On Asymptotic Exponential Family of Markov Sources and Exponential Family of Markov Kernels

Jun'ichi Takeuchi* Hiroshi Nagaoka†

May 26th, 2017; revised August 21st, 2017

1 Introduction

The notion of exponential family [2] was extended for families of Markov chains in an asymptotic manner by Ito & Amari [10], and some variants have been proposed. The notion of exponential family (e-family) of Markov kernels is the most important one among them. It was first introduced by Nakagawa and Kanaya [6] in the one-dimensional case. Then, Nagaoka [5] gave its established form, and Hayashi & Watanabe discussed it [4]. In [5], various concepts in information geometry including dually flat structure, are extended for families of Markov kernels. It is also remarkable that the notion does not need asymptotic setting.

On the other hand, notion of asymptotic exponential families of general stochastic processes was introduced in [8] as an extension of Ito and Amari's concept. Note that, only for asymptotic exponential families the Bayes mixture with Jeffreys prior asymptotically achieves the stochastic complexity of Rissanen [7], which is the most important notion in the minimum description length principle [3],

In this report, we show that both notions of exponential families of Markov sources are equivalent to each other. It means that the form of e-family of Markov kernels is unique to enjoy the same asymptotic properties of e-families as i.i.d. case.

2 Exponential Family of Markov Kernels

We review the notion of exponential family of Markov kernels, following [5]. Let $\mathcal{X}$ be a finite set and let $\mathcal{E}$ be a subset of $\mathcal{X}^2$. Assume that $(\mathcal{X}, \mathcal{E})$ is a strongly connected directed graph, i.e. for all $(x, y) \in \mathcal{X}^2$, there is a path from x to y. Let $\mathcal{W}(\mathcal{X}, \mathcal{E})$ denote the family of

* Kyushu University, Fukuoka, Japan
† The University of Elector-Communications, Tokyo, Japan

transition probability matrices as

$$\mathcal{W}(\mathcal{X},\mathcal{E}) = \Big\{ w : \forall (x,y) \in \mathcal{E},\ w_{xy} > 0,\ \sum_z w_{xz} = 1,\ \text{and}\ \ \forall (x,y) \in \mathcal{X}^2 \setminus \mathcal{E}, w_{xy} = 0 \Big\}.$$

We also use notation $w(y|x)$ to stand for w_{xy}.

Assume that the Markov kernels w_{xy} are parametrized by a parameter $\theta \in \mathbb{R}^m$ as $w_{xy} = w_\theta(y|x)$. If $w_\theta(y|x)$ is in a form of

$$w_\theta(y|x) = \begin{cases} \exp\big(C(xy) + \theta \cdot \boldsymbol{F}(xy) + K_\theta^+(xy) - K_\theta^-(xy) + \psi(\theta)\big), & \text{if } xy \in \mathcal{E}, \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

then the set $\mathcal{M} = \{w_\theta\}$ is referred to as an exponential family of Markov kernels, where $\boldsymbol{F}(xy) = (F_1(xy), ..., F_m(xy)) \in \mathbb{R}^m$, $K_\theta^+(xy) = K_\theta(y)$, $K_\theta^-(xy) = K_\theta(x)$, for each $(x,y) \in \mathcal{X}^2$, $K_\theta \in \mathbb{R}^{\mathcal{X}}$, and '$\cdot$' represents inner product.

Similarly as the i.i.d. case, an e-family of Markov kernels can be realized as the normalization of an affine space [5]. (See Section 2.6 of [1] for the i.i.d. case.) Now we introduce a normalization function $\Phi : \mathbb{R}^{\mathcal{E}} \to \mathcal{W}(\mathcal{X},\mathcal{E})$ by

$$\Phi(F) = \begin{cases} \exp(F + K^+ - K^- - c), & \text{over } \mathcal{E}, \\ 0, & \text{over } \mathcal{X}^2 \setminus \mathcal{E} \end{cases}$$

where $c \in \mathbb{R}$, $K \in \mathbb{R}^{\mathcal{X}}$, $K^+(xy) = K(y)$, and $K^-(xy) = K(x)$. This is well-defined, because the following argument is possible. Note that c and K, which make $\Phi(F)$ be an element of $\mathcal{W}(\mathcal{X},\mathcal{E})$, essentially uniquely exist for each $F \in \mathbb{R}^{\mathcal{E}}$, as discussed in [5]. This is shown as follows. Let $A = \exp(F)$ for $xy \in \mathcal{E}$, $A = 0$ for $xy \in \mathcal{X}^2 \setminus \mathcal{E}$, and $\nu = \exp(K)$. Then

$$\Phi(F)(y|x) = \frac{A(xy)\nu(y)}{\nu(x)e^c}. \tag{2}$$

To make $\Phi(F) \in \mathcal{W}(\mathcal{X},\mathcal{E})$, it is necessary and sufficient that $\sum_y \Phi(F)(y|x) = 1$ for all $x \in \mathcal{X}$, which is

$$\forall x,\ \sum_y A(xy)\nu(y) = e^c \nu(x). \tag{3}$$

It is clear that e^c is the Perron-Frobenius eigenvalue of the irreducible non-negative matrix $A(xy)$ and ν is its eigenvector, which uniquely exist. Define a linear subspace $\hat{\mathcal{K}}$ of $\mathbb{R}^{\mathcal{E}}$

$$\hat{\mathcal{K}} = \{\hat{K} \in \mathbb{R}^{\mathcal{E}} : \exists K \in \mathbb{R}^{\mathcal{X}}, \forall xy \in \mathcal{E},\ \hat{K}(xy) = K^+(y) - K^-(x)\}.$$

For $\mathcal{M} \subset \mathcal{W}(\mathcal{X},\mathcal{E})$, the inverse image of $\mathcal{M}$ by Φ is denoted as

$$\Phi^{-1}(\mathcal{M}) = \{\log w + \hat{K} - c : w \in \mathcal{M}, \hat{K} \in \hat{\mathcal{K}}, c \in \mathbb{R}\} \subset \mathbb{R}^{\mathcal{E}}.$$

Note that $\log w$ is an element of $\mathbb{R}^{\mathcal{E}}$.

Now, we show the following two lemmas characterizing the e-family of Markov kernels, which are stated in [5].

Lemma 1 Assume that $\mathcal{F}$ is an affine subspace of $\mathbb{R}^{\mathcal{E}}$. Then $\Phi(\mathcal{F})$ is an e-family of Markov kernels.

Proof: If $\mathcal{F}$ is an affine subspace of $\mathbb{R}^{\mathcal{E}}$, we can denote its element as $F_\theta = \theta \cdot \boldsymbol{F} + F_0$, where $\theta \in \mathbb{R}^m$, $\boldsymbol{F} = (F_1, ..., F_m)$, and $F_i \in \mathbb{R}^{\mathcal{E}}$ for $i = 0, 1, ..., m$. Then, $\Phi(F_\theta) = \exp(F_0 + \theta \cdot \boldsymbol{F} + K_\theta^+ - K_\theta^- - c(\theta))$ over $\mathcal{E}$. Hence $\Phi(\mathcal{F})$ is an e-family. *This completes the proof.*

Lemma 2 A family of Markov kernels $\mathcal{M} \subset \mathcal{W}(\mathcal{X}, \mathcal{E})$ is an e-family, iff $\Phi^{-1}(\mathcal{M})$ is an affine subspace of the linear space $\mathbb{R}^{\mathcal{E}}$.

Proof: When $\Phi^{-1}(\mathcal{M})$ is an affine space, then $\mathcal{M}$ is an e-family by Lemma 1.

Next, assume $\mathcal{M}$ is an e-family. We can denote its element as $w_\theta = \exp(F_0 + \theta \cdot \boldsymbol{F} + \hat{K}_\theta - \psi(\theta))\}$ over $\mathcal{E}$ by definition. Hence we have

$$\begin{aligned}\Phi^{-1}(\mathcal{M}) &= \{F_0 + \theta \cdot \boldsymbol{F} + \hat{K}_\theta - \psi(\theta) + \hat{K} - c : \theta \in \mathbb{R}^m, \hat{K} \in \hat{\mathcal{K}}, c \in \mathbb{R}\} \\ &= \{F_0 + \theta \cdot \boldsymbol{F} + \hat{K} - c : \theta \in \mathbb{R}^m, \hat{K} \in \hat{\mathcal{K}}, c \in \mathbb{R}\},\end{aligned}$$

which is an affine space. *This completes the proof.*

3 Asymptotic Exponential Family

We state the definition of an *asymptotic exponential family,* which is a refinement of the one given in [8, 9]. Let x_m^n denote a string $x_m x_{m+1} ... x_n \in \mathcal{X}^{n-m+1}$ $(m \leq n)$ and x^n a string x_1^n. For a parametric model $S = \{p_\theta(x^n | x_{1-k}^0) : \theta \in \Theta \subset \mathbb{R}^m\}$, assume that the probability density function is written as

$$p_\theta(x^n | x_{1-k}^0) = \exp\Big(n\big(\theta \cdot \boldsymbol{V}(x_{1-k}^n) + C(x_{1-k}^n) - \psi(\theta)\big) + U_\theta(x_{1-k}^n)\Big), \tag{4}$$

where $\boldsymbol{V} = (V_1, ...V_m)$, $V_i, U_\theta \in \mathbb{R}^{\mathcal{X}^{n+1}}$, for $n = 0, 1, \cdots$. Here, we suppose $U_\theta(x_{1-k}^0) = 0$ and $p_\theta(x_0 | x_{1-k}^0) = 1$. Assuming an appropriate distribution of x_{k-1}^0, define

$$\eta = \eta^{(n)} = E_\theta \boldsymbol{V}(x_{1-k}^n).$$

We expect that η has similar properties of expectation parameter of ordinary exponential families. Further define

$$\begin{aligned}H_n &= \mathrm{conv}(\{\boldsymbol{V}(x_{1-k}^n) : x_{1-k}^n \in \mathcal{X}^{n+k}\}), \\ H &= \bigcup_{k=1}^{\infty} \bigcap_{t=k}^{\infty} H_t.\end{aligned}$$

Then, $\eta^{(n)} \in H_n$ holds.

Assume that, for every compact set L interior to H°,

$$\lim_{n\to\infty}\max_{ij}\max_{\theta:\eta\in L}\max_{x_{1-k}^n:\boldsymbol{V}\in L}\frac{1}{n}\Big|\frac{\partial^2 U_\theta(x_{1-k}^n)}{\partial\theta_i\partial\theta_j}\Big|=0. \tag{5}$$

Then, we call S an asymptotic exponential family.

Remark: For i.i.d. case, Θ is usually defined as the set of θ for which $\psi(\theta)$ is finite, and is known to be a convex set. In particular for a finite $\mathcal{X}$, we can assume that Θ is $\mathbb{R}^m$.

Under the condition (5), twice integrating

$$\frac{1}{n}\frac{\partial^2 U_\theta(x_{1-k}^n)}{\partial\theta_i\partial\theta_j}$$

with respect to θ over L for each x_{1-k}^n, we have

$$\lim_{n\to\infty}\max_{i}\max_{\theta:\eta\in L}\max_{x_{1-k}^n:\boldsymbol{V}\in L}\Big|\frac{1}{n}\frac{\partial U_\theta(x_{1-k}^n)}{\partial\theta_i}-V_i'(x_{1-k}^n)\Big|=0,$$
$$\lim_{n\to\infty}\max_{\theta:\eta\in L}\max_{x_{1-k}^n:\boldsymbol{V}\in L}\Big|\frac{1}{n}U_\theta(x_{1-k}^n)-(\theta\cdot\boldsymbol{V}'(x_{1-k}^n)+C'(x_{1-k}^n))\Big|=0,$$

where $\boldsymbol{V}'(x_{1-k}^n)$ and $C'(x_{1-k}^n)$ are certain functions of x_{1-k}^n. Letting

$$\begin{aligned}\tilde{U}_\theta(x_{1-k}^n)&=nU_\theta(x_{1-k}^n)-n(\theta\cdot\boldsymbol{V}'(x_{1-k}^n)+C'(x_{1-k}^n))\\ \tilde{\boldsymbol{V}}(x_{1-k}^n)&=\boldsymbol{V}(x_{1-k}^n)+\boldsymbol{V}'(x_{1-k}^n)\\ \tilde{C}(x_{1-k}^n)&=C(x_{1-k}^n)+C'(x_{1-k}^n),\end{aligned}$$

we have

$$p_\theta(x^n|x_{1-k}^0)=\exp\Big(n\big(\theta\cdot\tilde{\boldsymbol{V}}(x_{1-k}^n)+\tilde{C}(x_{1-k}^n)-\psi(\theta)\big)+\tilde{U}_\theta(x_{1-k}^n)\Big), \tag{6}$$

for which

$$\lim_{n\to\infty}\max_{\theta:\eta\in L}\max_{x_{1-k}^n:\boldsymbol{V}\in L}\frac{1}{n}|\tilde{U}_\theta(x_{1-k}^n)|=0, \tag{7}$$

$$\lim_{n\to\infty}\max_{i}\max_{\theta:\eta\in L}\max_{x_{1-k}^n:\boldsymbol{V}\in L}\frac{1}{n}\Big|\frac{\partial\tilde{U}_\theta(x_{1-k}^n)}{\partial\theta_i}\Big|=0. \tag{8}$$

$$\lim_{n\to\infty}\max_{ij}\max_{\theta:\eta\in L}\max_{x_{1-k}^n:\boldsymbol{V}\in L}\frac{1}{n}\Big|\frac{\partial^2\tilde{U}_\theta(x_{1-k}^n)}{\partial\theta_i\partial\theta_j}\Big|=0$$

hold. That is, by rearranging the terms in (4), we can assume (7) and (8), which does not break generality. Hence, hereafter we assume

$$\lim_{n\to\infty}\max_{\theta:\eta\in L}\max_{x_{1-k}^n:\boldsymbol{V}\in L}U_\theta(x_{k-1}^n)=0 \tag{9}$$

$$\lim_{n\to\infty}\max_{i}\max_{\theta:\eta\in L}\max_{x_{1-k}^n:\boldsymbol{V}\in L}\frac{1}{n}\Big|\frac{\partial U_\theta(x_{1-k}^n)}{\partial\theta_i}\Big|=0. \tag{10}$$

in the definition of asymptotic exponential families.

Some Property of Asymptotic Exponential Families

We can show the following for asymptotic e-families:

$$\lim_{n\to\infty} \max_{x_{1-k}^n : \boldsymbol{V}\in L} |\hat{J}(\hat{\theta}, x_{1-k}^n) - J^n(\hat{\theta})| = 0, \tag{11}$$

where $\hat{J}$ and J^n is the empirical Fisher information and Fisher information defined as

$$\hat{J}_{ij}(\hat{\theta}, x_{1-k}^n) = \frac{-1}{n} \frac{\partial^2 \log p_\theta(x^n|x_{1-k}^0)}{\partial\theta_i \partial\theta_j},$$
$$J^n(\theta) = E_\theta \hat{J}(\theta, x_{1-k}^n).$$

For the expectation, we assume an appropriate distribution $p_\theta^{(0)}$ for the initial string x_{1-k}^0. Note that $\hat{J}(\hat{\theta}, x_{1-k}^n) - J^n(\hat{\theta})$ is essentially the exponential curvature of the model [1].

4 Equivalency of E-Family of Markov kernels and Asymptotic E-Family

Here we assume $k = 1$ to consider the 1st order Markov models with alphabet $\mathcal{X} = \{1, 2, ..., D\}$. For $S = \{p_\theta(x^n|x_0) : \theta \in \mathbb{R}^m\}$, assume $p_\theta(x^n|x_0)$ is defined by Markov kernels in $\mathcal{M} = \{w_\theta(y|x) : \theta \in \mathbb{R}^m\}$:

$$p_\theta(x^n|x_0) = \prod_{t=1}^{n} w_\theta(x_t|x_{t-1}). \tag{12}$$

Further, in taking expectation, we assume that the initial probability distribution is given as the stationary distribution determined by $w_\theta(y|x)$.

The following is our main result.

Theorem 1 A Markov model $S = \{p_\theta(x^n|x_0)\}$ is an asymptotic exponential family, iff the $\mathcal{M} = \{w_\theta(y|x)\}$ is an exponential family of Markov kernels.

We prove it based on a series of lemmas.

As preliminaries, we introduce some terminology for strings. For a string x^n, if $x_t = x_0$ and $x_{t-1}x_t \in \mathcal{E}$ for all $t : 2 \le t \le n$, then x^n is called a *loop* for $\mathcal{E}$, or simply a loop. For a path x^n, if $\{x_{t-1}x_t : 2 \le t \le n\} = \mathcal{E}$, then x^n is said to be *complete* for $\mathcal{E}$, or simply complete. Given two paths x^α and y^β with $x_\alpha = y_1$, we can connect them to obtain $x^{\alpha-1}y^\beta = x^\alpha y_2^\beta$, for which we denote as

$$x^\alpha \bullet y^\beta = x^{\alpha-1}y^\beta (= x^\alpha y_2^\beta).$$

Using this symbol, the infinite repetition of a loop x^n can be denoted as

$$x^n \bullet x^n \bullet x^n \bullet \cdots = x^{n-1}x^{n-1}x^{n-1}\ldots.$$

For each $xy \in \mathcal{E}$, define

$$\tau(xy|x_0^n) = \frac{\#\{t : x_{t-1}x_t = xy,\ 1 \le t \le n\}}{n}.$$

This is called a Markov type of a string x_0^n. Since $\sum_{xy} \tau(xy|x_0^n) = 1$, a type belongs to the probability simplex of order $m^2 - 1$.

We give some lemmas to show the main result.

Lemma 3 For $S = \{p_\theta(x^n|x_0)\}$, assume that p_θ is written as (4) with $k = 1$. Let $\{\bar{\theta}^{(j)}\}$ be the normal basis of $\mathbb{R}^m$, that is, $\bar{\theta}_i^{(j)} = \delta_{ij}$. Then, the following holds.

$$\log w_\theta(x_t|x_{t-1}) = \sum_i \theta_i \log \frac{w_{\bar{\theta}^{(i)}}(x_t|x_{t-1})}{w_0(x_t|x_{t-1})} + \log w_0(x_t|x_{t-1}) - \tilde{\psi}(\theta) + k_\theta(x_{t-1}x_t), \tag{13}$$

where k_θ is a certain element of $\mathbb{R}^{\mathcal{E}}$ for each θ, and

$$\tilde{\psi}(\theta) = \psi(\theta) - \sum_i \theta_i(\psi(\bar{\theta}^{(i)}) - \psi(0)) - \psi(0).$$

Proof: Let $f_\theta(xy) = \log w_\theta(y|x)$. Then from (4), we have

$$\sum_{t=1}^n f_\theta(x_{t-1}x_t) = \log \prod_{t=1}^n w_\theta(x_t|x_{t-1}) = n\theta \cdot \boldsymbol{V}(x_0^n) + nC(x_0^n) + U_\theta(x_0^n) - n\psi(\theta), \tag{14}$$

which is

$$n\theta \cdot \boldsymbol{V}(x_0^n) + nC(x_0^n) = \sum_{t=1}^n f_\theta(x_{t-1}x_t) - U_\theta(x_0^n) + n\psi(\theta).$$

Letting $\theta = 0$, we have

$$nC(x_0^n) = \sum_{t=1}^n f_0(x_{t-1}x_t) - U_0(x_0^n) + nc_0, \tag{15}$$

where $c_0 = \psi(0)$. Further, letting $\theta = \theta^{(i)}$ $(i = 1, ..., m)$, where $\theta_j^{(i)} = \delta_{ij}$, we have

$$n(V_i(x_0^n) + C(x_0^n)) = \sum_{t=1}^n f_{\theta^{(i)}}(x_{t-1}x_t) - U_{\theta^{(i)}}(x_0^n) + nc_i, \tag{16}$$

where c_i denotes $\psi(\theta^{(i)})$. Define $h_i = -U_{\theta^{(i)}} + U_0$. Then, recalling $f_\theta(xy) = \log w_\theta(y|x)$, from (15) and (16),

$$nV_i(x_0^n) = \sum_{t=1}^n \log \frac{w_{\bar{\theta}^{(i)}}(x_t|x_{t-1})}{w_0(x_t|x_{t-1})} + h_i(x_0^n) + n(c_i - c_0). \tag{17}$$

Plugging in (15) and (17) to (14), we have

$$w_\theta(x_t|x_{t-1}) = \log \frac{p_\theta(x^t|x_0)}{p_\theta(x^{t-1}|x_0)}$$
$$= \sum_i \theta_i \log \frac{w_{\bar{\theta}^{(i)}}(x_t|x_{t-1})}{w_0(x_t|x_{t-1})} + \log w_0(x_t|x_{t-1}) - \tilde{\psi}(\theta) + u_\theta(x_0^t) - u_\theta(x_0^{t-1}), \quad (18)$$

where $\tilde{\psi}(\theta) = \psi(\theta) - \sum_i \theta_i(c_i - c_0) - c_0$, and $u_\theta = \sum_i \theta_i h_i + U_\theta - U_0$. Since this is an identity, $u_\theta(x_0^t) - u_\theta(x_0^{t-1})$ is a function of $x_{t-1}x_t$. By letting $k_\theta(x_{t-1}, x_t)$ denote it, we have the claim of the Lemma. *The proof is completed.*

By Lemma 3, when the asymptotic e-family $p_\theta(x^t|x_0)$ is defined by Markov kernels, we have

$$V_i(x_0^n) = \frac{1}{n}\sum_{t=1}^n \log \frac{w_{\bar{\theta}^{(i)}}(x_t|x_{t-1})}{w_0(x_t|x_{t-1})} = \sum_{xy} \tau(xy|x_0^n) \log \frac{w_{\bar{\theta}^{(i)}}(y|x)}{w_0(y|x)}.$$

Note that this defines a linear mapping from the probability simplex $\Delta^{D^2-1} \subset \mathbb{R}^{D^2}$ to $\mathbb{R}^m$. Let g denote it:

$$\boldsymbol{V}(x_0^n) = g(\tau(\cdot|x_0^n)).$$

Then we can see $H_n = H = g(\Delta^{D^2-1})$.

Lemma 4 Assume that a Markov model $S = \{p_\theta(x^n|x_0)\}$ is an asymptotic exponential family with (9) and (10). Then, for every $\theta \in \mathbb{R}^m$, and for every sequence $\{x_t\}_{t\geq 0}$ which is the infinite repetition of a complete loop for $\mathcal{E}$, the following holds.

$$\lim_{n\to\infty} \frac{1}{n}|U_\theta(x_0^n)| = 0.$$

Proof: Let $\bar{x}_0^\alpha$ be a complete loop Suppose that $\{x_t\}_{t\geq 0}$ is the infinite repetition of a complete loop $\bar{x}_0^\alpha$. Then, we have

$$\forall xy \in \mathcal{E},\ \lim_{n\to\infty} \tau(xy|x_0^n) = \tau(xy|x_0^\alpha) > 0.$$

Since x_0^α is a complete loop, $\tau(\cdot|x_0^\alpha)$ is in the interior of Δ^{D^2-1} and $\boldsymbol{V}(x_0^\alpha) \in H^\circ$. Let L be a compact subset of H° such that $\boldsymbol{V}(x_0^\alpha) = g(\tau(\cdot|x_0^\alpha)) \in L^\circ$. Then, for all large n, $\boldsymbol{V}(x_0^n) \in L$ holds. Hence, by (9) we have obtained the claim of the Lemma. *The proof is completed.*

Lemma 5 If $\{U(x_0^n)\}_n$ satisfies

$$U(x_0^n) = \sum_{t=1}^n k(x_{t-1}, x_t), \quad (19)$$

$$\lim_{n\to\infty} \frac{1}{n}|U(x_0^n)| = 0 \quad (20)$$

hold for every sequence $\{x_t\}_{t\geq 0}$ which is the infinite repetition of a complete loop for $\mathcal{E}$. Then, the following holds.

$$\forall xy \in \mathcal{E},\ \ k(x,y) = \kappa(y) - \kappa(x).$$

Proof: Suppose that $\{x_t\}$ is the infinite repetition of a complete loop $\bar{x}_0^\alpha$. Then, $U(x_0^n) = o(n)$ holds by the assumption. It implies $U(x_0^\alpha) = 0$ for any complete loop x_0^α, since we have

$$U(x_0^{M\alpha}) = \sum_{N=0}^{M-1} \sum_{t=1+N\alpha}^{\alpha+N\alpha} k(x_{t-1}, x_t) = M \sum_{t=1}^{\alpha} k(x_{t-1}, x_t) = MU(x_0^\alpha).$$

Now, to each $(x, y) \in \mathcal{X}^2$, assigning a path from x to y, and let $\Pi(x, y)$ denote it. Suppose that $\Pi(z, y)$ with a fixed z is a complete path for all y. Then for any path x_0^n ($n \geq 0$), $\Pi(z, x_0) \bullet x_0^n \bullet \Pi(x_n, z)$ is a complete loop. In particular for $n = 1$, $\Pi(z, x_0) \bullet \Pi(x_0, z)$ is a complete loop. Hence

$$\forall x_0 \in \mathcal{X},\ U(\Pi(z, x_0) \bullet \Pi(x_0, z)) = 0,$$

which yields

$$\forall x_0 \in \mathcal{X},\ U(\Pi(z, x_0)) = -U(\Pi(x_0, z)). \tag{21}$$

For general $n \geq 1$, we have

$$\forall x_0^n \in \mathcal{X}^{n+1},\ U(\Pi(z, x_0) \bullet x_0^n \bullet \Pi(x_n, z)) = 0, \tag{22}$$

which yields

$$\forall x_0^n \in \mathcal{X}^{n+1},\ U(x_0^n) = -U(\Pi(z, x_0)) - U(\Pi(x_n, z)) = U(\Pi(z, x_n)) - U(\Pi(z, x_0)),$$

where we used (21) for the last equality. For the fixed z, defining

$$\forall x \in \mathcal{X},\ \kappa = U(\Pi(z, x))$$

we have

$$\sum_{t=1}^{n} k(x_{t-1}, x_t) = U(x_0^n) = \kappa(x_n) - \kappa(x_0),$$

which implies

$$\forall xy \in \mathcal{E},\ k(x, y) = \kappa(y) - \kappa(x).$$

The proof is completed.

Since Lemmas 3, 4 and 5, an asymptotic exponential family is an exponential family of Markov kernels. For the converse, we show thw following lemma.

Lemma 6 If a family of Markov kernels $\{w_\theta\}$ is an exponential family of Markov kernels, then the model $\{p_\theta(x^n|x_0)\}$ is an asymptotic exponential family.

Proof: By the assumption we have

$$\log w_\theta(y|x) = \theta \cdot \boldsymbol{F}(y|x) + C(xy) - \psi(\theta) + \kappa_\theta(y) - \kappa_\theta(x).$$

Hence

$$\log w_\theta(x^n|x_0) = \sum_{t=1}^{n} \big(\theta \cdot \boldsymbol{F}(x_{t-1}x_t) + C(x_{t-1}x_t) - \psi(\theta)\big) + \kappa_\theta(x_n) - \kappa_\theta(x_0).$$

Recall that $\exp(\kappa_\theta(x))$ is the Perron-Frobenuis eigenvector for the non-negative matrix

$$A_{xy} = \begin{cases} \exp(\theta \cdot \boldsymbol{F}(y|x) + C(xy)), & \text{for } xy \in \mathcal{E}, \\ 0, & \text{otherwise.} \end{cases}$$

Hence, $\exp(\kappa_\theta(x))$ is a rational function of positive entries of A, which are of class C^∞ over L and bounded from below by a positive number. Hence

$$\frac{\partial^2(\kappa_\theta(x_n) - \kappa_\theta(x_0))}{\partial\theta_i \partial\theta_j}$$

is bounded over L. *The proof is completed*

References

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

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

[3] P. Grünwald, *The minimum description length principle,* MIT Press, 2007.

[4] M. Hayashi and S. Watanabe, "Information Geometry Approach to Parameter Estimation in Markov Chains," *Annals of Statistics,* Vol. 44, No. 4, pp. 1495-1535, 2016.

[5] H. Nagaoka, "The exponential family of Markov chains and its information geometry," *Proc. of the 28th Symposium on Information Theory and its Applications* (SITA2005), 2005. Available at https://arxiv.org/abs/1701.06119

[6] K. Nakagawa and F. Kanaya, "On the converse theorem in statistical hypothesis testing for Markov chains," *IEEE Trans. Inform. Theory,* Vol. 39, No. 2, pp. 629-633, 1993.

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

[8] J. Takeuchi and A. R. Barron, "Asymptotically minimax regret by Bayes mixtures," *Proc. of the 1998 IEEE ISIT*, 1998.

[9] J. Takeuchi and T. Kawabata, "Exponential Curvature of Markov Models," *Proc. of 2007 IEEE ISIT International Symposium on Information Theory* , pp. 2891-2895, Nice, France, 2007.

[10] H. Ito and S. Amari, "Geometry of information sources (in Japanese)," *Proc. of SITA88*, pp. 57–60, 1988.