Theoretical Computer Science ••• (••••) •••–•••

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

On the teaching complexity of linear sets

Ziyuan Gao [a,*], Hans Ulrich Simon [b], Sandra Zilles [a]

[a] Department of Computer Science, University of Regina, Regina, SK, S4S 0A2, Canada
[b] Horst Görtz Institute for IT Security and Faculty of Mathematics, Ruhr-Universität Bochum, D-44780 Bochum, Germany

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Teaching complexity
Teaching dimension
Recursive teaching dimension
Linear sets

ABSTRACT

Linear sets are the building blocks of semilinear sets, which are in turn closely connected to automata theory and formal languages. Prior work has investigated the learnability of linear sets and semilinear sets in three models – Valiant's *PAC-learning* model, Gold's *learning in the limit* model, and Angluin's *query learning* model. This paper considers *teacher–learner* models of learning families of linear sets, in which a benevolent teacher presents a set of labelled examples to the learner. First, we study the classical teaching model, in which a teacher must successfully teach any consistent learner. Second, we will apply a generalisation of the recently introduced recursive teaching model to several infinite classes of linear sets, and show that thus the maximum sample complexity of teaching these classes can be drastically reduced compared to classical teaching. To this end, a major focus of the paper will be on determining two relevant teaching parameters, the *teaching dimension* and *recursive teaching dimension*, for various families of linear sets.

1. Introduction

A mathematical notion that has found numerous applications in Computer Science over the last five decades is the notion of *semilinear set*. A semilinear set is a set of vectors (m-tuples) of nonnegative integers following a specific structure. Consider for example fixed m-tuples $c = (1, 0, 0 \ldots, 0)$, $p_1 = (2, 2, 0, \ldots, 0)$, $p_2 = (0, 0, 1, \ldots, 1)$, for $m \geq 3$, and the set L of vectors that can be expressed as a sum $c + n_1 \cdot p_1 + n_2 \cdot p_2$, where n_1, n_2 are nonnegative integers. L consists of all vectors of the shape $(k_1, \ldots, k_m)$, where k_1 is any odd positive integer, $k_2 = k_1 - 1$, and $k_3 = \ldots = k_m$ are any nonnegative integers independent of k_1 and k_2. The set L is then called a *linear set* with constant c and periods p_1 and p_2. A linear set may have any finite number of periods, not just two as in the example above, and the components of the constant and the periods can be any nonnegative integers. A semilinear set is now simply a union of finitely many linear sets.

Semilinear sets generalise the class of ultimately periodic sets of integers to higher dimensions and have applications in formal language theory and database theory. One of the earliest and most important results on the connection between semilinear sets and context-free languages is *Parikh's theorem* [17], which states that any context-free language is mapped to a semilinear set via a function known as the Parikh vector[1] of a string. Another interesting result, due to Ibarra [14], characterises semilinear sets in terms of reversal-bounded multicounter machines. Moving beyond abstract theory, semilinear sets have also recently been applied in the fields of DNA self-assembly [6] and membrane computing [15].

* Corresponding author.
E-mail addresses: gao257@cs.uregina.ca (Z. Gao), hans.simon@rub.de (H.U. Simon), zilles@cs.uregina.ca (S. Zilles).

[1] For a fixed order $(a_1, \ldots, a_m)$ over a given finite alphabet $\Sigma = \{a_1, \ldots, a_m\}$, the Parikh vector of a finite string s over Σ is the vector $(q_1, \ldots, q_m) \in \mathbb{N}_0^m$, where q_i is the number of occurrences of the symbol a_i in s.

https://doi.org/10.1016/j.tcs.2017.11.024
0304-3975/

Please cite this article in press as: Z. Gao et al., On the teaching complexity of linear sets, Theoret. Comput. Sci. (2018), https://doi.org/10.1016/j.tcs.2017.11.024

Considering the variety of areas in which semilinear sets play a role, it is not surprising that the machine learning community has shown interest in the learnability of such sets in various formal models. Learning of semilinear sets has been investigated in Valiant's PAC-learning model [1], Gold's learning in the limit model [22], and Angluin's query learning model [22]. Abe [1] showed that when the integers are encoded in unary, the class of semilinear sets of dimension 1 or 2 is polynomially PAC-learnable; on the other hand, the question as to whether classes of semilinear sets of higher dimensions are PAC-learnable is open. Takada [22] established that for any fixed dimension, the family of linear sets is learnable in the limit from positive examples but the family of semilinear sets is not learnable in the limit from only positive examples. In addition, Takada showed the existence of a learning procedure via restricted subset and restricted superset queries that identifies any semilinear set and halts; however, he proved at the same time that any such algorithm must necessarily be time consuming. He also proved that for any variable dimension, if for any unknown linear set L and any conjectured semilinear set U, queries as to whether or not $L \subseteq U$ can be made, then the class of linear sets is learnable in polynomial time of the minimum size of representations (of each linear set) and the dimension.

Due to the complexity of the class of semilinear sets in general, several of the positive learnability results in the literature concern the special case of linear sets; further it is common to deal with restrictions on the dimensionality m. The present work is primarily concerned with classes of *linear* sets of nonnegative *integers* and *pairs of integers*, i.e., the cases when $m = 1$ and when $m = 2$. We study the learnability of such classes in settings different from those studied by Abe or Takada, namely *teaching* settings, in which the information presented to the learner consists of labelled examples[2] that are selected by a benevolent source called the teacher. After the teacher presents a set S of labelled examples, the learner makes a conjecture as to the identity of the target concept. If the learner's conjecture is correct, then it is said to have successfully learnt the target concept.

All teaching and learning models discussed in the present work involve a *protocol* between the teacher and learner. A protocol maps a concept class $\mathcal{C}$ to a teacher–learner pair; for any $c \in \mathcal{C}$ and any set S of labelled instances over the universe, the teacher maps c to a set of labelled instances and the learner maps S to some $c' \in \mathcal{C}$. As an illustration, consider the *teaching set protocol* [11,12,19]. According to this protocol, for any given $c \in \mathcal{C}$, the teacher selects a smallest possible set S_c of labelled instances such that c is the only concept in $\mathcal{C}$ consistent with S_c while the learner maps any set S of labelled instances to some $c' \in \mathcal{C}$ that is consistent with S.

Zilles, Lange, Holte and Zinkevich [24] later designed a protocol – the *recursive teaching protocol* – that only exploits an inherent hierarchical structure of any concept class. The RTD of a concept class is the maximum sample complexity derived by applying the recursive teaching protocol to the class. Our interest in the RTD partly stems from the fact that this parameter has a number of surprising connections to other central notions in computational learning theory, namely to the VC-dimension and to sample compression schemes [4,5,20]. In addition, the RTD possesses several regularity properties and has been fruitfully applied to the analysis of pattern languages [9,16], which, when defined over unary alphabets, exhibit a correspondence to linear sets.

When studying the sample complexity of concept learning from teachers, a core question is to determine the number of labelled examples that are needed in the worst case to teach any concept in a given concept class; the present paper deals with exactly this question for various classes of linear sets. Two combinatorial parameters (and some variants) will be used to measure the sample complexity, namely the *teaching dimension* (TD) [11,19], which refers to the sample complexity of the teaching set protocol mentioned above, and the *recursive teaching dimension* (RTD) [24], which measures the sample complexity of the recursive teaching protocol. In some cases, teaching with positive examples, i.e., using only m-tuples contained in the target set, appears to be a natural strategy. In order to study how restrictive such a choice of examples is, we compare the TD and RTD parameters to the corresponding parameters referring to settings when only examples labelled with "+" are allowed. We denote these parameters by TD^+ and RTD^+, respectively.

Our main results can be summarised as follows.

- We determine the exact values or almost matching upper and lower bounds of the TD, TD^+, RTD and RTD^+ for various classes of linear sets (see Table 1 and Section 2.1 for the definitions of these classes). For many classes of linear sets considered in the present work, the RTD is significantly smaller than the TD (Table 1).
- We show that there are natural classes of linear sets that cannot be taught in the recursive setting using only positive examples while the RTD is finite; these examples illustrate how supplying negative information may sometimes be indispensable to successful teaching and learning.
- We demonstrate that the RTD^+ is closely related to the notions of *weak* and *strong spanning sets*. In particular, the RTD^+ of a concept class $\mathcal{C}$ is upper bounded by the size of the largest minimum strong spanning set over all $L \in \mathcal{C}$, and lower bounded by the size of the largest minimum weak spanning set over all $L \in \mathcal{C}$ (Lemma 24). For intersection-closed concept classes such as the class of linear sets, weak spanning sets are also strong spanning sets and vice versa (Lemma 25).
- Coinfinite linear subsets of $\mathbb{N}_0$ with constant 0 do not have finite teaching sets (Corollary 15).
- The minimum teaching set of any cofinite linear subset L of $\mathbb{N}_0$ with constant 0 may be characterised by means of a partial order defined with respect to the set of periods of L (Corollary 16).

[2] A labelled example for a linear set L over $\mathbb{N}_0^m$ is a pair $((a_1, \ldots, a_m), \ell) \in \mathbb{N}_0^m \times \{+, -\}$ where $\ell = +$ if $(a_1, \ldots, a_m) \in L$ and $\ell = -$ otherwise.

Please cite this article in press as: Z. Gao et al., On the teaching complexity of linear sets, Theoret. Comput. Sci. (2018), https://doi.org/10.1016/j.tcs.2017.11.024