Theoretical Computer Science 412 (2011) 4650–4660

Contents lists available at ScienceDirect

Theoretical Computer Science

journal homepage: www.elsevier.com/locate/tcs

Learning Boolean functions in AC^0 on attribute and classification noise—Estimating an upper bound on attribute and classification noise

Akinobu Miyata *, Jun Tarui, Etsuji Tomita [1]

Department of Information and Communication Engineering, The University of Electro-Communications, Chofugaoka 1-5-1, Chofu, Tokyo 182-8585, Japan

ARTICLE INFO

Article history:
Received 27 September 2009
Received in revised form 26 January 2011
Accepted 27 April 2011
Communicated by G. Ausiello

Keywords:
Learning
Boolean functions
AC^0
Attribute noise
Classification noise

ABSTRACT

We study a procedure for estimating an upper bound of an unknown noise factor in the frequency domain. A learning algorithm using a Fourier transformation method was originally given by Linial, Mansour and Nisan. While Linial, Mansour and Nisan assumed that the learning algorithm estimates Fourier coefficients from noiseless data, Bshouty, Jackson, and Tamon, and also Ohtsuki and Tomita extended the algorithm to ones that are robust for noisy data. The noise process that we consider is as follows: for an example $\langle x, f(x)\rangle$, where $x \in \{0,1\}^n, f(x) \in \{-1,1\}$, each bit of x *and* $f(x)$ gets flipped independently with probability η during a learning process. The previous learning algorithms for noisy data all assume that the noise factor η or an upper bound of η is known in advance. The learning algorithm proposed in this paper works without this assumption. We estimate an upper bound of the noise factor by evaluating a noisy power spectrum in the frequency domain and by using a sampling trick. Combining this procedure with Ohtsuki and Tomita's algorithm, we obtain a quasi-polynomial-time learning algorithm that can cope with noise without knowing any information about the noise in advance.

1. Introduction

Linial et al. [5] gave a learning algorithm for Boolean functions in class AC^0 under a uniform distribution. The learning algorithm uses a Fourier transformation method via estimating Fourier coefficients from noiseless data.

Subsequently, Bshouty et al. [1] showed that learning is possible in cases where certain types of noise are mixed in. They considered two types of noise: attribute noise and classification noise. Using the Fourier-based algorithm of Linial et al. [5], they showed how learning can take place under the assumption that the precise values of these noise factors are made available to the learner in advance. Meanwhile, Ohtsuki and Tomita [7] gave a learning algorithm that works even when the noise factor itself remains unknown as long as an upper bound for the noise factor is provided to the learner in advance. Bshouty et al. [2] also demonstrated a method that works when only an upper bound for the noise factor is provided in advance with a sufficient approximation of the noise factor being provided by a certain type of special oracle that characterizes the noise process, thereby eliminating their above-mentioned assumption [1].

There have been many previous studies on learning in cases where noise gets mixed in. In such cases, successful learning critically depends on the availability of information about the noise. In many cases, the availability of suitable noise-related information allows the learner to cope with effects of noise. Consequently, by assuming that the learner has prior knowledge about the noise such as the noise factor or an upper bound of it, we are making things rather easy from the learner's

* Corresponding address: HIBIKI IP Law Firm, Asahi Bldg. 10th Floor, Tsuruya-cho 3-33-8, Kanagawa-ku, Yokohama, Kanagawa 221-0835, Japan. Tel.: +81 9014590113.

E-mail addresses: akinobu-m@sannet.ne.jp (A. Miyata), tarui@ice.uec.ac.jp (J. Tarui), tomita@ice.uec.ac.jp (E. Tomita).

[1] The author is also with the Research and Development Initiative, Chuo University, Kasuga 1-13-27, Bunkyo-ku, Tokyo 112-8551, Japan.

0304-3975/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2011.04.047

perspective. The noise factor η considered by Bshouty et al. [2] and by Ohtsuki and Tomita [7] is unknown to the learner, but it is assumed that $0 \leq \eta < 1/2$; that is, the bound of the noise factor is away from 1/2. Laird [4] showed that it is possible to estimate an upper bound for the noise factor when classification noise alone is mixed in.

It is preferable to restrict the noise-related information disclosed to the learner in advance, and if some way can be found of making an *a priori* estimate of this noise-related information, then learning can take place under more natural assumptions. Specifically, this would allow learning to take place under the assumption that noise is liable to be mixed in with a noise factor of less than 1/2 and would further reduce the amount of information required in advance for learning to take place.

This paper proposes a method for estimating an upper bound of the noise factor directly from noisy samples by considering the frequency domain. This method involves comparing the power spectrum measured from actual samples with values related to the power spectrum computed from a hypothetical value for a noise factor upper bound. With this method, an upper bound of the noise factor can easily be estimated simply by estimating the amount of decay across the power spectrum, so it is quite different from the method of Laird [4], where predicted and experiential mismatch factors of the hypothesis on samples are evaluated. Our method can be applied not only to AC^0 Boolean functions, but also to any function where the function's power spectrum is concentrated on lower frequencies.

Furthermore, by combining it with the learning algorithm of Ohtsuki and Tomita [7], we can obtain a more powerful learning algorithm that operates under the sole assumption that examples provided to the learner are varied by an unknown probability η $(0 \leq \eta < 1/2)$.

A preliminary version of this paper appeared in [6], in which all the necessary contents of [7] are quoted.

2. Definitions

A Boolean function f is a map $f : \{0, 1\}^n \rightarrow \{-1, 1\}$. For x in $\{0, 1\}^n$, $|x|$ represents the Hamming weight, i.e., the number of 1s in x, and $x[i]$ represents the ith bit of x. The bitwise exclusive-OR of two n-bit vectors x^1 and x^2 is denoted by $x^1 \oplus x^2$.

Valiant [8] introduced the Probably Approximately Correctly (PAC) learning model. In this paper, we consider a problem of PAC learning with respect to a uniform distribution over examples with attribute and classification noise. Let F_n be a set of Boolean functions (also called a concept class) and assume that $f,\ h \in F_n$. The learning model used in this paper is described below.

The learning algorithm stops after obtaining a number of examples selected at random under a uniform distribution, and then outputs a hypothesis h. An example $\langle x, f(x)\rangle$ comprises the combination of an input $x \in \{0, 1\}^n$ and its corresponding output $f(x) \in \{1, -1\}$. Under a uniform distribution, the probability that hypothesis h differs from the target f is $\Pr_U[f(x) \neq h(x)]$, which is expressed as $\Pr_U[f(x) \neq h(x)] = error_f(h)$. The success of learning is evaluated by two parameters: a precision parameter ϵ and an accuracy parameter δ. For any values of f, ϵ $(0 < \epsilon < 1)$, and δ $(0 < \delta < 1)$ in F_n, the learning algorithm L outputs hypothesis h in which the error rate $error(h)$ is at most ϵ with a probability of at least $1 - \delta$.

A hypothesis h for which the error rate $error_f(h)$ is less than or equal to ϵ is called an ϵ-good hypothesis; one whose error rate is greater than ϵ is called an ϵ-bad hypothesis. When a learning algorithm exists for concept class, the concept class is said to be "learnable" under the uniform distribution.

3. Learning algorithm based on Fourier transformation

Let f be a Boolean function from $\{0, 1\}^n$ to $\{-1, 1\}$. For α in $\{0, 1\}^n$, an orthonormal basis $\chi_\alpha(x)$ is given by $\chi_\alpha(x) = (-1)^{\alpha\cdot x}$ $(\alpha \cdot x = \sum_i \alpha[i]x[i])$. Using a Fourier transformation method, it is possible to express any Boolean function f by the following formula, where $\hat{f}(\alpha) = E_U[f(x)\chi_\alpha(x)]$ represents the Fourier coefficients of f in α, and $|\alpha|$ is called the frequency of α:

$$f(x) = \sum_\alpha \hat{f}(\alpha)\chi_\alpha(x).$$

3.1. Learning of constant-depth circuits

Let f be a Boolean function that can be computed by a constant-depth, polynomial-size circuit (AC^0 circuit). Linial et al. [5] proposed an algorithm that can learn a Boolean function f of this sort in $O(n^{polylog(n)})$ time under a uniform distribution by using a Fourier transformation method. They showed that the Fourier coefficients $\hat{f}(\alpha)$ of f that can be computed from an AC^0 circuit have the property that the majority of the components $\sum_\alpha \hat{f}(\alpha)^2$ are concentrated at low frequencies $(\sum_\alpha \hat{f}(\alpha)^2 \leq \epsilon/2$ for any α such that $|\alpha| \leq k)$. Let $h(x) = \sum_{\alpha:|\alpha|\leq k} \hat{h}(\alpha)\chi_\alpha(x)$ be a hypothesis of f, where $\hat{h}(\alpha)$ is an empirical value of $\hat{f}(\alpha)$, and let $k = O(\log^d n/\epsilon)$ be a frequency-related constant (where d is the depth of the AC^0 circuit). There are $\binom{n}{k}$ Fourier coefficients $\hat{h}(\alpha)$ with Hamming weight k. It is possible to construct an ϵ-good hypothesis $h(x)$ by making a good estimate of these Fourier coefficients.

4. Learning from noisy examples

4.1. Noise model

In this paper, we consider learning in environments where there are two types of noise, called “attribute noise” and “classification noise”, as discussed in parts of references [6,7,2]. Specifically, in a process that receives examples $\langle x, f(x)\rangle$ from an oracle $EX()$, the noise process changes the examples into noisy examples $\langle x \oplus x_N, f(x)y_N\rangle$ ($x_N \in \{0,1\}^n$, $y_N \in \{1,-1\}$) with a certain fixed probability. An oracle that provides these noisy examples is denoted by $EX()_\eta$, where η is called the noise factor ($0 \le \eta < 1/2$). The noisy oracle $EX()_\eta$ mixes examples with $x_N \in \{0,1\}^n$ and $y_N \in \{-1,+1\}$, which occur with respective probabilities

$$\prod_{x_N[i]=1} \eta \prod_{x_N[i]=0} (1-\eta), \qquad \prod_{y_N=-1} \eta \prod_{y_N=1} (1-\eta).$$

When this noise is mixed into the examples, the Fourier coefficients obtained from the sample are reduced according to the noise factor, as shown by the following lemma. (A proof of Lemma 1 can be found in [1,2,7], etc.)

Lemma 1 (*Bshouty et al. [1,2], Ohtsuki and Tomita [7]*)**.** *Let f be a Boolean function whose noise factor η lies in the range $0 \le \eta < 1/2$. Then, the expectation values of the Fourier coefficients obtained from noisy samples $\{\langle x^j \oplus x_N^j, f(x^j)y_N^j\rangle\}_{j=1}^m$ are given by*

$$E_U[f(x)y_N\chi_\alpha(x \oplus x_N)] = (1-2\eta)^{|\alpha|+1}\hat{f}(\alpha). \quad \square$$

According to Lemma 1, the Fourier coefficients obtained from noisy samples end up as Fourier coefficients with added bias terms $(1-2\eta)^{|\alpha|+1}$, so if these reduced Fourier coefficients are used directly in the construction of a hypothesis, it is not possible to obtain the desired ϵ-good hypothesis. It is thus necessary to somehow remove the bias terms $(1-2\eta)^{|\alpha|+1}$ by estimating η in order to recover the Fourier coefficients $\hat{f}(\alpha)$ before reduction.

4.2. Algorithm for learning from noisy examples

Bshouty et al. [2] demonstrated that learning is possible with this sort of noise model, and as one example of an application, they demonstrated a learning algorithm for class AC^0 Boolean functions under a uniform distribution. They showed that learning is possible if the noise-related information assumed to be known in advance consists of the upper bound of the noise factor and if a special oracle that can estimate η with adequate precision is provided.

On the other hand, as described in reference [6], Ohtsuki and Tomita [7] demonstrated a learning algorithm called Robust k-lowdegree (rkl) that estimates η from only examples provided by an oracle $EX_\eta()$ even when the upper bound of the noise factor is already known [7].

As demonstrated by Laird [4], in problems involving learning from noisy examples, the size of samples required for learning is generally inversely proportional to the size of the noise factor ($0 \le \eta < 1/2$). Specifically, the sample size is inversely proportional to $1-2\eta$, and as η approaches 1/2, the required sample size becomes larger.

Accordingly, when no prior information whatsoever is available beforehand when learning from a noisy input, the learning algorithm must achieve the complex task of estimating the noise factor while determining a suitable sample size through trial and error.

4.3. Estimation of upper bound when classification noise is included

Laird [4] also showed that it is possible to estimate an upper bound η_0 for the noise factor in the classification noise model. Classification noise refers to cases where, for an example, in $\langle x, f(x)\rangle$, only the $f(x)$ part changes with probability η.

The upper bound estimation algorithm described by Laird [4] evaluates the ratios at which a number of different hypotheses held by the learning algorithm form a mismatch with a suitably sized sample. This is based on the idea that the hypothesis with the smallest mismatch factor for the noise sample is likely to be an ϵ-good hypothesis. An attempt is then made to obtain an upper bound estimate value by comparing this minimum mismatch factor with a hypothetical mismatch factor predicted from a certain upper bound estimate value $\hat{\eta}$. A stable estimate of the unknown noise factor is made using a technique in which the noise factor's estimated value is gradually increased while the sample size is also gradually increased. For details, see Laird's paper [4].

If attribute noise is also mixed in, the task of evaluating the mismatch becomes harder. With Laird's method [4], the desired hypothesis (an ϵ-good hypothesis) may not have the smallest mismatch factor for the noise sample. Thus, it is currently impossible to apply Laird's method [4] directly.

In this paper, therefore, we estimate the upper bound of the noise factor η by using a method that involves evaluating the amount of reduction in the power spectrum obtained from the samples in the frequency domain.

5. Method for estimating the noise factor's upper bound

An algorithm for learning about noise factors under the assumption that the learner is provided in advance with an upper bound for the noise factor instead of accurate values of the actual noise factor has been demonstrated (Bshouty et al. [2], Ohtsuki and Tomita [7]). Specifically, if we assume that noise factor remains unknown but that the learner is provided with at least knowledge about its upper bound such that the noise factor is known to be less than or equal to a certain value, then this information is sufficient for constructing the desired hypothesis.

The reason why the desired hypothesis can be constructed by replacing the noise factor with its upper bound is as follows.

- Since the upper bound of the noise factor η is provided, the learning algorithm can determine in advance a suitable sample size for the learning process. In this way, the learning algorithm can determine the Fourier coefficients $(1-2\eta)^{|\alpha|+1}\hat{f}(\alpha)$ obtained from the noise sample within a fixed error range.
- In the limited range up to the upper bound η_0, the learning algorithm can estimate the unknown noise factor η within a fixed margin of error.

Furthermore, by using an upper bound, we can at least identify an adequate sample size necessary for learning, albeit not the bare minimum necessary sample size.

On this basis, in this paper we propose a method for estimating the upper bound of the noise factor, which is expected to be relatively effortless to estimate compared with the unknown noise factor itself.

5.1. Estimation idea

We know that the noise factor lies somewhere in the range $0 \le \eta < 1/2$, even if the noise factor mixed into the samples is unknown. Moreover, we know how the effects of noise appear in the Fourier domain. In a Fourier-based learning model, the degree of reduction of each Fourier coefficient is expressed by a bias term $(1-2\eta)^{|\alpha|+1}$. As the noise factor η increases or as the frequency $|\alpha|$ increases, the effects of noise become stronger and the observed power spectrum decreases.

Here, we introduce an estimated value η' of an unknown noise factor η. We compare the power spectra obtained with noise factors of η and η'. If η' is larger than or equal to η, we find that the power spectrum obtained with η' is lower than or equal to that obtained with η.

On the basis of this idea, we can estimate the upper bound of the unknown noise factor η. We then determine the (reduced) power spectrum predicted by this hypothetical value η' and compare this assumed power spectrum with the power spectrum actually measured by the learner from the samples. If the η'-based power spectrum is lower than the power spectrum actually measured from the samples, then η' is a possible upper bound of the unknown noise factor η.

In this power spectrum comparison, the sample size is an issue. We assume that due consideration is given to errors in the power spectrum caused by the noise factor's hypothetical value η' because although a sufficiently large sample can be used to correct the errors in these Fourier coefficients $(1-2\eta)^{|\alpha|+1}\hat{f}(\alpha)$, the size of this unknown sample depends on the unknown noise factor.

Theorem 1. *Let η' $(0 \le \eta' < 1/2)$ be an estimated value of the unknown noise factor η or its upper bound, and let δ $(0 \le \delta < 1)$ be a parameter expressing the level of confidence in the estimated noise factor η'. Let n be the input size, d the depth of AC^0 circuits, and k the maximum Hamming weight of α related to the Fourier coefficients being estimated, given as $k = O(\log^d(n))$ $(0 \le k \le n)$. Suppose that we obtained an empirical value c_α for the expectation value $(1-2\eta)^{|\alpha|+1}\hat{f}(\alpha)$ of the Fourier coefficients with bias terms from a sample of size $m(\eta')$, and let $s(\eta')$ be an upper bound of the error from the expectation value for the power spectrum $\sum_{\alpha:|\alpha|\le k} c_\alpha^2$, given as $s(\eta') = (1-2\eta')^{2k+2}/2n$. Then, when inequality (1) has been established for a specific frequency i $(0 \le i \le k)$, the estimated value η' is the unknown noise factor η or its upper bound $(\eta \le \eta' < 1/2)$ with a probability of at least $1-\delta$.*

$$(1-2\eta')^{2i+2} \le \sum_{\alpha:|\alpha|=i,i\le k} c_\alpha^2 - s(\eta'). \tag{1}$$

Here, the size $m(\eta')$ defined by the following formula is requested from the oracle $EX()_\eta$;

$$m(\eta') = 32n^{2k+2}\left(\frac{1}{1-2\eta'}\right)^{4k+4} \ln\left(\frac{2n^k}{\delta}\right).$$

Proof. Before we start proving, we recall the following version of Hoeffding bounds [3]. To prove Theorem 1, we will use Lemma 2.

Lemma 2 (*Hoeffding Bounds [3]*). *Let X_i $(1 \le i \le m)$ be independent, identically distributed random variables, where $E[X_i] = \mu$ and $|Xi| \le 1$. Then*

$$Pr\left[\left|\frac{1}{m}\sum_{i=1}^{m} X_i - \mu\right|\right] \le \delta,$$

whenever $m \ge (1/2\gamma^2)\ln(2/\delta)$. □

There are two points to be proved:

(A) That a sample of size $m(\eta')$ is large enough to obtain an empirical value $\sum_{\alpha:|\alpha|\le k} c_\alpha^2$ of the power spectrum with a probability of at least $1-\delta$ from the expectation value within the range of error $s(\eta') = (1-2\eta')^{2k+2}/2n$.

(B) That when η' satisfies inequality (1) of Theorem 1, it is the upper bound of the noise factor η.

We consider point (A) first.

A sample of size $m(\eta')$ is large enough to obtain an empirical value c_α with a probability of at least $1-\delta$ for the expectation values $(1-2\eta)^{|\alpha|+1}\hat{f}(\alpha)$ of each Fourier coefficient with a range of error $\gamma_c' = (1-2\eta')^{2k+2}/(4n^{k+1})$. This can easily be understood from Lemma 2 (Hoeffding bounds [3]). In other words, there is a high probability that $|(1-2\eta)^{|\alpha|+1}\hat{f}(\alpha) - c_\alpha| \le \gamma_c'$.

On the basis of this fact, as shown below, from a sample $m(\eta')$, at an arbitrary frequency $|\alpha|$ less than or equal to k, it is possible to obtain an empirical value $\sum_{\alpha:|\alpha|\le k} c_\alpha^2$ within a range of error $s(\eta') = (1-2\eta')^{2k+2}/2n$ for a power spectrum expectation value of $\sum_{\alpha:|\alpha|\le k}(1-2\eta)^{2|\alpha|+2}\hat{f}(\alpha)^2$. Here, we use the relationship $0 \le |(1-2\eta)^{|\alpha|+1}\hat{f}(\alpha)|$ and $|c_\alpha| \le 1$.

$$\because \sum_{\alpha:|\alpha|\le k} |(1-2\eta)^{2|\alpha|+2}\hat{f}(\alpha)^2 - c_\alpha^2| = \sum_{\alpha:|\alpha|\le k} |(1-2\eta)^{|\alpha|+1}\hat{f}(\alpha) + c_\alpha||(1-2\eta)^{|\alpha|+1}\hat{f}(\alpha) - c_\alpha|$$
$$\le n^k \times 2 \times \gamma_c' = \frac{(1-2\eta')^{2k+2}}{2n}.$$

It can thus be said that a sample of size $m(\eta')$ is large enough to obtain an empirical value $\sum_{\alpha:|\alpha|\le k} c_\alpha^2$ of the expectation value $\sum_{\alpha:|\alpha|\le k}(1-2\eta)^{2|\alpha|+2}\hat{f}(\alpha)^2$ within an error of $s(\eta')$, as expressed by

$$\Pr_U\left[\sum_{\alpha:|\alpha|\le k} |(1-2\eta)^{2|\alpha|+2}\hat{f}(\alpha)^2 - c_\alpha^2| \le s(\eta')\right] \ge 1-\delta.$$

At the same time, we can derive inequality (2), which happens with a probability of at least $1-\delta$. We use inequality (2) in the proof of point (B).

$$\sum_{\alpha:|\alpha|\le k} c_\alpha^2 - s(\eta') \le \sum_{\alpha:|\alpha|\le k} (1-2\eta)^{2|\alpha|+2}\hat{f}(\alpha)^2. \quad \Box \tag{2}$$

Point (B) is discussed next. A comparative value is introduced between the power spectra $\sum_{\alpha:|\alpha|\le k} c_\alpha^2$ obtained from the unknown noise factor η. For the empirical value $\sum_{\alpha:|\alpha|\le k} c_\alpha^2$ of the power spectra considering error $s(\eta')$, we assume a real number p that satisfies

$$p \le \sum_{\alpha:|\alpha|\le k} c_\alpha^2 - s(\eta').$$

This value p is a hypothetical value related to the power spectrum estimated from the noise factor's estimated value η', and is determined as $p = (1-2\eta')^{2i+2}$ for a specific frequency i $(0 \le i \le k)$. When $i = k$, this value p is equal to the value of the power spectrum $(1-2\eta)^{2k+2} \cdot 1$ (where $\sum_{\alpha:|\alpha|=k}\hat{f}(\alpha)^2 = 1$) when all the Fourier components are concentrated at frequency k, which is equivalent to the case where the power spectrum is reduced the most by the effects of noise. When we use the above inequality (2) and the relationship between this value p and the empirical value $\sum_{\alpha:|\alpha|\le k} c_\alpha^2$ of the power spectrum for a specific frequency i $(0 \le i \le k)$, the estimated value η' becomes the upper bound of η:

$$p \le \sum_{\alpha:|\alpha|=i,i\le k} c_\alpha^2 - s(\eta') \le (1-2\eta)^{2i+2} \sum_{\alpha:|\alpha|=i,i\le k} \hat{f}(\alpha)^2$$
$$\Rightarrow (1-2\eta')^{2i+2} \le (1-2\eta)^{2i+2} \sum_{\alpha:|\alpha|=i,i\le k} \hat{f}(\alpha)^2.$$

Thus, we can say that the estimated value η' is the upper bound of η, as shown below.

$$(1-2\eta')^{2i+2} \le (1-2\eta)^{2i+2} \sum_{\alpha:|\alpha|=i,i\le k} \hat{f}(\alpha)^2$$
$$\Rightarrow \left(\frac{1-2\eta'}{1-2\eta}\right)^{2i+2} \le \sum_{\alpha:|\alpha|=i,i\le k} \hat{f}(\alpha)^2$$
$$\Rightarrow 1-2\eta' \le 1-2\eta \; (\because\; 0 \le \sum_{\alpha:|\alpha|=i,i\le k} \hat{f}(\alpha)^2 \le 1)$$
$$\Rightarrow \eta \le \eta'. \quad \Box$$

These two points complete the proof of Theorem 1.

In the following section, we discuss a procedure for estimating the upper bound of the noise factor using criteria according to Theorem 1.

5.2. Procedure for estimating the noise factor's upper bound

The procedure for estimating the noise factor's upper bound involves (i) comparing the hypothetical value of the power spectrum obtained from η' (the upper bound estimate value of the noise factor) with the empirical value of the power spectrum actually observed from a sample including noisy examples and (ii) judging whether or not the estimated value η' is the upper bound of the unknown noise factor η.

Upper bound estimation procedure (δ):
Define $m_r(\eta')$ and $s(\eta')$ as follows:

$$m_r(\eta') = 32n^{2k+2}\left(\frac{1}{1-2\eta'}\right)^{4k+4} \ln\left(\frac{2^{r+1}n^k}{\delta}\right),$$

$$s(\eta') = \frac{(1-2\eta')^{2k+2}}{2n}.$$

Input: δ $(0 < \delta < 1)$, $k = O\left(\log^d n\right)$.
Output: η' such that $\eta' \geq \eta$
Procedure

```
1 $r := 1, \eta' := 0$
2 (round $r$) Repeat until the halt condition is satisfied.
  2.1 Request $m_r(\eta')$ examples.
  2.2 $m := m_r(\eta')$
  2.3 For each $\alpha$ such that $|\alpha| \leq k$, obtain
        $c_\alpha = \frac{1}{m}\sum_{j=1}^{m} f(x^j)y_N^j\chi_\alpha(x^j \oplus x_N^j).$
  2.4 $p := (1-2\eta'), s := s(\eta')$
  2.5 If the following inequality is satisfied for a specific frequency $i$ $(0 \leq i \leq k)$, then halt and output $\eta'$.
        $p^{2i+2} \leq \sum_{\alpha:|\alpha|=i} c_\alpha^2 - s.$
  2.6 Otherwise, perform the following steps and go to the next round.
        $r := r+1, \qquad \eta' := \frac{1}{2} - \frac{1}{2^{r+1}}.$
```

Here, we explain the situation in which the halt condition is satisfied by referring to Fig. 1. Fig. 1(a) shows the ideal Fourier spectra and Fig. 1(b) shows the empirical Fourier spectra obtained from noisy samples. Fig. 1(c) shows the situation when the halt condition is satisfied at $|\alpha| = 2$.

Theorem 2. *If the sample size $m_r(\eta')$ is declared as*

$$m_r(\eta') = 32n^{2k+2}\left(\frac{1}{1-2\eta'}\right)^{4k+4} \ln\left(\frac{2^{r+1}n^k}{\delta}\right),$$

then the procedure for estimating the noise factor's upper bound will stop at or before round $r_0 = 2 + \lceil \log_2(1-2\eta)^{-1} \rceil$ with a probability of at least $1-\delta$ and will output a value η' that is the upper bound of η $(0 \leq \eta \leq \eta' < 1/2)$.

Proof. The estimation procedure fails if one or both of the following situations occur. Therefore, in order to prove Theorem 2, we show that the sum of their occurrence probabilities is less than δ.

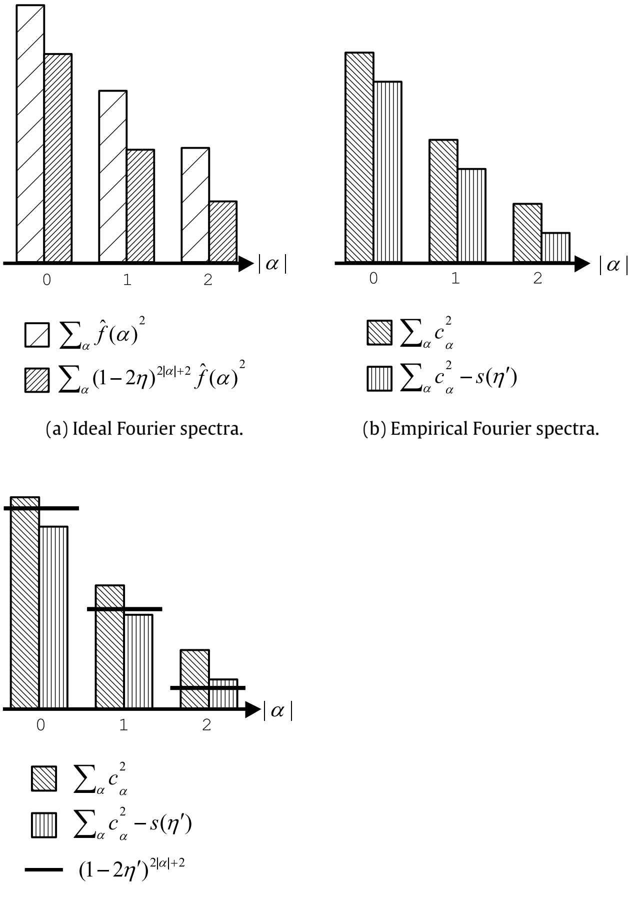
1. When the estimation procedure has stopped, the output estimation value η' does not constitute the upper bound of the noise factor η $(\eta' < \eta)$.
2. The estimation procedure does not stop at or before round $r_0 = 2 + \lceil \log_2(1-2\eta)^{-1} \rceil$.

Before discussing the two above-mentioned situations related to the proof of Theorem 2, let us first confirm the relationship between the sample size $m_r(\eta')$ and the error of the empirical values c_α of the Fourier coefficients used as a premise for the correct operation of the estimation procedure.

As the following formula shows, the sample size $m_r(\eta')$ is determined so that in any round, the probability of obtaining the empirical value c_α of any Fourier coefficient with a discrepancy greater than or equal to an error γ_c' from the expectation value $(1-2\eta)^{|\alpha|+1}\hat{f}(\alpha)$ is suppressed to less than or equal to $\delta/2$.

$$\Pr_U\left[\sum_{\alpha:|\alpha|\leq k} |(1-2\eta)^{|\alpha|+1}\hat{f}(\alpha) - c_\alpha| \geq \gamma_c'\right] \leq \frac{\delta/2}{2^r \times n^k} \times 2^r \times n^k = \frac{\delta}{2}.$$

The term 2^r in this formula represents the number of all possible rounds with regard to whether or not it was possible to estimate all the desired values of c_α in each round within the desired error range after r rounds of independent attempts.



(a) Ideal Fourier spectra.

(b) Empirical Fourier spectra.

(c) Satisfied halt condition.

Fig. 1. Fourier spectra and halt condition.

On the basis of this reasoning, which constitutes the premise of the estimation procedure, we now show that the above-mentioned two probabilities related to Theorem 2 are each less than or equal to $\delta/2$ for a sample of this size $m_r(\eta)$.

We start by considering the first of these situations.

If the estimation procedure has stopped at **Step 2** because it satisfies the halt condition, this means that the output estimated value η' represents the upper bound of η with a probability of at least $1 - \delta/2$.

Assume that the estimation procedure has stopped at round r_0. Here, the sample of size $m_r(\eta')$ is large enough to satisfy inequality E_1 for an estimated value of the power spectrum $\sum_{\alpha:|\alpha|\le k} c_\alpha^2$ with a probability of at least $1-\delta/2$ in the estimation procedure.

$$E_1\colon \sum_{\alpha:|\alpha|\le k} c_\alpha^2 \le (1-2\eta)^{2|\alpha|+2} \sum_{\alpha:|\alpha|\le k} \hat{f}(\alpha)^2 + s(\eta').$$

Moreover, since the estimation procedure comes to a halt, the halt condition E_2 holds for a specific frequency with regard to η', the estimated value of the noise factor's upper bound.

$$E_2\colon (1-2\eta')^{2|\alpha|+2} \le \sum_{\alpha:|\alpha|\le k} c_\alpha^2 - s(\eta')$$
$$\to (1-2\eta')^{2|\alpha|+2} + s(\eta') \le \sum_{\alpha:|\alpha|\le k} c_\alpha^2.$$

The following inequalities can thus be derived from inequalities E_1 and E_2, from which it follows that $\eta' \geq \eta$.

$$
\begin{aligned}
&(1-2\eta')^{2|\alpha|+2} + s(\eta') \leq (1-2\eta)^{2|\alpha|+2} \sum_{\alpha:|\alpha|\leq k} \hat{f}(\alpha)^2 + s(\eta') \\
&\rightarrow (1-2\eta')^{2|\alpha|+2} \leq (1-2\eta)^{2|\alpha|+2} \sum_{\alpha:|\alpha|\leq k} \hat{f}(\alpha)^2 \\
&\rightarrow \eta \leq \eta'. \quad \square
\end{aligned}
$$

Next, we consider the second situation.

We show that the estimation procedure will stop at or before round $r_0 = 2 + \lceil \log_2(1-2\eta)^{-1} \rceil$ with a probability of at least $1-\delta/2$.

In round r, the estimated value η'_r of the noise factor's upper bound becomes $\eta'_r = \frac{1}{2} - \frac{1}{2^{r+1}}$, where we assume that the relationship $\eta'_r \geq \eta$ holds so that Eq. (3) can be used for η'_r and η.

$$
(1-2\eta'_r)^{2|\alpha|+2} \leq (1-2\eta)^{2|\alpha|+2}. \tag{3}
$$

At round $r+1$, the estimated value of the upper bound of the noise factor η'_{r+1} becomes $\eta'_{r+1} = \frac{1}{2} - \frac{1}{2^{r+2}}$, and the hypothetical value of the estimated power spectrum can be derived as $(1-2\eta'_{r+1})^{2|\alpha|+2} = \left(\frac{1}{2}\right)^{2|\alpha|+2}(1-2\eta'_r)^{2|\alpha|+2}$. This is equal to $\left(\frac{1}{2}\right)^{2|\alpha|+2}$ times the left hand side of Eq. (3).

Then, in round $r+1$, it is shown that the hypothetical value $(1-2\eta'_{r+1})^{2|\alpha|+2}$ satisfies the halt condition in the estimation procedure. Consequently, for at least at one frequency $i\,(0 \leq i \leq k)$ where frequency $|\alpha|$ is less than or equal to k, it is shown that inequality (4) related to the stop condition is satisfied. This is because when inequality (4) is satisfied for the upper bound estimate η'_{r+1}, the hypothetical value $(1-2\eta'_{r+1})^{2|\alpha|+2}$ is less than or equal to the empirical value of the power spectrum, $\sum_{\alpha:|\alpha|\leq k} c_\alpha^2 - s(\eta'_{r+1})$, taking errors into consideration.

$$
(1-2\eta'_{r+1})^{2|\alpha|+2} \leq (1-2\eta)^{2|\alpha|+2} \sum_{\alpha:|\alpha|\leq k} \hat{f}(\alpha)^2 - 2s(\eta'_{r+1}). \tag{4}
$$

Since there is at least one frequency for which inequality (4) is satisfied, this shows that the value of formula (5) (obtained by modifying formula (4)) increases to greater than or equal to $2s(\eta'_{r+1})$.

$$
\begin{aligned}
&(1-2\eta)^{2|\alpha|+2} \sum_{\alpha:|\alpha|\leq k} \hat{f}(\alpha)^2 - (1-2\eta'_{r+1})^{2|\alpha|+2} \\
&= (1-2\eta)^{2|\alpha|+2} \sum_{\alpha:|\alpha|\leq k} \hat{f}(\alpha)^2 - (1-2\eta'_r)^{2|\alpha|+2} \left(\frac{1}{2}\right)^{2|\alpha|+2} \\
&= (1-2\eta_r)^{2|\alpha|+2} \left(\sum_{\alpha:|\alpha|\leq k} \hat{f}(\alpha)^2 - \left(\frac{1}{2}\right)^{2|\alpha|+2} \right).
\end{aligned} \tag{5}
$$

Since it is sufficient to show that the relationship holds for at least one frequency, we obtain the sum over all frequencies in the range from 0 to k for formula (5) and $2s(\eta'_{r+1})$ as shown below.

For Eq. (5), the sum is obtained as follows.

$$
\begin{aligned}
&(1-2\eta_r)^{2|\alpha|+2} \left(\sum_{\alpha:|\alpha|\leq k} \hat{f}(\alpha)^2 - \sum_{|\alpha|=0}^{k} \left(\frac{1}{2}\right)^{2|\alpha|+2} \right) \\
&\geq (1-2\eta_r)^{2|\alpha|+2} \left(\frac{3}{4} - \frac{1}{3} \right) = (1-2\eta_r)^{2|\alpha|+2} \left(\frac{5}{12} \right). \\
&\left(\because \begin{cases} \sum_{\alpha:|\alpha|\leq k} \hat{f}(\alpha)^2 \geq 1 - \frac{\epsilon}{4} \geq \frac{3}{4} \\ \sum_{|\alpha|=0}^{k} \left(\frac{1}{2}\right)^{2|\alpha|+2} = \frac{1}{3}\left(1 - \left(\frac{1}{4}\right)^{k+1}\right) < \frac{1}{3} \end{cases} \right).
\end{aligned}
$$

The sum for $2s(\eta'_{r+1})$ is also obtained as follows.

$$
2s(\eta'_{r+1}) \times k \leq 2 \times \frac{(1-2\eta'_{r+1})^{2k+2}}{14n} \times n = (1-2\eta'_r)^{2k+2} \left(\frac{1}{2}\right)^{2k+2} \left(\frac{1}{7}\right).
$$

As a result of obtaining the sum over the range up to k, we can say that the following formula holds for Eq. (5) and $2s(\eta'_{r+1})$; hence, inequality (4) holds for at least one frequency.

$$
(1-2\eta_r)^{2|\alpha|+2} \left(\sum_{\alpha:|\alpha|\leq k} \hat{f}(\alpha)^2 - \sum_{|\alpha|=0}^{k} \left(\frac{1}{2}\right)^{2|\alpha|+2} \right) \geq 2s(\eta'_{r+1}) \times k.
$$

Thus, a sample of size $m_r(\eta'_{r+1})$ is large enough for the following formula to hold with a probability of at least $1-\delta/2$. By subtracting $2s(\eta'_{r+1})$ from both sides, we obtain

$$(1-2\eta)^{2|\alpha|+2}\sum_{\alpha:|\alpha|\le k}\hat{f}(\alpha)^2 \le \sum_{\alpha:|\alpha|\le k} c_\alpha^2 + s(\eta'_{r+1})$$

$$(1-2\eta)^{2|\alpha|+2}\sum_{\alpha:|\alpha|\le k}\hat{f}(\alpha)^2 - 2s(\eta'_{r+1}) \le \sum_{\alpha:|\alpha|\le k} c_\alpha^2 - s(\eta'_{r+1}). \tag{6}$$

By combining inequalities (4) and (6), we can show that the halt condition $(1-2\eta'_{r+1})^{2|\alpha|+2} \le \sum_{\alpha:|\alpha|\le k} c_\alpha^2 - s(\eta'_{r+1})$ is met, thereby ending the estimation procedure.

By fulfilling the halt condition, the estimation procedure stops at or before round $r+1$. Up to round r, the estimated value of the upper bound η' of the noise factor was updated $1+\lceil\log_2(1-2\eta)^{-1}\rceil$ times, so the estimation procedure stops at or before round $r_0 = r+1 = 2+\lceil\log_2(1-2\eta)^{-1}\rceil$. □

Finally, since the two above-mentioned situations related to Theorem 2 both fail with a probability of at most $\delta/2$, the estimation procedure will fail in its estimation with a combined maximum probability of δ.

5.3. Combination of algorithm for learning from noisy examples and estimation procedure

In this section, we first present Theorem 3, which is related to the robust k-lowdegree algorithm [7] (referred to as the rkl algorithm below), and then describe the rkl algorithm itself. The rkl algorithm searches for values related to an unknown noise factor η by incrementing a natural number t in steps of Δt.

A proof of Theorem 3 can be found in the Appendix.

Theorem 3 (*Ohtsuki and Tomita [7]*). *Consider a Boolean function f, values ϵ and δ such that $0<\epsilon,\delta<1$, and a noise factor whose upper bound is η_0 $(0\le\eta\le\eta_0<1/2)$ applied to a learner. The rkl algorithm requests the m examples shown below, and outputs a hypothesis h in time of order $O(mn^k)$ with a probability of at least $1-\delta$ and an error rate of less than or equal to ϵ. Here, $k = O(\log^d(n)/((1-2\eta_0)\epsilon))$.*

$$m = 1568n^{k+2}\left(\frac{1}{1-2\eta_0}\right)^{2k+4}\left(\frac{1}{\epsilon}\right)\ln\left(\frac{2n^k}{\delta}\right). \quad \square$$

Robust k-lowdegree algorithm [7] (ϵ,δ,η_0):

$$\text{Define } m = 1568n^{k+2}\left(1/(1-2\eta_0)\right)^{2k+4}(1/\epsilon)\ln\left((2n^k)/\delta\right),$$
$$\gamma_u = \ln\left(1+\sqrt{\epsilon}/2\right)/(k+1),$$
$$\Delta t = \ln\left[(1-(1-2\eta_0)\gamma_u)^2 + ((1-2\eta_0)\sqrt{\epsilon})/(4n)\right]^{-1}/(4k+4).$$

Step 1. Request noisy examples $\{\langle x^j\oplus x_N^j, f(x^j)y_N^j\rangle\}_{j=1}^m$. For each α such that $|\alpha|\le k$, calculate $c_\alpha = \frac{1}{m}\sum_{j=1}^m f(x^j)y_N^j\chi_\alpha(x^j\oplus x_N^j)$.

Step 2. Initialize $t=1$ and then increase t by Δt until it satisfies the following inequality or until $t\ge\frac{1}{1-2\eta_0}$. Regard such a t as an estimated value of $\frac{1}{1-2\eta}$.

$$\sum_{i=0}^{k}\left(\sum_{|\alpha|=i}c_\alpha^2\right)t^{2i+2} \ge (1-(1-2\eta_0)\gamma_u)^2 + \frac{3(1-2\eta_0)\sqrt{\epsilon}}{28n}.$$

Step 3. Output the hypothesis $h(x) = \text{sign}(\sum_{|\alpha|\le k} c_\alpha t^{|\alpha|+1}\chi_\alpha(x))$ and halt.

In this rkl algorithm [7], we use the value of the noise factor's upper bound η_0 to predetermine a suitable sample size.

By combining the estimation procedure used in this paper with the above rkl algorithm [7], it is possible to obtain a learning algorithm that requires absolutely no prior information about the noise factor. This learning algorithm outputs a hypothesis h with an error rate of no more than ϵ with a probability of at least $1-\delta$. Since the estimation procedure requires a sample whose size is $O(rm)$, its time complexity is

$$O(rmn^k) = O\left(n^{2k+2}\left(\frac{1}{1-2\eta}\right)^{4k+4}\left(\frac{1}{\epsilon}\right)\ln\left(\frac{2n^k}{\delta(1-2\eta)}\right)\right).$$

6. Conclusion

In this paper, we proposed a procedure for estimating the upper bound of an unknown noise factor in the frequency domain. By simply evaluating the state of fluctuation of the frequency components obtained from the noise sample, this procedure can estimate the noise factor's upper bound in a noise model where attribute noise and classification noise are mixed in with a certain fixed probability, without having to evaluate the error rate of the hypothesis obtained from the noise sample. By combining this estimation procedure with the r*k*l algorithm [7], it is possible to obtain a more powerful learning algorithm that does not require any prior information about the noise factor.

Acknowledgements

We are grateful to Kazuhiro Ohtsuki for his great contribution in an early stage of this work.

We express our sincere gratitude to the referees for their careful reading of the previous version of this paper and giving us valuable comments on it.

This research was supported in part by Grants-in-Aid for Scientific Research Nos. 13680435, 16300001, 18500108, 19500010, 20500007, and 22500009 from the Ministry of Education, Culture, Sports, Science and Technology, Japan, and the Research Fund of the University of Electro-Communications. It also received a grant from the Funai Foundation for Information Technology.

Appendix. ([7])

A proof of Theorem 3 is given below.

Let f be a Boolean function that can be computed by a constant-depth, polynomial-size circuit (AC^0 circuit). Moreover, let η $(0 \le \eta < 1/2)$ be an unknown noise factor and let c_α be the empirical value for expectation values $(1-2\eta)^{|\alpha|+1}\hat{f}(\alpha)$.

Assume that the upper bound η_0 $(0 \le \eta \le \eta_0 < 1/2)$ of a known noise factor is applied to the learner. Then, we set $\gamma_c = ((1-2\eta_0)^{k+2}\sqrt{\epsilon})/(28n\sqrt{n^k})$ and $\gamma_u = \ln(1+\sqrt{\epsilon}/2)/(k+1)$, where $k = O(\log^d(n)/((1-2\eta_0)\epsilon))$($d$: depth of the AC^0 circuit).

For expectation values $(1-2\eta)^{|\alpha|+1}\hat{f}(\alpha)$ of Fourier coefficients with bias terms, c_α is the empirical value obtained from the noise sample, and the natural number t is the estimated value of $1/(1-2\eta)$ $(t > 1)$. For an α whose Hamming weight is larger than k, we assume that $\sum_{\alpha:|\alpha|\le k}\hat{f}(\alpha)^2 \simeq 0$, and t is estimated so that $\sum_{i=0}^{k}\left(\sum_{|\alpha|=i} {c_\alpha}^2\right)t^{2i+2} \simeq 1$. By estimating t according to the r*k*l algorithm, we can estimate η because, according to Parseval's equation, $\sum_{\alpha:|\alpha|\le k}\hat{f}(\alpha)^2 \simeq 1$, so this t should be a good approximation of $1/(1-2\eta)$.

Lemma 3. *Assume that the estimated value t satisfies* $|t - 1/(1-2\eta)| \le \gamma_u$ *and that* c_α *satisfies* $|c_\alpha - (1-2\eta)^{|\alpha|+1}\hat{f}(\alpha)| \le \gamma_c$. *Then it is possible to obtain a hypothesis* $h = \mathrm{sign}\,(\Sigma_{|\alpha|\le k} c_\alpha t^{|\alpha|+1}\chi_\alpha(x))$ *whose error rate is less than or equal to* ϵ.

Proof of Lemma 3. Since the value η_0 $(\eta_0 \ge \eta)$ is given, we determine the values of k and γ_c. Thus, we get $\sum_{\alpha:k<|\alpha|}\hat{f}(\alpha)^2 \le ((1-2\eta_0)\epsilon)/(28n)$. Consequently, by evaluating errors between c_α and $\hat{f}(\alpha)$, and considering the fact that $\sum_{\alpha:k<|\alpha|}\hat{f}(\alpha)^2 \le ((1-2\eta_0)\epsilon)/(28n)$, we get Lemma 3. □

Lemma 4. *Assume that* c_α *satisfies the relationship* $|c_\alpha - (1-2\eta)^{|\alpha|+1}\hat{f}(\alpha)| \le \gamma_c$. *When a natural number t satisfies*

$$\sum_{i=0}^{k}\left(\sum_{|\alpha|=i} c_\alpha^2\right)t^{2i+2} \ge (1-(1-2\eta_0)\gamma_u)^2 + \frac{3(1-2\eta_0)\sqrt{\epsilon}}{28n},$$

the relation $|t - 1/(1-2\eta)| \le \gamma_u$ *is established.*

Proof of Lemma 4. Consider the contraposition of Lemma 4. After plugging such a t $(t < 1/(1-2\eta) - \gamma_u)$ into $\sum_{i=0}^{k}(\sum_{|\alpha|=i} c_\alpha^2)t^{2i+2}$, we have

$$\sum_{i=0}^{k}\left(\sum_{|\alpha|=i} c_\alpha^2\right)t^{2i+2} < (1-(1-2\eta_0)\gamma_u)^2 + \frac{3(1-2\eta_0)\sqrt{\epsilon}}{28n}.$$

Hence, we have Lemma 4. Note that we do not consider the case where $1/(1-2\eta) + \gamma_u < t$ because we start searching for t from $t = 1$ $(1/(1-2\eta) \ge 1)$. □

The problem of estimating $1/(1-2\eta)$ thus reduces to the problem of finding t that satisfies the conditions in Lemma 3. To avoid estimating a value of t that radically departs from the true value $1/(1-2\eta)$ related to the noise factor, we assume that t is estimated by increasing its value in increments of Δt, where Δt is expressed as

$$\Delta t = \frac{1}{4k+4} \ln \left[(1-(1-2\eta_0)\gamma_u)^2 + \frac{(1-2\eta_0)\sqrt{\epsilon}}{4n} \right]^{-1}.$$

Moreover, according to the Hoeffding bounds [3], when m examples are provided, as stated below, the estimated value of the Fourier coefficients with bias terms is given by $|c_\alpha - (1-2\eta)^{|\alpha|+1}\hat{f}(\alpha)| \leq \gamma_c$ with a probability of at least $1-\delta$.

$$m = 1568 n^{k+2} \left(\frac{1}{1-2\eta_0} \right)^{2k+4} \left(\frac{1}{\epsilon} \right) \ln \left(\frac{2n^k}{\delta} \right).$$

As a whole, Lemma 3 holds for the above sample size m. This concludes the proof of Theorem 3.

References

[1] N.H. Bshouty, J.C. Jackson, C. Tamon, Uniform-distribution attribute noise learnability, COLT1999, 1999, pp. 75–80.

[2] N.H. Bshouty, J.C. Jackson, C. Tamon, Uniform-distribution attribute noise learnability, Information and Computation 187 (2) (2003) 277–290.

[3] W. Hoeffding, Probability inequalities for sums of bounded random variables, Journal of the American Statistical Association 58 (301) (1963) 13–30.

[4] P.D. Laird, Learning from Good and Bad Data, Kluwer Academic, 1988.

[5] L. Linial, Y. Mansour, N. Nisan, Constant depth circuits, Fourier transform, and learnability, Journal of the Association for Computing Machinery 40 (1993) 607–620.

[6] A. Miyata, J. Tarui, E. Tomita, Learning Boolean functions in AC^0 on attribute and classification noise, in: Proc. ALT 2004, LNAI 3244, 2004, pp. 142–155.

[7] K. Ohtsuki, E. Tomita, An algorithm for learning a certain Boolean function from noisy data, Technical Report of the The University of Electro-Communications, UEC-TR-CAS2, 2000 (in Japanese).

[8] L.G. Valiant, A theory of the learnable, Communication of the Association for Computing Machinery 27 (1984) 1134–1142.