2000 年情報論的学習理論ワークショップ
2000 Workshop on Information-Based Induction Sciences (IBIS2000)
Izu, Japan, July 17-18, 2000.

定常確率系列の族に関するミニマックスリグレットについて

On minimax regret with respect to families of stationary stochastic processes

竹内 純一*
Jun-ichi Takeuchi

Abstract: We study the problem of prediction, gambling and data compression of a sequence $x^n = x_1x_2...x_n$, in terms of regret with respect to the class of stationary stochastic processes. In particular, we try to generalize the result of [6] and evaluate lower bound on the minimax regret under the condition that Hellinger rate (generalized Hellinger distance) uniformly converges and other regularity conditions. Further, for Gaussian processes, we give a sufficient condition for the convergence of Hellinger rate.

1 まえがき

アルファベット $\mathcal{X}$ に値を取るデータ列 $x^n = x_1x_2...x_n$ に関する逐次予測, 賭け, データ圧縮の問題を, リグレット [5] を評価基準として考察する. リグレットの基準クラスには定常確率系列のパラメトリックな族 $S = \{p(\cdot|\theta)\}$ を仮定する. 特に minimax リグレットの下界を中心に, [6] の一般化を試み, ある仮定のもとで下界が $n \to \infty$ で

$$\frac{d}{2}\log\frac{n}{2\pi} + \log\int\sqrt{\det J(\theta)}d\theta + o(1) \qquad (1)$$

となることを示す. ここで, d はパラメータ θ の次元, J は θ に関する Fisher 情報行列である.

列 x^n の値を予想するゲームを考える. 競技者 (player) は x^n の各値を確率をもって予想する. 例えば, $\mathcal{X} = \{$ 晴, 雨, 曇 $\}$, $n = 2$ とし, 2 日間の天気を (雨, 晴) の確率 20%, (雨, 雨) が 10%... 等と予想する ($\mathcal{X}^n$ 上の確率分布を構成する). 実際の天気が (雨, 晴) だったならば, $-\log 0.2$ というペナルティ(対数損失) を与えられる. 出た結果に対して予想していた確率が大きければ対数損失は小さい. ただし, 対数損失の絶対値を小さくすることがゲームの目的ではない. ここでは競争相手 (adversary) があり, しかもそれは一人ではない. 競争相手の各々は競技者と同じように確率予想をする. 相手の成績は, 実際に出た結果 (雨, 晴) に最も高い確率を与えた相手のものが採用される. 例えば相手が 3 人あり, (雨, 晴) にそれぞれ 50%, 10%, 5%と予想していれば相手に課される対数損失は $-\log 0.5$ である. ここで,「自分の対数損失 – 相手の対数損失」をリグレットという. リグレットの値は実際に出た結果に依存するが, 最悪の場合を考え, それを最小にするというのが, 競技者の課題である. ただし競技者は, 相手の予想の仕方は前もって教えてもらえる. この条件では例えば相手の予想の平均をとるという戦略が考えられ, これを Bayes 戦略と呼ぶ.

以上において相手の集団を基準クラスと呼ぶ. ここではクラス $S = \{p(\cdot|\theta) : \theta \in \Theta\}$ がそれにあたる. x^n の予想について競技者が分布 q を構成した場合, リグレットは, $\hat{\theta}(x^n)$ を最尤推定値として $\log(p(x^n|\hat{\theta}(x^n))/q(x^n))$ と書くことが出来る. これを「S を基準クラスとする, 列 x^n に関する q のリグレット」と呼ぶ. このとき, 条件付き確率分布 $q(x_t|x^{t-1}) = q(x^t)/q(x^{t-1})$ は, 過去の結果 x^{t-1} を見た後で x_t を予測する確率として用いることが出来る. リグレットは, $\log(p(x^n|\theta)/q(x^n)) = \sum_{t=1}^{n-1}(-\log q(x_t|x^{t-1}) \ -(-\log p(x_t|x^{t-1},\theta)))$ のように, 対数損失の差の総和に書き直せる. すなわち, リグレットは逐次予測の評価基準でもある.

minimax リグレットは, $\mathcal{X}^n$ の部分集合 W_n をに対して定義する. すなわち, $x^n \in W_n$ に関するリグレットの最大値を最小にする予想法のリグレットを minimax リグレットという. 特にコンパクトな $\Theta_s \subset \Theta^\circ$ について, $W_n = \mathcal{X}^n(K) \stackrel{\text{def}}{=} \{x^n : \hat{\theta} \in \Theta_s\}$ というクラスを考えると, 一般に正規化最大尤度 $\hat{m}_n(x^n) = p(x^n|\hat{\theta}(x^n))/\int_{W_n} p(x^n|\hat{\theta}(x^n))dx^n$ が minimax リグレットを達成する [5]. これについて, Fruend[2] は Bernoulli 情報源について, Rissanen[4] はある条件を満たす確率系列の族について, $\hat{m}_n$ のリグレットが (1) となることを示した (第 2 項の積分の範囲は Θ_s). また, Xie & Barron[7] は多項 Bernoulli 情報源について, Takeuchi & Barron[6]

*RWCP[1] 理論 NEC 研究室 [2], 〒 216-8555 神奈川県川崎市宮前区宮崎 4-1-1
1) RWCP: Real World Computing Partnership (新情報処理開発機構)
2) NEC 情報通信メディア研究本部内
Theory NEC Lab., RWCP c/o Computer & Communication Media Research, NEC Corp., 4-1-1 Miyazaki, Miyamaeku, Kawasaki, Kanagawa 216-8555, Japan.

はある条件を満たす i.i.d. 系列および有限アルファベットの Markov 系列について，Jeffreys 事前分布を用いた Bayes 戦略が (1) を達成し，それが下界に一致することを示した．もう少し詳しくは [10] を参照されたい．データ圧縮との関連もそこで見ることが出来る．

本稿では，基準クラスとして一般の確率分布のパラメトリックモデル (i.i.d.)，有限アルファベットに関する Markov モデル，定常ガウス系列のパラメトリックな族 (例えば AR モデル) 等を扱える仮定のもとで議論する．特に下界を示すための条件として，Hellinger 距離の一般化である Hellinger レートを定義し，その収束が一様であると仮定する．この条件のもとでは，Bayes 混合の値の上限が，積分を真のパラメータの近傍に制限したもので近似出来ることが示せる．これに経験的 Fihser 情報量の分布および期待値の収束に関するある条件を加えると，minimax リグレットの下界が得られる．これらの仮定は [4] における「一様に成り立つ中心極限定理」より示しやすいと期待出来る．特に Hellinger レートに関する条件は，かなり広いクラスについて成り立つ．上界についても同様の評価が得られるが，上界評価のために仮定した条件はもう少し厳しい．

2 準備

$(\mathcal{X}, \mathcal{B}, \nu)$ を測度空間とする．$x_i (i = 1, 2, ...)$ を $\mathcal{X}$ の要素とし，$x^n = x_1 x_2 ...$ で長さ n の列を表す．$\nu(dx^n) = \prod_{i=1}^n \nu(dx_i)$ を基準測度とし，$p(x^n|\theta)$ $(\theta \in \Theta \in \Re^d, \bar{\Theta^\circ} = \bar{\Theta})$ を確率系列の密度の族とする．すなわち，$p(x^n|\theta)$ は $p(x^n|\theta) = \int_{\mathcal{X}} p(x^{n+1}|\theta)\nu(dx_{n+1})$ と $\int_{\mathcal{X}} p(x_1|\theta)\nu(dx_1) = 1$ を満たすものとするまた，定常性を仮定する．すなわち，長さ t の列 $x_s x_{s+1} ... x_{s+t}$ が従う周辺分布は s に依存しないものとする．E_θ で $p(\cdot|\theta)$ に関する期待値を，P_θ で $p(\cdot|\theta)$ に対応する確率分布を表す．また，経験的 Fisher 情報量 $(-(1/n) \cdot \log p(x^n|\theta)$ のヘッシアン) を $\hat{J}(x^n, \theta)$ で表す．ただし，log は自然対数を表す．

次に，リグレットを定義する．W_n で $\mathcal{X}^n$ の部分集合，$\mathcal{P}(W_n)$ で W_n 上の全ての確率密度の集合を表す．また，$\hat{\theta} = \hat{\theta}(x^n)$ で最尤推定値を表す．minimax リグレット $\bar{r}(W_n)$ と maximin リグレット $\underline{r}(W_n)$ を

$$\bar{r}(W_n) \stackrel{\rm def}{=} \inf_{q \in \mathcal{P}(W_n)} \sup_{x^n \in W_n} \log \frac{p(x^n|\hat{\theta})}{q(x^n)},$$

$$\underline{r}(W_n) \stackrel{\rm def}{=} \sup_{q \in \mathcal{P}(W_n)} \int_{W_n} q(x^n) \log \frac{p(x^n|\hat{\theta})}{q(x^n)} \nu(dx^n)$$

で定義する ([5, 7] 参照)．定義から $\bar{r}(W_n) \geq \underline{r}(W_n)$ となるが，$\bar{r}(W_n) = \underline{r}(W_n)$ であることが知られている [5]．

正則条件を与えるために Hellinger レートを導入する．

定義 1 密度が $p(\cdot|\theta)$ と $p(\cdot|\theta')$ とで与えられる確率系列間の *Hellinger* レート $d^{(0)}(\theta, \theta')$ を次式で定める．

$$d_n^{(0)}(\theta, \theta') \stackrel{\rm def}{=} 4\Big(1 - \Big(E_\theta\Big(\frac{p(x^n|\theta')}{p(x^n|\theta)}\Big)^{1/2}\Big)^{1/n}\Big),$$

$$d^{(0)}(\theta, \theta') \stackrel{\rm def}{=} \lim_{n\to\infty} d_n^{(0)}(\theta, \theta').$$

Hellinger レートは非負で対称性をもち，$\theta = \theta'$ のとき $d^{(0)}(\theta, \theta') = 0$ となり，距離のように用いることが出来る．特に，次の命題は有用である．

命題 1 全ての $n \geq 0$，任意の $\delta \in \Re$ について次式が成り立つ．

$$P_\theta(\frac{p(x^n|\theta')}{p(x^n|\theta)} \geq e^{-2\delta n}) \leq \exp(-n(\frac{d_n^{(0)}(\theta, \theta')}{4} - \delta)).$$

証明: $\sqrt{p(x^n|\theta')/p(x^n|\theta)}e^{\delta n} \geq 1$ を満たす x^n の集合を G で表すと，評価すべき確率は $P_\theta(G)$ である．ここに

$$\begin{aligned} P_\theta(G) &= \int_G p(x^n|\theta)\nu(dx^n) \\ &\leq \int_G \sqrt{p(x^n|\theta)p(x^n|\theta')}e^{\delta n}\nu(dx^n) \\ &\leq \int \sqrt{p(x^n|\theta)p(x^n|\theta')}e^{\delta n}\nu(dx^n) \\ &= (1 - \frac{d_n^{(0)}(\theta, \theta')}{4})^n e^{\delta n} \\ &\leq \exp(-n(\frac{d_n^{(0)}(\theta, \theta')}{4} - \delta)). \end{aligned}$$

が得られる． **(証明終わり)**

確率系列が i.i.d. の場合は，これは二乗 Hellinger 距離に一致し，その平方根は距離となるが，一般の場合は三角不等式が保証されない．また，後で示すように Gauss 系列の場合，Hellinger レートは

$$\begin{aligned} &d^{(0)}(\theta, \theta') \\ &= 4\Big(1 - \sqrt{2}\exp\Big(\frac{1}{8\pi}\int_{-\pi}^{\pi} \log \frac{S_\theta S_{\theta'}}{(S_\theta + S_{\theta'})^2} d\lambda\Big)\Big) \end{aligned} \quad (2)$$

となる．ただし，$S_\theta = S_\theta(\lambda)$ は $p(\cdot|\theta)$ のパワースペクトラムを表す．[8] では，スペクトラム間の α-ダイバージェンスを導入しているが，$\alpha = 0$ の場合，すなわち二乗 Hellinger 距離は $(1/4\pi)\int(\log S_\theta - \log S_{\theta'})^2 d\lambda$ となり，ここで定義した Hellinger レートとは異なる．なお，定義 1 も α-ダイバージェンスについて一般化出来る．

下界の評価に際し，以下の正則条件を仮定する．

1. ある $\delta > 0$ が存在し，すべての $\theta \in K$，すべての θ': $|\theta' - \theta| < \delta$ について一様に $\tilde{J}_\theta(\theta') \stackrel{\rm def}{=} \lim_{n\to} E_\theta \hat{J}(x^n, \theta')$ が収束し，θ' の関数の族 $\tilde{J}_\theta(\theta')$ は $\theta' = \theta$ において同程度連続である．また，Fisher 情報量 $J(\theta) \stackrel{\rm def}{=} \tilde{J}_\theta(\theta)$ は連続である．

2. 任意の ϵ について，ある $\delta > 0$ が存在し，$\theta \in K$ で一様に次式が成り立つ．

$$P_\theta\Big(\sup_{\theta':|\theta'-\theta|\le\delta} |\hat{J}_{ij}(x^n,\theta') - E_\theta \hat{J}_{ij}(x^n,\theta')| > \epsilon\Big) = o(1/\log n).$$

3. 任意の $\epsilon > 0$ について，$P_\theta(|\hat{\theta}-\theta| > \epsilon) = o(1/\log n)$ が $\theta \in K$ で一様に成り立つ．

4. $\theta, \theta' \in K$ について一様に $d_n^{(0)}(\theta,\theta')$ が Hellinger レートに収束し，$d^{(0)}(\theta,\theta')$ は K^2 で連続である．さらに，$\theta \neq \theta'$ のとき 0 でない．

5. $i = 1, ..., d$ について，次式が成り立つような $C_0 > 0$ が存在する．

$$\sup_{\theta^*\in K} P_{\theta^*}\Big(\sup_{\theta'\in K} |\frac{1}{n}\frac{\partial \log p(x^n|\theta)}{\partial\theta^i}| \ge C_0\Big) = o(\frac{1}{\log n}).$$

以上の条件のうち，条件 1,2 は，[1] で一般の i.i.d. 情報源の族に関する冗長度の評価に用いた条件の一般化である．条件 3 は，リグレットの定義に最尤推定値が現れるために必要となる条件である．条件 4,5 は，[1] と同様に，$\int_{N_\delta^c} p(x^n|\theta)w(\theta)d\theta \le \epsilon \int_{N_\delta} p(x^n|\theta)w(\theta)d\theta$ が $p(\cdot|\theta)$ のもとで高い確率で起こることを示すために用いる．ただし，N_δ は, θ の δ 近傍であり，ϵ と δ は小さな正の数である．条件 3, 4, 5 については，i.i.d. 系列以外に，有限アルファベットの Markov モデルと ARMA モデルについても成り立つ (4 について 5 節で論じる)．しかし，1, 2 はやや厳しく，現在までのところ，i.i.d. 系列以外では，AR モデルと有限アルファベットの Markov モデルについてしか分かっていない．

3 maximin リグレットの下界

以下は本稿の主結果である．

定理 1 条件 *1-5* を仮定する．Θ_s を Θ のコンパクトな部分集合とする．このとき，以下が成り立つ．

$$\liminf_{n\to\infty}(\underline{r}_n(\mathcal{X}^n(\Theta_s)) - \frac{d}{2}\log\frac{n}{2\pi}) \ge \log C_J(\Theta_s).$$

この節では，この定理の証明について述べる．以下，θ^* で真のモデルのパラメータを表す．すなわち，確率変数 x^n は，確率分布 P_{θ^*} に従うものとする．ここで，$\theta^* \in K$ を仮定する．. また，θ の近傍

$$N_\delta(\theta) \stackrel{\rm def}{=} \{\theta' : (\theta'-\theta)^t J(\theta^*)(\theta'-\theta) \le \delta\}$$

を定義する．$N_\delta \stackrel{\rm def}{=} N_\delta(\theta^*)$ と $\hat{N}_\delta \stackrel{\rm def}{=} N_\delta(\hat{\theta})$. なる記法を用いる．まず，次の補題を示す．

補題 1 条件 *4,5* を仮定する．Θ_s を Θ のコンパクトな部分集合，K を Θ_s° の任意の部分集合，w を Θ_s 上の確率密度で，連続とする．このとき，

$$A_n = A_n(\theta^*,\delta,\epsilon) \stackrel{\rm def}{=} \{\int_{N_\delta^c} p(x^n|\theta)w(\theta)d\theta \le \epsilon \int_{N_\delta} p(x^n|\theta)w(\theta)d\theta\}$$

とすると，任意の $\epsilon > 0$ と十分小さな全ての $\delta > 0$ について，$\sup_{\theta^*\in K} P_{\theta^*}({A_n}^c) = o(1/\log n)$ が成り立つ．

証明: まず，[1] に倣い，$P_{\theta^*}({A_n}^c)$ が以下の量で上から押さえられることに注意する (p. 49)．

$$\begin{aligned} & P_{\theta^*}\Big(\int_{N_\delta} p(x^n|\theta)w(\theta)d\theta < e^{nr}\int_{N_\delta^c} p(x^n|\theta)w(\theta)d\theta\Big) \\ \le\ & P_{\theta^*}\Big(p(x^n|\theta^*) < e^{n(r+r')}\int_{N_\delta^c} p(x^n|\theta)w(\theta)d\theta\Big) \\ +\ & P_{\theta^*}\Big(e^{nr'}\int_{N_\delta} p(x^n|\theta)w(\theta)d\theta < p(x^n|\theta^*)\Big). \end{aligned}$$

上式の第一項の P の中の事象を D_n^c，第二項の確率を $P_2(\theta^*)$ とおく．

まず，$P_{\theta^*}(D_n^c)$ を評価する．$\delta_1 \stackrel{\rm def}{=} \min_{\theta^*\in K,\theta\in N_\delta} |\theta^*-\theta|$，$\delta' \stackrel{\rm def}{=} \min_{\theta^*\in K,\theta:|\theta-\theta^*|\ge\delta_1/2} d^2(\theta^*,\theta)/9$ とおく．また，$\iota \stackrel{\rm def}{=} \min\{\delta'/C_0\sqrt{d}, \delta_1/2\}$ とする．今，K の有限個の点の集合 $\underline{K}$ を，$\forall\theta \in K, \exists\theta' \in \underline{K}, |\theta-\theta'| \le \iota$ を満たすように構成する．各 $\bar{\theta} \in \underline{K}$ について半径 ι のボールを $B(\bar{\theta})$ とし，$B(\bar{\theta})$ が N_δ^c と接触する全ての $\bar{\theta}$ の集合を V^* と書く．ι の定義から，$\min_{\theta^*\in K,\bar{\theta}\in V^*} d^2(\theta^*,\bar{\theta}) \ge 9\delta'$ が成り立つ．また，全ての $\theta \in N_\delta^c$ について，$|\theta-\bar{\theta}| \le \delta'/C_0\sqrt{d}$ なる $\bar{\theta} \in V^*$ が存在する．

ここで，次の二つの事象を定義する．

$$\begin{aligned} E_n &\stackrel{\rm def}{=} \{\forall\bar{\theta} \in V^*,\ \frac{p(x^n|\bar{\theta})}{p(x^n|\theta^*)} \le e^{-2\delta' n}\}, \\ F_n &\stackrel{\rm def}{=} \{\sup_{\theta\in K}\max_i |\frac{1}{n}\frac{\partial\log p(x^n|\theta)}{\partial\theta^i}| \le C_0\}. \end{aligned}$$

以下に，十分大きな n について，$E_n \cap F_n \subset D_n$ を示す．$x^n \in E_n \cap F_n$ を仮定すると，$|\theta-\bar{\theta}| \le \iota$ のとき，Taylor の定理により，

$$\begin{aligned} & \Big|\frac{\log p(x^n|\theta)}{n} - \frac{\log p(x^n|\bar{\theta})}{n}\Big| \\ =\ & \Big|\sum_i \frac{1}{n}\frac{\partial\log p(x^n|\tilde{\theta})}{\partial\tilde{\theta}^i}(\theta^i-\bar{\theta}^i)\Big| \le C_0\sqrt{d}\iota \le \delta' \end{aligned}$$

となる．ここで，ある $\alpha \in [0,1]$ について $\tilde{\theta} = \alpha\theta + (1-\alpha)\bar{\theta}$ であり，$x^n \in F_n$ と，$\iota \le \delta'/\sqrt{d}C_0$ を使った．よって，$p(x^n|\theta) \le p(x^n|\bar{\theta})e^{n\delta'}$ が成り立つ．これと $x^n \in E_n$

とから

$$\sup_{\theta\in N_\delta^c}\frac{p(x^n|\theta)}{p(x^n|\theta^*)} \leq \max_{\bar\theta\in V^*}\sup_{\theta:|\theta-\bar\theta|\leq\iota}\frac{p(x^n|\theta)}{p(x^n|\theta^*)} \leq \max_{\bar\theta\in V^*}\frac{p(x^n|\bar\theta)}{p(x^n|\theta^*)}e^{n\delta'} \leq e^{-n\delta'}$$

を得る．よって，$\int_{N_\delta^c} p(x^n|\theta)w(\theta)d\theta \leq e^{-n\delta'}\cdot p(x^n|\theta^*)$ が導かれる．ここで，$r, r' > 0$ は任意であるから，$r + r' < \delta'$ を仮定すると上式は $x^n \in D_n$ を意味する．よって，$P_{\theta^*}(D_n^c) \leq P_{\theta^*}(E_n^c) + P_{\theta^*}(F_n^c)$ が得られた．これらの確率を評価すればよいが，条件 5 により $P_{\theta^*}(F_n^c) = o(1/\log n)$ となる．また，E_n^c については $E_n^c = \{\exists\bar\theta \in V^*, p(x^n|\bar\theta)/p(x^n|\theta^*) > e^{-2\delta' n}\}$ であるから，命題 1 より，$P_{\theta^*}(E_n^c) \leq \sum_{\bar\theta\in V^*} e^{-(d_n^{(0)}(\theta^*,\bar\theta)/4-\delta')n}$ となる．条件 4 により，$d_n^2(\theta^*,\bar\theta) \to d^2(\theta^*,\bar\theta) \geq 9\delta'$ であるから，ある n' が存在して，全ての $n \geq n'$，全ての $\theta^* \in K$，全ての $\bar\theta \in V^*$ について $d_n^2(\theta^*,\bar\theta) \geq 8\delta'$ となる．よって，$n \geq n'$ のとき，$P_{\theta^*}(E_n^c) \leq Ce^{-\delta' n}$ となる．ここで C は V^* の要素数である．よって，$\theta^* \in K$ について一様に $P_{\theta^*}(E_n^c) = O(e^{-n\delta'})$ となり，$\sup_{\theta^*\in K} P_1(\theta^*) = o(1/\log n)$ を得た．

最後に P_2 を評価する．$\delta_n = 1/\sqrt{n}$ とおき，積分を θ^* の近傍 N_{δ_n} で評価する．条件 5 と Taylor の定理により，$1-o(1/\log n)$ 以上の確率で $\forall\theta \in N_{\delta_n}, p(x^n|\theta)/p(x^n|\theta^*) \geq e^{-C^2}$ が成り立ち，よって

$$\int_{N_\delta} p(x^n|\theta)w(\theta) \geq n^{-d/2}e^{-C'^2}p(x^n|\theta^*)$$

を得る．これから $P_2 \leq o(1/\log n)$ が出る．**(証明終わり)**

次に，これを用いて以下の補題を示す．以下，w_K は $C_J(K) \stackrel{\rm def}{=} \int_K \sqrt{\det(J(\theta))}d\theta$, $w_K(\theta) \stackrel{\rm def}{=} 1_K\sqrt{J(\theta)}/C_J(K)$ で定義される K 上の Jeffreys 事前分布である．ただし，1_A は集合 A の特性関数を表す．

補題 2 条件 *1-5* を仮定する．Θ_s を Θ のコンパクトな部分集合とする．$m_{\Theta_s} = \int_{\Theta_s} p(\cdot|\theta)w_{\Theta_s}(\theta)d\theta$ とおく．K を Θ_s° の部分集合とすると，以下が成り立つ．

$$\liminf_{n\to\infty}\Big(\inf_{\theta\in K} E_\theta 1_{\mathcal{X}^n(\Theta_s)}\log\frac{p(x^n|\hat\theta)}{m_{\Theta_s}(x^n)} - \frac{d}{2}\log\frac{n}{2\pi}\Big) \geq \log C_J(\Theta_s).$$

証明は付録 A で述べる．

定理 1 の証明は，[6] における確率系列の族に関する証明と同様に出来る．ここでは方針だけ述べる．

$$m_i(x^n) \stackrel{\rm def}{=} 1_{\mathcal{X}^n(K_i)}\int_K p(x^n|\theta)w_K(\theta)d\theta/M_i$$

(M_i は規格化定数) とする．ただし $\{K_i\}$ は，$K_{i+1} \supset K_i$ なる K の部分集合列で，$\int_{K_i} d\theta \to \int_K d\theta$ を満たすとする．このとき，

$$\underline{r}(\mathcal{X}^n(K)) \geq \int m_i(x^n)\log\frac{p(x^n|\hat\theta)}{m_i(x^n)}\nu(dx^n)$$

が成り立ち，条件 1-5 のもとで，補題 2 により

$$\liminf_{n\to\infty}(\int m_i(x^n)\log\frac{p(x^n|\hat\theta)}{m_i(x^n)}\nu(dx^n) - \frac{d}{2}\log\frac{n}{2\pi}) \geq \log C_J(K_i)$$

を示せる．そして，$C_J(K_i) \to C_J(K)$ $(i\to\infty)$ となるので，所望の下界を得る．

4 Hellinger レートの一様収束

条件 4 のうちの一様収束の部分ついて考察する．i.i.d. 系列の族の場合，それは自明に成り立つ．ここでは有限アルファベットの Markov 系列の場合と，定常 Gauss 系列の場合を考察する．

4.1 Markov 系列の場合

簡単のため $\mathcal{X} = \{0,1\}$ の場合の 1 次の Markov 系列を考える．η_{ij} で $x_t = i$ のときに $x_{t+1} = j$ となる確率を表す．$p(x^n|q)$ で η で定まる Markov 系列の密度を表すと，$p(x_1|q)$ は初期状態 x_1 の確率分布である．n_{ij} で系列 $x^n (n > 1)$ におけるパターン ij の出現回数を表す．$\sum_{ij} n_{ij} = n-1$ と $|n_{10} - n_{01}| \leq 1$ が成り立つ．$s = n_{01} - n_{10}$ とおくと，$n_{10} = n_{01} - s$，$n_{00} = s + n - n_{11} - 2n_{01} - 1$ を得る．ここで，x_1 が与えられたもとでの $x_2^n = x_2...x_n$ の条件付き確率を $p(x_2^n|x_1,q)$ で表すと，ある関数 $g(x|s)$ を用いて

$$\begin{aligned}&\log p(x_2^n|x_1,q)\\ =\ & n_{11}\log\frac{q_{11}}{q_{00}} + n_{01}\log\frac{q_{01}q_{10}}{{q_{00}}^2} + n\log q_{00} + g(s|q)\end{aligned}$$

と書ける．$\theta_{11} = \log(q_{11}/q_{00})$，$\theta_{01} = \log(q_{01}q_{10}/{q_{00}}^2)$，$\psi(\theta) = -\log q_{00}$，$t_{11} = n_{11}/n$，$t_{01} = n_{01}/n$，$r(x_1,s|q) = e^{g(s|q))}p(x_1|q)$ とおくと，

$$p(x^n|q) = \exp(n(\theta_{11}t_{11} + \theta_{01}t_{01} - \psi(\theta)))r(x_1,s|q)$$

と書ける．これは Markov 系列が漸近的に指数型になることを意味する．q の値をそのレンジの内部に包含される閉集合 K に制限すると，$0 < a \leq r(x_1,s|q) \leq b < \infty$ となる．よって，$e^{n\psi_n(\theta)} = \sum_{x^n}\exp(n(\theta_{11}t_{11}+\theta_{01}t_{01}))$ とおくと，$\sum_{x^n} p(x^n|q) = 1$ より，$e^{n\psi(\theta)}/b \leq e^{n\psi_n(\theta)} \leq e^{n\psi(\theta)}/a$ となる．よって，$\psi_n(\theta) \to \psi(\theta)$ $(n\to\infty)$ が $q \in K$ で一様に成り立つ．また，$e^{n(\psi_n((\theta+\theta')/2)-\psi(\theta)-\psi(\theta'))}/b \leq \sum_{x^n}\sqrt{p(x^n|q)p(x^n|q')} \leq e^{n(\psi_n((\theta+\theta')/2)-\psi(\theta)-\psi(\theta'))}/a$ が容易に確かめられる．よって，$(\sum\sqrt{p(x^n|q)p(x^n|q')})^{1/n}$ は $e^{\psi((\theta+\theta')/2)-\psi(\theta)-\psi(\theta')}$ に一様に収束する．

4.2 定常 Gauss 系列

$\mathcal{X} = \Re$, $E_\theta x_t = 0$ を仮定し，$p(\cdot|\theta)$ で定常 Gauss 系列の密度を表す．自己相関関数を $R_t(\theta) \stackrel{\rm def}{=} E_\theta x_s x_{s+t}$ $(t \geq 0,\ R_{-t} \stackrel{\rm def}{=} R_t)$ で表す．また，パワースペクトラムを $S_\theta(\lambda) \stackrel{\rm def}{=} (1/2\pi)\sum_{t=-\infty}^{\infty} R_t(\theta)e^{i\lambda t}$ とする．$\boldsymbol{x}$ でベクトル $(x_1, x_2, ..., x_n)^T$ を表し，Σ_n を，その ij 成分が $R_{|i-j|}$ である n 元正方行列とすると，

$$p(x^n|\theta) = \frac{1}{(2\pi)^{n/2}|\Sigma_n|^{1/2}} \exp(-\frac{\boldsymbol{x}^T {\Sigma_n}^{-1}\boldsymbol{x}}{2})$$

と書ける ($|\Sigma_n|$ は Σ_n の行列式)．$|i-j| = |k-l|$ のときに $A_{ij} = A_{kl}$ が成り立つ行列を Toeplitz 行列 ([3, 9] 参照) というが，Σ_n はその例である．

$$\begin{aligned}&\sqrt{p(x^n|\theta)p(x^n|\theta')}\\ =\ &\frac{1}{(2\pi)^{n/2}(|\Sigma_n||{\Sigma_n}'|)^{1/4}} \exp(-\frac{\boldsymbol{x}^T({\Sigma_n}^{-1} + {\Sigma_n}'^{-1})\boldsymbol{x}}{4})\end{aligned}$$

となるので

$$\int \sqrt{p(x^n|\theta)p(x^n|\theta')}dx^n = \frac{2^{n/2}|{\Sigma_n}^{-1} + {\Sigma'_n}^{-1}|^{-1/2}}{(|\Sigma_n||\Sigma'_n|)^{1/4}}$$

を得る．一般に $A(A^{-1} + B^{-1})B = A + B$ より $|A^{-1} + B^{-1}| = |A + B|/|A||B|$ が成り立つ．よって，

$$(\int \sqrt{p(x^n|\theta)p(x^n|\theta')}dx^n)^{1/n} = \frac{\sqrt{2}(|\Sigma_n||\Sigma'_n|)^{1/4n}}{|\Sigma_n + \Sigma'_n|^{1/2n}}$$

を得る．従って，$|\Sigma_n|^{1/n}$ の収束を調べればよいが，Toeplitz 行列について知られている公式

$$\lim_{n\to\infty} \frac{|\Sigma_{n+1}|}{|\Sigma_n|} = 2\pi \exp\big(\frac{1}{2\pi}\int_{-\pi}^{\pi} \log S_\theta(\lambda)d\lambda\big)$$

([3, 9]) を用いることが出来る．数列 $a_n(\theta)$ (> 0) について，$a_{n+1}(\theta)/a_n(\theta)$ が，θ について一様に $\lambda > 0$ に収束するならば $(a_n(\theta))^{1/n}$ も θ について一様に λ に収束する．よって上記から，$\log S$ が積分可能ならば Hellinger レートは各点で収束し，(2) が成り立つ．一様な収束のためには $|\Sigma_{n+1}|/|\Sigma_n|$ の一様な収束を示せばよい．それには，$1/S_\theta(\lambda)$ が Fourier 級数で一様に近似できればよい．これは，$1/S_\theta(\lambda)$ の Fourier 展開が q 項で打ち切れる場合は，$n \geq q$ において $|\Sigma_{n+1}|/|\Sigma_n|$ が定数になることから出る ([3]，p.71 および p.79 を参照)．例えば AR モデルの場合は $1/S_\theta(\lambda)$ の Fourier 展開は有限項で打ち切れるので条件を満たす．より一般には $1/S_\theta(\lambda)$ が θ に依存しない Lipschitz 条件を満たしていればよい．

5 minimax リグレットの上界

上界の評価に際しては下界の場合に仮定した条件 1 のほかに，以下を仮定する．

6. ある $b > 0$ について，$B = (-b, b)^{d^2}$ とし，$q_{id+j}(x^n|\theta) \stackrel{\rm def}{=} \hat{J}_{ij}(x^n, \theta) - E_\theta \hat{J}_{ij}(x^n, \theta)$ とおく．このとき，

$$\Big(\int p(x^n|\theta)\exp(nq(x^n|\theta)\cdot\xi)\nu(dx^n)\Big)^{1/n} < C_1$$

が，全ての $\theta \in K$，全ての $\xi \in B$，十分大きな全ての n について成り立つ．

7. 上記の表式の対数を $\psi_n(\theta, \xi)$ とおく．$|\partial\psi_n(\theta, \xi)/\partial\xi_l| < C_2$ が全ての $\theta \in K$，全ての $\xi \in B$，十分大きな全ての n について成り立つ．

これらの条件は指数型分布族や Markov モデルの場合は必要なく，曲指数型分布族の場合は容易に示せる．

この場合，通常の Bayes 混合では (1) は達成出来ないが，[6] で提案された，拡張したモデルに関する Bayes 混合をそのまま用いればよい．以下にそのに方式を示す．$\{G_n\}$ を $G_n \supset K$，$C_J(G_n) < \infty$ なる Θ の部分集合列とし，$\inf_{\theta\in\partial G_n, \theta'\in\partial K} |\theta - \theta'|$ と $\epsilon_n > 0$ はゆっくり 0 に近づくと仮定する．w_{G_n} を G_n 上の Jeffreys 事前分布とし，$m_{G_n} \stackrel{\rm def}{=} \int p(x^n|\theta)w_{G_n}(\theta)d\theta$ とする．このとき，

$$(1 - \epsilon_n)m_{G_n}(x^n) + \epsilon_n \int p_e(x^n|\theta, \xi)w(\theta, \xi)d\theta d\xi$$

なる分布は漸近的に minimax リグレットを達成する．ただし w は一様な事前分布であり，

$$p_e(x|\theta, \xi) \stackrel{\rm def}{=} p(x|\theta)\exp(q(x^n|\theta)\cdot\xi - \psi_n(\theta, \xi))$$

とおいた．密度 $p_e(x|\theta, \xi)$ が作る族を拡大モデルと呼ぶ．

残念なことに，一般には全ての $x^n \in \mathcal{X}^n(K)$ について一様に (1) を達成することは証明出来ない．しかし以下のような集合に制限すればよい ($\beta > \alpha > 0$)．

$$\begin{aligned}\mathcal{F}_n(K, \alpha, \beta) \stackrel{\rm def}{=} \{&x^n : \hat{\theta}(x^n) \in K\\ \wedge\ &\max_{i,j} \sup_{\theta:|\theta-\hat{\theta}|<n^{-\beta}} |q(x^n|\theta) - q(x^n|\hat{\theta})| \leq n^{-\alpha}\}.\end{aligned}$$

理由は以下のように説明出来る．Laplace 近似により，

$$\frac{m_{G_n}(x^n)}{p(x^n|\hat{\theta})} \sim \frac{(2\pi)^{d/2}}{n^{d/2}C_J(K)} \frac{|J(\hat{\theta})|^{1/2}}{|\hat{J}(x^n, \hat{\theta})|^{1/2}}$$

となり，m_{G_n} だけを用いた場合には，$|q(x^n|\hat{\theta})|$ が小さい場合にリグレットが minimax 値に近くなる．そうでない場合に，$\epsilon_n \int p_e(x^n|\theta, \xi)w(\theta, \xi)d\theta d\xi$ の部分がリグレットを下げる働きをする．というのは，$\tilde{\xi} = a_n q(x^n|\hat{\theta})/|q(x^n|\hat{\theta})|$ とおくと ($a_n = 2d(n^{-\alpha} + n^{-1/4})$)，$p_e(x^n|\hat{\theta}, \tilde{\xi})/p(x^n|\hat{\theta}) \geq e^{nCa_n^2} \geq e^{\sqrt{n}}$ となり，$(\hat{\theta}, \tilde{\xi})$ における尤度は最尤推定値の尤度よりもずっと高くなる．したがって，$(\hat{\theta}, \tilde{\xi})$ の $n^{-\beta}$ 近傍で積分 $\int p_e(x^n|\theta, \xi)w(\theta, \xi)d\theta d\xi/p((x^n|\hat{\theta}, \tilde{\xi})$ を

評価すると，通常は $n^{-(d+d^2)\beta}e^{\sqrt{n}}$ 以上となることが示せる．ここで通常とは,「$(\hat{\theta},\hat{\xi})$ 付近で尤度が急激に変化しない」ということである．これは $x^n \in \mathcal{F}_n(\alpha,\beta,K)$ ならば満たされる．しかし，例えば $\partial\hat{J}(x^n,\theta)/\partial\theta^i$ や，より高階の微分が有界であることは一般には期待できず，x^n に制限を加えなくてはならないのである．

しかし S が曲指数型分布族の場合には，S が埋め込まれている指数型分布族を拡大モデルとして利用すれば，その自然パラメータに関する 2 階微分は x^n に依存しないため，x^n への制限を $\hat{\theta} \in K$ だけにすることが出来る．また，$(1/n)\log p(x^n|\theta)$ の $k+1$ 階以上の微分が x^n に依存しない場合，$q(x^n|\theta)$ として，$\hat{J}-J$ だけでなく，3 階から k 階までの全ての微係数を成分とするベクトルを用いれば同様の議論が可能である．

$q=\hat{J}-J$ が急激に変化しないという条件は，[4] の Condition v) の後半 (p. 42) と類似のものであり，異なる二つの手法に共通に必要となったという意味で興味深い．[4] では，さらに $\hat{J}(x^n,\theta) < C_0 < \infty$ (Condition v) の前半) なる条件を加え，$p(x^n|\hat{\theta})$ を $p(x^n|\hat{\theta}_d)$ で近似することで，$\int p(x^n|\hat{\theta})\nu(dx^n)$ を評価している．ただし，$\hat{\theta}_d$ は離散化したパラメータ空間における最尤推定値である．従ってこれらの条件は全ての $x^n : \hat{\theta} \in K$ について成立しなくてはならない ($p(x^n|\hat{\theta}) \geq p(x^n|\hat{\theta}_d)$ なので下界については必要ない)．例えば曲指数型分布族の場合に，$\mathcal{X}$ が有界でない場合は満たされない．この条件や，本稿での x^n への制限を緩和することは今後の課題である．

謝辞: 定常ガウス系列に関する Hellinger レートの収束性について，電気通信大学の長岡浩司先生にご教示頂きましたことを感謝します．

参考文献

[1] B. Clarke & A. R. Barron, "Jeffreys prior is asymptotically least favorable under entropy risk," *J. Statistical Planning and Inference,* 41:37-60, 1994.

[2] Y. Freund, "Predicting a binary sequence almost as well as the optimal biased coin," *Proc. of the 9th Annual Workshp on Computational Learning Theory,* pp. 89-98, 1996.

[3] U. Grenander & G. Szegö, *Toepliz forms and their applications,* University of California Press, 1958.

[4] J. Rissanen, "Fisher information and stochastic complexity," *IEEE trans. IT,* vol. 40, pp. 40-47, 1996.

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

[6] J. Takeuchi & A. R. Barron, "Asymptotically minimax regret by Bayes mixtures," *in preparation* (an abstract appeared in *Proc. of 1998 IEEE ISIT*), 2000.

[7] Q. Xie & A. R. Barron, "Asymptotic minimax regret for data compression, gambling and prediction," *IEEE trans. IT,* vol. 46, no. 2, pp. 431-445, 2000.

[8] 甘利 俊一，長岡 浩司，**情報幾何の方法**，岩波書店，1993.

[9] 小倉 久直，**確率過程論**，コロナ社，1978.

[10] 竹内純一，"確率的コンプレキシティと Jeffreys 混合予測戦略"，**IBIS'98 予稿集**，pp. 9-16, 1998.

A 補題 2 の証明

集合 $K^{(\delta)}$ を $K^{(\delta)} \stackrel{\rm def}{=} \bigcup_{\theta\in K} N_\delta(\theta)$. で定める．$K^{(\delta)} \subset \Theta_s$ と $\forall\theta \in K^{(\delta)}$, $N_\delta(\theta) \subset \Theta_s$ が成り立つように，δ を小さくとる．ここで ϵ を一つ決め，仮定 1,2 を満たす δ を $\delta(K)$ で表し，δ を

$$\forall\theta^* \in K,\ \forall\hat{\theta} \in N_\delta(\theta^*),\ \forall\theta \in N_\delta(\hat{\theta}),\ |\theta-\theta^*| \leq \delta(K).$$

となるように小さくとる．また，$\bar{\delta}\ (<\delta)$ を

$$\forall\theta^* \in K,\ \forall\theta \in N_{\delta/2}(\theta^*),\ N_{\bar{\delta}}(\theta) \subseteq N_\delta(\theta^*)$$

を満たす値とする．事象 B_n，C_n を以下のように定義する:

$$B_n = B_n(\theta^*,\delta,\epsilon) \stackrel{\rm def}{=} \{\max_{i,j}\sup_{\theta:|\theta-\theta^*|\leq\delta}|\hat{J}_{ij}(x^n,\theta)-J_{ij}(\theta^*)| \leq \gamma\cdot\epsilon/d^2\},$$

$$C_n = C_n(\theta^*,\delta) \stackrel{\rm def}{=} \{\hat{\theta}(x^n) \in N_{\delta/2}\}.$$

ここで，$x^n \in B_n$ のとき，$\theta',\theta'' \in N_\delta$ ならば

$$\{1-\epsilon \leq \frac{(\theta'-\theta^*)^t\hat{J}(x^n,\theta'')(\theta'-\theta)}{(\theta'-\theta^*)^tJ(\theta^*)(\theta'-\theta^*)} \leq 1+\epsilon\}$$

となることに注意.

以下，$x^n \in A_n \cap B_n \cap C_n$ を仮定する．$x^n \in A_n$ より

$$\int_{\Theta_s}\frac{p(x^n|\theta)w_J(\theta)}{p(x^n|\hat{\theta})}d\theta \leq (1+\epsilon)\int_{N_\delta}\frac{p(x^n|\theta)w_J(\theta)}{p(x^n|\hat{\theta})}d\theta.$$

が成り立つので，$\int_{N_\delta} p(x^n|\theta)w_J(\theta)d\theta/p(x^n|\hat{\theta})$ を上から押さえればよい．すなわち，$\hat{\theta}$ のまわりで，積分を Laplace 近似して，

$$\int_{N_\delta}\frac{p(x^n|\theta)w_J(\theta)}{p(x^n|\hat{\theta})}d\theta \leq \frac{g(\theta^*,\delta)(2\pi)^{d/2}w_J(\theta^*)}{n^{d/2}\sqrt{(1-\epsilon)^{2d}\det(J(\theta^*))}} \qquad (3)$$

を得る．ここで，$x^n \in C_n$ より $\hat{\theta}$ が θ^* に近いこと，$x^n \in B_n$ より $\hat{J}$ が $J(\theta^*)$ に近いことを使っている．また，$g(\theta^*,\delta) \stackrel{\rm def}{=} \sup_{\theta\in N_\delta} w(\theta)/w(\theta^*)$ とおいた．よって

$$\frac{m_{\Theta_s}(x^n)}{p(x^n|\hat{\theta})} \leq \frac{(1+\epsilon)g(\theta^*,\delta)}{1-\epsilon}\frac{(2\pi)^{d/2}}{n^{d/2}C_J(\Theta_s)}$$

を得る．$x^n \in \mathcal{X}^n(\Theta_s)$ のとき $\log(p(x^n|\hat{\theta})/m_{\Theta_s}(x^n)) \geq 0$ であるから，

$$E_\theta 1_{\mathcal{X}^n(\Theta_s)}\log\frac{p(x^n|\hat{\theta})}{m_{\Theta_s}(x^n)} \geq \frac{d}{2}\log\frac{n}{2\pi} + \log C_J(\Theta_s) - 2\epsilon$$
$$- \quad P_{\theta^*}(D_n^c)1_{\mathcal{X}^n(\Theta_s)}(\frac{d}{2}\log\frac{n}{2\pi} + \log C_J(\Theta_s) - 2\epsilon)$$

となる．ただし $D_n \stackrel{\rm def}{=} A_n \cap B_n \cap C_n$ とした．ϵ は任意に小さく出来るので，$\sup_{\theta^*\in K} P_{\theta^*}(D_n^c) = o(1/\log n)$. を示せばよい．補題 1 より $\sup_{\theta^*\in K} P_{\theta^*}(A_n^c) = o(1/\log n)$ であり，条件 5 から $\sup_{\theta^*\in K} P_{\theta^*}(C_n^c) = o(1/\log n)$ が容易に示せる．

$P_{\theta^*}(B_n^c)$ の評価は [1, 6] と同様に出来て，$\sup_{\theta^*\in K} P_{\theta^*}(B_n^c) = o(1/\log n)$ が得られる． **(証明終わり)**