北海道大学工学部研究報告
第175号（平成7年）

Bulletin of the Faculty of Engineering
Hokkaido University, No. 175 (1995)

# 統計的学習と情報量の関係についての考察

塩谷浩之　小松隆行
伊達悼　佐藤義治

（平成7年6月23日受理）

# A Study on the Relations between the Stochastic Learning and Information Divergences

Hiroyuki SHIOYA, Takayuki KOMATSU, Tsutomu DA-TE and Yoshiharu SATO

(Received June 23, 1995)

**Abstract**

In computational learning theory, PAC learning (Probably Approximately Correct Learning) was suggested by Valiant in 1984. An improved PAC learning with stochastic rules is called stochastic PAC learning. The evaluation of this learning (the convergency and sample complexity) depends on the choice of the information divergences which are the measure of distance between the target rule and the hypothesis outputted by the learning algorithm.

In this paper, we study the evaluation of learning for each case of several information divergences, and discuss the mathematical relations between the evaluation of the learning and the choice of information divergences.

## 1. はじめに

計算論的学習理論において，近似同定学習のモデルとしてPAC学習(Probably Approximately Correct Learning)が，Valiantによって提唱された[12]。その学習は，真の規則を学習結果として求めるのではなく，ある程度の範囲の誤差を許容する規則を学習結果として求める，という基準を持つ。

PAC学習における学習対象が，決定的規則であったのに対して，学習の対象を確率的規則に拡張した統計的学習は，確率的PAC学習[13]と呼ばれる。この学習の評価（収束，サンプル複雑度）は，学習目的である真の規則と学習アルゴリズムの出力である仮説とを測る損失関数として用いる情報量の選び方に依存する。そこで用いられる情報量とは，二つの確率分布の違いを測る量である。そのような情報量には，カルバック情報量(Kullback divergence)[7]，ヘリンガー距離(Hellinger distance)，$L_1$距離(variation distance)などがあり，それらが，統計学や情報理論などにおいて果たしてきた役割は重要である。

本報告では，種々の情報量を損失関数として用いた場合の学習の評価について述べ，新たに各

システム情報工学専攻　数理情報工学講座

情報量に依存した学習係数のみを用いた学習評価の不等式を示し，学習の評価と情報量の関係について考察する。

## 2．情　報　量

### 2．1　種々の情報量

集合 X と X 上の $\sigma$-有限測度 $\mu$ に関する確率測度全体を $\Omega$ とおく，すなわち，

$$\Omega \stackrel{\text{def}}{=} \{p | p : X \to R, p(x) \geq 0 (\forall x \in X), \int_X p(x)\mu(dx) = 1\} \tag{1}$$

とおく。

情報量とは，確率密度関数 $p(x)$ から実数への汎関数のことであり，例えば，一つの確率密度関数の場合，シャノンのエントロピーやフィシャー情報量などが該当する。一方，二つの確率密度関数の場合，カルバック情報量，ヘリンガー距離，$L_1$ などが該当し，二つの確率密度関数の差異を測るもので，これらを総称してダイバージェンス(divergence)と呼ぶ。本研究では，後者の方の情報量を主に扱う。以下，log は底が 2 の対数で ln は自然対数となる。

カルバック情報量，ヘンリガー距離および $L_1$ 距離は，これらは，以下のように定義されている。

$$d_K(p,q) \stackrel{\text{def}}{=} \int_X p(x) \log \frac{q(x)}{p(x)} \mu(dx) \tag{2}$$

$$d_v(p,q) \stackrel{\text{def}}{=} \int_X |p(x) - q(x)| \mu(dx) \tag{3}$$

$$d_H(p,q) \stackrel{\text{def}}{=} \int_X |\sqrt{p(x)} - \sqrt{q(x)}|^2 \mu(dx) \tag{4}$$

これらの情報量を一般的な捉えるものとして，Csiszár の f-divergence がある。
それは次のように定義されている[1),3),4)]

$$d_f(p,q) \stackrel{\text{def}}{=} \int_X p(x) f\left(\frac{q(x)}{p(x)}\right) \mu(dx) \tag{5}$$

ただし，f(u)は(0,∞)上で定義される凸関数で，f(1)＝0 を満たし，u＝1 で狭義凸とする。また，以下の記法を用いることにする。

$$f^*(u) \stackrel{\text{def}}{=} uf\left(\frac{1}{u}\right),\ 0f\left(\frac{0}{0}\right) = 0,\ 0f\left(\frac{a}{0}\right) = \lim_{\epsilon \to +0} \epsilon f\left(\frac{a}{\epsilon}\right) = a \lim_{u \to \infty} \frac{f(u)}{u} \tag{6}$$

カルバック情報量，ヘリンガー距離および $L_1$ 距離は，それぞれ f(u)＝−log u, f(u)＝2(1−$\sqrt{u}$)，f(u)＝u−1とおくことで得られる。

### 2．2　情報量の評価不等式

これまでに，紹介した種々の情報量は，互いに異なる数字的性質を持つ。例えば $L_1$ 距離は，距離の公理（正値性，対称性，三角不等式）をすべて満たすが，カルバック情報量は，正値性のみ満たす。ヘリンガー距離は，正値性も対称性だけを満たす。

また，異なる二つの情報量を関係付ける，以下のような種々の評価不等式が示されている。た

だし，$e$ は，自然対数の底とし，$C_f$は，$f$ のみに依存する正定数とする。

$$d_H(p,q) \leq d_v(p,q) \tag{7}$$

$$\frac{d_v(p,q)^2}{4} \leq d_H(p,q) \tag{8}$$

$$d_H(p,q) \leq \frac{d_K(p,q)}{\log e} \tag{9}$$

$$d_f(p,q) \leq C_f\sqrt{d_v(p,q)} \tag{10}$$

さらに，f-divergence と $L_1$距離を用いた以下のような評価不等式を示す[10),11)]。

定理 I

$$d_f(p,q) \leq \left(\frac{f(0)+f^*(0)}{2}\right) d_v(p,q) \tag{11}$$

等号は，f(u)＝u－1つまり $d_f=d_v$の時に成り立つ。

この不等式は f-divergence が最大値およびその条件の導出を容易する，という有用性があることも示した。

## 3．統計的学習

ここでは，先に示した情報量の評価不等式の応用例として，統計学的学習の一つとして確率的 PAC 学習を取り上げる。

### 3．I　確率的 PAC 学習

Valiant[12)]の PAC 学習は，決定的規則を学習するものであったのに対して，確率的 PAC 学習とは，山西が確率的仮説を扱えるように，決定的規則を確率的規則に拡張したものである。

さらに，確率的規則を入力データ $X_1\cdots X_N$および出力データ $Y_1\cdots Y_N$によって定まる条件つき確率分布と見なすことによって，学習するということは，統計的パラメータ推定を行うことと同じことになる。

この学習の対称となるクラスは，次のように表される有限分割型の確率的規則のクラス，

$$\mathrm{H}_N \stackrel{\mathrm{def}}{=} \{P(Y|X:\theta \prec M):\theta \in \Theta_N(M), M \in \Lambda\} \tag{12}$$

である。そして，学習アルゴリズムには，MDL 規準(Minimum Descripton Length criterion)が導入されており，与えられた入出力データの組 $D^N=(X_1,Y_1)\cdots(X_N,Y_N)$に対して，このクラス中で式(13)(14)を満たすような仮説 $P(Y|X:\bar{\theta} \prec \hat{M}) \in \mathrm{H}_N$を出力する，ということである。

$$\hat{M} \stackrel{\mathrm{def}}{=} arg \inf_{M\in\Lambda}\{-\sum_{i=1}^{N}\log P(Y_i|X_i:\bar{\theta} \prec M) + 2l_N(\bar{\theta},M)\} \tag{13}$$

$$\bar{\theta} \stackrel{\mathrm{def}}{=} arg \min_{\theta\in\Theta_N(M)}\{-\sum_{i=1}^{N}\log P(Y_i|X_i:\theta \prec M)\} \tag{14}$$

ただし，$P(Y|X:\theta \prec M)$は $\theta$ と $M$ で規定される条件付き分布とし，$\Lambda$ は可算濃度のモデルの族とする。独立データ $D^N=(X_1,Y_1)\cdots(X_N,Y_N)$に対して，次元が m であるモデル $M$ の実数値パラメータの空間 $\Theta(M)\in[0,1]^m$を，データ数 $N$ に依存した形で，有限個の代表点で置き換えた

集合を，$\Theta_N(M)$とする。また，$\{H_N\}_{M\in\Lambda}$を仮説空間と呼ぶ。$X_i\in\Upsilon$, $Y_i\in\Delta$, $\Upsilon,\Delta$ は可算集合とし $l_N(\bar{\theta},M)$は $\bar{\theta}, M$ の記述長の和とする。$M$ の記述長を $l(M)$とすると，$l_N(\bar{\theta},M)=(m\log N)/2+(5/2)m+l(M)$となる。$l(M)$は，Kraft の不等式を満たす符号化関数で求めたモデル $M$ の符号長とする。

この学習アルゴリズムに導入されている MDL 規準は，確率分布のモデル選択の規準として用いられているもので，データ圧縮の理論から生まれたものである。この規準は，データに対する過剰適合を防ぐ性質を持っていることから，このアルゴリズムは，学習結果についても，過剰適合を防ぐ性質を持つ。

### 3．2 学習評価

学習の目標である真の規則 $P^*$は，真のモデル $M^*$に含まれているとし，$M^*$に付随する真の $m^*$ 次元のパラメータ $\theta^*$が存在して，$P(Y|X:\theta^*\prec M^*)$と書ける。データ $D^N$を入力した学習アルゴリズムの出力を $P_{DN}\in H_N$とする。$\theta^*$と $M^*$の記述長の和 $l_N(\theta^*,M^*)$を $g_NM^*$とおく。学習アルゴリズムにより得られた仮説 $\hat{P}D^N$と真の規則 $P^*$との距離を，f-divergence を用いると，次のように定義できる。

$$d_f(P^*,\bar{P}_{D^N})\stackrel{\text{def}}{=}\sum_{X\in\Upsilon}Q(X)\sum_{Y\in\Delta}P^*(Y|X)f\left(\frac{\hat{P}_{D^N}(Y|X)}{P^*(Y|X)}\right) \tag{15}$$

また以下では，$QP_N^*\stackrel{\text{def}}{=}\Pi_iQ(X_i)\Pi_iP(Y_i|X_i)$とおく。$N_f(\epsilon,\delta)$を，確率 $QP_N^*[D^N:d_f(P^*,\hat{P}_{D^N}]$が $\delta$ 以内になるための最小必要サンプル数とする。学習の評価において，これの上限が問題となる。$\hat{P}_{D^N}$と真の規則 $P^*$の距離に，ヘンガー距離と $L_1$距離を用いた場合は，次のように定義される。

$$d_H(P^*,\hat{P}_{D^N})\stackrel{\text{def}}{=}\sum_{X\in\Upsilon}Q(X)\sum_{Y\in\Delta}|\sqrt{P^*(Y|X)}-\sqrt{\hat{P}_{D^N}(Y|X)}|^2 \tag{16}$$

$$d_v(P^*,\hat{P}_{D^N})\stackrel{\text{def}}{=}\sum_{X\in\Upsilon}Q(X)\sum_{Y\in\Delta}|P^*(Y|X)-\hat{P}_{D^N}(Y|X)| \tag{17}$$

このような設定のもとで，山西は，確率的 PAC 学習の評価を次のように示した[13]。

**定理 2** 任意の $P(Y|X:\theta^*\prec M^*)$と任意の $\epsilon>0$ に対して，次が成り立つ。

$$QP_N^*[D^N\ Dd_H(P^*,\hat{P}_{D^N})\geq\epsilon]\leq\exp\left[-\frac{N\epsilon}{2}+\left(g_N(M^*)+\frac{m^*}{2}\right)ln2\right] \tag{18}$$

$$NP_N^*[D^N:d_v(P^*,\bar{P}_{D^N})\geq\epsilon]\leq QP_N^*\left[d_H(P^*,\hat{P}_{D^N})\geq\left(\frac{\epsilon}{2}\right)^2\right]$$

$$\leq\exp\left[-\frac{N\epsilon^2}{8}\left(g_N(M^*)+\frac{m^*}{2}\right)ln2\right] \tag{19}$$

$0<\delta<1$ に対して，ヘリンガー距離と $L_1$距離における最小必要サンプル数をそれぞれ $N_H(\epsilon,\delta)$ $N_v(\epsilon,\delta)$とすると次のようになる。

$$N_H(\epsilon,\delta)\leq\frac{e}{\epsilon(e-1)}\left(m^*\ln\frac{64m^*}{\epsilon}+(2\ln2)\,l(M^*)+2\ln\frac{1}{\delta}\right) \tag{20}$$

$$N_v(\epsilon,\delta)\leq\frac{e}{\epsilon^2(e-1)}\left(m^*\ln\frac{256m^*}{\epsilon^2}+(2\ln2)\,l(M^*)+2\ln\frac{1}{\delta}\right) \tag{21}$$

f-divergence を損失関数として用いた場合の学習評価について述べる。f-divergence の評価不

等式として，式(10)を用いた場合と式(11)を用いた場合の学習評価は，次のように示される。[10]。

定理 3

$$QP_N^*[D^N : d_f(P^*, \hat{P}_{D^N}) \geq \epsilon] \leq QP_N^*\left[ d_v(P^*, \hat{P}_{D^N}) \geq \left(\frac{\epsilon}{C_f}\right)^2 \right]$$

$$\leq \exp\left[ -\frac{N}{8}\left(\frac{\epsilon^4}{C_f^4}\right) + (g_N(M^*) + \frac{m^*}{2})\ln 2 \right] \quad (22)$$

$$\mathrm{QP}_{\mathrm{N}}^*[\mathrm{D}^{\mathrm{N}} : \mathrm{d}_f(\mathrm{P}^*, \hat{\mathrm{P}}_{\mathrm{D}^{\mathrm{N}}}) \geq \epsilon] \leq \exp\left[ -\frac{N}{2}\left(\frac{\epsilon}{f(0)+f^*(0)}\right)^2 + \left(g_N(M^*) + \frac{m^*}{2}\right)\ln 2 \right] \quad (23)$$

それぞれの学習評価での最小必要サンプル数を $N_{f_c}(\epsilon, \delta)$，$N_{f_s}(\epsilon, \delta)$ とすると，次のようになる。

$$N_{f_c}(\epsilon, \delta) \leq \frac{4e}{e-1}\left(\frac{C_f}{\epsilon}\right)^4 \left( m^* \ln \frac{256 m^*}{\left(\frac{C_a}{\epsilon}\right)^4} + (2\ln 2)\, l(M^*) + 2\ln\frac{1}{\delta} \right) \quad (24)$$

$$N_{f_s}(\epsilon, \delta) \leq \frac{e}{e-1}\left(\frac{(f(0)+f^*(0))^2}{\epsilon^2}\right)\left( m^* ln \frac{64 m^*}{\frac{(f(0)+f^*(0))^2}{\epsilon^2}} + (2\mathrm{lnb})\, l(M^*) + 2\ln\frac{1}{\delta} \right) \quad (25)$$

以上の定理の証明は，省略する。

### 3．3　学習評価の新たな表現

これまでに，得られた学習評価不等式から，データ数 $N$ に対して，学習結果 $\hat{P}$ と真の規則 $P^*$ との情報量が $\epsilon$ 以上離れている確率が，データ数 $N$ に対して，指数オーダーでゼロに収束するということがわかる。さらに，用いる情報量によって，学習の評価不等式の右辺の指数の巾の項の $\epsilon$ が異なっていることがわかる。

そこで，$\epsilon$ に関するオーダーを示す新たな学習の評価不等式として，次を示す。

定理 4

$$\lim_{N\to\infty}\frac{1}{N}\ln QP_N^*[D^N : d_H(P^*, \hat{P}_{D^N}) \geq \epsilon] \leq -\frac{\epsilon}{2} \quad (26)$$

$$\lim_{N\to\infty}\frac{1}{N}\ln QP_N^*[D^N : d_v(P^*, \hat{P}_{D^N}) \geq \epsilon] \leq -\frac{\epsilon^2}{8} \quad (27)$$

$$\lim_{N\to\infty}\frac{1}{N}\ln QP_N^*[D^N : d_f(P^*, \hat{P}_{D^N}) \geq \epsilon] \leq -\frac{\epsilon^2}{2(f(0)+f^*(0))^2} \quad (28)$$

$$\lim_{N\to\infty}\frac{1}{N}\ln QP_N^*[D^N : d_f(P^*, \hat{P}_{D^N}) < -\epsilon] \leq -\left(\frac{\epsilon^4}{8C_f^4}\right) \quad (29)$$

証明

先の結果から，

$$QP_N^*[D^N : d_H(P^*, \hat{P}_{D^N}) \geq \epsilon] \leq \exp\left[ -\frac{N\epsilon}{2} + (g_N(M^*) + \frac{m^*}{2}\ln 2 \right] \quad (30)$$

となる。そして，ln の単調性から，次が得られる。

$$\ln QP_N^*[D^N : d_H(P^*, \hat{P}_{D^N}) \geq \epsilon] \leq \left[ -\frac{N\epsilon}{2} + (g_N(M^*) + \frac{m^*}{2})\ln 2 \right] \quad (31)$$

両辺を $N$ で割って，極限($N \to \infty$)をとることで，次のように，定理の不等式(26)が得られる。

$$\lim_{N\to\infty}\frac{1}{N}\ln QP_N^*[D^N : d_H(P^*, P_{D^N}) \geq \epsilon] \leq -\frac{\epsilon}{2} \tag{32}$$

不等式(27)，(28)，(29)も同様の方法で得られる。(証明終)

これまでに得られた学習の評価不等式から，さらに，適用する情報量によって異なる学習の係数 $\epsilon$ に関するオーダーを表現する定理を導出した。これは，大偏差原理の方法を導入したものである。その原理の一例としては，中心極限定理がよく知られている。

## 4．ま と め

本論文では，確率的PAC学習の学習評価と情報量の関係について考察した。先の定理2および定理3において，学習の評価不等式に用いる情報量によって，学習の評価不等式が異なることを示した。

さらに，これまで，得られた学習評価から，用いる係数 $\epsilon$ に関する学習の評価不等式の収束のオーダーを直接に表現する，新たな補題を示した。その結果，$\lim_{N\to\infty}(1/N)\ln QP_N^*[D_N : d(P^*, \hat{P}) \geq \varepsilon]$ を評価する，という新たな学習の評価方法を示した。この学習の評価方法から，係数 $\epsilon$ のオーダーから評価不等式の良さがわかる。今後の課題として，大偏差原理を用いた学習の数理的性質の解明，学習の効率の改善，他の学習アルゴリズムの開発などが，挙げられる。

## 参考文献


1) Csiszár, I., "Eine informationstheoretische Ungleichung und ihre Anwendung auf Beweis der Ergodizitaet von Markoffschen Ketten, Magyer Tud. Akad. Mat. Kutato Int. Koezl. 8, pp. 85-108(1964).
2) Csiszár, I., "A note on Jensen's inequality, Studia Sci. Math. Hungar. 1, pp. 227-230(1966)
3) Csiszár, I., "On information-type measure of difference of probability distributions and indirect observations, "Studia Sci. Math. Hungar., 2, pp.299-318(1967).
4) Csiszár, I., "Topological property of f-divergence" Studia Sci. Math. Hungar. 2, pp.330-339(1967).
5) Haussler, D.,&Long, P., "A generalization of Sauer's lemma. Technical Report UCSCCRL-90-15, University of Califonia at Santa Cruz(1990).
6) Kraft, C. "Some conditions for consistency and uniform of statistical procedures" University of California Publication in Statistics, 2, pp. 125-141(1955).
7) Kullback, S. "Information Theory and Statistics." Wiley, New York(1959).
8) Pitman, E. J. G. "Some Basic Theory for Statistical Inference" London, Chapman and Hall(1979).
9) Rissanen, J. "Stochastic Complexity in Statistical Inquiry" World Scientific, Series in Comp. Sci., Vol. 15(1989).
10) 塩谷浩之，長岡浩司，伊達　惇："f-divergence に関する新しい不等式と最大値および学習問題への応用"，電子情報通信学会論文誌（A分冊）Vol. J77-A, No. 4, pp. 720-726(1994).
11) 塩谷浩之，伊達　惇：Lin情報量の一般化および新しい情報量の導出，電子情報通信学会論文誌(A分冊)Vol. J77-A, No. 12, pp. 1733-1737(1994).
12) Valiant, L. G. "A theory of the learnable," Commun. of the ACM. Vol.27, pp. 1134-1142(1984).
13) Yamanishi, K., "A learning criterion for stochastic rules." Machine Learning, 9, Special Issue for COLT-90, pp. 165-203(1992).