Pseudo Information Divergences Defined on the Family of Specific Probability Distributions

Hiroyuki Shioya† Tsutomu Da-te ‡

† Muroran Institute of Technology
Address: 27-1 Mizumoto Muroran, 050-8585, Japan
Tel & Fax: +81-143-46-5436, Email:shioya@csse.muroran-it.ac.jp

‡ Division of Systems and Information Engineering, Hokkaido University
Address: Kita-13 Nishi-8 Kita-ku Sapporo, 060-8628, Japan
Tel : +81-11-706-6855, Email:date@main.eng.hokudai.ac.jp

Abstract Several information measures have been used as the criteria in information theory, statistics and various fields of engineering. Especially an information divergence has been well used as the measure of the difference between two probability distributions. In this paper, we propose the pseudo information divergence, which functions as usual information divergence, if two measured probability distributions are in some family of specific distributions. We introduce an example of the pseudo information divergence, and apply it to the problem of training multi-layer perceptrons from the data with the gross error noise.
Keywords pseudo information divergence, f-divergence, direct on-line learning, multi-layer perceptron, gross error model.

1 Introduction

One of the most widely used learning procedures is the method of least squares used in many technical applications (neural networks, image processing, pattern recognition and so on). Its method has been well used and has been developed with adding several supplementary terms that bring an effective improvement. For example, one of these methods is the learning algorithm with the regularization term for multi-layer perceptrons. The theoretical background of the least squared method is the statistical estimation using Gaussian model, which can be regarded as the minimization of the difference between the target distribution and the model distribution. Information divergences represent such the difference in the space of all probability distributions. These measures well used in information theory are the pseudo distance between two probability distributions. The minimization of these divergences makes us to understand easily several statistical estimators in a space of all probability distributions topologically. The minimum Kullback divergence on the model of Gaussian distributions is corresponding to the method of least squares,

International Journal of Computing Anticipatory Systems, Volume 11, 2002
Edited by D. M. Dubois, CHAOS, Liège, Belgium, ISSN 1373-5411 ISBN 2-9600262-5-X

but the minimization of the other information divergences doesn't realize an effective on-line learning algorithm. For example, the direct on-line learning algorithm using the minimum α-divergence have not been derived. Thus, it will be important for the progress of the learning theory to examine the direct on-line learning method from the view of information theoretical measures.

In this paper, we introduce the theoretical framework of the pseudo information divergence (PID in short) defined on the family of specific probability distributions. We use the word *pseudo* in the meaning that its measure is not effective for all probability distributions but some specific family of probability distributions. We compose the axiom of pseudo information divergences. Using this framework, concretely we derive a kind of PIDs as a weak version of f-divergences (f-PID in short) and show the properties concerning to the family of specific probability distributions of the measure.

Moreover we introduce an example of pseudo information measures using α-divergence (α-PID in short), and we clearly derive the family of specific distributions of α-PID and the learning algorithm used the minimum α-PID. We show an experimental result using our learning method for training multi-layer perceptrons from the data with the gross error contamination.

2 Information divergence

Information measures are well used in information theory, which is a mathematical theory of communication, namely data compression and data transmission [4]. An information divergence is a kind of information measures and used in the space of all probability distributions space. So it has the properties of the spatial meaning. Information divergences are not used for generating the metric space but used for generating the natural structure in the differential manifold defined on all probability distributions [1].

Let μ be a finite measure dominating the probability measures defined on a set $\boldsymbol{\mathcal{X}}$. A set of all probabilty distributions defined on $\boldsymbol{\mathcal{X}}$ is:

$$\boldsymbol{\mathcal{P}} = \{p | \int_{\boldsymbol{\mathcal{X}}} p(x) d\mu(x) = 1, p(x) \geq 0 \ \forall x \in \boldsymbol{\mathcal{X}} \} \tag{1}$$

For $\forall p, q \in \boldsymbol{\mathcal{P}}$, the Kullback-Leibler divergence [10] is given by

$$D_k(p\|q) = \int_{\boldsymbol{\mathcal{X}}} p(x) \log \frac{p(x)}{q(x)} d\mu(x). \tag{2}$$

The Kullback divergence does not satisfy the axiom of metric. However, it's used as a natural distance-like measure in the space of probability distributions. Let us recall the fundamental property of an information divergence $D(\ \|\ )$ in the following: For p and q satisfying $\mu(B) > 0$ (where $B = \{x | p(x) \neq q(x), x \in \boldsymbol{\mathcal{X}} \}$, $D(p\|q)$ is positive, and $D(p\|q) = 0$ if and only if $p(x) = q(x)$ $(\forall x \in \boldsymbol{\mathcal{X}})$. There are many

kinds of functions satisfied its axiom, however few typical information divergences are used in information theory and statistics. The α-divergence was an originator of the generalized information divergence derived by A. Reniy[11].

$$D_\alpha(p\|q) \stackrel{\text{def}}{=} \frac{1}{\alpha(1-\alpha)}\left[1 - \int_{\boldsymbol{\mathcal{X}}} p(x)^{1-\alpha}q(x)^\alpha d\mu(x)\right]. \tag{3}$$

where $\alpha \in \boldsymbol{R}$. The relation between D_α and D_k is denoted by the properties $\lim_{\alpha\to 0} D_\alpha(p\|q) = D_k(p\|q)$ and $\lim_{\alpha\to 1} D_\alpha(p\|q) = D_k(q\|p)$.

A main research on the generalization of information divergences was started by I. Csiszár [6] [7]. He derived the following f-divergences as a generalized class of information divergences.

$$D_f(p\|q) \stackrel{\text{def}}{=} \int_{\boldsymbol{\mathcal{X}}} p(x) f\left(\frac{q(x)}{p(x)}\right) d\mu(x). \tag{4}$$

where $f(u)$ is a convex function defined on $(0,\infty)$, strictly convex at $u = 1$, and satisfies $f(1) = 0$. A class of f-divergences is useful for analyzing the general properties of information divergences. However, it is difficult to derive the meaningful results from the analysis of f-divergences, because of the generality of its formulation.

3 Pseudo Information Divergences

An information divergence is a kind of *pseudo* distance between two probability distributions. In this way, we consider a pseudo version of an information divergence. In this section, we define the concept of Pseudo Information Divergences (PID) and derive f-PID. We show an illustrative example of f-PIDs and its learning procedure.

3.1 Definition of PID

For the definition of a new information measure, the fundamental properties of the measure need to be provided. Therefore, we give the following definition of PIDs by which the fundamental properties of an information divergence are weakened.

Definition 1 *$PID(\ \|\ )$ is called the pseudo information divergence defined on $\boldsymbol{\mathcal{A}}$ ($\subset \boldsymbol{\mathcal{P}}$) if $PID(p\|q)$ functions as an information divergence in the case that $p, q \in \boldsymbol{\mathcal{A}}$. Then, $\boldsymbol{\mathcal{A}}$ is called a family of specific distributions for the $PID(\ \|\ )$.*

Definition 2 *Let $\boldsymbol{\mathcal{A}}$ be some family of specific distributions for $PID(\ \|\ )$. If any subset of $\boldsymbol{\mathcal{P}} \setminus \boldsymbol{\mathcal{A}}$ is not a family of specific distributions for $PID(\ \|\ )$, then $\boldsymbol{\mathcal{A}}$ is called the maximal family of specific distributions for $PID(\ \|\ )$.*

The word "pseudo" means that the maximal specific distributions family is a subset of $\boldsymbol{\mathcal{P}}$.

3.2 f-PID

In the previous, we gave the definition of PID, but we have not given its concrete formulation. For a present, it is difficult to give quite a new formulation without using every known divergence. Therefore, we define the following information measure as an extended version of f-divergences in the following.

$$D_f^g(p\|q) \stackrel{\text{def}}{=} \int_{\mathcal{X}} p(x)g[p(x)]f\left(\frac{q(x)}{p(x)}\right)d\mu(x), \tag{5}$$

where f is a convex function used in the definition of f-divergences. g is an non-negative and non-decreasing function on $[0,\infty)$, and $\int_{\mathcal{X}} q(x)g[p(x)]d\mu(x) < \infty$ is established for $\forall p, q \in \mathcal{P}$.

If $g(u)$ is a positive constant C for $\forall u \in [0,\infty)$, then $D_f^g(p\|q) = CD_f(p\|q)$. Thus we easily see that D_f^g is an extended version of f-divergences. Moreover we easily see that f-PIDs can be rewritten as the expectation of $f(q/p)$ with respect to the probability distribution p_g as follows,

$$D_f^g(p\|q) = \left[\int_{\mathcal{X}} p(s)g[p(s)]d\mu(s)\right] E_{p_g}[f\left(\frac{q}{p}\right)], \tag{6}$$

where

$$p_g(x) = \frac{1}{\int_{\mathcal{X}} p(s)g[p(s)]d\mu(s)}p(x)g[p(x)]. \tag{7}$$

The formulation using the expectation [the right side of eq. 6] becomes to be important in order to have the lower bound of f-PID using Jensen's inequality. Therefore, we derive the basic inequality of f-PID in the following lemma. It will be used for the derivation of the family of specific distributions for f-PIDs.

Lemma 1 (Basic Inequality of f-PID) *Let $f(u)$ be a convex function defined on $(0,\infty)$ and strictly convex at $u = 1$. And $f(1) = 0$ is satisfied. In addition, we suppose that f is a monotone decreasing function on the interval $[0,1]$. For p and q, we suppose that $\int p(x)g[p(x)]d\mu(x)$ and $\int q(x)g[p(x)]d\mu(x)$ are bounded. If*

$$\int_{\mathcal{X}} p(x)g[p(x)]d\mu(x) - \int_{\mathcal{X}} q(x)g[p(x)]d\mu(x) \geq 0, \tag{8}$$

then the following inequality is established.

$$D_f^g(p\|q) \geq \left(\int_{\mathcal{X}} p(x)g[p(x)]d\mu(x)\right) f\left(\frac{\int_{\mathcal{X}} q(x)g[p(x)]d\mu(x)}{\int_{\mathcal{X}} p(x)g[p(x)]d\mu(x)}\right) \geq 0. \tag{9}$$

The proof of this lemma is easily showed by using the representation of eq. 6 and Jensen's inequality [5]. Eq. 9 is regarded as an extended version of the log sum inequality. We need to examine the condition that the second term of eq. 9 is equal to 0, because the assumption of above lemma is helpful for giving a family of specific distributions. We have the following subset of $\mathcal{P}$ using eq. 8.

$$\mathcal{S}_p \stackrel{\text{def}}{=} \{q | \int_{\mathcal{X}} p(x)g[p(x)]d\mu(x) - \int_{\mathcal{X}} q(x)g[p(x)]d\mu(x) \geq 0, q \in \mathcal{P}\} \tag{10}$$

We easily see that $p \in \mathcal{S}_p$, because of the equality of eq. 10.

4 α-PID

In this section, we introduce an example of f-PID using α-divergence eq. 3.

4.1 Definition of α-PID

We consider the following case concerning to the pair f, g in the definition of f-PID:

$$f(u) = f_\alpha(u) \stackrel{\text{def}}{=} \frac{1}{\alpha(1-\alpha)}(1-u^\alpha), \quad g(v) = g_\alpha(v) \stackrel{\text{def}}{=} v^\alpha. \tag{11}$$

where $u, v \in [0, \infty)$ and $\alpha \in [0, 1)$. Using f_α and g_α, we define α-PID as follows,

$$D_{f_\alpha}^{g_\alpha}(p\|q) \stackrel{\text{def}}{=} \frac{1}{\alpha(1-\alpha)}\left(\int_{\mathcal{X}} [p(x)]^{1+\alpha}d\mu(x) - \int_{\mathcal{X}} p(x)[q(x)]^\alpha d\mu(x)\right). \tag{12}$$

Lemma 2 *$\mathcal{G}$ is a family of specific distributions for $D_{f_\alpha}^{g_\alpha}$, where*

$$\mathcal{G} \stackrel{\text{def}}{=} \{p_\theta | p_\theta(\boldsymbol{y}) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(\boldsymbol{y}-\theta)^2\right), \theta \in \boldsymbol{R}\}. \tag{13}$$

In order to prove this lemma, we rewrite eq. 12 using Gaussian ditsributions in the following.

$$\begin{aligned} D_{f_\alpha}^{g_\alpha}(p_{\theta_1}\|p_{\theta_2}) \stackrel{\text{def}}{=} & \frac{1}{\alpha(1-\alpha)}\Big(\int_{\boldsymbol{R}} [p_{\theta_1}(\boldsymbol{y})]^{1+\alpha}d\mu(\boldsymbol{y}) \\ & - \int_{\boldsymbol{R}} p_{\theta_1}(\boldsymbol{y})[p_{\theta_2}(\boldsymbol{y})]^\alpha d\mu(\boldsymbol{y})\Big), \end{aligned} \tag{14}$$

where p_{θ_1} and p_{θ_2} are Gaussian density functions with the variance 1, both θ_1 and θ_2 represent the expectation. The proof of this lemma is easy as follows:
[Proof]: If $\theta_1 = \theta_2$, then we easily have

$$D_{f_\alpha}^{g_\alpha}(p_{\theta_1}\|p_{\theta_2}) = 0. \tag{15}$$

And the following is established.

$$\begin{aligned} D_{f_\alpha}^{g_\alpha}(p_{\theta_1}\|p_{\theta_2}) &= \frac{1}{\alpha(1-\alpha)}\Big[\int_{\boldsymbol{R}}(p_{\theta_1}(\boldsymbol{y}))^{1+\alpha}\,d\mu(\boldsymbol{y}) \\ &- \exp\{-\frac{\alpha(\theta_1-\theta_2)^2}{2(1+\alpha)}\}\int_{\boldsymbol{R}}\left(p_{\theta_\alpha^{1,2}}(\boldsymbol{y})\right)^{1+\alpha}d\mu(\boldsymbol{y})\Big] \\ &= \frac{1}{\alpha(1-\alpha)}\left(1-\exp\{-\frac{\alpha(\theta_1-\theta_2)^2}{2(1+\alpha)}\}\right) \\ &\int_{\boldsymbol{R}}\left(p_{\theta_\alpha^{1,2}}(\boldsymbol{y})\right)^{1+\alpha}d\mu(\boldsymbol{y}) \\ &\geq 0, \end{aligned} \tag{16}$$

where $\theta_\alpha^{1,2} = \frac{\theta_1+\alpha\theta_2}{1+\alpha}$. Thus we have

$$D_{f_\alpha}^{g_\alpha}(p_{\theta_1}\|p_{\theta_2}) \geq 0 \tag{17}$$

with equality iff $\theta_1 = \theta_2$. Thus $\mathcal{G}$ is the family of specific distributions for α-PIDs. [Q.E.D]

We easily have the following corollary concerning to the relation between $D_{f_\alpha}^{g_\alpha}$ and D_k.

Corollary 1 *The relation between $D_{f_\alpha}^{g_\alpha}$ and D_k is:*

$$\lim_{\alpha\to 0} D_{f_\alpha}^{g_\alpha}(p_{\theta_1}\|p_{\theta_2}) = D_k(p_{\theta_1}\|p_{\theta_2}). \tag{18}$$

4.2 Learning Algorithm of α-PID

We consider the learning algorithm concerning to the minimum α-PID. Because the specific probability distributions family for α-PID is Gauss model, its learning algorithm will be applied to information systems used the least squared method.

Let $\boldsymbol{w}$ be the parameters vector of some input-output system, $\boldsymbol{w} = (w_1, \cdots, w_m) \in \boldsymbol{R}^m$. Let the system be denoted by $h(\boldsymbol{x} : \boldsymbol{w})$, where $\boldsymbol{x}$ $(\in \boldsymbol{R}^l)$ and $\boldsymbol{y}$ $(\in \boldsymbol{R})$ are an input vector and the output of the system respectively. Let $\boldsymbol{y}^*$ be the output of unknown target system $h(\boldsymbol{x}; \boldsymbol{w}^*)$ where $\boldsymbol{w}^*$ is unknown target parameters. Let $\boldsymbol{W}$ be a subset of $\boldsymbol{R}^m$.

When some input $\boldsymbol{x}$ is given at the system $h(\ :\ \boldsymbol{w})$, we have the conditional probability function with respect to the output $\boldsymbol{y}$ in the following:

$$p(\boldsymbol{y}|h(\boldsymbol{x} : \boldsymbol{w})) \stackrel{\text{def}}{=} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\{\boldsymbol{y} - h(\boldsymbol{x} : \boldsymbol{w})\}^2\right). \tag{19}$$

The negative log-likelihood function of eq. (19) is well-used loss function. In the fact, we easily have

$$-\log p(\boldsymbol{y}|h(\boldsymbol{x} : \boldsymbol{w})) = \frac{1}{2}\{\boldsymbol{y} - h(\boldsymbol{x} : \boldsymbol{w})\}^2 + \text{constant}. \tag{20}$$

From the view of the minimization of an information divergence, the minimum squared error loss can be explained by the minimum Kullback divergence, that is,

$$\arg\min_{\boldsymbol{w}\in\boldsymbol{W}} D_k(p_{\boldsymbol{w}^*}\|p_{\boldsymbol{w}}) = \hat{\boldsymbol{w}}_k, \tag{21}$$

where $p_{\boldsymbol{w}} = p(\boldsymbol{y}|h(\boldsymbol{x}:\boldsymbol{w}))$. As the same way, we easily have the minimization of α-PIDs with respect to the system parameters $\boldsymbol{w}$ in the following.

$$\arg\min_{\boldsymbol{w}\in\boldsymbol{W}} D^{g_\alpha}_{f_\alpha}(p_{\boldsymbol{w}^*}\|p_{\boldsymbol{w}}) = \hat{\boldsymbol{w}}_\alpha, \quad \alpha\in[0,1) \tag{22}$$

The gradient function of $D^{g_\alpha}_{f_\alpha}(p_{\boldsymbol{w}^*}\|p_{\boldsymbol{w}})$ with respect to the network weights vector $\boldsymbol{w}$ is:

$$\begin{aligned}\nabla_{\boldsymbol{w}} D^{g_\alpha}_{f_\alpha}(p_{\boldsymbol{w}^*}\|p_{\boldsymbol{w}}) &= \frac{-\nabla_{\boldsymbol{w}}}{\alpha(1-\alpha)}\int_{\boldsymbol{R}} p(\boldsymbol{y}|h(\boldsymbol{x}:\boldsymbol{w}^*))[p(\boldsymbol{y}|h(\boldsymbol{x}:\boldsymbol{w})]^\alpha d\mu(\boldsymbol{y}) \\ &= E_{p^*}\left[\frac{\exp\{-\alpha l(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})\}}{1-\alpha}\nabla_{\boldsymbol{w}} l(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})\right].\end{aligned} \tag{23}$$

Thus the learning algorithm using the minimum α-PID is the following.

$$\boldsymbol{w}^{n+1} = \boldsymbol{w}^n - \epsilon\left(\frac{1}{\sqrt{2\pi}}\right)^\alpha \frac{\exp\{-\alpha l\}}{(1-\alpha)}\nabla_{\boldsymbol{w}} l \tag{24}$$

where n is update frequency using the learning procedure and l is the square error loss.

By using the weights update [eq. 24], we easily have the direct on-line type learning algorithm. However, an effective on-line learning algorithm using the minimum α-divergence estimator has not been derived, because the gradient function $\nabla_{\boldsymbol{w}} D_\alpha(p_{\boldsymbol{w}^*}\|p_{\boldsymbol{w}})$ can not be expressed in the expectation with respect to the target probability distribution $p_{\boldsymbol{w}}$. That is, there does not exist the function F (defined on $\boldsymbol{R}$) satisfying $\nabla_{\boldsymbol{w}} D_\alpha(p_{\boldsymbol{w}^*}\|p_{\boldsymbol{w}}) = E_{p_{\boldsymbol{w}^*}}[F(p_{\boldsymbol{w}})]$.

We describe the Hellinger distance which is denoted by $D_H = 0.5D_{\alpha=0.5}$. In the theoretical framework of the minimum Hellinger distance estimation, the target data distribution is generated by the empirical data using the smoothing technique [2]. Its distribution is used as target probability distribution in the process of calculating the estimation function for the minimum Hellinger distance estimator.

5 Numerical Experiments

In the previous section, we introduced α-PID as an illustrative example of f-PID. We examine the effectiveness of the learning procedure using α-PIDs in this section.

5.1 Training Multi-layer Perceptrons

Let the input-output system be the multi-layer perceptron (MLP), which consists the input, hidden and output layers. We suppose that all MLPs used in this study have the same network architecture. The activation function of each network node is $(1 - \exp[-u])/(1 + \exp[-u])$, where u is the linear sum of the inputs and the network weight vectors.

As the input-output system used the least squared method is regarded as Gaussian distribution in the meaning of the stochastic modeling,

$$p_{h(\boldsymbol{x}:\boldsymbol{w})}(\boldsymbol{y}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\{\boldsymbol{y} - h(\boldsymbol{x} : \boldsymbol{w})\}^2\right), \tag{25}$$

where $h(\boldsymbol{x} : \boldsymbol{w})$ is the output of the MLP with the parameters $\boldsymbol{w}$ and the input $\boldsymbol{x}$. The least squared method is well used for training MLPs.

However, the model of MLPs is not parametric, that is, there exist various network weight vectors expressing the target MLP. Let us consider the learning of the training data from the MLP with the weights $\boldsymbol{w}^*$. We suppose that an initial network weights vector is in the neighborhood of $\boldsymbol{w}^*$. Let $\hat{\boldsymbol{w}}_N$ be the weights vector of the trained MLP obtained by either the least squared method or α-PID method, where N is the number of the training data. Then $\hat{\boldsymbol{w}}_N$ converges to $\boldsymbol{w}^*$ as $N \to \infty$. In the numerical experiments in this section, we do not see the difference between the $\boldsymbol{w}^*$ and the weights vector of the trained MLP, but examine the ability of representation of the trained MLP for the training data. In addition, we avoid an excessive adjustment to the training data with the proper setting of experiments.

The essential problems of the learning and generalization of multi-layer perceptrons are very important and have been studied from a mathematical approach ([12] and so on).

5.2 Mixture Model of MLPs

The learning problem of this simulation is to train MLPs from the gross error contaminated data. Let $\boldsymbol{w}^*$ be the weights vector of the target MLP and $\boldsymbol{w}_z$ the weights vector of the noise MLP.

The mixture model of two MLPs can be represented in the following:

$$p^t_{h(\boldsymbol{x}:\boldsymbol{w}),h(\boldsymbol{x}:\boldsymbol{w}_z)}(\boldsymbol{y}) = (1-t)p_{h(\boldsymbol{x}:\boldsymbol{w})}(\boldsymbol{y}) + tp_{h(\boldsymbol{x}:\boldsymbol{w}_z)}(\boldsymbol{y}) \tag{26}$$

where $0 \leq t \leq 1$. The mixture distribution [eq. 26] is called the gross error model [8]. The training data used in this simulation consists the target data and the noise data with the mixture probability $p(1-t, t)$. The number of the hidden unites is determined without an excessive adjustment to the training data.

Using the each noise rate t and the noise parameter $\boldsymbol{B}$, we examine the α-PID learning and the least squared learning from the noise contaminated data. We use 10 kinds of the initial weights vectors. The details of the setting of numerical experiments are showed in Table 1.

Table 1: Setting of experiments

Architecture of MLP (input-hidden-output)	2-2-1
Number of the taining data	100
α	$0 \leq \alpha \leq 1$
Mixture rate	t=0.3, 0.1, 0.01
Weights of the noise MLP	$\boldsymbol{w}_z = \boldsymbol{w}^* + \boldsymbol{B}$ $\boldsymbol{B} = (B, \cdots, B), B > 0$
Stop condition of the learning	$(\nabla D_{f_\alpha}^{g_\alpha})^2 < 0.001$
Training rate	$\epsilon = 0.01$

5.3 Results

We examine the error of the trained MLP for the target data. The results are showed in Figure 1, Figure 2 and Figure 3. We explain about the graph points in these figures using the general notation "**-##.dat". "**" indicates the percent rate of the target data in the training data. When "##" is "kl", the result is obtained by the least squared method. If "##" is a number ($\neq 0$), the result is obtained by the α-PID learning method with α =##. We use the log scale at the average error because of the large amount of the error.

In the case of using the least squared method, there is an immediate increase for the error in the proportion to $\boldsymbol{B}$. This means that the difference between the target output and the output of the noise MLP becomes to be large. However, the phenomenon of an excessive adjustment to the training data reduces such the difference. In this simulation, the setting of experiments is considered to avoid such the phenomenon.

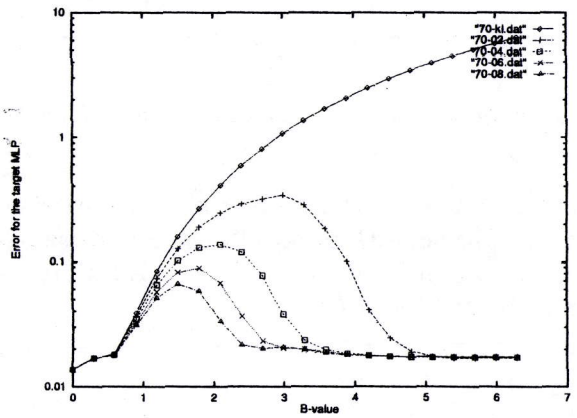

Fig. 1: average error for the target data in the case of the noise rate 1%.

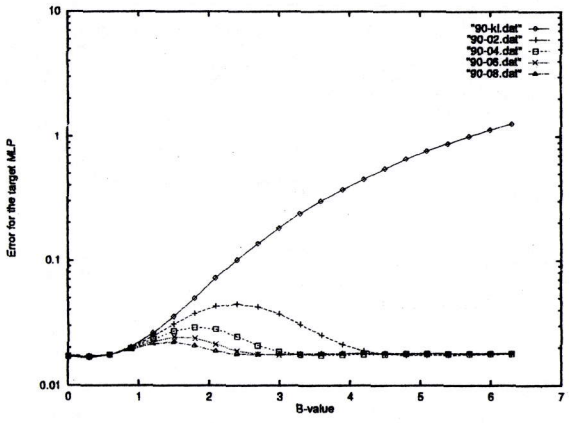

Fig. 2: average error for the target data in the case of the noise rate 10%.

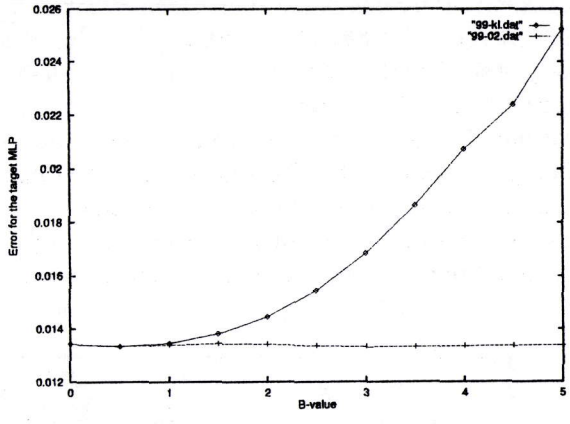

Fig. 3: average error for the target data in the case of the noise rate 1%.

We can see that the trained MLP by using the α-PID learning ($\alpha \neq 0$) becomes to fit the target data for the large $\boldsymbol{B}$. Especially there exists an influence in the case of the least squared learning in Figure 3, but the α-PID learning is successful to get rid of an influence of the gross error noise.

6 Conclusion

In this paper, we derive pseudo information divergences (PIDs), and derive the fundamental properties of the PIDs. We give a kind of PIDs using f-divergences. For showing the effectiveness of our measures, we introduce α-PIDs as an example

of f-PIDs. We derive the learning algorithm procedure with respect to the minimum α-PID, and show the effectiveness of our method using the problem concerning to the learning of MLPs from the gross error contaminated data. Statistically, the effectiveness of the minimum α-PID is the same as the robustness which was studied in the minimum Hellinger distance estimator [2]. The minimum Hellinger distance estimator has not given the direct on-line learning procedure, but however, in our framework of PIDs, an example of PIDs gives the direct on-line learning with such the effectiveness.

The concept of our pseudo information divergences have not been treated too much in information theory. It is seemed to give the proper criteria for various applications of engineering. In a word, it seems that this becomes one of the methods to generate a suitable criterion for engineering applications. As the first research we have mentioned the fundamental properties of PID, but the practicing application research will be one of our future works.

Acknowledgements

This research is supported by the Grant-in-Aid for Encouragement of Young Scientists (No.13780261) of Japan Society for the Promotion and Science.

References

[1] Amari, S., (1985) Differential Geometrical Method in Statistics, Lecture Note in Statistics ,28, Springer-Verlag.

[2] Beran. R. (1977) Minimum Hellinger distance estimators for parametric models, Ann. Statist. 5, pp. 445-463.

[3] Beran. R. (1978) A efficent and robust adaptive estimator of location, Ann. Statist. 5, pp.292-313.

[4] Cover, T. M.,& Thomas, J. A.(1991) Element of Information Theory, Wiley-Interscience publication.

[5] Csiszár, I., (1966) A note on Jensen's inequality, Studia Sci. Math. Hungar.1, pp. 227-230.

[6] Csiszár, I., (1967) On information-type measure of difference of probability distributions and indirect observations, Studia Sci. Math. Hungar. ,2, pp. 299-318.

[7] Csiszár,I., (1967) Topological property of f-divergence, Studia Sci. Math. Hungar. ,2 , pp. 330-339.

[8] Jureckova, J., & Sen, P. K. (1995) Robust Statistical Procedures Asymptotics and Interrelations, Wiley Inter Science.

[9] Hinton,G.E., and Sejinowski,T.J., (1986) Learning and Relearning in Boltzman Machines, in Parallel Distributed Processing Vol. 1, pp. 282-317, MIT Press.

[10] Kullback, S. (1959), Information Theory and Statistics. Wiley, New York.

[11] Rényi, A. (1961) On measuers of entropy and information, Proceeding of the Fourth Berkeley Symposium on Mathematical Statices and Probability, Vol.1, Berkeley, pp.541-561.

[12] Wananabe, S.,(2001) Algebraic analysis for non-identifiable learning machines, Neural Computation, Vol.13, No.4, pp.899-933.