PAPER *Special Section on Information Theory and Its Applications*

Almost Sure and Mean Convergence of Extended Stochastic Complexity

Masayuki GOTOH[†], **Toshiyasu MATSUSHIMA**[†], ***and*** **Shigeichi HIRASAWA**[†], ***Members***

SUMMARY We analyze the extended stochastic complexity (ESC) which has been proposed by K. Yamanishi. The ESC can be applied to learning algorithms for on-line prediction and batch-learning settings. Yamanishi derived the upper bound of ESC satisfying uniformly for all data sequences and that of the asymptotic expectation of ESC. However, Yamanishi concentrates mainly on the worst case performance and the lower bound has not been derived. In this paper, we show some interesting properties of ESC which are similar to Bayesian statistics: the Bayes rule and the asymptotic normality. We then derive the asymptotic formula of ESC in the meaning of almost sure and mean convergence within an error of $o(1)$ using these properties.

key words: *extended stochastic complexity, stochastic complexity, Bayesian statistics, asymptotic normality*

1. Introduction

J. Rissanen proposed the notion of stochastic complexity (SC) [14] which is a criterion to measure the amount of information in a given data sequence and is based on the concept of minimum description length [11]. This is defined by the minimum average codelength for a prior density on the parameter space when the parametric model class is given. It is known that SC is equivalent to the codelength of the mixture code on the parameter space and its loss is measured by the logarithmic function.

Recently, K. Yamanishi has proposed the extended stochastic complexity (ESC) which is a generalized version of SC and demonstrated its effectiveness in learning algorithms for on-line prediction and batch-learning settings [19]. Yamanishi concentrates mainly on the aggregating algorithm and the wost case performance. In [19], he derived the upper bound of ESC satisfying uniformly for all data sequences and that of the asymptotic expectation of ESC. However, the asymptotic formula of ESC satisfying uniformly for all data sequences within an error of order $o(1)$ have not been derived. And, Yamanishi has not derived the tight lower bound of ESC. This study is led and encouraged by the existance of this problem which was pointed out in [20] as a future research work.

In this paper, we show an interesting property in ESC like Bayesian statistics: the Bayes rule and the asymptotic normality. Using this property, we derive the asymptotic formulas of ESC within an error of $o(1)$ in the meaning of almost sure and mean convergence. Although K. Yamanishi discussed the upper bound of ESC which holds uniformly for all individual sequences, we discuss the almost sure convergence of ESC. From the discussion of almost sure convergence, we can derive the asymptotic formula of ESC within an error of $o(1)$. Assuming that the deta sequence is emitted from a source with the true probability distribution, the asymptotic formula which holds almost surely is an evaluation of ESC. Moreover, we derive the asymptotic expectation of ESC within an error of $o(1)$. From this result, we can show that Yamanishi's upper bound of asymptotic expectation of ESC is tight within an error of $o(1)$.

Manuscript received January 25, 1999.
Manuscript revised April 26, 1999.

[†]The authors are with the Department of the Industrial and Management Systems Engineering, School of Science and Engineering, Waseda University, Tokyo, 169–8555 Japan.

2. Preliminaries

2.1 Stochastic Complexity

We denote random variables on $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ by X, Y, and Z, where $Z = (X, Y)$. $Z^n = (X, Y)^n$ is a random variable on $\mathcal{Z}^n = \underbrace{\mathcal{Z} \times \mathcal{Z} \times \cdots \times \mathcal{Z}}_{n}$. Let $z^n = z_1 z_2 \cdots z_n \in \mathcal{Z}^n$ be an independently and identically distributed (i.i.d.) data sequence with length n from a source with the true probability distribution $p^*(z^n) = p^*(Z^n = z^n) = \prod_{i=1}^n p^*(Z_i = z_i)$, where $z_i = (x_i, y_i)$, $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$, and $z_i \in \mathcal{Z}$, $i = 1, 2, \cdots, n$. A class of probability models is given by $\mathcal{P}^k = \{p_{\theta^k}(Y|X) : \theta^k \in \Theta^k\}$, where θ^k is a k-dimensional parameter and Θ^k is a subset of the k-dimensional Euclidean space $\mathcal{R}^k$. $p_{\theta^k}(Y|X)$ is a probability mass or probability density of Y conditioned by X. Let $\pi(\theta^k)$ be a prior density on Θ^k.

Although several classes of SC have been proposed, we consider the following formula [13].

$$SC(y^n|x^n) = -\log q(y^n|x^n), \tag{1}$$

$$q(y^n|x^n) = \int p_{\theta^k}(y^n|x^n)\pi(\theta^k)d\theta^k. \tag{2}$$

We can regard $-\log q(y^n|x^n)$ as an ideal codelength of y^n conditioned by x^n. Here, we have

$$-\log q(y^n|x^n) = -\sum_{t=1}^{n} \log q(y_t|x_t, z^{t-1}), \tag{3}$$

$$q(y_t|x_t, z^{t-1}) = \int p_{\theta^k}(y_t|x_t)\pi(\theta^k|z^{t-1})d\theta^k, \tag{4}$$

where $\pi(\theta^k|z^{t-1})$ is the Bayes posterior density of θ^k which is given by

$$\pi(\theta^k|z^{t-1}) = \frac{\prod_{i=1}^{t-1} p_{\theta^k}(y_i|x_i)\pi(\theta^k)}{\int \prod_{i=1}^{t-1} p_{\theta^k}(y_i|x_i)\pi(\theta^k)d\theta^k}. \tag{5}$$

Therefore, the cumulative codelength of the predictive code is equivalent to that of the non-predictive code.

2.2 Extended Stochastic Complexity

At first, we define a hypothesis class $\mathcal{H}$, which may be a set of functions $f_{\theta^k}(X) : \mathcal{X} \to \mathcal{Y}$ written as $\mathcal{H} = \{f_{\theta^k}(X) : \theta^k \in \Theta^k\}$ or a set of conditional probability distributions $p_{\theta^k}(Y|X)$ written as $\mathcal{H} = \{p_{\theta^k}(Y|X) : \theta^k \in \Theta^k\}$. Here θ^k is a k-dimensional parameter and Θ^k is a subset of the k-dimensional Euclidean space $\mathcal{R}^k$.

Let L: $\mathcal{Z} \times \mathcal{H} \to [0, +\infty)$ be a loss function and $L(Y, D)$, a loss value for predicting Y with a decision D. Here, D is an element of a finite set, a subset of $\mathcal{R}^1$, or a set of probability density functions or probability mass functions over $\mathcal{Y}$. When a hypothesis class is a set of real valued functions $\mathcal{H} = \{f_{\theta^k}(X) : \theta^k \in \Theta^k\}$, the loss function is given by $L(Y, f_{\theta^k}(X))$. For example, the *α-loss function* is defined by

$$L(Y, f_{\theta^k}(X)) = |Y - f_{\theta^k}(X)|^\alpha, \quad (\alpha \geq 1), \tag{6}$$

and the *0-1 loss function* is defined by

$$L(Y, f_{\theta^k}(X)) = \begin{cases} 0 & ; Y = f_{\theta^k}(X), \\ 1 & ; Y \neq f_{\theta^k}(X). \end{cases} \tag{7}$$

When a hypothesis class is a set of conditional probability distributions $\mathcal{H} = \{p_{\theta^k}(Y|X) : \theta^k \in \Theta^k\}$, the loss function is given by $L(Y, p_{\theta^k}(\cdot|X))$. For example, the *$\alpha$-loss function* is defined by

$$L(Y, p_{\theta^k}(\cdot|X)) = \big(1 - p_{\theta^k}(Y|X)\big)^\alpha, \tag{8}$$

and the *logarithmic loss function* is given by

$$L(Y, p_{\theta^k}(\cdot|X)) = -\log p_{\theta^k}(Y|X). \tag{9}$$

Other examples of the loss functions are shown in [19]. Hereafter, we denote the loss $L(Y, f_{\theta^k}(X))$ or $L(Y, p_{\theta^k}(\cdot|X))$ by $L(Z : f_{\theta^k})$.

For a given number $\lambda > 0$, the extended stochastic complexity (ESC) of z^n is defined by [19], [20]

$$ESC(z^n) = -\frac{1}{\lambda}\log \int \exp\left(-\lambda \sum_{t=1}^{n} L(z_t : f_{\theta^k})\right)\pi(\theta^k)d\theta^k. \tag{10}$$

Here, the following lemma is shown in [19].

Lemma 1: For any z^n, the ESC of z^n for a given $(\mathcal{P}, \mathcal{H}, \pi, L)$ can be written as follows:

$$ESC(z^n) = \sum_{t=1}^{n} \bar{L}(y_t|x_t, z^{t-1}), \tag{11}$$

where

$$\bar{L}(y_t|x_t, z^{t-1}) = -\frac{1}{\lambda}\log \int \exp\big(-\lambda L(z_t : f_{\theta^k})\big)\pi(\theta^k|z^{t-1})d\theta^k, \tag{12}$$

$$\pi(\theta^k|z^{t-1}) = \frac{\exp\left(-\lambda \sum_{j=1}^{t-1} L(z_j : f_{\theta^k})\right)\pi(\theta^k)}{\int \exp\left(-\lambda \sum_{j=1}^{t-1} L(z_j : f_{\theta^k})\right)\pi(\theta^k)d\theta^k}, \tag{13}$$

where $\pi(\theta^k|z^0) = \pi(\theta^k)$ and $L(z_0 : f_{\theta^k}) = 0$. □

Using this lemma, the ESC can be applied to the aggregating algorithm for on-line learning. Although $\pi(\theta^k|z^{t-1})$ is a probability density on Θ^k, it is not generally equivalent to the Bayes posterior density. If we use the logarithmic loss function and $\lambda = 1$, then

$$\bar{L}(y_t|z^{t-1}, x_t) = -\log \int p_{\theta^k}(y_t|x_t)p(\theta^k|z^{t-1})d\theta^k, \tag{14}$$

$$p(\theta^k|z^{t-1}) = \frac{\prod_{j=1}^{t-1} p_{\theta^k}(y_j|x_j)\pi(\theta^k)}{\int \prod_{j=1}^{t-1} p_{\theta^k}(y_j|x_j)\pi(\theta^k)d\theta^k}, \tag{15}$$

where $p(\theta^k|z^{t-1})$ is the Bayesian posterior density of θ^k.

3. Main Results: Analysis of ESC

At first, we define $h(x^n)$ and $h(z^n|\theta^k)$ by

$$h(z^n) = \int \exp\left(-\lambda \sum_{t=1}^{n} L(z_t : f_{\theta^k})\right)\pi(\theta^k)d\theta^k, \tag{16}$$

and

$$h(z^n|\theta^k) = \exp\left(-\lambda \sum_{t=1}^{n} L(z_t : f_{\theta^k})\right), \tag{17}$$

respectively. Then, the following lemma can be derived.

Lemma 2: For $h(z^n)$, $h(z^n|\theta^k)$, and $\pi(\theta^k|z^n)$, the following equation is satisfied.

$$h(z^n) = \frac{h(z^n|\theta^k)\pi(\theta^k)}{\pi(\theta^k|z^n)}. \tag{18}$$

(Proof) This is obvious from definition. □

Equation (18) is a generalized version of the Bayes rule

$$p(z^n) = \frac{p(z^n|\theta^k)\pi(\theta^k)}{p(\theta^k|z^n)}. \tag{19}$$

From (18), we have

$$\begin{aligned} &ESC(z^n) \\ &\quad = -\frac{1}{\lambda}\log h(z^n|\theta^k)\pi(\theta^k) + \frac{1}{\lambda}\log \pi(\theta^k|z^n). \end{aligned} \tag{20}$$

Therefore if the asymptotic formula of $\log \pi(\theta^k|z^n)$ is shown, we can analyze $ESC(z^n)$.

Next, we show an interesting and important property of $\pi(\theta^k|z^n)$. For many practical probability models, the Bayes posterior density converges to the normal distribution [1], [3]. We call this property *asymptotic normality*. We can also show the asymptotic normality of $\pi(\theta^k|z^n)$.

At first, we define the information matrices $I^*(\theta^k)$ and $J^*(\theta^k)$ as follows:

$$\begin{aligned} I^*(\theta^k) &= -E^*\left[\frac{\partial^2 \log h(Z|\theta^k)}{\partial\theta^k(\partial\theta^k)^T}\right] \\ &= \lambda E^*\left[\frac{\partial^2 L(Z:f_{\theta^k})}{\partial\theta^k(\partial\theta^k)^T}\right], \end{aligned} \tag{21}$$

and

$$\begin{aligned} J^*(\theta^k) &= E^*\left[\frac{\partial \log h(Z|\theta^k)}{\partial\theta^k}\frac{\partial \log h(Z|\theta^k)}{(\partial\theta^k)^T}\right] \\ &= \lambda^2 E^*\left[\frac{\partial L(Z:f_{\theta^k})}{\partial\theta^k}\frac{\partial L(Z:f_{\theta^k})}{(\partial\theta^k)^T}\right], \end{aligned} \tag{22}$$

where $E^*[\cdot]$ means the expectation by the true distribution $p^*(\cdot)$, and T is the *transpose* of a vector. Although $I^*(\theta^k)$ and $J^*(\theta^k)$ may not exist for some ture distribution, we assume that these exist for $\forall\theta^k \in \Theta^k$.

The estimators, $\hat{\theta}^k$ and $\tilde{\theta}^k$, and the optimal parameter θ^{k^*} are defined as follows:

$$\hat{\theta}^k = \arg\min_{\theta^k} \frac{1}{n}\sum_{t=1}^{n} L(z_t : f_{\theta^k}), \tag{23}$$

$$\tilde{\theta}^k = \arg\max_{\theta^k} \pi(\theta^k|z^n), \tag{24}$$

and

$$\theta^{k*} = \arg\min_{\theta^k} E^*\left[L(Z:f_{\theta^k})\right]. \tag{25}$$

Let $B_\delta(\theta^{k*})$ be the ball $B_\delta(\theta^{k*}) = \left\{\theta^k \in \Theta^k \middle| \|\theta^k - \theta^{k*}\| < \delta\right\}$ on Θ^k for $\forall\delta > 0$. Similarly, we define $B_\delta(\tilde{\theta}^k) = \left\{\theta^k \in \Theta^k \middle| \|\theta^k - \tilde{\theta}^k\| < \delta\right\}$.

At first, we show a condition which we require for prior $\pi(\theta^k)$.

Condition 1: For $\forall\theta^k \in \Theta^k$, $\exists c_1 < \pi(\theta^k) < \exists c_2$. Here c_1 and c_2 are some positive value. $\pi(\theta^k)$ is twice continuously differentiable on Θ^k.

Next, we give a list of conditions which will be needed in our derivations. These conditions have appeared in [8], p.238.

Condition 2:

(1) Θ^k is compact.
(2) θ^{k*} is unique and in the interior of Θ^k.
(3) $I^*(\theta^k)$ is continuously differentiable with respect to θ^k and $I^*(\theta^{k*})$ is a positive definite.
(4) $\frac{1}{n}\sum_{t=1}^{n} L(z_t : f_{\theta^k})$ is almost surely (a.s.) continuous on Θ^k for $\forall n \in \{1, 2, \cdots\}$.
(5) $\frac{1}{n}\sum_{t=1}^{n} L(z_t : f_{\theta^k}) \to E^*\left[L(Z : f_{\theta^k})\right]$, *a.s.* uniformly on Θ^k.
(6) $E^*\left[L(Z : f_{\theta^k})\right]$ is twice continuously differentiable on Θ^k. $\frac{1}{n}\sum_{t=1}^{n} L(z_t : f_{\theta^k})$ for $\forall n \in \{1, 2, \cdots\}$ is almost surely twice continuously differentiable on Θ^k.
(7) $$\frac{1}{n}\frac{\partial}{\partial\theta^k}\sum_{t=1}^{n} L(z_t : f_{\theta^k}) \to \frac{\partial}{\partial\theta^k}E^*\left[L(Z : f_{\theta^k})\right], \quad a.s.$$
$$\frac{1}{n}\frac{\partial^2}{\partial\theta^k(\partial\theta^k)^T}\sum_{t=1}^{n} L(z_t : f_{\theta^k}) \to E^*\left[\frac{\partial^2 L(Z:f_{\theta^k})}{\partial\theta^k(\partial\theta^k)^T}\right] = \frac{1}{\lambda}I^*(\theta^k), \quad a.s.$$
uniformly on Θ^k.
(8) (The central limit theorem) $\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\frac{\partial L(Z_t:f_{\theta^{k*}})}{\partial\theta^k}$ converges in distribution to the normal distribution $N\left(\mathbf{0}, \frac{1}{\lambda^2}J^*(\theta^k)\right)$. □

Example 1 ([19]): Let $\mathcal{X} = \{X = (X_1, X_2, \cdots, X_k)^T \in [0,1]^k : X_1^2 + X_2^2 + \cdots X_k^n \le 1\}$ and $\mathcal{Y} = [0,1]$. Let $\Theta^k = \{\theta^k = (\theta_1, \theta_2, \cdots, \theta_k)^T \in [0,1]^k : \theta_1^2 + \theta_2^2 + \cdots \theta_k^2 \le 1\}$ and $\mathcal{H} = \{f_{\theta^k}(X) = (\theta^k)^T X : X \in \mathcal{X}, \theta^k \in \Theta^k\}$. In this case, Condition 2, (1) is satisfied. We apply the quadratic loss function: $L(Y, f_{\theta^k}(X)) = (Y - f_{\theta^k}(X))^2$. Then Condition 2, (4) is obviously satisfied. From

$$I^*(\theta^k) = \lambda E^*\left[\frac{\partial^2 (Y - (\theta^k)^T X)^2}{\partial\theta^k(\partial\theta^k)^T}\right], \tag{26}$$

we have

$$I^*_{i,j} = 2\lambda E^*\left[X_i X_j\right], \tag{27}$$

where $I^*_{i,j}$ is the $i-j$-th element of $I^*(\theta^k)$. That is, $I^*(\theta^k) = 2\lambda E^*\left[XX^T\right]$. If $0 < E^*\left[(X_i)^2\right]$ for $\forall i \in \{1, 2, \cdots, k\}$, then $I^*(\theta^k)$ is a positive definite and Condition 2, (3) is satisfied. Moreover, we have

$$\begin{aligned} &E^*\left[L(Y, f_{\theta^k}(X))\right] \\ &\quad = E^*\left[\left(Y - (\theta^k)^T X\right)^2\right] \\ &\quad = E^*\left[Y^2\right] - 2\sum_{i=1}^{k}\theta_i E^*\left[Y X_i\right] \end{aligned}$$

$$+\sum_{i=1}^{k}\sum_{j=1}^{k}\theta_i\theta_j E^*\left[X_iX_j\right]. \tag{28}$$

Then we have

$$\frac{\partial E^*\left[L(Y,f_{\theta^k}(X))\right]}{\partial\theta_i} = -2E^*\left[YX_i\right]+2\sum_{j=1}^{k}\theta_j E^*\left[X_iX_j\right], \tag{29}$$

for $i=1,2,\cdots,k$. That is,

$$\frac{\partial E^*\left[L(Y,f_{\theta^k}(X))\right]}{\partial\theta^k} = -2E^*[YX]+2E^*\left[XX^T\right]\theta^k, \tag{30}$$

Therefore, if θ^{k*} exists in the interior of Θ^k, then it satisfies the normal equation:

$$E^*[XX^T]\theta^{k*}=E^*[YX]. \tag{31}$$

If we can assume that $\{E^*[XX^T]\}^{-1}$ exists, then we have $\theta^{k*}=\{E^*[XX^T]\}^{-1}E^*[YX]$. Of course, $\{E^*[XX^T]\}^{-1}E^*[YX]\in\Theta^k$ is not always satisfied. If $\{E^*[XX^T]\}^{-1}E^*[YX]\notin\Theta^k$, then the hypothesis class $\mathcal{H}=\{f_{\theta^k}(X)=(\theta^k)^TX : X\in\mathcal{X},\theta^k\in\Theta^k\}$ is unsuitable for the true distribution. Here, we assume $0<\left\|\{E^*[XX^T]\}^{-1}E^*[YX]\right\|^2<1$. This means that the hypothesis class is suitable for the true distribution. On this case, θ^{k*} uniquely exists in Θ^k and Condition 2, (2) is satisfied.

Since X and Y have the finite variances, $\frac{1}{n}\sum_{t=1}^{n}x_t(x_t)^T\to E^*[XX^T]$, *a.s.*, $\frac{1}{n}\sum_{t=1}^{n}y_tx_t\to E^*[YX]$, *a.s.*, and $\frac{1}{n}\sum_{t=1}^{n}(y_t)^2\to E^*[Y^2]$, *a.s.* [4]. On the other hand, $\frac{1}{n}\sum_{t=1}^{n}L(z_t:f_{\theta^k})$ is given by

$$\frac{1}{n}\sum_{t=1}^{n}L(z_t:f_{\theta^k})=\frac{1}{n}\sum_{t=1}^{n}\left(y_t-f_{\theta^k}(x_t)\right)^2. \tag{32}$$

Therefore, Condition 2, (5)∼(7) are obviously satisfied. Moreover, since $E^*\left[\frac{1}{\sqrt{n}}\sum_{t=1}^{n}L(Z_t:f_{\theta^{k*}})\right]=0$ and the variance of $L(Z_t:f_{\theta^{k*}})$ is finite, Condition 2, (8) is satisfied from the central limit theorem [4]. □

Example 2: Let $\mathcal{X}=\{X=(X_1,X_2,\cdots,X_k)^T\in\mathcal{R}^k\}$ and $\mathcal{Y}=\mathcal{R}^1$. Let $\Theta^k=\{\theta^k=(\theta_1,\theta_2,\cdots,\theta_k)^T\in\mathcal{R}^k : \theta_1^2+\theta_2^2+\cdots\theta_k^2\le C_\theta\}$ and $\mathcal{H}=\{f_{\theta^k}(X)=(\theta^k)^TX : X\in\mathcal{X},\theta^k\in\Theta^k\}$. We apply the quadratic loss function: $L(Y,f_{\theta^k}(X))=(Y-f_{\theta^k}(X))^2$.

From the discussion similar to Example 1, we have $\theta^{k*}=\{E^*[XX^T]\}^{-1}E^*[YX]$ when $\{E^*[XX^T]\}^{-1}$ exists. Setting that C_θ is appropriately large, θ^{k*} is unique in the interior of Θ^k. We can see that this example satisfys Condition 2 from the same discussion as Example 1. □

For other examples satisfying the above conditions, see [8]. Many practical model classes with suitable loss function satisfy the above conditions. Under these conditions, the following properties have been shown [8].

Lemma 3 ([8], p.238): If Condition 2 holds, then $\hat{\theta}^k$ uniquely exists in Θ^k and is a consistent estimator. That is,

$$\hat{\theta}^k\to\theta^{k*},\quad a.s. \tag{33}$$

□

Next we consider the distribution of $\hat{\theta}^k$. It is well known that $\hat{\theta}^k$ converges in law to the normal distribution for many practical model class. For example, see [7], [8].

Lemma 4 ([8], p.239): If Condition 2 holds, then the distribution of $\sqrt{n}(\hat{\theta}^k-\theta^{k*})$ converges in law to the normal distribution:

$$N\Big(\mathbf{0},\ \{I^*(\theta^{k*})\}^{-1}J^*(\theta^{k*})\{I^*(\theta^{k*})\}^{-1}\Big).$$

□

Next, we show the important property of $\pi(\theta^k|z^n)$.

Theorem 1: Assuming Conditions 1 and 2 and defining $\xi^k=\sqrt{n}(\theta^k-\hat{\theta}^k)$ and

$$\pi_\xi(\xi^k|z^n)=\frac{1}{\sqrt{n}^k}\pi(\theta^k|z^n), \tag{34}$$

the following equation is satisfied.

$$\pi_\xi(\hat{\xi}^k|z^n)\to\left(\frac{1}{2\pi}\right)^{k/2}\sqrt{\det I^*(\hat{\theta}^k)}\quad a.s. \tag{35}$$

where $\hat{\xi}^k=\mathbf{0}^\dagger$.

(Proof) See Appendix A. □

Moreover, we can replace $\hat{\theta}^k$ by $\tilde{\theta}^k$. That is, we have the following theorem.

Theorem 2: Assuming Conditions 1 and 2 and defining $\eta^k=\sqrt{n}(\theta^k-\tilde{\theta}^k)$ and

$$\pi_\eta(\eta^k|z^n)=\frac{1}{\sqrt{n}^k}\pi(\theta^k|z^n), \tag{37}$$

then the following equation is satisfied.

$$\pi_\eta(\tilde{\eta}^k|z^n)\to\left(\frac{1}{2\pi}\right)^{k/2}\sqrt{\det I^*(\tilde{\theta}^k)}\quad a.s. \tag{38}$$

---

†Moreover, the sequences of $h_\xi(\xi^k|z^n)$ converge to the normal distribution almost surely. That is, letting E be arbitrary rectangle on ξ^k-space, the following equation is satisfied.

$$\int_E\pi_\xi(\xi^k|z^n)d\xi^k \to\frac{\sqrt{\det I^*(\hat{\theta}^k)}}{(2\pi)^{k/2}}\int_E\exp\left\{-\frac{1}{2}\|\xi^k\|^2_{I^*(\hat{\theta}^k)}\right\}d\xi^k,\quad a.s. \tag{36}$$

where $\tilde{\eta}^k = \mathbf{0}^\dagger$.
(Proof) See Appendix B. □

Therefore, the sequence of distributions $\pi(\theta^k|z^n)$, $n = 1, 2, \cdots$, each of which is an extended version of the Bayes posterior distribution, has the property of asymptotic normality. Its asymptotic variance - covariance matrix is given by $\{I^*(\hat{\theta}^k)\}^{-1}$, which is essential to this problem. In this paper, we do not assume that the true distribution exists in the parametric model class $\mathcal{P}^k$. In this case, the distribution of the minimum loss estimator converges in law to the normal distribution with variance - covariance matrix $\{I^*(\theta^{k*})\}^{-1}J^*(\theta^{k*})\{I^*(\theta^{k*})\}^{-1}$. However, only $\{I^*(\hat{\theta}^k)\}^{-1}$ appears in the asymptotic distribution $\pi(\theta^k|z^n)$. This is the same property as the Bayesian posterior density.

Using this theorem, we have the following theorem.

Theorem 3: Under Conditions 1 and 2, the asymptotic formula of $ESC(z^n)$ is given by

$$ESC(z^n) = \sum_{t=1}^{n} L(z_t : f_{\hat{\theta}^k}) + \frac{k}{2\lambda}\log\frac{n}{2\pi} + \frac{1}{\lambda}\log\frac{\sqrt{\det I^*(\hat{\theta}^k)}}{\pi(\hat{\theta}^k)} + o(1), \ a.s. \quad (40)$$

$$= \sum_{t=1}^{n} L(z_t : f_{\tilde{\theta}^k}) + \frac{k}{2\lambda}\log\frac{n}{2\pi} + \frac{1}{\lambda}\log\frac{\sqrt{\det I^*(\tilde{\theta}^k)}}{\pi(\tilde{\theta}^k)} + o(1), \ a.s. \quad (41)$$

(Proof) This is obvious from Eqs. (17), (20), (35), and (38). □

The above theorem shows an asymptotic formula for the sequences which are almost surely emitted from the true distribution. Here, $\hat{\theta}^k$ and $\tilde{\theta}^k$ are the random variables based on the true distribution.

Using Lemmas 3 and 4, we have the following lemma.

Lemma 5 ([8], pp.240–241): Under Condition 2, we have

$$E^*\left[\log\frac{h(Z^n|\hat{\theta}^k)}{h(Z^n|\theta^{k*})}\right] \to \frac{TrJ^*(\theta^{k*})\{I^*(\theta^{k*})\}^{-1}}{2\lambda}, \quad (42)$$

where Tr means the *trace* of a matrices. □

From the above discussion, we have the asymptotic formula of the expectation of $ESC(z^n)$, $E^*[ESC(Z^n)]$.

Theorem 4: Under Conditions 1 and 2, $E^*[ESC(Z^n)]$ satisfies

$$E^*[ESC(Z^n)] = E^*[L(Z : f_{\theta^{k*}})] + \frac{k}{2\lambda}\log\frac{n}{2\pi} - \frac{TrJ^*(\theta^{k*})\{I^*(\theta^{k*})\}^{-1}}{2\lambda} + \frac{1}{\lambda}\log\frac{\sqrt{\det I^*(\theta^{k*})}}{\pi(\theta^{k*})} + o(1). \quad (43)$$

(Proof) Applying the bounded convergence theorem to Eq. (40) and using Eq. (42), we have (43). □

4. Discussion

We have derived the asymptotic formula of ESC which holds almost surely. Although Yamanishi derived an asymptotic bound of ESC satisfied uniformly for all individual sequences, we have discussed almost sure convergence of ESC. Therefore, $J^*(\hat{\theta}^k)$ and $I^*(\hat{\theta}^k)$ which are defined using expectation by the true distribution appear in our asymptotic formula of ESC.

If the true distribution exists in the parametric model class $\mathcal{P}^k$, then $J^*(\theta^{k*}) = cI^*(\theta^{k*})$ is satisfied for some loss functions [8], where c is some constant. Then, we have $TrJ^*(\theta^{k*})\{I^*(\theta^{k*})\}^{-1} = c^k k$ and $E^*[ESC(Z^n)]$ satisfies the following equation:

$$E^*[ESC(Z^n)] = E^*[L(Z : f_{\theta^{k*}})] + \frac{k}{2\lambda}\log\frac{n}{2\pi e^C} + \frac{1}{\lambda}\log\frac{\sqrt{\det I^*(\theta^{k*})}}{\pi(\theta^{k*})} + o(1), \quad (44)$$

where $C = c^k$. This is a similar equation with the result of Clarke and Barron's asymptotics for the Bayes method. If we assume the logarithmic loss function and $\lambda = 1$, then $c = 1$. On this case, Eq. (44) is identical with the asymptotic formula derived by Clarke and Barron.

In [19], Yamanishi discussed the upper bound of ESC which holds uniformly for all data sequences. Assuming a source with the true distribution, it may be reasonable to apply the asymptotics for data sequences which is almost surely emitted from the source to evaluate the performancs of prediction. However, for the purpose of inductive learning, it is also desired to evaluate the performance of learning algorithm for each of all data sequences. It will be a future work to derive

†Moreover, the sequences of $h_\eta(\eta^k|z^n)$ converge to the normal distribution almost surely. That is, letting E be arbitrary rectangle on η^k-space, the following equation is satisfied.

$$\int_E \pi_\eta(\eta^k|z^n)d\eta^k \to \frac{\sqrt{\det I^*(\tilde{\theta}^k)}}{(2\pi)^{k/2}}\int_E \exp\left\{-\frac{1}{2}\|\eta^k\|^2_{I^*(\tilde{\theta}^k)}\right\}d\eta^k, \quad a.s. \quad (39)$$

the correct asymptotic formula within an error of $o(1)$ which holds uniformly for all individual sequences.

5. Conclusion

In this paper, we have analyzed the almost sure and mean convergence of ESC. Although Yamanishi concentrated mainly on the aggregating algorithm and the wost case performance and derived the upper bound of ESC satisfying uniformly for all data sequences and that of the asymptotic expectation of ESC, we have derived the asymptotic formulas of ESC in the meaning of almost sure and mean convergence. As a result, we have the similar type of asymptotics with Bayes method within an error of $o(1)$. It has been shown that Yamanishi's upper bound of asymptotic expectation of ESC is tight within an error of $o(1)$. The asymptotic normality is essential in this analysis.

References

[1] J.M. Bernardo and A.F.M. Smith, Bayesian Theory, John Wiley & Sons, 1994.
[2] B.S. Clarke and A.R. Barron, "Jeffreys' prior is asymptotically least favorable under entropy risk," J. Statistical Planning and Inference, vol.41, pp.37–60, 1994.
[3] C.-F. Chen, "On asymptotic normality of limiting density function with Bayesian implications," J.R. Statist. Soc. B, vol.47, no.3, pp.540–546, 1985.
[4] W. Feller, An Introduction to Probability and Its Applications, vol.1 and 2, John Wiley & Sons, New York, 1957, 1966.
[5] M. Gotoh, T. Matsushima, and S. Hirasawa, "A generalization of B.S. Clarke and A.R. Barron's asymptotics of Bayes codes for FSMX sources," IEICE Trans. Fundamentals, vol.E81-A, no.10, pp.2123–2132, 1998.
[6] T.S. Han and K. Kobayashi, Mathematics of Information and Coding, Iwanami Shoten, 1994.
[7] K. Ikeda, "Asymptotic analysis of incremental learning," IEICE Trans., vol.J80-D-II, no.7, pp.1913–1918, 1997.
[8] H. Linhart and W. Zucchini, Model Selection, John Wiley & Sons, New York, 1986.
[9] T. Matsushima, H. Inazumi, and S. Hirasawa, "A class of distortionless codes designed by Bayes decision theory," IEEE Trans. Inf. Theory, vol.37, no.5, pp.1288–1293, 1991.
[10] G. Qian, G. Gabor, and R.P. Gupta, "On stochastic complexity estimation: A decision-theoretic approach," IEEE Trans. Inf. Theory, vol.40, no.4, pp.1181–1191, 1994.
[11] J. Rissanen, "Modeling by shortest data description," Automatica, vol.46, pp.465–471, 1978.
[12] J. Rissanen, "Universal coding, information, prediction, and estimation," IEEE Trans. Inf. Theory, vol.IT-30, no.4, 1984.
[13] J. Rissanen, "Stochastic complexity," J.R. Statist., Soc. B, vol.49, no.3, pp.223–239, 1987.
[14] J. Rissanen, "Fisher information and stochastic complexity," IEEE Trans. Inf. Theory, vol.42, no.1, pp.40–47, 1996.
[15] C. Schwarz, "Estimating the dimension of a model," Ann. Statist., vol.6, pp.461–464, 1978.
[16] J. Takeuchi, "Characterization of the Bayes estimator and the MDL estimator for exponential families," IEEE Trans. Inf. Theory, vol.43, no.4, pp.1165–1174, 1997.
[17] C.S. Wallace and P.R. Freeman, "Estimation and inference by compact coding," J.R. Statist., Soc. B, vol.49, no.3, pp.240–265, 1987.
[18] K. Yamanishi, "A loss bound model for on-line stochastic prediction algorithms," Information and Computation, vol.119, pp.39–54, 1995.
[19] K. Yamanishi, "A decision-theoretic extension of stochastic complexity and its applications to learning," IEEE Trans. Inf. Theory, vol.44, no.4, pp.1424–1439, 1998.
[20] K. Yamanishi, "Extended stochastic complexity and learning theory," Proc. IBIS '98, pp.33–40, 1998.

Appendix A: Proof of Theorem 1

The similar discussion in order to show the asymptotic normality of the Bayes posterior density appeared in [1], pp.285–297, and [5]. We discuss the asymptotic normality around the minimum loss estimator although Bernard and Smith discussed that around the maximum posterior estimator. Since their conditions (c.1) $\sim$(c.3) in [1] don't directly lead Theorem 1, we should give the proof of Theorem 1.

At first, we show the first part of the theorem, Eq. (35). Define

$$K_n(\theta^k) = \log \pi(\theta^k|z^n), \tag{A·1}$$

$$K_n''(\theta^k) = \frac{\partial^2 K_n(\theta^k)}{\partial\theta^k(\partial\theta^k)^T}, \tag{A·2}$$

From the Taylor expansion with respect to θ^k, we have

$$\begin{aligned}\pi(\theta^k|z^n) &= \pi(\hat\theta^k|z^n)\exp\left\{(\theta^k-\hat\theta^k)^T K_n'(\hat\theta^k)\right\} \\ &\quad\cdot\exp\left\{\frac{1}{2}(\theta^k-\hat\theta^k)^T(I+R_n)K_n''(\hat\theta^k)(\theta^k-\hat\theta^k)\right\},\end{aligned} \tag{A·3}$$

where R_n is given by

$$R_n = K_n''(\theta^{k+})\{K_n''(\hat\theta^k)\}^{-1} - I, \tag{A·4}$$

for some θ^{k+} lying between θ^k and $\hat\theta^k$.

Since $\hat\theta^k$ minimizing $h(z^n|\theta^k)$ is almost surely unique in the interior of Θ^k when $n\to\infty$ from Condition 2, (2) and Lemma 3, $K_n'(\hat\theta^k)$ is given by

$$\begin{aligned}K_n'(\hat\theta^k) &= \left.\frac{\partial\log h(z^n|\theta^k)}{\partial\theta^k}\right|_{\theta^k=\hat\theta^k} + \left.\frac{\partial\log\pi(\theta^k)}{\partial\theta^k}\right|_{\theta^k=\hat\theta^k} \\ &= \left.\frac{\partial\log\pi(\theta^k)}{\partial\theta^k}\right|_{\theta^k=\hat\theta^k}, \quad a.s.\end{aligned} \tag{A·5}$$

when $n\to\infty$. On the other hand, from

$$K_n''(\theta^k) = \frac{\partial^2\log h(z^n|\theta^k)}{\partial\theta^k(\partial\theta^k)^T} + \frac{\partial^2\log\pi(\theta^k)}{\partial\theta^k(\partial\theta^k)^T}, \tag{A·6}$$

we have

$$-\frac{1}{n}K_n''(\theta^k) \to I^*(\theta^k),\ a.s. \tag{A·7}$$

uniformly for $\forall\theta^k\in\Theta^k$ because of Condition 2, (7).

Since

$$\hat{\theta}^k \to \theta^{k*}, \quad a.s. \tag{A·8}$$

is satisfied from Lemma 3, we have $B_{\delta'}(\hat{\theta}^k) \subset B_{\delta''}(\theta^{k*})$ $a.s.$ when $n \to \infty$ for $0 < \forall\delta' < \forall\delta''$. Since (A·8) is satisfied and $I^*(\theta^k)$ is continuously differentiable with respect to θ^k, we have

$$K_n''(\theta^{k+})\{K_n''(\hat{\theta}^k)\}^{-1} \to I^*(\theta^{k+})\{I^*(\hat{\theta}^k)\}^{-1}, \quad a.s. \tag{A·9}$$

uniformly for $\forall\theta^k \in B_\delta(\theta^{k*})$ for $\forall\delta > 0$. On the other hand, from the continuity of $I^*(\theta^k)$, there exists some positive number $\exists\delta_\epsilon > 0$ for $\forall\epsilon > 0$ such that

$$I - A(\epsilon) \le I^*(\theta^k)\{I^*(\theta^{k*})\}^{-1} \le I + A(\epsilon), \tag{A·10}$$

for $\forall\theta^k \in B_{\delta_\epsilon}(\theta^{k*})$. Here, I is the $k \times k$ identity matrix and $A(\epsilon)$ is a $k \times k$ symmetric positive-definite matrix whose largest eigenvalue tends to 0 as $\epsilon \to 0$. Here, δ_ϵ depends on ϵ. Since $I^*(\theta^k)$ is continuous, $\delta_\epsilon \to 0$ when $\epsilon \to 0$.

Therefore, since (A·9) and (A·10) are satisfied and $B_{\delta'}(\hat{\theta}^k) \subset B_{\delta''}(\theta^{k*})$ $a.s.$ when $n \to \infty$ for $0 < \forall\delta' < \forall\delta''$, there exists $\exists\delta_\epsilon > 0$ for $\forall\epsilon > 0$ such that

$$I - A(\epsilon) \le K_n''(\theta^k)\{K_n''(\hat{\theta}^k)\}^{-1} \le I + A(\epsilon), \quad a.s. \tag{A·11}$$

for $\forall\theta^k \in B_{\delta_\epsilon}(\theta^{k*})$ when $n \to \infty$[†]. Here, $\delta_\epsilon \to 0$ when $\epsilon \to 0$.

On the other hand, defining

$$\bar{c}_f = \max_{\theta^k \in \Theta^k, 1 \le i \le k} \left| \frac{\partial \log \pi(\theta^k)}{\partial \theta_i} \right|, \tag{A·12}$$

we have

$$\begin{aligned} \exp\{-\bar{c}_f\delta_\epsilon\} &\le \exp\left\{(\theta^k - \hat{\theta}^k)^T K_n'(\hat{\theta}^k)\right\} \\ &\le \exp\{\bar{c}_f\delta_\epsilon\}, \quad a.s. \end{aligned} \tag{A·13}$$

for $\forall\theta^k \in \overline{B_{\delta_\epsilon}}(\hat{\theta}^k)$ when $n \to \infty$ from Eq. (A·5).

Let $\overline{a(\epsilon)}$ and $\underline{a(\epsilon)}$ be the maximum and the minimum eigenvalues of $A(\epsilon)$ respectively. And we define $\overline{l_n}$ and $\underline{l_n}$ as follows:

$$\overline{l_n} = \sup_{\theta^k \in B_\delta(\theta^{k*})} \overline{\lambda_n}(\theta^k), \tag{A·14}$$

and

$$\underline{l_n} = \inf_{\theta^k \in B_\delta(\theta^{k*})} \underline{\lambda_n}(\theta^k), \tag{A·15}$$

for some $\delta > \delta_\epsilon$, where $\overline{\lambda_n}(\theta^k)$ and $\underline{\lambda_n}(\theta^k)$ are the maximum and the minimum eigenvalues of $K''(\theta^k)$ respectively. Defining

$$u^k = \left\{(I + R_n)K_n''(\hat{\theta}^k)\right\}^{1/2}(\theta^k - \hat{\theta}^k), \tag{A·16}$$

$$t_n = \delta_\epsilon(1 - \underline{a(\epsilon)})^{1/2}\underline{l_n}^{1/2}, \tag{A·17}$$

and

$$s_n = \delta_\epsilon(1 - \overline{a(\epsilon)})^{1/2}\overline{l_n}^{1/2}, \tag{A·18}$$

we have

$$\left\{\theta^k \middle| \|u^k\| < t_n\right\} \subset B_{\delta_\epsilon}(\hat{\theta}^k) \subset \left\{\theta^k \middle| \|u^k\| < s_n\right\}. \tag{A·19}$$

From (A·13)∼(A·19),

$$P_n(\delta_\epsilon) = \int_{B_{\delta_\epsilon}(\hat{\theta}^k)} \pi(\theta^k|z^n)d\theta^k, \tag{A·20}$$

is almost surely upper bounded by

$$\begin{aligned} &\pi(\hat{\theta}^k|z^n)\exp\{\bar{c}_f\delta_\epsilon\}\left\{\det\big(I - A(\epsilon)\big)\right\}^{-1/2} \\ &\quad \cdot \{\det K_n''(\hat{\theta}^k)\}^{-1/2} \\ &\quad \cdot \int_{\|u^k\| < s_n} \exp\left\{-\frac{1}{2}(u^k)^T u^k\right\} du^k, \end{aligned} \tag{A·21}$$

and lower bounded by

$$\begin{aligned} &\pi(\hat{\theta}^k|z^n)\exp\{-\bar{c}_f\delta_\epsilon\}\left\{\det\big(I + A(\epsilon)\big)\right\}^{-1/2} \\ &\quad \cdot \{\det K_n''(\hat{\theta}^k)\}^{-1/2} \\ &\quad \cdot \int_{\|u^k\| < t_n} \exp\left\{-\frac{1}{2}(u^k)^T u^k\right\} du^k, \end{aligned} \tag{A·22}$$

when $n \to \infty$.

From Eqs. (A·7) and (A·8), we have $\overline{l_n} \to \infty$, $a.s.$ and $\underline{l_n} \to \infty$, $a.s.$ when $n \to \infty$. Therefore, $s_n \to \infty$, $a.s.$ and $t_n \to \infty$, $a.s.$ are satisfied when $n \to \infty$. Then, when $n \to \infty$, we have

$$\begin{aligned} &\left\{\det\big(I - A(\epsilon)\big)\right\}^{1/2}\exp\{-\bar{c}_f\delta_\epsilon\}\lim_{n\to\infty} P_n(\delta_\epsilon) \\ &\le \lim_{n\to\infty} \frac{\pi_\xi(\hat{\xi}^k|z^n)(2\pi)^{k/2}}{\left\{\det I^*(\hat{\theta}^k)\right\}^{1/2}} \\ &\le \left\{\det\big(I + A(\epsilon)\big)\right\}^{1/2}\exp\{\bar{c}_f\delta_\epsilon\}\lim_{n\to\infty} P_n(\delta_\epsilon), \ a.s. \end{aligned} \tag{A·23}$$

from Eqs. (34) and (A·7).

Since $\delta_\epsilon \to 0$ when $\epsilon \to 0$, we can set that the positive constants $\epsilon > 0$ and $\delta_\epsilon > 0$ are abitrary small. Therefore, since $\lim_{n\to\infty} P_n(\delta_\epsilon) \le 1$ is satisfied for $\forall\delta_\epsilon > 0$, we have

$$\lim_{n\to\infty} \pi_\xi(\hat{\xi}^k|z^n) \le \frac{\{\det I^*(\hat{\theta}^k)\}^{1/2}}{(2\pi)^{k/2}}, \quad a.s. \tag{A·24}$$

If $\lim_{n\to\infty} P_n(\delta_\epsilon) \to 1$, $a.s.$ for $\forall\delta_\epsilon > 0$ is satisfied, then

$$\pi_\xi(\hat{\xi}^k|z^n) \to \frac{\{\det I^*(\hat{\theta}^k)\}^{1/2}}{(2\pi)^{k/2}}, \quad a.s. \tag{A·25}$$

[†]In this paper, for some events $\mathcal{A}_n$, $n = 1, 2, \cdots$, we denote "$P^*(\cup_{n=1}^\infty \cap_{k=n}^\infty \mathcal{A}_k) = 1$" as "$\mathcal{A}$, $a.s.$ when $n \to \infty$".

can be derived.

Then, we prove $\lim_{n\to\infty} P_n(\delta_\epsilon) \to 1$, *a.s.* for $\forall\delta_\epsilon > 0$ at last. We have

$$\begin{aligned}
&\frac{1}{n}\{K_n(\theta^k) - K_n(\theta^{k*})\} \\
&= \frac{1}{n}\log\frac{h(z^n|\theta^k)}{h(z^n|\theta^{k*})} + \frac{1}{n}\log\frac{\pi(\theta^k)}{\pi(\theta^{k*})} \\
&= -\frac{\lambda}{n}\sum_{t=1}^{n} L(z_t : f_{\theta^k}) + \frac{\lambda}{n}\sum_{t=1}^{n} L(z_t : f_{\theta^{k*}}) \\
&\quad + \frac{1}{n}\log\frac{\pi(\theta^k)}{\pi(\theta^{k*})} \\
&\to -\lambda E^*\left[L(Z : f_{\theta^k})\right] + \lambda E^*\left[L(Z : f_{\theta^{k*}})\right], \quad a.s.
\end{aligned} \tag{A·26}$$

uniformly for $\forall\theta^k \notin B_{\delta_\epsilon}(\theta^{k*})$ for $\forall\delta_\epsilon > 0$ from Condition 1 and Condition 2, (5). Because $E^*\left[L(Z : f_{\theta^{k*}})\right]$ is an unique minimum of $E^*\left[L(Z : f_{\theta^k})\right]$ from Condition 2, (2), there exists some positive constant $\exists C_{\delta_\epsilon} > 0$ for $\forall\delta_\epsilon > 0$ such that

$$\frac{1}{n}\{K_n(\theta^k) - K_n(\theta^{k*})\} < -C_{\delta_\epsilon},\ a.s. \tag{A·27}$$

uniformly for $\forall\theta^k \notin B_{\delta_\epsilon}(\theta^{k*})$ when $n \to \infty$. Therefore,

$$\frac{\pi(\theta^k|z^n)}{\pi(\theta^{k*}|z^n)} < \exp\left\{-nC_{\delta_\epsilon}\right\},\ a.s. \tag{A·28}$$

is satisfied uniformly for $\forall\theta^k \notin B_{\delta_\epsilon}(\theta^{k*})$ when $n \to \infty$ for $\forall\delta_\epsilon > 0$. On the other hand, we have

$$\pi_\xi(\xi^{k*}|z^n) < \frac{\{\det I^*(\hat\theta^k)\}^{1/2}}{(2\pi)^{k/2}}, \quad a.s. \tag{A·29}$$

for $\forall\theta^k \in B_{\delta_\epsilon}(\theta^{k*})$ when $n \to \infty$ from (A·8) and (A·24), where $\xi^{k*} = \sqrt{n}(\theta^{k*} - \hat\theta^k)$. Here, $\det I^*(\hat\theta^k) \to \det I^*(\theta^{k*})$, *a.s.* is satisfied because of Condition 2, (3) and Lemma 3. Therefore, there exists some positive constant $C^* > 0$ such that

$$\pi(\theta^{k*}|z^n) < C^*\sqrt{n}^k, \quad a.s. \tag{A·30}$$

when $n \to \infty$ because of Eq. (34). Therefore, for $\forall\delta_\epsilon > 0$,

$$\begin{aligned}
\pi(\theta^k|z^n) &< \pi(\theta^{k*}|z^n)\exp\{-nC_{\delta_\epsilon}\}, \quad a.s. \\
&< C^*\sqrt{n}^k \exp\{-nC_{\delta_\epsilon}\} \to 0, \quad a.s.
\end{aligned} \tag{A·31}$$

is satisfied uniformly for $\theta^k \notin B_{\delta_\epsilon}(\theta^{k*})$ from inequations (A·28) and (A·30). Since Θ^k is compact from Condition 2, (1), we have

$$\int_{\theta^k \notin B_{\delta_\epsilon}(\theta^{k*})} \pi(\theta^k|z^n)d\theta^k \to 0, \quad a.s. \tag{A·32}$$

for $\forall\delta_\epsilon > 0$. This means

$$\int_{\theta^k \in B_{\delta_\epsilon}(\theta^{k*})} \pi(\theta^k|z^n)d\theta^k \to 1, \quad a.s. \tag{A·33}$$

for $\forall\delta_\epsilon > 0$.

From (A·8), we have $B_{\delta'}(\theta^{k*}) \subset B_{\delta''}(\hat\theta^k)$, *a.s.* when $n \to \infty$ for $0 < \forall\delta' < \forall\delta''$. Therefore we have $\lim_{n\to\infty} P_n(\delta_\epsilon) = 1$, *a.s.*, then we have (35)†. □

Appendix B: Proof of Theorem 2

From the Taylor expansion with respect to θ^k around $\tilde\theta^k$, we have

$$\begin{aligned}
\pi(\theta^k|z^n) &= \pi(\tilde\theta^k|z^n) \\
&\quad \cdot \exp\left\{-\frac{1}{2}(\theta^k - \tilde\theta^k)^T(I + \tilde R_n)K''(\tilde\theta^k)(\theta^k - \tilde\theta^k)\right\},
\end{aligned} \tag{A·35}$$

where $\tilde R_n$ is given by

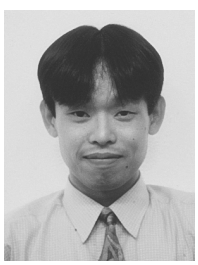
$$\tilde R_n = K_n''(\theta^{k+})\{K_n''(\tilde\theta^k)\}^{-1} - I, \tag{A·36}$$

for some θ^{k+} lying between θ^k and $\tilde\theta^k$. Therefore, we can see that the identical discussion with Appendix A leads Theorem 2. □

†It is obvious that (36) can be similarly proved by bounding

$$\int_E \pi_\xi(\xi^k|x^n)d\xi^k, \tag{A·34}$$

for arbitrary rectangle $\forall E$.



Masayuki Gotoh was born in Tokyo, Japan, on Jan. 1, 1969. He received his B.E. and M.E. degrees from Musashi Institute of Technology, Tokyo, Japan, in 1992 and 1994, respectively. He is now a student of doctoral program and a research associate in Industrial and Management Systems Engineering at Waseda University, Tokyo, Japan. His research interests include intelligent control, machine learning theory, model selection, and Bayesian statistics. He is a member of the Society of Information Theory and Its Applications, the Japan Industrial Management Association, and the Japan Society for Artificial Intelligence.

Toshiyasu Matsushima was born in Tokyo, Japan, on Nov. 26, 1955. He received the B.E. degree, M.E. degree and Dr.E. degree in Industrial Engineering and Management from Waseda University, Tokyo, Japan, in 1978, 1980 and 1991, respectively. From 1980 to 1986, he was with the Nippon Electric Corporation, Kanagawa, Japan. From 1986 to 1992, he was a lecture to the Department of Management Information, Yokohama College of Commerce. From 1993, he was a associate professor and since 1996 has been a professor of School of Science and Engineering, Waseda University, Tokyo, Japan. His research interests are information theory and its application, statistics and artificial intelligence. He is a member of the Society of Information Theory and Its Applications, the Japan Society for Quality Control, the Japan Industrial Management Association, the Japan Society for Artificial Intelligence, and IEEE.

Shigeichi Hirasawa was born in Kobe, Japan, on Oct. 2, 1938. He received the B.S. degree in mathematics and the B.E. degree in electrical communication engineering from Waseda University, Tokyo, Japan, 1961 and 1963, respectively, and the Dr.E. degree in electrical communication engineering from Osaka University, Osaka, Japan, in 1975. From 1963 to 1981, he was with the Mitsubishi Electric corporation, Hyogo, Japan. Since 1981, he has been a professor of School of Science and Engineering, Waseda University, Tokyo, Japan. In 1979, he was a Visiting Researcher in the Computer Science Department at the University of California, Los Angels, CA. He was a Visiting Researcher at the Hungarian Academy of Science, Hungary, in 1985, and at the University of Trieste, Italy, in 1986. From 1987 to 1989, he was the Chairman of Technical Group on Information Theory of IEICE. He received the 1993 Achievement Award, and the 1993 Kobayashi-Memorial Achievement Award from IEICE. In 1996, he was the President of the Society of Information Theory and Its Applications (Soc. of ITA). His research interests are information theory and its applications, and information processing systems. He is a member of Soc. of ITA, the Operations Research Society of Japan, the Information Processing Society of Japan, the Japan Industrial Management Association, IEEE, and Informs.