Machine Learning, 18, 23–50 (1995)

Probably Almost Discriminative Learning

KENJI YAMANISHI yamanisi@research.nj.nec.com
NEC Research Institute, Inc., 4 Independence Way, Princeton NJ 08540

Editor: David Haussler

Abstract. This paper develops a new computational model for learning stochastic rules, called PAD (Probably Almost Discriminative)-learning model, based on statistical hypothesis testing theory. The model deals with the problem of designing a discrimination algorithm to test whether or not any given test sequence of examples of pairs of (instance, label) has come from a given stochastic rule P^*. Here a composite hypothesis $\tilde{P}$ is unknown other than it belongs to a given class $\mathcal{C}$.

In this model, we propose a new discrimination algorithm on the basis of the MDL (Minimum Description Length) principle, and then derive upper bounds on the least test sample size required by the algorithm to guarantee that two types of error probabilities are respectively less than δ_1 and δ_2 provided that the distance between the two rules to be discriminated is not less than ε.

For the parametric case where $\mathcal{C}$ is a parametric class, this paper shows that an upper bound on test sample size is given by $O(\frac{1}{\varepsilon}\ln\frac{1}{\delta_1}+\frac{1}{\varepsilon^2}\ln\frac{1}{\delta_2}+\frac{\tilde{k}}{\varepsilon}\ln\frac{\tilde{k}}{\varepsilon}+\frac{\ell(\tilde{M})}{\varepsilon})$. Here $\tilde{k}$ is the number of real-valued parameters for the composite hypothesis $\tilde{P}$, and $\ell(\tilde{M})$ is the description length for the countable model for $\tilde{P}$. Further this paper shows that the MDL-based discrimination algorithm performs well in the sense of sample complexity efficiency, comparing it with other kinds of information-criteria-based discrimination algorithms. This paper also shows how to transform any stochastic PAC (Probably Approximately Correct)-learning algorithm into a PAD-learning algorithm.

For the non-parametric case where $\mathcal{C}$ is a non-parametric class but the discrimination algorithm uses a parametric class, this paper demonstrates that the sample complexity bound for the MDL-based discrimination algorithm is essentially related to Barron and Cover's index of resolvability. The sample complexity bound gives a new view at the relationship between the index of resolvability and the MDL principle from the PAD-learning viewpoint.

Keywords: Computational learning theory, universal hypothesis testing, stochastic rule, PAD-learning, MDL principle

1. Introduction

1.1. Basic problem

The problem of learning stochastic rules has recently come to be widely discussed in computational learning theory (see for example, Kearns & Schapire, 1994, Yamanishi, 1992a, Yamanishi, 1991, Rissanen & Yu, 1991). A stochastic rule here refers to a conditional probability distribution over the set of labels $\mathcal{Y}=\{0,1\}$ given an instance $\mathbf{X}$, and is, for example, expressed as a rule of the form: "if $\mathbf{X}$ makes a Boolean formula $f(\mathbf{X})$ true, then $Y=1$ with probability p_1 and $Y=0$ with probability $1-p_1$, else if ...." In other words, a stochastic rule refers to a rule which probabilistically assigns a number of labels to each instance. Models of learning stochastic rules enable us to deal with learning under noise

An extended abstract appeared in Proceedings of the Fifth ACM Workshop on Computational Learning Theory (Yamanishi, 1992a). A part of this paper appeared in Proceedings of the Sixth ACM Conference on Computational Learning Theory (Yamanishi, 1993).

or such uncertainty as the occurrence of noise and ambiguity induced by lack of relevant attributes in data.

In the community of computational learning theory, several relevant issues concerning stochastic rules have extensively been studied, including "estimation" (Kearns & Schapire, 1994, Yamanishi, 1992a, Rissanen & Yu, 1991) and "on-line prediction" (DeSantis, Markowsky & Wegman, 1988, Yamanishi, 1991, Haussler & Barron, 1992). In addition to these issues, the field of statistical inference studies a third important issue called *hypothesis testing* [or the *discrimination problem* (Hand, 1981)], which deals with the problem of discriminating between two probability distributions, by testing which of them generated a given sequence of test examples. This paper considers this issue from the computational learning aspect.

Let us briefly describe the basic problem which we address in this paper. Let $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ be a countable set which we call a domain and $\mathcal{Y}$ be $\{0,1\}$ which we call a range. Here n is the size of the domain. Let $Q(\mathbf{X})$ denote a probability distribution over $\mathcal{X}$, and both $P^*(Y \mid \mathbf{X})$ and $\tilde{P}(Y \mid \mathbf{X})$ denote stochastic rules. Let $D^m = (\mathbf{X}_1, Y_1) \cdots (\mathbf{X}_m, Y_m)$ be a given sequence of test examples. Here we assume that each $(\mathbf{X}_i, Y_i)$ is independently generated according to $Q(\mathbf{X})P^*(Y \mid \mathbf{X})$ or $Q(\mathbf{X})\tilde{P}(Y \mid \mathbf{X})$. We further suppose that P^* is known, and that Q and $\tilde{P}$ are unknown other than $\tilde{P}$ is assumed to belong to a predetermined class, which we call the target class.

We may then design a discrimination algorithm $\mathcal{A}$ that takes as input a given sequence D^m and a class of stochastic rules, called a hypothesis class, and outputs a decision about whether D^m has originated from QP^* or not. For a given discrimination algorithm $\mathcal{A}$, we define Type 1 error probability for $\mathcal{A}$ as the probability that D^m is generated according to P^* even though $\mathcal{A}$ outputs a decision that D^m has not come from P^*. Similarly we define Type 2 error probability for $\mathcal{A}$ as the probability that D^m is not generated according to P^* although $\mathcal{A}$ outputs a decision that D^m has come from P^*. We wish to design a discrimination algorithm such that both Type 1 and 2 error probabilities approach to zero as fast as possible as sample size increases. This kind of hypothesis testing problem has been called a *universal hypothesis testing problem* (Zeitouni & Gutman, 1991).

In this study, we focus on "computational efficiency," most specifically "sample-size efficiency" for the universal testing problem, and our technical approach may be characterized by "finite-sample-size analysis" of discrimination performance. Speaking more precisely, we address the issue of how large a test sample size and how much computation time are required to guarantee that for a given target class $\mathcal{C}$ and a given hypothesis class $\mathcal{H}$, for some discrimination algorithm using $\mathcal{H}$, for any $0 < \varepsilon < 1$, for any $0 < \delta_1, \delta_2 < 1$, for all $\tilde{P} \in \mathcal{C}$, and for all P^* such that $d(\tilde{P}, P^*) > \varepsilon$, Type 1 and Type 2 error probabilities for the discrimination algorithm are respectively not more than δ_1 and δ_2, where $d(\tilde{P}, P^*)$ denotes a distance between $\tilde{P}$ and P^* (e.g. the Kullback-Leibler divergence). We are then interested in the question of which classes of stochastic rules are PAD (Probably Almost Discriminatively)-learnable in the sense that the sample size and computation time as described above are polynomial in $1/\varepsilon, 1/\delta_1, 1/\delta_2$, and n. This learning framework is inspired by Valiant's PAC-learning model (Valiant, 1984) in the sense that the main concern is to evaluate sample and time complexity required for "probably almost correct" learning.

1.2. Purposes of this paper

This paper has two purposes. The first is to explore a new computational model of learning, which we call PAD-learning model, based on the universal hypothesis testing theory. This model enables us to determine whether any given class of stochastic rules is PAD-learnable in the sense that there exists a polynomial-time algorithm that can discriminate with high probability a known hypothesis from an unknown composite hypothesis by testing from which of the two a given test sequence has originated.

The second purpose is to introduce a discrimination algorithm that performs well within our model and to derive upper bounds on the test sample size required for PAD-learning with the proposed algorithm. We analyze the sample complexity issue for both the "parametric case" and "non-parametric case." Here the parametric case refers to the case where the target class is parametric and is identical to the hypothesis class. The non-parametric case refers to the case where the target class is non-parametric while the hypothesis class is parametric.

1.3. Related work

Related to our problem setting, for the special case where both hypotheses P^* and $\tilde{P}$ are known, it was shown by Hoeffding (Hoeffding, 1965) that the discrimination algorithm based on the likelihood ratio (of P^* to $\tilde{P}$) for test examples is optimal in the sense that Type 2 error probability is minimum over all discrimination algorithms for any fixed Type 1 error probability. In this case the asymptotically best error exponent is given in Stein's lemma (see e.g., Blahut, 1988, Cover & Thomas, 1991), which relates the error exponent to the Kullback-Leibler divergence between P^* and $\tilde{P}$.

The universal testing problem that we consider in this paper has extensively been discussed in the context of information theory (see e.g., Ziv, 1988, Gutman, 1989, Zeitouni & Gutman, 1991). In particular, Ziv proposed a discrimination algorithm based on a universal coding scheme (Ziv, 1988), e.g. Lempel-Ziv universal coding (Ziv & Lempel, 1978), and proved that his proposed discrimination algorithm is asymptotically optimal with respect to the Neyman-Pearson criterion [see e.g. (Hoeffding 1965)], i.e., it maximizes the rate of decrease in Type 2 error probability in the limit under the constraint that the rate of decrease in Type 1 error probability is bounded by a fixed number. However, it has not yet been reported how well his proposed discrimination algorithm works for finite sample size. Further note that his algorithm makes no use of any hypothesis class as we do.

Our own technical approach is unique in this regard in that a discrimination algorithm is designed using a parametric hypothesis class and the *MDL (Minimum Description Length) principle* (Wallace & Boulton, 1968, Schwarz, 1978, Rissanen, 1978, Rissanen, 1987, Rissanen, 1989, Barron & Cover, 1991) rather than universal coding schemes, and that it offers a method of finite test sample analysis. Notice here that while Ziv's analysis concentrated on the issue of asymptotic optimality, we instead consider the issue of how many examples are required to achieve given error probabilities for a given pair of $\tilde{P}$ and P^* such that $d(\tilde{P}, P^*) > \varepsilon$. We are further interested in determining whether any given class is PAD-learnable with sample size polynomial in the size of the domain and

other relevant parameters. We stress that asymptotic optimality does not always imply polynomial-sample-size PAD-learnability.

The application of the MDL principle to hypothesis testing problems was first suggested by Rissanen in (Rissanen, 1987) (pp. 236–238), (Rissanen, 1989), (pp. 109–121). He proposed an MDL-based approach to hypothesis testing for a number of classes of distributions including binomial distributions, gaussian distributions, and two-way contigency tables. He has not yet reported, however, any general theory for finite-sample-size behavior of the MDL-based approach in the universal hypothesis testing setting.

1.4. Summary of results

Let us summarize the results shown in this paper. Let $m_0(\varepsilon, \delta_1, \delta_2, \tilde{P})$ be the test sample size required by the MDL-based discrimination algorithm (for short, the MDL discrimination algorithm) to guarantee that for a given target class $\mathcal{C}$, for an unknown $\tilde{P} \in \mathcal{C}$ and for the known P^* such that $d(\tilde{P}, P^*) > \varepsilon$, Type 1 and 2 error probabilities are respectively at most δ_1 and δ_2, where d is the Kullback-Leibler divergence.

For the parametric case in which the target class $\mathcal{C}$ is a parametric class called a class of stochastic rules with finite partitioning [(Yamanishi, 1992a), each of which takes a form of a piecewise constant conditional probability distribution] with $\sup_{P \in \mathcal{C}} \sup_{\mathbf{X},Y} \{1/P(Y \mid \mathbf{X})\} < \infty$, ignoring time complexity, we give the following upper bound on $m_0(\varepsilon, \delta_1, \delta_2, \tilde{P})$:

$$m_0(\varepsilon, \delta_1, \delta_2, \tilde{P}) = O\left(\frac{1}{\varepsilon} \ln \frac{1}{\delta_1} + \frac{1}{\varepsilon^2} \ln \frac{1}{\delta_2} + \frac{\tilde{k}}{\varepsilon} \ln \frac{\tilde{k}}{\varepsilon} + \frac{\ell(\tilde{M})}{\varepsilon}\right),$$

where $\tilde{k}$ is the number of real-valued parameters in $\tilde{P}$, and $\ell(\tilde{M})$ is the code-length for the countable model specifying $\tilde{P}$. This paper shows that the upper bound on test sample size for the MDL discrimination algorithm is the least reported to date for any information-criteria-based discrimination algorithm, including the maximum likelihood principle-based algorithm.

In addition to the above target-dependent sample size bound, we derive a worst-case bound where the worst-case is taken over the set of all possible $\tilde{P}$ in the given class. Thereby we derive worst-case sample size bounds for PAD-learning of stochastic decision lists (Yamanishi, 1992a, Kearns & Shapire, 1994) with at most s literals in each term (s is fixed) and of stochastic decision trees with depth of at most $s \ln n$ (s is fixed) in order to demonstrate their polynomial-sample-size PAD-learnability.

Further we give a relationship between PAD-learnability and stochastic PAC-learnability. Here the stochastic PAC-learning criterion determines whether any given class of stochastic rules is learnable in the sense that there exists a polynomial-time algorithm that with high probability produces an approximately correct hypothesis from a given training sequence (Kearns & Shapire, 1994, Yamanishi, 1992a). The criterion can be regarded as a stochastic analogue of Valiant's PAC-learning criterion (Valiant, 1984). We show that when given any class of stochastic rules with finite partitioning, if there exists a polynomial-time stochastic PAC-learning algorithm for it, then the class is polynomial-time PAD-learnable

under some conditions. This theorem is proven illustrating how to transform a stochastic PAC learning algorithm to a PAD-learning algorithm.

For the non-parametric case where the non-parametric target class $\mathcal{C}$ satisfies some smoothness conditions and the class of stochastic rules with finite partitioning is employed as a hypothesis class, we give the following upper bound on test sample size $m_0(\varepsilon, \delta_1, \delta_2, \tilde{P})$:

$$m_0(\varepsilon, \delta_1, \delta_2, \tilde{P}) = O\left(\frac{1}{\varepsilon}\ln\frac{1}{\delta_1} + \frac{1}{\varepsilon^2}\ln\frac{1}{\delta_2} + \left(\frac{1}{\varepsilon}\right)^{\frac{\alpha+1}{\alpha}}\ln\frac{1}{\varepsilon}\right),$$

for some $\alpha > 0$. We show that a general test sample size bound for the non-parametric case is essentially related to Barron and Cover's index of resolvability (Barron & Cover, 1991), which is the sample-size dependent measure of the optimal balance between the approximation error of the parametric hypothesis class (to the non-parametric target rule) and the descriptional complexity of the hypothesis class itself.

1.5. Organization of paper

The rest of this paper is organized as follows. Section 2 gives a formal definition of the PAD-learnability criterion. Section 3 reviews a number of notions of stochastic rule learning. Section 4 derives upper bounds on test sample complexity of PAD-learning for the parametric case. Section 5 yields a relationship between PAD-learnability and stochastic PAC-learnability. Section 6 derives upper bounds on test sample complexity for the non-parametric case. Section 7 gives concluding remarks.

2. PAD-learning model

Although the basic outline of the model dealt with in this paper was briefly described in Introduction, this section gives a more precise formal definition of the PAD-learnability criterion. Hereafter all logarithms used are natural logarithms.

For a positive integer n, let $\mathcal{X}_i$ be a measurable space for $i = 1, \ldots, n$. Let $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ be a measurable space which we call a *domain* and $\mathcal{Y} = \{0, 1\}$ be a *range*. We may call n the *size of the domain*. Let $q(\mathbf{X})$ be a probability density function over $\mathcal{X}$ in the case where $\mathcal{X}$ is continuous and let it be a probability mass function over $\mathcal{X}$ in the case where $\mathcal{X}$ is discrete. $Q(\mathbf{X})$ denotes the probability distribution corresponding to $q(\mathbf{X})$. Let $\mathcal{C}_{\text{all}}$ be a set of all stochastic rules defined over $\mathcal{X} \times \mathcal{Y}$. Let $\mathcal{C}$ be a parametric or non-parametric subclass of $\mathcal{C}_{\text{all}}$. Here a *parametric class* is a class in which each rule is specified by a finite number of real-valued parameters, and a *non-parametric class* is a class in which each rule is not specified by any parameter and is constrained only by some conditions such as smoothness or differentiability. Let $P^*(Y \mid \mathbf{X})$ be a stochastic rule belonging to $\mathcal{C}_{\text{all}}$ and $\tilde{P}(Y \mid \mathbf{X})$ be a stochastic rule belonging to $\mathcal{C}$.

We observe the sequence $D^m = D_1 \cdots D_m (D_i = (\mathbf{X}_i, Y_i), i = 1, \ldots, m)$, which we call a *test sequence* (each of which we call a *test example*), and based on the observation, we wish to decide on a correct hypothesis among the two: D^m has originated from $Q(\mathbf{X})P^*(Y \mid$

$\mathbf{X})$ or from $Q(\mathbf{X})\tilde{P}(Y \mid \mathbf{X})$. Here we assume that each $D_i (i = 1, \ldots, m)$ is independently generated from the identical source. We are specifically interested in the situation where P^* is known but Q and $\tilde{P}$ are unknown to us other than that $\tilde{P} \in \mathcal{C}$. Hereafter, according to the convention of statistics, we may call P^* the *null hypothesis* and $\tilde{P}$ the *composite hypothesis*.

Let $\mathcal{H}$ be a parametric class of stochastic rules. Let $(\mathcal{X} \times \mathcal{Y})^*$ denote the set of all finite sequences of elements of $\mathcal{X} \times \mathcal{Y}$, and let $\mathbf{R}$ ($\mathbf{R}^+$) denote the set of real numbers (the set of all positive real numbers). A *discrimination algorithm using* $\mathcal{H}$, which we write as $\mathcal{A}$, is an algorithm that takes as input $P^* \in \mathcal{C}_{\text{all}}, \mathcal{H}, D^m \in (\mathcal{X} \times \mathcal{Y})^*, \varepsilon \in (0, \infty)$ and outputs "+1" or "−1." Here "+1" means the decision that D^m has originated from P^*, and "−1" means the decision that D^m has not originated from P^*. Since any discrimination algorithm $\mathcal{A}$ can also be regarded as a function $\mathcal{C}_{\text{all}} \times \mathcal{H} \times (\mathcal{X} \times \mathcal{Y})^* \times \mathbf{R}^+ \to \{+1, -1\}$, we write $\mathcal{A}(P^*, \mathcal{H}, D^m, \varepsilon) = +1\ (-1)$ when the outputs of $\mathcal{A}$ is "+1" ("−1").

Hereafter we refer to the class $\mathcal{C}$ to which the composite hypothesis belongs as a *target class*. We refer to the class $\mathcal{H}$ which a discrimination algorithm uses as a *hypothesis class*.

For a given discrimination algorithm $\mathcal{A}$, we define *Type 1 error probability* for $\mathcal{A}$ by

$$(QP^*)^m[D^m \colon \mathcal{A}(P^*, \mathcal{H}, D^m, \varepsilon) = -1],$$

where $(QP^*)^m$ denotes the distribution $(Q(\mathbf{X})P^*(Y \mid \mathbf{X}))^m$ over $(\mathcal{X} \times \mathcal{Y})^m$. That is, Type 1 error probability is the probability that the test sequence has originated from P^* but $\mathcal{A}$ determines that it has not been generated from P^*. Further we define *Type 2 error probability* for $\mathcal{A}$ by

$$(Q\tilde{P})^m[D^m \colon \mathcal{A}(P^*, \mathcal{H}, D^m, \varepsilon) = +1],$$

where $(Q\tilde{P})^m$ denotes the distribution $(Q(\mathbf{X})\tilde{P}(Y \mid \mathbf{X}))^m$ over $(\mathcal{X} \times \mathcal{Y})^m$. That is, Type 2 error probability is the probability that the test sequence has originated from $\tilde{P}$ but $\mathcal{A}$ determines that it has been generated from P^*.

Now we are ready to define PAD-learnability.

Definition 1 (*PAD-Learnability*). Let three classes of stochastic rules, $\mathcal{C}, \mathcal{D}$ and $\mathcal{H}$ be given where $\mathcal{H}$ is a parametric class. Let a distance measure d be given. We say that $\mathcal{C}$ is *statistically PAD (Probably Almost Discriminatively)-learnable (with respect to d) in terms of $\mathcal{H}$ under $\mathcal{D}$-constraint*, if there exists a discrimination algorithm $\mathcal{A}$ using $\mathcal{H}$ such that for some polynomial *poly* $(\cdot, \cdot, \cdot, \cdot)$, for all n, for all $\varepsilon > 0$, for all $0 < \delta_1 < 1$, for all $0 < \delta_2 < 1$, for all $m \geq poly(\frac{1}{\varepsilon}, \frac{1}{\delta_1}, \frac{1}{\delta_2}, n)$, for all $q(\mathbf{X})$ on $\mathcal{X}$, for all $\tilde{P}(Y \mid \mathbf{X}) \in \mathcal{C}$, for all $P^*(Y \mid \mathbf{X}) \in \mathcal{D}$ such that $d(\tilde{P}, P^*) > \varepsilon$, Type 1 and Type 2 error probabilities for $\mathcal{A}$ are respectively at most δ_1 and δ_2, i.e.,

$$(QP^*)^m[D^m \colon \mathcal{A}(P^*, \mathcal{H}, D^m, \varepsilon) = -1] \leq \delta_1, \tag{1}$$

$$(Q\tilde{P})^m[D^m \colon \mathcal{A}(P^*, \mathcal{H}, D^m, \varepsilon) = +1] \leq \delta_2. \tag{2}$$

If, in addition, $\mathcal{A}$ runs in time polynomial in $\frac{1}{\varepsilon}, \frac{1}{\delta_1}, \frac{1}{\delta_2}$, and n, then we say that $\mathcal{C}$ is *polynomial-time PAD-learnable (with respect to d) in terms of $\mathcal{H}$ under $\mathcal{D}$-constraint.*

In particular, we say that $\mathcal{C}$ *is (statistically/polynomial-time) PAD-learnable under $\mathcal{D}$-constraint* when $\mathcal{C}$ is (statistically/polynomial-time) PAD-learnable in terms of $\mathcal{C}$ itself under $\mathcal{D}$-constraint.

Hereafter, we use as a distance measure the Kullback-Leibler divergence $d(*,*)$. When $\mathcal{X}$ is continuous, $d(*,*)$ is defined as follows:

$$d(\tilde{P}, P^*) \stackrel{\text{def}}{=} \int d\mathbf{X} q(\mathbf{X}) \sum_{Y \in \mathcal{Y}} \tilde{P}(Y \mid \mathbf{X}) \ln \frac{\tilde{P}(Y \mid \mathbf{X})}{P^*(Y \mid \mathbf{X})}.$$

When $\mathcal{X}$ is discrete, the integral is replaced with the summation over $\mathcal{X}$, and $q(\mathbf{X})$ is considered as a probability mass function.

For a given target class $\mathcal{C}$ and a hypothesis class $\mathcal{H}$, for a given discrimination algorithm $\mathcal{A}$, let $m_{\mathcal{A}}(\varepsilon, \delta_1, \delta_2, n)$ be the least test sample size required by $\mathcal{A}$ to guarantee that Type 1 and 2 error probabilities for $\mathcal{A}$ do not exceed δ_1 and δ_2 respectively for all $\tilde{P} \in \mathcal{C}$, for all $P^* \in \mathcal{D}$ such that $d(\tilde{P}, P^*) > \varepsilon$. We define the *test sample complexity of PAD-learning of $\mathcal{C}$ in terms of $\mathcal{H}$ under $\mathcal{D}$-constraint* as $\min_{\mathcal{A}} m_{\mathcal{A}}(\varepsilon, \delta_1, \delta_2, n)$. Our main concern is here to design a discrimination algorithm that requires the least test sample size.

Hereafter, a *parametric case* refers to the case where $\mathcal{C} = \mathcal{H}$ and $\mathcal{C}$ is a parametric class. A *non-parametric case* refers to the case where $\mathcal{C} \neq \mathcal{H}$ and $\mathcal{C}$ is a non-parametric class.

3. Preliminaries

In order to define the MDL discrimination algorithm in the next section, we need to introduce into our discussion the following two notions with regard to stochastic rule learning: "stochastic rules with finite partitioning" and a "minimum description length." This section briefly reviews them.

Let n be the size of the domain. Let k be a positive integer. Let $\{S_i\}_{i=1,2,\ldots,k}$ be a finite set of disjoint cells of $\mathcal{X}$ (non-empty subsets of $\mathcal{X}$) such that $\cup_{i=1}^{k} S_i = \mathcal{X}$, $S_i \cap S_j = \phi\ (i \neq j)$, and let $p_i \in [0,1] (i = 1, \ldots, k)$ be a real-valued parameter. Let us consider a stochastic rule of the following form:

$$\begin{aligned} &\text{"For any given } \mathbf{X} \in \mathcal{X}, \\ &Y = 1 \text{ (with probability } p_i) \quad \text{and } Y = 0 \text{ (with probability } 1 - p_i), \text{"} \end{aligned} \tag{3}$$

where i denotes the index of the cell into which $\mathbf{X}$ falls. The set of disjoint cells, $\{S_i\}_{i=1,\ldots,k}$, is assumed to be specified by a countable parameter, called a *countable model,* which we denote as M. We call a k-tuple vector $\theta = (p_1, \ldots, p_k)$ a *probability parameter vector.* We denote a stochastic rule specified by θ and M as $P(Y \mid \mathbf{X}\text{: } \theta \prec M)$. A rule of this form is called a *stochastic rule with finite partitioning* (Yamanishi, 1992a).

Letting $\mathcal{M}$ be a finite set of all countable models and $\Theta(M)$ be a set of real-valued parameter vectors associated with a fixed $M \in \mathcal{M}$, we denote a class $\mathcal{C}$ of stochastic rules with finite partitioning as

$$\mathcal{C} = \{P(Y \mid \mathbf{X}\text{: } \theta \prec M)\text{: } M \in \mathcal{M}, \theta \in \Theta(M)\}.$$

Examples of such classes include *stochastic decision lists* and *stochastic decision trees* [see (Yamanishi, 1992a)]. Notice here that $\Theta(M) = [0,1]^k$ when a k-dimensional probability parameter vector is associated with M. We denote the dimension of $\Theta(M)$ as $\dim \Theta(M)$. That is, if $\Theta(M) = [0,1]^k$, then $\dim \Theta(M) = k$.

$\Theta_m(M)$ denotes a set obtained by quantizing $\Theta(M)$ with respect to m so that $\Theta_1(M) \subset \Theta_2(M) \subset \cdots$. In the discussion to follow, we quantize $\Theta(M)$ for each $M \in \mathcal{M}$ with width of $\delta = (\delta_1, \ldots, \delta_k)$ so that every component of any element in $\Theta_m(M)$ is not less than $1/(2m\sqrt{m})$, and δ_j is $\sqrt{(t_j(1-t_j))/2m}$ $(j = 1, \ldots, k)$ where t_j denotes a one-dimensional parameter on $[0, 1]$. It turns out that this gives an optimal quantization scale for minimum-length coding for the sum of quantization scale and data itself [see (Yamanishi, 1992a), p. 180, Eq. (19) for the details].

For a given class $\mathcal{C} = \{P(Y \mid \mathbf{X}: \theta \prec M): M \in \mathcal{M}, \theta \in \Theta(M)\}$, for sample size m, $\mathcal{C}_m$ denotes a class obtained by replacing $\Theta(M)$ with $\Theta_m(M)$, i.e.,

$$\mathcal{C}_m = \{P(Y \mid \mathbf{X}: \theta \prec M): M \in \mathcal{M}, \theta \in \Theta_m(M)\}.$$

Let $D^m = D_1 \cdots D_m (D_i = (\mathbf{X}_i, Y_i), i = 1, \ldots, m)$ be given. Hereafter, we denote $\mathbf{X}_1 \cdots \mathbf{X}_m$ as $\mathbf{X}^m$ and $Y_1 \cdots Y_m$ as Y^m. We define the *minimum description length (MDL) (of Y^m for given $\mathbf{X}^m$) relative to $\mathcal{C}_m$*, denoted as $L_{\mathrm{MDL}}(Y^m \mid \mathbf{X}^m: \mathcal{C}_m)$, by

$$L_{\mathrm{MDL}}(Y^m \mid \mathbf{X}^m: \mathcal{C}_m) \stackrel{\mathrm{def}}{=} \min_{M \in \mathcal{M}} \min_{\theta \in \Theta_m(M)} \{-\ln P(Y^m \mid \mathbf{X}^m: \theta \prec M) + \ell_m(\theta, M)\},$$

where we denote $\prod_{i=1}^m P(Y_i \mid \mathbf{X}_i: \theta \prec M)$ as $P(Y^m \mid \mathbf{X}^m: \theta \prec M)$. ℓ_m is a function $\Theta_m \times \mathcal{M} \to \mathbf{R}^+ \cup \{0\}$ satisfying the following inequality:

$$\sum_{M \in \mathcal{M}} \sum_{\theta \in \Theta_m(M)} e^{-\ell_m(\theta, M)} \leq 1. \tag{4}$$

It is known from (Yamanishi, 1992a) [P. 183, Eq. (25)] that if we use the above method of quantizing $\mathcal{C}$, for any stochastic rule with finite partitioning, $\ell_m(\theta, M)$ is calculated as

$$\ell_m(\theta, M) = \frac{k \ln m}{2} + \frac{(5 \ln 2)k}{2} + \ell(M), \tag{5}$$

where $k = \dim \Theta(M)$, and ℓ is an arbitrary function $\mathcal{M} \to \mathbf{R}^+ \cup \{0\}$ such that $\sum_{M \in \mathcal{M}} e^{-\ell(M)} \leq 1$. Throughout this paper we fix function (5) as $\ell_m(\theta, M)$.

When given a set S, a *code* for S is defined as a mapping from S to a set of all binary sequences, and a *codeword* for s over S is an image of a code for $s \in S$. A code for S is said to be a *prefix code* for S if no codeword is a prefix of any other codeword, and then a codeword for a prefix code is said to be a *prefix codeword* (see e.g. Cover & Thomas, 1991, p. 81). Hereafter, a *code-length* for s (over S) refers to a length of a prefix codeword for s. Although a code-length is usually measured in bits, for the sake of mathematical convenience, we measure it in nats, and allow it to take a non-integer value throughout this paper.

In general it is known (see e.g., Cover & Thomas, 1991, p. 82, Theorem 5.2.1.) that there exists a prefix code such that $\ell: S \to \mathbf{R}^+ \cup \{0\}$ is a code-length assignment function (for short, a *code-length function*) over S if and only if it holds $\sum_{s \in S} e^{-\ell(s)} \leq 1$ (Kraft, 1949), which is called *Kraft's inequality* (over S). Hence, by (4), $\ell_m(\theta, M) = \frac{k \ln m}{2} + \frac{(5 \ln 2)k}{2} + \ell(M)$ can be interpreted as a code-length for (θ, M) over $\cup_{M \in \mathcal{M}} \Theta_m(M) \times \mathcal{M}$ where $\ell(M)$ is a code-length for M over $\mathcal{M}$, and $\frac{k \ln m}{2} + \frac{(5 \ln 2)k}{2}$ is a code-length for θ over

$\Theta_m(M)$. When θ and M are given, following the argument by Shannon (1988, 1948), $-\ln P(Y^m \mid \mathbf{X}^m\colon \theta \prec M)$ is also interpreted as a code-length for Y^m over $\mathcal{Y}^m$ when given $\mathbf{X}^m$ under the assumption that each Y is generated according to $P(Y \mid \mathbf{X}\colon \theta \prec M)$. Thus $L_{\mathrm{MDL}}(Y^m \mid \mathbf{X}^m\colon \mathcal{C}_m)$ can be interpreted as the minimum of the total code-length for Y^m over $\mathcal{Y}^m$ for given $\mathbf{X}^m$ under the condition that only the class $\mathcal{C}_m$ of distributions is known.

A remarkable property of $L_{\mathrm{MDL}}(Y^m \mid \mathbf{X}^m\colon \mathcal{C}_m)$ is that it satisfies Kraft's inequality over $\mathcal{Y}^m$, i.e., for given $\mathbf{X}^m$,

$$\sum_{Y^m \in \mathcal{Y}^m} e^{-L_{\mathrm{MDL}}(Y^m \mid \mathbf{X}^m\colon \mathcal{C}_m)} \leq 1. \tag{6}$$

This holds because $L_{\mathrm{MDL}}(Y^m \mid \mathbf{X}^m\colon \mathcal{C}_m)$ is a code-length for the prefix codeword for Y^m when given $\mathbf{X}^m$.

4. Sample complexity bounds for PAD-learning: Parametric case

This section introduces the MDL discrimination algorithm and analyzes its PAD-learning performance for the parametric case.

4.1. Sample size bounds for PAD-learning: Parametric case

Here is a definition of the MDL discrimination algorithm.

Definition 2 *(MDL Discrimination Algorithm).* Let $\mathcal{H} = \{\, P\,(Y \mid \mathbf{X} : \theta \prec M) : M \in \mathcal{M}, \theta \in \Theta\,(M)\}$ be a class of stochastic rules with finite partitioning. For a sequence of independent test examples $D^m = (\mathbf{X}_1, Y_1) \cdots (\mathbf{X}_m, Y_m)$, let $L_{\mathrm{MDL}}(Y^m \mid \mathbf{X}^m\colon \mathcal{H}_m)$ be the MDL of $Y^m = Y_1 \cdots Y_m$ for given $\mathbf{X}^m = \mathbf{X}_1 \cdots \mathbf{X}_m$ relative to $\mathcal{H}_m$, where $\mathcal{H}_m$ is the class obtained by quantizing $\mathcal{H}$ depending on m (see Section 3). For $P^* \in \mathcal{C}_{\mathrm{all}}$, we denote $\prod_{i=1}^m P^*(Y_i \mid \mathbf{X}_i)$ as $P^*(Y^m \mid \mathbf{X}^m)$. We define a function $h_{\mathrm{MDL}}\colon \mathcal{C}_{\mathrm{all}} \times \mathcal{H} \times (\mathcal{X} \times \mathcal{Y})^* \times \mathbf{R}^+ \to \mathbf{R}$ as follows:

$$h_{\mathrm{MDL}}(P^*, \mathcal{H}, D^m, \varepsilon) = \ln P^*(Y^m \mid \mathbf{X}^m) + L_{\mathrm{MDL}}(Y^m \mid \mathbf{X}^m\colon \mathcal{H}_m) + \frac{m\varepsilon}{2}. \tag{7}$$

The *MDL discrimination algorithm*, which we write as $\mathcal{A}_{\mathrm{MDL}}$, is an algorithm that takes as input $P^*, \mathcal{H}, D^m, \varepsilon$ and outputs "+1" if $h_{\mathrm{MDL}}(P^*, \mathcal{H}, D^m, \varepsilon) > 0$; otherwise "−1."

The following theorem shows hypothesis-dependent and worst-case upper bounds on test sample size for PAD-learning with the MDL discrimination algorithm for the parametric case. Hereafter, for $1 < \gamma < \infty$, we denote as $\mathcal{C}_\gamma$ the set of stochastic rules such that for all $P^* \in \mathcal{C}_\gamma$, $\sup_{\mathbf{X},Y}(1/P^*(Y \mid \mathbf{X})) \leq \gamma$.

Theorem 3 (A) *Hypothesis-Dependent Sample Size Bound. Let* $\mathcal{C} = \{P(Y \mid \mathbf{X}\colon \theta \prec M)\colon M \in \mathcal{M}, \theta \in \Theta(M)\}$ *be a class of stochastic rules with finite partitioning. For any*

n, for any $\varepsilon > 0$, for any $0 < \delta_1, \delta_2 < 1$, for any Q over $\mathcal{X}$, for any $\tilde{P} \in \mathcal{C}$ such that $\tilde{\gamma} \stackrel{\text{def}}{=} \sup_{\mathbf{X}\in\mathcal{X}, Y\in\mathcal{Y}}\{1/\tilde{P}(Y \mid \mathbf{X})\} < \infty$, for any $P^* \in \mathcal{C}_{all}$ such that $d(\tilde{P}, P^*) > \varepsilon$ and $\gamma^* \stackrel{\text{def}}{=} \sup_{\mathbf{X}\in\mathcal{X}, Y\in\mathcal{Y}}\{1/P^*(Y \mid \mathbf{X})\} < \infty$, whenever sample size satisfies

$$m \geq \max\left\{\frac{2}{\varepsilon}\ln\frac{1}{\delta_1}, \frac{8(\ln\tilde{\gamma}\gamma^*)^2}{\varepsilon^2}\ln\frac{1}{\delta_2}, \frac{2e}{\varepsilon(e-1)}\left(\tilde{k}\ln\frac{256\tilde{k}}{\varepsilon} + 2\ell(\tilde{M})\right)\right\}, \tag{8}$$

Type 1 and 2 error probabilities for the MDL discrimination algorithm using $\mathcal{C}$ are then at most δ_1 and δ_2 respectively. Here $\tilde{k} = \dim\Theta(\tilde{M})$, and $\ell(\tilde{M})$ is the code-length for $\tilde{M}$ (the countable model specifying $\tilde{P}$).

(B) *Worst-Case Sample Size Bound. For a class $\mathcal{C}$ of stochastic rules with finite partitioning, assume that $\gamma(\mathcal{C}) \stackrel{\text{def}}{=} \sup_{P\in\mathcal{C}}\sup_{\mathbf{X}\in\mathcal{X}, Y\in\mathcal{Y}}\{1/P(Y \mid \mathbf{X})\} < \infty$. Then for any $1 < \gamma < \infty$, we have the following upper bound on the test sample complexity $m_0(\varepsilon, \delta_1, \delta_2, n)$ of PAD-learning of $\mathcal{C}$ under $\mathcal{C}_\gamma$-constraint.*

$$m_0(\varepsilon, \delta_1, \delta_2, n) = O\left(\frac{1}{\varepsilon}\ln\frac{1}{\delta_1} + \frac{1}{\varepsilon^2}\ln\frac{1}{\delta_2} + \frac{\mu(\mathcal{C})}{\varepsilon}\ln\frac{\mu(\mathcal{C})}{\varepsilon} + \frac{\ln|\mathcal{M}|}{\varepsilon}\right), \tag{9}$$

where $\mu(\mathcal{C}) \stackrel{\text{def}}{=} \max_{M\in\mathcal{M}}\dim\Theta(M)$ is the largest number of real-valued parameters in any rule in $\mathcal{C}$, and $|\mathcal{M}|$ is the total number of countable models for $\mathcal{C}$.

Therefore, for any given class $\mathcal{C}$ of stochastic rules with finite partitioning, for any $1 < \gamma < \infty$, if $\gamma(\mathcal{C}) < \infty$ and both $\mu(\mathcal{C})$ and $\ln|\mathcal{M}|$ are polynomial in n, then $\mathcal{C}$ is statistically PAD-learnable with respect to the Kullback-Leibler divergence under $\mathcal{C}_\gamma$-constraint.

Proof of (A). First we evaluate Type 1 error probability for the MDL discrimination algorithm. Letting $(*)$ be the event that $h_{\mathrm{MDL}}(P^*, \mathcal{C}, D^m, \varepsilon) \leq 0$ (i.e., $P^*(Y^m \mid \mathbf{X}^m) \leq e^{-L_{\mathrm{MDL}}(Y^m|\mathbf{X}^m:\, \mathcal{C}_m) - \frac{m\varepsilon}{2}}$), we have

$$\begin{aligned}
&(QP^*)^m[D^m\colon \mathcal{A}_{\mathrm{MDL}}(P^*, \mathcal{C}, D^m, \varepsilon) = -1] \\
&\quad = (QP^*)^m[D^m\colon h_{\mathrm{MDL}}(P^*, \mathcal{C}, D^m, \varepsilon) \leq 0] \\
&\quad = \sum_{D^m\cdots(*)} (QP^*)(D^m) \\
&\quad \leq \sum_{D^m\cdots(*)} Q(\mathbf{X}^m)e^{-L_{\mathrm{MDL}}(Y^m|\mathbf{X}^m:\, \mathcal{C}_m) - \frac{m\varepsilon}{2}} \\
&\quad \leq e^{-\frac{m\varepsilon}{2}}\sum_{\mathbf{X}^m} Q(\mathbf{X}^m)\sum_{Y^m} e^{-L_{\mathrm{MDL}}(Y^m|\mathbf{X}^m:\, \mathcal{C}_m)} \\
&\quad \leq e^{-\frac{m\varepsilon}{2}}.
\end{aligned}$$

Here we have used the property (6) of the MDL to derive the last inequality. Thus, for $0 < \delta_1 < 1$, the following sample size is sufficient to guarantee that Type 1 error probability is at most δ_1.

$$m \geq \frac{2}{\varepsilon}\ln\frac{1}{\delta_1}. \tag{10}$$

Next we evaluate Type 2 error probability for the MDL discrimination algorithm. Hereafter, for the sake of notational simplicity, we denote $\prod_{i=1}^{m} P^*(Y_i \mid \mathbf{X}_i)$ as $P^*(Y^m \mid \mathbf{X}^m)$ and $P(Y^m \mid \mathbf{X}^m; \tilde{\theta} \prec \tilde{M}) = \prod_{i=1}^{m} P(Y_i \mid \mathbf{X}_i; \tilde{\theta} \prec \tilde{M})$ as $\tilde{P}(Y^m \mid \mathbf{X}^m)$. Here $\tilde{\theta}$ and $\tilde{M}$ respectively denote the probability parameter vector and the countable model for $\tilde{P}$. We have the following inequalities for Type 2 error probability: if $d(\tilde{P}, P^*) > \varepsilon$, then

$$
\begin{aligned}
&(Q\tilde{P})^m[D^m: \mathcal{A}_{\mathrm{MDL}}(P^*, \mathcal{C}, D^m, \varepsilon) = +1] \\
&\quad = (Q\tilde{P})^m[D^m: h_{\mathrm{MDL}}(P^*, \mathcal{C}, D^m, \varepsilon) > 0] \\
&\quad = (Q\tilde{P})^m[D^m: \ln P^*(Y^m \mid \mathbf{X}^m) + L_{\mathrm{MDL}}(Y^m \mid \mathbf{X}^m; \mathcal{C}_m) + \frac{m\varepsilon}{2} > 0] \\
&\quad = (Q\tilde{P})^m \left[\frac{1}{m} \ln \frac{P^*(Y^m \mid \mathbf{X}^m)}{\tilde{P}(Y^m \mid \mathbf{X}^m)} + d(\tilde{P}, P^*) \right. \\
&\qquad \left. + \frac{1}{m}(\ln \tilde{P}(Y^m \mid \mathbf{X}^m) + L_{\mathrm{MDL}}(Y^m \mid \mathbf{X}^m; \mathcal{C}_m)) > d(\tilde{P}, P^*) - \frac{\varepsilon}{2} \right] \\
&\quad \leq (Q\tilde{P})^m \left[\frac{1}{m} \ln \frac{P^*(Y^m \mid \mathbf{X}^m)}{\tilde{P}(Y^m \mid \mathbf{X}^m)} + d(\tilde{P}, P^*) > \frac{d(\tilde{P}, P^*) - \varepsilon/2}{2} \right] \qquad (11) \\
&\qquad + (Q\tilde{P})^m \left[\frac{1}{m}(\ln \tilde{P}(Y^m \mid \mathbf{X}^m) + L_{\mathrm{MDL}}(Y^m \mid \mathbf{X}^m; \mathcal{C}_m)) \right. \\
&\qquad \left. > \frac{d(\tilde{P}, P^*) - \varepsilon/2}{2} \right] \\
&\quad \leq (Q\tilde{P})^m \left[\frac{1}{m} \ln \frac{P^*(Y^m \mid \mathbf{X}^m)}{\tilde{P}(Y^m \mid \mathbf{X}^m)} + d(\tilde{P}, P^*) > \frac{\varepsilon}{4} \right] \qquad (12) \\
&\qquad + (Q\tilde{P})^m \left[\frac{1}{m}(\ln \tilde{P}(Y^m \mid \mathbf{X}^m) + L_{\mathrm{MDL}}(Y^m \mid \mathbf{X}^m; \mathcal{C}_m)) > \frac{\varepsilon}{4} \right]. \qquad (13)
\end{aligned}
$$

Here in order to derive inequality (11) we have used the general inequality: $Prob[A + B > \varepsilon] \leq Prob[A > \frac{\varepsilon}{2}] + Prob[B > \frac{\varepsilon}{2}]$.

To further evaluate the probability (12), we prepare the following lemma.

Lemma 4. *Letting* $\tilde{\gamma} = \sup_{\mathbf{X} \in \mathcal{X}, Y \in \mathcal{Y}} \{1/\tilde{P}(Y \mid \mathbf{X})\} < \infty$ *and* $\gamma = \sup_{\mathbf{X} \in \mathcal{X}, Y \in \mathcal{Y}} \{1/P^*(Y \mid \mathbf{X})\} < \infty$, *for* $0 < \delta_2 < 1$, *the probability* (12) *is at most* δ_2 *for all sample size satisfying*

$$
m \geq \frac{8(\ln \tilde{\gamma}\gamma^*)^2}{\varepsilon^2} \ln \frac{1}{\delta_2}. \qquad (14)
$$

Proof of Lemma 4. We prepare the following sublemma in order to prove Lemma 4.

Sublemma 5 (*Hoeffding 1963*). *Letting* $Z_1, \ldots, Z_m$ *be independent random variables with bounded ranges:* $a \leq Z_i \leq b$, *for each* $\eta > 0$, *we have*

$$
Prob\left[\frac{1}{m} \sum_{i=1}^{m} Z_i - E[Z] > \eta \right] \leq e^{-\frac{2\eta^2 m}{(b-a)^2}}.
$$

Let $E_{Q,\tilde{P}}$ be the expectation taken with respect to $Q(\mathbf{X})\tilde{P}(Y \mid \mathbf{X})$. Since $-\ln\tilde{\gamma} < \ln\frac{\tilde{P}(Y|\mathbf{X})}{P^*(Y|\mathbf{X})} < \ln\gamma^*$ for all $\mathbf{X} \in \mathcal{X}$ and $Y \in \mathcal{Y}$, and $E_{Q,\tilde{P}}[\ln\frac{\tilde{P}(Y|\mathbf{X})}{P^*(Y|\mathbf{X})}] = d(\tilde{P}, P^*)$, we have the following inequality using Hoeffding's inequality.

$$
\begin{aligned}
&(Q\tilde{P})^m\left[\frac{1}{m}\ln\frac{P^*(Y^m \mid \mathbf{X}^m)}{\tilde{P}(Y^m \mid \mathbf{X}^m)} + d(\tilde{P}, P^*) > \frac{\varepsilon}{4}\right] \\
&\quad = (Q\tilde{P})^m\left[-\frac{1}{m}\sum_{i=1}^{m}\ln\frac{\tilde{P}(Y_i \mid \mathbf{X}_i)}{P^*(Y_i \mid \mathbf{X}_i)} - E_{Q,P^*}\left[-\ln\frac{\tilde{P}(Y \mid \mathbf{X})}{P^*(Y \mid \mathbf{X})}\right] > \frac{\varepsilon}{4}\right] \\
&\quad \le e^{-\frac{m\varepsilon^2}{8(\ln\tilde{\gamma}\gamma^*)^2}}. \qquad (15)
\end{aligned}
$$

Letting the righthand side of (15) be δ_2 and solving it for m_2 yield (14). This completes the proof of Lemma 4. □

To evaluate the probability (13), we prepare the following lemma.

Lemma 6. *The probability* (13) *is* 0 *for all sample size satisfying*

$$
m \ge \frac{2e}{\varepsilon(e-1)}\left(\tilde{k}\ln\frac{256\tilde{k}}{\varepsilon} + 2\ell(\tilde{M})\right). \qquad (16)
$$

Proof of Lemma 6. We define a maximum likelihood estimator $\hat{\theta}$ for a fixed $\tilde{M}$ as $\hat{\theta} \stackrel{\text{def}}{=} \arg\max_{\theta\in\Theta(\tilde{M})} P(Y^m \mid \mathbf{X}^m\colon \theta \prec \tilde{M})$. We denote the truncated vector (in $\Theta_m(\tilde{M})$) of $\hat{\theta}$ as $\bar{\theta}$. From the definition of the MDL and (5), we have the following inequalities:

$$
\begin{aligned}
&L_{\mathrm{MDL}}(Y^m \mid \mathbf{X}^m\colon C_m) \\
&\quad \le -\ln P(Y^m \mid \mathbf{X}^m\colon \bar{\theta} \prec \tilde{M}) + \ell_m(\bar{\theta}, \tilde{M}) \\
&\quad < -\ln P(Y^m \mid \mathbf{X}^m\colon \hat{\theta} \prec \tilde{M}) + \ell_m(\bar{\theta}, \tilde{M}) + \tilde{k}\ln 2 \qquad (17) \\
&\quad = -\ln P(Y^m \mid \mathbf{X}^m\colon \tilde{\theta} \prec \tilde{M}) + \frac{\tilde{k}\ln m}{2} + \frac{(5\ln 2)\tilde{k}}{2} + \ell(\tilde{M}) + \tilde{k}\ln 2,
\end{aligned}
$$

where $\tilde{k} = \dim\Theta(\tilde{M})$, and $\ell(\tilde{M})$ is the code-length for $\tilde{M}$ over $\mathcal{M}$. To derive (17) we have used the following general relationship between the likelihood for $\hat{\theta}$ and that for $\bar{\theta}$ with regard to stochastic rules with finite partitioning (see Yamanishi, 1992a, p. 184, Eq. (27)):

$$
-\ln\tilde{P}(Y^m \mid \mathbf{X}^m\colon \bar{\theta} \prec \tilde{M}) < -\ln\tilde{P}(Y^m \mid \mathbf{X}^m\colon \hat{\theta} \prec \tilde{M}) + \tilde{k}\ln 2.
$$

Hence we obtain

$$
\ln\tilde{P}(Y^m \mid \mathbf{X}^m) + L_{\mathrm{MDL}}(Y^m \mid \mathbf{X}^m\colon C_m) < \frac{\tilde{k}\ln m}{2} + \frac{(5\ln 2)\tilde{k}}{2} + \tilde{k}\ln 2 + \ell(\tilde{M}). \qquad (18)
$$

Thus, if m satisfies

$$
\frac{1}{m}\left(\frac{\tilde{k}\ln m}{2} + \frac{(5\ln 2)\tilde{k}}{2} + \tilde{k}\ln 2 + \ell(\tilde{M})\right) \le \frac{\varepsilon}{4}, \qquad (19)
$$

then

$$(Q\tilde{P})^m\left[\frac{1}{m}(\ln \tilde{P}(Y^m \mid \mathbf{X}^m) + L_{\mathrm{MDL}}(Y^m \mid \mathbf{X}^m\colon \mathcal{C}_m)) > \frac{\varepsilon}{4}\right] = 0.$$

(19) yields the following equivalent expression:

$$m \geq \frac{2\tilde{k}\ln m}{\varepsilon} + \frac{(14\ln 2)\tilde{k}}{\varepsilon} + \frac{4\ell(\tilde{M})}{\varepsilon}. \tag{20}$$

Notice here that the following general inequality holds for $x, y \in \mathbf{R}^+$ [see, (Haussler & Long, 1990) for the proof].

$$\ln x \leq xy - \ln ey.$$

Thus, for any $0 < \nu < 1$, $\ln m \leq \frac{\nu\varepsilon m}{2\tilde{k}} + \ln \frac{2\tilde{k}}{\nu\varepsilon e}$. Hence the following sample size is sufficient to guarantee (20).

$$m \geq \frac{2\tilde{k}}{\varepsilon}\left(\frac{\nu\varepsilon m}{2\tilde{k}} + \ln\frac{2\tilde{k}}{\nu\varepsilon e}\right) + \frac{(14\ln 2)\tilde{k}}{\varepsilon} + \frac{4\ell(\tilde{M})}{\varepsilon}.$$

Letting $\nu = 1/e$ and solving this for m, we see that

$$m \geq \frac{2e}{\varepsilon(e-1)}\left(\tilde{k}\ln\frac{256\tilde{k}}{\varepsilon} + 2\ell(\tilde{M})\right).$$

suffices to guarantee that (19) holds. This completes the proof of Lemma 6. □

From (10), (14), and (16), we see that sample size satisfying

$$m \geq \max\left\{\frac{2}{\varepsilon}\ln\frac{1}{\delta_1}, \frac{8(\ln\tilde{\gamma}\gamma^*)^2}{\varepsilon^2}\ln\frac{1}{\delta_2}, \frac{2e}{\varepsilon(e-1)}\left(\tilde{k}\ln\frac{256\tilde{k}}{\varepsilon} + 2\ell(\tilde{M})\right)\right\}$$

is sufficient to guarantee that Type 1 and 2 error probabilities are at most δ_1 and δ_2 respectively if $d(\tilde{P}, P^*) > \varepsilon$. This completes the proof of (A) in Theorem 3. □

Proof of (B). First note that the righthand side of (8) depends on $\tilde{P}$ and P^* but does not depend on Q. Applying the worst-case analysis, we see that the following two inequalities must be satisfied in order to guarantee that for all Q over $\mathcal{X}$, for all $\tilde{P} \in \mathcal{C}$, for all $P^* \in \mathcal{C}_\gamma$ such that $d(\tilde{P}, P^*) > \varepsilon$, Type 1 and 2 error probabilities are respectively not larger than δ_1 and δ_2.

$$\sup_{P^*\in\mathcal{C}_\gamma}(QP^*)^m[D^m\colon h_{MDL}(P^*, \mathcal{C}, D^m, \varepsilon) \leq 0\,] \leq \delta_1, \tag{21}$$

$$\sup_{P^*\in\mathcal{C}_\gamma}\sup_{\tilde{P}\in\mathcal{C}, d(\tilde{P},P^*)>\varepsilon}(Q\tilde{P})^m[D^m\colon h_{MDL}(P^*, \mathcal{C}, D^m, \varepsilon) > 0\,] \leq \delta_2. \tag{22}$$

Notice that the righthand side of (8) is a monotone increasing function with respect to $\tilde{\gamma}, \gamma^*$, and $\tilde{k}$. Thus, by replacing $\tilde{\gamma}, \gamma^*, \tilde{k}$ in (8) with $\gamma(\mathcal{C}), \gamma, \mu(\mathcal{C})$ respectively, we can prove that (21) is satisfied for

$$m \geq \frac{2}{\varepsilon} \ln \frac{1}{\delta_1},$$

and that (22) is satisfied for all sample size satisfying

$$m \geq \max \left\{ \frac{8(\ln \gamma(\mathcal{C})\gamma)^2}{\varepsilon^2} \ln \frac{1}{\delta_2}, \frac{2e}{\varepsilon(e-1)} \left(\mu(\mathcal{C}) \ln \frac{256\mu(\mathcal{C})}{\varepsilon} + 2 \ln |\mathcal{M}| \right) \right\}, \quad (23)$$

where we have used the following code-length function over $\mathcal{M}$:

$$\ell(\tilde{M}) = \ln |\mathcal{M}|$$

for all $\tilde{M} \in \mathcal{M}$. Hence (23) yields (9). It immediately follows from (9) that for any $1 < \gamma < \infty$, if $\gamma(\mathcal{C}) < \infty$ and both $\mu(\mathcal{C})$ and $\ln |\mathcal{M}|$ are polynomial in n, then $\mathcal{C}$ is statistically PAD-learnable with respect to the Kullback-Leibler divergence under $\mathcal{C}_\gamma$-constraint. This completes the proof of (B) in Theorem 3. □

In the proof of Theorem 3, we have used two notable properties of the MDL. One is that the MDL satisfies Kraft's inequality (see (6)), which we have used to derive (10). The other is that the MDL determines the minimum of the total code-length for test examples over the hypothesis class, which we have used to derive (16).

From bound (8) we see that the order of sample complexity (with respect to either $\frac{1}{\varepsilon}, \frac{1}{\delta_1}$, $\frac{1}{\delta_2}, \tilde{k}$, or $\ell(\tilde{M})$) is at most a square of the parameter.

As corollaries of Theorem 3, we have results on PAD-learnability of specific classes of stochastic rules with finite partitioning.

Corollary 7. *Let $\mathcal{X} = \{0,1\}^n$ and $\mathcal{Y} = \{0,1\}$. Consider classes of stochastic decision lists (see Yamanishi, 1992a, Kearns & Shapire, 1994). Let a positive integer $1 \leq s \leq n$ be given. We denote the set of all terms with at most s-literals as T_s^n. Letting $t_1, \ldots, t_k \in T_s^n, (1 \leq k \leq |T_s^n|)$, each stochastic decision list (with at most s literals in each term) is defined as a stochastic rule of the form: $(t_1, p_1) \cdots (t_{k-1}, p_{k-1}), (\mathbf{true}, p_k)$, with the following semantics: for any given $\mathbf{X} \in \mathcal{X}$, $Y = 1$ with probability p_i and $Y = 0$ with probability $1 - p_i$ where i is the least index such that $\mathbf{X}$ makes t_i **true.** We denote the class of stochastic decision lists (with at most s literals in each term) as $\mathcal{C}_{DL}^s$. Let Γ_s^n be the set of all countable models specifying $\mathcal{C}_{DL}^s$. That is, each finite partitioning specifying a stochastic decision list is parametrized by an element in Γ_s^n. Then $\mathcal{C}_{DL}^s$ can be written as $\mathcal{C}_{DL}^s = \{P(Y \mid \mathbf{X}: \theta \prec M): M \in \Gamma_s^n, \theta \in \Theta(M)\}$, where if $\dim \Theta(M) = k$, then $\theta = (p_1, \ldots, p_k)$ and $\Theta(M) = [0,1]^k$. When we wish to emphasize the number of attributes, n, we will indicate this in parentheses after the class name, as in $\mathcal{C}_{DL}^s(n)$.*

For given $0 < \nu < 1$, let $\mathcal{C}_{DL}^{s,\nu}$ be a subclass of $\mathcal{C}_{DL}^s$ such that for each probability parameter vector $\theta = (p_1, \ldots, p_k)$, $\nu \leq p_i \leq 1 - \nu (i = 1, \ldots, k)$. Then for fixed s and ν, for any $1 < \gamma < \infty$, $\mathcal{C}_{DL}^{s,\nu}$ is statistically PAD-learnable with respect to the Kullback-Leibler

divergence under $\mathcal{C}_\gamma$-constraint. We have the following worst-case upper bound on the test sample complexity $m_0(\varepsilon, \delta_1, \delta_2, n)$ for PAD-learning of $\mathcal{C}_{DL}^{s,\nu}(n)$:

$$m_0(\varepsilon, \delta_1, \delta_2, n) = O\left(\frac{1}{\varepsilon^2} \ln \frac{1}{\delta_1} + \frac{1}{\varepsilon^2} \ln \frac{1}{\delta_2} + \frac{n^s}{\varepsilon} \ln \frac{n}{\varepsilon}\right). \tag{24}$$

Proof. (24) is immediately obtained from (9) and the facts that $\mu(\mathcal{C}_{\mathrm{DL}}^{s,\nu}(n)) = O(n^s)$ and $O(\ln |\Gamma_s^n|) = O(n^s \ln n)$ [see (Yamanishi, 1992a)]. It follows from (24) that $m_0(\varepsilon, \delta_1, \delta_2, n)$ is polynomial in $\frac{1}{\varepsilon}, \frac{1}{\delta_1}, \frac{1}{\delta_2}$, and n. Thus for fixed s and ν, for any $1 < \gamma < \infty$, $\mathcal{C}_{\mathrm{DL}}^{s,\nu}$ is PAD-learnable with respect to the Kullback-Leibler divergence under $\mathcal{C}_\gamma$-constraint. This completes the proof of Corollary 7. □

Corollary 8. *Let $\mathcal{X} = \{0,1\}^n$ and $\mathcal{Y} = \{0,1\}$. Let s be a fixed positive integer. Let $\Omega_{DT}^{s \ln n}(n)$ be the set of all countable models each of which specifies finite partitioning for a stochastic decision tree (see Yamanishi, 1992a) (with at most $s \ln n$ depth). We denote the class of stochastic decision trees with at most $s \ln n$ depth as $\mathcal{C}_{DT}^{s \ln n} = \{P(Y \mid \mathbf{X}\text{: } \theta \prec M)\text{: } M \in \Omega_{DT}^{s \ln n}(n), \theta \in \Theta(M)\}$. Assume that all leaves of a decision tree is appropriately indexed. When $P(Y \mid \mathbf{X}\text{: } \theta \prec M)$ denotes a stochastic decision tree with k leaves, for $\theta = (p_1, \ldots, p_k) \in \Theta(M)$, each p_i denotes the probability that $Y = 1$ for any $\mathbf{X}$ that reaches the ith leaf. That is, a stochastic decision tree is a stochastic rule which has the following semantics: for any given $\mathbf{X} \in \mathcal{X}$, $Y = 1$ with probability p_i and $Y = 0$ with probability $1 - p_i$ where i is the index of the leaf that $\mathbf{X}$ reaches. When we wish to emphasize the number of attributes, n, we will indicate this in parentheses after the class name, as in $\mathcal{C}_{DT}^{s \ln n}(n)$.*

For given $0 < \nu < 1$, let $\mathcal{C}_{DT}^{s \ln n, \nu}$ be a subclass of $\mathcal{C}_{DT}^{s \ln n}$ such that for each probability parameter vector $\theta = (p_1, \ldots, p_k)$, $\nu \leq p_i \leq 1 - \nu, (i = 1, \ldots, k)$. Then for fixed s and ν, for any $1 < \gamma < \infty$, $\mathcal{C}_{DT}^{s \ln n, \nu}$ is statistically PAD-learnable with respect to the Kullback-Leibler divergence under $\mathcal{C}_\gamma$-constraint. We have the following worst-case upper bound on the test sample complexity $m_0(\varepsilon, \delta_1, \delta_2, n)$ for PAD-learning of $\mathcal{C}_{DT}^{s \ln n, \nu}(n)$:

$$m_0(\varepsilon, \delta_1, \delta_2, n) = O\left(\frac{1}{\varepsilon^2} \ln \frac{1}{\delta_1} + \frac{1}{\varepsilon^2} \ln \frac{1}{\delta_2} + \frac{n^s \ln n}{\varepsilon} \ln \frac{n}{\varepsilon}\right). \tag{25}$$

Proof. (25) can be immediately obtained from (9) and the facts that $\mu(\mathcal{C}_{\mathrm{DT}}^{s \ln n, \nu}(n)) = O(n^s)$ and $O(\ln \mid \Omega_{\mathrm{DT}}^{s \ln n}(n)|) = 2^{s \ln n - 1} \ln(n+1) \cdots (n - s \ln n + 1) = O(n^s (\ln n)^2)$. It follows from (25) that $m_0(\varepsilon, \delta_1, \delta_2, n)$ is polynomial in $\frac{1}{\varepsilon}, \frac{1}{\delta_1}, \frac{1}{\delta_2}$, and n. Thus for fixed s and ν, for any $1 < \gamma < \infty$, $\mathcal{C}_{\mathrm{DT}}^{s \ln n, \nu}$ is statistically PAD-learnable with respect to the Kullback-Leibler divergence under $\mathcal{C}_\gamma$-constraint. This completes the proof of Corollary 8. □

4.2. *Comparison of MDL discrimination with other information-criteria-based discrimination*

To see how well the MDL discrimination algorithm performs within the PAD-learning model, let us consider a more general family of information-criteria-based discrimination algorithms and compare their discriminative performance with that of the MDL discrimination algorithm. Here an *information-criteria-based discrimination algorithm* is a discrimination algorithm $\mathcal{A}$ that takes as input $P^*, \mathcal{H}, D^m, \varepsilon$, and outputs "+1" if the following $h(P^*, \mathcal{C}, D^m, \varepsilon)$ is positive; otherwise "−1."

$$h(P^*, \mathcal{C}, D^m, \varepsilon) = \ln P^*(Y^m \mid \mathbf{X}^m) + \min_{M \in \mathcal{M}} \min_{\theta \in \Theta_m(M)} \{-\ln P(Y^m \mid \mathbf{X}^m : \theta \prec M) + f_m(\theta, M)\}, \quad (26)$$

where $f_m(\theta, M)$ is a function of $\theta \in \Theta_m(M)$ and M, depending on m.

We evaluate the discrimination performance of the general information-criteria-based discrimination algorithm in terms of the sample size bound required for (1) and (2) to be satisfied for given ε, δ_1, and δ_2. We consider the following three cases.

Case 1. $f_m(\theta, M) = O(\ln^\alpha m)(0 \leq \alpha < 1)$.

Specifically, if we let $f_m(\theta, M) = 0$, the second term of the righthand side of (26) can be written as $\min_{M \in \mathcal{M}} \min_{\theta \in \Theta_m(M)} \{-\ln P(Y^m \mid \mathbf{X}^m : \theta \prec M)\} = -\ln \max_{M \in \mathcal{M}} \max_{\theta \in \Theta_m(M)} P(Y^m \mid \mathbf{X}^m : \theta \prec M)$, which is a logarithm of the maximum likelihood for Y^m for given $\mathbf{X}^m$ relative to $\mathcal{C}_m$.

First let us investigate the case where $f_m(\theta, M) = C \ln^\alpha m + g(M)$ for some $C > 0$ and where for some code-length function ℓ satisfying Kraft's inequality over $\mathcal{M}$, $g(M) \geq \ell(M)$ for all M. In Case 1, we cannot any longer apply Kraft's inequality in Type 1 error probability evaluation as with the derivation of (10) in the proof of Theorem 3. It can be easily verified that Type 1 error probability is at most δ_1 for sample size $O(\frac{1}{\varepsilon} \ln \frac{1}{\delta_1} + \frac{\mu(\mathcal{C})}{\varepsilon} \ln \frac{\mu(\mathcal{C})}{\varepsilon})$, where $\mu(\mathcal{C})$ is the largest number of parameters in any rule in $\mathcal{C}$. Using the same type of proof technique as in Theorem 3, it is easily verified that the least sample size required for Type 2 error probability to be at most δ_2 is $O(\frac{1}{\varepsilon^2} \ln \frac{1}{\delta_2} + \frac{\tilde{k}}{\varepsilon} \ln^\alpha (\frac{\tilde{k}}{\varepsilon}) + \frac{g(\tilde{M})}{\varepsilon})$. Hence we obtain the following upper bound on the test sample size:

$$O\left(\frac{1}{\varepsilon} \ln \frac{1}{\delta_1} + \frac{1}{\varepsilon^2} \ln \frac{1}{\delta_2} + \frac{\mu(\mathcal{C})}{\varepsilon} \ln \frac{\mu(\mathcal{C})}{\varepsilon} + \frac{g(\tilde{M})}{\varepsilon}\right), \quad (27)$$

which is larger than (8) except in the case where $\tilde{P}$ has the largest number of parameters in $\mathcal{C}$.

If we consider the case where $f_m(\theta, M) = C \ln^\alpha m + g(M)$ and where there exists no code-length function ℓ over $\mathcal{M}$ such that $g(M) \geq \ell(M)$ for all M, then the fourth term in (27) may be replaced with the term larger than $\frac{g(\tilde{M})}{\varepsilon}$ and is at most $\frac{\ln|\mathcal{M}|}{\varepsilon}$. This still yields a bound of higher order than (8).

Case 2. $f_m(\theta, M) = O(\ln^\alpha m)(\alpha = 1)$.

In this case, letting $f_m(\theta, M) = \frac{k \ln m}{2} + \frac{(5 \ln 2)k}{2} + \ell(M)$, we have the MDL discrimination function. Then we have the best upper bound (8) on test sample size in this case.

Case 3. $f_m(\theta, M) = O(\ln^\alpha m)(\alpha > 1)$.

Specifically we consider the case where $f_m(\theta, M) = C \ln^\alpha m + g(M)$ for some $C > 0$ and where for some code-length function ℓ satisfying Kraft's inequality over $\mathcal{M}$, $g(M) \geq \ell(M)$ for all $M \in \mathcal{M}$. In this case we can apply Kraft's inequality in Type 1 error probability evaluation as with the derivation of (10), and thus the least sample size required for Type 1 error probability to be at most δ_1 is $O(\frac{1}{\varepsilon} \ln \frac{1}{\delta_1})$. On the other hand, the least sample size required to guarantee that the probability corresponding to (13) becomes zero is $O(\frac{\tilde{k}}{\varepsilon} \ln^\alpha(\frac{\tilde{k}}{\varepsilon}) + \frac{g(\tilde{M})}{\varepsilon})$, and thus the least sample size required for Type 2 error probability to be at most δ_2 is $O(\frac{1}{\varepsilon^2} \ln \frac{1}{\delta_2} + \frac{\tilde{k}}{\varepsilon} \ln^\alpha(\frac{\tilde{k}}{\varepsilon}) + \frac{g(\tilde{M})}{\varepsilon})$. Hence, using the same type of proof technique as in Theorem 3, we obtain the following upper bound on the test sample size:

$$O\left(\frac{1}{\varepsilon} \ln \frac{1}{\delta_1} + \frac{1}{\varepsilon^2} \ln \frac{1}{\delta_2} + \frac{\tilde{k}}{\varepsilon} \ln^\alpha \left(\frac{\tilde{k}}{\varepsilon}\right) + \frac{g(\tilde{M})}{\varepsilon}\right), \tag{28}$$

which is of higher order with respect to $\frac{1}{\varepsilon}$ and $\tilde{k}$ than (8) since α is larger than 1.

If we consider the case where $f_m(\theta, M) = C \ln^\alpha m + g(M)$ and where there exists no code-length function ℓ over $\mathcal{M}$ such that $g(M) \geq \ell(M)$ for all M, then the fourth term in (27) may be replaced with the term larger than $\frac{\ell(\tilde{M})}{\varepsilon}$ and is at most $\frac{\ln|\mathcal{M}|}{\varepsilon}$. This still yields a bound of higher order than (8).

From the above comparison of the upper bounds on test sample size, we can say that the upper bound on the test sample size given for the MDL discrimination algorithm is the least reported to date for any information-criteria-based discrimination algorithm. Although this is a comparison of upper bounds and any lower bounds to be compared have not yet been obtained, this analysis shows that the MDL principle effectively works in the universal hypothesis testing problem.

5. Relationship between PAD-learnability and stochastic PAC-learnability

In this section we give a theorem relating PAD-learnability to "stochastic PAC-learnability." Before describing the theorem, let us review the definition of stochastic PAC-learning algorithms. The following definition follows (Yamanishi, 1992a).

Definition 9 (*Stochastic PAC-Learning Algorithm*). Let a class $\mathcal{C}$ of stochastic rules and a distance measure d be given. We say that $\mathcal{A}$ is a *stochastic PAC-learning algorithm for $\mathcal{C}$ (with respect to d)* if for some polynomial *poly*$(*, *, *)$, for all $\varepsilon > 0$, for all $0 < \delta < 1$, for all Q over $\mathcal{X}$, for all $P^* \in \mathcal{C}$, for all positive integer $m \geq$ *poly*$(\frac{1}{\varepsilon}, \frac{1}{\delta}, n)$, $\mathcal{A}$ takes as input $D^m = D_1 \cdots D_m (D_i = (\mathbf{X}_i, Y_i) \in \mathcal{X} \times \mathcal{Y}, i = 1, \ldots, m)$, each D_i of which is independently drawn according to $Q(\mathbf{X})P^*(Y \mid \mathbf{X})$ (Q and P^* are unknown to $\mathcal{A}$), and outputs a hypothesis $\hat{P}_{[D^m]} \in \mathcal{C}$ such that

$$Prob[d(P^*, \hat{P}_{[D^m]}) > \varepsilon] \leq \delta,$$

and $\mathcal{A}$ runs in time polynomial in $\frac{1}{\varepsilon}$, $\frac{1}{\delta}$, and n, where *Prob* denotes the probability taken with respect to the probability distribution $(Q(\mathbf{X})P^*(Y \mid \mathbf{X}))^m$ over $(\mathcal{X} \times \mathcal{Y})^m$ and any coin-tossing that $\mathcal{A}$ may make.

Theorem 10. *For a class $\mathcal{C} = \{P(Y \mid \mathbf{X}\colon \theta \prec M)\colon M \in \mathcal{M}, \theta \in \Theta(M)\}$ of stochastic rules with finite partitioning, assume that $\gamma(\mathcal{C}) = \sup_{P \in \mathcal{C}} \sup_{\mathbf{X} \in \mathcal{X}, Y \in \mathcal{Y}} \{1/P(Y \mid \mathbf{X})\} < \infty$, and that $\mu(\mathcal{C}) = \dim_{M \in \mathcal{M}} \Theta(M)$ and $\ln |\mathcal{M}|$ are both polynomial in n. If there exists a stochastic PAC-learning algorithm for $\mathcal{C}$ with respect to the Kullback-Leibler divergence, then for any $1 < \gamma < \infty$, $\mathcal{C}$ is polynomial-time PAD-learnable with respect to the Kullback-Leibler divergence under $\mathcal{C}_\gamma$-constraint.*

Proof. For a given class $\mathcal{C}$, assume that $\gamma(\mathcal{C}) < \infty$, and that $\mu(\mathcal{C})$ and $\ln |\mathcal{M}|$ are both polynomial in n. Suppose that there exists a stochastic PAC-learning algorithm for $\mathcal{C}$ with respect to the Kullback-Leibler divergence d and denote it as $\mathcal{A}$. Letting $D^m = D_1 \cdots D_m (D_i = (\mathbf{X}_i, Y_i), i = 1, \ldots, m)$ be a test sequence, we denote an output of $\mathcal{A}$ from D^m as $\hat{P}_{[D^m]}$. We define a discrimination algorithm $\mathcal{B}$ as an algorithm that takes as input $P^*, \mathcal{H}, D^m, \varepsilon$ and outputs "+1" if the following function $h_{\mathcal{B}}$ is positive; otherwise "−1." Here $h_{\mathcal{B}}$ is defined as

$$\begin{aligned} &h_{\mathcal{B}}(P^*, \mathcal{C}, D^m, \varepsilon) \\ &\quad = \ln P^*(Y^m \mid \mathbf{X}^m) - \ln \hat{P}_{[D^m]}(Y^m \mid \mathbf{X}^m) \\ &\qquad + \frac{\mu(\mathcal{C}) \ln m}{2} + \frac{(5 \ln 2)\mu(\mathcal{C})}{2} + \ln |\mathcal{M}| + \frac{m\varepsilon}{2}. \end{aligned} \tag{29}$$

First let us evaluate Type 1 error probability for $\mathcal{B}$. We prepare the following lemma.

Lemma 11. *For any fixed $\mathbf{X}^m$, the following inequality holds.*

$$\sum_{Y^m} e^{-(-\ln \hat{P}_{[D^m]}(Y^m \mid \mathbf{X}^m) + \frac{\mu(\mathcal{C}) \ln m}{2} + \frac{(5 \ln 2)\mu(\mathcal{C})}{2} + \ln |\mathcal{M}|)} \leq 1. \tag{30}$$

Proof of Lemma 11. First notice that, for a given class $\mathcal{C} = \{P(Y \mid \mathbf{X}\colon \theta \prec M)\colon M \in \mathcal{M}, \theta \in \Theta(M)\}$ of stochastic rules with finite partitioning, there exists a prefix code with code-length of $\frac{k \ln m}{2} + \frac{(5 \ln 2)k}{2} + \ell(M)$ for $P(Y \mid \mathbf{X}\colon \theta \prec M) \in \mathcal{C}$ (see Yamanishi, 1992a). Here $k = \dim \Theta(M)$ and ℓ is an arbitrary code-length function over $\mathcal{M}$.

Hence, letting $\ell(M) = \ln |\mathcal{M}|$ for all $M \in \mathcal{M}$, there exists a prefix code over $\mathcal{C}$ with code-length of $\frac{\mu(\mathcal{C}) \ln m}{2} + \frac{(5 \ln 2)\mu(\mathcal{C})}{2} + \ln |\mathcal{M}|$ for all $P \in \mathcal{C}$, since it allows even the rule with the highest number of real-valued parameters over $\mathcal{C}$ to be encoded into a prefix codeword.

Next observe that when given $\mathbf{X}^m$, each Y^m can be encoded into a prefix codeword with code-length of at most $-\ln \hat{P}_{[D^m]}(Y^m \mid \mathbf{X}^m) + \frac{\mu(\mathcal{C}) \ln m}{2} + \frac{(5 \ln 2)\mu(\mathcal{C})}{2} + \ln |\mathcal{M}|$ in the following two steps. First $\hat{P}_{[D^m]}$ itself is encoded into a prefix codeword with code-length

of $\frac{\mu(\mathcal{C})\ln m}{2} + \frac{(5\ln 2)\mu(\mathcal{C})}{2} + \ln|\mathcal{M}|$, and then Y^m is encoded using the arithmetic coding scheme (Rissanen, 1989) into a prefix codeword with code-length of $-\ln \hat{P}_{[D^m]}(Y^m \mid \mathbf{X}^m)$ under the condition that $\hat{P}_{[D^m]}$ has already been known.

Since the existence of a prefix code over any countable set implies that its code-length function satisfies Kraft's inequality over the set (see Section 3), the code-length function defined above satisfies Kraft's inequality. This completes the proof of Lemma 11. □

Using Lemma 11, Type 1 error probability for $\mathcal{B}$ is bounded as follows. Letting $(*)$ be the event that $h_{\mathcal{B}}(P^*, \mathcal{C}, D^m, \varepsilon) \le 0$ (i.e., $P^*(Y^m \mid \mathbf{X}^m) \le e^{-(-\ln \hat{P}_{[D^m]}(Y^m|\mathbf{X}^m) + \frac{\mu(\mathcal{C})\ln m}{2} + \frac{(5\ln 2)\mu(\mathcal{C})}{2} + \ln|\mathcal{M}|) - \frac{m\varepsilon}{2}}$), we have

$$
\begin{aligned}
&(QP^*)^m[D^m \colon h_{\mathcal{B}}(P^*, \mathcal{C}, D^m, \varepsilon) \le 0] \\
&\quad = \sum_{D^m \cdots (*)} (QP^*)(D^m) \\
&\quad \le \sum_{D^m \cdots (*)} Q(\mathbf{X}^m) e^{-(-\ln \hat{P}_{[D^m]}(Y^m|\mathbf{X}^m) + \frac{\mu(\mathcal{C})\ln m}{2} + \frac{(5\ln 2)\mu(\mathcal{C})}{2} + \ln|\mathcal{M}|) - \frac{m\varepsilon}{2}} \\
&\quad \le e^{-\frac{m\varepsilon}{2}} \sum_{\mathbf{X}^m} Q(\mathbf{X}^m) \sum_{Y^m} e^{-(-\ln \hat{P}_{[D^m]}(Y^m|\mathbf{X}) + \frac{\mu(\mathcal{C})\ln m}{2} + \frac{(5\ln 2)\mu(\mathcal{C})}{2} + \ln|\mathcal{M}|)} \\
&\quad \le e^{-\frac{m\varepsilon}{2}}.
\end{aligned}
$$

Here we have used Lemma 11 to derive the last inequality. Thus, for $0 < \delta_1 < 1$, the following sample size is sufficient to guarantee that Type 1 error probability is at most δ_1.

$$
m \ge \frac{2}{\varepsilon} \ln \frac{1}{\delta_1}. \tag{31}
$$

Next let us evaluate Type 2 error probability for $\mathcal{B}$. First notice that the following inequalities hold if $d(\tilde{P}, P^*) > \varepsilon$.

$$
\begin{aligned}
&(Q\tilde{P})^m[h_{\mathcal{B}}(P^*, \mathcal{C}, D^m, \varepsilon) > 0] \\
&\quad \le (Q\tilde{P})^m \left[\frac{1}{m} \ln \frac{P^*(Y^m \mid \mathbf{X}^m)}{\tilde{P}(Y^m \mid \mathbf{X}^m)} + d(\tilde{P}, P^*) > \frac{d(\tilde{P}, P^*) - \varepsilon/2}{4} \right] \\
&\qquad + (Q\tilde{P})^m \left[\frac{1}{m} \ln \frac{\tilde{P}(Y^m \mid \mathbf{X}^m)}{\hat{P}_{[D^m]}(Y^m \mid \mathbf{X}^m)} - d(\tilde{P}, \hat{P}_{[D^m]}) > \frac{d(\tilde{P}, P^*) - \varepsilon/2}{4} \right] \\
&\qquad + (Q\tilde{P})^m \left[d(\tilde{P}, \hat{P}_{[D^m]}) > \frac{d(\tilde{P}, P^*) - \varepsilon/2}{4} \right] \\
&\qquad + (Q\tilde{P})^m \left[\frac{1}{m} \left(\frac{\mu(\mathcal{C})\ln m}{2} + \frac{(5\ln 2)\mu(\mathcal{C})}{2} + \ln|\mathcal{M}| \right) > \frac{d(\tilde{P}, P^*) - \varepsilon/2}{4} \right] \\
&\quad \le (Q\tilde{P})^m \left[\frac{1}{m} \sum_{i=1}^m \ln \frac{P^*(Y_i \mid \mathbf{X}_i)}{\tilde{P}(Y_i \mid \mathbf{X}_i)} + d(\tilde{P}, P^*) > \frac{\varepsilon}{8} \right] \qquad (32) \\
&\qquad + (Q\tilde{P})^m \left[\frac{1}{m} \sum_{i=1}^m \ln \frac{\tilde{P}(Y_i \mid \mathbf{X}_i)}{\hat{P}_{[D^m]}(Y_i \mid \mathbf{X}_i)} - d(\tilde{P}, \hat{P}_{[D^m]}) > \frac{\varepsilon}{8} \right] \qquad (33) \\
&\qquad + (Q\tilde{P})^m \left[d(\tilde{P}, \hat{P}_{[D^m]}) > \frac{\varepsilon}{8} \right] \qquad (34)
\end{aligned}
$$

$$+(Q\tilde{P})^m\left[\frac{1}{m}\left(\frac{\mu(\mathcal{C})\ln m}{2}+\frac{(5\ln 2)\mu(\mathcal{C})}{2}+\ln|\mathcal{M}|\right)>\frac{\varepsilon}{8}\right]. \tag{35}$$

Using Hoeffding's inequality as with the derivation of (14), we see that for $0<\delta_2<1$, if $\sup_{\mathbf{X},Y}(1/P(Y\mid\mathbf{X}))<\gamma<\infty$, the probabilities (32) and (33) are at most $\delta_2/3$ for all sample size satisfying

$$m\geq\frac{32(\ln\gamma(\mathcal{C})\gamma)^2}{\varepsilon^2}\ln\frac{3}{\delta_2}. \tag{36}$$

From the assumption that $\mathcal{A}$ is a stochastic PAC-learning algorithm, we see that the least sample size and computation time required for the probability (34) to be upper bounded by $\delta_2/3$ are both polynomial in $\frac{1}{\varepsilon}$, $\frac{1}{\delta_2}$ and n.

It can be proven as with the derivation of (16) that the probability (35) becomes zero for all sample size satisfying

$$m\geq\frac{4e}{\varepsilon(e-1)}\left(\mu(\mathcal{C})\ln\frac{512\mu(\mathcal{C})}{\varepsilon}+2\ln|\mathcal{M}|\right), \tag{37}$$

which is polynomial in $\frac{1}{\varepsilon}$ and n since by the assumption, $\mu(\mathcal{C})$ and $\ln|\mathcal{M}|$ are both polynomial in n.

From (31), (36), (37) and the polynomiality of sample size and computation time for $\mathcal{A}$, we see that the least sample size and computation time required for Type 1 and 2 error probabilities for $\mathcal{B}$ to be respectively upper bounded by δ_1 and δ_2 are polynomial in $\frac{1}{\varepsilon}$, $\frac{1}{\delta_1}$, $\frac{1}{\delta_2}$ and n. This implies that $\mathcal{B}$ is a polynomial-time PAD-learning algorithm, and thus for any $1<\gamma<\infty$, $\mathcal{C}$ is polynomial-time PAD-learnable with respect to the Kullback-Leibler divergence under $\mathcal{C}_\gamma$-constraint. This completes the proof of Theorem 10. □

Theorem 10 shows that for any given class $\mathcal{C}$ of stochastic rules with finite partitioning such that $\mu(\mathcal{C})$ and $\ln|\mathcal{M}|$ are polynomial in n and that $\gamma(\mathcal{C})<\infty$, the existence of stochastic PAC-learning algorithms for $\mathcal{C}$ immediately implies polynomial-time PAD-learnability of $\mathcal{C}$. However, it is an open problem whether the converse holds. We may conjecture that the converse doesn't hold because one may not efficiently produce a good hypothesis approximately achieving the MDL even if one can approximate the MDL itself in polynomial time.

6. Sample complexity bounds for PAD-learning: Non-parametric case

This section derives test sample size bounds for PAD-learning for the non-parametric case where the target class $\mathcal{C}$ is non-parametric and the hypothesis class $\mathcal{H}$ is written as a union of an infinite number of finite dimensional parametric classes of stochastic rules.

Hereafter let $\mathcal{X}$ be continuous. We assume that a parametric hypothesis class $\mathcal{H}$ is written as $\mathcal{H}=\cup_{d=1}^{\infty}\mathcal{H}^{(d)}$ and $\mathcal{H}^{(d)}=\{P(Y\mid\mathbf{X}\colon\theta\prec M)\colon M\in\mathcal{M}^{(d)},\theta\in\Theta(M)\}$ is a class

of stochastic rules with finite partitioning each of which is specified by a d-dimensional real-valued probability parameter vector and a countable model belonging to a finite set $\mathcal{M}^{(d)}$. We define $\mathcal{H}_m^{(d)}$ by $\mathcal{H}_m^{(d)} \stackrel{\text{def}}{=} \mathcal{H}_m \cap \mathcal{H}^{(d)}$ where $\mathcal{H}_m$ is a quantized subset of $\mathcal{H}$. For any given set $\mathcal{C}$ of stochastic rules such that $\gamma(\mathcal{C}) = \sup_{P\in\mathcal{C}} \sup_{\mathbf{X},Y}\{1/P(Y \mid \mathbf{X})\} < \infty$, for any $\tilde{P} \in \mathcal{C}$, we define the *projection of* $\tilde{P}$ *on* $\mathcal{H}^{(d)}$, which we denote as $\tilde{P}^{(d)}$, by $\tilde{P}^{(d)} \stackrel{\text{def}}{=} \arg\min_P d(\tilde{P}, P)$, where the minimum is taken over all Ps in $\mathcal{H}^{(d)}$ such that $\sup_{\mathbf{X},Y}(1/P(Y \mid \mathbf{X})) \leq \gamma(\mathcal{C})$. We let $\tilde{P}_m^{(d)}$ be the function obtained by truncating $\tilde{P}^{(d)} \in \mathcal{H}^{(d)}$ in $\mathcal{H}_m^{(d)}$. We write $\tilde{P}_m^{(d)}(Y \mid \mathbf{X})$ as $P(Y \mid \mathbf{X}\colon \tilde{\theta}^{(d)} \prec \tilde{M}^{(d)})$ where $\tilde{\theta}^{(d)}$ is a d-dimensional probability vector and $\tilde{M}^{(d)} \in \mathcal{M}^{(d)}$.

In this setting, when D^m is given, the MDL of Y^m given $\mathbf{X}^m$ relative to $\mathcal{H}_m$ is calculated as follows:

$$L_{\text{MDL}}(Y^m \mid \mathbf{X}^m\colon \mathcal{H}_m) = \min_{1\leq d\leq m} \min_{M\in\mathcal{M}^{(d)}} \min_{\theta\in\Theta_m(M)} \{-\ln P(Y^m \mid \mathbf{X}^m\colon \theta \prec M) + \ell_m(\theta, M)\}, \tag{38}$$

where $\ell_m(\theta, M) \stackrel{\text{def}}{=} \frac{d\ln m}{2} + \frac{5d\ln 2}{2} + \ell(M) + \ell^*(d)$. Here $\ell(M)$ is an arbitrary code-length function over $\mathcal{M}^{(d)}$. Hereafter, we assume that for each d, $\ell(M)$ is constant over $\mathcal{M}^{(d)}$. $\ell^*(d)$ is the code-length required for encoding of the integer d and is calculated using Rissanen's integer coding scheme [see (Rissanen, 1983), (Rissanen, 1989)] as follows: $\ell^*(d) = (\ln 2)(\log_2 c + \log_2 d + \log_2\log_2 d + \cdots)$ where $c = 2.865$. Note that the range of d in (38) could be improved, which will be discussed after the proof of Theorem 12.

The following theorem shows an upper bound on test sample size for PAD-learning of a non-parametric class in terms of a parametric class.

Theorem 12. *Let $\mathcal{C}$ be a non-parametric class of stochastic rules such that $\gamma(\mathcal{C}) = \sup_{P\in\mathcal{C}} \sup_{\mathbf{X},Y}\{1/P(Y \mid \mathbf{X})\} < \infty$. Let $\mathcal{H} = \cup_{d=1}^{\infty}\mathcal{H}^{(d)}$ be a class of stochastic rules with finite partitioning where $\mathcal{H}^{(d)} = \{P(Y \mid \mathbf{X}\colon \theta \prec M)\colon M \in \mathcal{M}^{(d)}, \theta \in \Theta(M)\}$ as described above. For any n, for any $0 < \varepsilon < 1$, for any $0 < \delta_1, \delta_2 < 1$, for any density q over $\mathcal{X}$, for any $\tilde{P} \in \mathcal{C}$ such that $\tilde{\gamma} = \sup_{\mathbf{X},Y}(1/\tilde{P}(Y \mid \mathbf{X})) < \infty$, for any $P^* \in \mathcal{C}_{all}$ such that $d(\tilde{P}, P^*) > \varepsilon$ and $\gamma^* = \sup_{\mathbf{X},Y}(1/P^*(Y \mid \mathbf{X})) < \infty$, whenever sample size satisfies*

$$m \geq \max\left\{\frac{2}{\varepsilon}\ln\frac{1}{\delta_1}, \frac{18(\ln\tilde{\gamma}\gamma^*)^2}{\varepsilon^2}\ln\frac{2}{\delta_2}, m_1(\varepsilon, \tilde{P}\colon \mathcal{H})\right\}, \tag{39}$$

Type 1 and 2 error probabilities for the MDL discrimination algorithm using $\mathcal{H}$ are then at most δ_1 and δ_2 respectively. Here $m_1(\varepsilon, \tilde{P}\colon \mathcal{H})$ is the least sample size such that

$$\min_d \left\{d(\tilde{P}, \tilde{P}^{(d)}) + \frac{\ell_m(\tilde{P}_m^{(d)})}{m} + \frac{d\ln 2}{m}\right\} \leq \frac{\varepsilon}{6}, \tag{40}$$

where for $\tilde{P}_m^{(d)}(Y \mid \mathbf{X}) = P(Y \mid \mathbf{X}\colon \tilde{\theta}^{(d)} \prec \tilde{M}^{(d)})$, $\ell_m(\tilde{P}_m^{(d)}) \stackrel{\text{def}}{=} \frac{d\ln m}{2} + \frac{5d\ln 2}{2} + \ell(\tilde{M}^{(d)}) + \ell^*(d)$.

Proof. It can be proven as with the derivation of (10) that Type 1 error probability for the MDL discrimination algorithm using $\mathcal{H}$ is at most δ_1 for all sample size satisfying

$$m \geq \frac{2}{\varepsilon} \ln \frac{1}{\delta_1}. \tag{41}$$

Next we evaluate Type 2 error probability for the MDL discrimination algorithm using $\mathcal{H}$. If $d(\tilde{P}, P^*) > \varepsilon$, then for any $1 \leq d \leq m$, we have the following inequalities.

$$\begin{aligned}
&(Q\tilde{P})^m[D^m\colon h_{\mathrm{MDL}}(P^*, \mathcal{H}, D^m, \varepsilon) > 0] \\
&\leq (Q\tilde{P})^m\left[\frac{1}{m}\ln\frac{P^*(Y^m \mid \mathbf{X}^m)}{\tilde{P}(Y^m \mid \mathbf{X}^m)} + d(\tilde{P}, P^*) > \frac{d(\tilde{P}, P^*) - \varepsilon/2}{3}\right] \\
&\quad + (Q\tilde{P})^m\left[\frac{1}{m}\ln\frac{\tilde{P}(Y^m \mid \mathbf{X}^m)}{\tilde{P}^{(d)}(Y^m \mid \mathbf{X}^m)} - d(\tilde{P}, \tilde{P}^{(d)}) > \frac{d(\tilde{P}, P^*) - \varepsilon/2}{3}\right] \\
&\quad + (Q\tilde{P})^m\left[\frac{1}{m}\ln\tilde{P}^{(d)}(Y^m \mid \mathbf{X}^m) + d(\tilde{P}, \tilde{P}^{(d)})\right. \\
&\qquad\qquad \left. + \frac{1}{m}L_{\mathrm{MDL}}(Y^m \mid \mathbf{X}^m\colon \mathcal{H}_m) > \frac{d(\tilde{P}, P^*) - \varepsilon/2}{3}\right] \\
&\leq (Q\tilde{P})^m\left[\frac{1}{m}\sum_{i=1}^{m}\ln\frac{P^*(Y_i \mid \mathbf{X}_i)}{\tilde{P}(Y_i \mid \mathbf{X}_i)} + d(\tilde{P}, P^*) > \frac{\varepsilon}{6}\right] \qquad (42) \\
&\quad + (Q\tilde{P})^m\left[\frac{1}{m}\sum_{i=1}^{m}\ln\frac{\tilde{P}(Y_i \mid \mathbf{X}_i)}{\tilde{P}^{(d)}(Y_i \mid \mathbf{X}_i)} - d(\tilde{P}, \tilde{P}^{(d)}) > \frac{\varepsilon}{6}\right] \qquad (43) \\
&\quad + (Q\tilde{P})^m\left[\frac{1}{m}\ln\tilde{P}^{(d)}(Y^m \mid \mathbf{X}^m) + d(\tilde{P}, \tilde{P}^{(d)})\right. \\
&\qquad\qquad \left. + \frac{1}{m}L_{\mathrm{MDL}}(Y^m \mid \mathbf{X}^m\colon \mathcal{H}_m) > \frac{\varepsilon}{6}\right]. \qquad (44)
\end{aligned}$$

It can be proven using Hoeffding's inequality as with the derivation of (14) that the probabilities (42) and (43) are both at most $\delta_2/2$ for all sample size satisfying

$$m \geq \frac{18(\ln \tilde{\gamma}\gamma^*)^2}{\varepsilon^2} \ln \frac{2}{\delta_2}. \tag{45}$$

Next we upper bound (44). Since the MDL is a lower bound for the total code-length of examples, as with (17), for any d such that $1 \leq d \leq m$, we have

$$L_{\mathrm{MDL}}(Y^m \mid \mathbf{X}^m\colon \mathcal{H}_m) < -\ln \tilde{P}^{(d)}(Y^m \mid \mathbf{X}^m) + \ell_m(\tilde{P}_m^{(d)}) + d\ln 2$$

Hence we obtain

$$\begin{aligned}
&\frac{1}{m}\ln\tilde{P}^{(d)}(Y^m \mid \mathbf{X}^m) + d(\tilde{P}, \tilde{P}^{(d)}) + \frac{1}{m}L_{\mathrm{MDL}}(Y^m \mid \mathbf{X}^m\colon \mathcal{H}_m) \\
&\quad < d(\tilde{P}, \tilde{P}^{(d)}) + \frac{\ell_m(\tilde{P}_m^{(d)})}{m} + \frac{d\ln 2}{m}.
\end{aligned}$$

Thus if m exceeds the least sample size such that

$$\min_d \left\{ d(\tilde{P}, \tilde{P}^{(d)}) + \frac{\ell_m(\tilde{P}_m^{(d)})}{m} + \frac{d \ln 2}{m} \right\} \leq \frac{\varepsilon}{6}, \tag{46}$$

then

$$(Q\tilde{P})^m \left[\frac{1}{m} \ln \tilde{P}^{(d)}(Y^m \mid \mathbf{X}^m) + d(\tilde{P}, \tilde{P}^{(d)}) + \frac{1}{m} L_{\mathrm{MDL}}(Y^m \mid \mathbf{X}^m \colon \mathcal{H}_m) > \frac{\varepsilon}{6} \right] = 0.$$

From (41), (45), and (46), we see that Type 1 and 2 error probabilities are respectively at most δ_1 and δ_2 for all sample size satisfying

$$m \geq \max \left\{ \frac{2}{\varepsilon} \ln \frac{1}{\delta_1}, \frac{18(\ln \tilde{\gamma}\gamma^*)^2}{\varepsilon^2} \ln \frac{2}{\delta_2}, m_1(\varepsilon, \tilde{P} \colon \mathcal{H}) \right\},$$

which yields the bound (39). This completes the proof of Theorem 12. □

Let $d_0(m) \stackrel{\mathrm{def}}{=} \arg\min_d \sup_{\tilde{P} \in \mathcal{C}} \{ d(\tilde{P}, \tilde{P}^{(d)}) + \frac{\ell_m(\tilde{P}_m^{(d)})}{m} + \frac{d \ln 2}{m} \}$. As seen from the proof of Theorem 12, even if we replace the range "$1 \leq d \leq m$" for the minimization in (38) with respect to d with "$1 \leq d \leq d_0(m)$," we obtain the same upper bound on test sample size as (39). Thus, hereafter, we may change the range of d in (38) into $1 \leq d \leq d_0(m)$. This range reduces computation-time greatly since $d_0(m)$ becomes much smaller than m when m is sufficiently large. $d_0(m)$ is hard to estimate in actual cases, however, as shown after Corollary 13, there exist classes for which $d_0(m)$ can be obtained as a simple form.

The lefthand side of (40) is called the *index of resolvability* (Barron & Cover, 1991), which is considered to be the optimal balance of trade-off between the approximation error $d(\tilde{P}, \tilde{P}^{(d)})$ of the d-dimensional hypothesis class $\mathcal{H}^{(d)}$ to $\tilde{P}$ and the complexity $\frac{\ell_m(\tilde{P}_m^{(d)})}{m}$ of the projection of $\tilde{P}$ on $\mathcal{H}_m^{(d)}$ (measured by the ratio of the code-length for $\tilde{P}_m^{(d)}$ to sample size). Note that (40) is not truly the same as Barron and Cover's original definition of index of resolvability, in which $d(\tilde{P}, \tilde{P}^{(d)})$ is replaced with $d(\tilde{P}, \tilde{P}_m^{(d)})$ and $d(\ln 2)/m$ does not appear. In most cases, however, they have the same order with respect to m.

It is known from (Barron & Cover, 1991) that in the context of density estimation, the rate of convergence of the MDL estimator to the target rule is asymptotically governed by the index of resolvability. Theorem 12 shows that even in the context of universal hypothesis testing, the test sample size bound is also essentially related to the rate of convergence of the index of resolvability to zero.

Let $\mathcal{Q}$ be a class of densities over $\mathcal{X}$. In particular, if $\mathcal{Q}$ and $\mathcal{C}$ are constrained so that for some $\alpha > 0$, for all $q \in \mathcal{Q}$, for all $\tilde{P} \in \mathcal{C}$, for any d,

$$d(\tilde{P}, \tilde{P}^{(d)}) \leq O\left(\frac{1}{d^\alpha} \right),$$

we have

$$d(\tilde{P}, \tilde{P}^{(d)}) + \frac{\ell_m(\tilde{P}_m^{(d)})}{m} + \frac{d \ln 2}{m} = O\left(\frac{1}{d^\alpha}\right) + O\left(\frac{d \ln m}{m}\right), \tag{47}$$

since $\frac{\ell_m(\tilde{P}_m^{(d)})}{m} = O(\frac{d \ln m}{m})$.

The minimum of (47) is attained by

$$d = \left\lceil c\left(\frac{m}{\ln m}\right)^{\frac{1}{\alpha+1}} \right\rceil,$$

where c is a positive constant. The minimum is then given by $O(\frac{1}{d^\alpha}) + O(\frac{d \ln m}{m}) = O((\frac{\ln m}{m})^{\frac{\alpha}{\alpha+1}})$. Hence we see

$$\min_d \left\{ d(\tilde{P}, \tilde{P}^{(d)}) + \frac{\ell_m(\tilde{P}_m^{(d)})}{m} + \frac{d \ln 2}{m} \right\} = O\left(\left(\frac{\ln m}{m}\right)^{\frac{\alpha}{\alpha+1}}\right). \tag{48}$$

Thus the least sample size needed for the lefthand side of (48) to be at most $\frac{\varepsilon}{6}$ has the following upper bound:

$$O\left(\left(\frac{1}{\varepsilon}\right)^{\frac{\alpha+1}{\alpha}} \ln \frac{1}{\varepsilon}\right).$$

The method of optimizing d by balancing $d(\tilde{P}, \tilde{P}^{(d)})$ and $\frac{\ell_m(\tilde{P}_m^{(d)})}{m}$ follows from (Barron & Cover, 1991). More generally, α can be replaced with a function $\alpha(n)$ of n.

Combining the above argument with Theorem 12 and applying the worst-case analysis as in (B) of Theorem 3 to sample size evaluation, we have the following corollary.

Corollary 13. *Let $\mathcal{H} = \cup_{d=1}^{\infty} \mathcal{H}^{(d)}$ be a class of stochastic rules with finite partitioning as in Theorem 12. Let $\mathcal{Q}$ be a class of densities on $\mathcal{X}$ and $\mathcal{C}$ be a non-parametric class of stochastic rules such that for some $\alpha > 0$, $\sup_{q \in \mathcal{Q}} \sup_{\tilde{P} \in \mathcal{C}} d(\tilde{P}, \tilde{P}^{(d)}) = \sup_{q \in \mathcal{Q}} \sup_{\tilde{P} \in \mathcal{C}} \min_{P \in \mathcal{H}^{(d)}} d(\tilde{P}, P) = O(1/d^\alpha)$ and $\sup_{P \in \mathcal{C}} \sup_{\mathbf{X},Y} \{1/P(Y \mid \mathbf{X})\} < \infty$. Then for any $1 < \gamma < \infty$, for any $q \in \mathcal{Q}$, we have the following upper bound on the sample complexity of PAD-learning of $\mathcal{C}$ in terms of $\mathcal{H}$ with respect to the Kullback-Leibler divergence under $\mathcal{C}_\gamma$-constraint.*

$$m_0(\varepsilon, \delta_1, \delta_2, n) = O\left(\frac{1}{\varepsilon} \ln \frac{1}{\delta_1} + \frac{1}{\varepsilon^2} \ln \frac{1}{\delta_2} + \left(\frac{1}{\varepsilon}\right)^{\frac{\alpha+1}{\alpha}} \ln \frac{1}{\varepsilon}\right). \tag{49}$$

We see that if the target class $\mathcal{C}$ is constrained as in Corollary 13, the range $1 \leq d \leq m$ for the minimization in (38) can be replaced with $1 \leq d \leq \lceil c(\frac{m}{\ln m})^{\frac{1}{\alpha+1}} \rceil$ $(c > 0)$.

Next let us derive a test sample complexity bound for a more concrete non-parametric class of stochastic rules. We prepare the following lemma.

Lemma 14 [Rissanen & Yu, 1991]. *Let $\mathcal{X} = [0,1]$ and $\mathcal{Y} = \{0,1\}$. Let $\mathcal{Q}$ be a class of densities over $\mathcal{X}$ such that for some $0 < c_0 < 1$, $1 < c_1 < \infty$, and $0 < c_2 < \infty$, for all $q \in \mathcal{Q}$, for all $\mathbf{X} \in \mathcal{X}$, $0 < c_0 < q(\mathbf{X}) < c_1$ and $|\frac{d}{d\mathbf{X}} q(\mathbf{X})| < c_2$. Let $\mathcal{C}$ be a class of stochastic rules such that for some $0 < c_3 < c_4 < 1$, $0 < c_5 < \infty$, for all $P \in \mathcal{C}$, for all $\mathbf{X} \in \mathcal{X}$, $0 < c_3 < P(0 \mid \mathbf{X}) < c_4 < 1$, and $|\frac{d}{d\mathbf{X}} P(0|\mathbf{X}) | < c_5$. Let $\mathcal{H} = \cup_{d=1}^{\infty} \mathcal{H}^{(d)}$ be a class of stochastic rules with finite partitioning with equal-length cells where $\mathcal{H}^{(d)}$ is a d-dimensional parametric class, i.e., the set of disjoint cells of $\mathcal{X} = [0,1]$ for each element in $\mathcal{H}^{(d)}$ consists of d equal-length cells with length $1/d$. Then for any $q \in \mathcal{Q}$, for any $\tilde{P} \in \mathcal{C}$, there exists $P \in \mathcal{H}^{(d)}$ such that*

$$d(\tilde{P}, P) = O\Big(\frac{1}{d^2}\Big). \tag{50}$$

Combining Corollary 13 with Lemma 14, we have the following theorem.

Theorem 15. *Let $\mathcal{X} = [0,1]$ and $\mathcal{Y} = \{0,1\}$. Let $\mathcal{Q}$ be a class of densities over $\mathcal{X}$ as in Lemma 14. Let $\mathcal{C}$ be a non-parametric class of stochastic rules as in Lemma 14. Let $\mathcal{H} = \cup_{d=1}^{\infty} \mathcal{H}^{(d)}$ be a class of stochastic rules with finite partitioning with equal-length cells as in Lemma 14. Then for any $1 < \gamma < \infty$, for any $q \in \mathcal{Q}$, we have the following upper bound on the sample complexity of PAD-learning of $\mathcal{C}$ in terms of $\mathcal{H}$ with respect to the Kullback-Leibler divergence under $\mathcal{C}_\gamma$-constraint.*

$$\begin{aligned} m_0(\varepsilon, \delta_1, \delta_2, n) &= O\Big(\frac{1}{\varepsilon} \ln \frac{1}{\delta_1} + \frac{1}{\varepsilon^2} \ln \frac{1}{\delta_2} + \Big(\frac{1}{\varepsilon}\Big)^{\frac{3}{2}} \ln \frac{1}{\varepsilon}\Big) \\ &= O\Big(\frac{1}{\varepsilon} \ln \frac{1}{\delta_1} + \frac{1}{\varepsilon^2} \ln \frac{1}{\delta_2}\Big). \end{aligned} \tag{51}$$

We see that if the target class $\mathcal{C}$ is constrained as in Lemma 40, the range $1 \le d \le m$ for the minimization in (40) can be replaced with $1 \le d \le \lceil c(\frac{m}{\ln m})^{\frac{1}{3}} \rceil$ $(c > 0)$.

The bound (51) shows that for the non-parametric class as in Corollary 13, the upper bound on test sample size is governed by the first and second terms of (51) only, and the third term of (51) is asymptotically ignored compared to the second term.

Theorem 15 implies that if $\mathcal{C}$ is a non-parametric class specified by some smoothness conditions over the real line, then $\mathcal{C}$ is also polynomial-time PAD-learnable in terms of the class of stochastic rules with finite partitioning because the computation time for the MDL discrimination algorithm is polynomial in $1/\varepsilon$ and $1/\delta$ in the case where the size of domain is fixed to be 1.

7. Conclusion

In this paper, we have developed a PAD-learning model based on the universal hypothesis testing theory. Unlike the conventional Neyman-Pearson type hypothesis testing theory, our concern is not to find an asymptotically optimal test but to derive tight bounds on test sample size. The discrimination algorithm which we have proposed based on the MDL

principle has turned out to perform well within the PAD model in the sense of sample complexity efficiency.

For the parametric case, an upper bound on test sample size for PAD-learning with the MDL discrimination algorithm is given by $O(\frac{1}{\varepsilon}\ln\frac{1}{\delta_1}+\frac{1}{\varepsilon^2}\ln\frac{1}{\delta_2}+\frac{\tilde{k}}{\varepsilon}\ln\frac{\tilde{k}}{\varepsilon}+\frac{\ell(\tilde{M})}{\varepsilon})$ where $\tilde{k}$ is the number of parmeters for the composite hypothesis, and $\ell(\tilde{M})$ is the code-length for the countable model for the composite hypothesis. Further we have demonstrated that this upper bound is the least reported to date for any information-criteria-based discrimination algorithm. This sample complexity analysis might give a rationale for the effectiveness of the MDL principle in the PAD-learning framework.

For the non-parametric case, we have shown the test sample size bound for PAD-learning by the MDL discrimination algorithm is essentially related to Barron and Cover's index of resolvability. This analysis might give a new view at the index of resolvability from the universal hypothesis testing viewpoint, whereas Barron and Cover related it to the rate of convergence of the MDL estimation only. Further we have shown that when a non-parametric target class of hypotheses is constrained under some smoothness conditions and the family of stochastic rules with finite partitioning is taken as a hypothesis class, an upper bound on test sample size is given by $O(\frac{1}{\varepsilon}\ln\frac{1}{\delta_1}+\frac{1}{\varepsilon^2}\ln\frac{1}{\delta_2}+(\frac{1}{\varepsilon})^{\frac{\alpha+1}{\alpha}}\ln\frac{1}{\varepsilon})$ for some $\alpha>0$.

The following issues remain open.

1) *How can we design an efficient algorithm for approximating the MDL discrimination algorithm?* This paper has shown the sample size efficiency of the MDL discrimination algorithm, but it is computationally inefficient to calculate the MDL itself in some cases where an important class (e.g. stochastic decision lists, stochastic decision trees, etc.) is employed as a hypothesis class. Hence the development of an efficient algorithm for approximating the MDL is an important issue which we have to address to get polynomial-time PAD-learnability results. It is shown in (Yamanishi, 1993) that a large family of classes of stochastic rules including k-stochastic decision lists is polynomial-time PAD-learnable for some limited range of accuracy parameter. However it still remains open whether they are polynomial-time PAD-learnable for arbitrary $\varepsilon>0$.

2) *How can we separate stochastic PAC-learnable classes from PAD-learnable ones?* In Section 5 it has been shown that under some conditions polynomial-time stochastic PAC-learnability of any class is a sufficient condition for polynomial-time PAD-learnability of the class. We may conjecture that PAD-learnability is weaker than stochastic PAC-learnability, but it still remains open whether there exists a class which is polynomial-time PAD-learnable but is not polynomial-time stochastic PAC-learnable.

3) *How can we improve the upper bounds on sample complexity of PAD-learning?* All the bounds derived in this paper include an $O(\frac{1}{\varepsilon^2}\ln\frac{1}{\delta_2})$ term, which was derived using Hoeffding's inequality. We expect that this can be improved to $O(\frac{1}{\varepsilon}\ln\frac{1}{\delta_2})$ using more sophisticated techniques. It would be interesting to derive a lower bound on test sample size to compare it with our upper bounds.

These issues will be dealt in future study.

Acknowledgment

The author appreciates Jason Catlett for his reading the earlier draft of this paper. He also thanks anonymous reviewers for their helpful comments.

References

Barron, A.R., & Cover T.M. (1991). Minimum complexity density estimation. *IEEE Trans. on Information Theory, IT-37,* 1034–1054.

Blahut, R.E. (1988). *Principle and Practice of Information Theory.* Addison-Wesley.

Cover, T.M., & Thomas, J.A. (1991). *Elements of Information Theory.* Wiley-Interscience.

DeSantis, A., Markowsky, G., & Wegman, M.N. (1988). Learning probabilistic prediction functions. *Proceedings of the First Annual Workshop on Computational Learning Theory* (pp. 312–328), Morgan Kaufmann.

Gutman, M. (1989). Asymptotically optimal classification for multiple tests with empirically observed statistics. *IEEE Trans. on Information Theory, IT-35,* 2, 401–408.

Hand, D.J. (1981). *Discrimination and Classification.* New York: Wiley.

Haussler, D., & Barron, A. (1992). How well does the Bayes method work in on-line predictions of $\{+1, -1\}$-values. *Proceedings of the Third NEC Symposium* (pp. 74–100): SIAM.

Haussler, D., & Long P. (1990). A generalization of Sauer's lemma. *Technical Report UCSC CRL-90-15,* University of California at Santa Cruz.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Jr. Amer. Stat. Assoc.,* 58, 13–30.

Hoeffding, W. (1965). Asymptotically optimal test for multinomial distributions. *Annals of Mathematical Statistics,* 36, 369–400.

Kearns, M., & Schapire, R. (1994). Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences,* 48, 3, 464–497.

Kraft, C. (1949). A device for quantizing, grouping, and coding amplitude modulated pulses. *M.S. Thesis,* Department of Electrical Engineering, MIT, Cambridge, MA.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica,* 14, 465–471.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Annals of Statistics,* 11, 416–431.

Rissanen, J. (1987). Stochastic complexity. *J.R. Statist. Soc. B, 49,* 3, 223–239.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry.* World Scientific, *Series in Computer Science,* 15.

Rissanen, J., & Yu, B. (1991). MDL learning. *Progress in Automation and Information Systems,* Springer Verlag.

Rivest. R.L. (1987). Learning decision lists. *Machine Learning,* 2, 229–246.

Schwarz, G. (1978). Estimation of the dimension of a model. *Annals of Statistics,* 6, 416–446.

Shannon, C.E. (1948). A mathematical theory of communications. *Bell Syst. Tech. J.* 47:147–157.

Valiant, L.G. (1984). A theory of the learnable. *Communications. of the ACM,* 27, 1134–1142.

Wallace, C.S., & Boulton, D.M. (1968). An information measure for classification. *Computer Journal,* 185–194.

Yamanishi, K. (1991). A loss bound model for on-line stochastic prediction strategies. *Proceedings of the Fourth Annual Workshop on Computational Learning Theory* (pp. 290–302), Morgan Kaufmann.

Yamanishi, K. (1992a). A learning criterion for stochastic rules. *Machine Learning: Special Issues for COLT-90,* 9, 165–203.

Yamanishi, K. (1992b). Probably almost discriminative learning. *Proceedings of the Fifth ACM Workshop on Computational Learning Theory* (pp. 164–171), ACM Press.

Yamanishi, K. (1993). On polynomial-time probably almost discriminative learnability. *Proceedings of the Sixth ACM Conference on Computational Learning Theory* (pp. 94–100), ACM Press.

Zeitouni, O., & Gutman, M. (1991). On universal hypothesis testing via large deviations. *IEEE Trans. on Information Theory, IT-37,* 285–290.

Ziv, J. (1988). On classification with empirically observed statistics and universal data compression. *IEEE Trans. on Information Theory, IT-34,* 278–286.

Ziv, J., & Lempel, A. (1978). Compression of individual sequences via variable-rate coding. *IEEE Trans. on Information Theory, IT-24,* 530–536.

Received March 11, 1993
Accepted September 13, 1993
Final Manuscript November 12, 1993