2004 年情報論的学習理論ワークショップ
2004 Workshop on Information-Based Induction Sciences (IBIS2004)
Tokyo, Japan, November 8 - 10, 2004.

Markov モデルの指数曲率と Jeffreys 混合予測
Exponential Curvature Jeffreys Mixture Prediction Strategy for Markov Model

竹内純一* 川端勉†
Jun-ichi Takeuchi Tsutomu Kawabata

Abstract: We consider the problem of sequential prediction for classes of Markov sources and their geometrical structures. A class of Markov sources defined by a context tree (tree model) is an exponential family when it drops in an FSMX model. In that case, the Jeffreys mixture is minimax in terms of coding regret and an approximation formula for Jeffreys mixture is known. As for the case in which a tree model is not FSMX, we conjecture that it is not exponential type. We have partly solved this problem, i.e. we show that a certain type of non FSMX tree models is curved.

1 まえがき

文脈木によって定まる Markov 情報源の族 (tree model と呼ぶ) に関する逐次予測問題と微分幾何的性質について考察する．情報源のクラスが i.i.d. の指数型分布族であるとき，Jeffreys 事前分布を用いた Bayes 予測 (以下，Jeffreys 予測) は確率的コンプレキシティを漸近的に達成することが知られている [7]．これは符号長 regret を評価基準としたとき，漸近的に minimax 予測になると言い換えられる．より一般の指数型でないモデルの場合には，Jeffreys 予測は minimax 予測とはならない [7]．これは，Jeffreys 予測の regret の中に埋め込み指数曲率に関係した項が含まれることによっている．この事情は，i.i.d. ではない時系列モデルの場合も同じである．特に有限アルファベットの Markov 連鎖は漸近的に指数型分布族に近づくことがよく知られており ([13] など)，Jeffreys 予測によって minimax regret が達成できることも示されている [9](正確には FSMX モデルいついて示されたもの)．この場合 Jeffreys 事前分布には Markov モデルの各状態の定常確率がふくまれており，混合を計算することは困難である．これに関し，Laplace 推定量を補正する形で表す近似式が知られている [8, 9]．これは定常確率の遷移確率による微分係数を利用するもので，ナイーブな実装では 1 シンボルあたり状態数の 3 乗と，計算コストがまだ高い．しかし，最も単純な二状態の Markov 連鎖の場合には，シミュレーションによって minimax リグレットに近い性能を示すことが報告されている．よってその適用可能範囲を広げることは興味深い課題である．本稿では，これについて 1) 原理的に適用可能な範囲を明らかにするという観点と，2) 適用可能な場面が増やせるよう，計算コストを下げる工夫をするという観点から考察する．

1) についてはまず，[9] で示された Jeffreys 予測が minimax という命題は，FSMX モデルに限定されたものであることに注意する．本稿で取り上げる tree model は，データ圧縮アルゴリズム [5, 11, 3] で用いられもので，文脈木で定義される Markov 連鎖の部分空間である．よって一般には曲指数型分布族となるが，木の形によって，FSM(Finite State Machine) になる場合 (FSMX と呼ばれる [5]) とそうでない場合の二種類に分かれる．ここで，FSM になる場合とは，状態遷移関数が存在する場合であり，対応する tree model は指数型分布族になる．[9] の結果は，これに関して Jeffreys 予測が漸近的に minimax となることを示したものである．一方，Jeffreys 予測の近似式も，明示的には FSMX モデルについて示されているが，簡単な操作で一般の tree model に適用可能である．よって一般の tree model の場合に Jeffreys 予測の regret がどうなるかは興味深い問題である．本稿で

*NEC インターネットシステム研究所，〒 211-8666 神奈川県川崎市中原区下沼部 1753
Internet Systems Research Laboratories, NEC Corporation, 1753 Shimonumabe, Nakahara-ku, Kawasaki, Kanagawa 211-8666, Japan.
†電気通信大学 情報通信工学科，〒 182-8585 東京都調布市調布ヶ丘 1-5-1
Department of Information and Communication Engineering, University of Electro-Communications, 1-5-1 Chofugaoka, Chofu, Tokyo 182-8585, Japan

はこれに対し否定的な解決を与える．すなわち，ある種の tree model は，漸近的にも指数型でないことを示す．従って，FSM でない tree model については Jeffreys 予測の近似式を用いても minimax 予測の近似とはならない．また，この結果は tree model が漸近的に指数型であることは，FSM であることと等価であることを示唆していると考えられる．

次に 2) については，FSMX モデルに関する Jeffreys 予測の近似式について，より効率的に計算する二つのアルゴリズムを提案する．一つは Shermann-Morrisson の公式 (例えば [4]) という逆行列の高速更新アルゴリズムを利用するもので，状態数の 2 乗で計算が可能である．もう一つは遷移確率行列の冪乗を利用した近似アルゴリズムである．この場合必要な掛け算の回数は遷移確率の下限が規定されていれば，状態数にはよらない．また，掛け算に要するコストは tree model の場合は状態数について線形になるため，結局状態数の線形で済むアルゴリズムが得られる．

以下本稿は次のように構成する．2 節で，符号長 regret を評価基準とした逐次予測問題を解説する．3 節で tree source を導入し，その背景を説明する．4 節では，FSMX model が漸近的に指数型となることを説明し，Jeffreys 予測のための近似式を示す．5 節で FSM でない tree source が指数型でない空間を構成することを示す．6 節で近似式を効率的に計算するアルゴリズムについて論じる．これらで 4 節までは解説であり，5,6 節は新しい内容を含む．

2 逐次予測と minimax regret

各時点 t において逐次的に文字 x_t を読み込みながら，それまでに読んだ文字列に基づいて次の時刻に出現する文字を確率的に予測する逐次予測問題を考える [6, 12, 7]．例えば，x^t に基づいて $x = 1$ となる確率を推定し，推定した確率を予測分布とする方法がある．x^t の中に 1 が k 個含まれているならば，k/t の確率で次に 1 が出ると予測するのである．こうした予測戦略の一つを $q(x_{t+1}|x^t)$ という形で書こう．ここでは，予測戦略 q の評価を対数損失で行う．すなわち，実際に x_{t+1} が出現したときの損失を $\log(1/q(x_{t+1}|x^t))$ で計量する．x_{t+1} に付与した確率の値が小さければ損失は大きくなる．このような対数損失を逐次的に総和したものを累積対数損失と呼ぶ．すなわち，$\sum_{t=0}^{n-1} \log(1/q(x_{t+1}|x^t))$ である．さらに予測評価の基準となる基準クラスを用意する．基準クラスとは通常径数付きの情報源の集合であり，$\mathcal{C} = \{r(\cdot|\cdot, u) : u \in U\}$ と書く．これを用いて q の列 x^n に対する regret を以下のように定義する．

$$R(q, x^n, \mathcal{C}) = \sum_{t=0}^{n-1} \Big(\log \frac{1}{q(x_{t+1}|x^t)} - \log \frac{1}{r(x_{t+1}|x^t, \hat{u})} \Big)$$

ただし，$\hat{u}$ は x^n のもとでの u の最尤推定値である．ここで，$\sum_{t=0}^{n-1} \log(1/r(x_{t+1}|x^t, u)) = \log(1/r(x^n|u))$ となるので，$r(\cdot|\cdot, \hat{u})$ は基準クラスの中で最も累積対数損失の小さい情報源である．これは x^n をあらかじめ知っていないと選べないため，手の届かない理想的予測戦略である．リグレットとは理想からの隔たりを意味している．次に q の最悪 regret を $r(q, \mathcal{C}) = \sup_{x^n} R(q, x^n, \mathcal{C})$ で定める．これを最小にする q を minimax 予測戦略とよび，そのときの regret を minimax regret と呼ぶ．q^* で minimax 予測戦略を表すと，$r(q^*, x^n, \mathcal{C})$ が x^n によらないという重要な性質がある．

このような予測問題について $\mathcal{C}$ が指数型分布族の場合は，Jeffreys 事前分布を用いた Bayes 混合によって漸近的に minimax regret が達成出来ることが知られている [7]．パラメータ u に関する Jeffreys 事前分布とは Fisher 情報量 $J(u)$ を用いて密度関数が $w_J(u) = \sqrt{\det J(u)}/C_J$ として定義される事前分布である ($C_J \stackrel{\text{def}}{=} \int \sqrt{\det J(u)} du$ は規格化定数)．これを用いた Bayes 混合は $\rho(x^n) = \int r(x^n|u) w_J(u) du$ として与えられる．これが漸近的 minimax 予測であることは次の様に Laplace 近似を使って示される．B_n を, 中心が $\hat{u}$, 半径が $\log n/\sqrt{n}$ のボールとするとき，

$$\begin{aligned}
\frac{\rho(x^n)}{r(x^n|\hat{u})} &\sim \frac{\int_{B_n} r(x^n|u) w_J(u) du}{r(x^n|\hat{u})} \\
&\sim \int_{B_n} \exp\Big(\frac{-n u^\dagger \hat{J}(x^n, \hat{u}) u}{2}\Big) w_J(u) du \\
&\sim \frac{w_J(\hat{u})(2\pi)^{d/2}}{n^{d/2} (\det(\hat{J}(x^n, \hat{u})))^{1/2}} \\
&\sim \frac{(\det(J(\hat{u})))^{1/2} (2\pi)^{d/2}}{C_J n^{d/2} (\det(\hat{J}(x^n, \hat{u})))^{1/2}}
\end{aligned}$$

となる．ここで，指数型分布族の場合は $J(\hat{u}) = \hat{J}(x^n, \hat{u})$ であるから，

$$\frac{\rho(x^n)}{r(x^n|\hat{u})} \sim \frac{(2\pi)^{d/2}}{C_J n^{d/2}}$$

すなわち，

$$\log \frac{1}{\rho(x^n)} - \log \frac{1}{r(x^n|\hat{u})} \sim \frac{d}{2} \log \frac{n}{2\pi} + \log C_J$$

を得る．これは ρ の regret が x^n によらないことを意味する．また，別に示されている下界に一致することからもこれが minimax regret を達成することが分かる．逆に $\mathcal{C}$ が指数型でない場合は，

$$\frac{(\det(J(\hat{u})))^{1/2}}{(\det(\hat{J}(x^n, \hat{u})))^{1/2}}$$

がキャンセルせず，Jeffreys 予測は minimax とならないことも分かる．ただし，regret ではなく，その期待値版である冗長度 (redundancy) を基準とした場合 ([2] 等参照)，Jeffresy 予測は依然 minimax 予測であると予想される．

3 tree source

tree source[10] を導入する．ここではアルファベットを $\mathcal{X} = \{0,1\}$ に固定する．$\mathcal{X}' \stackrel{\rm def}{=} \mathcal{X} \setminus \{0\}$ とする．T を $\mathcal{X}^* \stackrel{\rm def}{=} \{\lambda\} \cup \mathcal{X} \cup \mathcal{X}^2 \cup ...,$ の有限部分集合とする．ただし λ は空列である．全ての $s \in T$ に対し，そのポストフィックスが T に属すると仮定する (例えば x_1x_2 のポストフィックスとは x_1x_2，x_2，λ である)．このような T を文脈木と呼ぶ．文脈木 T について，∂T を次のように定義する．

$$\partial T \stackrel{\rm def}{=} \{xs : x \in \mathcal{X}, s \in T\} \cup \{\lambda\} \setminus T.$$

∂T の各要素を T の葉と呼ぶ．∂T は，$\mathcal{X}$ の完全ポストフックス集合であることが示せる．すなわち，∂T の要素が ∂T の他の要素のポストフィックスになることはなく，かつ ∂T の要素の長さの集合は Kraft の不等式を等式で満たす．列 $s \in \mathcal{X}^*$ について，$\tau(s)$ で s のポストフィックスのうち ∂T の要素となるものを表す．τ を文脈関数と呼ぶ．$d \stackrel{\rm def}{=} \max_{s\in\partial T} |s|$ ($|s|$ は s の長さ) を T の深さと呼ぶ．もし $|s| \geq d$ ならば $\tau(s)$ は一意に定まる．列 s ($|s| \geq d$) について，$\tau(s)$ を**文脈**または**状態**と呼ぶ．このような文脈木で定義される文脈ごとに，次の文字の発生する確率分布が定義されている Markov 情報源を tree source と呼ぶ [10]．特に，∂T について，任意の $s \in \partial T$ と任意の $x \in \mathcal{X}$ に対して，($|sx| < d$ であっても)$\tau(sx)$ が一意に存在する場合，その tree source を FSMX (Finite State Machine X) source と呼ぶ．この条件は,「状態遷移関数が存在する」と言い替えられる．以下に，FSMX source を定義する T と，FSMX source でない tree source を定義する T の例を与える．

例 1 $T_{e1} = \{\lambda, 0, 1, 10\}$ ($\mathcal{X} = \{0,1\}$) とおくと，$\partial T_{e1} = \{00, 11, 01, 110, 010\}$ となる (図 *1*)．この文脈木については状態遷移関数が存在するので，*FSMX source* になる．

例 2 ∂T_{e1} から '11' と '01' とを除く (T_{e1} から '1' を除く) と, 図 *2* の T_{e2} が得られる．文脈 '1' のもとで '0' が発生すると，次の文脈が '110' なのか '010' なのか決定出来ないのでこれは *FSMX* にならない．

次に，文脈木によって定まる情報源の径数付き族を定義しよう．p_s^x で，文脈 $s \in \partial T$ において $x \in \mathcal{X}$ が発生する確率を表し，p によってベクトル $(p_{s_1}, p_{s_2}, ..., p_{s_{|\partial T|}})$ を表す．パラメータ p_s^1 の動く範囲を $H_s \stackrel{\rm def}{=} [0,1]$ で表し，$H = H(T) \stackrel{\rm def}{=} \prod_{s\in\partial T} H_s$ と定義する．本稿では x_m^n で列 $x_m x_{m+1} ... x_n$ を表し，特に x^n は x_1^n を表すものとする．また，初期列 x_{-d+1}^0 が与えられていると仮定し，s_0 で $\tau(x_{-d+1}^0)$ を表す．p によって定まる列 x^n の発生確率を $q(x^n|p, x_{-d+1}^0)$ で表す．n_s^x で列 x^n における sx の出現回数を表すと，

$$\log q(x^n|p, x_{-d+1}^0) = \sum_{s\in\partial T,\, x\in\mathcal{X}} n_s^x \log p_s^x \qquad (1)$$

となる．また，以後 $n_s \stackrel{\rm def}{=} \sum_{x\in\mathcal{X}} n_s^x$ なる記号も用いる．ここで，以下の $S(T)$ を，T で定まる tree model と呼ぶ.

$$S(T) \stackrel{\rm def}{=} \{q(\cdot|p,\cdot) : p \in H(T)\}. \qquad (2)$$

特に状態遷移関数が存在する場合は $S(T)$ を FSMX model と呼ぶ.

ここで，tree model をデータ圧縮に利用する例として，CTW(Context Tree Weighting) 法 [11] と CONTEXT [5] を見ておこう．まず，予測戦略 q に基づいて，データ列 x^n を $-\log q(x^n)$ の符号長で符号化する符号を構成することが出来る [15] ので, 予測戦略 q は符号器とみなせることに注意する．CTW 法は, 深さ d 以下の文脈木からなる集合 $\mathcal{T}$ に対し，$w(T) = 2^{-|\partial T\cup T|+|\partial T\cap\mathcal{X}^d|}$ で事前確率を定義し，それぞれの木 T のもとでの予測確

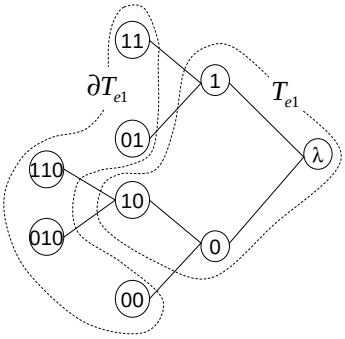

図 1: FSMX source を定義する文脈木

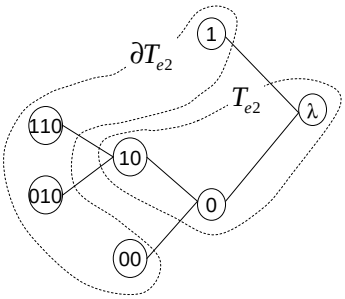

図 2: 一般の文脈木

率を木に関する事後確率で平均して，次の記号の予測確率を求める方法である．ただし，各 T のもとでの予測確率は，

$$\frac{n^x_{\tau(x^n)} + 0.5}{n_{\tau(x^n)} + 1}$$

で計算している．CTW 法の特長は，事前確率 $w(T)$ の性質をうまく使って事後分布を効率的に計算するところにある．もう一方の CONTEXT では, 文字列を順に読み込みながら, 読み込んだ文字列のなかで現れた文脈をもとに文脈木 T を構成していく．出来た文脈木 T は次の文字の予測にそのまま使われるのではない．現在までに読み込んだ文字列を x^i, 出来た文脈木を $T(i)$ とする．CONTEXT は，$T(i) \cup \partial T(i)$ の要素のうち，x^i のポストフィックスに一致するものの中から, MDL 基準に類似した基準で文脈を一つ選択し，その文脈のもとで推定される予測確率で次の文字を予測する．真のモデルの文脈木を $T^\star$ とすると, 系列が無限に長くなるとき, CONTEXT が $\partial T^\star$ に属する文脈を選択する確率が 1 に収束することが Rissanen によって示されており [5]，モデル選択の意味で最適性をもつ．tree source はこうした流れの中で用いられてきた情報源のクラスである．

4 FSMX model と指数型分布族

FSMX model が漸近的に指数型分布族となることを説明しよう．一般に指数型分布族とは $p(z|\theta) = \exp(\theta^i z_i - \psi(\theta))$ となるような確率密度の族を指す [1]．ここで z は k 次元の (実数) ベクトル値の確率変数であり，θ は k 次元パラメータとする．$\theta^i z_i = \sum_i \theta^i z_i$ とする和の規約を用いている．また，$\partial_i \stackrel{\rm def}{=} \partial/\partial\theta^i$ とする．$\eta_i \stackrel{\rm def}{=} \partial_i \psi(\theta)$ とするとき $\eta_i = E_\theta z_i$ が成り立つ．η_i を期待値パラメータ, θ^i を正準パラメータと呼ぶ．また，$J_{ij} = \partial_j \eta_i$ は θ に関する Fisher 情報量に一致する．z の列 z^n の密度関数は $p(z^n|\theta) \stackrel{\rm def}{=} \prod_t p(z_t|\theta) = \exp(n(\theta^i \bar{z}_i - \psi(\theta)))$ となる．ただし $\bar{z} \stackrel{\rm def}{=} (1/n)\sum_t z_t$ である．ここに $\partial_i \log p(z^n|\theta) = n(\bar{z}_i - \eta_i)$ なので，η_i の最尤推定値は十分統計量 $\bar{z}_i$ に一致する．

次に，$S = \{p(\cdot|\theta) : \theta \in \Theta\}$ の部分空間として曲指数型分布族を定義する．$u \in U$ を $k' < k$ 次元のベクトルとし，$\phi : U \to \Theta$ を C^∞ 級の関数とし，そのヤコビアンのランクは常に k' とする．このとき，$M = \{r(\cdot|u) = p(\cdot|\phi(u)) : u \in U\}$ を曲指数型分布族と呼ぶ．u の各成分は u^a で表す．M が指数型ならば，最尤推定値における Fisher 情報量と経験的 Fisher 情報量が一致する．これは

$$\begin{aligned}
\frac{\partial}{\partial u^a}\log r(z^n|u) &= \frac{\partial\theta^i}{\partial u^a}(\bar{z}_i - \eta_i) \\
\hat{J}_{ab}(z^n, u) &\stackrel{\rm def}{=} -\frac{1}{n}\frac{\partial^2}{\partial u^a \partial u^b}\log r(z^n|u) \qquad (3) \\
&= -\frac{\partial^2\theta^i}{\partial u^a \partial u^b}(\bar{z}_i - \eta_i) + J_{ab}(u)
\end{aligned}$$

に注意すると分かる ($J_{ab}(u)$ は u に関する Fisher 情報量，$\hat{J}_{ab}(z^n, u)$ は経験的 Fisher 情報量を表す)．すなわち，最尤推定値においては $(\partial\theta^i/\partial u^a)(\bar{z}_i - \eta_i) = 0$ が成り立つが，$\phi(u)$ が平面をなす場合は $\partial^2\theta^i/\partial u^a\partial u^b$ は $\partial\theta^i/\partial u^a$ が張る空間に属するため (3) 第 3 辺の第 1 項も 0 になる．逆にあらゆるデータ列 z^n について最尤推定値における Fisher 情報量と経験的 Fisher 情報量が一致するならば，M が指数型であることも分かる．実は $J_{ab}(u) - \hat{J}_{ab}(z^n, u)$ は M の (確率密度関数全体の空間に対する) 埋め込み指数曲率と密接に関係した量である．

では，FSMX model が漸近的に指数型となることを見てみよう．これはよく知られた事実であり，例えば [13] で論じられている．[13] では，正準パラメータを書き下すことで示しているが，ここでは $|J_{st}(\hat{p}) - \hat{J}_{st}(z^n_{-d+1}, \hat{p})| \to 0$ となることを示す．まず，経験的 Fisher 情報量の p^1_s と p^1_t に関する成分が

$$\hat{J}_{st}(x^n, p) = \delta_{st}\hat{p}_s\Big(\frac{\hat{p}^1_s}{(p^1_s)^2} + \frac{\hat{p}^0_s}{(p^0_s)^2}\Big)$$

となることに注意する．ここで $\hat{p}_s \stackrel{\rm def}{=} (n^0_s + n^1_s)/n$, $\hat{p}^0_s \stackrel{\rm def}{=} n^0_s/n_s$, $\hat{p}^1_s \stackrel{\rm def}{=} n^1_s/n_s$ とした．よって，Fisher 情報量は $J_{st}(p) = \delta_{st}\mu_s/p^1_s p^0_s$ である．ただし，μ_s は p^1_s によって決まる文脈 s の定常確率であり，$E_p\hat{p}_s$ に一致する．ただし，E_p は $q(\cdot|\cdot, p)$ に関する期待値を表す．また，$\hat{J}_{st}(x^n, \hat{p}) = \delta_{st}\hat{p}_s/\hat{p}^1_s\hat{p}^0_s$ となる．よって殆どの列 x^n について $|\hat{p}_s - \mu_s(\hat{p})| \to 0$ となることを示せばよい．ただし，$\hat{p}$ は p の最尤推定値である．さて，p^x_s と τ によって $s \in \partial T$ から $t \in \partial T$ への遷移確率が与えられるが，それを Π_{ts} と書く．このとき，$\sum_s \Pi_{ts}\mu_s = \mu_t$ が成り立つ．次に $s_i = \tau(x^i_{-d+1})$ として，列 x^n_{-d+1} から文脈の列 s^n_0 を構成し，s^n_0 における st の出現回数を n^t_s と書く．今，$s_0 = s_n$ を仮定すると，$\sum_s n^t_s = \sum_s n^s_t$ が成り立つが，これから $\sum_s n^t_s/n_s \cdot n_s = n_t$ となり，$\sum_s \hat{\Pi}_{ts}\hat{p}_s = \hat{p}_t$ を得る．ここに $\hat{\Pi}_{ts} \stackrel{\rm def}{=} n^t_s/n_s$ としたが，これは $\hat{p}^x_s$ から構成した遷移確率行列であり，$\mu_s(\hat{p}) = \hat{p}_s$ が成り立つことが分かる．$s_0 = s_n$ でない場合は，まず $\tau(x^{n+d}) = s_0$ となるように x^n の後ろに列 x^{n+d}_{n+1} をつなげ，$J(\hat{p}(x^{n+d}_{-d})) = \hat{J}(x^{n+d}_{-d}, \hat{p}(x^{n+d}_{-d}))$ を示す．ただし，$\hat{p}(x^{n+d}_{-d+1})$ は x^{n+d}_{-d+1} に基づく p の最尤推定値とする．次に $J(\hat{p}(x^{n+d}_{-d})) = J(\hat{p}(x^n_{-d})) + O(1/n)$ と $\hat{J}(x^n_{-d}, \hat{p}(x^n_{-d})) = \hat{J}(x^{n+d}_{-d}, \hat{p}(x^{n+d}_{-d})) + O(1/n)$ を示

ことで,

$$|J(\hat{p}(x_{-d}^{n+d})) - \hat{J}(x_{-d}^{n}, \hat{p}(x_{-d}^{n}))| \to 0 \quad (n \to \infty)$$

が分かる．本稿ではこのことをさして「漸近的に指数型」と言っている．より一般的には,

$$q(x^n|\theta) = \exp(n(\theta^i T_i(x_{-d+1}^n) - \psi(\theta)) + U(x_{-d+1}^n|\theta))$$

となるとき，q を漸近的に指数型と定義すればよい．ただし，U の θ による二階微分が有界であるとする．

FSMX モデルの場合にも，Jeffreys 混合が minimax regret の意味でほぼ最適であることは保証される [9]. しかし，FSMX モデルの場合の Jeffreys 事前分布は，既に見たように定常確率分布を含み，その積分計算は容易ではない．そこで，Laplace 推定を補正する形で Jeffreys 予測を近似する以下の式を提案されている [8].

$$\begin{aligned} \tilde{p}_s^x &= \frac{n_s^x + 0.5}{n_s + (k+1)/2} \\ &- \sum_{t \in \partial T} \frac{k\hat{p}_s^1(\delta_{x1} - \hat{p}_s^1)}{2n_s + k + 1} \frac{\partial \log \mu_t(\hat{p})}{\partial p_s^1} \end{aligned} \tag{4}$$

アルファベットが $\{0,1\}$ で，$\partial T = \{0,1\}$ の場合には，数値実験によって minimax regret に近い性能を達成することが示されている．

5 tree model の指数曲率

ここでは FSMX ではない tree model が指数型ではないことを示す．具体的には次の定理が成り立つ.

定理 1 文脈木 T は状態遷移関数をもち，$1s, 0s \in \partial T$ の親ノード $s(\neq \lambda)$ を T から取り除いて得られる文脈木 T' は状態遷移関数をもたないと仮定する．このとき $S(T')$ は指数型ではない.

ここで $S(T')$ は $S(T)$ における超曲面である．定理はその埋め込み指数曲率が 0 でないことを主張している．T と T' の具体例は，例 1 の T_{e1} と例 2 の T_{e2} に与えてある．すなわち，T_{e1} からノード 1 を除いたものが T_{e2} であり，T_{e2} は状態遷移関数を持たない.

定理の証明には次の補題が鍵となる.

補題 1 T を *FSMX source* を定義する文脈木，τ をその文脈関数とする．T から $1\bar{s} \in \partial T$ なるノード $\bar{s}$ を取り除いて得られる文脈木を T' とする．このとき，T' が *FSMX source* を定義するのは $\forall x \in \mathcal{X}, \tau(0\bar{s}x) = \tau(1\bar{s}x)$ のときであり，かつそのときに限る.

証明: τ' で T' で定まる文脈関数を表す．$\forall x \in \mathcal{X}, \tau(0\bar{s}x) = \tau(1\bar{s}x)$ を仮定すると $\forall x \in \mathcal{X}, \tau'(0\bar{s}x) = \tau'(1\bar{s}x)$ が成り立つ．よって $\tau'(\bar{s}x)$ は一意に定まる．$\bar{s}$ 以外の $s \in \partial T'$ は，∂T にも属するので，$\tau(sx)$ は一意に定まる．よって $\tau'(sx)$ も一意に定まる．次に $\exists x \in \mathcal{X}, \tau(0\bar{s}x) \neq \tau(1\bar{s}x)$ を仮定すると，$\tau(0\bar{s}x) \neq \tau(1\bar{s}x)$ なので，$0\bar{s}x$ と $1\bar{s}x$ は ∂T の要素である．よって $\bar{s}x$ は T の要素であり，$\bar{s}$ とは異なる．よって $\bar{s}x$ は T' の要素である．よって，$0\bar{s}x$ と $1\bar{s}x$ は $\partial T'$ の要素であり，$\tau'(0\bar{s}x) \neq \tau'(1\bar{s}x)$ を得る．これは T' の文脈 $\bar{s}$ について $\tau'(\bar{s}x)$ を決定できないことを意味する．従って状態遷移関数は存在しない．**Q.E.D.**

以下に定理 1 の証明を述べる.

定理 1 の証明: ここでは，τ，τ'，$\bar{s}$ を上記補題の証明と同様に用いる．$\partial T = \{s_1, s_2, ..., s_w\}$ とし，$s_1 = 1\bar{s}$, $s_2 = 0\bar{s}$ とおく．すなわち，$\partial T' = \{\bar{s}, s_3, ..., s_w\}$ である．$S(T)$ の点を $q(\cdot|\cdot, p)$ と書き，$S(T')$ の点を $r(\cdot|\cdot, u) = q(\cdot|\cdot, \phi(u))$ と書く．ただし,

$$\begin{aligned} p &= (p_{s_1}^1, p_{s_2}^1, ..., p_{s_w}^1), \\ u &= (u_{\bar{s}}^1, u_{s_3}^1, ..., u_{s_w}^1) \end{aligned}$$

であり，ϕ は $u \mapsto p$ なる関数で,

$$\phi(u) = (u_{\bar{s}}^1, u_{\bar{s}}^1, u_{s_3}^1, ..., u_{s_w}^1)$$

とする．また，(p, T) で定義される s から t への遷移確率を Π_{ts} と書く．$t = \tau(s1)$ ならば $\Pi_{ts} = p_s^1$ である．今，列 x_{-d+1}^n に τ を適用して文脈の列 s_0^n を構成し，その中で現れる st の個数を n_s^t と書き，$n_s \stackrel{\text{def}}{=} \sum_t n_s^t$ とする．このとき,

$$\begin{aligned} &\log r(x^n|x_{-d+1}^0, u) \\ &= \sum_{s \in \{s_1, s_2\}} (n_s^{\tau(s1)} \log u_{\bar{s}}^1 + n_s^{\tau(s0)} \log u_{\bar{s}}^0) \\ &+ \sum_{s \notin \{s_1, s_2\}} (n_s^{\tau(s1)} \log u_s^1 + n_s^{\tau(s0)} \log u_s^0) \end{aligned}$$

となる．右辺第 1 項は

$$\begin{aligned} &n \sum_{s \in \{s_1, s_2\}} (\frac{n_s}{n} \frac{n_s^{\tau(s1)}}{n_s} \log u_{\bar{s}}^1 + \frac{n_s}{n} \frac{n_s^{\tau(s0)}}{n_s} \log u_{\bar{s}}^0) \\ = \; &n \sum_{s \in \{s_1, s_2\}} (\hat{p}_s \hat{p}_s^1 \log u_{\bar{s}}^1 + \hat{p}_s \hat{p}_s^0 \log u_{\bar{s}}^0) \\ = \; &n\hat{p}_{\bar{s}} (\hat{u}_{\bar{s}}^1 \log u_{\bar{s}}^1 + \hat{u}_{\bar{s}}^0 \log u_{\bar{s}}^0) \end{aligned}$$

と書き直せる．ただし，$\hat{p}_{\bar{s}} = \hat{p}_{s_1} + \hat{p}_{s_2}$，$\hat{u}_{\bar{s}}^1 = (\hat{p}_{s_1} \hat{p}_{s_1}^1 + \hat{p}_{s_2} \hat{p}_{s_2}^1)/\hat{p}_{\bar{s}}$ である．同様に第 2 項は

$$n \sum_{s \notin \{s_1, s_2\}} \hat{p}_s (\hat{u}_s^1 \log u_s^1 + \hat{u}_s^0 \log u_s^0)$$

となる．ただし $\hat{u}_s^x = \hat{p}_s^x$ である．よって,

$$\log r(x^n|x_{-d+1}^0, u) = n \sum_{s \in \partial T'} \hat{p}_s (\hat{u}_s^1 \log u_s^1 + \hat{u}_s^0 \log u_s^0)$$

と書ける．ここで経験的 Fisher 情報量の u_s^1 と u_t^1 に関する成分を $\hat{J}_{st}(x^n,u)$ と書くと，

$$\hat{J}_{st}(x^n,u)=\delta_{st}\hat{p}_s\Big(\frac{\hat{u}_s^1}{(u_s^1)^2}+\frac{\hat{u}_s^0}{(u_s^0)^2}\Big)$$

となり，$\hat{J}_{st}(x^n,\hat{u})=\delta_{st}\hat{p}_s/\hat{u}_s^1\hat{u}_s^0$ を得る．また，Fihser 情報量を $J_{st}(u)$ で書くと $J_{st}(\hat{u})=\delta_{st}\mu_s(\phi(\hat{u}))/\hat{u}_s^1\hat{u}_s^0$ となる．これによって，$\hat{J}_{st}(x^n,\hat{u})-J_{st}(\hat{u})$ を調べるには $\mu_s(\phi(\hat{u}))$ と $\hat{p}_s$ の違いを調べればよいことが分かる．今，$\tau(x^s_{-d+1})=\tau(x^n)$ を仮定すると，$\hat{p}_s=\mu_s(\hat{p})(s\in\partial T)$ が成り立つ．よって結局，$\mu_s(\phi(\hat{u}))$ と $\mu_s(\hat{p})$ の違いを調べることに帰着する．すなわち，$\mu_s(\phi(\hat{u}))-\mu_s(\hat{p})$ $(s\in\partial T')$ が 0 でないことを示せばよい．ただし，$\mu_{\bar{s}}=\mu_{s_1}+\mu_{s_2}$ としている．これを行うために，$\mu(\hat{p}+\epsilon(\phi(\hat{u})-\hat{p}))$ を ϵ の関数として微分する．今，$p=\hat{p}+\epsilon(\phi(\hat{u})-\hat{p})$ に対応する遷移確率の行列を $\Pi^{(p)}$ とおくと，$(\Pi^{(p)}-I)\mu(p)=0$ である (I は単位行列)．よって $\epsilon=0$ において微分すると $\dot{\Pi}^{(\hat{p})}\mu(\hat{p})+(\Pi^{(\hat{p})}-I)\dot{\mu}(\hat{p})=0$ となる．ここで $\dot{\mu}$ などは ϵ による微分を表す．これを変形して

$$\dot{\Pi}^{(\hat{p})}\mu(\hat{p})=(I-\Pi^{(\hat{p})})\dot{\mu}(\hat{p}) \tag{5}$$

と書く．ここで，$\delta p\stackrel{\rm def}{=}\phi(\hat{u})-\hat{p}$ の各成分のうち 0 でないのは s_1 と s_2 に対応する成分のみである．すなわち

$$\begin{aligned}\delta p^1_{s_1} &= \hat{u}^1_{\bar{s}}-\hat{p}^1_{s_1}=\frac{\hat{p}_{s_1}\hat{p}^1_{s_1}+\hat{p}_{s_2}\hat{p}^1_{s_2}}{\hat{p}_{\bar{s}}}-\hat{p}^1_{s_1}\\ &= \frac{\hat{p}_{s_2}(\hat{p}^1_{s_2}-\hat{p}^1_{s_1})}{\hat{p}_{\bar{s}}}\end{aligned}$$

であり，同様に $\delta p^1_{s_2}=\hat{p}_{s_1}(\hat{p}^1_{s_1}-\hat{p}^1_{s_2})/\hat{p}_{\bar{s}}$ である．これらから，$\hat{p}_{s_2}\delta p^1_{s_2}=-\hat{p}_{s_1}\delta p^1_{s_1}$ を得る．今 $\Pi^{(p)}_{ij}$ で s_j から s_i への遷移確率を表すことにすると，$\dot{\Pi}^{(\hat{p})}$ のうち 0 でないのは 1 列目と 2 列目である．1 列目は s_1 からの遷移を表すが，$\tau(s_11)=s_a$，$\tau(s_10)=s_b$ とすると，1 列目で 0 でないのは a 行目と b 行目である．同様に 2 列目については，$\tau(s_21)=s_c$，$\tau(s_20)=s_d$ とおく．このとき補題 1 より，$a\neq c$ または $b\neq d$ が成り立つ (もとより $a\neq d$ かつ $b\neq c$ は成り立つ)．一般性を失うことなく $a\neq c$ が仮定出来る．このとき，$a\neq d$ とから，

$$\begin{aligned}(\dot{\Pi}^{(\hat{p})}\mu(\hat{p}))_a &= \delta p^1_{s_1}\mu_{s_1}(\hat{p})\\ (\dot{\Pi}^{(\hat{p})}\mu(\hat{p}))_c &= \delta p^1_{s_2}\mu_{s_2}(\hat{p})=-\delta p^1_{s_1}\mu_{s_1}(\hat{p})\end{aligned}$$

となる．よって，$\delta p^1_{s_1}\neq 0$ であれば (5) の左辺は 0 でない．すなわち，$\dot{\mu}(\hat{p})$ は $\mu(\hat{p})$ と平行でないことが分かる．最後に，$\dot{\mu}(\hat{p})=\lambda(1,-1,0,...,0)^\dagger$ ではないことを示そう．これが成り立つと仮定すると

$$\begin{aligned}((I-\Pi^{(\hat{p})})\dot{\mu}(\hat{p}))_a &= \lambda((\delta_{1a}-\hat{p}^1_{s_1})\\ ((I-\Pi^{(\hat{p})})\dot{\mu}(\hat{p}))_c &= -\lambda(\delta_{2c}-\hat{p}^1_{s_2})\end{aligned}$$

となる．従って $\delta_{1a}-\hat{p}^1_{s_1}=\delta_{2c}-\hat{p}^1_{s_2}$ でなければならないが，殆ど全ての x^n_{-d+1} について成り立たない．これにより $\mu_s(\phi(\hat{u}))-\mu_s(\hat{p})$ $(s\in\partial T')$ が 0 でないことが分かる． **Q.E.D.**

6 効率的アルゴリズム

近似式 (4) に現れる $\partial\log\mu_t/\partial\eta^y_s$ の計算を行う方法を考えよう．

$|\partial T|-1$ 次元のベクトル ν を $\nu_i\stackrel{\rm def}{=}\mu_i/\mu_{|}\partial T|(i\leq|\partial T|-1)$ で定義し，$I-\Pi$ の $|\partial T|$ 番目の行と列を除いて得られる行列を $\tilde{B}$ とおくと，$(I-\Pi)\mu=0$ から $\tilde{B}\nu=\beta$ を得る．ただし，$\beta_i=(I-\Pi)_{|\partial T|i}(i\leq|\partial T|-1)$ である．よって，$\tilde{B}^{-1}\beta$ の微分係数を計算すれば，容易に定常確率の微分係数を求めることが出来る．これはナイーブなアルゴリズムでは $O(|\partial T|^3)$ の時間がかかる．この節ではこの計算量を減らす方法を提示する．

6.1 Shermann-Morrison の公式の利用

ここでは $\tilde{B}^{-1}\beta$ を p^1_s で微分したときの微分係数を直接計算する方法を検討する．$\tilde{B}$ と β の各成分は $-p^1_s$ または $1-p^1_s$ であるから，$\tilde{B}_{kl}$ で微分する問題を考えればよい．特に β の微分は容易なので $\tilde{B}^{-1}$ の微分係数のみ考える．$\tilde{B}^{-1}\tilde{B}=I$ を $\tilde{B}_{kl}$ で微分すると，

$$\frac{\partial\tilde{B}^{-1}}{\partial\tilde{B}_{kl}}\tilde{B}+\tilde{B}^{-1}D^{(kl)}=0$$

となる．ただし $D^{(kl)}$ は (k,l) 成分が 1 で他の成分は 0 である正方行列を表す．これから，

$$\frac{\partial\tilde{B}^{-1}}{\partial\tilde{B}_{kl}}=-\tilde{B}^{-1}D^{(kl)}\tilde{B}^{-1}.$$

を得る．これから，$\tilde{B}^{-1}$ からその微分係数が $O(|\partial T|^2)$ の計算量で得られることが分かる．

さて，我々の問題は単に $\tilde{B}^{-1}$ を求めるのではなく，$\hat{p}^1_s$ などの統計量が逐次更新されていくのに応じて $\tilde{B}^{-1}$ を更新していく問題であることに注意しよう．こうした状況では次の Sherman-Morisson の公式 (例えば [4]) が有用である．

$$(L+vw^t)^{-1}=L^{-1}-\frac{L^{-1}vw^tL^{-1}}{1+v^tL^{-1}w}. \tag{6}$$

ただし，L は正方行列で v と w は列ベクトルである．右辺は行列とベクトルの掛け算までしか現れないので計算量は行列の次数の二乗である．この公式を用いると，行列 L とその逆行列が得られていれば，L が $L'=L+vw^t$ と変化したときにその逆行列を次数の二乗の計算量で求めることが出来る．我々の問題の場合，データを一つ読

み込んだときの $\tilde{B}$ の変化は，ただ一つの列だけに起こる．それゆえ，この公式が適用可能である．

6.2 Markov の極限定理の利用

定常確率分布を求めるよく知られた方法に，Markov の極限定理 (例えば [16]) を用いるものがある．これは Π^m の各列が，m を大きくしたときに定常確率分布のなすベクトルに収束するというものである．従って，各成分が $1/|\partial T|$ である列ベクトル v に対して，Π を繰り返し掛けることで定常確率分布を近似出来る．この方法の計算コストは，必要な掛け算の回数を N としたとき，一般には $O(T|\partial T|^2)$ となる．しかし，tree model の場合，Π の各成分のうち 0 でないものは $k|\partial T|$ 個のみである (k はアルファベット数とする)．従って，一度の掛け算に要する計算量は $O(k|\partial T|)$ となる．

次に N の評価を行う．$|(\Pi^m)_{st} - \mu_s| < C\rho^m$ $(0 < \rho < 1)$ なる評価が知られている．ただし，ρ はある適当な M について

$$\rho = \Big(\frac{1}{2}\max_{s,t}\sum_{\sigma}|(\Pi^M)_{\sigma s} - (\Pi^M)_{\sigma t}|\Big)^{1/M}$$

と評価される．ここで，Π^M の各成分が ϵ 以上ならば，$(1/2)\max_{s,t}\sum_{\sigma}|(\Pi^M)_{\sigma s} - (\Pi^M)_{\sigma t}|$ は $1-2\epsilon$ 以下となる．今，p^x_s が全て ϵ 以上であると仮定しよう．すると，Π^d(d は T の深さ) の各成分は，ϵ^d 以上となる．よって，我々の問題の場合は $M = d$ とおいて，$\rho = (1-2\epsilon^d)^{1/d}$ となる．従って誤差を δ 以下にするための繰り返し回数は，$\rho^N = (1-2\epsilon^d)^{N/d} < \delta$ で求められ，ϵ^d が小さいという仮定のもとで，$N > (d/\epsilon^d)\log\delta$ を得る．従って，1 シンボルあたりの計算量は $O((dk|\partial T|/\epsilon^d)\log\delta)$ となる．入力に任意の列を許す場合，微分係数を最尤推定値ではなく Laplace 推定値において評価することにしても，ϵ は $1/n$ 程度まで小さくなりうる．これを代入すると n^d に比例することになり，大きな計算量を要する．しかし，これに関する部分は，固定した ϵ に対して $\bar{p}^x_s < \epsilon$ になったときには近似式を使わないという選択をすれば，定数とみなすことができ，実用上 $O(dk|\partial T|)$ の計算量のアルゴリズムとみなせる．

7 むすび

tree model に関する逐次予測問題を考察し，FSMX model でないある種の tree model は指数型とはならず，従って Jeffreys 混合が minimax regret を達成しないことを示した．これによって，regret を評価基準とする場合には，FSMX でない tree model については Jeffreys 混合を近似する努力は無駄であることが分かる．またこのことは，tree model について，“木 T が FSMX source を規定する $\Leftrightarrow$ $S(T)$ が指数型” という命題を示唆していると考えられる．これを示すことは今後の課題である．

謝辞: 有益な助言を与えて下さった二人の査読者に感謝する．

参考文献

[1] S. Amari & H. Nagaoka, *Methods of Information Geometry*, AMS & Oxford University Press, 2000.

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

[3] T. Kawabata & F. Willems, “A context tree weighting algorithm with an incremental context set,” *IEICE Trans. on Fundamentals*, vol. E83-A, No. 10, pp. 1898–1903, 2000.

[4] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C : The Art of Scientific Computing*, (*2nd edition*), Cambridge University Press, 1993.

[5] J. Rissanen, “A universal data compression system,” *IEEE trans. Inform. Theory,* Vol. 29, No. 5, pp. 656-664, 1983.

[6] Yu M. Shtarkov, “Universal sequential coding of single messages,” *Problems of Information Transmission*, vol. 23, pp. 3-17, July 1987.

[7] J. Takeuchi & A. R. Barron, “Asymptotically minimax regret by Bayes mixtures”, *Proc. of IEEE International Symposium on Inform. Theory*, 1998.

[8] J. Takeuchi & T. Kawabata, “Approximation of Bayes code for Markov sources”, *Proc. of 1995 IEEE International Symposium on Inform. Theory*, p. 391, 1995.

[9] J. Takeuchi, T. Kawabata, and A. R. Barron, “Properties of Jeffreys mixture for Markov sources”, *Proc. of the fourth Workshop on Information-Based Induction Sciences* (*IBIS2001)*, pp. 327-332, 2001.

[10] M. J. Weinberger, J. Rissanen and M. Feder, “A universal finite memory source”, *IEEE trans. Inform. Theory,* Vol. 41. No. 3, pp. 643-652, 1995.

[11] F. Willems, Y. Shtarkov and T. Tjalkens, “The context tree weighting method: basic properties,” *IEEE trans. Inform. Theory,* Vol. 41. No. 3, pp. 653-664, 1995.

[12] Q. Xie & A. R. Barron, “Asymptotic minimax regret for data compression, gambling and prediction”, *IEEE trans. Inform. Theory,* vol. 46, no. 2, pp. 431-445, 2000.

[13] H. Itoh & S. Amari, Geometry of information sources (in Japanese), *Proc. of SITA88*, pp. 57–60, 1988.

[14] J. Takeuchi & T. Kawabata, “On data compression algorithms by Bayes coding for Markov sources (in Japanese),” *Proc. of SITA94,* pp. 513–516, 1994.

[15] 韓, 小林, **情報と符号化の数理**, 岩波書店, 1994.

[16] 小倉久直, **続確率過程論**, コロナ社，1985.