Machine Learning manuscript No.
(will be inserted by the editor)

Learning figures with the Hausdorff metric by fractals — towards computable binary classification

Mahito Sugiyama · Eiju Hirowatari · Hideki Tsuiki · Akihiro Yamamoto

Received: date / Accepted: date

Abstract We present learning of *figures*, nonempty compact sets in Euclidean space, based on *Gold's learning model* aiming at a *computable* foundation for binary classification of multivariate data. Encoding real vectors with no numerical error requires *infinite* sequences, resulting in a gap between each real vector and its *discretized* representation used for the actual machine learning process. Our motivation is to provide an analysis of machine learning problems that explicitly tackles this aspect which has been glossed over in the literature on binary classification as well as in other machine learning tasks such as regression and clustering. In this paper, we amalgamate two processes: discretization and binary classification. Each learning target, the set of real vectors classified as positive, is treated as a figure. A learning machine receives discretized vectors as input data and outputs a sequence of discrete representations of the target figure in the form of *self-similar sets*, known as *fractals*. The generalization error of each output is measured by the *Hausdorff metric*. Using this learning framework, we reveal a hierarchy of learnable classes under various learning criteria in the track of traditional analysis based on Gold's learning model, and show a mathematical connection between machine learning and fractal geometry by measuring the complexity of learning using the *Hausdorff dimension* and the *VC dimension*. Moreover, we analyze computability aspects of learning of figures using the framework of Type-2 Theory of Effectivity (TTE).

Keywords Binary classification · Discretization · Self-similar set · Gold's learning model · Hausdorff metric · Type-2 Theory of Effectivity

Mahito Sugiyama
The Institute of Scientific and Industrial Research (ISIR), Osaka University, Mihogaoka 8-1, Ibaraki, Osaka, 567-0047, Japan
Tel: +81-6-6879-8540
Fax: +81-3-6879-8544
E-mail: mahito@ar.sanken.osaka-u.ac.jp

Eiju Hirowatari
Center for Fundamental Education, The University of Kitakyushu

Hideki Tsuiki
Graduate School of Human and Environmental Studies, Kyoto University

Akihiro Yamamoto
Graduate School of Informatics, Kyoto University

* Most of this work was accomplished when the first author was at Kyoto University.

1 Introduction

Discretization is a fundamental process in machine learning from analog data. For example, Fourier analysis is one of the most essential signal processing methods and its discrete version, *discrete Fourier analysis*, is used for learning or recognition on a computer from continuous signals. However, in the method, only the direction of the time axis is discretized, so each data point is not purely discretized. That is to say, continuous (electrical) waves are essentially treated as finite/infinite sequences of *real numbers*, hence each value is still continuous (analog). The gap between analog and digital data therefore remains.

This problem appears all over machine learning from observed multivariate data. The reason is that an infinite sequence is needed to encode a real vector exactly without any numerical error, since the cardinality of the set of real numbers, which is the same as that of infinite sequences, is much larger than that of the set of finite sequences. Thus to treat each data point on a computer, it has to be *discretized* and considered as an approximate value with some numerical error. However, to date, most machine learning algorithms ignore the gap between the original value and its discretized representation. This gap could result in some unexpected numerical errors[1]. Since now machine learning algorithms can be applied to massive datasets, it is urgent to give a theoretical foundation for learning, such as classification, regression, and clustering, from multivariate data, in a fully computational manner to guarantee the soundness of the results of learning.

In the field of computational learning theory, *Valiant's learning model* (also called *PAC, Probably Approximately Correct, learning model*), proposed by Valiant (1984), is used for theoretical analysis of machine learning algorithms. In this model, we can analyze the robustness of a learning algorithm in the face of noise or inaccurate data and the complexity of learning with respect to the rate of convergence or the size of the input using the concept of probability. Blumer et al (1989) and Ehrenfeucht et al (1989) provided the crucial conditions for learnability, that is, the lower and upper bounds for the sample size, using the *VC* (*Vapnik-Chervonenkis*) *dimension* (Vapnik and Chervonenkis 1971). These results can be applied to various concept representations that handle real-valued inputs and use real-valued parameters, for example, to analyze learning of neural networks (Baum and Haussler 1989). However, this learning model is not in line with discrete and computational analysis of machine learning. We cannot know which class of continuous objects is exactly learnable and what kind of data are needed to learn from a finite expression of discretized multivariate data. Although PAC learning from axis-parallel rectangles has already been investigated (Blumer et al 1989; Kearns and Vazirani 1994; Long and Tan 1998), which can be viewed as a variant of learning from multivariate data with numerical error, it is not applicable in the study. Our goal is to investigate computational learning, focusing on a common ground between "learning" and "computation" of real numbers based on the behavior of Turing machines, without any reference to probability distributions. For the purpose of the investigation, we need to distinguish abstract mathematical objects such as real numbers and their concrete representations, or codes, on a computer.

Instead, in this paper we use *Gold's learning model* (also called *identification in the limit*), which is originally designed for learning of recursive functions (Gold 1965) and languages (Gold 1967). In the model, a learning machine is assumed to be a procedure, *i.e.*, a Turing machine (Turing 1937) which never halts, that receives training data from time to time, and outputs representations (hypotheses) of the target from time to time. All data are usually assumed to be given at some point in the future. Starting from this learning model, learnability of classes of discrete objects, such as languages and recursive functions, has been analyzed in detail under various learning criteria (Jain et al 1999). However, analysis of learning for continuous objects, such as classification, regression, and clustering for multivariate data, with Gold's model is still under development, despite such settings being typical in modern machine learning. To the best of our knowledge, the only line of studies devoted to learning of real-valued functions was by Hirowatari and Arikawa (1997); Apsītis et al (1999); Hirowatari and Arikawa (2001); Hirowatari et al (2003, 2005, 2006), where they addressed the analysis of learnable classes of real-valued functions using computable representations of real numbers[2]. We therefore need a new theoretical and computational framework for modern machine learning based on Gold's learning model with discretization of numerical data.

[1] Müller (2001) and Schröder (2002a) give some interesting examples in the study of computation for real numbers.

[2] Sugiyama et al (2006, 2009) have also contributed to the area, but their work was only presented at closed workshops.

In this paper we consider the problem of *binary classification* for multivariate data, which is one of the most fundamental problems in machine learning and pattern recognition. In this task, a training dataset consists of a set of pairs $\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$, where $x_i \in \mathbb{R}^d$ is a *feature vector*, $y_i \in \{0,1\}$ is a *label*, and the d-dimensional Euclidean space $\mathbb{R}^d$ is a *feature space*. The goal is to learn a *classifier* from the given training dataset, that is, to find a mapping $h : \mathbb{R}^d \to \{0,1\}$ such that, for all $x \in \mathbb{R}^d$, $h(x)$ is expected to be the same as the true label of x. In other words, such a classifier h is the *characteristic function* of a subset $L = \{x \in \mathbb{R}^d \mid h(x) = 1\}$ of $\mathbb{R}^d$, which has to be similar to the true set $K = \{x \in \mathbb{R}^d \mid$ the true label of x is $1\}$ as far as possible. Throughout the paper, we assume that each feature is normalized by some data preprocessing such as min-max normalization for simplicity, that is, the feature space is the unit interval (cube) $\mathscr{I}^d = [0,1] \times \cdots \times [0,1]$ in the d-dimensional Euclidean space $\mathbb{R}^d$. In many realistic scenarios, each target K is a closed and bounded subset of $\mathscr{I}^d$, *i.e.*, a nonempty compact subset of $\mathscr{I}^d$, called a *figure*. Thus here we address the problem of binary classification by treating it as "learning of figures".

In this machine learning process, we implicitly treat any feature vector through its *representation*, or *code* on a computer, that is, each feature vector $x \in \mathscr{I}^d$ is represented by a sequence p over some alphabet Σ using an encoding scheme ρ. Here such a surjective mapping ρ is called a *representation* and should map the set of "infinite" sequences Σ^ω to $\mathscr{I}^d$ since there is no one-to-one correspondence between finite sequences and real numbers (or real vectors). In this paper, we use the *binary representation* $\rho : \Sigma^\omega \to [0,1]$ with $\Sigma = \{\mathtt{0},\mathtt{1}\}$, which is defined by $\rho(p) := \sum p_i \cdot 2^{-(i+1)}$ for an infinite sequence $p = p_0p_1p_2\ldots$. For example, $\rho(\mathtt{0100}\ldots) = 0.25$, $\rho(\mathtt{1000}\ldots) = 0.5$, and $\rho(\mathtt{0111}\ldots) = 0.5$. However, we cannot treat infinite sequences on a computer in finite time and, instead, we have to use *discretized* values, *i.e.*, *truncated finite sequences* in any actual machine learning process. Thus in learning of a classifier h for the target figure K, we cannot use an exact data point $x \in K$ but have to use a discretized finite sequence $w \in \Sigma^*$ which tells us that x takes one of the values in the set $\{\rho(p) \mid w \sqsubset p\}$ ($w \sqsubset p$ means that w is a *prefix* of p). For instance, if $w = \mathtt{01}$, then x should be in the interval $[0.25,0.5]$. For a finite sequence $w \in \Sigma^*$, we define $\rho(w) := \{\rho(p) \mid w \sqsubset p$ with $p \in \Sigma^\omega\}$ using the same symbol ρ. From a geometric point of view, $\rho(w)$ means a hyper-rectangle whose sides are parallel to the axes in the space $\mathscr{I}^d$. For example, for the binary representation ρ, we have $\rho(\mathtt{0}) = [0,0.5]$, $\rho(\mathtt{1}) = [0.5,1]$, $\rho(\mathtt{01}) = [0.25,0.5]$, and so on. Therefore in the actual learning process, while a target set K and each point $x \in K$ exist mathematically, a learning machine can only treat finite sequences as training data.

Here the problem of binary classification is stated in a computational manner as follows: Given a training dataset $\{(w_1,y_1),(w_2,y_2),\ldots,(w_n,y_n)\}$ ($w_i \in \Sigma^*$ for each $i \in \{1,2,\ldots,n\}$), where $y_i = 1$ if $\rho(w_i) \cap K \neq \emptyset$ for a target figure $K \subseteq \mathscr{I}^d$ and $y_i = 0$ otherwise, learn a classifier $h : \Sigma^* \to \{0,1\}$ for which $h(w)$ should be the same as the true label of w for all $w \in \Sigma^*$. Each training datum (w_i,y_i) is called a *positive example* if $y_i = 1$ and a *negative example* if $y_i = 0$.

Assume that a figure K is represented by a set P of infinite sequences, *i.e.*, $\{\rho(p) \mid p \in P\} = K$, using the binary representation ρ. Then learning the figure is different from learning the well-known *prefix closed set* $\mathrm{Pref}(P)$, defined as $\mathrm{Pref}(P) := \{w \in \Sigma^* \mid w \sqsubset p$ for some $p \in P\}$, since generally $\mathrm{Pref}(P) \neq \{w \in \Sigma^* \mid \rho(w) \cap K \neq \emptyset\}$ holds. For example, if $P = \{p \in \Sigma^\omega \mid \mathtt{1} \sqsubset p\}$, the corresponding figure K is the interval $[0.5,1]$. Then, the infinite sequence $\mathtt{0111}\ldots$ is a positive example since $\rho(\mathtt{0111}\ldots) = 0.5$ and $\rho(\mathtt{0111}\ldots) \cap K \neq \emptyset$, but it is not contained in $\mathrm{Pref}(P)$. This problem is fundamentally due to rational numbers having two representations, for example, both $\mathtt{0111}\ldots$ and $\mathtt{1000}\ldots$ represent 0.5. Solving this mismatch between objects of learning and their representations is one of the challenging problems of learning continuous objects based on their representation in a computational manner.

For finite expression of classifiers, we use *self-similar sets* known as *fractals* (Mandelbrot 1982) to exploit their simplicity and the power of expression theoretically provided by the field of fractal geometry. Specifically, we can approximate any figure by some self-similar set arbitrarily closely (derived from the Collage Theorem given by Falconer (2003)) and can compute it by a simple recursive algorithm, called an *IFS* (*Iterated Function System*) (Barnsley 1993; Falconer 2003). This approach can be viewed as the analog of the discrete Fourier analysis, where *FFT* (*Fast Fourier Transformation*) is used as the fundamental recursive algorithm. Moreover, in the process of sampling from analog data in discrete Fourier analysis, *scalability* is a desirable property. It requires that when the sample resolution increases, the accuracy of the result is monotonically refined. We formalize this property as *effective learning* of figures, which is inspired by *effective computing* in the framework

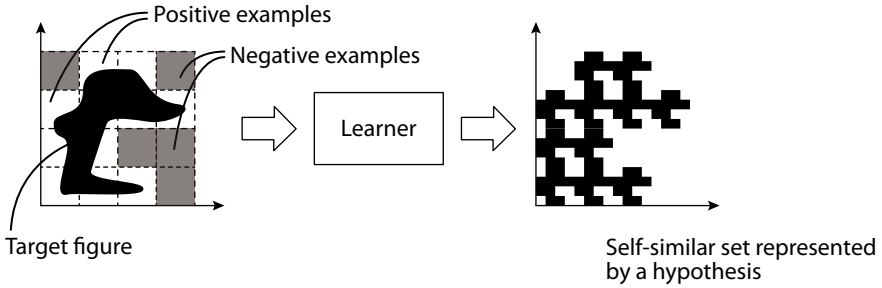

Fig. 1 Our framework of learning figures.

of Type-2 Theory of Effectivity (TTE) studied in computable analysis (Schröder 2002b; Weihrauch 2000). This model guarantees that as a computer reads more and more precise information of the input, it produces more and more accurate approximations of the result. Here we adapt this model from computation to learning, where if a learner (learning machine) receives more and more accurate training data, it learns better and better classifiers (self-similar sets) approximating the target figure.

To summarize, our framework of learning figures (shown in Figure 1) is as follows: Positive examples are axis-parallel rectangles intersecting the target figure, and negative examples are those disjoint with the target. A learner reads a presentation (infinite sequence of examples), and generates hypotheses. Hypotheses are finite sequences (codes) that are discrete expressions of self-similar sets. To evaluate "goodness" of each classifier, we use the concept of *generalization error* and measure the error by the *Hausdorff metric* since it induces the standard topology on the set of figures (Beer 1993).

The main contributions of this paper are as follows:

1. We formalize the learning of figures using self-similar sets based on Gold's learning model towards realizing fully computable binary classification (Section 3). We construct a representational system for learning using self-similar sets based on the binary representation of real numbers, and show desirable properties of it (Lemmas 3.2, 3.3, and 3.4).
2. We construct a learnability hierarchy under various learning criteria, summarized in Figure 3 (Section 4 and 5). We consider five criteria for learning: explanatory learning (Section 4.1), consistent learning (Section 4.2), reliable and refutable learning (Section 4.3), and effective learning (Section 5).
3. We show a mathematical connection between learning and fractal geometry by measuring the complexity of learning using the Hausdorff dimension and the VC dimension (Section 6). Specifically, we give a lower bound on the number of positive examples using the dimensions.
4. We also show a connection between computability of figures studied in computable analysis and learnability of figures discussed in this paper using TTE (Section 7). Learning can be viewed as computable realization of the identity from the set of figures to the same set equipped with a finer topology.

The rest of the paper is organized as follows: We review related work in comparison to the present work in Section 2. We formalize computable binary classification as learning of figures in Section 3 and analyze the learnability hierarchy induced by variants of our model in Section 4 and Section 5. The mathematical connection between fractal geometry and Gold's model with the Hausdorff and the VC dimensions is presented in Section 6 and between computability and learnability of figures in Section 7. Section 8 gives the conclusion.

A preliminary version of this paper was presented at the 21st International Conference on Algorithmic Learning Theory (Sugiyama et al. 2010). In this paper, formalization of learning in Section 3 is completely updated for clarity and simplicity, and all theorems and lemmas have formal proofs (they were omitted in the conference paper). Furthermore, discussion about related work in Section 2 and TTE analysis in Section 7 are new contributions. In addition, several examples and figures are added for readability.

Table 1 Notation.

$\mathbb{N}$	The set of natural numbers including 0
$\mathbb{N}^+$	The set of positive natural numbers, *i.e.*, $\mathbb{N}^+ = \mathbb{N} \setminus \{0\}$
$\mathbb{Q}$	The set of rational numbers
$\mathbb{R}$	The set of real numbers
$\mathbb{R}^+$	The set of positive real numbers
d	The number of dimensions ($d \in \mathbb{N}^+$)
$\mathbb{R}^d$	d-dimensional Euclidean space
$\mathscr{K}^*$	The set of figures (nonempty compact subsets of $\mathbb{R}^d$)
$\mathscr{I}^d$	The unit interval $[0,1] \times \cdots \times [0,1]$
K, L	Figures (nonempty compact sets)
$\#X$	The number of elements in X
$\mathscr{F}$	Set of figures
φ	Contraction for real numbers
C	Finite set of contractions
Φ	Contraction for figures
Σ	Alphabet
Σ^d	The set of finite sequences whose length are d, *i.e.*, $\Sigma^d = \{a_1 a_2 \dots a_d \mid a_i \in \Sigma\}$
Σ^*	The set of finite sequences
Σ^+	The set of finite sequences without the empty string λ
Σ^ω	The set of infinite sequences
λ	The empty string
u, v, w	Finite sequences
$w \sqsubseteq p$	w means a prefix of p ($w \sqsubset p$ is $w \sqsubseteq p$ and $w \neq p$)
$\uparrow w$	The set $\{p \in \Sigma^\omega \mid w \sqsubset p\}$
$\langle \cdot \rangle$	The tupling function, *i.e.*, $\langle p^1, p^2, \dots, p^d \rangle := p^1_0 p^2_0 \dots p^d_0 p^1_1 p^2_1 \dots p^d_1 p^1_2 p^2_2 \dots p^d_2 \dots$
$\lvert w \rvert$	The length of w. If $w = \langle w^1, \dots, w^d \rangle \in (\Sigma^d)^*$, $\lvert w \rvert = \lvert w^1 \rvert = \cdots = \lvert w^d \rvert$
$\mathrm{diam}(k)$	The diameter of the set $\rho(w)$ with $\lvert w \rvert = k$, *i.e.*, $\mathrm{diam}(k) = \sqrt{d} \cdot 2^{-k}$
p, q	Infinite sequences
V, W	Set of finite or infinite sequences
ρ	Binary representation
ξ, ζ	Representation, *i.e.*, a mapping from finite or infinite sequences to some objects
$\xi \leqslant \zeta$	ξ is reducible to ζ
$\xi \equiv \zeta$	ξ is equivalent to ζ
$\nu_{\mathbb{Q}^d}$	Representation for rational numbers
$\nu_{\mathscr{Q}}$	Representation for finite sets of rational numbers
$\mathscr{H}$	The hypothesis space (The set of finite sets of finite sequences)
H	Hypothesis
h	Classifier of hypothesis H
κ	The mapping from hypotheses to figures
$\mathbf{M}$	Learner
σ	Presentation (informant or text)
$\mathrm{Pos}(K)$	The set of finite sequences of positive examples of K, *i.e.*, $\{w \mid \rho(w) \cap K \neq \emptyset\}$
$\mathrm{Pos}_k(K)$	The set $\{w \in \mathrm{Pos}(K) \mid \lvert w \rvert = k\}$
$\mathrm{Neg}(K)$	The set of finite sequences of negative examples of K, *i.e.*, $\{w \mid \rho(w) \cap K = \emptyset\}$
d_{E}	The Euclidean distance
d_{H}	The Hausdorff distance
$\mathfrak{H}$	The Hausdorff measure
$\dim_{\mathrm{H}}$	The Hausdorff dimension
$\dim_{\mathrm{B}}$	The box-counting dimension
$\dim_{\mathrm{S}}$	The similarity dimension
$\dim_{\mathrm{VC}}$	The VC dimension

2 Related Work

Statistical approaches to machine learning are now achieving great success since they are originally designed for analyzing observed multivariate data and, to date, many statistical methods have been proposed to treat continuous objects such as real-valued functions (Bishop 2007). However, most methods pay no attention to discretization and the finite representation of analog data on a computer. For example, multi-layer perceptrons

are used to learn real-valued functions, since they can approximate every continuous function arbitrarily and accurately. However, a perceptron is based on the idea of regulating analog wiring (Rosenblatt 1958), hence such learning is not purely computable, *i.e.*, it ignores the gap between analog raw data and digital discretized data. Furthermore, although several discretization techniques have been proposed by Elomaa and Rousu (2003); Fayyad and Irani (1993); Gama and Pinto (2006); Kontkanen et al (1997); Li et al (2003); Lin et al (2003); Liu et al (2002); Skubacz and Hollmén (2000), they treat discretization as data preprocessing for improving the accuracy or efficiency of machine learning algorithms. The process of discretization is therefore not considered from a computational point of view, and "computability" of machine learning algorithms is not discussed at sufficient depth.

There are several related articles considering learning under various restrictions in Gold's model (Goldman et al 2003), Valiant's model (Ben-David and Dichterman 1998; Decatur and Gennaro 1995), and other learning context (Khardon and Roth 1999). Moreover, recently learning from partial examples, or examples with missing information, has attracted much attention in Valiant's learning model (Michael 2010, 2011). In this paper we also consider learning from examples with missing information, which are truncated finite sequences. However, our model is different from the cited work, since the "missing information" in this paper corresponds to *measurement error* of real-valued data. Our motivation comes from actual measurement/observation of a physical object, where every datum obtained by an experimental instrument must have some numerical error in principle (Baird 1994). For example, if we measure the size of a cell by a microscope equipped with micrometers, we cannot know the true value of the size but an approximate value with numerical error, which depends on the degree of magnification by the micrometers. In this paper we try to treat this process as learning from multivariate data, where an approximate value corresponds to a truncated finite sequence and error becomes small as the length of the sequence increases. The model of computation for real numbers within the framework of TTE, as mentioned in the introduction, fits our motivation, and this approach is unique in computational learning theory.

Self-similar sets can be viewed as a geometric interpretation of languages recognized by ω-*automata* (Perrin and Pin 2004), first introduced by Büchi (1960), and learning of such languages has been investigated by De La Higuera and Janodet (2001); Jain et al (2011). Both works focus on learning ω-languages from their prefixes, *i.e.* texts (positive data), and show several learnable classes. This approach is different from ours since our motivation is to address computability issues in the field of machine learning from numerical data, and hence there is a gap between prefixes of ω-languages and positive data for learning in our setting as mentioned in the introduction. Moreover, we consider learning from both positive and negative data, which is a new approach in the context of learning of infinite words.

Recently, two of the authors, Sugiyama and Yamamoto (2010), have addressed discretization of real vectors in a computational approach and proposed a new similarity measure, called *coding divergence*. It evaluates the similarity between two sets of real vectors and can be applied to many machine learning tasks such as classification and clustering. However, it does not address the issue of the learnability or complexity of learning of continuous objects.

3 Formalization of Learning

To analyze binary classification in a computable approach, we first formalize learning of figures based on Gold's model. Specifically, we define targets of learning, representations of classifiers produced by a learning machine, and a protocol for learning. In the following, let $\mathbb{N}$ be the set of natural numbers including 0, $\mathbb{Q}$ the set of rational numbers, and $\mathbb{R}$ the set of real numbers. The set $\mathbb{N}^+$ (resp. $\mathbb{R}^+$) is the set of positive natural (resp. real) numbers. The d-fold product of $\mathbb{R}$ is denoted by $\mathbb{R}^d$ and the set of nonempty compact subsets of $\mathbb{R}^d$ is denoted by $\mathscr{K}^*$. Notations used in this paper are summarized in Table 1.

Throughout this paper, we use the *binary representation* $\rho^d : (\Sigma^d)^\omega \to \mathscr{I}^d$ as the canonical representation for real numbers. If $d = 1$, this is defined as follows: $\Sigma = \{\mathtt{0}, \mathtt{1}\}$ and

$$\rho^1(p) := \sum_{i=0}^{\infty} p_i \cdot 2^{-(i+1)} \tag{1}$$

for an infinite sequence $p = p_0p_1p_2\ldots$. Note that Σ^d denotes the set $\{a_1a_2\ldots a_d \mid a_i \in \Sigma\}$ and $\Sigma^1 = \Sigma$. For example, $\rho^1(\mathtt{0100}\ldots) = 0.25$, $\rho^1(\mathtt{1000}\ldots) = 0.5$, and so on. Moreover, by using the same symbol ρ, we introduce a representation $\rho^1 : \Sigma^* \to \mathscr{K}^*$ for finite sequences defined as follows:

$$\rho^1(w) := \rho^1(\uparrow w) = [\rho^1(w\mathtt{000}\ldots), \rho^1(w\mathtt{111}\ldots)] = \left[\sum w_i \cdot 2^{-(i+1)},\ \sum w_i \cdot 2^{-(i+1)} + 2^{-|w|}\right], \tag{2}$$

where $\uparrow w = \{p \in \Sigma^\omega \mid w \sqsubset p\}$. For instance, $\rho^1(\mathtt{01}) = [0.25, 0.5]$ and $\rho^1(\mathtt{10}) = [0.5, 0.75]$.

In a d-dimensional space with $d > 1$, we use the *d-dimensional binary representation* $\rho^d : (\Sigma^d)^\omega \to \mathscr{I}^d$ defined in the following manner.

$$\rho^d(\langle p^1, p^2, \ldots, p^d\rangle) := \left(\rho^1(p^1), \rho^1(p^2), \ldots, \rho^1(p^d)\right), \tag{3}$$

where d infinite sequences p^1, p^2, …, and p^d are concatenated using the *tupling function* $\langle\cdot\rangle$ such that

$$\langle p^1, p^2, \ldots, p^d\rangle := p^1_0p^2_0\ldots p^d_0p^1_1p^2_1\ldots p^d_1p^1_2p^2_2\ldots p^d_2\ldots.$$

Similarly, we define a representation $\rho^d : (\Sigma^d)^* \to \mathscr{K}^*$ by

$$\rho^d(\langle w^1, w^2, \ldots, w^d\rangle) := \rho^d(\uparrow\langle w^1, w^2, \ldots, w^d\rangle),$$

where

$$\langle w^1, w^2, \ldots, w^d\rangle := w^1_0w^2_0\ldots w^d_0w^1_1w^2_1\ldots w^d_1\ldots w^1_nw^2_n\ldots w^d_n.$$

with $|w^1| = |w^2| = \cdots = |w^d| = n$. Note that, for any $w = \langle w^1, \ldots, w^d\rangle \in (\Sigma^d)^*$, $|w^1| = |w^2| = \cdots = |w^d|$ always holds, and we denote the length by $|w|$ in this paper. For a set of finite sequences, *i.e.*, a *language* $L \subset (\Sigma^d)^*$, we define

$$\rho^d(L) := \{\rho^d(w) \mid w \in L\}.$$

We omit the superscript d of ρ^d if it is understood from the context.

A target set of learning is a set of figures $\mathscr{F} \subseteq \mathscr{K}^*$ fixed *a priori*, and one of them is chosen as a target in each learning phase. A learning machine uses *self-similar sets*, known as fractals and defined by finite sets of contractions. This approach is one of the key ideas in this paper. Here, a *contraction* is a mapping $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ such that, for all $x, y \in X$, $d(\varphi(x), \varphi(y)) \leqslant cd(x,y)$ for some real number c with $0 < c < 1$. For a finite set of contractions C, a nonempty compact set F satisfying

$$F = \bigcup_{\varphi\in C} \varphi(F)$$

is determined uniquely (see (Falconer 2003) for a formal proof). The set F is called the *self-similar set* of C. Moreover, if we define a mapping $\Phi : \mathscr{K}^* \to \mathscr{K}^*$ by

$$\Phi(K) := \bigcup_{\varphi\in C} \varphi(K) \tag{4}$$

and define

$$\Phi^0(K) := K \text{ and } \Phi^{k+1}(K) := \Phi(\Phi^k(K)) \tag{5}$$

for each $k \in \mathbb{N}$ recursively, then

$$F = \bigcap_{k=0}^{\infty} \Phi^k(K)$$

for every $K \in \mathscr{K}^*$ such that $\varphi(K) \subset K$ for every $\varphi \in C$. This means that we have a level-wise construction algorithm with Φ to obtain the self-similar set F.

A learning machine produces *hypotheses*, each of which is a finite language and becomes a finite expression of a self-similar set that works as a classifier. Formally, for a finite language $H \subset (\Sigma^d)^*$, we consider $H^0, H^1, H^2, \ldots$ such that H^k is recursively defined as follows:

$$\begin{cases} H^0 := \{\lambda\}, \\ H^k := \{\langle w^1u^1, w^2u^2, \ldots, w^du^d\rangle \mid \langle w^1, w^2, \ldots, w^d\rangle \in H^{k-1} \text{ and } \langle u^1, u^2, \ldots, u^d\rangle \in H\}, \end{cases}$$

We can easily construct a fixed program $P(\cdot)$ which generates $H^0, H^1, H^2, \ldots$ when receiving a hypothesis H. We give the semantics of a hypothesis H by the following equation:

$$\kappa(H) := \bigcap_{k=0}^{\infty} \bigcup \rho(H^k). \tag{6}$$

Since $\bigcup \rho(H^k) \supset \bigcup \rho(H^{k+1})$ holds for all $k \in \mathbb{N}$, $\kappa(H) = \lim_{k\to\infty} \bigcup \rho(H^k)$. We denote the set of hypotheses $\{H \subset (\Sigma^d)^* \mid H \text{ is finite}\}$ by $\mathscr{H}$ and call it the *hypothesis space*. We use this hypothesis space throughout the paper. Note that, for a pair of hypotheses H and L, $H = L$ implies $\kappa(H) = \kappa(L)$, but the converse may not hold.

Example 3.1 Assume $d = 2$ and let a hypothesis H be the set $\{\langle 0,0\rangle, \langle 0,1\rangle, \langle 1,1\rangle\} = \{00, 01, 11\}$. We have

$$\begin{aligned} H^0 &= \emptyset, \quad H^1 = \{\langle 0,0\rangle, \langle 0,1\rangle, \langle 1,1\rangle\} = \{00, 01, 11\}, \\ H^2 &= \{\langle 00,00\rangle, \langle 00,01\rangle, \langle 01,01\rangle, \langle 00,10\rangle, \langle 00,11\rangle, \langle 01,11\rangle, \langle 10,10\rangle, \langle 10,11\rangle, \langle 11,11\rangle\} \\ &= \{0000, 0001, 0011, 0100, 0101, 0111, 1100, 1101, 1111\}, \ldots \end{aligned}$$

and the figure $\kappa(H)$ defined in the equation (6) is the *Sierpiński triangle* (Figure 2). If we consider the following three mappings:

$$\varphi_1 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \ \varphi_2 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1/2 \end{bmatrix}, \ \varphi_3 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix},$$

the three squares $\varphi_1(\mathscr{I}^d)$, $\varphi_2(\mathscr{I}^d)$, and $\varphi_3(\mathscr{I}^d)$ are exactly the same as $\rho(00)$, $\rho(01)$, and $\rho(11)$, respectively. Thus each sequence in a hypothesis can be viewed as a representation of one of these squares, which are called *generators* for a self-similar set since if we have the initial set $\mathscr{I}^d$ and generators $\varphi_1(\mathscr{I}^d)$, $\varphi_2(\mathscr{I}^d)$, and $\varphi_3(\mathscr{I}^d)$, we can reproduce the three mappings φ_1, φ_2, and φ_3 and construct the self-similar set from them. Note that there exist infinitely many hypotheses L such that $\kappa(H) = \kappa(L)$ and $H \neq L$. For example, $L = \{\langle 0,0\rangle$, $\langle 1,1\rangle$, $\langle 00,10\rangle$, $\langle 00,11\rangle$, $\langle 01,11\rangle\}$.

Lemma 3.2 (Soundness of hypotheses) *For every hypothesis $H \in \mathscr{H}$, the set $\kappa(H)$ defined by the equation* (6) *is a self-similar set.*

Proof Let $H = \{w_1, w_2, \ldots, w_n\}$. We can easily check that the set of rectangles $\rho(w_1), \rho(w_2), \ldots, \rho(w_n)$ is a generator defined by the mappings $\varphi_1, \varphi_2, \ldots, \varphi_n$, where each φ_i maps the unit interval $\mathscr{I}^d$ to the figure $\rho(w_i)$. Define Φ and Φ^k in the same way as the equations (4) and (5). For each $k \in \mathbb{N}$,

$$\bigcup \rho(H^k) = \Phi^k(\mathscr{I}^d)$$

holds. It therefore follows that the set $\kappa(H)$ is exactly the same as the self-similar set defined by the mappings $\varphi_1, \varphi_2, \ldots, \varphi_n$, that is, $\kappa(H) = \bigcup \varphi_i(\kappa(H))$ holds. □

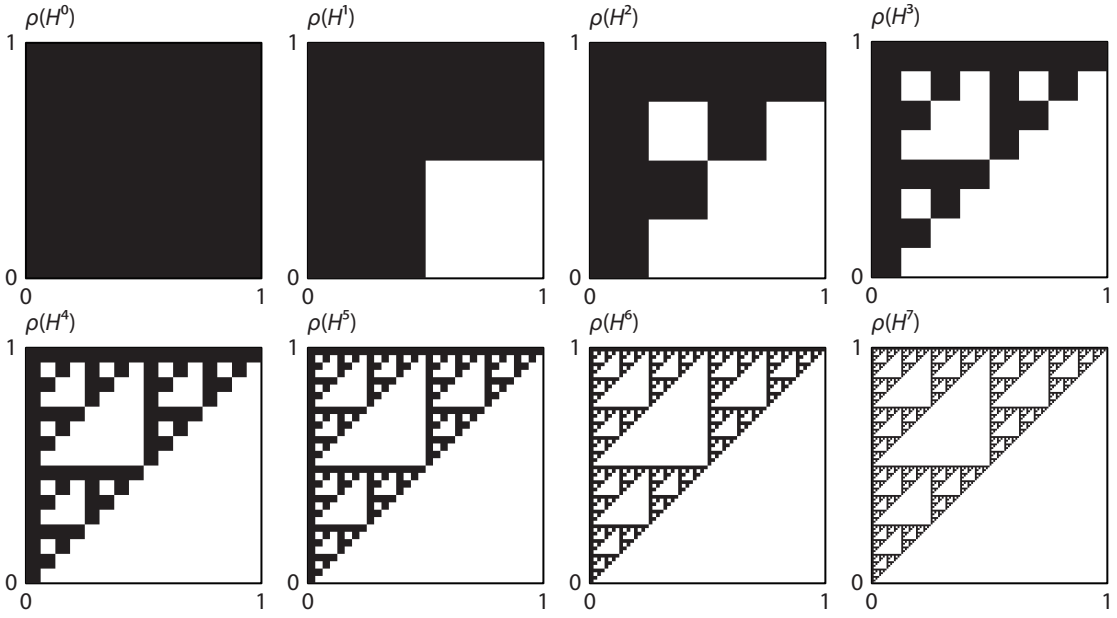

Fig. 2 Generation of the Sierpiński triangle from the hypothesis $H = \{\langle 0,0\rangle, \langle 0,1\rangle, \langle 1,1\rangle\}$ (Example 3.1).

To evaluate the "goodness" of each hypothesis, we use the concept of *generalization error*, which is usually used to score the quality of hypotheses in a machine learning context. The generalization error of a hypothesis H for a target figure K, written by $\mathrm{GE}(K,H)$, is defined by the *Hausdorff metric* d_{H} on the space of figures, *i.e.*,

$$\mathrm{GE}(K,H) := d_{\mathrm{H}}(K,\kappa(H)) = \inf\{\,\delta \mid K \subseteq \kappa(H)_\delta \text{ and } \kappa(H) \subseteq K_\delta\,\},$$

where K_δ is the δ-*neighborhood* of K defined by

$$K_\delta := \{\,x \in \mathbb{R}^d \mid d_{\mathrm{E}}(x,a) \leqslant \delta \text{ for some } a \in K\,\}.$$

The metric d_{E} is the Euclidean metric such that

$$d_{\mathrm{E}}(x,a) = \sqrt{\sum_{i=1}^{d}(x^i - a^i)^2}$$

for $x = (x^1,\ldots,x^d), a = (a^1,\ldots,a^d) \in \mathbb{R}^d$. The Hausdorff metric is one of the standard metrics on the space since the metric space $(\mathscr{K}^*, d_{\mathrm{H}})$ is complete (in the sense of topology) and $\mathrm{GE}(K,H) = 0$ if and only if $K = \kappa(H)$ (Beer 1993; Kechris 1995). The topology on $\mathscr{K}^*$ induced by the Hausdorff metric is called the *Vietoris topology*. Since the cardinality of the set of hypotheses $\mathscr{H}$ is smaller than that of the set of figures $\mathscr{K}^*$, we often cannot find the exact hypothesis H for a figure K such that $\mathrm{GE}(K,H) = 0$. However, following the Collage Theorem given by Falconer (2003), we show that the power of representation of hypotheses is still sufficient, that is, we always can approximate a given figure arbitrarily closely by some hypothesis.

Lemma 3.3 (Representational power of hypotheses) *For any $\delta \in \mathbb{R}$ and for every figure $K \in \mathscr{K}^*$, there exists a hypothesis H such that* $\mathrm{GE}(K,H) < \delta$.

Proof Fix a figure K and the parameter δ. Here we denote the diameter of the set $\rho(w)$ with $|w| = k$ by $\mathrm{diam}(k)$. Then we have

$$\mathrm{diam}(k) = \sqrt{d} \cdot 2^{-k}.$$

For example, $\mathrm{diam}(1) = 1/2$ and $\mathrm{diam}(2) = 1/4$ if $d = 1$, and $\mathrm{diam}(1) = 1/\sqrt{2}$ and $\mathrm{diam}(2) = 1/\sqrt{8}$ if $d = 2$. For k with $\mathrm{diam}(k) < \delta$, let

$$H = \{\, w \in (\Sigma^d)^* \mid |w| = k \text{ and } \rho(w) \cap K \neq \emptyset \,\}.$$

We can easily check that the $\mathrm{diam}(k)$-neighborhood of K contains $\kappa(H)$ and the $\mathrm{diam}(k)$-neighborhood of $\kappa(H)$ contains K. Therefore we have $\mathrm{GE}(K, H) < \delta$. □

Moreover, to work as a classifier, every hypothesis H has to be *computable*, that is, the function $h : (\Sigma^d)^* \to \{0, 1\}$ such that, for all $w \in (\Sigma^d)^*$,

$$h(w) = \begin{cases} 1 & \text{if } \rho(w) \cap \kappa(H) \neq \emptyset, \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

should be computable. We say that such h is the *classifier* of H. The computability of h is not trivial, since for a finite sequence w, the two conditions $h(w) = 1$ and $w \in H^k$ are not equivalent. Intuitively, this is because each interval represented by a finite sequence is *closed*. For example, in the case of Example 3.1, $h(\mathtt{10}) = 1$ because $\rho(\mathtt{10}) = [0.5, 1] \times [0, 0.5]$ and $\rho(\mathtt{10}) \cap \kappa(H) = \{(0.5, 0.5)\} \neq \emptyset$ whereas $\mathtt{10} \notin H^k$ for any $k \in \mathbb{N}$. Here we guarantee this property of computability.

Lemma 3.4 (Computability of classifiers) *For every hypothesis $H \in \mathscr{H}$, the classifier h of H defined by the equation* (7) *is computable.*

Proof First we consider whether or not the boundary of an interval is contained in $\kappa(H)$. Suppose $d = 1$ and let C be a finite set of contractions and F be the self-similar set of C. We have the following property: Let $[x, y] = \varphi_1 \circ \varphi_2 \circ \cdots \circ \varphi_n(\mathscr{I}^1)$ for some $\varphi_1, \varphi_2, \ldots, \varphi_n \in C$ and let $I = \varphi'_1 \circ \varphi'_2 \circ \cdots \circ \varphi'_{n'}(\mathscr{I}^1)$ for $\varphi'_1, \varphi'_2, \ldots, \varphi'_{n'} \in C$. Assume that, if n' is large enough, there is no such I satisfying $x \in I$ and $\min I < x$ (resp. $\max I > y$). Then, we have $x \in F$ (resp. $y \in F$) if and only if $0 \in \varphi(\mathscr{I}^1)$ (resp. $1 \in \varphi(\mathscr{I}^1)$) for some $\varphi \in C$. This means that if $[x, y] = \rho(v)$ with a sequence $v \in H^k$ ($k \in \mathbb{N}$) for a hypothesis H, where there is no sequence $v' \in H^{k'}$ with $x \in \rho(v')$ and $\min \rho(v') < x$ (resp. $\max \rho(v') > y$) when k' is large enough, we have $x \in \kappa(H)$ (resp. $y \in \kappa(H)$) if and only if $u \in \{\mathtt{0}\}^+$ (resp. $u \in \{\mathtt{1}\}^+$) for some $u \in H$.

We show a pseudo-code of the classifier h in Algorithm 1 and prove that the output of the algorithm is 1 if and only if $h(w) = 1$, *i.e.*, $\rho(w) \cap \kappa(H) \neq \emptyset$. In the algorithm, $\underline{v^s}$ and $\overline{v^s}$ denote the previous and subsequent binary sequences of v^s with $|v^s| = |\underline{v^s}| = |\overline{v^s}|$ in the lexicographic order, respectively. For example, if $v^s = \mathtt{001}$, $\underline{v^s} = \mathtt{000}$ and $\overline{v^s} = \mathtt{010}$. Moreover, we use the special symbol $\perp$ meaning undefinedness, that is, $v = w$ if and only if $v_i = w_i$ for all $i \in \{0, 1, \ldots, |v| - 1\}$ with $v_i \neq \perp$ and $w_i \neq \perp$.

The "if" part: For an input of a finite sequence w and a hypothesis H, if $h(w) = 1$, there are two possibilities as follows:

1. For some $k \in \mathbb{N}$, there exists $v \in H^k$ such that $w \sqsubseteq v$. This is because $\rho(w) \supseteq \rho(v)$ and $\rho(v) \cap \kappa(H) \neq \emptyset$.
2. The above condition does not hold, but $\rho(w) \cap \kappa(H) \neq \emptyset$.

In the first case, the algorithm goes to line 7 and stops with outputting 1. The second case means that the algorithm uses the function CHECKBOUNDARY. Since $h(w) = 1$, there should exist a sequence $v \in H$ such that $u = aaa \ldots a$ for some $u \in H$, where a is obtained in lines 1–10. CHECKBOUNDARY therefore returns 1.

The "only if" part: In Algorithm 1, if $v \in H^k$ satisfies conditions in line 6 or line 8, $h(w) \cap \kappa(H) \neq \emptyset$. Thus $h(w) = 1$ holds. □

Algorithm 1: Classifier h of hypothesis H

Input: Finite sequence w and hypothesis H
Output: Class label 1 or 0 of w

$k \leftarrow 0$
repeat
  $k \leftarrow k+1$
until $\min_{v \in H^k} |v| > |w|$
for each $v \in H^k$
  if $w \sqsubseteq v$ **then**
    output 1 and halt
  else if CHECKBOUNDARY$(w,v,H) = 1$ **then**
    output 1 and halt
  end if
end for
output 0

function CHECKBOUNDARY(w, v, H)

for each s in $\{1,2,\ldots,d\}$
  if $w^s \sqsubseteq v^s$ **then** $a^s \leftarrow \bot$
  else
    if $w^s \sqsubseteq \underline{v^s}$ **then** $a^s \leftarrow \mathtt{0}$
    else if $w^s \sqsubseteq \overline{v^s}$ **then** $a^s \leftarrow \mathtt{1}$
    else return 0
    end if
  end if
end for
$a \leftarrow a^1 a^2 \ldots a^d$ // a is a finite sequence whose length is d
for each $u \in H$
  if $u = aaa\ldots a$ **then return** 1
end for
return 0

The set $\{\kappa(H) \mid H \subset (\Sigma^d)^*$ and the classifier h of H is computable$\}$ exactly corresponds to an *indexed family of recursive concepts / languages* discussed in computational learning theory (Angluin 1980), which is a common assumption for learning of languages. On the other hand, there exists some class of figures $\mathscr{F} \subseteq \mathscr{K}^*$ that is not an indexed family of recursive concepts. This means that, for some figure K, there is no *computable* classifier which classifies all data correctly. Therefore we address the problems of both exact and approximate learning of figures to obtain a computable classifier for any target figure.

We consider two types of input data stream, one includes both positive and negative data and the other includes only positive data, to analyze learning based on Gold's learning model. Formally, each training datum is called an *example* and is defined as a pair (w,l) of a finite sequence $w \in (\Sigma^d)^*$ and a label $l \in \{0,1\}$. For a target figure K,

$$l = \begin{cases} 1 & \text{if } \rho(w) \cap K \neq \emptyset \ \ (\textit{positive example}), \\ 0 & \text{otherwise} \ \ (\textit{negative example}). \end{cases}$$

In the following, for a target figure K, we denote the set of finite sequences of positive examples $\{w \in (\Sigma^d)^* \mid \rho(w) \cap K \neq \emptyset\}$ by $\mathrm{Pos}(K)$ and that of negative examples by $\mathrm{Neg}(K)$. Moreover, we denote $\mathrm{Pos}_k(K) = \{w \in \mathrm{Pos}(K) \mid |w| = k\}$. From the geometric nature of figures, we obtain the following *monotonicity* of examples:

Table 2 Relationship between the conditions for each finite sequence $w \in \Sigma^*$ and the standard notation of binary classification.

		Target figure K	
		$w \in \mathrm{Pos}(K)$ $(\rho(w) \cap K \neq \emptyset)$	$w \in \mathrm{Neg}(K)$ $(\rho(w) \cap K = \emptyset)$
Hypothesis H	$h(w) = 1$ $(\rho(w) \cap \kappa(H) \neq \emptyset)$	True positive	False positive (Type I error)
	$h(w) = 0$ $(\rho(w) \cap \kappa(H) = \emptyset)$	False negative (Type II error)	True negative

Lemma 3.5 (Monotonicity of examples) *If $(v,1)$ is an example of K, then $(w,1)$ is an example of K for all prefixes $w \sqsubseteq v$, and $(va,1)$ is an example of K for some $a \in \Sigma^d$. If $(w,0)$ is an example of K, then $(wv,0)$ is an example of K for all $v \in (\Sigma^d)^*$.*

Proof From the definition of the representation ρ in the equations (1) and (3), if $w \sqsubseteq v$, we have $\rho(w) \supseteq \rho(v)$, hence $(w,1)$ is an example of K. Moreover,

$$\bigcup_{a \in \Sigma^d} \rho(va) = \rho(u)$$

holds. Thus there should exist an example $(va,1)$ for some $a \in \Sigma^d$. Furthermore, for all $v \in \Sigma^*$, $\rho(wv) \subset \rho(w)$. Therefore if $K \cap \rho(w) = \emptyset$, then $K \cap \rho(wv) = \emptyset$ for all $v \in (\Sigma^d)^*$, and $(wv,0)$ is an example of K. □

We say that an infinite sequence σ of examples of a figure K is a *presentation* of K. The ith example is denoted by $\sigma(i-1)$, and the set of all examples occurring in σ is denoted by $\mathrm{range}(\sigma)$[3]. The initial segment of σ of length n, *i.e.*, the sequence $\sigma(0), \sigma(1), \ldots, \sigma(n-1)$, is denoted by $\sigma[n-1]$. A *text* of a figure K is a presentation σ such that

$$\{\, w \mid (w,1) \in \mathrm{range}(\sigma) \,\} = \mathrm{Pos}(K)\ (= \{\, w \mid \rho(w) \cap K \neq \emptyset \,\}),$$

and an *informant* is a presentation σ such that

$$\{\, w \mid (w,1) \in \mathrm{range}(\sigma) \,\} = \mathrm{Pos}(K) \text{ and}$$
$$\{\, w \mid (w,0) \in \mathrm{range}(\sigma) \,\} = \mathrm{Neg}(K).$$

Table 2 shows the relationship between the standard terminology in classification and our definitions. For a target figure K and the classifier h of a hypothesis H, the set $\{w \in \mathrm{Pos}(K) \mid h(w) = 1\}$ corresponds to *true positive*, $\{w \in \mathrm{Neg}(K) \mid h(w) = 1\}$ *false positive* (type I error), $\{w \in \mathrm{Pos}(K) \mid h(w) = 0\}$ *false negative* (type II error), and $\{w \in \mathrm{Neg}(K) \mid h(w) = 0\}$ *true negative*.

Let h be the classifier of a hypothesis H. We say that the hypothesis H is *consistent* with an example (w,l) if $l = 1$ implies $h(w) = 1$ and $l = 0$ implies $h(w) = 0$, and consistent with a set of examples E if H is consistent with all examples in E.

A learning machine, called a *learner*, is a procedure, (*i.e.* a Turing machine that never halts) that reads a presentation of a target figure from time to time, and outputs hypotheses from time to time. In the following, we denote a learner by $\mathbf{M}$ and an infinite sequence of hypotheses produced by $\mathbf{M}$ on the input σ by $\mathbf{M}_\sigma$, and $\mathbf{M}_\sigma(i-1)$ denotes the ith hypothesis produced by $\mathbf{M}$. Assume that $\mathbf{M}$ receives j examples $\sigma(0), \sigma(1), \ldots, \sigma(j-1)$ so far when it outputs the ith hypothesis $\mathbf{M}_\sigma(i-1)$. We do not require the condition $i = j$, that is, the inequality $i \leqslant j$ usually holds since $\mathbf{M}$ can "wait" until it receives enough examples. We say that an infinite sequence of hypotheses $\mathbf{M}_\sigma$ *converges* to a hypothesis H if there exists $n \in \mathbb{N}$ such that $\mathbf{M}_\sigma(i) = H$ for all $i \geqslant n$.

[3] The reason for this notation is that σ can be viewed as a mapping from $\mathbb{N}$ (including 0) to the set of examples.

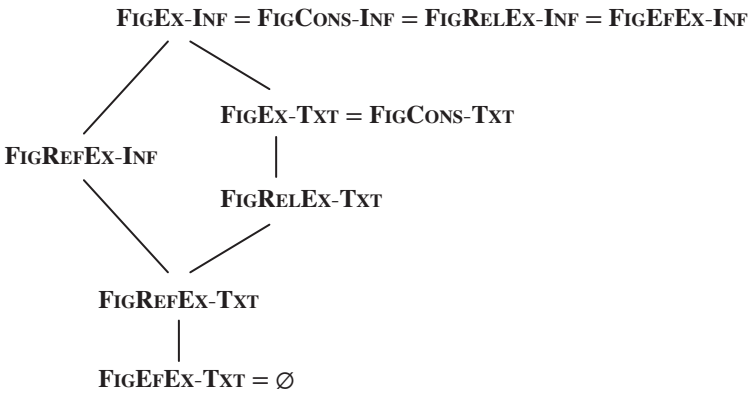

Fig. 3 Learnability hierarchy. For each line, the lower set is a proper subset of the upper set.

4 Exact Learning of Figures

We analyze "exact" learning of figures. This means that, for any target figure K, there should be a hypothesis H such that the generalization error is zero (*i.e.*, $K = \kappa(H)$), hence the classifier h of H can classify all data correctly with no error, that is, h satisfies the equation (7). The goal is to find such a hypothesis H from examples (training data) of K.

In the following two sections (Sections 4 and 5), we follow the standard path of studies in computational learning theory (Jain et al 1999; Jain 2011; Zeugmann and Zilles 2008), that is, we define learning criteria to understand various learning situations and construct a learnability hierarchy under the criteria. We summarize our results in Fig. 3.

4.1 Explanatory Learning

The most basic learning criterion in Gold's model is **Ex**-learning (EX means EXplain), *i.e.*, learning in the limit proposed by Gold (1967). We call these criteria **FIGEX-INF**- (INF means an informant) and **FIGEX-TXT**-learning (TXT means a text) for **Ex**-learning from informants and texts, respectively. We introduce these criteria into the learning of figures, and analyze the learnability of figures.

Definition 4.1 (Explanatory learning) A learner **M** **FIGEX-INF**-*learns* (resp. **FIGEX-TXT**-*learns*) a set of figures $\mathscr{F} \subseteq \mathscr{K}^*$ if for all figures $K \in \mathscr{F}$ and all informants (resp. texts) σ of K, the outputs $\mathbf{M}_\sigma$ converge to a hypothesis H such that $\mathrm{GE}(K,H) = 0$.

For every learning criterion **CR** introduced in the following, we say that a set of figures $\mathscr{F}$ is **CR**-*learnable* if there exists a learner that **CR**-learns $\mathscr{F}$, and denote by **CR** the collection of **CR**-learnable sets of figures following the standard notation of this field (Jain et al 1999).

First, we consider **FIGEX-INF**-learning. Informally, a learner can **FIGEX-INF**-learn a set of figures if it has the ability to enumerate all hypotheses and to judge whether or not each hypothesis is consistent with the received examples (Gold 1967). Here we introduce a convenient enumeration of hypotheses. An infinite sequence of hypotheses $H_0, H_1, \ldots$ is called a *normal enumeration* if $\{H_i \mid i \in \mathbb{N}\} = \mathscr{H}$ and, for all $i, j \in \mathbb{N}$, $i < j$ implies

$$\max_{v \in H_i} |v| \leqslant \max_{w \in H_j} |w|.$$

We can easily implement a procedure that enumerates $\mathscr{H}$ through a normal enumeration.

Procedure 1: Learning procedure that **FIGEX-INF**-learns $\kappa(\mathscr{H})$

Input: Informant $\sigma = (w_0, l_0), (w_1, l_1), \ldots$ of figure $K \in \kappa(\mathscr{H})$
Output: Infinite sequence of hypotheses $\mathbf{M}_\sigma(0), \mathbf{M}_\sigma(1), \ldots$

```
1: i ← 0
2: E ← ∅    // E is a set of received examples
3: repeat
4:    read σ(i) and add to E    // σ(i) = (w_i, l_i)
5:    search the first hypothesis H consistent with E through a normal enumeration
6:    output H    // M_σ(i) = H
7:    i ← i + 1
8: until forever
```

Theorem 4.2 *The set of figures* $\kappa(\mathscr{H}) = \{\, \kappa(H) \mid H \in \mathscr{H} \,\}$ *is* **FIGEX-INF**-*learnable.*

Proof This learning can be done by the well-known strategy of identification by enumeration. We show a pseudo-code of a learner $\mathbf{M}$ that **FIGEX-INF**-learns $\kappa(\mathscr{H})$ in Procedure 1. The learner $\mathbf{M}$ generates hypotheses through normal enumeration. If $\mathbf{M}$ outputs a wrong hypothesis H, there must exist a positive or negative example that is not consistent with the hypothesis since, for a target figure K_*,

$$\mathrm{Pos}(K_*) \ominus \mathrm{Pos}(\kappa(H)) \neq \emptyset$$

for every hypothesis H with $\kappa(H) \neq K_*$, where $X \ominus Y$ denotes the *symmetric difference*, *i.e.*, $X \ominus Y = (X \cup Y) \setminus (X \cap Y)$. Thus the learner $\mathbf{M}$ changes the wrong hypothesis and reaches a correct hypothesis H_* such that $\kappa(H_*) = K_*$ in finite time. If $\mathbf{M}$ produces a correct hypothesis, it never changes the hypothesis, since every example is consistent with it. Therefore the learner $\mathbf{M}$ **FIGEX-INF**-learns $\kappa(\mathscr{H})$. □

Next, we consider **FIGEX-TXT**-learning. In learning of languages from texts, the necessary and sufficient conditions for learning have been studied in detail by Angluin (1980, 1982); Kobayashi (1996); Lange et al (2008); Motoki et al (1991); Wright (1989), and characterization of learnability using finite tell-tale sets is one of the crucial results. We adapt these results into the learning of figures and show the **FIGEX-TXT**-learnability.

Definition 4.3 (Finite tell-tale set, cf. Angluin (1980)) Let $\mathscr{F}$ be a set of figures. For a figure $K \in \mathscr{F}$, a finite subset $\mathscr{T}$ of the set of positive examples $\mathrm{Pos}(K)$ is a *finite tell-tale set of* K *with respect to* $\mathscr{F}$ if for all figures $L \in \mathscr{F}$, $\mathscr{T} \subset \mathrm{Pos}(L)$ implies $\mathrm{Pos}(L) \not\subset \mathrm{Pos}(K)$ (*i.e.*, $L \not\subset K$). If every figure $K \in \mathscr{F}$ has finite tell-tale sets with respect to $\mathscr{F}$, we say that $\mathscr{F}$ has finite tell-tale sets.

Theorem 4.4 *Let* $\mathscr{F}$ *be a subset of* $\kappa(\mathscr{H})$*. Then* $\mathscr{F}$ *is* **FIGEX-TXT**-*learnable if and only if there is a procedure that, for every figure* $K \in \mathscr{F}$*, enumerates a finite tell-tale set* W *of* K *with respect to* $\mathscr{F}$*.*

This theorem can be proved in exactly the same way as that for learning of languages given by Angluin (1980). Note that such procedure does not need to stop. Using this theorem, we show that the set $\kappa(\mathscr{H})$ is not **FIGEX-TXT**-learnable.

Theorem 4.5 *The set* $\kappa(\mathscr{H})$ *does not have finite tell-tale sets.*

Proof Fix a figure $K = \kappa(H) \in \kappa(\mathscr{H})$, where there exists a pair $v, w \in H$ such that $\rho(vvv\ldots) \neq \rho(www\ldots)$, and fix a finite set $T = \{w_1, w_2, \ldots, w_n\}$ contained in $\mathrm{Pos}(K)$. Suppose that $\#\mathrm{Pos}_m(K) > n$ holds for a natural number m. For each finite sequence w_i, there exists $u_i \in \mathrm{Pos}(K)$ such that $|u_i| > m$, $w_i \sqsubset u_i$, and $u_i \in H^k$ for some k. For the figure $L = \kappa(U)$ with $U = \{u_1, u_2, \ldots, u_n\}$, $T \subset \mathrm{Pos}(L)$ and $\mathrm{Pos}(L) \subset \mathrm{Pos}(K)$ hold. Therefore K has no finite tell-tale set with respect to $\kappa(\mathscr{H})$. □

Corollary 4.6 *The set of figures* $\kappa(\mathscr{H})$ *is not* **FIGEX-TXT**-*learnable.*

In any realistic scenarios of machine learning, however, this set $\kappa(\mathscr{H})$ is too large to search for the best hypothesis since we usually want to obtain a "compact" representation of a target figure. Thus we (implicitly) have an upper bound on the number of elements in a hypothesis. Here we give a positive result for the above situation, that is, if we fix the number of elements $\#H$ in each hypothesis H *a priori*, the resulting set of figures becomes **FigEx-Txt**-learnable. Intuitively, this is because if we take k large enough, the set $\{w \in \mathrm{Pos}(K) \mid |w| \leqslant k\}$ becomes a finite tell-tale set of K. Here we denote by $\mathrm{Red}(H)$ the hypothesis in which for every pair $v, w \in H$ with $|v| \leqslant |w|$, w is removed if $\rho(vvv\dots) = \rho(www\dots)$. For a finite subset of natural numbers $N \subset \mathbb{N}$, we define the set of hypotheses $\mathscr{H}_N := \{H \in \mathscr{H} \mid \#\mathrm{Red}(H) \in N\}$.

Theorem 4.7 *There exists a procedure that, for all finite subsets $N \subset \mathbb{N}$ and all figures $K \in \kappa(\mathscr{H}_N)$, enumerates a finite tell-tale set of K with respect to $\kappa(\mathscr{H}_N)$.*

Proof First, we assume that $N = \{1\}$. It is trivial that there exists a procedure that, for an arbitrary figure $K \in \kappa(\mathscr{H}_N)$, enumerates a finite tell-tale set of K with respect to $\kappa(\mathscr{H}_N)$, since we always have $L \not\subset K$ for all pairs of figures $K, L \in \kappa(\mathscr{H}_N)$.

Next, fix $N \subset \mathbb{N}$ with $N \neq \{1\}$. Let us consider the procedure that enumerates elements of the sets

$$\mathrm{Pos}_1(K), \mathrm{Pos}_2(K), \mathrm{Pos}_3(K), \dots.$$

We show that this procedure enumerates a finite tell-tale set of K with respect to $\kappa(\mathscr{H}_N)$. It is enough to show that there exists a natural number m, where there is no hypothesis H such that $\kappa(H) \subset K$, $\#H \leqslant \max N$, and $\mathrm{Pos}(\kappa(H)) \supset \mathrm{Pos}_m(K)$.

We construct a tree as follows (the similar technique called d-explorer was used by Jain and Sharma (1997)). Each node has a pair (H, w) as its label, where $\kappa(H) \subset K$ and $w \in \mathrm{Pos}(K) \setminus \mathrm{Pos}(\kappa(H))$. The root node is labeled $(\emptyset, v)$ with a finite sequence $v \in \mathrm{Pos}(K)$. The tree is constructed iteratively by adding children for each node of the tree, whose depth (the length to the root) is at most $\max N - 1$. Let the label of such a node be (H, w). For every finite sequence w' with $|w'| \leqslant |w|$, if there exists a finite sequence w'' satisfying $|w''| > |w|$ and $w'' \in \mathrm{Pos}(K) \setminus \kappa(H \cup \{w'\})$, add a child labeled $(H \cup \{w'\}, w'')$ to the node.

The above tree is bounded in depth $\max N$ and the number of children for any node is always finite, hence the number of nodes of the tree is finite. Let m be the length of the longest w such that (H, w) is the label of a node of the tree. Then, we can easily check that there is no hypothesis H' such that $\kappa(H') \subset K$, $\#H' \leqslant \max N$, and $\mathrm{Pos}(\kappa(H')) \supset \mathrm{Pos}_m(K)$. □

Corollary 4.8 *For all finite subsets of natural numbers $N \subset \mathbb{N}$, the set of figures $\kappa(\mathscr{H}_N)$ is* **FigEx-Txt**-*learnable.*

4.2 Consistent Learning

In a learning process, it is natural that every hypothesis generated by a learner is consistent with the examples received by it so far. Here we introduce **FigCons-Inf**- and **FigCons-Txt**-learning (CONS means CONSistent). These criteria correspond to **Cons**-learning that was first introduced by Blum and Blum (1975)[4]. This model was also used (but implicitly) in the Model Inference System (MIS) proposed by Shapiro (1981, 1983), and studied in the computational learning of formal languages and recursive functions (Jain et al 1999).

Definition 4.9 (Consistent learning) A learner $\mathbf{M}$ **FigCons-Inf**-*learns* (resp. **FigCons-Txt**-*learns*) a set of figures $\mathscr{F} \subseteq \mathscr{K}^*$ if $\mathbf{M}$ **FigEx-Inf**-learns (resp. **FigEx-Txt**-learns) $\mathscr{F}$ and for all figures $K \in \mathscr{F}$ and all informants (resp. texts) σ of K, each hypothesis $\mathbf{M}_\sigma(i)$ is consistent with E_i that is the set of examples received by $\mathbf{M}$ until just before it generates the hypothesis $\mathbf{M}_\sigma(i)$.

Assume that a learner $\mathbf{M}$ achieves **FigEx-Inf**-learning of $\kappa(\mathscr{H})$ using Procedure 1. We can easily check that $\mathbf{M}$ always generates a hypothesis that is consistent with the received examples.

[4] Consistency was also studied in the same form by Barzdin (1974).

Corollary 4.10 **FIGEX-INF** = **FIGCONS-INF**.

Suppose that $\mathscr{F} \subset \kappa(\mathscr{H})$ is **FIGEX-TXT**-learnable. We can construct a learner **M** in the same way as in the case of **Ex**-learning of languages from texts (Angluin 1980), where **M** always outputs a hypothesis that is consistent with received examples.

Corollary 4.11 **FIGEX-TXT** = **FIGCONS-TXT**.

4.3 Reliable and Refutable Learning

In this subsection, we consider target figures that might not be represented exactly by any hypothesis since there are infinitely many such figures, and if we have no background knowledge, there is no guarantee of the existence of an exact hypothesis. Thus in practice this approach is more convenient than the explanatory or consistent learning considered in the previous two subsections.

To realize the above case, we use two concepts, *reliability* and *refutability*. The aim of the concepts is to introduce targets which cannot be exactly represented by any hypotheses. Reliable learning was introduced by Blum and Blum (1975); Minicozzi (1976) and refutable learning by Mukouchi and Arikawa (1995); Sakurai (1991) in computational learning of languages and recursive functions, and developed by Jain et al (2001); Merkle and Stephan (2003); Mukouchi and Sato (2003). Here we introduce these concepts into the learning of figures and analyze learnability.

First, we treat reliable learning of figures. Intuitively, reliability requires that an infinite sequence of hypotheses only converges to a correct hypothesis.

Definition 4.12 **(Reliable learning)** A learner **M** **FIGRELEX-INF**-*learns* (resp. **FIGRELEX-TXT**-*learns*) a set of figures $\mathscr{F} \subseteq \mathscr{K}^*$ if **M** satisfies the following conditions:

1. The learner **M** **FIGEX-INF**-learns (resp. **FIGEX-TXT**-learns) $\mathscr{F}$.
2. For any target figure $K \in \mathscr{K}^*$ and its informants (resp. texts) σ, the infinite sequence of hypotheses $\mathbf{M}_\sigma$ does not converge to a wrong hypothesis H such that $\mathrm{GE}(K, \kappa(H)) \neq 0$.

We analyze reliable learning of figures from informants. Intuitively, for any target figure $K \in \mathscr{F}$, if a learner can judge whether or not the current hypothesis H is consistent with the target, *i.e.*, $\kappa(H) = K$ or not in finite time, then the set $\mathscr{F}$ is reliably learnable.

Theorem 4.13 **FIGEX-INF** = **FIGRELEX-INF**.

Proof The statement **FIGRELEX-INF** $\subseteq$ **FIGEX-INF** is trivial, thus we prove **FIGEX-INF** $\subseteq$ **FIGRELEX-INF**. Fix a set of figures $\mathscr{F} \subseteq \kappa(\mathscr{H})$ with $\mathscr{F} \in$ **FIGEX-INF**, and suppose that a learner **M** **FIGEX-INF**-learns $\mathscr{F}$ using Procedure 1. The goal is to show that $\mathscr{F} \in$ **FIGRELEX-INF**. Assume that a target figure K belongs to $\mathscr{K}^* \setminus \mathscr{F}$. Here we have the following property: for all figures $L \in \mathscr{F}$, there must exist a finite sequences $w \in (\Sigma^d)^*$ such that

$$w \in \mathrm{Pos}(K) \ominus \mathrm{Pos}(L),$$

hence for any **M**'s current hypothesis H, **M** changes H if it receives a positive or negative example (w, l) such that $w \in \mathrm{Pos}(K) \ominus \mathrm{Pos}(\kappa(H))$. This means that an infinite sequence of hypotheses does not converge to any hypothesis. Thus we have $\mathscr{F} \in$ **FIGRELEX-INF**. □

In contrast, we have an interesting result on reliable learning from texts. We show in the following that **FIGEX-TXT** $\neq$ **FIGRELEX-TXT** holds and that a set of figures $\mathscr{F}$ is reliably learnable from positive data only if any figure $K \in \mathscr{F}$ is a singleton. Remember that $\mathscr{H}_N$ denotes the set of hypotheses $\{H \in \mathscr{H} \mid \#H \in N\}$ for a subset $N \subset \mathbb{N}$ and, for simplicity, we denote $\mathscr{H}_{\{n\}}$ by $\mathscr{H}_n$ for a natural number $n \in \mathbb{N}$.

Theorem 4.14 *The set of figures* $\kappa(\mathscr{H}_N)$ *is* **FIGRELEX-TXT**-*learnable if and only if* $N = \{1\}$.

Proof First we show that the set of figures $\kappa(\mathscr{H}_1)$ is **FigRelEx-Txt**-learnable. From the self-similar sets property of hypotheses, we have the following: A figure $K \in \kappa(\mathscr{H})$ is a singleton if and only if $K \in \kappa(\mathscr{H}_1)$. Let $K \in \mathscr{K}^* \setminus \kappa(\mathscr{H}_1)$, and assume that a learner $\mathbf{M}$ **FigEx-Txt**-learns $\kappa(\mathscr{H}_1)$. We can naturally suppose that $\mathbf{M}$ changes the current hypothesis H whenever it receives a positive example $(w,1)$ such that $w \notin \mathrm{Pos}(\kappa(H))$ without loss of generality. For any hypothesis $H \in \mathscr{H}_1$, there exists $w \in (\Sigma^d)^*$ such that

$$w \in \mathrm{Pos}(K) \setminus \mathrm{Pos}(\kappa(H)).$$

Thus if the learner $\mathbf{M}$ receives such a positive example $(w,1)$, it changes the hypothesis H. This means that an infinite sequence of hypotheses does not converge to any hypothesis. Therefore $\kappa(\mathscr{H}_1)$ is **FigRelEx-Txt**-learnable.

Next, we prove that $\kappa(\mathscr{H}_n)$ is not **FigRelEx-Txt**-learnable for any $n > 1$. Fix such $n \in \mathbb{N}$ with $n > 1$. We can easily check that, for a figure $K \in \kappa(\mathscr{H}_n)$ and any of its finite tell-tale sets $\mathscr{T}$ with respect to $\kappa(\mathscr{H}_n)$, there exists a figure $L \in \mathscr{K}^* \setminus \kappa(\mathscr{H}_n)$ such that $L \subset K$ and $\mathscr{T} \subset \mathrm{Pos}(L)$. This means that

$$\mathrm{Pos}(L) \subseteq \mathrm{Pos}(K) \text{ and } \mathscr{T} \subseteq \mathrm{Pos}(L)$$

hold. Thus if a learner $\mathbf{M}$ **FigEx-Txt**-learns $\kappa(\mathscr{H}_n)$, $\mathbf{M}_\sigma$ for some presentation σ of such L must converge to some hypothesis in $\mathscr{H}_n$. Consequently, we have $\kappa(\mathscr{H}_n) \notin$ **FigRelEx-Txt**. □

Corollary 4.15 **FigRelEx-Txt** $\subset$ **FigEx-Txt**.

Sakurai (1991) proved that a set of concepts $\mathscr{C}$ is reliably **Ex**-learnable from texts if and only if $\mathscr{C}$ contains no infinite concept (p. 182, Theorem 3.1)[5]. However, we have shown that the set $\kappa(\mathscr{H}_1)$ is **FigRelEx-Txt**-learnable, though all figures $K \in \kappa(\mathscr{H}_1)$ correspond to infinite concepts since $\mathrm{Pos}(K)$ is infinite for all $K \in \kappa(\mathscr{H}_1)$. The monotonicity of the set $\mathrm{Pos}(K)$ (Lemma 3.5), which is a constraint naturally derived from the geometric property of examples, causes this difference.

Next, we extend **FigEx-Inf**- and **FigEx-Txt**-learning by paying our attention to *refutability*. In refutable learning, a learner tries to learn figures in the limit, but it understands that it cannot find a correct hypothesis in finite time, that is, outputs the refutation symbol $\triangle$ and stops if the target figure is not in the considered space.

Definition 4.16 (Refutable learning) A learner $\mathbf{M}$ **FigRefEx-Inf**-*learns* (resp. **FigRefEx-Txt**-*learns*) a set of figures $\mathscr{F} \subseteq \mathscr{K}^*$ if $\mathbf{M}$ satisfies the following conditions. Here, $\triangle$ is the *refutation symbol*.

1. The learner $\mathbf{M}$ **FigEx-Inf**-learns (resp. **FigEx-Txt**-learns) $\mathscr{F}$.
2. If $K \in \mathscr{F}$, then for all informants (resp. texts) σ of K, $\mathbf{M}_\sigma(i) \neq \triangle$ for all $i \in \mathbb{N}$.
3. If $K \in \mathscr{K}^* \setminus \mathscr{F}$, then for all informants (resp. texts) σ of K, there exists $m \in \mathbb{N}$ such that $\mathbf{M}_\sigma(i) \neq \triangle$ for all $i < m$, and $\mathbf{M}_\sigma(i) = \triangle$ for all $i \geqslant m$.

Conditions 2 and 3 in the above definition mean that a learner $\mathbf{M}$ refutes the set $\mathscr{F}$ in finite time if and only if a target figure $K \in \mathscr{K}^* \setminus \mathscr{F}$. We compare **FigRefEx-Inf**-learnability with other learning criteria.

Theorem 4.17 **FigRefEx-Inf** $\not\subseteq$ **FigEx-Txt** *and* **FigEx-Txt** $\not\subseteq$ **FigRefEx-Inf**.

Proof First we consider **FigRefEx-Inf** $\not\subseteq$ **FigEx-Txt**. We show an example of a set of figures $\mathscr{F}$ with $\mathscr{F} \in$ **FigRefEx-Inf** and $\mathscr{F} \notin$ **FigEx-Txt** in the case of $d = 2$. Let $K_0 = \kappa(\{\langle 0,0\rangle, \langle 1,1\rangle\})$, $K_i = \kappa(\{\langle w,w\rangle \mid w \in \Sigma^i \setminus \{1\}^i\})$ for every $i \geqslant 1$, and $\mathscr{F} = \{K_i \mid i \in \mathbb{N}\}$. Note that K_0 is the line $y = x$ and $K_i \subset K_0$ for all $i \geqslant 1$.

We prove that $\mathscr{F} \in$ **FigRefEx-Inf**. It is trivial that $\mathscr{F} \in$ **FigEx-Inf**, thereby assume that a target figure $K \in \mathscr{K}^* \setminus \mathscr{F}$. If a target figure $K \supset K_0$, it is trivial that, for any informant σ of K, the set of examples $\mathrm{range}(\sigma[n])$ for some $n \in \mathbb{N}$ is not consistent with any $K_i \in \mathscr{F}$ (consider a positive example for a point $x \in K \setminus K_0$). Otherwise if $K \subset K_0$, there should exist a negative example $\langle v,v\rangle \in \mathrm{Neg}(K)$. Then we have $K \neq K_i$ for all $i > |v|$. Thus a learner can refute candidates $\{K_1, K_2, \ldots, K_{|v|}\}$ in finite time. Therefore $\mathscr{F} \in$ **FigRefEx-Inf** holds.

[5] The article (Sakurai 1991) is written in Japanese. The same theorem is mentioned by Mukouchi and Arikawa (1995, p. 60, Theorem 3).

Next we show that $\mathscr{F} \notin$ **FigEx-Txt**. Let K_0 be the target figure. For any finite set of positive examples $\mathscr{T} \subset \mathrm{Pos}(K_0)$, there exists a figure $K_i \in \mathscr{F}$ such that $K_i \subset K_0$ and $\mathscr{T}$ is consistent with K_i. Therefore it has no finite tell-tale set with respect to $\mathscr{F}$ and hence $\mathscr{F} \notin$ **FigEx-Txt** from Theorem 4.4.

Second we check **FigEx-Txt** $\not\subseteq$ **FigRefEx-Inf**. Assume that $\mathscr{F} = \kappa(\mathscr{H}_{\{1\}})$ and a target figure K is a singleton $\{x\}$ with $K \notin \mathscr{F}$. It is clear that, for any informant σ of K and $n \in \mathbb{N}$, $\mathrm{range}(\sigma[n])$ is consistent with some figure $L \in \mathscr{F}$. Thus $\mathscr{F} \notin$ **FigRefEx-Inf** whereas $\mathscr{F} \in$ **FigEx-Txt**. □

Corollary 4.18 **FigRelEx-Txt** $\not\subseteq$ **FigRefEx-Inf** *and* **FigRefEx-Inf** $\not\subseteq$ **FigRelEx-Txt**.

Note that it is trivial that **FigRelEx-Txt** $\not\subseteq$ **FigRefEx-Inf** since we have $\kappa(\mathscr{H}_{\{1\}}) \notin$ **FigRefEx-Inf** in the above proof and $\kappa(\mathscr{H}_{\{1\}}) \in$ **FigRelEx-Txt** from Theorem 4.14. Moreover, the condition **FigRefEx-Inf** $\not\subseteq$ **FigRelEx-Txt** holds since **FigRefEx-Inf** $\not\subseteq$ **FigEx-Txt** and **FigRelEx-Txt** $\subset$ **FigEx-Txt**. These results mean that both **FigRefEx-Inf**- and **FigRelEx-Txt**-learning are difficult, but they are incomparable in terms of learnability. Furthermore, we have the following hierarchy.

Theorem 4.19 **FigRefEx-Txt** $\neq \emptyset$ *and* **FigRefEx-Txt** $\subset$ **FigRefEx-Inf**.

Proof Let a set of figures $\mathscr{F}$ be a singleton $\{K\}$ such that $K = \kappa(w)$ for some $w \in (\Sigma^d)^*$. Then there exists a learner $\mathbf{M}$ that **FigRefEx-Txt**-learns $\mathscr{F}$, *i.e.*, $\mathscr{F} \in$ **FigRefEx-Txt**, since all $\mathbf{M}$ has to do is to check whether or not, for a given positive example $(v,1)$, $v \sqsubseteq u$ for some $u \in \mathrm{Pos}(K) = \{x \mid x \sqsubseteq www\dots\}$.

Next, let $\mathscr{F} = \{K\}$ such that $K = \kappa(H)$ with $\#\mathrm{Red}(H) \geqslant 2$. We can easily check that $\mathscr{F} \notin$ **FigRefEx-Txt** because if a target figure L is a proper subset of K, no learner can refute $\mathscr{F}$ in finite time. Conversely, $\mathscr{F} \in$ **FigRefEx-Inf** since for all L with $L \neq K$, there exists an example with which the hypothesis H is not consistent. □

Corollary 4.20 **FigRefEx-Txt** $\subset$ **FigRelEx-Txt**.

5 Effective Learning of Figures

In learning under the proposed criteria, *i.e.* explanatory, consistent, reliable, and refutable learning, each hypothesis is just considered as exactly "correct" or not, that is, for a target figure K and for a hypothesis H, H is correct if $\mathrm{GE}(K,H) = 0$ and is not correct if $\mathrm{GE}(K,H) \neq 0$. Thus we cannot know the rate of convergence to the target figure and how far it is from the recent hypothesis to the target. It is therefore more useful if we consider *approximate* hypotheses by taking various *generalization errors* into account in the learning process.

We define novel learning criteria, **FigEfEx-Inf**- and **FigEfEx-Txt**-learning (EF means EFfective), to introduce into learning the concept of *effectivity*, which has been analyzed in computation of real numbers in the area of computable analysis (Weihrauch 2000). Intuitively, these criteria guarantee that for any target figure, a generalization error becomes smaller and smaller monotonically and converges to zero. Thus we can know when the learner learns the target figure "well enough". Furthermore, if a target figure is learnable in the limit, then the generalization error goes to zero in finite time.

Definition 5.1 (Effective learning) A learner $\mathbf{M}$ **FigEfEx-Inf**-*learns* (resp. **FigEfEx-Txt**-*learns*) a set of figures $\mathscr{F} \subseteq \mathscr{K}^*$ if $\mathbf{M}$ satisfies the following conditions:

1. The learner $\mathbf{M}$ **FigEx-Inf**-learns (resp. **FigEx-Txt**-learns) $\mathscr{F}$.
2. For an arbitrary target figure $K \in \mathscr{K}^*$ and all informants (resp. texts) σ of K, for all $i \in \mathbb{N}$,

$$\mathrm{GE}(K, \mathbf{M}_\sigma(i)) \leqslant 2^{-i}.$$

This definition is inspired by the *Cauchy representation* of real numbers (Weihrauch 2000, Definition 4.1.5).

Effective learning is related to *monotonic* learning (Lange and Zeugmann 1993, 1994; Kinber 1994; Zeugmann et al 1995) originally introduced by Jantke (1991); Wiehagen (1991), since both learning models consider monotonic convergence of hypotheses. In contrast to their approach, where various monotonicity over

languages was considered, we geometrically measure the generalization error of a hypothesis by the Hausdorff metric. On the other hand, the effective learning is different from **BC**-learning developed in the learning of languages and recursive functions (Jain et al 1999) since **BC**-learning only guarantees that generalization errors go to zero in finite time. This means that **BC**-learning is *not* effective.

First we show that we can bound the generalization error of the hypothesis H using the diameter $\mathrm{diam}(k)$ of the set $\rho(w)$ with $|w| = k$. Recall that we have

$$\mathrm{diam}(k) = \sqrt{d} \cdot 2^{-k}$$

(see Proof of Lemma 3.3). In the following, we denote the set of examples $\{(w,l) \mid |w| = k\}$ in σ by E^k and call each example in it a *level-k example*.

Lemma 5.2 *Let σ be an informant of a figure K and H be a hypothesis that is consistent with the set of examples $E^k = \{(w,l) \mid |w| = k\}$. We have the inequality*

$$\mathrm{GE}(K,H) \leqslant \mathrm{diam}(k).$$

Proof Since H is consistent with E^k,

$$\kappa(H) \cap \rho(w) \begin{cases} \neq \emptyset & \text{if } (w,1) \in E^k, \\ = \emptyset & \text{if } (w,0) \in E^k. \end{cases}$$

Thus for $\delta = \mathrm{diam}(k)$, the δ-neighborhood of $\kappa(H)$ contains K and the δ-neighborhood of K contains $\kappa(H)$. It therefore follows that $\mathrm{GE}(K,H) = d_{\mathrm{H}}(K, \kappa(H)) \leqslant \mathrm{diam}(k)$. □

Theorem 5.3 *The set of figures $\kappa(\mathscr{H})$ is* **FIGEFEX-INF***-learnable.*

Proof We show the learner $\mathbf{M}$ that **FIGEFEX-INF**-learns $\kappa(\mathscr{H})$ in Procedure 2. We use the function

$$g(k) = \lceil k + \log_2 \sqrt{d} \rceil.$$

Then for all $k \in \mathbb{N}$, we have

$$\mathrm{diam}(g(k)) = \sqrt{d} \cdot 2^{-g(k)} \leqslant 2^{-k}.$$

The learner $\mathbf{M}$ stores examples, and when it receives all examples at the level $g(k)$, it outputs a hypothesis. Every kth hypothesis $\mathbf{M}_\sigma(k)$ is consistent with the set of examples $E^{g(k)}$. Thus we have

$$\mathrm{GE}(K, \mathbf{M}_\sigma(k)) \leqslant \mathrm{diam}(g(k)) \leqslant 2^{-k}$$

for all $k \in \mathbb{N}$ from Lemma 5.2.

Assume that $K \in \kappa(\mathscr{H})$. If $\mathbf{M}$ outputs a wrong hypothesis, there must be a positive or negative example that is not consistent with the hypothesis, and it changes the wrong hypothesis. If it produces a correct hypothesis, then it never changes the correct hypothesis, since every example is consistent with the hypothesis. Thus there exists $n \in \mathbb{N}$ with $\mathrm{GE}(K, \mathbf{M}_\sigma(i)) = 0$ for all $i \geqslant n$. Therefore $\mathbf{M}$ **FIGEFEX-INF**-learns $\kappa(\mathscr{H})$. □

Corollary 5.4 **FIGEFEX-INF** = **FIGRELEX-INF** = **FIGEX-INF**.

Thus the learner with Procedure 2 can treat the set of *all* figures $\mathscr{K}^*$ as learning targets, since for any figure $K \in \mathscr{K}^*$, it can approximate the figure arbitrarily closely using only the figures represented by hypotheses in the hypothesis space $\mathscr{H}$.

In contrast to **FIGEX-TXT**-learning, there is no set of figures that is **FIGEFEX-TXT**-learnable.

Theorem 5.5 **FIGEFEX-TXT** = $\emptyset$.

Procedure 2: Learning procedure that **FIGEFEX-INF**-learns $\kappa(\mathscr{H})$

Input: Informant $\sigma = (w_0, l_0), (w_1, l_1), \ldots$ of figure $K \in \kappa(\mathscr{H})$
Output: Infinite sequence of hypotheses $\mathbf{M}_\sigma(0), \mathbf{M}_\sigma(1), \ldots$

```
1:  i ← 0
2:  k ← 0
3:  E ← ∅    // E is a set of received examples
4:  repeat
5:    read σ(i) and add to E    // σ(i) = (w_i, l_i)
6:    if E^{g(k)} ⊆ E then    // E^{g(k)} = { (w,l) ∈ range(σ) | |w| = g(k) } and g(k) = ⌈k + log_2 √d⌉
7:      search the first H that is consistent with E through a normal enumeration
8:      output P    // M_σ(i) = H
9:      k ← k + 1
10:   end if
11:   i ← i + 1
12: until forever
```

Proof We show a counterexample of a target figure which no learner $\mathbf{M}$ can approximate effectively. Assume that $d = 2$ and a learner $\mathbf{M}$ **FIGEFEX-TXT**-learns a set of figures $\mathscr{F} \subseteq \mathscr{K}^*$. Let us consider two target figures $K = \{(0,0),(1,1)\}$ and $L = \{(0,0)\}$. For a text σ of L, for all examples $(w,1) \in \mathrm{range}(\sigma)$, $w \in \{\mathtt{00}\}^*$. Since $\mathbf{M}$ **FIGEFEX-TXT**-learns $\mathscr{F}$, it should output the hypothesis H as $\mathbf{M}_\sigma(2)$ such that $\mathrm{GE}(L,H) < 1/4$. Suppose that $\mathbf{M}$ receives n examples before outputting the hypothesis H. Then there exists a presentation τ of the figure K such that $\tau[n-1] = \sigma[n-1]$, and $\mathbf{M}$ outputs the hypothesis H with receiving $\tau[n-1]$. However, $\mathrm{GE}(K,H) \geqslant \sqrt{2} - 1/4$ holds from the triangle inequality, contradicting our assumption that $\mathbf{M}$ **FIGEFEX-TXT**-learns $\mathscr{F}$. This proof can be applied for any $\mathscr{F} \subseteq \mathscr{K}^*$, thereby we have **FIGEFEX-TXT** $= \emptyset$. □

Since **FIGREFEX-TXT** $\neq \emptyset$, we have the relation

$$\textbf{FIGEFEX-TXT} \subset \textbf{FIGREFEX-TXT}.$$

This result means that we cannot learn any figures "effectively" by using only positive examples.

6 Evaluation of Learning Using Dimensions

Here we show a novel mathematical connection between fractal geometry and Gold's learning under the proposed learning model described in Section 3. More precisely, we bound the number of positive examples, one of the complexities of learning, using the Hausdorff dimension and the VC dimension. The Hausdorff dimension is known as the central concept of fractal geometry, which measures the density of figures, and VC dimension is the central concept of Valiant's model (PAC learning model) (Kearns and Vazirani 1994), which measures the complexity of classes of hypotheses.

6.1 Preliminaries for Dimensions

First we introduce the Hausdorff dimension and related dimensions: the box-counting dimension, the similarity dimension, and also introduce the VC dimension.

For $X \subseteq \mathbb{R}^n$ and $s \in \mathbb{R}$ with $s > 0$, define

$$\mathfrak{H}^s_\delta(X) := \inf \left\{ \sum_{U \in \mathscr{U}} |U|^s \;\middle|\; \mathscr{U} \text{ is a } \delta\text{-cover of } X \right\}.$$

The *s-dimensional Hausdorff measure* of X is $\lim_{\delta\to 0}\mathfrak{H}^s_\delta(X)$, denoted by $\mathfrak{H}^s(X)$. We say that $\mathscr{U}$ is a δ-cover of X if $\mathscr{U}$ is countable, $X\subseteq\bigcup_{U\in\mathscr{U}}U$, and $|U|\leqslant\delta$ for all $U\in\mathscr{U}$. When we fix a set X and view $\mathfrak{H}^s(X)$ as a function with respect to s, it has at most one value where the value $\mathfrak{H}^s(X)$ changes from ∞ to 0 (Federer 1996). This value is called the *Hausdorff dimension* of X. Formally, the Hausdorff dimension of a set X, written as $\dim_{\mathrm{H}}X$, is defined by

$$\dim_{\mathrm{H}}X:=\sup\{\,s\mid\mathfrak{H}^s(X)=\infty\,\}=\inf\{\,s\geqslant 0\mid\mathfrak{H}^s(X)=0\,\}.$$

The box-counting dimension, also known as the Minkowski-Bouligand dimension, is one of the most widely used dimensions since its mathematical calculation and empirical estimation are relatively easy compared to the Hausdorff dimension. Moreover, if we try to calculate the box-counting dimension, which is given as the limit of the following equation (8) by decreasing δ, the values obtained often converge to the Hausdorff dimension at the same time. Thus we can obtain an approximate value of the Hausdorff dimension by an empirical method. Let X be a nonempty bounded subset of $\mathbb{R}^n$ and $N_\delta(X)$ be the smallest cardinality of a δ-cover of X. The *box-counting dimension* $\dim_{\mathrm{B}}X$ of X is defined by

$$\dim_{\mathrm{B}}X:=\lim_{\delta\to 0}\frac{\log N_\delta(X)}{-\log\delta}\tag{8}$$

if this limit exists. Falconer (2003, Equivalent definitions 3.1, P.43) also shows that we have the equivalent box-counting dimension $\dim_{\mathrm{B}}X$ if $N_\delta(K)$ is the smallest number of cubes of side δ that cover K, or the number of δ-mesh cubes that intersect K. We have

$$\dim_{\mathrm{H}}X\leqslant\dim_{\mathrm{B}}X$$

for all $X\subseteq\mathbb{R}^n$.

It is usually difficult to find the Hausdorff dimension of a given set. However, we can obtain the dimension of a certain class of self-similar sets in the following manner. Let C be a finite set of contractions, and F be the self-similar set of C. The *similarity dimension* of F, denoted by $\dim_{\mathrm{S}}F$, is defined by the equation

$$\sum_{\varphi\in C}L(\varphi)^{\dim_{\mathrm{S}}F}=1,$$

where $L(\varphi)$ is the *contractivity factor* of φ, which is defined by the infimum of all real numbers c with $0<c<1$ such that $d(\varphi(x),\varphi(y))\leqslant cd(x,y)$ for all $x,y\in X$. We have

$$\dim_{\mathrm{H}}F\leqslant\dim_{\mathrm{S}}F$$

and if C satisfies the open set condition,

$$\dim_{\mathrm{H}}F=\dim_{\mathrm{B}}F=\dim_{\mathrm{S}}F$$

(Falconer 2003). Here, a finite set of contractions C satisfies the *open set condition* if there exists a nonempty bounded open set $O\subset\mathbb{R}^n$ such that $\varphi(O)\subset O$ for all $\varphi\in C$ and $\varphi(O)\cap\varphi'(O)=\emptyset$ for all $\varphi,\varphi'\in C$ with $\varphi\neq\varphi'$.

Intuitively, the Vapnik-Chervonenkis (VC) dimension (Blumer et al 1989; Valiant 1984; Vapnik and Chervonenkis 1971) is a parameter of separability and it gives lower and upper bounds for the sample size in Valiant's (PAC) learning model (Kearns and Vazirani 1994). For all $\mathscr{R}\subseteq\mathscr{H}$ and $W\subseteq\Sigma^*$, define

$$\Pi_{\mathscr{R}}(W):=\{\,\mathrm{Pos}(\kappa(H))\cap W\mid H\in\mathscr{R}\,\}.$$

If $\Pi_{\mathscr{R}}(W)=2^W$, we say that W is *shattered* by $\mathscr{R}$. Here the *VC dimension* of $\mathscr{R}$, denoted by $\dim_{\mathrm{VC}}\mathscr{R}$, is the cardinality of the largest set W shattered by $\mathscr{R}$.

6.2 Measuring the Complexity of Learning with Dimensions

We show that the Hausdorff dimension of a target figure gives a lower bound to the number of positive examples. Remember that $\mathrm{Pos}_k(K) = \{w \in \mathrm{Pos}(K) \mid |w| = k\}$ and the diameter $\mathrm{diam}(k)$ of the set $\rho(w)$ with $|w| = k$ is $\sqrt{d}2^{-k}$. Moreover, the size $\#\{w \in (\Sigma^d)^* \mid |w| = k\} = 2^{kd}$ for all $k \in \mathbb{N}$.

Theorem 6.1 *For every figure $K \in \mathscr{K}^*$ and for any $s < \dim_{\mathrm{H}} K$, if we take k large enough,*

$$\#\mathrm{Pos}_k(K) \geqslant 2^{ks}.$$

Proof Fix $s < \dim_{\mathrm{H}} K$. From the definition of the Hausdorff measure,

$$\mathfrak{H}^s_{\mathrm{diam}(k)}(K) \leqslant \#\mathrm{Pos}_k(K) \cdot (\sqrt{d}2^{-k})^s$$

since $\mathrm{diam}(k) = \sqrt{d}2^{-k}$. If we take k large enough,

$$\mathfrak{H}^s_{\mathrm{diam}(k)}(K) \geqslant (\sqrt{d})^s$$

because $\mathfrak{H}^s_\delta(K)$ is monotonically increasing with decreasing δ, and goes to ∞. Thus

$$\#\mathrm{Pos}_k(K) \geqslant \mathfrak{H}^s_{\mathrm{diam}(k)}(K)(\sqrt{d}2^{-k})^{-s} \geqslant (\sqrt{d})^s(\sqrt{d}2^{-k})^{-s} = 2^{ks}. \qquad \square$$

Moreover, if a target figure K can be represented by some hypothesis, that is, $K \in \kappa(\mathscr{H})$, we can use the exact dimension $\dim_{\mathrm{H}} K$ as a bound for the number of positive examples $\#\mathrm{Pos}_k(K)$.

Theorem 6.2 *For every figure $K \in \kappa(\mathscr{H})$, if we take k large enough,*

$$\#\mathrm{Pos}_k(K) \geqslant 2^{k \dim_{\mathrm{H}} K}.$$

Proof Since the set of contractions encoded by a hypothesis H meets the open set condition, $\dim_{\mathrm{H}} \kappa(H) = \dim_{\mathrm{B}} \kappa(H) = \dim_{\mathrm{S}} \kappa(H)$ holds. Thus we have

$$\dim_{\mathrm{H}} K = \dim_{\mathrm{B}} K = \lim_{\delta \to 0} \frac{\log N_\delta(X)}{-\log \delta} \leqslant \lim_{k \to \infty} \frac{\log \#\mathrm{Pos}_k(K)}{-\log 2^{-k}},$$

where 2^{-k} is the length of one side of an interval $\rho(w)$ with $|w| = k$. The above inequality is trivial from the definition of the box-counting dimension since $N_\delta(X) \leqslant \#\mathrm{Pos}_k(K)$. Therefore if we take k large enough,

$$\log \#\mathrm{Pos}_k(K) \geqslant \dim_{\mathrm{H}} K \cdot -\log 2^{-k},$$
$$\#\mathrm{Pos}_k(K) \geqslant 2^{k \dim_{\mathrm{H}} K}. \qquad \square$$

Example 6.3 Let us consider the figure K in Example 3.1. It is known that $\dim_{\mathrm{H}} K = \log 3 / \log 2 = 1.584\ldots$. From Theorem 6.2,

$$\mathrm{Pos}_1(K) \geqslant 2^{\dim_{\mathrm{H}} K} = 3$$

holds at level 1 and

$$\mathrm{Pos}_2(K) \geqslant 4^{\dim_{\mathrm{H}} K} = 9$$

holds at level 2. Actually, $\mathrm{Pos}_1(K) = 4$ and $\mathrm{Pos}_2(K) = 13$. Note that K is already covered by 3 and 9 intervals at level 1 and 2, respectively (Fig. 4).

The VC dimension can also be used to characterize the number of positive examples. Define

$$\mathscr{H}^k := \{\, H \in \mathscr{H} \mid |w| = k \text{ for all } w \in H \,\},$$

and call each hypothesis in the set a *level-k hypothesis*. We show that the VC dimension of the set of level k hypotheses $\mathscr{H}^k$ is equal to $\#\{\, w \in (\Sigma^d)^* \mid |w| = k \,\} = 2^{kd}$.

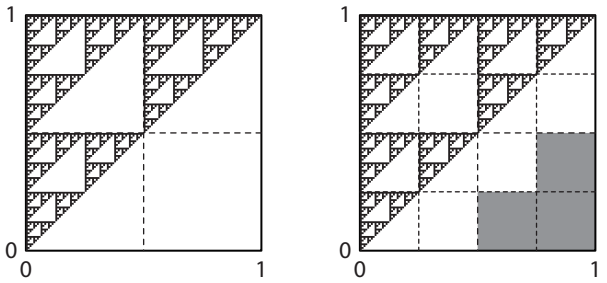

Fig. 4 Positive and negative examples for the Sierpiński triangle at level 1 and 2. White (resp. gray) squares mean positive (resp. negative) examples.

Lemma 6.4 *At each level k, we have* $\dim_{\mathrm{VC}} \mathscr{H}^k = 2^{kd}$.

Proof First of all,

$$\dim_{\mathrm{VC}} \mathscr{H}^k \leqslant 2^{kd}$$

is trivial since $\#\mathscr{H}^k = 2^{2^{kd}}$. Let $\mathscr{H}^k_n$ denote the set $\{H \in \mathscr{H}^k \mid \#\mathrm{Red}(H) = n\}$. For all $H \in \mathscr{H}^k_1$, there exists $w \in \mathrm{Pos}(\kappa(H))$ such that $w \notin \mathrm{Pos}(\kappa(G))$ for all $G \in \mathscr{H}^k_1$ with $H \neq G$. Thus if we assume $\mathscr{H}^k_1 = \{H_1, \ldots, H_{2^{kd}}\}$, there exists the set of finite sequences $W = \{w_1, \ldots, w_{2^{kd}}\}$ such that for all $i \in \{1, \ldots, 2^{kd}\}$, $w_i \in \mathrm{Pos}(\kappa(H_i))$ and $w_i \notin \mathrm{Pos}(\kappa(H_j))$ for all $j \in \{1, \ldots, 2^{kd}\}$ with $i \neq j$. For every pair $V, W \subset (\Sigma^d)^*$, $V \subset W$ implies $\kappa(V) \subset \kappa(W)$. Therefore the set W is shattered by $\mathscr{H}^k$, meaning that we have $\dim_{\mathrm{VC}} \mathscr{H}^k = 2^{kd}$. □

Therefore we can rewrite Theorems 6.1 and 6.2 as follows.

Theorem 6.5 *For every figure $K \in \mathscr{K}^*$ and for any $s < \dim_{\mathrm{H}} K$, if we take k large enough,*

$$\#\mathrm{Pos}_k(K) \geqslant (\dim_{\mathrm{VC}} \mathscr{H}^k)^{s/d}.$$

Moreover, when $K \in \kappa(\mathscr{H})$, if we take k large enough,

$$\#\mathrm{Pos}_k(K) \geqslant (\dim_{\mathrm{VC}} \mathscr{H}^k)^{\dim_{\mathrm{H}} K/d}.$$

These results demonstrate a relationship among the complexities of learning figures (numbers of positive examples), classes of hypotheses (VC dimension), and target figures (Hausdorff dimension).

6.3 Learning the Box-Counting Dimension through Effective Learning

One may think that **FigEfEx-Inf**-learning can be achieved without the proposed hypothesis space. For instance, if a learner just outputs figures represented by a set of received positive examples, the generalization error becomes smaller and smaller. Here we show that one "quality" of a target figure, the box-counting dimension, is also learned in **FigEfEx-Inf**-learning, whereas if a learner outputs figures represented by a set of received positive examples, the box-counting dimension (and also the Hausdorff dimension) of any figure represented by a hypothesis is always d.

Recall that for all hypotheses $H \in \mathscr{H}$, $\dim_{\mathrm{H}} \kappa(H) = \dim_{\mathrm{B}} \kappa(H) = \dim_{\mathrm{S}} \kappa(H)$, since the set of contractions encoded by the hypothesis H meets the open set condition.

Theorem 6.6 *Assume that a learner* **M** **FigEfEx-Inf**-*learns $\kappa(\mathscr{H})$. For every target figure $K \in \mathscr{K}^*$,*

$$\lim_{k\to\infty} \dim_{\mathrm{B}} \kappa(\mathbf{M}_\sigma(k)) = \dim_{\mathrm{B}} K.$$

Proof First, we assume that a target figure $K \in \kappa(\mathscr{H})$. For every informant σ of K, $\mathbf{M}_\sigma$ converges to a hypothesis H with $\kappa(H) = K$. Thus

$$\lim_{k\to\infty} \dim_{\mathrm{B}} \kappa(\mathbf{M}_\sigma(k)) = \dim_{\mathrm{B}} K = \dim_{\mathrm{H}} K.$$

Next, we assume $K \in \mathscr{K}^* \setminus \kappa(\mathscr{H})$. Since $\mathrm{GE}(K, \mathbf{M}_\sigma(i)) \leqslant 2^{-i}$ holds for every $i \in \mathbb{N}$, for each $k \in \mathbb{N}$ we have some $i \geqslant k$ such that the hypothesis $\mathbf{M}_\sigma(i)$ is consistent with the set of level-k examples $E^k = \{(w,l) \in \mathrm{range}(\sigma) \mid |w| = k\}$. Thus

$$\dim_{\mathrm{B}} \kappa(\mathbf{M}_\sigma(i)) = \lim_{k\to\infty} \frac{\log \#\mathrm{Pos}_k(K)}{-\log 2^{-k}}.$$

Falconer (2003, Equivalent definitions 3.1, P.43) shows that the box-counting dimension $\dim_{\mathrm{B}} K$ is defined equivalently by

$$\dim_{\mathrm{B}} K = \lim_{k\to\infty} \frac{\log \#\mathrm{Pos}_k(K)}{-\log 2^{-k}}.$$

Therefore from the definition of the box-counting dimension, we have

$$\lim_{i\to\infty} \dim_{\mathrm{B}} \kappa(\mathbf{M}_\sigma(i)) = \lim_{k\to\infty} \frac{\log \#\mathrm{Pos}_k(K)}{-\log 2^{-k}} = \dim_{\mathrm{B}} K. \qquad \square$$

7 Computational Interpretation of Learning

Recently, the concept of "computability" for continuous objects has been introduced in the framework of Type-2 Theory of Effectivity (TTE) (Schröder 2002b; Weihrauch 2000, 2008; Weihrauch and Grubba 2009; Tavana and Weihrauch 2011), where we treat an uncountable set X as objects for computing through infinite sequences over a given alphabet Σ. Using the framework, we analyze our learning model from the computational point of view. Some studies by de Brecht and Yamamoto (2009); de Brecht (2010) have already demonstrated a close connection between TTE and Gold's model, and our analysis becomes an instance and extension of their analysis.

7.1 Preliminaries for Type-2 Theory of Effectivity

We prepare mathematical notations for TTE. In the following in this section, we assume $\Sigma = \{\mathtt{0}, \mathtt{1}, [,], \|, \diamondsuit\}$. A partial (resp. total) function g from a set A to a set B is denoted by $g :\subseteq A \to B$ (resp. $g : A \to B$). A *representation* of a set X is a surjection $\xi :\subseteq C \to X$, where C is Σ^* or Σ^ω. We see $p \in \mathrm{dom}(\xi)$ as a name of the encoded element $\xi(p)$.

Computability of string functions $f :\subseteq X \to Y$, where X and Y are Σ^* or Σ^ω, is defined via a *Type-2 machine*, which is a usual Turing machine with one-way input tapes, some work tapes, and a one-way output tape (Weihrauch 2000). The function $f_M :\subseteq X \to Y$ computed by a Type-2 machine M is defined as follows: When Y is Σ^*, $f_M(p) := q$ if M with input p halts with q on the output tape, and when Y is Σ^ω, $f_M(p) := q$ if M with input p writes step by step q onto the output tape. We say that a function $f :\subseteq C \to D$ is *computable* if there is a Type-2 machine that computes f, and a finite or infinite sequence p is computable if the constant function f which outputs p is computable. A Type-2 machine never changes symbols that have already been written onto the output tape, thus each prefix of the output depends only on a prefix of the input.

By treating a Type-2 machine as a translator between names of some objects, a hierarchy of representations is introduced. A representation ξ is *reducible* to ζ, denoted by $\xi \leqslant \zeta$, if there exists a computable function f such that $\xi(p) = \zeta(f(p))$ for all $p \in \mathrm{dom}(\xi)$. Two representations ξ and ζ are *equivalent*, denoted by $\xi \equiv \zeta$, if both $\xi \leqslant \zeta$ and $\zeta \leqslant \xi$ hold. As usual, $\xi < \zeta$ means $\xi \leqslant \zeta$ and not $\zeta \leqslant \xi$.

Computability for functions is defined through representations and computability of string functions.

Definition 7.1 Let ξ and ζ be representations of X and Y, respectively. An element $x \in X$ is ξ-*computable* if there is some computable p such that $\xi(p) = x$. A function $f :\subseteq X \to Y$ is (ξ,ζ)-*computable* if there is some computable function g such that

$$f \circ \xi(p) = \zeta \circ g(p)$$

for all $p \in \mathrm{dom}(\xi)$. This g is called a (ξ,ζ)-*realization* of f.

Thus the abstract function f is "realized" by the concrete function (Type-2 machine) g through the two representations ξ and ζ.

Various representations of the set of nonempty compact sets $\mathscr{K}^*$ are well studied by Brattka and Weihrauch (1999); Brattka and Presser (2003). Let

$$\mathscr{Q} = \{A \subset \mathbb{Q}^d \mid A \text{ is finite and nonempty}\}$$

and define a representation $\nu_{\mathscr{Q}} :\subseteq \Sigma^* \to \mathscr{Q}$ by

$$\nu_{\mathscr{Q}}([w_0\|w_1\|\ldots\|w_n]) := \{\nu_{\mathbb{Q}^d}(w_0),\ldots,\nu_{\mathbb{Q}^d}(w_n)\},$$

where $\nu_{\mathbb{Q}^d} :\subseteq (\Sigma^d)^* \to \mathbb{Q}^d$ is the standard binary notation of rational numbers defined by

$$\nu_{\mathbb{Q}^d}(\langle w^1, w^2, \ldots, w^d\rangle) := \left(\sum_{i=0}^{|w^1|-1} w_i^1 \cdot 2^{-(i+1)}, \sum_{i=0}^{|w^2|-1} w_i^2 \cdot 2^{-(i+1)}, \ldots, \sum_{i=0}^{|w^d|-1} w_i^d \cdot 2^{-(i+1)}\right)$$

and "[", "]", and "‖" are special symbols used to separate two finite sequences. For a finite set of finite sequences $\{w_0,\ldots,w_m\}$, for convenience we introduce the mapping ι which translates the set into a finite sequence defined by $\iota(w_0,\ldots,w_m) := [w_0\|\ldots\|w_m]$. Note that $\nu_{\mathbb{Q}^d}(\langle w^1,\ldots,w^d\rangle) = (\min\rho(w^1), \ldots, \min\rho(w^d))$ for our representation ρ introduced in equation (2). The standard representation of the topological space $(\mathscr{K}^*, d_{\mathrm{H}})$, given by Brattka and Weihrauch (1999, Definition 4.8), is defined in the following manner.

Definition 7.2 (Standard representation of figures) Define the representation $\kappa_{\mathrm{H}} :\subseteq \Sigma^\omega \to \mathscr{K}^*$ of figures by $\kappa_{\mathrm{H}}(p) = K$ if $p = w_0\diamondsuit w_1\diamondsuit w_2\diamondsuit\ldots$,

$$d_{\mathrm{H}}(K, \nu_{\mathscr{Q}}(w_i)) < 2^{-i}$$

for each $i \in \mathbb{N}$, and $\lim_{i\to\infty}\nu_{\mathscr{Q}}(w_i) = K$, where $\diamondsuit$ denotes a separator of two finite sequences.

This representation κ_{H} is known to be an *admissible representation* of the space $(\mathscr{K}^*, d_{\mathrm{H}})$, which is the key concept in TTE (Schröder 2002b; Weihrauch 2000), and is also known as the Σ^0_1-admissible representation proposed by de Brecht and Yamamoto (2009).

7.2 Computability and Learnability of Figures

First, we show computability of figures in $\kappa(\mathscr{H})$.

Theorem 7.3 *For every figure $K \in \kappa(\mathscr{H})$, K is κ_{H}-computable.*

Proof It is enough to prove that there exists a computable function f such that $\kappa(H) = \kappa_{\mathrm{H}}(f(H))$ for all $H \in \mathscr{H}$. Fix a hypothesis $H \in \mathscr{H}$ such that $\kappa(H) = K$. For all $k \in \mathbb{N}$ and for H_k defined by

$$H_k := \{w \in (\Sigma^d)^* \mid w \sqsubseteq v \text{ with } v \in H^m \text{ for some } m, \text{ and } |w| = k\},$$

we can easily check that

$$d_{\mathrm{H}}(K, \nu_{\mathscr{Q}}(\iota(H_k))) < \mathrm{diam}(k) = \sqrt{d}\cdot 2^{-k}.$$

Moreover, for each k, $\sqrt{d} \cdot 2^{-g(k)} < 2^{-k}$, where

$$g(k) = \lceil k + \log_2 \sqrt{d} \rceil.$$

Therefore there exists a computable function f which translates H into a representation of K given as follows: $f(H) = p$ with $p = w_0 \diamondsuit w_1 \diamondsuit \dots$ such that $\iota(H_{g(k)}) = w_k$ for all $k \in \mathbb{N}$. □

Thus a hypothesis H can be viewed as a "program" of a Type-2 machine that produces a κ_{H}-representation of the figure $\kappa(H)$.

Both informants and texts are also representations (in the sense of TTE) of compact sets. Define the mapping η_{INF} by $\eta_{\mathrm{INF}}(\sigma) := K$ for every $K \in \mathscr{K}^*$ and informant σ of K, and the mapping η_{TXT} by $\eta_{\mathrm{TXT}}(\sigma) := K$ for every $K \in \mathscr{K}^*$ and text σ of K. Trivially $\eta_{\mathrm{INF}} < \eta_{\mathrm{TXT}}$ holds, that is, some Type-2 machine can translate η_{INF} to η_{TXT}, but no machine can translate η_{TXT} to η_{INF}. Moreover, we have the following hierarchy of representations.

Lemma 7.4 *$\eta_{\mathrm{INF}} < \kappa_{\mathrm{H}}$, $\eta_{\mathrm{TXT}} \not\leqslant \kappa_{\mathrm{H}}$, and $\kappa_{\mathrm{H}} \not\leqslant \eta_{\mathrm{TXT}}$.*

Proof First we prove $\eta_{\mathrm{INF}} \leqslant \kappa_{\mathrm{H}}$, that is, there is some computable function f such that $\eta_{\mathrm{INF}}(\sigma) = \kappa_{\mathrm{H}}(f(\sigma))$. Fix a figure K and its informant $\sigma \in \mathrm{dom}(\eta_{\mathrm{INF}})$. For all $k \in \mathbb{N}$, we have

$$d_{\mathrm{H}}(K, \mathrm{Pos}_k(K)) \leqslant \mathrm{diam}(k) = \sqrt{d} \cdot 2^{-k}$$

and $\mathrm{Pos}_k(K)$ can be obtained from σ. Moreover, for each k, $\sqrt{d} \cdot 2^{-g(k)} < 2^{-k}$, where

$$g(k) = \lceil k + \log_2 \sqrt{d} \rceil.$$

Therefore there exists a computable function f that translates σ into a representation of K as follows: $f(\sigma) = p$, where $p = w_0 \diamondsuit w_1 \diamondsuit \dots$ such that $w_k = \iota(\mathrm{Pos}_{g(k)}(K))$ for all $k \in \mathbb{N}$.

Second, we prove $\eta_{\mathrm{TXT}} \not\leqslant \kappa_{\mathrm{H}}$. Assume that the opposite, $\eta_{\mathrm{TXT}} \leqslant \kappa_{\mathrm{H}}$ holds. Then there exists a computable function f such that $\eta_{\mathrm{TXT}}(\sigma) = \kappa_{\mathrm{H}}(f(\sigma))$ for every figure $K \in \mathscr{K}^*$. Fix a figure K and its text $\sigma \in \mathrm{dom}(\eta_{\mathrm{TXT}})$. This means that for any small $\varepsilon \in \mathbb{R}$, f can pick up finite sequences $w_1, w_2, \dots, w_n$ from $\mathrm{Pos}(K)$ such that $d_{\mathrm{H}}(K, \nu_{\mathscr{Q}}(\iota(w_1, w_2, \dots, w_n))) \leqslant \varepsilon$. However, if such f exists, we can easily check that $\{K\} \in$ **FigEfEx-Txt**, contradicting to our result (Theorem 5.5). It follows that $\eta_{\mathrm{TXT}} \not\leqslant \kappa_{\mathrm{H}}$.

Third, we prove $\kappa_{\mathrm{H}} \not\leqslant \eta_{\mathrm{INF}}$ and $\kappa_{\mathrm{H}} \not\leqslant \eta_{\mathrm{TXT}}$. There is a figure K such that $K \cap \rho(w) = \{x\}$ for some $w \in \Sigma^*$, *i.e.*, K and $\rho(w)$ intersect in only one point x. Such a w must be in σ as a positive example, that is, $w \in \mathrm{Pos}(K)$. However, a representation of K can be constructed without w. There exists an infinite sequence $p \in \kappa_{\mathrm{H}}$ with $p = w_0 \diamondsuit w_1 \diamondsuit \dots$ such that $x \notin \nu_{\mathscr{Q}}(w_k)$ for all $k \in \mathbb{N}$. Thus, if there exists a computable f which outputs an example $(w, 1)$ from such a sequence after only seeing $w_0 \diamondsuit w_1 \diamondsuit \dots \diamondsuit w_n$, one can extend the sequence in such a way for some figure L with $w \notin \mathrm{Pos}(L)$, in contradiction to the reduction. Therefore there is no computable function that outputs an example $(w, 1)$ from p, meaning that $\kappa_{\mathrm{H}} \not\leqslant \eta_{\mathrm{INF}}$ and $\kappa_{\mathrm{H}} \not\leqslant \eta_{\mathrm{TXT}}$. □

Here we interpret *learning* of figures as *computation* based on TTE. If we see the output of a learner, *i.e.*, an infinite sequence of hypotheses, as an infinite sequence encoding a figure, the learner can be viewed as a translator of codes of figures. Naturally, we can assume that the hypothesis space $\mathscr{H}$ is a discrete topological space, that is, every hypothesis $H \in \mathscr{H}$ is isolated and is an open set itself. Define the mapping $\lim_{\mathscr{H}} : \mathscr{H}^{\omega} \to \mathscr{H}$, where $\mathscr{H}^{\omega}$ is the set of infinite sequences of hypotheses in $\mathscr{H}$, by $\lim_{\mathscr{H}}(\tau) := H$ if τ is an infinite sequence of hypotheses that converges to H, *i.e.*, there exists $n \in \mathbb{N}$ such that $\tau(i) = \tau(n)$ for all $i \geqslant n$. This coincides with the *naïve Cauchy representation* given by Weihrauch (2000) and Σ^0_2*-admissible representation* of hypotheses introduced by de Brecht and Yamamoto (2009). For any set $\mathscr{F} \subseteq \mathscr{K}^*$, let $\mathscr{F}_{\mathrm{D}}$ denote the space $\mathscr{F}$ equipped with the discrete topology, that is, every subset of $\mathscr{F}$ is open, and the mapping $\mathrm{id}_{\mathscr{F}} : \mathscr{F} \to \mathscr{F}_{\mathrm{D}}$ be the identity on $\mathscr{F}$. The computability of this identity is not trivial, since the topology of $\mathscr{F}_{\mathrm{D}}$ is finer than that of $\mathscr{F}$. Intuitively, this means that $\mathscr{F}_{\mathrm{D}}$ is more informative than $\mathscr{F}$. We can interpret learnability of $\mathscr{F}$ as computability of the identity $\mathrm{id}_{\mathscr{F}}$. The results in the following are summarized in Figure 5.

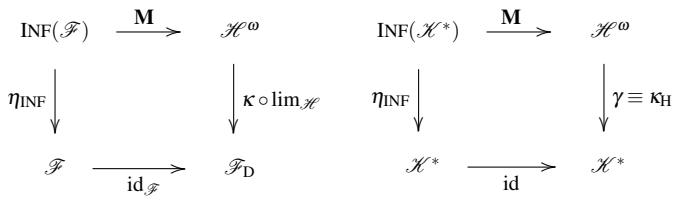

Fig. 5 The commutative diagram representing **FIGEX-INF**-learning of $\mathscr{F}$ (left), and **FIGEFEX-INF**-learning of $\mathscr{F}$ (both left and right). In this diagram, INF($\mathscr{F}$) denotes the set of informants of $K\in\mathscr{F}$.

Theorem 7.5 *A set $\mathscr{F}\subseteq\mathscr{K}^*$ is* **FIGEX-INF**-*learnable* (*resp.* **FIGEX-TXT**-*learnable*) *if and only if the identity* $\mathrm{id}_{\mathscr{F}}$ *is* $(\eta_{\mathrm{INF}},\kappa\circ\lim_{\mathscr{H}})$-*computable* (*resp.* $(\eta_{\mathrm{TXT}},\kappa\circ\lim_{\mathscr{H}})$-*computable*).

Proof We only prove the case of **FIGEX-INF**-learning, since we can prove the case of **FIGEX-TXT**-learning in exactly the same way.

The "only if" part: There is a learner $\mathbf{M}$ that **FIGEX-INF**-learns $\mathscr{F}$, hence for all $K\in\mathscr{F}$ and all $\sigma\in\mathrm{dom}(\eta_{\mathrm{INF}})$, $\mathbf{M}_\sigma$ converges to a hypothesis $H\in\mathscr{H}$ such that $\kappa(H)=K$. Thus

$$\mathrm{id}_{\mathscr{F}}\circ\eta_{\mathrm{INF}}(\sigma)=\kappa\circ\lim_{\mathscr{H}}(\mathbf{M}_\sigma), \tag{9}$$

and this means that $\mathrm{id}_{\mathscr{F}}$ is $(\eta_{\mathrm{INF}},\kappa\circ\lim_{\mathscr{H}})$-computable.

The "if" part: For some $\mathbf{M}$, the above equation (9) holds for all $\sigma\in\mathrm{dom}(\eta_{\mathrm{INF}})$. This means that $\mathbf{M}$ is a learner that **FIGEX-INF**-learns $\mathscr{F}$. □

Here we consider two more learning criteria, **FIGFIN-INF**- and **FIGFIN-TXT**-learning, where the learner generates only one correct hypothesis and halts. This learning corresponds to *finite learning* or *one shot learning* introduced by Gold (1967); Trakhtenbrot and Barzdin (1970) and it is a special case of learning with a bound of *mind change complexity*, the number of changes of hypothesis, introduced by Freivalds and Smith (1993) and used to measure the complexity of learning classes (Jain et al 1999). We obtain the following theorem.

Theorem 7.6 *A set $\mathscr{F}\subseteq\mathscr{K}^*$ is* **FIGFIN-INF**-*learnable* (*resp.* **FIGFIN-TXT**-*learnable*) *if and only if the identity* $\mathrm{id}_{\mathscr{F}}$ *is* $(\eta_{\mathrm{INF}},\kappa)$-*computable* (*resp.* $(\eta_{\mathrm{TXT}},\kappa)$-*computable*).

Proof We only prove the case of **FIGFIN-INF**-learning, since we can prove the case of **FIGFIN-TXT**-learning in exactly the same way.

The "only if" part: There is a learner $\mathbf{M}$ that **FIGFIN-INF**-learns $\mathscr{F}$, hence for all $K\in\mathscr{F}$ and all $\sigma\in\mathrm{dom}(\eta_{\mathrm{INF}})$ of K, we can assume that $\mathbf{M}_\sigma=H$ such that $\kappa(H)=K$. Thus we have

$$\mathrm{id}_{\mathscr{F}}\circ\eta_{\mathrm{INF}}(\sigma)=\kappa(\mathbf{M}_\sigma). \tag{10}$$

This means that $\mathrm{id}_{\mathscr{F}}$ is $(\eta_{\mathrm{INF}},\kappa)$-computable.

The "if" part: For some $\mathbf{M}$, the above equation (10) holds for all $\sigma\in\mathrm{dom}(\eta_{\mathrm{INF}})$. This means that $\mathbf{M}$ is a learner that **FIGFIN-INF**-learns $\mathscr{F}$. □

Finally, we show a connection between effective learning of figures and the computability of figures. Since **FIGEFEX-TXT** $=\emptyset$ (Theorem 5.5), we only treat effective learning from informants. We define the representation $\gamma:\subseteq\mathscr{H}^\omega\to\mathscr{K}^*$ by $\gamma(p):=K$ if $p=H_0,H_1,\ldots$ such that $H_i\in\mathscr{H}$ and $d_{\mathrm{H}}(K,\kappa(H_i))\leqslant 2^{-i}$ for all $i\in\mathbb{N}$.

Lemma 7.7 $\gamma\equiv\kappa_{\mathrm{H}}$.

Proof First we prove $\gamma \leqslant \kappa_{\mathrm{H}}$. For the function $g : \mathbb{N} \to \mathbb{R}$ such that

$$g(i) = \lceil i + \log_2 \sqrt{d} \rceil,$$

we have $\mathrm{diam}(g(i)) = \sqrt{d} \cdot 2^{-g(i)} \leqslant 2^{-i}$ for all $i \in \mathbb{N}$. Thus there exists a computable function f such that, for all $p \in \mathrm{dom}(\gamma)$, $f(p)$ is a representation of κ_{H} since, for an infinite sequence of hypotheses $p = H_0, H_1, \ldots$, all f has to do is to generate an infinite sequence $q = w_0 \diamondsuit w_1 \diamondsuit w_2 \diamondsuit \ldots$ such that $w_i = \iota(H^{g(i)}_{g(i)})$ for all $i \in \mathbb{N}$, which results in

$$d_{\mathrm{H}}(K, \nu_{\mathscr{D}}(w_i)) \leqslant \mathrm{diam}(g(i)) = \sqrt{d} \cdot 2^{-g(i)} \leqslant 2^{-i}$$

for all $i \in \mathbb{N}$.

Next, we prove $\kappa_{\mathrm{H}} \leqslant \gamma$. Fix $q \in \mathrm{dom}(\kappa_{\mathrm{H}})$ with $q = w_0 \diamondsuit w_1 \diamondsuit \ldots$. For each $i \in \mathbb{N}$, let $w_i = \iota(w_{i,0}, w_{i,1}, \ldots, w_{i,n})$. Then the set $\{w_{i,0}, \ldots, w_{i,n}\}$, which we denote H_i, becomes a hypothesis. From the definition of κ_{H},

$$d_{\mathrm{H}}(K, \kappa(H_i)) \leqslant 2^{-i}$$

holds for all $i \in \mathbb{N}$. This means that, for the sequence $p = w_0, w_1, \ldots$, $\gamma(p) = K$. We therefore have $\gamma \equiv \kappa_{\mathrm{H}}$. □

By using this lemma, we interpret effective learning of figures as the computability of two identities (Fig. 5).

Theorem 7.8 *A set $\mathscr{F} \subseteq \mathscr{K}^*$ is* **FIGEFEX-INF***-learnable if and only if there exists a computable function f such that f is a $(\eta_{\mathrm{INF}}, \kappa \circ \lim_{\mathscr{H}})$-realization of the identity $\mathrm{id}_{\mathscr{F}}$, and f is also a $(\eta_{\mathrm{INF}}, \gamma)$-realization of the identity* $\mathrm{id} : \mathscr{K}^* \to \mathscr{K}^*$.

Proof We prove the latter half of the theorem, since the former part can be proved exactly as for Theorem 7.5.

The "only if" part: We assume that a learner $\mathbf{M}$ **FIGEFEX-INF**-learns $\mathscr{F}$. For all $K \in \mathscr{K}^*$ and all $\sigma \in \mathrm{dom}(\eta_{\mathrm{INF}})$,

$$\mathrm{id} \circ \eta_{\mathrm{INF}}(\sigma) = \gamma(\mathbf{M}_\sigma)$$

holds since the identity id is $(\eta_{\mathrm{INF}}, \gamma)$-computable.

The "if" part: For some $\mathbf{M}$, $\mathrm{id} \circ \eta_{\mathrm{INF}}(\sigma) = \gamma(\mathbf{M}_\sigma)$ for all $\sigma \in \mathrm{dom}(\eta_{\mathrm{INF}})$. It follows that $\mathbf{M}$ is a learner that **FIGEFEX-INF**-learns $\mathscr{F}$. □

Thus in **FIGEFEX-INF**- and **FIGEFEX-TXT**-learning of a set of figures $\mathscr{F}$, a learner $\mathbf{M}$ outputs a hypothesis H with $\kappa(H) = K$ in finite time if $K \in \mathscr{F}$, and $\mathbf{M}$ outputs the "standard" representation of K if $K \in \mathscr{K}^* \setminus \mathscr{F}$ since we prove that $\gamma \equiv \kappa_{\mathrm{H}}$ in Lemma 7.7. Informally, this means that there is not too much loss of information of figures even if they are not explanatorily learnable.

8 Conclusion

We have proposed the learning of figures using self-similar sets based on Gold's learning model towards a new theoretical framework of binary classification focusing on computability, and demonstrated a learnability hierarchy under various learning criteria (Fig. 3). The key to the computable approach is the amalgamation of discretization of data and the learning process. We showed a novel mathematical connection between fractal geometry and Gold's model by measuring the lower bound of the size of training data with the Hausdorff dimension and the VC dimension. Furthermore, we analyzed our learning model using TTE (Type-2 Theory of Effectivity) and presented several mathematical connections between computability and learnability.

Many recent methods in machine learning are based on a statistical approach (Bishop 2007). The reason is that many data in the real world are in analog (real-valued) form, and the statistical approach can treat such analog data directly in theory. However, all learning methods are performed on computers. This means that all machine learning algorithms actually treat discretized digital data and, now, most research pays no attention to the gap between analog and digital data. In this paper we have proposed a novel and completely computable learning method for analog data, and have analyzed the method precisely. This work provides a theoretical foundation for computable learning from analog data, such as classification, regression, and clustering.

Acknowledgements The authors sincerely thank to the editor and anonymous reviewers for their lots of useful comments and suggestions which have led to invaluable improvements of this paper. This work was partly supported by Grant-in-Aid for Scientific Research (A) 22240010 and for JSPS Fellows 22·5714.

References

Angluin D (1980) Inductive inference of formal languages from positive data. Information and Control 45(2):117–135
Angluin D (1982) Inference of reversible languages. Journal of the ACM 29(3):741–765
Apsītis K, Arikawa S, Freivalds R, Hirowatari E, Smith CH (1999) On the inductive inference of recursive real-valued functions. Theoretical Computer Science 219(1–2):3–12
Baird DC (1994) Experimentation: An Introduction to Measurement Theory and Experiment Design, 3rd edn. Benjamin Cummings
Barnsley MF (1993) Fractals Everywhere, 2nd edn. Morgan Kaufmann
Barzdin YM (1974) Inductive inference of automata, languages and programs (in Russian). In: Proceedings of the International Congress of Mathematicians, vol 2, pp 455–460
Baum EB, Haussler D (1989) What size net gives valid generalization? Neural computation 1(1):151–160
Beer GA (1993) Topologies on Closed and Closed Convex Sets, Mathematics and Its Applications, vol 268. Kluwer Academic Publishers
Ben-David S, Dichterman E (1998) Learning with restricted focus of attention. Journal of Computer and System Sciences 56(3):277–298
Bishop C (2007) Pattern Recognition and Machine Learning (Information Science and Statistics). Springer
Blum L, Blum M (1975) Toward a mathematical theory of inductive inference. Information and Control 28(2):125–155
Blumer A, Ehrenfeucht A, Haussler D, Warmuth MK (1989) Learnability and the Vapnik-Chervonenkis dimension. Journal of the ACM 36(4):929–965
Brattka V, Presser G (2003) Computability on subsets of metric spaces. Theoretical Computer Science 305(1-3):43–76
Brattka V, Weihrauch K (1999) Computability on subsets of Euclidean space I: Closed and compact subsets. Theoretical Computer Science 219(1-2):65–93
de Brecht M (2010) Topological and algebraic aspects of algorithmic learning theory. PhD thesis, Graduate School of Informatics, Kyoto University
de Brecht M, Yamamoto A (2009) Σ^0_α-admissible representations. In: Proceedings of the 6th International Conference on Computability and Complexity in Analysis
Büchi JR (1960) On a decision method in restricted second order arithmetic. In: Proceedings of International Congress on Logic, Methodology and Philosophy of Science, pp 1–12
De La Higuera C, Janodet JC (2001) Inference of ω-languages from prefixes. In: Abe N, Khardon R, Zeugmann T (eds) Algorithmic Learning Theory, Springer, Lecture Notes in Computer Science, vol 2225, pp 364–377
Decatur SE, Gennaro R (1995) On learning from noisy and incomplete examples. In: Proceedings of the 8th Annual Conference on Computational Learning Theory, pp 353–360
Ehrenfeucht A, Haussler D, Kearns M, Valiant L (1989) A general lower bound on the number of examples needed for learning. Information and Computation 82(3):247–261
Elomaa T, Rousu J (2003) Necessary and sufficient pre-processing in numerical range discretization. Knowledge and Information Systems 5(2):162–182
Falconer K (2003) Fractal Geometry: Mathematical Foundations and Applications. Wiley
Fayyad UM, Irani KB (1993) Multi-interval discretization of continuous-valued attributes for classification learning. In: Proceedings of the 13th International Joint Conference on Artificial Intelligence, pp 1022–1029
Federer H (1996) Geometric Measure Theory. Springer
Freivalds R, Smith CH (1993) On the role of procrastination in machine learning. Information and Computation 107(2):237–271
Gama J, Pinto C (2006) Discretization from data streams: applications to histograms and data mining. In: Proceedings of the 21st Annual ACM Symposium on Applied Computing, pp 23–27
Gold EM (1965) Limiting recursion. The Journal of Symbolic Logic 30(1):28–48
Gold EM (1967) Language identification in the limit. Information and Control 10(5):447–474
Goldman SA, Kwek SS, Scott SD (2003) Learning from examples with unspecified attribute values. Information and Computation 180(2):82–100
Hirowatari E, Arikawa S (1997) Inferability of recursive real-valued functions. In: Li M, Maruoka A (eds) Algorithmic Learning Theory, Springer, Lecture Notes in Computer Science, vol 1316, pp 18–31
Hirowatari E, Arikawa S (2001) A comparison of identification criteria for inductive inference of recursive real-valued functions. Theoretical Computer Science 268(2):351–366
Hirowatari E, Hirata K, Miyahara T, Arikawa S (2003) Criteria for inductive inference with mind changes and anomalies of recursive real-valued functions. IEICE Transactions on Information and Systems 86(2):219–227
Hirowatari E, Hirata K, Miyahara T, Arikawa S (2005) Refutability and reliability for inductive inference of recursive real-valued functions. IPSJ Digital Courier 1:141–152
Hirowatari E, Hirata K, Miyahara T (2006) Prediction of recursive real-valued functions from finite examples. In: Washio T, Sakurai A, Nakajima K, Takeda H, Tojo S, Yokoo M (eds) New Frontiers in Artificial Intelligence, Springer, Lecture Notes in Computer Science, vol 4012, pp 224–234

Jain S, Sharma A (1997) Elementary formal systems, intrinsic complexity, and procrastination. Information and Computation 132(1):65–84
Jain S (2011) Hypothesis spaces for learning. Information and Computation 209(3):513–527
Jain S, Osherson D, Royer JS, Sharma A (1999) Systems That Learn, 2nd edn. The MIT Press
Jain S, Kinber E, Wiehagen R, Zeugmann T (2001) Learning recursive functions refutably. In: Abe N, Khardon R, Zeugmann T (eds) Algorithmic Learning Theory, Lecture Notes in Computer Science, vol 2225, pp 283–298
Jain S, Luo Q, Semukhin P, Stephan F (2011) Uncountable automatic classes and learning. Theoretical Computer Science 412(19):1805–1820
Jantke KP (1991) Monotonic and non-monotonic inductive inference. New Generation Computing 8(4):349–360
Kearns MJ, Vazirani UV (1994) An Introduction to Computational Learning Theory. The MIT Press
Kechris AS (1995) Classical Descriptive Set Theory. Springer
Khardon R, Roth D (1999) Learning to reason with a restricted view. Machine Learning 35(2):95–116
Kinber E (1994) Monotonicity versus efficiency for learning languages from texts. In: Algorithmic Learning Theory, Springer, Lecture Notes in Computer Science, vol 872, pp 395–406
Kobayashi S (1996) Approximate identification, finite elasticity and lattice structure of hypothesis space. Tech. Rep. CSIM 96-04, Department of Computer Science and Information Mathematics, The University of Electro-Communications
Kontkanen P, Myllymäki P, Silander T, Tirri H (1997) A bayesian approach to discretization. In: Proceedings of the European Symposium on Intelligent Techniques, pp 265–268
Lange S, Zeugmann T (1993) Monotonic versus non-monotonic language learning. In: Nonmonotonic and Inductive Logic, Springer, Lecture Notes in Computer Science, vol 659, pp 254–269
Lange S, Zeugmann T (1994) Characterization of language learning front informant under various monotonicity constraints. Journal of Experimental & Theoretical Artificial Intelligence 6(1):73–94
Lange S, Zeugmann T, Zilles S (2008) Learning indexed families of recursive languages from positive data: A survey. Theoretical Computer Science 397(1–3):194–232
Li M, Chen X, Li X, Ma B, Vitányi P (2003) The similarity metric. In: Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms, pp 863–872
Lin J, Keogh E, Lonardi S, Chiu B (2003) A symbolic representation of time series, with implications for streaming algorithms. In: Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pp 1–11
Liu H, Hussain F, Tan CL, Dash M (2002) Discretization: An enabling technique. Data Mining and Knowledge Discovery 6(4):393–423
Long PM, Tan L (1998) PAC learning axis-aligned rectangles with respect to product distributions from multiple-instance examples. Machine Learning 30(1):7–21
Mandelbrot BB (1982) The Fractal Geometry of Nature. W. H. Freeman
Merkle W, Stephan F (2003) Refuting learning revisited. Theoretical Computer Science 298(1):145–177
Michael L (2010) Partial observability and learnability. Artificial Intelligence 174(11):639–669
Michael L (2011) Missing information impediments to learnability. In: 24th Annual Conference on Learning Theory, pp 1–2
Minicozzi E (1976) Some natural properties of strong-identification in inductive inference. Theoretical Computer Science 2(3):345–360
Motoki T, Shinohara T, Wright K (1991) The correct definition of finite elasticity: Corrigendum to identification of unions. In: Proceedings of the 4th Annual Workshop on Computational Learning Theory, p 375
Mukouchi Y, Arikawa S (1995) Towards a mathematical theory of machine discovery from facts. Theoretical Computer Science 137(1):53–84
Mukouchi Y, Sato M (2003) Refutable language learning with a neighbor system. Theoretical Computer Science 298(1):89–110
Müller N (2001) The iRRAM: Exact arithmetic in C++. In: Blanck J, Brattka V, Hertling P (eds) Computability and Complexity in Analysis, Springer, Lecture Notes in Computer Science, vol 2064, pp 222–252
Perrin D, Pin JÉ (2004) Infinite words. Elsevier
Rosenblatt F (1958) The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review 65(6):386–408
Sakurai A (1991) Inductive inference of formal languages from positive data enumerated primitive-recursively. In: Algorithmic Learning Theory, JSAI, pp 73–83
Schröder M (2002a) Admissible representations for continuous computations. PhD thesis, dem Fachbereich Informatik, der FernUniversit at – Gesamthochschule in Hagen
Schröder M (2002b) Extended admissibility. Theoretical Computer Science 284(2):519–538
Shapiro EY (1981) Inductive inference of theories from facts. Tech. rep., Department of Computer Science, Yale University
Shapiro EY (1983) Algorithmic Program Debugging. The MIT Press
Skubacz M, Hollmén J (2000) Quantization of continuous input variables for binary classification. In: Intelligent Data Engineering and Automated Learning — IDEAL 2000. Data Mining, Financial Engineering, and Intelligent Agents, Springer, Lecture Notes in Computer Science, vol 1983, pp 42–47
Sugiyama M, Yamamoto A (2010) The coding divergence for measuring the complexity of separating two sets. In: Proceedings of 2nd Asian Conference on Machine Learning, JMLR Workshop and Conference Proceedings, vol 13, pp 127–143
Sugiyama M, Hirowatari E, Tsuiki H, Yamamoto A (2006) Learning from real-valued data with the model inference mechanism through the Gray-code embedding. In: Proceedings of 4th Workshop on Learning with Logics and Logics for Learning (LLLL2006), pp 31–37

Sugiyama M, Hirowatari E, Tsuiki H, Yamamoto A (2009) Learning figures with the Hausdorff metric by self-similar sets. In: Proceedings of 6th Workshop on Learning with Logics and Logics for Learning (LLLL2009), pp 27–34

Sugiyama, M., Hirowatari, E., Tsuiki, H., & Yamamoto, A. (2010). Learning figures with the Hausdorff metric by fractals. In M. Hutter, F. Stephan, V. Vovk, & T. Zeugmann (Eds.), *Lecture notes in computer science: Vol. 6331. Algorithmic learning theory* (pp. 315–329). Canberra: Springer.

Tavana NR, Weihrauch K (2011) Turing machines on represented sets, a model of computation for analysis. Logical Methods in Computer Science 7(2):1–21

Trakhtenbrot B, Barzdin YM (1970) Konetschnyje awtomaty (powedenie i sintez). English Translation: Finite automata-behavior and synthesis, Fundamental Studies in Computer Science 1, 1975

Turing AM (1937) On computable numbers, with the application to the entscheidungsproblem. Proceedings of the London Mathematical Society 1(42):230–265

Valiant LG (1984) A theory of the learnable. Communications of the ACM 27(11):1134–1142

Vapnik V, Chervonenkis A (1971) On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and its Applications 16(2):264–280

Weihrauch K (2000) Computable Analysis: An Introduction. Springer

Weihrauch K (2008) The computable multi-functions on multi-represented sets are closed under programming. Journal of Universal Computer Science 14(6):801–844

Weihrauch K, Grubba T (2009) Elementary computable topology. Journal of Universal Computer Science 15(6):1381–1422

Wiehagen R (1991) A thesis in inductive inference. In: Dix J, Jantke KP, Schmitt PH (eds) Nonmonotonic and Inductive Logic, Springer, Lecture Notes in Computer Science, vol 543, pp 184–207

Wright K (1989) Identification of unions of languages drawn from an identifiable class. In: Proceedings of the 2nd Annual Workshop on Computational Learning Theory, pp 328–333

Zeugmann T, Zilles S (2008) Learning recursive functions: A survey. Theoretical Computer Science 397(1-3):4–56

Zeugmann T, Lange S, Kapur S (1995) Characterizations of monotonic and dual monotonic language learning. Information and Computation 120(2):155–173