* $\nu$*-learners?*

*Proof.* (*Theorem* 8). The proof is by reduction from `3-SAT`. Let $F = K_1 \wedge \cdots \wedge K_m$ be a formula in 3-CNF with clauses $K_1, \ldots, K_m$ and variables $v_1, \ldots, v_n$. Define $X_F, \mathcal{C}_F, R_F$ and $\nu_F$ as follows. $X_F$ contains instances $x_1, \bar{x}_1, \ldots, x_n, \bar{x}_n$ and $y_1, y'_1, \ldots, y_n, y'_n$ and an instance $w$, hence $\mathrm{card}(X_F) = 4n + 1$. $R_F$ contains the representations $r_0, r_1, \ldots, r_m, s_0, s_1, \ldots, s_n, s^*$. The initial hypothesis $r_0$ represents $\{x_1, \bar{x}_1, \ldots, x_n, \bar{x}_n\}$, the target concept $c^* := \{x_1, \bar{x}_1, \ldots, x_n, \bar{x}_n, w\}$ will be represented by $s^*$. For each clause $K_i$ we use $r_i$ to represent a concept that consists of all instances except of $x_j$ for all $v_j$ in $K_i$ and except of $\bar{x}_j$ for all $\bar{v}_j$ in $K_i$. Finally $s_0$ represents $X_F$ and $s_i$ represents the concept $\{x_{i+1}, \bar{x}_{i+1} \ldots x_n, \bar{x}_n, y_{i+1}, y'_{i+1}, \ldots, y_n, y'_n, w\}$ for $i = 1, \ldots, n$ (see Fig. 1).

The relation $\nu_F$ contains $(r_0, r_i)$ for all $i = 1, \ldots, n$ and $(r_0, s_0)$ and $(s_i, s_{i+1})$ for $i = 0, \ldots, n-1$, as well as $(s_n, s^*)$. The only path from the initial to the target

hypothesis is $r_0, s_0, s_1, \ldots, s_n, s^*$. If one of the $r_i$'s is reached, teaching has failed as these representations are dead ends.

$\mathcal{C}_F$ and $\nu_F$ can easily be computed and encoded as a $(4n + 1 + m + 1 + n + 2) \cdot (m + 1 + n + 2) = O((n + m)^2)$ size matrix. Therefore the reduction is polynomial.

Let $F$ be satisfied by an assignment $\beta\colon \{v_1, \ldots, v_n\} \to \{0, 1\}$. We have to show that $c^*$ is teachable in the environment defined above. A successful example sequence consists of (1) for all $i = 1, \ldots, n$ the positive examples $x_i$, if $\beta(v_i) = 1$, or $\bar{x}_i$, if $\beta(v_i) = 0$; (2) the positive example $w$; (3) the sequence $y_1, y_1', \ldots, y_n, y_n'$ of negative examples; (4) any positive example not yet presented. All examples before $w$ are consistent with $r_0$, hence no mind change can take place. A mind change is then triggered by teaching $(w, 1)$. By their definition all $r_i$'s are inconsistent with the examples taught at step (1), whereas $s_0$ certainly is consistent. Therefore all $\nu_F$-learners will hypothesize $s_0$ after step (2). Teaching $y_1$ and $y_1'$ causes two inconsistencies with $s_0$, but $s_1$ has only one error (either $x_1$ or $\bar{x}_1$, depending on $\beta$). It follows that all learners are forced to $s_1$. Similarly one can see that after teaching $y_2, y_2', \ldots, y_n, y_n'$ all learners have reached $s_n$. Now each missing positive example triggers a mind change to $s^*$. This shows that $c^*$ is teachable.

Let $F$ be a formula such that $c^* \in \mathcal{C}_F$ is teachable to all $\nu_F$-learners. Let $z_1, \ldots, z_\ell \in \mathcal{X}$ be a sequence of examples such that all $\nu_F$-learners starting at $r_0$ end up in $s^*$. We have to show that $F$ can be satisfied.

The idea of the proof is as follows. First we show that after a certain example all learners must have reached $s_0$. At this point, for all $i = 1, \ldots, n$ *not* both $x_i$ and $\bar{x}_i$ have been taught. To prove this we show that if for some $i$ both $x_i$ and $\bar{x}_i$ have been taught, then it is impossible to force all learners to reach $s^*$. Finally we define a satisfying assignment $\beta$ depending, for each $i$, on whether $x_i$ or $\bar{x}_i$ occurs in the sample.

As long as the teacher teaches examples different from $w$, $r_0$ is consistent and no mind change happens. Therefore, for some $k$, $z_k = (w, 1)$. At this point a mind change must happen. If there were no mind change, all neighbors of $r_0$, i.e., $r_1, \ldots, r_m, s_0$, had more errors than $r_0$. This cannot be repaired, thus all learners would remain in $r_0$ forever.

Since the example sequence eventually leads to $s^*$, the hypothesis after example $z_k$ must be $s_0$. Furthermore none of the $y$-examples can have been taught: Otherwise all neighbors of $r_0$ had at least one error (the $y$-example) and $r_0$ had exactly one error (the example $w$), hence no change from $r_0$ had occured.

Since the only way to $s^*$ is via $s_1, \ldots, s_n$, the teacher must now provide examples that make all learners switch to $s_1$. The point now is that if $x_1$ *and* $\bar{x}_1$ occur in the sample, $s_1$ has two errors, but if only one of these examples occurs, $s_1$ has only one error. If the hypothesis is to be switched to $s_1$, the teacher must provide examples such that $s_0$ has at least two errors (otherwise there were no better hypothesis in the neighborhood). Since $s_0$ and $s_1$ are identical with respect to all instances except

$x_1, \bar{x}_1, y_1, y_1'$, such errors can only be generated by teaching $y_1$ as well as $y_1'$. But even then, the mind change can only be performed if $s_1$ has less than two errors. Thus, since $s_1$ is reached via the example sequence, it follows that not both $x_1$ and $\bar{x}_1$ have been taught.

In a similar way it can be shown that $s_{i+1}$ can only be reached from $s_i$ if not both $x_{i+1}$ and $\bar{x}_{i+1}$ appear in the sample. Note that teaching $x_i$ or $\bar{x}_i$ *after* the learners have reached $s_i$ is possible, but does not influence the following mind changes, because the concepts $s_{i+1}, \ldots, s_n$ are identical wrt $x_1, \bar{x}_1 \ldots, x_i, \bar{x}_i$.

Altogether it follows that when all learners changed to $s_0$ for all $i$ either $x_i$ or $\bar{x}_i$ had not been in the sample taught so far. Therefore the assignment $\beta$ is well-defined by $\beta(v_i) = 1$ iff $x_i$ appears among the examples $z_1, \ldots, z_k$.

It remains to show that $\beta$ satisfies $F$. This is clear from the definition of the $r_i$'s. If $\beta$ did not satisfy a clause $K_j$ then $r_j$ is consistent with whatever $x$-examples have been taught before $w$. Thus, $r_j$ is an equally good neighbor as $s_0$ and there will be a $\nu_F$-learner choosing $r_j$ instead of $s_0$. But this is a contradiction to the assumption that all such learners reach $s^*$. ∎

For infinite instance spaces or infinite classes (and infinite $\nu$) the following theorem applies.

THEOREM 9. *The following function is not computable:*

Input*: Algorithms computing total functions deciding $\mathcal{C}$ and $\nu$.*

Output*: 1, if $\mathcal{C}$ can be taught to $\nu$-learners; 0 otherwise.*

*Proof.* Within this proof, we use $\mathbb{N}$ as instance space and as representation language. Let $\mathcal{C} = \{\{0, \ldots, i\} \mid i \geq 1\} \cup \{\mathbb{N}\}$ a concept class. A concept $\{0, \ldots, i\}$ is represented by $i$, and 0 represents the concept $\mathbb{N}$. Let $(\varphi_i)_{i \in \mathbb{N}}$ be an effective enumeration of all partial recursive functions. For all $j \in \mathbb{N}$ we define an effective enumeration $(\nu_j)_{j \in \mathbb{N}}$ by

$$\nu_j(r, s) = \begin{cases} 1, & \text{if } r + 1 = s \text{ or } (s = 0 \text{ and } \varphi_j(j) \text{ is defined after} \leq r \text{ steps}), \\ 0, & \text{otherwise.} \end{cases}$$

It suffices to show that $\mathcal{C}$ is teachable to $\nu_j$-learners iff $\varphi_j(j)$ is defined. Let $\mathcal{C}$ be teachable to $\nu_j$-learners. Then $\mathbb{N}$ can be taught, hence the representation 0 must be reachable in the graph $(\mathbb{N}, \nu_j)$. From the definition of $\nu_j$ it follows that $\varphi_j(j)$ is defined.

For the opposite direction, let $\varphi_j(j)$ be defined after $r$ steps. Then $\mathbb{N}$ can be taught by the example sequence $(2, 1), \ldots, (r, 1), (r + 2, 1)$, where the last example ensures that the only consistent neighbor of $r$ is 0. Concepts $\{0, \ldots, i\}$ can be taught by $(i + 1, 0), (2, 1), \ldots, (i, 1)$, where the first example prohibits a transition to the hypothesis 0. ∎

## 5. Teaching Without Feedback

A teacher $T$ without feedback knows all learners' initial hypotheses $h_0$, but can quickly lose track of them during teaching. On the other hand, $T$ can rule out neighbors $r$ of $h_0$ by giving examples consistent with $h_0$, but inconsistent with $r$. If in such a way $T$ can eliminate all but one neighbor $r'$, he effectively forces all learners to switch to $r'$. By continuing in this manner, $T$ always knows all learners' hypotheses even without feedback. If the enforced hypotheses approach the target, $T$ will be successful. Figure 2 describes this strategy more formally.

1 $r := h_0$;

2 **while** $\mathcal{C}_r \neq c^*$ **do**:

2.1 Find $s \in Nb(r)$, $S \subseteq \mathcal{X}$, and $z \in \mathcal{X}$ such that (1) $\mathcal{C}_r$ is consistent with $S$, but not with $z$, (2) $s$ is the only neighbor of $r$ consistent with $S \cup \{z\}$, and (3) $dist(s, r^*) < dist(r, r^*)$;

2.2 Teach $S$ in arbitrary order and then $z$;

2.3 $r := s$;

Figure 2: A simple general strategy for teaching without feedback by forcing all learners to make the same mind changes. The initial hypothesis is $h_0$, $r^*$ represents the target.

The feasibility of this strategy depends on Step 2.1. If teaching does not need to be finite, the condition in Step 2 does not need to be checked. Albeit simple, the strategy works surprisingly often for natural concept classes and $\nu$-restrictions. In the following we give some examples.

First, we consider the class of all monomials over $n$ variables. Let $R = \{0, 1, *\}^n$ and define $(r, s) \in \nu$ iff $r$ and $s$ differ only in one "bit." As initial hypothesis $h_0 = *^n$ is used.

FACT 10. *Monomials are finitely teachable without feedback. The teaching time for each concept equals its teaching dimension.*

*Proof.* Let $c^*$ be a concept represented by $r^*$. We use the "standard" minimum teaching set for monomials that can be constructed in time $O(n^2)$ (see [8, 16]). Let $k_1, \ldots, k_\ell$ be the positions of all constants in $r^*$. The teaching set consists of two positive examples $x_0^+, x_1^+$ which result from substituting all $*$'s with zeroes and ones, respectively. Furthermore it contains one negative example $x_i^-$ for each $k_i$ where the $k_i$-th bit is inverted and all $*$'s are replaced by zeroes. Let $T$ teach the sequence $\langle x_0^+, x_1^+, x_1^-, \ldots, x_\ell^- \rangle$.

$T$ follows the strategy of Fig. 2: After the first inconsistent example, $x_1^-$, all $\nu$-learners are forced to a consistent hypothesis in the neighborhood of $*^n$. The only such hypothesis is obtained from $*^n$ by setting the $k_1$-th "bit" to the correct value.

This reduces the distance from the target by one. Each of the remaining examples forces all learners to set one $*$-bit of their hypothesis to a constant. After $x_\ell^-$ all constants are set correctly and the target is reached. ∎

At first glance, the new and the TD-model show little difference with regard to monomials, since we can use teaching sets also for $\nu$-learners. However, not every teaching set could be used for teaching. Even the same teaching set might fail if the examples are given in the wrong order. For example, consider $r^* = 11**$ which has a teaching set with $x_0^+ = 1100$, $x_1^+ = 1111$, $x_1^- = 0100$, $x_2^- = 1000$. Teaching those examples in reverse order can lead to the following hypothesis sequence: $0***$, $00**$, $00**$, $00**$. The last hypothesis is not only incorrect, it is even impossible to reach $r^*$ from it (given the examples taught so far).

As another natural concept class, together with a representation, we consider the class of all Boolean functions of $n$ variables represented by *decision trees*. A decision tree is a binary tree whose internal nodes are labeled with a variable and whose leaves are labeled either as positive or as negative. An instance $x \in \{0,1\}^n$ traverses the tree beginning at the root and at each internal node choosing the left child if that node's variable is satisfied and the right child otherwise, until a leaf is reached. Thus each tree represents a concept $c \subseteq \{0,1\}^n$ containing all positively classified instances.

Each learner starts at the tree consisting of only one negative leaf. In each round one leaf may be substituted by an internal node that has two differently labeled leaves as children. This specifies a relation $\nu_{DT}$ over all decision trees.

Fact 11. *The class of Boolean functions represented as decision trees can be taught without feedback to $\nu_{DT}$-learners. The teaching time is linear in the size of the tree representation.*

*Proof.* Let $c^*$ be a Boolean function over $n$ variables $v_1, \ldots, v_n$ represented by a decision tree $D^*$. Without loss of generality we can assume that (1) for a node with two leaves as children, both leaves are differently labeled, (2) on each path from the root to a leaf each variable occurs at most once, and (3) for all internal nodes both subtrees are semantically different. From this follows that for all internal nodes $t$ containing a variable $v_k$ exist $x_-, x_+ \in \{0,1\}^n$ such that $x_+ \in c^*$, $x_- \notin c^*$; $x_-$ and $x_+$ reach $t$ and differ only in the $k$-th bit.

The basic idea of the teaching algorithm is to force all learners to build their hypothesis identical to $D^*$. Beginning with the root node, with each hypothesis change a leaf in the hypothesis is substituted by the "correct" node of $D^*$. Substitution of a leaf by a node $t$ can be caused by teaching $x_-$ and $x_+$ as defined above. More precisely:

1. $r :=$ tree consisting of a negative leaf only;
2. $G := \{$the only node of $r\}$;
3. **while** $r \neq D^*$:

3.1 Let $g \in G$ be a leaf of $r$ with label $b$;

3.2 Let $t$ be the root of the subtree of $D^*$ by which $g$ must be substituted in order to construct $D^*$ from $r$;

3.3 Let $v_k$ be the variable in $t$;

3.4 Find $x_-, x_+$ as described above;

3.5 Let $\hat{g}$ be the tree with root $t$ and two child leaves that correctly classifies $x_-$ and $x_+$;

3.6 $r := r$ after replacing $g$ by $\hat{g}$;

3.7 Teach the $x \in \{x_i, x_+\}$ with $c^*(x) = b$;

3.8 Teach the $x \in \{x_i, x_+\}$ with $c^*(x) \neq b$;

3.9 $G := (G \setminus \{g\}) \cup \{\ell |\ \ell$ is a leaf of $\hat{g}$ that is not in $D^*\}$;

The correctness is implied by the following invariants which hold at the beginning of each iteration of the while loop:

1. $G \neq \emptyset$;
2. $D^*$ can be built from $r$ by substituting all leaves in $G$ by suitable subtrees of $D^*$;
3. $r$ is a hypothesis consistent with all examples taught so far and is also the hypothesis of all $\nu$-learners.

The algorithm terminates since $r$ strictly grows in size with each iteration until it equals $D^*$. Note that $r$ cannot "go astray" because of invariant 2. Therefore, the number of iterations is two times the number of internal nodes of $D^*$.

We sketch the proof of the invariants which trivially hold before the first iteration. Assume the invariants hold at the beginning of an iteration. Since $G \neq \emptyset$, there is a leaf $g \in G$ of $r$ with a label $b$. From invariant 2 follows that there is a subtree of $D^*$ with root $t$ by which $g$ must be substituted in order to build $D^*$. Then the path from the root of $r$ to $t$ also is in $D^*$, hence there are instances $x_-, x_+$ reaching $t$ such that $x_- \notin c^*$, $x_+ \in c^*$, and only differing in the $k$-th bit. Both instances also reach $g$ in $r$ and are classified as $b$. Therefore, the example taught in step 3.7 is consistent with $r$, but the example from 3.8 is not.

Denote by $r'$ the tree resulting from replacing $g$ in $r$ by $\hat{g}$ (step 3.6). We show that $r'$ is the only $\nu$-neighbor of $r$ consistent with all examples including $x_-$ and $x_+$. Replacing any other node than $g$ does not change the way $x_-$ and $x_+$ are classified. Such a hypothesis would remain inconsistent with one of these examples. Replacing $g$ by a node with different variable than $v_k$ would also classify one of the instances $x_-, x_+$ wrong since they differ only wrt. variable $v_k$. Hence, all $\nu$-learners assume $r'$ as their next hypothesis, which proves invariant 3.

Since $t$ is a node from $D^*$ which has been put "in the right place" in $r$, it is still possible to complete $r$ with subtrees of $D^*$ which shows invariant 2.

$G = \emptyset$ can only happen, if there are no more leaves of $r$ to be replaced, that is if $r = D^*$. In this case the loop terminates and there is no next iteration, hence nothing to prove for invariant 1. ∎

The teaching dimension with respect to all Boolean functions is $2^n$ for all concepts. As we have seen, for $\nu$-learners based on decision trees, teaching can often be successful with much fewer examples.

One can think of three situations where the above strategy either fails or is inefficient due to lack of feedback: (1) it is impossible to enforce a certain mind change by ruling out all but one neighbor; (2) correcting a wrong hypothesis afterwards is cheaper than preventing all possible errors beforehand; (3) there are several equivalent, but syntactically different hypotheses in the neighborhood.

We have already seen examples of situations (1) in Fact 6, and of situation (2) in Facts 3 and 4. For completing the picture, in the following we construct an example for situation (3).

We consider *monotone 1-decision lists* $\langle (y_1, b_1), \ldots, (y_m, b_m), (*, 0)\rangle$ of variables $y_1, \ldots, y_m$ and bits $b_i \in \{0, 1\}$. An instance $x \in \{0, 1\}^n$ runs through the list starting at the node $(y_1, b_1)$ until it satisfies a variable, say $y_j$, in which case it is classified as $b_j$. The default node $(*, 0)$ classifies all instances as negative that do not satisfy any of the variables $y_1, \ldots, y_m$.

We use two kinds of learners obeying different neighborhood relations. Both start at a decision list consisting of only a positive default node $(*, 1)$ whose only neighbor is the list $\langle (*, 0)\rangle$ with a negative default node. All learners may insert nodes of the form $(y, 0)$ in any position of the list. However, restrictions apply with regard to nodes of the form $(y, 1)$. Learners of the first kind are allowed to substitute the *first* node of the hypothesis by an arbitrary positive node or to insert such a node at the *beginning* of the list. Learners of the second kind may only substitute the *last* node or insert at the *end* of the list. In both cases, the default node must not be substituted.

To distinguish the hypotheses of both kinds of learners we label the decision lists with either `B` or `E` specifying whether modifications are allowed at the beginning or at the end of the list, respectively. We have therefore two relations, $\nu_B$ and $\nu_E$, with exactly one common representation, the initital hypothesis $\langle (*, 1)\rangle$. If we join both relations at this point, we get a relation $\nu_{DL}$. Thus, a $\nu_{DL}$-learner will, after receiving the first negative example, switch to either $\langle (*, 0)\rangle_{\mathtt{B}}$ or $\langle (*, 0)\rangle_{\mathtt{E}}$ and then act like a $\nu_B$- or a $\nu_E$-learner.

Intuitively, examples suitable for $\nu_B$-learners can lead $\nu_E$-learners into a dead end hypothesis and vice versa. Hence, it is important for the teacher to know what type

of learner he teaches. This can be recognized by the B/E-extension of the hypotheses, which requires feedback.

FACT 12. *The class of monotone 1-decision lists can be taught to $\nu_{DL}$-learners with feedback using $m+1$ examples for a list of length $m$. It cannot be taught without feedback.*

*Proof.* We present an outline of the proof. We first describe a teacher $T$ using feedback. Let $c^*$ be a Boolean function over $n$ variables represented by a monotone 1-decision list $D^* = \langle (y_1, b_1), \ldots, (y_m, b_m), (*, 0)\rangle$. Without loss of generality we assume that $D^*$ is in reduced form, that is (1) each variable occurs at most once (either as positive or negative literal), (2) the default node is negative, (3) $b_m = 1$ (otherwise $(y_m, b_m)$ can be removed without changing the represented concept).

$T$ first teaches $(0^n, 0)$. This makes all $\nu_{DL}$-learners switch to $\langle (*, 0)\rangle_{\mathtt{B}}$ or $\langle (*, 0)\rangle_{\mathtt{E}}$. The behavior of $T$ depends on the type (B/E) of the learner.

*Type* B: For all $j = m, \ldots, 1$ the teacher gives examples $(x_j, b_j)$ such that $x_j$ satisfies $y_j$ and $y_k$, where $k > j$ is the minimum index with $b_k \neq b_j$, i.e., the next node to the right with a different bit (or the default node, if $j = m$).

It can be shown by induction over $j$ that these examples enforce the construction of the list beginning from the rightmost node. Intuitively an example $x_j$ proves that $y_j$ is a relevant variable and triggers a mind change by being inconsistent.

*Type* E: For all $j = 1, \ldots, m$ with $b_j = 1$ the teacher gives an example $(x_j, 1)$ where $x_j$ satisfies only $y_j$. On these examples all learners build a decision list containing all positive nodes of $D^*$ in the right order.

To force the inclusion of the negative nodes, $T$ presents a sequence of negative examples. For each negative node $(y_j, 0)$ (from left to right) $T$ teaches an example $(x_j, 0)$ such that $x_j$ satisfies $y_j$ and $y_k$, where $(y_k, 1)$ is next positive node to the right. This example causes the inclusion of the node $(y_j, 0)$ at the correct position in the hypothesis.

For both types the teaching time is $m + 1$.

Assume for a contradiction that there is a teacher $T$ not using feedback. Let $D^* = \langle (v_1, 1), (v_2, 0), (v_3, 1), (*, 0)\rangle$ be the target decision list. Note that there is no other monotone 1-decision list equivalent with $D^*$.

Since both neighbors of the initial hypothesis represent the same concept, $T$ cannot enforce just one of them. Hence, there exist $\nu_{DL}$-learners $L_B$ and $L_E$ going to $\langle (*, 0)\rangle_{\mathtt{B}}$ and $\langle (*, 0)\rangle_{\mathtt{E}}$, respectively, on the example sequence taught. Now, $L_E$ must continue with hypothesis $D'_{\mathtt{E}} := \langle (v_1, 1), (*, 0)\rangle_{\mathtt{E}}$ since there is no other way to reach $D^*$ for a type-E learner. However, $D'_{\mathtt{B}}$ is also a neighbor of $L_B$'s hypothesis, because inserting at the beginning is equivalent to inserting at the end of the list $\langle (*, 0)\rangle$. Thus, $L_B$ can reach $D'_{\mathtt{E}}$, but this is a dead end, as $(v_3, 1)$ has to be inserted *after* $(v_1, 1)$ which is impossible for $L_B$. ∎

## 6. Comparison with Learning

Such comparisons have been done in the mistake bound model between teacher-directed learning and self-directed learning. In many natural concept classes, the best learner can always learn with fewer mistakes than the best teacher needs to teach all consistent learners [11, 8, 10]. Rivest und Yin [15] use cryptographic assumptions to construct a concept class where a teacher needs less examples than the best learner, if both are restricted to polynomial time algorithms. Ben-David and Eiron [6] construct such classes without relying on cryptographic assumptions.

Teaching and learning can also be compared according to the sample complexity instead of the mistake bound. This amounts to a comparison of the teaching dimension $TD$ with the number $MEMB$ of membership queries necessary. Goldman and Kearns [8] observed that for all $\mathcal{C}$, $MEMB(\mathcal{C}) \geq TD(\mathcal{C})$, i.e., being taught is generally simpler than learning by oneself. This contrasts with the mistake bound model.

We will have a brief look at how the introduction of the $\nu$-relation influences the relationship between teaching and learning. To do so, we give the $\nu$-learners access to a membership oracle. Note that still all conditions of Definition 1 apply. For example, a $\nu$-learner must try to change his mind when the oracle's answer is inconsistent with the current hypothesis.

The next two facts demonstrate that in our model teachability and learnability can be rather different.

FACT 13. *There are a class $\mathcal{C}$ with representation language $R$ and a $\nu \subseteq R \times R$ such that $\mathcal{C}$ can be taught to all $\nu$-learners, but no $\nu$-learner can learn it.*

*Proof.* Let $X = \{x_1, x_2, x_3\}$, $c_0 = \emptyset$, $c_1 = \{x_1\}, c_2 = \{x_1, x_3\}$, $c_3 = \{x_2\}$, $c_4 = \{x_2, x_3\}$ and $R = \{r_0, r_1, r_2, r_3, r_4\}$ with $r_i$ representing $c_i$. Finally, $\nu$ contains $(r_0, r_i)$ for $i = 1, 2, 3, 4$.

The concept $c_1$ can be taught using the instances $x_3, x_1$; $c_2$ by $x_2, x_3$; $c_3$ by $x_3, x_2$; and $c_4$ by $x_1, x_3$. Thus $\mathcal{C}$ can be taught without feedback to $\nu$-learners.

Assume there is a $\nu$-learner $L$ with access to a membership oracle.

*Case* 1. *$L$ first queries $x_1$.* On answer "1", $L$ must change its hypothesis to either $r_1$ or $r_2$. If $L$ chooses $r_1$ than it cannot learn $c_2$ since there is no way back to $r_0$. Similar, if $r_2$ is chosen, $L$ cannot learn $c_1$ any more.

*Case* 2. *$L$ first queries $x_2$.* Analogous to Case 1 with concepts $c_3$ and $c_4$.

*Case* 3. *$L$ first queries $x_3$.* Analogous to Case 1 with concepts $c_2$ and $c_4$. ∎

FACT 14. *There are a class $\mathcal{C}$ with representation language $R$ and a $\nu \subseteq R \times R$ such that $\mathcal{C}$ can be learned by a $\nu$-learner, but cannot be taught to all $\nu$-learners.*

*Proof.* Let $X = \{x_1, x_2\}$, and let $c_0 = \emptyset$, $c_1 = \{x_1\}$, $c_2 = \{x_1, x_2\}$. Furthermore, let $R = \{r_0, r_1, r_1', r_2\}$ with $r_i$ representing $c_i$ and additionally $r_1'$ representing $c_1$. Let $\nu = \{(r_0, r_1), (r_0, r_1'), (r_1, r_2)\}$.

A $\nu$-learner works as follows. First query $x_1$. If the answer is "0", then the target must be $c_0$ and $L$ stops. If the answer is "1", change to hypothesis $r_1$ and query $x_2$. If the answer is "0", the target is $c_1$ and $L$ stops, otherwise $L$ switches to $r_2$ and stops. Hence, this $\nu$-learner learns $\mathcal{C}$.

Let $T$ be a teacher. We show that $T$ cannot teach $c_2$. Let $z$ be the first example taught. If $z = (x_1, 1)$ there is a $\nu$-learner going to $r_1'$ from where $r_2$ cannot be reached. Consequently, $T$ has to begin with $z = (x_2, 1)$ which causes no hypothesis change. As soon as $T$ teaches $(x_2, 1)$ there is a learner switching to $r_1'$. This learner will never reach $r_2$. Thus, $\mathcal{C}$ cannot be taught. ∎

## 7. Conclusion and Further Research

In our model several effects regarding feedback can be observed. Feedback can be useless, helpful, or even indispensable for teaching. In addition, natural infinite concept classes can be taught in this model and the relationship between teachability and learnability is more diverse than in the TD-model.

The variety of possible results stems mostly from the ability to define $\nu$ arbitrarily. We have also used rather artificial $\nu$'s in some places. It would therefore be interesting to put some natural restrictions on $\nu$, e.g., some relation between syntax (distance in the $(R, \nu)$-graph) and semantics (number of errors).

The strategy of Section 5, which often makes teaching without feedback possible, relies on the (somewhat unrealistic) feature of our models that all learners remember all examples (especially the consistent ones). It seems natural to study feedback for learners with some sort of memory limitation.

Further directions of research include adding computability restrictions to the teachers and/or learners, teaching with only positive examples, and other types of feedback, e.g., answering teacher's questions.

**Acknowledgments.** The second author heartily acknowledges support by the *21st Century COE Program C01.*

# References

[1] D. Angluin. Queries and concept learning. *Machine Learning*, 2(4):319–342, 1988.

[2] D. Angluin. Queries revisited. In *Algorithmic Learning Theory, 12th International Conference, ALT 2001, Proc.*, vol. 2225 of *Lecture Notes in Artificial Intelligence*, pages 12–31. Springer, 2001.

[3] D. Angluin and M. Krikis. Teachers, learners and black boxes. In *Proc. 10th Annual Conference on Computational Learning Theory*, pages 285–297, ACM Press, New York, NY, 1997.

[4] D. Angluin and M. Krikis. Learning from different teachers. *Machine Learning*, 51(2):137–163, 2003.

[5] M. Anthony, G. Brightwell, D. Cohen, and J. Shawe-Taylor. On exact specification by examples. In *Proc. 5th Annual ACM Workshop on Computational Learning Theory*, pages 311–318. ACM Press, New York, NY, 1992.

[6] 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.

[7] R. Freivalds, E. B. Kinber, and R. Wiehagen. Learning from good examples. In *Algorithmic Learning for Knowledge-Based Systems*, vol. 961 of *Lecture Notes in Artificial Intelligence*, pages 49–62. Springer, 1995.

[8] S. A. Goldman and M. J. Kearns. On the complexity of teaching. *J. of Comput. Syst. Sci.*, 50(1):20–31, 1995.

[9] S. A. Goldman and H. D. Mathias. Teaching a smarter learner. *J. of Comput. Syst. Sci.*, 52(2):255–267, 1996.

[10] 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.

[11] S. A. Goldman and R. H. Sloan. The power of self-directed learning. *Machine Learning*, 14(3):271–294, 1994.

[12] J. Jackson and A. Tomkins. A computational model of teaching. In *Proc. 5th Annual ACM Workshop on Computational Learning Theory*, pages 319–326. ACM Press, New York, NY, 1992.

[13] S. Jain, S. Lange, and J. Nessel. Learning of r.e. languages from good examples. In *Algorithmic Learning Theory, 8th International Workshop, ALT '97, Proc.*, vol. 1316 of *Lecture Notes in Artificial Intelligence*, pages 32–47. Springer, 1997.

[14] H. D. Mathias. A model of interactive teaching. *J. of Comput. Syst. Sci.*, 54(3):487–501, 1997.

[15] R. L. Rivest and Y. L. Yin. Being taught can be faster than asking questions. In *Proc. 8th Annual Conference on Computational Learning Theory*, pages 144–151. ACM Press, New York, NY, 1995.

[16] A. Shinohara and S. Miyano. Teachability in computational learning. *New Generation Computing*, 8(4):337–348, 1991.