1998 年情報論的学習理論ワークショップ
1998 Workshop on Information-Based Induction Sciences (IBIS'98)
Hakone, Japan, July 11-12, 1998.

確率的コンプレキシティと Jeffreys 混合予測戦略
Stochastic Complexity and Jeffreys Mixture Prediction Strategies

竹内 純一*
Jun-ichi Takeuchi

Abstract: We review the notion of Rissanen's stochastic complexity and the method of modified Jeffreys mixtures for achieving the stochastic complexity. The stochastic complexity is defined as the code length of the maximum likelihood code and plays an important role in statistical inference and data compression. Appropriate variants of Jeffreys mixtures asymptotically achieve the stochastic complexity and give asymptotically minimax prediction strategies with respect to logarithmic regret.

1 まえがき

本稿では, 統計的推論やデータ圧縮の問題において大きな役割を果たす Rissanen の確率的コンプレキシティの概念 [12] を紹介し, それを漸近的に達成する方法として Jeffreys 混合分布について論じる.

モデル (d 次元の経数が付いた確率過程の族を指すものとする) が一つ与えられている時, そのモデルに関する確率的コンプレキシティはモデル上の最尤符号 [13] の符号長として定義される [12]. その値はモデル選択の問題における情報量規準として利用される. また, 確率的コンプレキシティを達成する符号や予測戦略は, それ自体ある最適性を持つ.

Rissanen はモデルのパラメータを離散化して最尤符号を近似することで, モデルに関する確率的コンプレキシティの漸近的評価式を与えた [12]. これは定数オーダまで同定する精密なものだが, 適用出来るモデルの範囲があまり広くない. 一方, Barron, Xie, Takeuchi らは, より緩い条件のもとで, 修正した Jeffreys 事前分布による Bayes 混合に基づく符号 (以下, 修正 Jeffreys 符号と呼ぶ) が, 最尤符号の符号長を漸近的に達成することを示し, 確率的コンプレキシティの漸近的評価を与えた [20, 21, 15, 6, 16]. そもそも最尤符号は, 対数リグレット (各点冗長度, [13, 19] 等で用いられている) に関する minimax 符号として導入されたもので, 上記の成果は修正 Jeffreys 符号がリグレットについて minimax 性を持つことを主張する. また, 修正 Jeffreys 符号が, 逐次型予測問題に用いた時も良い性質を持つことを意味する.

以下, 次のような構成で上記の事情を解説する. 2 節で確率的コンプレキシティの定義の経緯を簡単に振り返り, 3 節で逐次型学習問題との関連を論じる. 4 節で基本的定義や仮定の説明をした後, 5 節で Rissanen による確率的コンプレキシティの評価式についてやや詳しく述べ, その問題点を指摘する. 6 節では, 修正 Jeffreys 符号を用いて上記の問題点を解決する方法について解説する.

2 確率的コンプレキシティ

統計的推論において, データをなるべく短い符号長で記述しようという思想— 本稿ではこれを MDL 原理 (Minimum Description Length principle) と呼ぶ— は推論方法を設計するときに指導的な役割を果たす. それが最も明確な形で表れたのは, モデル選択に用いられる MDL 基準の提唱 [10] においてであろう. これは, データと, そのデータを支配する法則の候補となるモデルが複数与えられている時, '全記述長'= 'モデル記述長'+'そのモデルによるデータ記述長' が最小になるモデルを選択せよという基準である. モデル $S = \{p(\cdot|u) : u \in U \subseteq \Re^d\}$[1] について, $L(S)$ で S を指定するための記述長を表そう. 長さ n のデータ列 $x^n = x_1x_2...x_n$ の, モデル S に関する全記述長は

$$-\log p(x^n|\hat{u}(x^n)) + L(S) + \frac{d}{2}\log n + O(1) \qquad (1)$$

で与えられる [10, 3, 18, 14]. ただし $\hat{u}(x^n)$ は u の最尤推定量, log は自然対数を表す. ここで第一項はデータ記述長である. これは所謂情報量規準の一つであり, その有効性は, 様々な文献で示されている [10, 3, 4, 22]. こ

*NEC C&C メディア研究所, 〒 216-8555 神奈川県川崎市宮前区宮崎 4-1-1 tel. 044-856-2143
C&C Media Research Laboratories, NEC Corporation, 4-1-1 Miyazaki, Miyamae, Kawasaki, Kanagawa 216-8555, Japan

[1] アルファベットを $\mathcal{X} \subseteq \Re^s$ とし, ν を $\Re^s$ 上の測度, $p(x^n|u)$ を測度 $\nu(dx^n) \stackrel{\text{def}}{=} \prod_{t=1}^n \nu(dx_t)$ に関して定義された確率密度とする.

のような段階的なデータの記述は二段階符号化と呼ばれる．(1) はモデルを一つ固定したもとで全記述長を最小化することで得られる．すなわち, モデルの選択基準が「出来るだけ短い記述長」というだけでなく, その選択に用いる記述長そのものが, MDL 原理により導かれているのである．ここで, 我々が追い求めている出来るだけ短い符号長の限界こそが, Rissanen がモデルに関する確率的コンプレキシティと呼んだものである．

二段階符号化を用いた MDL 基準の有効性は保証されたが, その後 Bayes 符号が二段階符号化よりも短い符号長を達成することが指摘された．これにより, MDL 原理に従う立場に立つとき, 二段階符号化による MDL 基準は最善の方法ではないと認識されるようになった．モデル S に関する Bayes 符号とは, Bayes 混合

$$m(x^n) = \int p(x^n|u)w(du)$$

に基づく符号である．ただし, $w(du)$ は U の上の確率測度であり, 事前分布と呼ばれる．Rissanen 自身, Bayes 符号の符号長 $-\log m(x^n)$ をモデル S に関する確率的コンプレキシティと名付けたことがある [11]．この定義では, 確率的コンプレキシティは事前分布 $w(du)$ に依存することに注意されたい．

これに対し Clarke と Barron は, 上記の任意性をある意味で取り除いた [9]．すなわち, S が i.i.d. 過程のモデルである場合, Jeffreys 事前分布 $w_J(u)du$ を用いた Bayes 混合 m_J の, $p(\cdot|u)$ に関する平均符号長が

$$E_u[\log \frac{1}{p(x^n|u)}] + \frac{d}{2}\log\frac{n}{2\pi e} + \log C_J(U) + o(1) \quad (2)$$

となることを示した．ただし, $I(u)$ は u の Fisher 情報行列, $C_J(K) \stackrel{\rm def}{=} \int_K \sqrt{\det(I(u))}du$ $(K \subseteq U)$, $w_J(u) \stackrel{\rm def}{=} \sqrt{\det(I(u))}/C_J(U)$ である．また, $o(1)$ は $n \to \infty$ のとき 0 に収束する量を表す．

ここで, U の内部に包含されるコンパクト集合 K を一つ固定すると, (2) はすべての $u \in K$ について一様に成り立つ．次式で定義される $R_n(q,u)$ は冗長度と呼ばれるが, (2) は冗長度 $R_n(m_J,u)$ が漸近的には u に依存しないことを示す．

$$R_n(q,u) \stackrel{\rm def}{=} E_u[\log\frac{1}{q(x^n)} - \log\frac{1}{p(x^n|u)}]. \quad (3)$$

これから, Jeffreys 事前分布に弱収束するようなある事前分布の列 $\{w_n\}$ を用いて定義される Bayes 符号 m_n(修正 Jeffreys 符号) をデータ x^n の符号化に用いると, その冗長度の ($u \in K$ についての) 最悪値は, $n \to \infty$ のとき漸近的に最小になることが示せる．これは冗長度に関する漸近的 minimax 符号と呼ばれ, 平均符号長の基準を $E_u[-\log p(x^n|u)]$ においたとき, 最小の平均符号長を漸近的に達成するということが出来る．

Rissanen はこの結果に刺激され, [12] において確率的コンプレキシティの「最終的な」定義を得た．ここで, 冗長度に関する minimax 符号の符号長はそのままでは不十分である．それは平均の意味での最小符号長であるが, 例えばモデル選択といった統計的推論に用いるには, 個々のデータ列について最小の符号長が望ましいのである．そこで Rissanen が着目したのは最尤符号 [13] である．長さ n のデータに関する最尤符号を $\hat{m}_n(x^n)$ と書くと, それは, 最大尤度 $p(x^n|\hat{u}(x^n))$ を規格化して得られる確率密度として定義される．ここで, $\hat{u}(x^n)$ は x^n が与えられたもとでの u の最尤推定値である．最尤符号は Shtarkov によって,

$$r(q,x^n) \stackrel{\rm def}{=} \log\frac{1}{q(x^n)} - \log\frac{1}{p(x^n|\hat{u}(x^n))} \quad (4)$$

で定義されるリグレット r について minimax を達成する符号として導入された．先の定義からすぐ分かるように, $\hat{m}_n$ については, どんな x^n についてもリグレットの値は厳密に同じになる．これは, リグレットの最悪値が最小になるような符号であることを意味する [13, 21]．すなわち最尤符号とは, データ x^n に対して, S に属する符号で達成出来る最小の符号長 $-\log p(x^n|\hat{u}(x^n))$ を基準としたとき, x^n の値に関わらず, なるべく基準に近い符号長を実現する符号である．

結局 Rissanen は $\hat{m}_n$ による x^n の記述長 $-\log\hat{m}_n(x^n)$ を x^n の S に関する確率的コンプレキシティと定義し, ある条件のもとで以下の値になることを示した [12]．

$$\log\frac{1}{p(x^n|\hat{u}(x^n))} + \frac{d}{2}\log\frac{n}{2\pi} + \log C_J(U) + o(1). \quad (5)$$

これは, $\hat{u}(x^n) \in K$ を満たす x^n について一様に成り立つ．(5) は (2) と密接な関係がある．それは, ある条件のもとで

$$E_u[\log\frac{p(x^n|\hat{u}(x^n))}{p(x^n|u)}] = \frac{d}{2} + o(1) \quad (6)$$

が $u \in K$ に対して一様に成り立つことが証明出来る [9] が, これを用いると, (5) から (2) が導出されるからである．すなわち, 最尤符号については (2) も成り立つ．ただし, (2) は必ずしも (5) を意味しないので, Jeffreys 符号について (5) が成り立つ保証は無い．実際, 後で示すように, モデルが指数型分布族で無い場合は成り立たない．

評価式 (5) は [12] の主要結果であり, Rissanen はこれをもって,「20 年にわたる模索に終止符をうつ」と宣言したのである．その後, 修正 Jeffreys 符号の方法を工夫することで, より広い範囲のモデルに対して漸近的に

mimimax リグレットが達成されること, そしてその値はやはり (5) で与えられることが示された.

3 逐次型予測問題と確率的コンプレキシティ

データ列を先頭から順に読み込みながら, 次に現れるデータを逐次的に予測していく問題を逐次型予測問題と呼ぶ. 確率過程 q について, x^t が与えられたもとでの x_{t+1} の条件付き確率密度 $q(x_{t+1}|x^t)$ は, x^t を知った後で次のデータ x_{t+1} の発生を予測する確率分布とみなすことが出来る. この認識のもとで q を予測戦略と呼ぶ.

予測戦略 q の x^n に関するリグレットは

$$r(q, x^n) = \sum_{t=1}^{n-1} \log \frac{1}{q(x_{t+1}|x^t)} - \sum_{t=1}^{n-1} \log \frac{1}{p(x_{t+1}|x^t, \hat{\theta}(x^n))}$$

と書き直せる. この式における $\log(1/q(x_{t+1}|x^t))$ は, 予測分布が $q(\cdot|x^t)$ であるのに対し, 実際は x_{t+1} が現れた場合の損失を表現している (x_{t+1} に付与される確率が大きいほど小さくなる). これを対数損失と呼ぶ. $\sum_{t=1}^{n-1} \log(1/q(x_{t+1}|x^t))$ は, それを逐次的に積み上げていった量で, 累積対数損失と呼ばれる. すると, $p(\cdot|\hat{u}(x^n))$ は「モデル S に属する予測戦略の中で, x^n に関する累積対数損失を最も小さくするもの」であることが分かる. もちろん, x^n はあらかじめ分かっている訳ではないので, 実際にそれを使うことは出来ない. すなわち $r(q, x^n)$ は, そうしたある意味理想的な競争相手と比較したときの q の評価尺度となっている.

この評価尺度を採用するとき, もし予測問題の長さが有限の n で予め与えられているならば, リグレットに関する minimax 符号である最尤符号 $\hat{m}_n$ は, 確率過程ではないが, 予測戦略として使うことが出来る. しかも長さ n の予測問題に関しては最適な予測戦略となる. 一方, 修正 Jeffreys 符号 m_n は漸近的に minimax となる確率過程なので, 漸近的に最適な予測戦略を構成する.

以上のことは, 確率的コンプレキシティを追求することと最適な予測戦略の設計は表裏一体となっていることを示している[2].

ここで, 修正 Jeffreys 予測戦略などは必要ない, 最尤符号を使えばよいと思われるかもしれない. しかし, 最尤符号は n が分かっていないときには使えない. また, n が分かっていても, 周辺密度 $\hat{m}_n(x^t)$ を求めるのが難しいため, 条件付き密度 $\hat{m}_n(x_{t+1}|x^t)$ の計算が難しい. 一方, 修正 Jeffreys 予測戦略も n が分かっていないと決められないのは同じだが, 周辺密度は $m_n(x^t) = \int p(x^t|u)w_n(du)$ から計算出来るので, 条件付き確率密度は $m_n(x_{t+1}|x^t) = m_n(x^{t+1})/m_n(x^t)$ で計算出来る. また, $t > n$ に対しても $m_n(x^t)$ が定義されているので, 最適性こそ保証されないが, 予測戦略 m_n を, n より長いデータ列の予測問題に使うことも出来るのである.

[2] 詳しくは [25, 5] 等を参照されたい. また, 対数損失以外の損失関数に対応した確率的コンプレキシティの一般化と, その逐次型学習への適用が [26] で論じられている.

4 定義と仮定

モデルは一般に $S = \{p(\cdot|u) : u \in U\}$ で表す. 先に述べたように $p(\cdot|u)$ はアルファベットを $\mathcal{X} \subseteq \Re^s$ とする確率過程, すなわち無限列 $x_1x_2...$ の確率測度の, 測度 $\nu(dx_1dx_2...) \stackrel{\text{def}}{=} \prod_{t=1}^{n} \nu(dx_t)$ に関する密度関数であるとする. また, 単に確率過程 q といったときも, q はそのような確率密度関数であるとする. $\hat{u}(x^n)$ で, x^n が与えられた時の u の最尤推定値を表すが, しばしば x^n を省略して $\hat{u}$ と書く.

パラメータの範囲 U の条件として特に, $U \subseteq \Re^d$ とし, $\bar{U^\circ} = \bar{U}$ であるとする ($A \subseteq \Re^d$ について, A° で A の内部を, $\bar{A}$ で A の閉包を表す). これは U の本質的な次元が d であることを要求する.

$\mathcal{G}$ で, U の部分集合で $\bar{\mathcal{G}^\circ} = \bar{\mathcal{G}}$ を満たすものを表す. $S(\mathcal{G}) \stackrel{\text{def}}{=} \{p(\cdot|u) : u \in \mathcal{G}\}$ という記法を用いる. また, $\mathcal{X}^n(\mathcal{G}) \stackrel{\text{def}}{=} \{x^n : \hat{u}(x^n) \in \mathcal{G}\}$ と定義する.

q を確率過程とする. このとき q に基づいて, x^n の符号長が (ほぼ)$-\log q(x^n)$ である符号が構成出来るので, q のことを符号とも呼ぶ. q の $p(\cdot|u)$ に関する冗長度は (3) で定義される. $S(\mathcal{G})$ に関する minimax 冗長度は

$$\bar{R}_n(S(\mathcal{G})) \stackrel{\text{def}}{=} \inf_q \sup_{u \in \mathcal{G}} R_n(q, u)$$

で定義される. ただし, inf は $\mathcal{X}^n$ 上のすべての確率密度関数 q についてとる.

符号 q の, データ列 x^n とモデル S に関するリグレットは (4) で定義される. $W_n \subseteq \mathcal{X}^n$ とモデル S に関する minimax リグレットは

$$\bar{r}(W_n) \stackrel{\text{def}}{=} \inf_q \sup_{x^n \in W_n} r(q, x^n)$$

で定義される. 以下, $W_n = \mathcal{X}^n(\mathcal{G})$ の場合だけを考える.

$\mathcal{X}^n(\mathcal{G})$ と S に関する最尤符号とは

$$\hat{m}_{n,\mathcal{G}}(x^n) \stackrel{\text{def}}{=} p(x^n|\hat{u}(x^n))/C_n(\mathcal{G})$$

で定義される $\hat{m}_{n,\mathcal{G}}$ である. ただし以下のようにおいた.

$$C_n(\mathcal{G}) \stackrel{\text{def}}{=} \int_{\mathcal{X}^n(\mathcal{G})} p(x^n|\hat{u}(x^n))\nu(dx^n).$$

$\hat{m}_{n,\mathcal{G}}$ は $\mathcal{X}^n(\mathcal{G})$ と S に関して minimax である [13, 21].

U において, 以下に記す正則条件を仮定する. $\hat{I}(x^n,u)$ で u の経験的 Fisher 情報行列を表す. すなわち,

$$\hat{I}_{ij}(x^n,u) \stackrel{\rm def}{=} -\frac{1}{n}\frac{\partial^2 \log p(x^n|u)}{\partial u^i \partial u^j}$$

である. これが U° において連続であるとする. Fisher 情報行列を $I(u) \stackrel{\rm def}{=} \lim_{n\to\infty} E_u[\hat{I}(x^n,u)]$ で定める. ただし, S が i.i.d. 過程のクラスである場合は極限操作は必要ない. これが U° において存在し, 連続でかつ正定値であると仮定する. i.i.d. の場合は次の仮定もおく.

$$I_{ij}(u) = E_u\big[\frac{\partial \log p(x|u)}{\partial u^i}\frac{\partial \log p(x|u)}{\partial u^j}\big]. \qquad (7)$$

次に

$$C_J(\mathcal{G}) \stackrel{\rm def}{=} \int_{\mathcal{G}} \sqrt{\det(I(u))}du$$

とし, $\mathcal{G}$ の上の Jeffreys 事前分布を次式で定義する.

$$w_{\mathcal{G}}(u)du \stackrel{\rm def}{=} \sqrt{\det(I(u))}du/C_J(\mathcal{G}).$$

6 節で用いる指数型分布族と曲指数型分布族を定義しておく (詳しくは [8, 1, 23] 参照). 前者はとても良い性質をもったモデルであり, 例えば正規分布の作るモデル, Bernoulli モデル, Poisson 分布の作るモデルは指数型分布族となる. 後者は, 指数型分布族に埋め込まれた超曲面に相当するモデルである. 今, $x \in \Re^d$ とし, ν の台が $\mathcal{X}$ であり, $\mathcal{X}$ の凸包は $\Re^d$ で体積をもっていると仮定する. $\theta \in \Re^d$ に対して $\psi(\theta) \stackrel{\rm def}{=} \log \int \exp(\theta\cdot x)\nu(dx)$ とおく. また $\Theta \stackrel{\rm def}{=} \{\theta : \psi(\theta) < \infty\}$ と定める.

$$p(x|\theta) \stackrel{\rm def}{=} \exp(\theta\cdot x - \psi(\theta))$$

とおくと, $p(x|\theta)$ $(\theta \in \Theta)$ は ν に関する確率密度となる. $S = \{p(\cdot|\theta) : \theta \in \Theta\}$ を指数型分布族と呼ぶ. θ は自然パラメータと呼ばれ, 特別な意味を持っている. Θ は凸集合であり, ψ は Θ° で解析的かつ凸関数となることが知られている. ここでは Θ の次元が d であると仮定する. Θ が開集合であるとき S は regular といわれる. また, $\forall\theta \in \Theta - \Theta^\circ$, $\psi(\theta) = \infty$ を仮定する. これを満たす指数型分布族は steep であるといわれる. 定義から regular ならば steep であることが分かる. S が steep ならば, $\theta \mapsto \eta(\theta) \stackrel{\rm def}{=} E_\theta[x]$ なる関数は Θ° で単射となる. (指数型分布族について) 経験的 Fisher 情報行列を $\hat{J}(x^n,\theta)$, θ の Fisher 情報行列を $J(\theta)$ と書く. すると $\eta_i = \partial\psi(\theta)/\partial\theta^i$, $J_{ij}(\theta) = \partial^2\psi(\theta)/\partial\theta^i\partial\theta^j$ が成り立つ (二番目の式は J の定義により簡単に確かめられる). $\bar{x} = \sum_{t=1}^n x_t/n$ とおくと, $p(x^n|\theta) = \exp(n(\bar{x}\cdot\theta - \psi(\theta)))$ と書ける. これらから $\hat{J}(x^n,\theta) = J(\theta)$ が得られる.

曲指数型分布族を定義する. 今, $\bar{S}$ を $\bar{d}$ 次元 $(\bar{d} > d)$ の指数型分布族とする. U を $\Re^d$ の開集合とし ϕ を $U \to \Theta$ なる関数とする. ただし ϕ は U で単射で適当な回数微分可能であり, ϕ のヤコビアンのランクはいたるところ d であると仮定する. このとき, $S = \{p_c(\cdot|u) \stackrel{\rm def}{=} p(\cdot|\phi(u)) : u \in U\}$ を $\bar{S}$ に埋め込まれた曲指数型分布族と呼ぶ.

5 最尤符号の符号長評価

ここでは評価式 (5) を詳細に検討する. これはいくつかの条件のもとに証明されている. まず, 前節で述べた条件の他に, $\mathcal{G}$ は有界な開集合とされる. この他

i) $\forall u \in \mathcal{G}$, $0 < c_1 \le \det(I(u)) \le c_2 < \infty$.

ii) 最尤推定値 $\hat{u}(x^n)$ に関する中心極限定理が, すべての $u \in \mathcal{G}$ について一様に成り立つ. すなわち, $\xi = \sqrt{n}(\hat{u}-u)$ とおき, R_r で, 各辺の長さが $2r$ で, 中心が原点にある立方体を表すとき, 次式が一様に成り立つ.

$$\mathrm{P}_u(\xi \in R_r) = \frac{(\det(I(u)))^{1/2}}{(2\pi)^{d/2}}\int_{R_r} e^{-\xi\cdot I(u)\xi/2}d\xi + o(1).$$

iii) $\hat{I}(x^n,\hat{u}) < M < \infty$ が全ての n, 全ての $x^n \in \mathcal{X}(\mathcal{G})$ について成り立つ. ただし M はある正定値の定数行列とする. また, $u(\xi) = \hat{u} + \xi/\sqrt{n}$ とおくとき, 全ての $n \ge 1$, 全ての $x^n \in \mathcal{X}^n(\mathcal{G})$, 全ての i,j についての ξ の関数の族 $I_{ij}(x^n,u(\xi))$ は, $\xi = 0$ において同程度連続である.

を用いる. これらの条件のもとで

$$\log\frac{p(x^n|\hat{u})}{\hat{m}_{n,\mathcal{G}}(x^n)} = \frac{d}{2}\log\frac{n}{2\pi} + \log C_J(\mathcal{G}) + o(1) \qquad (8)$$

が, すべての $x^n \in \mathcal{X}^n(\mathcal{G})$ について一様に成り立つ.

以下に, [12] による証明のアイデアを述べる. $C_n(\mathcal{G}) \sim C_J(\mathcal{G})n^{d/2}/(2\pi)^{d/2}$ を示せばよいことに注意しよう. まず, パラメータ集合 $\mathcal{G}$ を, 各辺の長さが $2r/\sqrt{n}$ の立方体で離散化する. そのときの一つのセルを b_i と書くと

$$C_n(\mathcal{G}) = \sum_i \int_{x^n:\hat{u}\in b_i} p(x^n|\hat{u})\nu(dx^n)$$

が成り立つ. 今, i を固定し b_i の重心を $\bar{u}$ で表す. すると, 条件 iii) 等により, $\hat{u} \in b_i$ のとき $p(x^n|\hat{u}) \sim p(x^n|\bar{u})$ が成り立つ. よって,

$$\begin{aligned}\int_{x^n:\hat{u}\in b_i} p(x^n|\hat{u})\nu(dx^n) &\sim \int_{x^n:\hat{u}\in b_i} p(x^n|\bar{u})\nu(dx^n)\\ &= \mathrm{P}_{\bar{u}}(\hat{u}\in b_i)\\ &= \mathrm{P}_{\bar{u}}(\xi \in R_r)\end{aligned}$$

となる. ここで条件 ii) を使うと次式が得られる.

$$C_n(\mathcal{G}) \sim n^{d/2}\int_{\mathcal{G}} \frac{(\det(I(u)))^{1/2}}{(2\pi)^{d/2}}du = \frac{C_J(\mathcal{G})n^{d/2}}{(2\pi)^{d/2}}.$$

次に前提条件 i)-iii) について検討する. 条件 i) は, 例えば $\bar{\mathcal{G}} \subseteq U^\circ$ という仮定をおけば成り立つ. 従ってこれはある程度妥当な仮定である. 実際, [9] では同様の仮定をおいている. もちろん, これを取り除くことは重要な課題である.

条件 ii) と iii) は簡単には保証出来ない. ii) について言うと, u に関する一様性が問題となる. というのは, 通常, 中心極限定理は, 各点での収束の形で示されているからである. すなわち, 各 u においての収束は保証されるが, その収束の速さが u に対して一様である保証が無いのである. けれども, S が i.i.d. のモデルあり, しかも曲指数型分布族である場合, $\bar{\mathcal{G}} \subseteq U^\circ$ ならば ii) を証明することが出来る. 詳しくは述べないが, 標本平均についての中心極限定理の収束の速さを特徴付ける不等式 [7] を用いればよい.

iii) については, $x^n \in \mathcal{X}^n(\mathcal{G})$ に関する一様性が厄介である. 例えば指数型分布族の場合は経験的 Fisher 情報行列は x^n に依存しないので問題ない. しかし, S が i.i.d. の曲指数型分布族であっても $\mathcal{X}$ が有界で無ければ保証出来ない. 例えば正規分布 (二次元) に埋め込まれた (曲がった) 一次元のモデルでは成り立たない. 証明を見直して条件を緩和することが望まれる.

6 Jeffreys 符号の符号長の評価

Xie と Barron は多項 Bernoulli モデルについて, 修正 Jeffreys 符号で $\mathcal{X}^n$ に関する minimax リグレットを漸近的に達成できることを示した [21]. その後, Takeuchi と Barron は, やはり修正 Jeffreys 符号によって, $\bar{\mathcal{G}} \subseteq U^\circ$ のときの一般の滑らかな i.i.d. のモデル, および $\mathcal{G}$ に関する条件が無いときの一次元の指数型分布族について同様の命題を示した [15, 6, 16]. さらに, ある場合には i.i.d. という条件も取り除いている. この節では, それらの解説を行う.

6.1 下界について

$C_J(\mathcal{G}) < \infty$ を仮定する. ある符号 q_s について

$$\liminf_{n\to\infty} \sup_{x^n \in \mathcal{X}^n(\mathcal{G})} \Big(r(q_s, x^n) - \frac{d}{2} \log \frac{n}{2\pi} \Big) \le \log C_J(\mathcal{G}) - \gamma$$

が成り立つと仮定する ($\gamma > 0$). ここで $\mathcal{G}' \subset \mathcal{G}^\circ$ かつ $\log C_J(\mathcal{G}') \ge \log C_J(\mathcal{G}) - \gamma/2$ なる閉集合をとる. すると, (6) を利用して, q_s から

$$\liminf_{n\to\infty} \sup_{u \in \mathcal{G}'} \Big(R_n(q, u) - \frac{d}{2} \log \frac{n}{2\pi e} \Big) \le \log C_J(\mathcal{G}') - \gamma/4$$

を満たす q を構成出来ることが示せる. これは, minimax 冗長度の値 [9] よりも小さいので矛盾である. これより, minimax リグレットの下界として次式を得る.

$$\liminf_{n\to\infty} \Big(\bar{r}(\mathcal{X}^n(\mathcal{G})) - \frac{d}{2} \log \frac{n}{2\pi} \Big) \ge \log C_J(\mathcal{G}) \qquad (9)$$

$C_J(\mathcal{G}) = \infty$ の場合は, $C_J(K) < \infty$ なる $\mathcal{G}$ の任意の部分集合 K について同じ議論が成り立つことから, 依然 (9) が成り立つことが分かる.

6.2 指数型分布族についての minimax 符号

$S = \{p(\cdot|\theta) : \theta \in \Theta\}$ を d 次元の指数型分布族, $\mathcal{G} \subseteq \Theta^\circ$ を $\bar{\mathcal{G}^\circ} = \bar{\mathcal{G}}$ なる有界閉集合とすると, $C_J(\mathcal{G}) < \infty$ である. ここで, $\{\mathcal{G}_i\}$ を $\mathcal{G}_i^\circ \supset \mathcal{G}$ を満たす閉集合の列とする. また $C_J(\mathcal{G}_i) \to C_J(\mathcal{G})$ $(i \to \infty)$ とし, さらに $\inf\{|x-y| : x \in \partial\mathcal{G}, y \in \partial\mathcal{G}_i\}$ がゆっくりと 0 に収束すると仮定する ($\partial\mathcal{G}$ は $\mathcal{G}$ の境界). $m_{\mathcal{G}_n}$ で, $\mathcal{G}_n$ における $p(x^n|\theta)$ の Jeffreys 混合を表す. このとき,

$$\limsup_{n\to\infty} \sup_{x^n \in \mathcal{X}^n(\mathcal{G})} \Big(r(m_{\mathcal{G}_n}, x^n) - \frac{d}{2} \log \frac{n}{2\pi} \Big) \le \log C_J(\mathcal{G})$$

が成り立つ. これは, (9) と併せて列 $\{m_{\mathcal{G}_n}\}$ が漸近的に minimax リグレットを達成することを意味する.

これは次の様に Laplace 近似を使って示される. B_n を, 中心が $\hat{\theta}$, 半径が $\log n/\sqrt{n}$ のボールとするとき、

$$\begin{aligned}
\frac{m_{\mathcal{G}_n}(x^n)}{p(x^n|\hat{\theta})} &\ge \frac{\int_{B_n} p(x^n|\theta) w_{\mathcal{G}_n}(\theta) d\theta}{p(x^n|\hat{\theta})} \\
&\sim \int_{B_n} \exp\Big(\frac{-n\theta \cdot \hat{J}(x^n, \hat{\theta})\theta}{2}\Big) w_{\mathcal{G}_n}(\theta) d\theta \\
&\sim \frac{w_{\mathcal{G}_n}(\hat{\theta})(2\pi)^{d/2}}{n^{d/2}(\det(\hat{J}(x^n, \hat{\theta})))^{1/2}} \\
&\sim \frac{(\det(J(\hat{\theta})))^{1/2}(2\pi)^{d/2}}{C_J(\mathcal{G}_n) n^{d/2}(\det(\hat{J}(x^n, \hat{\theta})))^{1/2}}
\end{aligned}$$

となる. ここで, 指数型分布族の場合は $J(\hat{\theta}) = \hat{J}(x^n, \hat{\theta})$ であるから, 次式を得る (実は逆向きの評価も出来る).

$$\frac{m_{\mathcal{G}_n}(x^n)}{p(x^n|\hat{\theta})} \gtrsim \frac{(2\pi)^{d/2}}{C_J(\mathcal{G}_n) n^{d/2}}.$$

6.3 一般のモデルについての minimax 符号

一般のモデルの場合は, 指数型分布族の場合のような $m_{\mathcal{G}_n}$ は漸近的 minimax にならない. なぜならこの場合も

$$\frac{m_{\mathcal{G}_n}(x^n)}{p(x^n|\hat{\theta})} \sim \frac{(\det(I(\hat{u})))^{1/2}(2\pi)^{d/2}}{C_J(\mathcal{G}_n) n^{d/2}(\det(\hat{I}(x^n, \hat{u})))^{1/2}}$$

が成り立つが, $\det(I(\hat{u}))$ と $\det(\hat{I}(x^n, \hat{u}))$ の比が 1 から有限の値だけずれるような x^n が存在するからである. この場合, S をある方法で高い次元に拡大して得られる

モデル S_e に関する Bayes 混合を $m_{\mathcal{G}_n}$ に少しだけ混ぜることにより mimimax 符号が得られる. S_e は

$$p_e(x|u,v) \stackrel{\rm def}{=} p(x|u)\exp\bigl(\sum_{i\le j}\bigl(\hat{I}_{ij}(x,u)-I_{ij}(u)\bigr)v_{ij}\bigr)/\Lambda(u,v)$$

を用いて, $S_e \stackrel{\rm def}{=} \{p_e(\cdot|u,v) : u\in U, |v|\le c_3\}$ と定義する. ここで, v は $v_{ij}(i\le j)$ を成分とする $d(d+1)/2$ 次元のパラメータ, $\Lambda(u,v)$ は

$$\Lambda(u,v) \stackrel{\rm def}{=} \int p(x|u)\exp(\sum_{i\le j}(\hat{I}_{ij}(x,u)-I_{ij}(u)\bigr)v_{ij}))\nu(dx).$$

で定まる規格化定数である. このとき,

$$m_n(x^n) \stackrel{\rm def}{=} (1-\epsilon_n)m_{\mathcal{G}_n}(x^n)+\epsilon_n\int p_e(x^n|u,v)w(u,v)dudv$$

に基づく符号は漸近的に minimax となる. ただし, $m_{\mathcal{G}_n}$ は前節と同様に定義した $\mathcal{G}_n\subseteq U$ に関する Jeffreys 符号, $w(u,v)$ は $\mathcal{G}_1\times\{v:|v|\le c_3\}$ で常に正の値をとる滑らかな事前分布, ϵ_n は多項式の速さで 0 に収束する数列である.

証明は次のようにする. まず性質がよいデータ列の集合を $G_n \stackrel{\rm def}{=} \{x^n : |\hat{I}(x^n|\hat{u})-I(\hat{u})|\le a_n\}$ と定義する. ただし $a_n = n^{-1/4}$ とする. $x^n\in G_n$ に対しては,

$$\frac{m_n(x^n)}{p(x^n|\hat{u})} \ge \frac{(1-\epsilon_n)m_{\mathcal{G}_n}(x^n)}{p(x^n|\hat{u})} \gtrsim \frac{(2\pi)^{d/2}}{C_J(\mathcal{G}_n)n^{d/2}} \qquad (10)$$

が成り立つ.

$x^n\notin G_n$ の場合, $|\hat{I}(x^n|\hat{u})-I(\hat{u})|>a_n$ を利用して

$$\begin{aligned}\frac{p_e(x^n|\hat{u},\tilde{v})}{p(x^n|\hat{u})} &= \frac{\exp\Bigl(n\sum_{i\le j}\bigl(\hat{I}_{ij}(x^n,\hat{u})-I_{ij}(\hat{u})\bigr)\tilde{v}_{ij}\bigr)\Bigr)}{\Lambda(\hat{u},\tilde{v})}\\ &\ge \exp(c_4 n a_n^2) \qquad (11)\end{aligned}$$

なる $\tilde{v}$ が存在することが示せる. また, 簡単な計算により

$$\frac{\bar{m}(x^n)}{p_e(x^n|\hat{u},\tilde{v})} \ge \frac{c_5}{n^{d+d(d+1)/2}}$$

が示せる. ただし, $\bar{m}(x^n) \stackrel{\rm def}{=} \int p_e(x^n|u,v)w(u,v)dudv$ とした. よって (11) を用いて

$$\begin{aligned}\frac{\bar{m}(x^n)}{p(x^n|\hat{u})} &= \frac{\bar{m}(x^n)}{p_e(x^n|\hat{u},\tilde{v})}\frac{p_e(x^n|\hat{u},\tilde{v})}{p(x^n|\hat{u})}\\ &\ge \frac{c_5\exp(c_4 n a_n^2)}{n^{d+d(d+1)/2}} \ge \frac{c_5\exp(c_4\sqrt{n})}{n^{d+d(d+1)/2}}\end{aligned}$$

となり, 結局 $x^n\notin G_n$ については

$$\frac{m_n(x^n)}{p(x^n|\hat{u})} \ge \frac{\epsilon_n\bar{m}(x^n)}{p(x^n|\hat{u})} \ge \frac{\exp(\sqrt{n})}{\mathrm{poly}(n)}\to\infty$$

を得る (リグレットが負になる). よって, 結局すべての $x^n\in\mathcal{X}^n(\mathcal{G})$ について (10) が一様に成り立つので m_n は漸近的に minimax となる. 上記の解析において, 全ての $u\in\mathcal{G}_1$, 全ての v ($|v|\le c_6\le c_3$) について $\Lambda(u,v)$ が有限であることを仮定している.

確率変数 $\hat{I}(x,u)-I(u)$ は, S の埋め込み e-曲率 (指数曲率 [2]) と密接な関係がある (これが 0 ならば e-曲率は 0). e-曲率とはモデルがどの程度指数型分布族からずれているかを表す指標であり, もし至るところで 0 ならば S は指数型となる. 上記で導入した S_e は, 局所指数族バンドルという構造 [23] と本質的に同じである. それは S の各点に, 局所的な指数型分布族を付与した構造である.

この事情を反映して, S が曲指数型分布族である場合は, 漸近的 minimax 符号の構成は著しく簡単になる. すなわち, 拡大モデル S_e として, S が埋め込まれているところの指数型分布族 $\bar{S}$ を用いればよい.

6.4 境界の問題

これまで述べた最尤符号や修正 Jeffreys 符号の符号長の評価は, いずれも $\bar{\mathcal{G}}\subseteq U^\circ$ (あるいは $\bar{\mathcal{G}}\subseteq\Theta^\circ$) を満たす $\mathcal{X}^n(\mathcal{G})$ に関するものであった. この制限のもとでは, 最尤推定値が U の境界に近づくようなデータ列は除外しなければならない.

これに対し, Xie and Barron は多項 Bernoulli モデルについて, $\mathcal{X}^n$ について漸近的に minimax となる修正 Jeffreys 符号を構成してみせた [21][3].

ここでは簡単のため, Bernoulli モデルの場合について, その符号を示す. $\mathcal{X}=\{0,1\}$, $p(1|u)=u$, $p(0|u)=1-u$, $U=[0,1]$ とする. $I(u)=1/u(1-u)$ であるから, Jeffreys 事前分布は $w_J(u)=1/C_J(U)\sqrt{u(1-u)}$ となる. この場合, Jeffreys 符号 $m_J(x^n)=\int p(x^n|u)w_J(u)du$ は $\mathcal{X}^n$ で漸近的 minimax とはならない. というのは $\hat{u}=0$ または 1 である列については, 漸近的にリグレットが $(1/2)\log 2$ だけ大きくなるのである. ところが $w_{3/4}(u)=c_7(u(1-u))^{-3/4}$ (c_7 は規格化定数) と $\epsilon_n=n^{-\beta}(0<\beta<1/4)$ を用いて定義される

$$m_n(x^n)=(1-\epsilon_n)m_J(x^n)+\epsilon_n\int p(x^n|u)w_{3/4}(u)du \qquad (12)$$

は漸近的に minimax である. $\epsilon_n\cdot w_{3/4}$ は w_J に比べて境界近くで大きな値をとり, 結果として境界近くのリグレットを minimax 値にまで下げられるからである.

類似の方法は, 一次元の指数型分布族についても構成出来る [15, 16]. Θ は凸集合だから, 今の場合は区間になる. $\mathcal{X}^n(\Theta)$ に関する問題を考えたいところだが, $C_J(\Theta)<\infty$ となる重要な例は少ない. そこで, Θ を右半分 Θ_r と左半分 Θ_l に分けて, どちらかについて $\sqrt{J(\theta)}$ が積分可能

[3] [21] では Laplace 近似ではなく, $\int p(x^n|u)w_J(u)du$ を解析的に求め, それを近似するという手法を用いている.

である場合を考えよう. 対称性から $\mathcal{X}^n(\Theta_r)$ についてだけ考えれば十分である. 今 λ ($\in \Theta^\circ$) で Θ_l と Θ_r の境界を, b で Θ の右端の点を表す. すなわち $b \stackrel{\rm def}{=} \sup \Theta_r$ とする. ここで, 二つの場合が考えられる. すなわち, b が有限の場合 ($b < \infty$) とそうでない場合である ($b = \infty$). 以下 $\mathcal{G} = \Theta_r$ と書く.

$b = \infty$ の場合, $\sqrt{J(\theta)}$ の積分可能性から, 典型的には $J(\theta) \to 0$ ($\theta \to \infty$) になる. ここでは, 小さな $\epsilon > 0$ について $(J(\theta))^{1/2-\epsilon}$ が $\mathcal{G}$ で積分可能であると仮定する. そして $\mathcal{G}_n$ を, 前節で用いたのと同様の列とし, α_n を $1/\log n$ よりもゆっくり 0 に収束する列とする. このとき事前分布

$$w_n(\theta) \stackrel{\rm def}{=} \frac{(J(\theta))^{(1-\alpha_n)/2}}{\int_{\mathcal{G}_n} (J(\theta))^{(1-\alpha_n)/2} d\theta} \tag{13}$$

を用いた Bayes 符号は漸近的に minimax となる[4]. このケースは Bernoulli モデルと Poisson 分布の場合を含んでいる. 特に Bernoulli モデルの場合は, $C_J(\Theta) < \infty$ となるので, $\mathcal{G} = \Theta$ と出来る. また, 多項 Bernoulli モデルに拡張するのは容易である.

注意 1 *Bernoulli* モデルの場合は, *(13)* と *(12)* という二つの修正した *Jeffreys* 符号が得られる *(*実は *[21]* では他にも同等の符号が出ている*)*. だが, *(13)* には予測確率 $m_n(x_{t+1}|x^t)$ が求め易いという利点がある. m_∞ で *Jeffreys* 符号を表すと, よく知られているように, $m_\infty(1|x^t) = (s + 0.5)/(t + 1)$ となる *(Laplace* 推定値*)*. ただし s は x^t における 1 の数である. *(13)* を用いると $m_n(1|x^t) = (s + 0.5(1 - \alpha_n))/(t + 1 - \alpha_n)$ となる.

$b < \infty$ の場合は, さらに二つの場合に別れる.

一つ目は右端が閉じている場合で, S は regular ではない. この場合は

$$w_n(d\theta) = (1 - n^{-\gamma}) w_{\mathcal{G}_n}(\theta) d\theta + n^{-\gamma} \delta_b(d\theta) \tag{14}$$

という事前分布を用いると漸近的 minimax 符号が得られる. ただし, $w_{\mathcal{G}_n}$ は $\mathcal{G}_n$ の上の Jeffreys 事前分布である. また δ_b は, 端点 b に集中している単位測度であり, $0 < \gamma < 1/2$ とする. この場合は Inverse Gaussian family を含む.

注意 2 b において絶対連続な測度 $w(d\theta)$ *(*例えば *Jeffreys* 事前分布*)* については, 各 n について

$$\lim_{\hat\theta \to b-0} \left(\log p(x^n|\hat\theta) - \log \int p(x^n|\theta) w(d\theta) \right) = \infty$$

が示せる. これはリグレットが無限大になることを意味する. したがって, *(14)* の二番目の項には必然性がある.

[4]興味深いことに, これは α-平行事前分布 [24] の例になっている.

最後に $b < \infty$ かつ Θ の右端が開いている場合 (S が regular ならそうなる) は, ある緩い条件のもとで $\int_{\mathcal{G}} \sqrt{J(\theta)} d\theta = \infty$ となることが示せる.

注意 3 これらの解析を見て分かるように, 境界の問題を考えるには, 境界付近での $J(\theta)$ の振る舞いや境界の形状を特徴付けることが重要である. 従って, 上記の結果を, 一般的な設定で多次元に拡張することは難しい.

注意 4 *6.4* で述べたリグレットに関する全ての漸近的 *minimax* 符号は, 対応する *minimax* 冗長度 $\bar{R}_n(\mathcal{G})$ も漸近的に達成する *[20, 15, 16]*. これは *[9]* の拡張である.

6.5 Markov 情報源についての拡張

i.i.d. の指数型分布族に関する結果を, i.i.d. でない指数型分布族について拡張するのは難しくない. それについてはほぼ同じ命題が成立する.

特に, 以下に述べるように, $\mathcal{X} = \{0, 1\}$ のときの一次 Markov 情報源のクラスについては $\mathcal{X}^n$ に関する漸近的 minimax 符号が得られている. α で 0 の次に 1 が発生する確率を, β で 1 の次に 0 が発生する確率を表す. このとき, (α, β) の Fisher 情報行列の行列式は

$$\frac{\alpha}{\alpha+\beta}\frac{\beta}{\alpha+\beta}\frac{1}{\alpha(1-\alpha)\beta(1-\beta)} = \frac{1}{(\alpha+\beta)^2(1-\alpha)(1-\beta)}$$

で与えられる. ここで $\alpha/(\alpha + \beta)$ と $\beta/(\alpha + \beta)$ は, それぞれ 1 と 0 の出現する定常確率である. w_J でこれを用いて構成される Jeffreys 事前分布の密度を表す. また, $w_{3/4}(\alpha, \beta) \stackrel{\rm def}{=} c_8(\alpha(1-\alpha)\beta(1-\beta))^{-3/4}$ とおき (c_8 は規格化定数), $\gamma \in (0, 1/8)$ とする. このとき $(1-n^{-\gamma})w_J + n^{-\gamma} w_{3/4}$ を用いた Bayes 符号は (8) と同様なリグレットを達成し, 漸近的に minimax となる.

注意 5 この場合は, *Jeffreys* 符号や修正 *Jeffreys* 符号を用いた予測確率は *i.i.d.* の場合のような簡単な形には書けない. 例えば *CTW* 法 *[19]* 等で用いる *Laplace* 推定値は, *Fisher* 情報行列における定常確率を定数に置き換えてしまった事前分布で得られる. しかし *Takeuchi* と *Kawabata* による近似式 *[17]* を使うと, *Jeffreys* 符号による予測確率は

$$m_\infty(1|x^t) = \tilde\alpha + \frac{1}{t_0+1}\left(\frac{1-\tilde\alpha}{2} - \frac{\tilde\alpha(1-\tilde\alpha)}{\tilde\alpha+\tilde\beta}\right) + O\left(\frac{\sqrt{\log t}}{t\sqrt{t}}\right)$$

で与えられる. ただし, $x_t = 0$ とし, t_0 は x^t の中の 0 の数, $\tilde\alpha$ と $\tilde\beta$ は, それぞれ 0 の次に 1 が出る確率と 1 の次に 0 が出る確率の *Laplace* 推定値である.

7 むすび

Rissanen 自身が述べているように, [12] による確率的コンプレキシティの定義は最終的なものであろう. その値は, 最尤符号の符号長を直接評価することによって, あるいは修正 Jeffreys 符号の方法によって, 様々なケースについて

$$\log \frac{1}{p(x^n|\hat{u})} + \frac{d}{2}\log\frac{n}{2\pi} + \log\int_{\mathcal{G}} \sqrt{\det(I(u))}du$$

と評価される. 最尤推定値がパラメータの範囲の境界に近づく場合は一般にはどうなるのか分からない. しかし一次元の指数型分布族の場合には, 境界の性質やその近傍での Fisher 情報量の振る舞いに関わらず, やはり上記の符号長を得ることが確かめられている. このことは, 多次元の一般の場合にもやはり上記の符号長が確率的コンプレキシティであることを予想させる.

こうして得られた確率的コンプレキシティの値は, モデル選択への適用を考えたときに重要であるのみならず, この値を達成するために構成した修正 Jeffreys 符号は, データ圧縮や逐次型予測問題について, minimax 性という最適な性能をもった統計的推測法を提供している. これはまさに, MDL 原理に導かれた成果に他ならない.

謝辞 6.3 で現れる修正 Jeffreys 符号のための拡大モデルが局所指数族バンドルと等価であることは, 甘利俊一先生に教えて頂きました. ご指摘に感謝致します.

参考文献

[1] S. Amari, *Differential-geometrical methods in statistics (2nd pr.)*, Springer-Verlag, 1990.

[2] S. Amari, "Statistical curvature," *Encyclopedia of Statistical Sciences,* vol. 8, pp. 642-646, Willey & Sons, 1994.

[3] A. R. Barron, *Logically smooth density estimation*, Ph.D. thesis, Stanford Univ., 1985.

[4] A. R. Barron & T. M. Cover, "Minimum complexity density estimation," *IEEE trans. IT,* vol. 37, no. 4, pp. 1034-1054, 1991.

[5] A. Barron, J. Rissanen and B. Yu, "The minimum description length principle in coding and modeling," to appear, *IEEE trans. IT,* 1998.

[6] A. R. Barron & J. Takeuchi, "Mixture models achieving optimal coding regret," *Proc. of 1998 Inform. Theory Workshop,* 1998.

[7] R. N. Bhattacharya and R. Rao, *Normal approximation and asymptotic expansions,* John Wiley & Sons, New York, 1976.

[8] L. Brown, *Fundamentals of statistical exponential families*, Institute of Mathematical Statistics, 1986.

[9] B. Clarke & A. R. Barron, "Jeffreys prior is asymptotically least favorable under entropy risk," *JSPI,* 41:37-60, 1994.

[10] J. Rissanen, "Modeling by shortest data description," *Automatica.* 14, pp. 465-471, 1978.

[11] J. Rissanen, "Stochastic complexity," *J. R. Statist. Soc. B,* vol. 49, No. 3. pp. 223-239 and 252-265, 1987.

[12] J. Rissanen, "Fisher information and stochastic complexity," *IEEE trans. IT,* vol. 40, no. 1, pp. 40-47, 1996.

[13] Yu M. Shtarkov, "Universal sequential coding of single messages," *Problems of Information Transmission*, vol. 23, pp. 3-17, July 1988.

[14] J. Takeuchi, "Characterization of the Bayes estimator and the MDL estimator for exponential families," *IEEE trans. IT,* vol. 43, no. 5, pp. 1165-1174, 1997.

[15] J. Takeuchi & A. R. Barron, "Asymptotically minimax regret for exponential families," *Proc. of SITA'97,* pp. 665-668, 1997.

[16] J. Takeuchi & A. R. Barron, "Asymptotically minimax regret by Bayes mixtures," to appear, *Proc. of 1998 IEEE ISIT,* 1998.

[17] J. Takeuchi & T. Kawabata, "Approximation of Bayes code for Markov sources," *Proc. of 1995 IEEE ISIT,* 1995.

[18] C. Wallace & P. Freeman, "Estimating and inference by compact coding," *J. Roy. Statist. Soc. B,* vol. 49, no. 3, pp. 240-265, 1987.

[19] F. Willems, Y. Shtarkov & T. Tjalkens, "The context tree weighting method: basic properties," *IEEE trans. IT,* vol. 41, no. 3, pp. 653-664, 1995.

[20] Q. Xie & A. R. Barron, "Minimax redundancy for the class of memoryless sources", *IEEE trans. IT,* vol. 43, no. 2, pp. 646-657, 1997.

[21] Q. Xie & A. R. Barron, "Asymptotic minimax regret for data compression, gambling and prediction," to appear, *IEEE trans. IT,* 1998.

[22] K. Yamanishi, "A learning criterion for stochastic rules," *Machine Learning,* a special issue for COLT'90, 9(2/3), 1992.

[23] 甘利 俊一, 長岡 浩司, **情報幾何の方法**, 岩波講座応用数学, 岩波書店, 1993.

[24] 竹内 純一, 甘利 俊一, "α-平行事前分布とその性質," **電子情報通信学会技術研究報告**, IT26-20, pp. 61-66, 1996.

[25] 山西 健司, "確率的コンプレキシティと学習理論," **オペレーションズ・リサーチ**, 1996 年 7 月号, pp. 379-386, 1996.

[26] 山西 健司, "拡張型確率的コンプレキシティと学習理論," **IBIS'98 予稿集**, 1998.