社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.
IT2014-40(2014-7)

Geometry of Markov Chains, Finite State Machines, and Tree Models

Jun'ichi TAKEUCHI†

† Faculty of Information Science and Electrical Engineering, Kyushu University
744 Motooka, Nishiku, Fukuoka-city, Fukuoka 819-0395, Japan

Abstract We discuss information geometrical aspects of models of Markov chains. It is known that the models of Markov chains defined by a strongly connected graph are exponential families in asymptotic sense. In this manuscript, we try to understand this fact in the view point of the embedding exponential curvature (e-curvature for short). In particular, we show that the e-curvature of the models of Markov chains defined by a strongly connected graph is of order $O(1/n)$, where n is the data size.

Key words exponential family, Markov model, exponential connection

1. Introduction

We discuss information geometrical aspects of models of Markov chains. It is known that the models of Markov chains defined by a strongly connected graph are exponential families in asymptotic sense [6], [8], [13]. In this manuscript, we try to understand this facts in the view point of the embedding exponential curvature.

In 1988, Itoh & Amari [13] showed that the family of kth order Markov chains is an exponential family in asymptotic sense by giving its canonical parameter concretely. In 2005, Nagaoka generalized it to the Markov model with a strongly connected graph which specifies non-zero entries of its transition matrix. He defines the notion of exponential family of Markov chains non in asymptotic form. In [6] (Hayashi & Watanabe) the essentially same definition was given.

On the other hand, the author has given discussions on this topic in the view point of the embedding curvature in the joint works with Tsutomu Kawabata and Andrew Barron [9]〜[11], but their discussion was not well organized. In this manuscript, we give clearer discussions and add some new view point about that the model of stochastic process is an exponential family. We also discuss the fact that the tree model [12] is an exponential family in asymptotic sense, if and only if it is specified by a finite state machine. This is a review of [10].

2. Exponential and Curved Exponential Families

Here we review the exponential family of i.i.d. probability densities. If the probability density function is given as

$$p(z|\theta) = \exp(\theta \cdot T(z) - \psi(\theta)), \tag{1}$$

then the family $\{p(\cdot|\theta) : \theta \in \Theta\}$ is referred to as an exponential family of probability densities [4] [3]. Here, $T(z)$ is a d-dimensional real valued random variable, θ a d-dimensional parameter, and $\theta \cdot T(z) = \sum_{i=1}^{d} \theta_i T_i(z)$. Let $\partial_i \stackrel{\rm def}{=} \partial/\partial\theta_i$. Define $\eta_i \stackrel{\rm def}{=} \partial_i \psi(\theta)$, then we have $\eta_i = E_\theta T_i(z)$. Further, $J_{ij}(\theta) = \partial_j \eta_i$ holds. The parameters η and θ are referred to as expectation parameter and canonical parameter, respectively. For the exponential family (1), the probability density function of the string $T(z^n) = T(z_1)T(z_2)...T(z_n)$ is given as

$$p(z^n|\theta) \stackrel{\rm def}{=} \prod_t p(z_t|\theta) = \exp(n(\theta \cdot \bar{T} - \psi(\theta))), \tag{2}$$

where $\bar{T} \stackrel{\rm def}{=} (1/n)\sum_t T(z_t)$. It means that the joint model $S^n = \{p(z^n|\theta) : \theta \in \Theta\}$ is an exponential family again. Since $\partial_i \log p(z^n|\theta) = n(\bar{T}_i - \eta_i)$ holds, the MLE of η given z^n, equals the sufficient statistics $\bar{T}$.

Next we review the notion of curved exponential families. Consider the model embedded in a $\bar{d}$-dimensional exponential family $\bar{S} = \{\bar{p}(\cdot|\theta) : \theta \in \Theta\}$, where θ is the canonical parameter. Let $u \in \mathcal{U}$ be a d-dimensional vector ($d < \bar{d}$) and $\phi : \mathcal{U} \to \Theta$ a class C^∞ function, provided the rank of its Jacobian is d over $\mathcal{U}$. Define $p(\cdot|u) = \bar{p}(\cdot|\phi(u))$ then, we refer to $S = \{p(\cdot|u) : u \in \mathcal{U}\}$ as a curved exponential family embedded in $\bar{S}$.

When S is an exponential family too, the Fisher information equals the empirical Fisher information at MLE. It can be confirmed as follows. First note the following.

$$\frac{\partial}{\partial u_a} \log p(z^n|u) = \sum_i \frac{\partial \phi_i(u)}{\partial u_a}(\bar{T}_i - \eta_i)$$

$$\hat{J}_{ab}(u, z^n) \stackrel{\rm def}{=} -\frac{1}{n}\frac{\partial^2}{\partial u_a \partial u_b} \log p(z^n|u) \tag{3}$$

This is a slightly corrected one of the article published in IEICE Technical Reports IT2014-40.

$$= -\sum_i \frac{\partial^2 \phi_i}{\partial u_a \partial u_b}(\bar{T}_i - \eta_i) + J_{ab}(u),$$

where $J_{ab}(u)$ and $\hat{J}_{ab}(u, z^n)$ are Fisher information and empirical Fisher information with respect to u, respectively. Since the log likelihood is maximized at $\hat{u}$, $\sum_i (\partial\phi_i/\partial u_a)(\bar{T}_i - \eta_i) = 0$ holds at $u = \hat{u}$. If S is an exponential family, $\phi(u)$ forms an affine subspace in Θ. Hence, $\partial^2\phi/\partial u_a \partial u_b$ belongs to the linear subspace spanned by $\{\partial\phi/\partial u_a\}_{a=1,\cdots,d}$. Therefore, the first term of the third side of (3) equals zero.

Conversely, it can be shown that when $J(\hat{\theta}) = \hat{J}(\hat{\theta}, z^n)$ holds for every $z^n \in Z^n$ for all n, then M is an exponential family. To understand it, we review the notion of the embedding exponential curvature in the next section.

3. Exponential Curvature

Following [2], we introduce the notion of embedding curvature of general smooth families with respect to the exponential-connection (α-connection with $\alpha = 1$), which we refer to as the e-curvature in this paper.

Let $S = \{p(x|u) : u \in \mathcal{U}\}$ and let $\mathcal{P} = \mathcal{P}(\mathcal{X})$ denote the totality of probability densities over $\mathcal{X}$.

Let $\mathcal{T}_{p(\cdot|u)}(S)$ be the linear space spanned by $\{\partial_a \log p(x|u)\}_a$, which is referred to as exponential representation (e-representation) of the tangent space of S at $p(\cdot|u)$. An element of a tangent space is referred to as a tangent vector.

To generalize tangent vectors, we employ the variation operator Δ, which is a generalization of the differential operators ∂_a's. The Δ is a linear mapping from $p \in \mathcal{P}$ to its variation Δp. The variation Δp is a real-valued function over $\mathcal{X}$. such that $\int \Delta p(x) dx = 0$. Note that the operator Δ admits the same calculation rule as for ∂_a's. In particular, $\Delta \log p(x) = \Delta p(x)/p(x)$ holds. Define for each $p \in \mathcal{P}$

$$\mathcal{T}_p(\mathcal{P}) = \{\Delta \log p : E_p \Delta \log p(x) = 0\}.$$

This is the e-representation of the tangent space of $\mathcal{P}$ at p. Note that $\mathcal{T}_p(\mathcal{P})$ is a linear space and $\mathcal{T}_p(S)$ is a linear subspace of $\mathcal{T}_p(\mathcal{P})$.

Let us introduce the Fisher metric for $\mathcal{P}$ by

$$J_{\Delta_1,\Delta_2}(p) = \langle \Delta_1, \Delta_2 \rangle_p = E_p \frac{\Delta_1 p(x)}{p(x)} \frac{\Delta_2 p(x)}{p(x)}. \tag{4}$$

Note that $J_{\partial_a,\partial_b}(p(\cdot|u))$ equals the Fisher information matrix $J_{ab}(u)$.

The exponential covariant derivative of ∂_b to the direction of ∂_b in $\mathcal{P}$ is defined by giving its e-representation as

$$(\overset{\mathrm{e}}{\nabla}{}^{(\mathcal{P})}_{\partial_a} \partial_b) \log p(x|u) = \partial_a \partial_b \log p(x|u) + J_{ab}(u),$$

where $J_{ab}(u)$ is the term to make its expectation to be zero and corresponds to the parallel displacement. (See [1].) Then, $(\overset{\mathrm{e}}{\nabla}{}^{(\mathcal{P})}_{\partial_a} \partial_b) \log p(\cdot|u)$ belongs to $\mathcal{T}_p(\mathcal{P})$, but it does not belong to $\mathcal{T}_p(S)$ in general. Here, its projection by the Fisher metric onto $\mathcal{T}_p(S)$ is the covariant derivative of ∂_b to the direction of ∂_a, which we denote by $(\overset{\mathrm{e}}{\nabla}{}^{(S)}_{\partial_a} \partial_b) \log p(\cdot|u)$. Here, coefficients of exponential connection Γ_{abc} (the α-connection with $\alpha = 1$) is obtained by

$$\begin{aligned}\Gamma^{(\mathrm{e})}_{abc}(u) &= \langle \overset{\mathrm{e}}{\nabla}{}^{(\mathcal{P})}_{\partial_a} \partial_b, \partial_c \rangle_{p(\cdot|u)} = E_{p(\cdot|u)}(\partial_a \partial_b l(x|u)) \partial_c l(x|u) \\ &= \Gamma^{(0)}_{abc}(u) - \frac{E_{p(\cdot|u)} \partial_a l(x|u) \partial_b l(x|u) \partial_c l(x|u)}{2}, \end{aligned} \tag{5}$$

where $\Gamma^{(0)}$ represents the Levi-Civita connection.

Now we define e-embedding curvature. It is a variation operator on $p(x|u)$ defined for every pair in $(\mathcal{T}_p(S))^2$ as

$$\overset{(\mathrm{e})}{H}(\partial_a, \partial_b) = \overset{\mathrm{e}}{\nabla}{}^{(\mathcal{P})}_{\partial_a} \partial_b - \overset{\mathrm{e}}{\nabla}{}^{(S)}_{\partial_a} \partial_b,$$

whose e-representation is

$$\overset{(\mathrm{e})}{H}(\partial_a, \partial_b) l = \partial_a \partial_b l + J_{ab}(u) - \sum_{d,e} \Gamma^{(\mathrm{e})}_{abd} (J(u))^{-1}_{de} \partial_e l.$$

If S is an exponential family, it is easy to see the e-curvature vanishes everywhere.

Now we can understand the relation between the e-curvature and $\hat{J}(\hat{u}, x^n) - J(\hat{u})$. Given a data string x^n, let $\hat{p}$ denote the empirical distribution over $\mathcal{X}$ and let $\hat{u}$ the MLE of u. Let $\Delta_{p(\cdot|\hat{u})\hat{p}}$ denote the variation operator which maps $p(\cdot|\hat{u})$ to $\hat{p}$. Noting $\hat{u}$ is the maximizer of $\log p(x^n|u) = nE_{\hat{p}} \log p(x|u)$ and $nE_{p(\cdot|\hat{u})} \log p(x|u)$, we have

$$\frac{\partial E_{\hat{p}} \log p(x|u)}{\partial u_a}\Big|_{u=\hat{u}} = \frac{\partial E_{p(\cdot|\hat{u})} \log p(x|u)}{\partial u_a}\Big|_{u=\hat{u}} = 0,$$

which implies

$$\begin{aligned} 0 &= \int (\hat{p}(x) - p(x|\hat{u}))(\partial_a l(x|\hat{u})) p(x|\hat{u}) dx \\ &= \langle \Delta_{p(\cdot|\hat{u})\hat{p}}, \partial_a \rangle_{p(\cdot|\hat{u})}. \end{aligned}$$

From this we have

$$\langle \overset{(\mathrm{e})}{H}(\partial_a, \partial_b), \Delta_{p(\cdot|\hat{u})\hat{p}} \rangle_{p(\cdot|\hat{u})} = -\hat{J}_{ab}(\hat{u}, x^n) + J_{ab}(\hat{u}), \tag{6}$$

which implies that $J(\hat{u}) - \hat{J}(\hat{u}, x^n)$ is the coefficient of e-curvature to the direction of $\hat{p} - p(\cdot|\hat{u})$.

If S is a curved exponential family in $\bar{S}$,$p(\cdot|u)$ an element of the exponential family S,

$$\begin{aligned}\log p(x^n|u) &= \log \bar{p}(x^n|\phi(u)) = n(\theta \cdot \bar{T} - \psi(\phi(u))) \\ &= nE_{\bar{p}(\cdot|\hat{\theta})} \log p(x|u), \end{aligned}$$

where $\hat{\theta}$ is the MLE of θ in $\bar{S}$, which corresponds to the point $\eta = \bar{T}$. This is the mixture projection of $\hat{p}$ onto S. Hence, we can replace $\Delta_{p(\cdot|\hat{u})\hat{p}}$ by the variation operator $\Delta_{p(\cdot|\hat{u})\bar{p}(\cdot|\hat{\theta})}$ in (6). This means that the e-curvature of S in $\mathcal{P}$ is essentially same as that in $\bar{S}$.

From (6), the statement that $\hat{J}(\hat{u}, x^n) = J(\hat{u})$ holds for all x^n implies that the components of the e-curvature at $\hat{u}$ to

This is a slightly corrected one of the article published in IEICE Technical Reports IT2014-40.

the particular directions are zero. Further note that $\hat{u}$ is in a discrete subset if $\mathcal{X}$ is discrete. Hence, the above proposition is not sufficient for the purpose of deciding whether a model is an exponential family or not.

For the above point, we can obtain the lemma below.

[*Lemma 1*] For a curved exponential family Sf embeeded in an exponential family $\bar{S}$, the exponential curvature at u^* is zero, if and only if for any $p \in \bar{S}$ such that $E_p \log p(x|u)$ is maximized at u^*, $E_p \hat{J}(u^*, x) = J(u^*)$ holds.

Proof: Necessity is obvious. Similarly to (6), we can show

$$\langle \overset{(e)}{H}(\partial_a, \partial_b), \Delta_{p(\cdot|u^*)p} \rangle_{p(\cdot|u^*)} = J_{ab}(u^*) - E_p \hat{J}_{ab}(u^*, x).$$

Since $\{\Delta_{p(\cdot|u^*)p} l(x|u^*)\}_p$ spans the orthogonal complement of $\mathcal{T}_{p(\cdot|u^*)}(S)$, the sufficiency follows. The claim is shown.

4. E-curvature of Models of Stochastic Processes

Hereafter, we assume $\mathcal{X} = \{0, 1, 2, .., D\}$. Let $S_n = \{p^n(x^n|u)\}$ be a family of probability mass functions over $\mathcal{X}^n$ of stochastic processes, i.e. we assume

$$p^n(x^n|u) = \sum_{x_{n+1}} p^{n+1}(x^n x_{n+1}|u).$$

Now we discuss geometry of S_n. In order to measure information per symbol, we use normalized form of Fisher metric at $p^n \in S_n$ as

$$J^n_{\Delta_1, \Delta_2}(p^n) = \frac{\langle \Delta_1, \Delta_2 \rangle_{p^n}}{n} = \frac{E_{p^n} \frac{\Delta_1 p^n(x^n)}{p^n(x^n)} \frac{\Delta_2 p^n(x^n)}{p^n(x^n)}}{n}. \quad (7)$$

Correspondingly, define the Fisher information of u as

$$\begin{aligned} J^n_{ab}(u) &= J^n_{\partial_a, \partial_b}(p^n(\cdot|u)) \\ &= n E_{p(\cdot|u)} \partial_a l_n \partial_b l_n = -E_{p(\cdot|u)} \partial_a \partial_b l_n, \end{aligned}$$

where $l_n = (1/n) \log p^n(x^n|u)$.

We consider the e-curvature of S_n relative to $\mathcal{P}_n = \mathcal{P}(\mathcal{X}^n)$ (multinomial model of alphabet $\mathcal{X}^n$). Here the point is that $\mathcal{P}(\mathcal{X}^n)$ is a $((1+D)^n - 1)$-dimensional exponential family. Hence, S_n is a curved exponential family in general. Then, define the e-curvature of S_n by the following. This is the e-representation.

$$\overset{(e)}{H}(\partial_a, \partial_b) l_n = (\overset{e\ (\mathcal{P}_n)}{\nabla_{\partial_a}} \partial_b - \overset{e\ (S_n)}{\nabla_{\partial_a}} \partial_b) l_n.$$

If this is zero everywhere in $\mathcal{U}$, then S_n is an exponential family. An example is the i.i.d. multinomial model. It is known that models of Markov chains are an exponential family in asymptotic sense.

Now, similarly to the i.i.d. case, we relate the e-curvature to the difference $\hat{J}^n(u, x^n) - J^n(u)$, where $\hat{J}^n_{ab}(u, x^n) = -\partial_a \partial_b l_n$ and $\hat{J}^n_{ab}(u, x^n) = -E_{p^n(\cdot|u)} \partial_a \partial_b l_n$.

Given x^n, let δ_{x^n} be the element of $\mathcal{P}(\mathcal{X}^n)$ such that $\delta_{x^n}(x^n) = 1$ ($\delta_{x^n}(y^n) = 0$ for $y^n \neq x^n$) and let $\hat{u}$ is the maximizer of $E_{\delta_{x^n}} \log p(x^n|u)$. Then, similarly to the verification of (6)

$$\frac{1}{n} \langle \overset{(e)}{H}(\partial_a, \partial_b), \Delta_{p(\cdot|\hat{u})\delta_{x^n}} \rangle_{p(\cdot|\hat{u})} = J^n_{ab}(\hat{u}) - \hat{J}^n_{ab}(\hat{u}, x^n), \quad (8)$$

Further, corresponding to Lemma 1, we have the following Lemma.

[*Lemma 2*] Let p^n be an arbitrary element of $\mathcal{P}_n$. Then, for a family S_n, the following holds, where u^* is the maximizer of $E_{p^n} \log p^n(x^n|u)$.

$$\frac{1}{n} \langle \overset{(e)}{H}(\partial_a, \partial_b), \Delta_{p(\cdot|u^*)p^n} \rangle_{p(\cdot|u^*)} = J^n_{ab}(u^*) - E_{p^n} \hat{J}^n_{ab}(u^*, x^n).$$

5. Asymptotic Exponential Families

The following is a definition of an *asymptotic exponential family,* which is a refinement of the one given in [10].

[*Definition 1*] (*Asymptotic Exponential Family*) For a parametric model $S_n = \{p^n(x^n|\theta) : \theta \in \Theta\}$ over $\mathcal{X}^n$, assume that the probability mass function is written as

$$p^n(x^n|\theta) = \exp(n(\theta \cdot T(x^n) - \psi(\theta)) + U(x^n|\theta)),$$

where T and U are certain functions of x^n ($n = 1, 2, \cdots$). If for every compact set K interior to Θ,

$$\max_{i,j,x^n} \max_{\theta \in K} \frac{1}{n} \Big| \frac{\partial^2 U(x^n|\theta)}{\partial \theta_i \partial \theta_j} \Big| \to 0 \ (n \to \infty)$$

holds, then we refer to S_n as an asymptotic exponential family.

Remark1: The (ordinary) exponential family is an asymptotic exponential family. Hence, when we say "a model is not an asymptotic exponential family," it implies that the model is not an exponential family.

Remark2: This definition is equivalent to the definition of the exponential family of Markov sources given by Nagaoka [8] and employed by [6].

We have the following Lemma.

[*Lemma 3*] For an asymptotic exponential family $S_n = \{p^n(x^n|\theta) : \theta \in \Theta\}$, for every $K \subset \Theta^\circ$,

$$\lim_{n \to \infty} \max_{\theta \in K} \max_{a,b} | \overset{(e)}{H}(\partial_a, \partial_b) l_n | = 0.$$

The proof is almost trivial by definition of S_n.

6. Markov models

Here, we show that simple Markov models are an asymptotic exponential family based on *Lemma* 2. Following [8], we define models of irreducible Markov chains. Let $\mathcal{E} \subseteq \mathcal{X}^2$ be a strongly connected directed graph, i.e. for every $(x, y) \in \mathcal{X}^2$, there is a path from x to y. Define a range of transition probability matrices $w = (w_{y|x})_{xy \in \mathcal{X}^2}$

This is a slightly corrected one of the article published in IEICE Technical Reports IT2014-40.

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

Here assume that $w_{0|x}$'s are dependent variables. Define a probability mass function as

$$p^n(x^n|w) = \mu_{x_1}(w) \prod_{t=1}^{n-1} w_{x_{t+1}|x_t},$$

where $\mu_x(w)$ is the stationary probability determined by w:

$$\mu_y(w) = \sum_x w_{y|x}\mu_x(w).$$

Define the Markov model as $S_n = \{p^n(x^n|w) : w \in \mathcal{E}\}$.

Let $\mathcal{Y}_n \subset \mathcal{X}^n$ be the set of the strings x^n such that $x_t x_{t+1} \in \mathcal{E}$ for all $t : 1 \leqq t < n$. For every $x^n \in \mathcal{Y}_n$, define type of x^n as

$$\tau_{xy} = \tau_{xy}(x^n) = \#\{t : t \in [1, n-1],\ xy = x_t x_{t+1}\}$$

for every $xy \in \mathcal{E}$. We also use the notation $\tau_x = \sum_y \tau_{xy}$. Then, we have

$$p^n(x^n|w) = \mu_{x_1}(w) \prod_{(x,y)\in\mathcal{E}} (w_{y|x})^{\tau_{xy}}.$$

Define $\bar{w}_{y|x} = \tau_{y|x}/\tau_x$. then $\bar{w} \in \mathcal{W}$ and $\hat{w}_{y|x} - \bar{w}_{y|x} = O(1/n)$ holds. Using this notation, the empirical Fisher information with respect to $w_{x'|x}$ and $w_{y'|y}$ is given as

$$\begin{aligned} &\hat{J}^n_{xx',yy'}(w, x^n) \\ &= \frac{\delta_{xy}}{n-1}\Big(\frac{\delta_{x'y'}\tau_{xx'}}{(w_{x'|x})^2} + \frac{\tau_{x0}}{(w_{0|x})^2}\Big) + O\Big(\frac{1}{n}\Big) \qquad (9) \\ &= \frac{\delta_{xy}\tau_x}{n-1}\Big(\frac{\delta_{x'y'}\hat{w}_{x'|x}}{(w_{x'|x})^2} + \frac{\hat{w}_{0|x}}{(w_{0|x})^2}\Big) + O\Big(\frac{1}{n}\Big), \end{aligned}$$

where δ_{xy} is Kronecker's delta. Also, Fisher information is

$$J^n_{xx',yy'}(w) = \delta_{xy}\mu_x(w)\Big(\frac{\delta_{x'y'}}{w_{x'|x}} + \frac{1}{w_{0|x}}\Big).$$

Hence, we have

$$\begin{aligned} &\hat{J}^n_{xx',yy'}(\hat{w}, x^n) - J^n_{xx',yy'}(\hat{w}) \\ &= \delta_{xy}\Big(\frac{\tau_x}{n-1} - \mu_x(\hat{w})\Big)\Big(\frac{\delta_{x'y'}}{\hat{w}_{x'|x}} + \frac{1}{\hat{w}_{0|x}}\Big) + O(1/n). \end{aligned}$$

Here, we can show

$$\Big|\frac{\tau_x}{n-1} - \mu_x(\hat{w})\Big| = O\Big(\frac{1}{n}\Big) \qquad (10)$$

as follows. Note that

$$\forall x \in \mathcal{X},\ \sum_{y:xy\in\mathcal{E}} \tau_{xy}(x^n) = \sum_{y:yx\in\mathcal{E}} \tau_{yx}(x^n) + \alpha \qquad (11)$$

holds for all $x^n \in \mathcal{Y}_n$, where $\alpha = -1$, 0, or 1. Then,

$$\begin{aligned} \tau_x &= \sum_y \tau_{xy} = \sum_y \tau_{yx} + \alpha = \sum_y \frac{\tau_{yx}}{\tau_y}\tau_y + \alpha \\ &= \sum_y \bar{w}_{x|y}\tau_y + \alpha = \sum_y \hat{w}_{x|y}\tau_y + \alpha + O(1/n), \end{aligned}$$

which implies

$$\frac{\tau_x}{n-1} = \sum_y \hat{w}_{x|y}\frac{\tau_y}{n-1} + O(1/n).$$

Noting that $\mathcal{E}$ is strongly connected, (10) is obtained by the Perron-Frobenius theorem.

From (10), $|\hat{J}^n(\hat{w}, x^n) - J^n(\hat{w})| = O(1/n)$. Here note that the order $O(1/n)$ term is uniform for all strings with $\hat{w} \in K$. This means that the component of e-curvature to the direction of $\Delta_{p^n(\cdot|\hat{w})\delta_{x^n}}$ is of order $O(1/n)$. Extending this fact, below we show that S_n is an asymptotic exponential family.

Let $\mathcal{Q}_n = \mathcal{P}(\mathcal{Y}_n)$. Then by *Lemma* 2,

$$\begin{aligned} &\frac{1}{n}\langle \overset{(\mathrm{e})}{H}(\partial_{xx'}, \partial_{yy'}), \Delta_{p(\cdot|u^*)q^n}\rangle_{p(\cdot|u^*)} \qquad (12) \\ &= J^n_{xx',yy'}(w^*) - E_{q^n}\hat{J}^n_{xx',yy'}(w^*, x^n) \end{aligned}$$

for every $q^n \in \mathcal{Q}_n$, where w^* is the w which maximizes $E_{q^n}\log p^n(x^n|w)$. Below we prove

$$E_{q^n}\hat{J}^n_{xx',yy'}(w^*, x^n) = J^n_{xx',yy'}(w^*) + O(1/n). \qquad (13)$$

Define

$$\tilde{\tau}_{xy} = \tilde{\tau}_{xy}(q^n) = E_{q^n}\tau_{xy}(x^n).$$

Then by (9), we have

$$\begin{aligned} &E_{q^n}\hat{J}^n_{xx',yy'}(w^*, x^n) \\ &= \frac{\delta_{xy}}{n-1}\Big(\frac{\delta_{x'y'}\tilde{\tau}_{xx'}}{(w^*_{x'|x})^2} + \frac{\tilde{\tau}_{x0}}{(w^*_{0|x})^2}\Big) + O\Big(\frac{1}{n}\Big) \\ &= \frac{\delta_{xy}\tilde{\tau}_x}{n-1}\Big(\frac{\delta_{x'y'}\tilde{w}_{x'|x}}{(w^*_{x'|x})^2} + \frac{\tilde{w}_{0|x}}{(w^*_{0|x})^2}\Big) + O\Big(\frac{1}{n}\Big), \qquad (14) \end{aligned}$$

where $\tilde{w}_{y|x} = \tilde{\tau}_{xy}/\tilde{\tau}_x$. Since

$$E_{q^n}\log p^n(x^n|w) = \sum_{xy\in\mathcal{E}} \tilde{\tau}_{xy}\log w_{y|x} + O(1/n),$$

we have $\tilde{w}_{y|x} = w^*_{y|x} + O(1/n)$. Plugging this into (14), we have

$$\begin{aligned} &E_{q^n}\hat{J}^n_{xx',yy'}(w^*, x^n) \qquad (15) \\ &= \frac{\delta_{xy}\tilde{\tau}_x}{n-1}\Big(\frac{\delta_{x'y'}}{w^*_{x'|x}} + \frac{1}{w^*_{0|x}}\Big) + O\Big(\frac{1}{n}\Big). \end{aligned}$$

From (11), we have

$$\forall x \in \mathcal{X},\ \sum_{y:xy\in\mathcal{E}} \tilde{\tau}_{xy} = \sum_{y:yx\in\mathcal{E}} \tilde{\tau}_{yx} + \alpha, \qquad (16)$$

Hence,

$$\frac{\tilde{\tau}_x}{n-1} = \mu_x(\tilde{w}) + O\Big(\frac{1}{n}\Big) = \mu_x(w^*) + O\Big(\frac{1}{n}\Big)$$

holds. Then by (15),

$$\begin{aligned} E_{q^n}\hat{J}^n_{xx',yy'}(w^*, x^n) &= \delta_{xy}\mu_x(w^*)\Big(\frac{\delta_{x'y'}}{w^*_{x'|x}} + \frac{1}{w^*_{0|x}}\Big) + O\Big(\frac{1}{n}\Big) \\ &= J^n_{xx',yy'}(w^*) + O\Big(\frac{1}{n}\Big). \qquad (17) \end{aligned}$$

This is a slightly corrected one of the article published in IEICE Technical Reports IT2014-40.

This is (13). Hence from (12)

$$\frac{1}{n}\langle \overset{(e)}{H}(\partial_{xx'}, \partial_{yy'}), \Delta_{p(\cdot|w^*)q^n} \rangle_{p(\cdot|w^*)} = O(1/n). \tag{18}$$

That is, the e-curvature per symbol is of order $O(1/n)$. We can prove that this bound holds uniformly for $w^* \in K$. Recall that S_n is a curved exponential family embedded in $\mathcal{P}(\mathcal{X}^n)$. Let $\theta = \phi(w)$ be the embedding function. Then, (18) means that $\partial_{xx'}\partial_{yy'}\phi(w)$ is approximately spanned by $\{\partial_{xx'}\phi(w)\}$ everywhere in $K \in \mathcal{W}$. In other words, when w is changing constantly, angular velocity of the tangent space is bounded by $O(1/n)$. Hence we can conclude that $\phi(w)$ converges to a plane in the canonical parameter's space. This implies S_n is an asymptotic exponential family.

7. Tree Models

We review the definition of tree source [12]. Let T be a finite subset of $\mathcal{X}^* \overset{\text{def}}{=} \{\lambda\} \cup \mathcal{X} \cup \mathcal{X}^2 \cup ...$, where λ denotes a null string. Assume that for all $s \in T$, any postfix of s belongs to T (e.g., the postfixes of x_1x_2 are x_1x_2, x_2 and λ). Such a set T is referred to as a context tree. For a context tree T, define ∂T as

$$\partial T \overset{\text{def}}{=} \{xs : x \in \mathcal{X}, s \in T\} \setminus T.$$

It can be shown that ∂T is a complete postfix set of $\mathcal{X}$, i.e. no element of ∂T is a postfix of another element and their length satisfies Kraft inequality with equality. For a string $s \in \mathcal{X}^*$, let $c(s)$ denote the element of ∂T which matches a postfix of s, if it exists. We refer to $c(s)$ as the context of s (or the state for s). Let $k \overset{\text{def}}{=} \max_{s\in\partial T} |s|$ ($|s|$ is length of s). When $|s| \geqq k$, $c(s)$ exists and is unique. An information source in which the probability of the successive character is defined for each context, is referred to as a tree source. If the set of contexts ∂T satisfies a condition that $c(sx)$ for any $s \in \partial T$ and any $x \in \mathcal{X}$ is determined (i.e. c defines a state transition function), then the tree source is referred to as an FSMX source. We give examples of context trees for an FSMX source and a non-FSMX tree source.

[*Example 1*] Assume $\mathcal{X} = \{0, 1\}$. Let $T_1 = \{\lambda, 1, 10, 0\}$, then we have $\partial T_1 = \{00, 11, 01, 110, 010\}$, which is a complete postfix set (See Figure 1). This tree defines state transition function. Hence the source defined with this tree is an FSMX source.

[*Example 2*] Removing '11' and '01' from ∂T_1 ('1' from T_1), we obtain T_2 in Figure 2. If '0' is generated at the context '1', we cannot determine whether the machine has transferred to the context '110' or '010'.

Let us introduce the tree model given a context tree T. Let $\ell = |\partial T|$ (the number of contexts), and $w_{x|s}$ denote the probability that x is generated at the context s. For each $s \in \partial T$, define a D-dimensional vector $w_s = (w_{1|s}, w_{2|s}, ..., w_{D|s})^t$. Let w denote a $D\ell$-dimensional vector $(w_{s_1}^t, w_{s_2}^t, , ..., w_{s_{|\partial T|}}^t)^t$. Define the range of parameter w as

$$\mathcal{W}_s = \{w_s : \forall x \in \mathcal{X},\ w_{x|s} \geqq 0,\ \sum_{x=1}^{D} w_{x|s} \leqq 1\}$$

and $\mathcal{W} = \mathcal{W}(T) \overset{\text{def}}{=} \prod_{s\in\partial T} \mathcal{W}_s$.

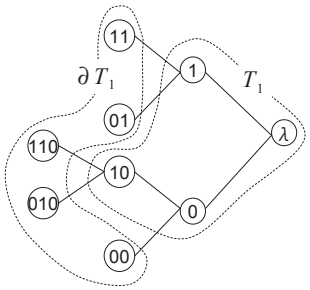

Figure 1 A Context Tree for an FSMX model

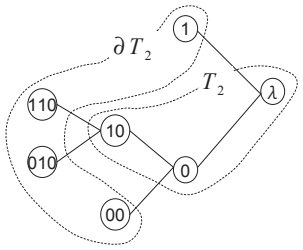

Figure 2 A Context Tree for a non-FSMX tree source

Let x_m^n denote a string $x_m x_{m+1}...x_n \in \mathcal{X}^{n-m+1}$ ($m \leqq n$) and x^n a string x_1^n. Assume that we have an initial string x_{-d+1}^0 in advance. Define the probability mass function for the sequence x^n as

$$p_T(x^n|x_{-d+1}^0, w) = \prod_{i=0}^{n-1} w_{x_{i+1}|c(x_{-d+1}^i)}.$$

and the tree model [5], [12] based on T as

$$S(T) \overset{\text{def}}{=} \{p_T(\cdot|\cdot, w) : w \in \mathcal{W}(T)\}. \tag{19}$$

When c defines a state transition function, then we refer to $S(T)$ as an FSMX model [5], [12]. The class of kth order Markov chains (case of $T = T^{(k)} \overset{\text{def}}{=} \bigcup_{i=0}^{k-1} \mathcal{X}^i$) is an example of FSMX model. For every $sx \in \partial T \times \mathcal{X}$, let τ'_{sx} denote the number of appearance of the pattern sx in x^n after the initial string x_{-d+1}^0 and let $\tau'_s = \sum_{x\in\mathcal{X}} \tau'_{sx}$. Then

$$p_T(x^n|x_{-d+1}^0, w) = \prod_{sx\in\partial T\times\mathcal{X}} (w_{x|s})^{\tau'_{sx}},$$

holds. By this, MLE of w is denoted as $\hat{w}_{x|s} = \tau'_{sx}/\tau'_s$, where we define $\hat{w} = \arg\max_{w\in\mathcal{W}} p_T(x^n|x_{-d+1}^0, w)$.

For an FSMX model, define $\sigma_i = c(x_{-d+1}^i)$, then $\sigma_i = c(\sigma_{i-1}x_i)$ and we have the sequence of contexts $\sigma_0^n = \sigma_0\sigma_1...\sigma_n$ induced by x_{-d+1}^n. Let τ_{st} for $st \in (\partial T)^2$ denote the number of appearance of the pattern st in σ_0^n and

This is a slightly corrected one of the article published in IEICE Technical Reports IT2014-40.

$\tau_s = \sum_t \tau_{st}$. Then, $\tau'_{sx} = \tau_{sc(sx)}$ and $\tau'_s = \tau_s$ hold. Here for each $st \in (\partial T)^2$ we define $w_{t|s} = w_{x|s}$, if there exists an $x \in \mathcal{X}$ such that $c(sx)$ equals t, $= 0$, otherwise. Then the probability mass function can be denoted as

$$p_T(x^n|\sigma_0, w) = \prod_{sx \in \partial T \times \mathcal{X}} (w_{x|s})^{\tau'_{sx}} = \prod_{st \in (\partial T)^2} (w_{t|s})^{\tau_{st}},$$

where we define $0^0 = 1$. Hence we can define $\hat{w}_{t|s} = \tau_{st}/\tau_s$. Then $\hat{w}_{c(sx)|s} = \hat{w}_{x|s}$ holds. This means that $S(T)$ can be transformed to a model of first order Markov chains, hence we can conclude $S(T)$ is an asymptotic exponential family.

Note that any tree model is a subspace of a certain FSMX model. In fact, a tree model $S(T)$ is a subspace of FSMX model $S(T^{(k)})$, where $k \stackrel{\text{def}}{=} \max_{s \in \partial T} |s|$. The following provides a non-trivial example.

[*Example 3*] Consider T_1 and T_2 in Examples 1 and 2. We have $S(T_2) = \{p(\cdot|\cdot, w) \in S(T_1) : w_{1|01} = w_{1|11}\}$. That is, $S(T_2)$ is a subspace of $S(T_1)$.

It was shown in [10] that the non-FSMX tree model is not an asymptotic exponential family. The following theorems holds provided $\mathcal{X}$ is binary. Hence in this section, we assume $\mathcal{X}$ is binary, but we think it is not difficult to extend these theorems to finite alphabet case.

[*Theorem 1*] (*Takeuchi & Kawabata 2007*) For a context tree T, $S(T)$ is an asymptotic exponential family, iff $S(T)$ is an FSMX model.

This theorem is proved by showing that $\hat{J}(\hat{w}) - J(\hat{w})$ for the non-FSMX tree model does not converge to zero.

Note that $S(T)$ is an FSMX model, iff sx does not belong to T for every $s \in \partial T$ and every $x \in \mathcal{X}$. Now, consider the tree $T = \{\lambda\} \cup \{s1 : s \in T_1\} \cup \{s0 : s \in T_2\}$, where T_1 and T_2 are context trees such that $T_1 \supset T^{(d)}$ and $T_2 \subset T^{(d-2)}$. Then for any $s0 \in \partial T$, its successor $s01$ belongs to T, hence $s01$'s context is not determined. This suggests that almost every tree model is not an asymptotic exponential family.

Acknowledgment

The author thanks Shun-ichi Amari, Hiroshi Nagaoka and Masahito Hayashi for fruitful discussions. This research was supported in part by JSPS KAKENHI Grant Number 24500018.

References

[1] S. Amari, *Differential-geometrical methods in statistics (2nd pr.)*, Lecture Notes in Statistics, Vol.28, Springer-Verlag, 1990.

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

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

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

[5] 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.

[6] M. Hayashi & S. Watanabe, "Information Geometry Approach to Parameter Estimation in Markov Chains," arXiv:1401.3814v1, 2014.

[7] A. Martìn, G. Seroussi, & M. J. Weinberger, "Linear time universal coding and time reversal of tree sources via FSM closure," *IEEE Trans. Inf. Theory,* vol. 50, no. 7, pp. 1442-1468, July 2004.

[8] 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.

[9] J. Takeuchi & A. R. Barron, "Asymptotically minimax regret by Bayes mixtures," *Proc. of 1998 IEEE ISIT,* p. 318, 1998.

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

[11] J. Takeuchi, T. Kawabata, & A. R. Barron, "Properties of Jeffreys mixture for Markov sources," *IEEE trans. Inform. Theory,* vol. 59, no. 1, pp. 438-457, 2013.

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

[13] H. Itoh & S. Amari, "Geometry of information sources (in Japanese)," *Proc. of the 11th Symp. on Inform. Theory and Its Apps.*, pp. 57–60, 1988.

This is a slightly corrected one of the article published in IEICE Technical Reports IT2014-40.