INFORMATION GEOMETRY APPROACH TO PARAMETER ESTIMATION IN MARKOV CHAINS

BY MASAHITO HAYASHI[*,†,1] AND SHUN WATANABE[‡,2]

Nagoya University, National University of Singapore†
and Tokyo University of Agriculture and Technology‡*

We consider the parameter estimation of Markov chain when the unknown transition matrix belongs to an exponential family of transition matrices. Then we show that the sample mean of the generator of the exponential family is an asymptotically efficient estimator. Further, we also define a curved exponential family of transition matrices. Using a transition matrix version of the Pythagorean theorem, we give an asymptotically efficient estimator for a curved exponential family.

1. Introduction. Information geometry established by Amari and Nagaoka [2] is an elegant method for statistical inference. This method provides us a very general approach to statistical parameter estimation. Under this framework, we easily find that the efficient estimator can be given with less calculation complexity for exponential families and a curved exponential families under the independent and identical distributed case. Therefore, we can expect a similar structure in the Markov chains.

The preceding studies [3, 4, 10, 13, 19, 20, 35, 36] introduced the concept of exponential families of transition matrices. However, in their definition, although the maximum likelihood estimator has the asymptotic efficiency, that is, attains the Cramér–Rao bound asymptotically, the maximum likelihood estimator is not necessarily calculated with less calculation complexity. That is, the maximum likelihood estimator has a complex form so that it requires long calculation time in their model. Further, it is quite difficult to calculate the Cramér–Rao bound even with the asymptotic first-order coefficient because these papers focused only on the limit of the inverse of the Fisher information. From a practical viewpoint, it is needed to calculate the asymptotic first-order coefficient. So, it is strongly required to resolve these two problems for the estimation of Markovian process,

that is, (1) to give an asymptotically efficient estimator with small calculation and (2) to derive a formula for the asymptotic Cramér–Rao bound with small calculation.

The purpose of this paper is giving the answers for these two problems. For this purpose, we notice another type of exponential family of transition matrices by Nakagawa and Kanaya [27] and Nagaoka [26]. They defined the Fisher information matrix in their sense. On the other hand, for the estimation of the probability distribution, the class of curved exponential families plays an important role as a wider class of distribution families than the class of exponential families. That is, when the unknown distribution belongs to a curved exponential family, the asymptotic efficient estimator can be treated in the information-geometrical framework. Therefore, to deal with these problems in a wider class of families of transition matrices, we introduce a curved exponential family of transition matrices as a subset of an exponential family of transition matrices in the sense of [26, 27]. Since any exponential family of transition matrices is a curved exponential family, the class of curved exponential families is a larger class of families of transition matrices than the class of exponential families. Especially, any smooth subset of transition matrices on a finite-size system forms a curved exponential family of transition matrices. Our purpose is resolving the above two problems for a curved exponential family as well as for an exponential family. Since any smooth parametric subfamily of transition matrices on a finite-size system forms a curved exponential family, our treatment for curved exponential families has a wide applicability for the estimation of Markovian process. This is reason why we adopted the definition of an exponential family by [26, 27].

First we show that, for an exponential family of transition matrices in the sense of [26, 27], an estimator of a simple form asymptotically attains the Cramér–Rao bound, which is given as the inverse of Fisher information matrix. That is, the estimator for the expectation parameter is asymptotically efficient and is written as the sample mean of $n + 1$-observations. Since it requires only a small amount of calculation, the problem (1) is resolved. Additionally, the problem (2) is also resolved for an exponential family of transition matrices because Fisher information matrix is computable.

To show the above items, we discuss the behavior of the sample mean of $n + 1$ observations. Indeed, while the existing papers [17, 31] derived the form of the asymptotic variance, this paper shows that the asymptotic variance can be written by using the second derivative of the potential function of the generated exponential family. Using this relation, we show that the sample mean asymptotically attains the Cramér–Rao bound for the expectation parameter.

Next, we define the Fisher information matrix for a curved exponential family with a computable form. Then, using a transition matrix version of the Pythagorean theorem, we give an asymptotically efficient estimator for a curved exponential family, in which, the estimator is given as a function of the above estimator in the larger exponential family. Since the asymptotic mean square error is the inverse

of the Fisher information matrix, the problems (1) and (2) are resolved jointly. In the above way, we resolve the problems that were unsolved in existing papers [3, 4, 10, 13, 19, 20, 35, 36]. Further, during this derivation, we also obtain a notable evaluation for variance of sample mean as a by product, which is summarized in Section 2.1.

For the above discussion, we need the description of an exponential family of transition matrices. Since the information geometrical structure for probability distributions plays important roles in several topics in information theory as well as statistics, it is better to describe the information geometry of transition matrices so that it can be easily applied to these topics. In fact, the authors applied it to finite-length evaluations of the tail probability, the error probability in simple hypothesis testing, source coding, channel coding and random number generation in Markov chain as well as the estimation error of parametric family of transition matrices [11, 12]. Thus, we revisit the exponential family of transition matrices [26, 27] in a manner consistent with the above purpose by using Bregman divergence [1, 6]. In particular, the relative Rényi entropy for transition matrices plays an important role in the finite-length analysis; we define the relative entropy for transition matrices so that it is a special case of the relative Rényi entropy, which is different from the definitions in the literatures [26, 27]. Although some of results in this paper have been already stated in [26] (without detailed proof), we restate those results and give proofs since the logical order of arguments are different from [26] and we want to keep the paper self-contained. In particular, although the paper [26] is written with differential geometrical terminologies, for example, Christoffel symbols, this paper is written only with terminologies of convex functions and linear algebra.

The remaining of this paper is organized as follows. Section 2 gives the brief summary of obtained results, which is crucial for understanding the structure of this paper. In Section 3, we define the relative entropy and the relative Rényi entropy between two transition matrices In Section 4, we revisit an exponential family of transition matrices and its properties. In Section 5, we focus on the joint distribution when a transition matrix is given as an element of a one-parameter exponential family and the input distribution is given as the stationary distribution. Then we characterize the quantities given in Sections 3 and 4 by using the joint distribution. In Section 6, we proceed to the $n + 1$ observation Markov process when the initial distribution is the stationary distribution. Then we show that the sample mean of the generator is an unbiased and asymptotically efficient estimator under a one-parameter exponential family. In Section 7, we proceed to the $n + 1$ observation Markov process when the initial distribution is a nonstationary distribution. We show a similar fact in this case. Section 8 extends a part of these results to the multiparameter case and the case of a curved exponential family. In the Appendix, we address the relations with existing results by Nakagawa and Kanaya [26], Nagaoka [26] and Natarajan [28].

2. Summary of results. Here, we prepare notation and definitions. For two given transition matrices W and W_Y over $\mathcal{X}$ and $\mathcal{Y}$, we define

$$W \times W_Y(x, y|x', y') := W(x|x')W_Y(y|y'),$$

$$W^{\times n}(x_n, x_{n-1}, \ldots, x_1|x') := W(x_n|x_{n-1})W(x_{n-1}|x_{n-2}) \cdots W(x_1|x'),$$

$$W^n(x|x') := \sum_{x_{n-1}, \ldots, x_1} W^{\times n}(x, x_{n-1}, \ldots, x_1|x').$$

For a given distribution P on $\mathcal{X}$ and a transition matrix V from $\mathcal{X}$ to $\mathcal{Y}$, we define $V \times P(y, x) := V(y|x)P(x)$ and $VP(y) := \sum_x V \times P(y, x)$.

A nonnegative matrix W is called *irreducible* when for each $x, x' \in \mathcal{X}$, there exists a natural number n such that $W^n(x|x') > 0$ [25]. An irreducible matrix W is called *ergodic* when there are no input x' and no integer n' such that $W^n(x'|x') = 0$ unless n is divisible by n' [25]. The irreducibility and the ergodicity depend only on the support $\mathcal{X}_W^2 := \{(x, x') \in \mathcal{X}^2 | W(x|x') > 0\}$ for a nonnegative matrix W over $\mathcal{X}$. Hence, we say that $\mathcal{X}_W^2$ is *irreducible* and *ergodic* when a nonnegative matrix W is irreducible and ergodic, respectively. Indeed, when a subset of $\mathcal{X}_W^2$ is irreducible and ergodic, the set $\mathcal{X}_W^2$ is also irreducible and ergodic, respectively. It is known that the output distribution $W^n P$ converges to the stationary distribution of W for a given ergodic transition matrix W [8, 17, 25]. Although the main result is asymptotic estimation for an exponential family and a curved exponential family, we also have additional results as Sections 2.1 and 2.2.

2.1. *Asymptotic behavior of sample mean.* Assume that the random variables $X^{n+1} := (X_{n+1}, \ldots, X_1)$ obey the Markov process with the irreducible and ergodic transition matrix $W(x|x')$. In this paper, for an arbitrary two-input function $g(x, x')$, we focus on the sample mean $S_n := \frac{g^n(X^{n+1})}{n}$ where $g^n(X^{n+1}) := \sum_{i=1}^n g(X_{i+1}, X_i)$. This is because a two-input function $g(x, x')$ is closely related to an exponential family of transition matrices. Indeed, the simple sample mean can be treated in this formulation by choosing $g(x, x')$ as x or x'. Since the function $g(x, x')$ can be chosen arbitrary, the following discussion can handle the sample mean of the hidden Markov process.

Then the expectation $\mathsf{E}[S_n]$ and the variance $\mathsf{V}[S_n]$ are characterized as follows. We denote the normalized Perron–Frobenius eigenvector of $W(x|x')$ by P_W and define the limiting expectation $\mathsf{E}[g(X, X')] := \sum_{x,x'} g(x, x')W(x|x')P_W(x')$. We denote the Perron–Frobenius eigenvalue of $W(x|x')e^{\theta g(x,x')}$ by λ_θ and define the cumulant generating function $\phi(\theta) := \log \lambda_\theta$. Then, when the transition matrix W is irreducible and ergodic, the relation

$$(2.1) \qquad\qquad \mathsf{E}[S_n] \to \mathsf{E}[g(X, X')]$$

is known. In Sections 6 and 7 of this paper, we show

$$(2.2) \qquad\qquad n\mathsf{V}[S_n] \to \frac{d^2\phi}{d\theta^2}(0)$$

while existing papers [17, 31] characterized the asymptotic variance by using the fundamental matrix (see [12], Section 6).

In particular, when the initial distribution is the stationary distribution P_W, we have $\mathsf{E}[S_n] = \mathsf{E}[g(X, X')]$. Then, in Section 6, using a constant C, we show that

$$(2.3) \qquad \frac{d^2\phi}{d\theta^2}(0)\left(1 - \frac{C}{\sqrt{n}}\right)^2 \le n\mathsf{V}[S_n] \le \frac{d^2\phi}{d\theta^2}(0)\left(1 + \frac{C}{\sqrt{n}}\right)^2$$

for the stationary case. The concrete form of C is also given in Section 6. This analysis is obtained via evaluations of Fisher information given in Sections 5, 6 and 7.

2.2. Cramér–Rao bound and asymptotically efficient estimator.

First, for simplicity, we summarize our obtained results for the one-parameter case while this paper addresses a multiparameter exponential family. In Section 4, for a given two-input function $g(x, x')$ and an irreducible and ergodic transition matrix W, we define the potential function $\phi(\theta)$ and exponential family of transition matrices $\{W_\theta\}$ with the generator $g(x, x')$. We also define its Fisher information matrix $\frac{d^2\phi}{d\theta^2}(\theta)$ and the expectation parameter $\eta(\theta) := \frac{d\phi}{d\theta}(\theta)$. Then we focus on the distribution family of Markov chains generated by the family of transition matrices $\{W_\theta\}$ with arbitrary initial distributions. We show that the Fisher information of the expectation parameter under the distribution family is asymptotically equal to $n\frac{d^2\phi}{d\theta^2}(\theta(\eta))^{-1} + o(n)$ even for the nonstationary case in Section 7. Then we show that the random variable S_n is the asymptotically efficient estimator, that is, the mean square error is $\frac{d^2\phi}{d\theta^2}(\theta(\eta))/n + o(1/n)$. In Section 6, we give more detailed analysis for the stationary case. To derive the results in Sections 6 and 7, we prepare evaluations of Fisher information in Section 5.

Now, we address the multiparameter case. In Section 4, we also define a multiparameter exponential family $W_{\vec{\theta}}$ of transition matrices, and show the Pythagorean theorem. Then we show the asymptotic efficiency of the sample mean in the multiparameter case in Sections 8.1 and 8.2. We also show that the set of all positive transition matrices on a finite-size system forms an exponential family in Example 1. Further, we define a curved exponential family of transition matrices, and give its asymptotically efficient estimator in Section 8.3. Since any smooth parametric family of transition matrices on a finite-size system forms a curved exponential family, this result has a wide applicability. These results require the technical preparations given in Sections 3, 4 and 5.

2.3. Relative entropy and relative Rényi entropy.

In this paper, given two transition matrices W and V, we define the relative entropy $D(W\|V)$ and the relative Rényi entropy $D_{1+s}(W\|V)$ in Section 3. In Section 8.3, the relative entropy $D(W\|V)$ plays a crucial role in our estimator in a curved exponential family. We also show that the Fisher information is given as the limits of the relative

entropy and the relative Rényi entropy, which plays important roles in the proof of the asymptotic efficiency of our estimator in a curved exponential family in Section 8.3. Also, as discussed in [12], the relative Rényi entropy $D_{1+s}(W\|V)$ plays a central role in simple hypothesis testing as well as the relative entropy $D(W\|V)$. Further, these information quantities play an central role in random number generation, data compression and channel coding [11]. In Section 3, we also give their properties that are useful in the above applications.

For these applications, we need to address the relative entropy $D(W\|V)$ and the relative Rényi entropy $D_{1+s}(W\|V)$ in a unified way. More precisely, the relative entropy $D(W\|V)$ is needed to be defined as the limit of the relative Rényi entropy $D_{1+s}(W\|V)$. Indeed, the existing paper [26] defined the relative entropy $D(W\|V)$ in a different way. However, the definition by [26] cannot yield the definition of the relative Rényi entropy in a unified way. Appendix A summarizes the detailed relation between the results in this part and existing results.

3. Relative entropy and relative Rényi entropy.   In this section, in order to investigate geometric structure for transition matrices, we define the relative entropy and the relative Rényi entropy. For this purpose, we prepare the following lemma, which is shown after Lemma 5.2.

LEMMA 3.1.   *Consider an irreducible transition matrix W over $\mathcal{X}$ and a real-valued function g on $\mathcal{X} \times \mathcal{X}$. Define $\phi(\theta)$ as the logarithm of the Perron–Frobenius eigenvalue of the matrix*:

$$(3.1) \qquad \overline{W}_\theta(x|x') := W(x|x')e^{\theta g(x,x')}.$$

Then the function $\phi(\theta)$ is convex. Further, the following conditions are equivalent:

(1) *No real-valued function f on $\mathcal{X}$ satisfies that $g(x,x') = f(x) - f(x') + c$ for any $(x,x') \in \mathcal{X}_W^2$ with a constant $c \in \mathbb{R}$.*

(2) *The function $\phi(\theta)$ is strictly convex, that is, $\frac{d^2\phi}{d\theta^2}(\theta) > 0$ for any θ.*

(3) *$\frac{d^2\phi}{d\theta^2}(\theta)|_{\theta=0} > 0$.*

Using Lemma 3.1, given two distinct transition matrices W and V, we define the relative entropy $D(W\|V)$ and the relative Rényi entropy $D_{1+s}(W\|V)$ as follows. For this purpose, we denote the logarithm of the Perron–Frobenius eigenvalue of the matrix $W(x|x')^{1+s}V(x|x')^{-s}$ by $\varphi(1+s)$ under the condition given below. When $\mathcal{X}_W^2 \subset \mathcal{X}_V^2$ and $\mathcal{X}_W^2$ is irreducible, we define

$$(3.2) \qquad D(W\|V) := \frac{d\varphi}{ds}(1), \qquad D_{1+s}(W\|V) := \frac{\varphi(1+s)}{s}$$

for $s > 0$. The relative Rényi entropy $D_{1+s}(W\|V)$ with $s \in (-1,0)$ is defined by (3.2) when $\mathcal{X}_W^2 \cap \mathcal{X}_V^2$ is irreducible, which is a weaker assumption. When $\mathcal{X}_W^2 \cap$

$\mathcal{X}_V^2$ is irreducible and the condition $\mathcal{X}_W^2 \subset \mathcal{X}_V^2$ does not hold, the relative entropy $D(W\|V)$ and the relative Rényi entropy $D_{1+s}(W\|V)$ with $s > 0$ are regarded as the infinity. Note that the limit $\lim_{s\to 0} D_{1+s}(W\|W')$ equals $D(W\|W')$. When $\mathcal{X}_W^2 \subset \mathcal{X}_V^2$ and $\mathcal{X}_W^2$ is irreducible, the function $\log \frac{W(x|x')}{V(x|x')}$ satisfies the condition for the function g in Lemma 3.1 because W and V are distinct. Hence, the function $s \mapsto s D_{1+s}(W\|V)$ is strictly convex. So, the relative Rényi entropy $D_{1+s}(W\|V)$ is strictly monotone increasing with respect to s.

From the property of Perron–Frobenius eigenvalue, we immediately obtain the following lemma.

LEMMA 3.2.   *Given two transition matrices W_X and V_X (W_Y and V_Y) on $\mathcal{X}$ ($\mathcal{Y}$), respectively, we have*

$$D(W_X\|V_X) + D(W_Y\|V_Y) = D(W_X \times W_Y \| V_X \times V_Y),$$

$$D_{1+s}(W_X\|V_X) + D_{1+s}(W_Y\|V_Y) = D_{1+s}(W_X \times W_Y \| V_X \times V_Y)$$

for $s \in (-1, 0) \cup (0, \infty)$.

THEOREM 3.3.   *Transition matrices W_1, W_2 and W satisfy*

$$(3.3) \qquad p D(W_1\|W) + (1-p) D(W_2\|W) \geq D(pW_1 + (1-p)W_2 \| W),$$

$$(3.4) \qquad p D(W\|W_1) + (1-p) D(W\|W_2) \geq D(W \| pW_1 + (1-p)W_2)$$

for $p \in (0, 1)$.

Equation (3.3) can be directly shown from Lemma 4.5 given later. The proof of (3.4) will be given after (5.5).

4. Information geometry for transition matrices.

4.1. *Exponential family.*   In the following, we treat only irreducible transition matrices. Hence, an irreducible transition matrix is simply called a transition matrix. We define an exponential family for transition matrices. We focus on a transition matrix $W(x|x')$ from $\mathcal{X}$ to $\mathcal{X}$. Then a set of real-valued functions $\{g_j\}$ on $\mathcal{X} \times \mathcal{X}$ is called *linearly independent* under the transition matrix $W(x|x')$ when any linear nonzero combination of $\{g_j\}$ satisfies the condition in Lemma 3.1. For $\vec{\theta} = (\theta^1, \ldots, \theta^d)$ and linearly independent functions $\{g_j\}$, we define the matrix $W_{\vec{\theta}}(x|x')$ from $\mathcal{X}$ to $\mathcal{X}$ in the following way:

$$(4.1) \qquad \overline{W}_{\vec{\theta}}(x|x') := W(x|x') e^{\sum_{j=1}^d \theta^j g_j(x,x')}.$$

Using the Perron–Frobenius eigenvalue $\lambda_{\vec{\theta}}$ of $\overline{W}_{\vec{\theta}}$, we define the potential function $\phi(\vec{\theta}) := \log \lambda_{\vec{\theta}}$.

Note that, since the value $\sum_x \overline{W}_{\vec{\theta}}(x|x')$ generally depends on x', we cannot make a transition matrix by simply multiplying a constant with the matrix $\overline{W}_{\vec{\theta}}$. To make a transition matrix from the matrix $\overline{W}_{\vec{\theta}}$, we recall that a nonnegative matrix V from $\mathcal{X}$ to $\mathcal{X}$ is a transition matrix if and only if the vector $(1, \dots, 1)^T$ is an eigenvector of the transpose V^T. In order to resolve this problem, we focus on the structure of the matrix $\overline{W}_{\vec{\theta}}$. We denote the Perron–Frobenius eigenvectors of $\overline{W}_{\vec{\theta}}$ and its transpose $\overline{W}_{\vec{\theta}}^T$ by $\overline{P}_{\vec{\theta}}^2$ and $\overline{P}_{\vec{\theta}}^3$. Then, similar to [27], (16); [26], (2), we define the matrix $W_{\vec{\theta}}(x|x')$ as

$$(4.2) \qquad W_{\vec{\theta}}(x|x') := \lambda_{\vec{\theta}}^{-1} \overline{P}_{\vec{\theta}}^3(x) \overline{W}_{\vec{\theta}}(x|x') \overline{P}_{\vec{\theta}}^3(x')^{-1}.$$

The matrix $W_{\vec{\theta}}(x|x')$ is a transition matrix because the vector $(1, \dots, 1)^T$ is an eigenvector of the transpose $W_{\vec{\theta}}^T$. The stationary distribution of the given transition matrix $W_{\vec{\theta}}$ is the Perron–Frobenius normalized eigenvector of the transition matrix $\overline{W}_{\vec{\theta}}$, which is given as

$$(4.3) \qquad \overline{P}_{\vec{\theta}}^1(x) := \frac{\overline{P}_{\vec{\theta}}^3(x) \overline{P}_{\vec{\theta}}^2(x)}{\sum_{x''} \overline{P}_{\vec{\theta}}^3(x'') \overline{P}_{\vec{\theta}}^2(x'')}$$

because

$$\sum_{x'} W_{\vec{\theta}}(x|x') \overline{P}_{\vec{\theta}}^1(x') = \frac{\overline{P}_{\vec{\theta}}^3(x)}{\lambda_{\vec{\theta}} \sum_{x''} \overline{P}_{\vec{\theta}}^3(x'') \overline{P}_{\vec{\theta}}^2(x'')} \sum_{x'} \overline{W}_{\vec{\theta}}(x|x') \overline{P}_{\vec{\theta}}^2(x')$$

$$= \frac{\overline{P}_{\vec{\theta}}^3(x) \overline{P}_{\vec{\theta}}^2(x)}{\sum_{x''} \overline{P}_{\vec{\theta}}^3(x'') \overline{P}_{\vec{\theta}}^2(x'')} = \overline{P}_{\vec{\theta}}^1(x).$$

In the following, we call the family of transition matrices $\mathcal{E} := \{W_{\vec{\theta}}\}$ an *exponential family* of transition matrices generated by W with the generator $\{g_1, \dots, g_d\}$.

Since the generator $\{g_1, \dots, g_d\}$ is linearly independent, due to Lemma 3.1, $\sum_{i,j} c_i c_j \frac{\partial^2 \phi}{\partial \theta^i \partial \theta^j} = \frac{d^2 \phi(\vec{c}t)}{dt^2}$ is strictly positive for an arbitrary nonzero vector $\vec{c} = (c_1, \dots, c_d)$. That is, the Hesse matrix $\mathsf{H}_{\vec{\theta}}[\phi] = [\frac{\partial^2 \phi}{\partial \theta^i \partial \theta^j}]_{i,j}$ is positive.

Using the potential function $\phi(\theta)$, we discuss several concepts for transition matrices based on Lemma 3.1, formally. We call the parameter $(\theta^1, \dots, \theta^d)$ the *natural parameter*, and the parameter $\eta_j(\vec{\theta}) := \frac{\partial \phi}{\partial \theta^j}(\vec{\theta})$ the *expectation parameter*. For $\vec{\eta} = (\eta_1, \dots, \eta_d)$, we define $\theta^1(\vec{\eta}), \dots, \theta^d(\vec{\eta})$ as $\eta_j(\theta^1(\vec{\eta}), \dots, \theta^d(\vec{\eta})) = \eta_j$.

For a given transition matrix W, we define a linear subspace $\mathcal{N}(\mathcal{X}_W^2)$ of the space $\mathcal{G}(\mathcal{X}_W^2)$ of all two-input functions as the set of functions $f(x) - f(x') + c$. Then we obtain the following lemma.

LEMMA 4.1. *The following are equivalent for the generator $\{g_j\}$ and the transition matrix W:*

(1) *The set of functions $\{g_j\}$ are linearly independent in the quotient space $\mathcal{G}(\mathcal{X}_W^2)/\mathcal{N}(\mathcal{X}_W^2)$.*

(2) *The map $\vec{\theta} \to \vec{\eta}(\vec{\theta})$ is one-to-one.*

(3) *The Hesse matrix $\mathsf{H}_{\vec{\theta}}[\phi]$ is strictly positive for any $\vec{\theta}$, which implies the strict convexity of the potential function $\phi(\vec{\theta})$.*

(4) *The Hesse matrix $\mathsf{H}_{\vec{\theta}}[\phi]|_{\vec{\theta}=0}$ is strictly positive.*

(5) *The parametrization $\vec{\theta} \mapsto W_{\vec{\theta}}$ is faithful for any $\vec{\theta}$.*

PROOF. Applying Lemma 3.1 to $\phi(\vec{c}t)$ for an arbitrary nonzero vector $\vec{c} = (c_1, \ldots, c_d)$, we obtain the equivalence among (1), (3) and (4). (3) $\Rightarrow$ (2) is trivial.

Now, we show (2) $\Rightarrow$ (1) by showing the contraposition. If (1) does not holds. There exists a nonzero vector $\vec{c} = (c_1, \ldots, c_d)$ such that $\sum_i c_i g_i(x, x') = f(x) - f(x') + C$. Hence, we have $\frac{d^2\phi(\vec{c}t)}{dt^2} = 0$. Hence, (2) does not hold.

Now, we show (1) $\Rightarrow$ (5) by showing the contraposition. When $W_{\vec{\theta}'} = W_{\vec{\theta}}$, considering the logarithm, there exist a function f and a constant C such that $\sum_j \theta'^j g_j(x, x') - \sum_j \theta^j g_j(x, x') = f(x) - f(x') + C$ for $(x, x') \in \mathcal{X}_W^2$.

Now, we show (5) $\Rightarrow$ (1) by showing the contraposition. If a set of real-valued functions $\{g_j\}$ on $\mathcal{X} \times \mathcal{X}$ is not linearly independent, there exist a function f and a constant C such that $\sum_j \theta'^j g_j(x, x') - \sum_j \theta^j g_j(x, x') = f(x) - f(x') + C$. In this case, choosing $\overline{P}_{\vec{\theta}'}^3(x) = \overline{P}_{\vec{\theta}}^3(x)e^{f(x)}$ and $\lambda_{\vec{\theta}'} = \lambda_{\vec{\theta}}e^{-C}$, $\overline{P}_{\vec{\theta}'}^3$ and $\lambda_{\vec{\theta}'}$ are the Perron–Frobenius eigenvector and eigenvalue of the transition matrix $W_{\vec{\theta}'}$. Then we have $W_{\vec{\theta}'} = W_{\vec{\theta}}$. $\square$

Now, we introduce the notation $\mathcal{W}_{\mathcal{X}, W} := \{V | V \text{ is a transition matrix and } \mathcal{X}_W^2 = \mathcal{X}_V^2\}$. Any element $W' \in \mathcal{W}_{\mathcal{X}, W}$ can be written as $W'(x|x') = W(x|x')e^{g(x,x')}$ by using an element $g \in \mathcal{G}(\mathcal{X}_W^2)$ because of $\log \frac{W'(x|x')}{W(x|x')} \in \mathcal{G}(\mathcal{X}_W^2)$. Hence, if and only if the set of two-input functions $\{g_j\}$ form a basis of the quotient space $\mathcal{G}(\mathcal{X}_W^2)/\mathcal{N}(\mathcal{X}_W^2)$, the set $\mathcal{W}_{\mathcal{X}, W}$ coincides with the exponential family generated by W with the generator $\{g_j\}$. This fact shows that $\mathcal{W}_{\mathcal{X}, W}$ is an exponential family.

In particular, when W is a positive transition matrix, the subspace $\mathcal{N}(\mathcal{X}_W^2)$ does not depend on W and is abbreviated to $\mathcal{N}(\mathcal{X}^2)$. In this case, $\mathcal{W}_{\mathcal{X}, W}$ is the set of positive transition matrices. Then it does not depend on W, and is abbreviated to $\mathcal{W}_{\mathcal{X}}$.

We define the Fisher information matrix for the natural parameter by the Hesse matrix $\mathsf{H}_{\vec{\theta}}[\phi] := [\frac{\partial^2\phi}{\partial\theta^i \partial\theta^j}(\vec{\theta})]_{i,j}$. The Fisher information matrix for the expectation parameter is given as $\mathsf{H}_{\vec{\theta}}[\phi]^{-1}$. Further, for fixed values $\theta_o^{k+1}, \ldots, \theta_o^d$, we call the subset $\{W_{\vec{\theta}} \in \mathcal{E} | \vec{\theta} = (\theta^1, \ldots, \theta^k, \theta_o^{k+1}, \ldots, \theta_o^d)\}$ an exponential subfamily of $\mathcal{E}$. The following are examples of an exponential family.

EXAMPLE 1. Now, we assume that $\mathcal{X} = \{0, 1, \ldots, m\}$ and W is a positive transition matrix, that is, $\mathcal{X}_W^2 = \mathcal{X}^2$. Define $g_{i,j}(x, x') = \delta_{x,i}\delta_{x',j}$ for $i = 1, \ldots, m$

and $j = 0, 1, \ldots, m$. Then the $m^2 + m$ functions $g_{i,j}$ form a basis of the quotient space $\mathcal{G}(\mathcal{X}^2)/\mathcal{N}(\mathcal{X}^2)$. Therefore, the set of positive transition matrices forms an exponential family with the above choice of $g_{i,j}$.

EXAMPLE 2. For a given subset $\mathcal{S} \subset \mathcal{X}^2$ for $\mathcal{X} = \{0, 1, \ldots, m\}$, we choose a transition matrix W whose support is $\mathcal{S}$. Define the subset $\tilde{\mathcal{S}}$ as $\{(i, j) \in S | i$ is not minimum integer satisfying $(i, j) \in \mathcal{S}$ for a fixed $j\}$. We define $g_{i,j}(x, x') = \delta_{x,i}\delta_{x',j}$ for $(i, j) \in \tilde{\mathcal{S}}$. Then the set $\mathcal{W}_{\mathcal{X},W}$ is an exponential family generated by $\{g_{i,j}\}_{(i,j)\in\tilde{\mathcal{S}}}$. However, the set $\mathcal{W}_{\mathcal{X},W}$ is not an exponential subfamily of the set of positive transition matrices because it is not included in the set of positive transition matrices.

REMARK 1. The above-defined exponential families contain exponential families of distributions as follows. For a given exponential family of distributions P_θ on $\mathcal{X}$ with the generator $f(x)$, we define the transition matrix $W(x|x')$ as $P_0(x)$ and the generator $g(x, x')$ as $f(x)$. Then the exponential family $W_\theta(x|x')$ is $P_\theta(x)$. The given potential function and the given expectation parameter (defined in the next subsection) are the same as those in the case with the exponential family of distributions $\{P_\theta\}$.

REMARK 2. The papers [3, 10, 13, 20] called a family of transition matrices $\{W_\theta(x|x')\}$ an exponential family when $W_\theta(x|x')$ has the form

(4.4) $W_\theta(x|x') = e^{C(x,x')+\theta g(x,x')-\psi(\theta,x')}$.

The papers [19, 35, 36] extended the above definition to the continuous-time case. However, our exponential family is written as [26]

(4.5) $W_\theta(x|x') = e^{C(x,x')+\theta g(x,x')+\psi(\theta,x)-\psi(\theta,x')-\phi(\theta)}$

by choosing $C(x, x')$ and $\psi(\theta, x)$ as $\log W(x|x')$ and $\log \overline{P}_\theta^3(x)$, respectively. So, the traditional definition (4.4) is different from ours. The advantage of our model over their model is explained in Remark 3.

4.2. *Mixture family.* In the following, we assume that the functions $\{g_j\}$ satisfies the condition of Lemma 4.1. For fixed values $\eta_{o,1}, \ldots, \eta_{o,k}$, we call the subset $\{W_{\vec{\theta}} \in \mathcal{E} | \vec{\eta}(\vec{\theta}) = (\eta_{o,1}, \ldots, \eta_{o,k}, \eta_{k+1}, \ldots, \eta_d)\}$ a *mixture subfamily of* $\mathcal{E}$. Given a transition matrix W, real-valued functions g_j on $\mathcal{X}^2$, and real numbers b_j, we say that the set $\{V \in \mathcal{W}_{\mathcal{X},W} | \sum_{x,x'} g_j(x, x')V(x|x')P_V(x') = b_j \ \forall j\}$ is a *mixture family on* $\mathcal{X}_W^2$ *generated by the constraints* $\{g_j = b_j\}$. Note that a mixture family on $\mathcal{X}_W^2$ does not necessarily contain W because its definition depends on the real numbers b_j. When W is a positive transition matrix, it is simply called a *mixture family generated by the constraints* $\{g_j = b_j\}$ because $\mathcal{W}_{\mathcal{X},W}$ is the set of positive transition matrices. For a given transition matrix W and two mixture families $\mathcal{M}_1$ and $\mathcal{M}_2$ on $\mathcal{X}_W^2$, the intersection $\mathcal{M}_1 \cap \mathcal{M}_2$ is also a mixture family on $\mathcal{X}_W^2$.

LEMMA 4.2. *The intersection of the mixture family on $\mathcal{X}_W^2$ generated by the constraints $\{g_j = b_j\}_{j=1,\ldots,k}$ and the exponential family $\mathcal{W}_{\mathcal{X},W}$ is the mixture subfamily $\{W_{\vec{\theta}} \in \mathcal{W}_{\mathcal{X},W} | \vec{\eta}(\vec{\theta}) = (b_1, \ldots, b_k, \eta_{k+1}, \ldots, \eta_d)\}$ of the exponential family $\mathcal{W}_{\mathcal{X},W}$.*

Lemma 4.2 will be shown after Lemma 5.1 in Section 5. Here, we give examples for mixture families.

EXAMPLE 3. A transition matrix W on $\mathcal{X} \times \mathcal{Y}$ is called *nonhidden* for $\mathcal{X}$ when $W_X(x|x') := \sum_{y \in \mathcal{Y}} W(x, y|x', y')$ does not depend on $y' \in \mathcal{Y}$. For a transition matrix W on $\mathcal{X} \times \mathcal{Y}$, the set $\mathcal{W}_{\mathcal{X}|\mathcal{X} \times \mathcal{Y}, W} := \{V \in \mathcal{W}_{\mathcal{X} \times \mathcal{Y}, W} | V \text{ is non-hidden for } \mathcal{X} \text{ on } \mathcal{X} \times \mathcal{Y}\}$ is a mixture family on $(\mathcal{X} \times \mathcal{Y})_W^2$. Hence, the set $\mathcal{W}_{\mathcal{X}|\mathcal{X} \times \mathcal{Y}, W} \cap \mathcal{W}_{\mathcal{Y}|\mathcal{X} \times \mathcal{Y}, W}$ is also a mixture family on $\mathcal{X}_W^2$.

EXAMPLE 4. The set of bistochastic matrices on $\mathcal{X} = \{0, 1, \ldots, m\}$ forms a mixture family as follows. For a permutation σ, we define the transition matrix $W_\sigma(x|x') = \delta_{x, \sigma x'}$. Then we focus on the set T of transpositions (i, j) and the subset H of cyclic permutations with length 3 defined by $H := \{(0, i, j)|0 < i < j \leq m\}$. Then $|T \cup H| = |T| + |H| = \frac{m(m+1)}{2} + \frac{m(m-1)}{2} = m^2$. As will be shown in Appendix B, The set of bistochastic matrices on $\mathcal{X} = \{0, 1, \ldots, m\}$ is parametrized as $\{W_{\vec{\eta}}\}_{\vec{\eta} \in E}$, where

$$(4.6) \qquad W_{\vec{\eta}} := \sum_{\sigma \in T \cup H} \eta_\sigma W_\sigma + \left(1 - \sum_{\sigma \in T \cup H} \eta_\sigma\right) W_{id},$$

$$(4.7) \qquad E := \{\eta \in \mathbb{R}^{m^2} | W_{\vec{\eta}}(x|x') \geq 0 \text{ for } \forall x, x' \in \mathcal{X}\}.$$

We define the functions

$$(4.8) \qquad g_i(x, x') := \delta_{x,i} - \delta_{x,0} \qquad \text{for } i = 1, \ldots, m,$$

$$(4.9) \qquad \hat{g}_\sigma(x, x') := W_\sigma(x|x') - W_{id}(x|x').$$

As will be shown in Appendix B, the set $\{g_i\}_{i=1}^m \cup \{\hat{g}_\sigma\}_{\sigma \in T \cup H}$ is linearly independent. Then the matrix $A = (a_{\sigma,\sigma'})$ given as follows is invertible:

$$(4.10) \qquad a_{\sigma,\sigma'} := \sum_{x,x'} \hat{g}_{\sigma'}(x, x') \hat{g}_\sigma(x, x') \frac{1}{m+1}.$$

Then, using the inverse matrix $B = A^{-1}$, we can define the functions $\{g_\sigma\}_{\sigma \in T \cup H}$ as the dual basis in the following way:

$$(4.11) \qquad g_{\sigma'} := \sum_{\sigma \in T \cup H} b_{\sigma,\sigma'} \hat{g}_\sigma,$$

which implies that

$$(4.12) \qquad \sum_{x,x'} g_{\sigma'}(x,x') \hat{g}_\sigma(x,x') \frac{1}{m+1} = \delta_{\sigma,\sigma'}.$$

Hence, the set of functions $\{g_i\}_{i=1}^m \cup \{g_\sigma\}_{\sigma \in T \cup H}$ is linearly independent. We can employ the mixture parameter under the above set of functions. Since the stationary distribution of $W_{\vec{\eta}}$ is the uniform distribution and

$$(4.13) \qquad \sum g_i(x,x') W_{\vec{\eta}}(x|x') \frac{1}{m+1} = 0 \qquad \text{for } i = 1, \dots, m,$$

$$(4.14) \qquad \sum g_\sigma(x,x') W_{\vec{\eta}}(x|x') \frac{1}{m+1} = \eta_\sigma \qquad \text{for } \sigma \in T \cup H,$$

the transition matrix $W_{\vec{\eta}}(x|x')$ is the expectation parameter $(0, \dots, 0, \eta_\sigma)$. That is, the set of bistochastic matrices on $\mathcal{X}$ is the mixture family generated by the constraints $\{g_j = 0\}_{j=1,\dots,m}$.

4.3. *Relation with relative entropy and relative Rényi entropies.* The relative entropy and the relative Rényi entropies are characterized by using the potential function $\phi(\vec{\theta})$ as follows.

LEMMA 4.3. *Two transition matrices $W_{\vec{\theta}}$ and $W_{\vec{\theta}'}$ satisfy*

$$(4.15) \qquad D(W_{\vec{\theta}} \| W_{\vec{\theta}'}) = \sum_{j=1}^d (\theta^j - \theta'^j) \frac{\partial \phi}{\partial \theta^j}(\vec{\theta}) - \phi(\vec{\theta}) + \phi(\vec{\theta}'),$$

$$(4.16) \qquad D_{1+s}(W_{\vec{\theta}} \| W_{\vec{\theta}'}) = \frac{\phi((1+s)\vec{\theta} - s\vec{\theta}') - (1+s)\phi(\vec{\theta}) + s\phi(\vec{\theta}')}{s}.$$

PROOF. Let $\varphi(1+s)$ be the logarithm of the Perron–Frobenius eigenvalue of the matrix $W_{\vec{\theta}}(x|x')^{1+s} W_{\vec{\theta}'}(x|x')^{-s}$. Since $W_{\vec{\theta}}(x|x')^{1+s} W_{\vec{\theta}'}(x|x')^{-s} \exp((1+s)\phi(\vec{\theta}) - s\phi(\vec{\theta}'))$ equals $\overline{W}_{\vec{\theta}}(x|x')$, we have $\varphi(1+s) = \phi((1+s)\vec{\theta} - s\vec{\theta}') - (1+s)\phi(\vec{\theta}) + s\phi(\vec{\theta}')$. Hence, we obtain (4.16). Taking the limit $s \to 0$, we obtain (4.15). $\square$

The Fisher information matrix $\mathsf{H}_{\vec{\theta}}[\phi]$ can be characterized by the limits of the relative entropy and relative Rényi entropy as follows. That is, taking the limits in (4.15) and (4.16) in Lemma 4.3, we can show the following lemma.

LEMMA 4.4. *For $\vec{c} = (c_1, \dots, c_d)$, we have*

$$(4.17) \qquad \lim_{t \to 0} \frac{2}{t^2} D(W_{\vec{\theta}} \| W_{\vec{\theta}+\vec{c}t}) = \lim_{t \to 0} \frac{2}{t^2} D(W_{\vec{\theta}+\vec{c}t} \| W_{\vec{\theta}}) = \sum_{i,j} \mathsf{H}_{\vec{\theta}}[\phi]_{i,j} c^i c^j,$$

$$\lim_{t\to 0}\frac{2}{t^2}D_{1+s}(W_{\vec{\theta}}\|W_{\vec{\theta}+\vec{c}t}) = \lim_{t\to 0}\frac{2}{t^2}D_{1+s}(W_{\vec{\theta}+\vec{c}t}\|W_{\vec{\theta}})$$

(4.18)

$$= (1+s)\sum_{i,j}\mathsf{H}_{\vec{\theta}}[\phi]_{i,j}c^i c^j.$$

The right-hand side of (4.15) can be regarded as the Bregman divergence [6][3] of the strictly convex function $\phi(\vec{\theta})$. In the following, we derive several properties of the relative entropy by using Bregman divergence. That is, the following properties follow only from the strong convexity of $\phi(\vec{\theta})$ and the properties of Bregman divergence.

Using [1], (40), we have another expression of $D(W_{\vec{\theta}}\|W_{\vec{\theta}'})$ as

(4.19) $$D(W_{\vec{\theta}(\vec{\eta})}\|W_{\vec{\theta}(\vec{\eta}')}) = \sum_j \theta(\vec{\eta}')^j(\eta'_j - \eta_j) - \nu(\vec{\eta}') + \nu(\vec{\eta}),$$

where $\nu(\vec{\eta})$ is defined as Legendre transform of $\phi(\vec{\theta})$ as

$$\nu(\vec{\eta}) := \max_{\vec{\theta}}\sum_i \theta^i \eta_i - \phi(\vec{\theta}) = \sum_i \theta^i(\vec{\eta})\eta_i - \phi(\vec{\theta}(\vec{\eta})).$$

Since $\nu(\vec{\eta})$ is convex as well as $\phi(\vec{\theta})$, we have the following lemma.

LEMMA 4.5.   (1) *For a fixed $\vec{\theta}$, the maps $\vec{\theta}' \mapsto D(W_{\vec{\theta}}\|W_{\vec{\theta}'})$ and $D_{1+s}(W_{\vec{\theta}}\| W_{\vec{\theta}'})$ are convex for $s > 0$. (2) For a fixed $\vec{\theta}'$, the map $\vec{\eta} \mapsto D(W_{\vec{\theta}(\vec{\eta})}\|W_{\vec{\theta}'})$ is convex.*

4.4. *Pythagorean theorem.*   It is known that Bregman divergence satisfies the Pythagorean theorem for [1], (34). Applying this fact, we have the following proposition as the Pythagorean theorem.

PROPOSITION 4.6 (Nagaoka [26], (23)).   *We focus on two points $\vec{\theta}' = (\theta'^1, \ldots, \theta'^d)$ and $\vec{\theta}'' = (\theta''^1, \ldots, \theta''^d)$. We choose the exponential subfamily of $\mathcal{E}$ whose natural parameters $\theta^{k+1}, \ldots, \theta^d$ are fixed to $\theta''^{k+1}, \ldots, \theta''^d$, and the mixture subfamily of $\mathcal{E}$ whose expectation parameters $\eta_1, \ldots, \eta_k$ are fixed to $\eta(\vec{\theta}')^1, \ldots, \eta(\vec{\theta}')^k$. Let $\tilde{\vec{\theta}} = (\tilde{\theta}^1, \ldots, \tilde{\theta}^d)$ be the natural parameter of the intersection of these two subfamilies of $\mathcal{E}$. That is, $\tilde{\theta}^j = \theta''^j$ for $j = k+1, \ldots, d$ and $\eta_j(\tilde{\vec{\theta}}) = \eta_j(\vec{\theta}')$ for $k = 1, \ldots, k$. Then we have*

(4.20) $$D(W_{\vec{\theta}'}\|W_{\vec{\theta}''}) = D(W_{\vec{\theta}'}\|W_{\tilde{\vec{\theta}}}) + D(W_{\tilde{\vec{\theta}}}\|W_{\vec{\theta}''}).$$

---

[3]Amari–Nagaoka [2] also defined the same quantity as the Bregman divergence with the name "canonical divergence." They showed that the canonical divergence satisfies the Pythagorean theorem and (4.19) via the concept of the dually flat. Recently, Amari [1] showed these properties by a calculation of the convex function $\phi(\vec{\theta})$, which does not require Christoffel symbols calculation. Since the derivations by [1] more directly explain the relation between the convex function $\phi(\vec{\theta})$ and these properties, we refer the paper [1] for these properties.

Indeed, Nagaoka [26] showed (4.20) in a more general form by showing the dually flat structure [2] via Christoffel symbols calculation. Using (4.20) and Lemma 4.2, we obtain the following corollary.

COROLLARY 4.7. *Given a transition matrix V and a mixture family $\mathcal{M}$ on $\mathcal{X}_V^2$ with constraints $\{g_j = b_j\}_{j=1}^k$, we define $V^* := \operatorname{argmin}_{W \in \mathcal{M}} D(W \| V)$.*

(1) *Any transition matrix $W \in \mathcal{M}$ satisfies $D(W \| V) = D(W \| V^*) + D(V^* \| V)$.*

(2) *The transition matrix V^* is the intersection of the mixture family $\mathcal{M}$ on $\mathcal{X}_V^2$ and the exponential family generated by V and the generator $\{g_j\}_{j=1}^k$.*

PROOF. First, we notice that the exponential family $\mathcal{W}_{\mathcal{X},V}$ contains V and includes $\mathcal{M}$. Choose an element $\tilde{V}$ in the intersection of the mixture family $\mathcal{M}$ on $\mathcal{X}_V^2$ and the exponential family $\mathcal{E}_V$ generated by V and the generator $\{g_j\}_{j=1}^k$. We apply (4.20) to the mixture family $\mathcal{M}$ and the exponential family $\mathcal{E}_V$. Then any transition matrix $W \in \mathcal{M}$ satisfies that $D(W \| V) = D(W \| \tilde{V}) + D(\tilde{V} \| V)$. Since $D(W \| \tilde{V}) > 0$ except for $W = \tilde{V}$, we have $\min_{W \in \mathcal{M}} D(W \| V) = D(\tilde{V} \| V)$, which implies that $V^* = \tilde{V}$, that is, (2). Hence, we obtain (1). $\square$

Similarly, we have another version of the above corollary.

COROLLARY 4.8. *Given a transition matrix W and an exponential family $\mathcal{E} \subset \mathcal{W}_{\mathcal{X},W}$ with the generator $\{g_j\}$, we define $W_* := \operatorname{argmin}_{V \in \mathcal{E}} D(W \| V)$. Assume that $\sum_{x,x'} g_j(x, x') W_*(x|x') P_{W_*}(x') = b_j$.*

(1) *Any transition matrix $V \in \mathcal{E}$ satisfies $D(W \| V) = D(W \| W_*) + D(W_* \| V)$.*

(2) *The transition matrix W_* is the intersection of the exponential family $\mathcal{E}$ and the mixture family on $\mathcal{X}_W^2$ with the constraints $\{g_j = b_j\}$.*

EXAMPLE 5. We choose transition matrices V_X and V_Y on $\mathcal{X}$ and $\mathcal{Y}$, respectively. We also choose a transition matrix W on $\mathcal{X} \times \mathcal{Y}$ whose support is $(\mathcal{X} \times \mathcal{Y})_{V_X \times V_Y}^2$. When a set of two-input functions $\{g_{X|i}\}$ forms a basis of $\mathcal{G}(\mathcal{X}_{V_X}^2)/\mathcal{N}(\mathcal{X}_{V_X}^2)$, the exponential family generated by $V_X \times V_Y$ with the generator $\{g_{X|i}\}$ is $\{V_X' \times V_Y | V_X' \in \mathcal{W}_{\mathcal{X},V_X}\}$. When a set of two-input functions $\{g_{Y|j}\}$ forms a basis of $\mathcal{G}(\mathcal{Y}_{V_Y}^2)/\mathcal{N}(\mathcal{Y}_{V_Y}^2)$, the exponential family generated by $V_X \times V_Y$ with the generator $\{g_{X|i}\} \cup \{g_{Y|j}\}$ is $\{V_X' \times V_Y' | V_X' \in \mathcal{E}_{\mathcal{X},V_X}, V_Y' \in \mathcal{E}_{\mathcal{Y},V_Y}\}$. Hence, when a transition matrix W belongs to a mixture family with the constraints $\{g_{X|i} = a_i\} \cup \{g_{Y|j} = b_j\}$, the intersection between the exponential family and the mixture family consists of one points, which is denoted by $W_X' \times W_Y'$. Applying (4.20), we obtain

$$(4.21) \quad D(W \| V_X \times V_Y) = D(W \| W_X' \times W_Y') + D(W_X' \times W_Y' \| V_X \times V_Y).$$

In particular, when W is nonhidden for $\mathcal{X}$ (for the definition, see Example 3), W_X satisfies the same constraint $\{g_{X|i} = a_i\}$ because the stationary distribution P_{W_X} is the marginal distribution of the stationary distribution P_W. Hence, $W'_X = W_X$. Thus, W'_X can be regarded as a marginalization of a transition matrix W that is not necessarily nonhidden.

5. Stationary two-observation case.

5.1. *Relative entropies and expectation.* In the previous section, we formally defined several information quantities from the convex function $\phi(\vec{\theta})$ in the multi-parameter case. In this section, we consider the relation with the structure of probabilities in the one-parameter case. That is, we will see how the information quantities reflect the conventional information quantities. For this purpose, we assume that the input distribution is the stationary distribution of the given transition matrix.

Since the stationary distribution of the given transition matrix W_θ is $\overline{P}^1_\theta$ given in (4.1), we can define the joint distribution

$$(5.1) \qquad W_\theta \times \overline{P}^1_\theta(x, x') := W_\theta(x|x')\overline{P}^1_\theta(x') = \frac{\overline{P}^3_\theta(x)\overline{W}_\theta(x|x')\overline{P}^2_\theta(x')}{\lambda_\theta \sum_{x''} \overline{P}^3_\theta(x'')\overline{P}^2_\theta(x'')}$$

on $\mathcal{X} \times \mathcal{X}$. Now, we focus on the probability distribution family $\{W_\theta \times \overline{P}^1_\theta\}$, and denote the expectation and the variance under the distribution $W_\theta \times \overline{P}^1_\theta$ by E_θ and V_θ. These are simplified to E and V when $\theta = 0$.

LEMMA 5.1 ([26], Theorem 4, [27], (28)). *For $\theta \in \mathbb{R}$, we have*

$$(5.2) \qquad \eta(\theta) = \frac{d\phi}{d\theta}(\theta) = \mathsf{E}_\theta[g(X, X')] = \sum_{x,x'} \overline{P}^1_\theta(x')W_\theta(x|x')g(x, x').$$

The lemma shows the reason why we call the parameter η the expectation parameter.

PROOF OF LEMMA 5.1. From the definition of W_θ, we have

$$(5.3) \qquad \frac{d}{d\theta}\log W_\theta(x|x') = -\frac{d}{d\theta}\log\lambda_\theta + \frac{d}{d\theta}\log\frac{\overline{P}^3_\theta(x)}{\overline{P}^3_\theta(x')} + g(x, x').$$

Taking the average of the both hand sides with respect to the distribution $W_\theta \times \overline{P}^1_\theta$, we have $0 = -\frac{d}{d\theta}\log\lambda_\theta + \sum_{x,x'} \overline{P}^1_\theta(x')W_\theta(x|x')g(x, x')$. □

Lemma 5.1 shows Lemma 4.2 as follows.

PROOF OF LEMMA 4.2.   In this proof, we consider the multiparameter case. Replacing the derivative by the partial derivative in Lemma 5.1, we have

$$(5.4) \qquad \eta_j(\vec{\theta}) = \frac{\partial \phi}{\partial \theta_j}(\vec{\theta}) = \sum_{x,x'} \overline{P}_{\vec{\theta}}^1(x') W_{\vec{\theta}}(x|x') g_j(x, x').$$

Choose the generator $\{g_1, \ldots, g_k\}$ of the mixture family on $\mathcal{X}_W^2$. There exist two-input functions $g_{k+1}, \ldots, g_l$ such that the set of two-input functions $\{g_1, \ldots, g_l\}$ form a basis of $\mathcal{G}(\mathcal{X}_W^2)/\mathcal{N}(\mathcal{X}_W^2)$, Hence, due to (5.4), we see that the intersection of the mixture family on $\mathcal{X}_W^2$ generated by the constraints $\{g_j = b_j\}_{j=1,\ldots,k}$ and the exponential family $\mathcal{W}_{\mathcal{X},W}$ is the mixture subfamily $\{W_{\vec{\theta}} \in \mathcal{W}_{\mathcal{X},W} | \vec{\eta}(\vec{\theta}) = (b_1, \ldots, b_k, \eta_{k+1}, \ldots, \eta_d)\}$ of the exponential family $\mathcal{W}_{\mathcal{X},W}$. $\square$

Now, we introduce the conditional relative entropy for transition matrices W and V from $\mathcal{X}$ to $\mathcal{Y}$ and a distribution P on $\mathcal{X}$ as follows:

$$D(W\|V|P) := D(W \times P \| V \times P),$$

where the relative entropy between two distributions P and P' is defined in the conventional way as $D(P\|P') := \sum_x P(x) \log \frac{P(x)}{P'(x)}$. Hence, the relative entropy defined in the previous section is characterized as follows [26], (24):

$$
\begin{aligned}
D(W_\theta \| W_{\theta'}) &= (\theta - \theta') \frac{d\phi}{d\theta}(\theta) - \phi(\theta) + \phi(\theta') \\
&= \sum_{x,x'} \overline{P}_\theta^1(x') W_\theta(x|x') \log \frac{\overline{W}_\theta(x|x')}{\overline{W}_{\theta'}(x|x')} - \phi(\theta) + \phi(\theta') \\
&= \sum_{x,x'} \overline{P}_\theta^1(x') W_\theta(x|x') \log \frac{W_\theta(x|x')}{W_{\theta'}(x|x')} - \log \frac{\overline{P}_\theta^3(x)}{\overline{P}_{\theta'}^3(x)} + \log \frac{\overline{P}_\theta^3(x')}{\overline{P}_{\theta'}^3(x')} \\
&\stackrel{(a)}{=} \sum_{x,x'} \overline{P}_\theta^1(x') W_\theta(x|x') \log \frac{W_\theta(x|x')}{W_{\theta'}(x|x')} = D(W_\theta \| W_{\theta'} | \overline{P}_\theta^1),
\end{aligned}
$$
(5.5)

where (a) follows from the fact that $\overline{P}_\theta^1(x) = \sum_{x'} W_\theta(x|x') \overline{P}_\theta^1(x')$.

PROOF OF (3.4).   Since the map $W' \mapsto -\sum_{x,x'} \overline{P}_\theta^1(x') W_\theta(x|x') \log W'(x|x')$ is convex for a given θ, (5.5) guarantees (3.4). $\square$

5.2. *Fisher information and variance.*   Using the Fisher information J_θ^1 of the family $\{\overline{P}_\theta^1\}_\theta$ of stationary distributions, we discuss the Fisher information J_θ^2 of the family $\{W_\theta \times \overline{P}_\theta^1\}_\theta$ of joint distributions in the following lemma.

LEMMA 5.2.   *The Fisher information J_θ^2 can be written as*

$$(5.6) \qquad J_\theta^2 = \frac{d^2\phi}{d\theta^2}(\theta) + J_\theta^1.$$

LEMMA 5.3.   *The second derivative $\frac{d^2\phi}{d\theta^2}(\theta)$ is calculated as*

$$(5.7) \quad \frac{d^2\phi}{d\theta^2}(\theta) = \mathsf{V}_\theta\left[g(X, X') - \frac{d\phi}{d\theta}(\theta) + \frac{d}{d\theta}\log \overline{P}_\theta^3(X) - \frac{d}{d\theta}\log \overline{P}_\theta^3(X') \right].$$

In particular, when $\theta = 0$,

$$(5.8) \qquad \frac{d^2\phi}{d\theta^2}(0) = \mathsf{V}_0[g(X, X')] + 2\sum_{x,x'} W(x|x')g(x, x')\frac{d\overline{P}_\theta^2(x')}{d\theta}\bigg|_{\theta=0}.$$

Proofs of Lemmas 5.2 and 5.3 are given in Appendix C. Further, the quantity $\frac{d^2\phi}{d\theta^2}(0)$ has another form [12], Theorem 6.6. Using Lemma 5.3, we can show Lemma 3.1 as follows.

PROOF OF LEMMA 3.1.   Due to (5.7), the nonnegativity of variance implies that $\phi(\theta)$ is convex. Since Condition (2) trivially implies Condition (3), it is enough to show that Condition (1) implies Condition (2) and Condition (3) implies Condition (1).

Assume Condition (1). Then, the random variable $g(X, X') - \frac{d\phi}{d\theta}(\theta) + \frac{d}{d\theta}\log \overline{P}_\theta^3(X) - \frac{d}{d\theta}\log \overline{P}_\theta^3(X')$ is not a constant on $\mathcal{X}_W^2$. Hence, the variance in (5.7) is strictly greater than zero, which implies Condition (2).

Conversely, we assume that Condition (1) does not hold, that is, $g(x, x') = f(x) - f(x') + C$ for any $(x, x') \in \mathcal{X}_W^2$ with a constant $C \in \mathbb{R}$. Then, we can find that the Perron–Frobenius eigenvalue of $\overline{W}_\theta(x|x') = W(x|x')e^{\theta f(x) - \theta f(x') + \theta C}$ is $\lambda_\theta = e^{\theta C}$ and its right eigenvector is $\overline{P}_\theta^2$. Thus, we have $\frac{d^2\phi(\theta)}{d\theta^2} = 0$, that is, Condition (3) does not hold. Hence, Condition (3) implies Condition (1).   □

6. Stationary $n + 1$-observation case.

6.1. *Information quantities.*   Similar to the previous section, this section also discusses the one-parameter case with the stationary initial distribution $\overline{P}_\theta^1$. Now, we consider the distribution $W_\theta^{\times n} \times \overline{P}_\theta^1$ on $\mathcal{X}^n$, which is defined as

$$(6.1) \qquad W_\theta^{\times n} \times \overline{P}_\theta^1(x_n, \ldots, x_1) := W_\theta(x_{n+1}|x_n)\cdots W_\theta(x_2|x_1)\overline{P}_\theta^1(x_1).$$

We also define the random variable $g^n(X^{n+1}) := \sum_{k=1}^n g(X_{k+1}, X_k)$ for $X^{n+1} := (X_{n+1}, \ldots, X_1)$. In this section, we denote the expectation and the variance under

the distribution $W_\theta^n \times \overline{P}_\theta^1$ by E_θ and V_θ. Then the cumulant generating function $\phi_n(\theta) := \log \mathsf{E}_0[\exp(\theta g^n(X^{n+1}))]$ satisfies

$$(6.2) \qquad \frac{d\phi_n}{d\theta}(\theta) = \mathsf{E}_\theta[g^n(X^{n+1})] = n\eta(\theta).$$

Now, we calculate information quantities. Similar to Lemma 5.2, the Fisher information can be calculated as follows.

LEMMA 6.1. *The Fisher information J_θ^{n+1} of the family $\{W_\theta^{\times n} \times \overline{P}_\theta^1\}_\theta$ can be written as*

$$(6.3) \qquad J_\theta^{n+1} = n \frac{d^2\phi}{d\theta^2}(\theta) + J_\theta^1.$$

The proof can be done in the same way as Lemma 5.2. The conditional relative entropy is characterized by the Bregman divergence defined by the convex function $\phi(\theta)$ as follows:

$$
\begin{aligned}
D\big(W_\theta^{\times n} \| W_{\theta'}^{\times n} | \overline{P}_\theta^1\big) &:= D\big(W_\theta^{\times n} \times \overline{P}_\theta^1 \| W_{\theta'}^{\times n} \times \overline{P}_\theta^1\big) \\
&= n\left((\theta - \theta')\frac{d\phi}{d\theta}(\theta) - \phi(\theta) + \phi(\theta')\right) \\
&= nD(W_\theta \| W_{\theta'}).
\end{aligned}
$$
(6.4)

6.2. *Asymptotically efficient estimator.* The relation (6.2) implies that $\frac{g^n(X^{n+1})}{n}$ is an unbiased estimator for the parameter η. The variance of $g^n(X^{n+1})$ is evaluated as follows.

LEMMA 6.2. *The inequalities*

$$
n\frac{d^2\phi}{d\theta^2}(\theta)\left(1 - 2\sqrt{\hat{\mathsf{V}}_\theta \Big/ n\frac{d^2\phi}{d\theta^2}(\theta)}\right)^2 \le \mathsf{V}_\theta[g^n(X^{n+1})]
$$
(6.5)
$$
\le n\frac{d^2\phi}{d\theta^2}(\theta)\left(1 + 2\sqrt{\hat{\mathsf{V}}_\theta \Big/ n\frac{d^2\phi}{d\theta^2}(\theta)}\right)^2
$$

hold, where $\hat{\mathsf{V}}_\theta := \mathsf{V}_\theta[\frac{d}{d\theta}\log \overline{P}_\theta^3(X)] = \sum_x \overline{P}_\theta^1(x)(\frac{d}{d\theta}\log \overline{P}_\theta^3(x))^2$.

Hence, we obtain

$$(6.6) \qquad \mathsf{V}_\theta\left[\frac{g^n(X^{n+1})}{n}\right] = \frac{\mathsf{V}_\theta[g^n(X^{n+1})]}{n^2} = \frac{d^2\phi}{d\theta^2}(\theta)\Big/n + O\left(\frac{1}{n\sqrt{n}}\right).$$

The Fisher information $\tilde{J}_{\eta(\theta)}^{n+1}$ for the expectation parameter η of the family $\{W_\theta^{\times n} \times \overline{P}_\theta^1\}_\theta$ is

$$
\tilde{J}_{\eta(\theta)}^{n+1} = J_\theta^{n+1} \Big/ \left(\frac{d\eta(\theta)}{d\theta}\right)^2 = \left(n\frac{d^2\phi}{d\theta^2}(\theta) + J_\theta^1\right)\left(\frac{d^2\phi}{d\theta^2}(\theta)\right)^{-2}
$$

$$
= \frac{n(1 + J_\theta^1/n\frac{d^2\phi}{d\theta^2}(\theta))}{(\frac{d^2\phi}{d\theta^2}(\theta))}.
$$

That is, the lower bound of the variance of the unbiased estimator given by Cramér–Rao inequality is $\frac{d^2\phi}{d\theta^2}(\theta)/n(1 + J_\theta^1/n\frac{d^2\phi}{d\theta^2}(\theta))$. Hence, any unbiased estimator Z_n for the expectation parameter η satisfies

(6.7) $\qquad V_\theta[Z_n] \geq \dfrac{\frac{d^2\phi}{d\theta^2}(\theta)}{n(1 + J_\theta^1/n\frac{d^2\phi}{d\theta^2}(\theta))} = \dfrac{\frac{d^2\phi}{d\theta^2}(\theta)}{n} - \dfrac{J_\theta^1}{n^2} + o\left(\dfrac{1}{n^2}\right).$

The relation (6.6) shows that the unbiased estimator $\frac{g^n(X^{n+1})}{n}$ realizes the optimal performance with the order $\frac{1}{n}$.

PROOF OF LEMMA 6.2.   Remember that Lemma 5.3 was shown by the combination of (C.1) and (C.2). In the $n+1$-observation case, combining Lemma 6.1 and the $n+1$-observation version of (C.2), we can similarly show that $n\frac{d^2\phi}{d\theta^2}(\theta)$ is the variance of $[-n\frac{d\phi}{d\theta}(\theta) + \frac{d}{d\theta}\log\overline{P}_\theta^3(X_{n+1}) - \frac{d}{d\theta}\log\overline{P}_\theta^3(X_1) + g^n(X^{n+1})]$ under the distribution $W_\theta^n \times \overline{P}_\theta^1$.

Now, we define the 2-norm of the random variable $f(X^{n+1})$ as $\|f\|_2 := \sqrt{\sum_{x^{n+1}} W_\theta^n \times P_\theta(x^{n+1})f(x^{n+1})^2}$. Then we have

$$
\sqrt{n\frac{d^2\phi}{d\theta^2}(\theta)} = \left\| g^n(X^{n+1}) - n\frac{d\phi}{d\theta}(\theta) + \frac{d}{d\theta}\log\overline{P}_\theta^3(X_{n+1}) - \frac{d}{d\theta}\log\overline{P}_\theta^3(X_1) \right\|_2
$$

$$
\leq \left\| g^n(X^{n+1}) - n\frac{d\phi}{d\theta}(\theta) \right\|_2 + \left\| \frac{d}{d\theta}\log\overline{P}_\theta^3(X_{n+1}) \right\|_2
$$

$$
+ \left\| \frac{d}{d\theta}\log\overline{P}_\theta^3(X_1) \right\|_2
$$

$$
= \sqrt{V_\theta[g(X^{n+1})]} + 2\sqrt{\hat{V}_\theta},
$$

which implies $(\sqrt{n\frac{d^2\phi}{d\theta^2}(\theta)} - 2\sqrt{\hat{V}_\theta})^2 \leq V_\theta[g(X^{n+1})]$. Then we obtain the first inequality because $n\frac{d^2\phi}{d\theta^2}(\theta)(1 - 2\sqrt{\hat{V}_\theta/n\frac{d^2\phi}{d\theta^2}(\theta)})^2 = (\sqrt{n\frac{d^2\phi}{d\theta^2}(\theta)} - 2\sqrt{\hat{V}_\theta})^2$. Similarly, since $\|g^n(X^{n+1}) - n\frac{d\phi}{d\theta}(\theta) + \frac{d}{d\theta}\log\overline{P}_\theta^3(X_{n+1}) - \frac{d}{d\theta}\log\overline{P}_\theta^3(X_1)\|_2 \geq$

$\|g^n(X^{n+1}) - n\frac{d\phi}{d\theta}(\theta)\|_2 - \|\frac{d}{d\theta}\log\overline{P}_\theta^3(X_{n+1})\|_2 - \|\frac{d}{d\theta}\log\overline{P}_\theta^3(X_1)\|_2$, we obtain the second inequality because $\frac{d^2\phi}{d\theta^2}(\theta)(1 + 2\sqrt{\hat{V}_\theta/n\frac{d^2\phi}{d\theta^2}(\theta)})^2 = (\sqrt{n\frac{d^2\phi}{d\theta^2}(\theta)} + 2\hat{V}_\theta)^2$.

$\square$

7. Nonstationary $n + 1$-observation case. Similar to the previous section, this section also discusses the one-parameter case. Now, we consider the non-stationary case. Since the convergence to the stationary distribution is required, we assume that the transition matrices W_θ are ergodic as well as irreducible. Then we fix an arbitrary initial distributions P_θ on $\mathcal{X}$ such that the distribution P_θ is smoothly parameterized by the parameter θ. In this section, we assume that W_θ is the exponential family generated by the generator $g(x, x')$ and the random variable $X^{n+1} := (X_{n+1}, \ldots, X_1)$ is subject to $W_\theta^{\times n} \times P_\theta$ with the unknown parameter θ. Then we denote the expectation and the variance under the distribution $W_\theta^{\times n} \times P_\theta$ by E_θ and V_θ. In this general case, the relation (6.4) does not hold. Instead of these relations, as is shown in [12], Lemma 5.4, we have

$$(7.1) \qquad \lim_{n\to\infty} \frac{1}{n} D(W_\theta^{\times n} \times P_\theta \| W_{\theta'}^{\times n} \times P_{\theta'}) = D(W_\theta \| W_{\theta'}),$$

$$(7.2) \qquad \lim_{n\to\infty} \frac{1}{n} D_{1+s}(W_\theta^{\times n} \times P_\theta \| W_{\theta'}^{\times n} \times P_{\theta'}) = D_{1+s}(W_\theta \| W_{\theta'}).$$

For a function h on $\mathbb{R}$, we define the random variable $\tilde{g}^n(X^{n+1}) := g^n(X_{n+1}) + h(X_1)$. When we use the random variable $\tilde{g}^n(X^{n+1})/n$ as an estimator of the parameter $\eta(\theta)$, the error is measured by the mean square error:

$$(7.3) \qquad \mathsf{MSE}_\theta[\tilde{g}^n(X^{n+1})] := \mathsf{E}_\theta\left[\left(\frac{\tilde{g}^n(X^{n+1})}{n} - \eta(\theta)\right)^2\right].$$

Then we have $\mathsf{E}_\theta[\tilde{g}^n(X^{n+1})] = \mathsf{E}_\theta[g^n(X^{n+1})] + \mathsf{E}_\theta[h(X_1)]$. In the following discussion, we employ the norm $\|f(X^{n+1})\|_2 := \sqrt{\mathsf{E}_\theta[f(X^{n+1})^2]}$ for a function f on $\mathbb{R}^{n+1}$. Using the triangle inequality for this norm, we have

$$(7.4) \quad \sqrt{\mathsf{V}_\theta[g^n(X^{n+1})]} - \sqrt{\mathsf{V}_\theta[h(X_1)]} \le \sqrt{\mathsf{V}_\theta[\tilde{g}^n(X^{n+1})]}$$
$$\le \sqrt{\mathsf{V}_\theta[g^n(X^{n+1})]} + \sqrt{\mathsf{V}_\theta[h(X_1)]},$$

and

$$\sqrt{\mathsf{E}_\theta\left[\left(\frac{g^n(X^{n+1})}{n} - \mathsf{E}_\theta\left[\frac{g^n(X^{n+1})}{n}\right]\right)^2\right]}$$
$$- \sqrt{\mathsf{E}_\theta\left[\left(\frac{h(X_1)}{n} + \mathsf{E}_\theta\left[\frac{g^n(X^{n+1})}{n}\right] - \eta(\theta)\right)^2\right]}$$

$$(7.5) \qquad \leq \sqrt{\mathsf{E}_\theta\left[\left(\frac{\tilde{g}^n(X^{n+1})}{n} - \eta(\theta)\right)^2\right]}$$

$$\leq \sqrt{\mathsf{E}_\theta\left[\left(\frac{g^n(X^{n+1})}{n} - \mathsf{E}_\theta\left[\frac{g^n(X^{n+1})}{n}\right]\right)^2\right]}$$

$$+ \sqrt{\mathsf{E}_\theta\left[\left(\frac{h(X_1)}{n} + \mathsf{E}_\theta\left[\frac{g^n(X^{n+1})}{n}\right] - \eta(\theta)\right)^2\right]}.$$

It is known that for a given compact subset $U \subset \mathbb{R}$, there exist constants $1 > \rho > 0$ and $\gamma > 0$ such that

$$(7.6) \qquad \left|W_\theta^n(x|x') - P_\theta\right| < \gamma\rho^n$$

for $\theta \in U$ [9], page 173. That is, the distribution of X_n converges to the stationary distribution in terms of variational distance, which is compact uniform with respect to θ. So, we find that the expectation $\mathsf{E}_\theta[\frac{g^n(X^{n+1})}{n}]$ and the variance $\mathsf{V}_\theta[\frac{g^n(X^{n+1})}{\sqrt{n}}]$ converge to those under the stationary distribution, whose convergences are compact uniform with respect to θ. This is because $\max_{x^{n+1}} \frac{g^n(x^{n+1})}{n}\gamma\rho^n$ and $\max_{x^{n+1}} (\frac{g^n(x^{n+1})}{\sqrt{n}})^2\gamma\rho^n$ go to zero. Hence, we have

$$(7.7) \qquad \lim_{n\to\infty} \mathsf{E}_\theta\left[\frac{\tilde{g}^n(X^{n+1})}{n}\right] = \lim_{n\to\infty} \mathsf{E}_\theta\left[\frac{g^n(X^{n+1})}{n}\right] \overset{(a)}{=} \eta(\theta) = \frac{d\phi}{d\theta}(\theta),$$

$$\lim_{n\to\infty} n\mathsf{MSE}_\theta\left[\frac{\tilde{g}^n(X^{n+1})}{n}\right] \overset{(b)}{=} \lim_{n\to\infty} \mathsf{V}_\theta\left[\frac{\tilde{g}^n(X^{n+1})}{\sqrt{n}}\right]$$

$$(7.8) \qquad \overset{(c)}{=} \lim_{n\to\infty} \mathsf{V}_\theta\left[\frac{g^n(X^{n+1})}{\sqrt{n}}\right] \overset{(d)}{=} \frac{d^2\phi}{d\theta^2}(\theta),$$

where (a), (b), (c) and (d) follow from (6.2), (7.5), (7.4) and (6.5), respectively. The above convergences are compact uniform with respect to θ. The relation (7.7) shows that the estimator $\frac{\tilde{g}^n(X^{n+1})}{n}$ is asymptotically unbiased for the parameter η. The mean square error is $\frac{d^2\phi}{d\theta^2}(\theta)\frac{1}{n} + o(\frac{1}{n})$, which implies (2.2). Further, it is shown that the random variable $\sqrt{n}(\frac{g^n(X^{n+1})}{n} - \eta(0))$ asymptotically obeys the Gaussian distribution with the variance $\frac{d^2\phi}{d\theta^2}(0)$ at $\theta = 0$ [12], Corollary 6.2. Replacing W_0 by W_θ, we find that the random variable $\sqrt{n}(\frac{g^n(X^{n+1})}{n} - \eta(\theta))$ asymptotically obeys the Gaussian distribution with the variance $\frac{d^2\phi}{d\theta^2}(\theta)$.

Next, for the family $\{W_\theta^{\times n} \times P_\theta\}_\theta$, we consider the Fisher information J_θ^n for the natural parameter θ and the Fisher information $\tilde{J}_\theta^n$ for the expectation parameter η.

LEMMA 7.1. *The limit of the Fisher information J_θ^n for the natural parameter θ is characterized as*

$$
(7.9) \qquad \lim_{n \to \infty} \frac{J_\theta^n}{n} = \frac{d^2\phi}{d\theta^2}(\theta).
$$

Hence, the limit of the Fisher information $\tilde{J}_\theta^n$ for the expectation parameter η is characterized as $\lim_{n \to \infty} \frac{\tilde{J}_\theta^n}{n} = \frac{d^2\phi}{d\theta^2}(\theta)^{-1}$.

Lemma 7.1 implies that the lower bound of the Cramér–Rao inequality is $\frac{d^2\phi}{d\theta^2}(\theta(\eta))\frac{1}{n} + o(\frac{1}{n})$. Therefore, the estimator $\frac{\tilde{g}^n(X^{n+1})}{n}$ attains the lower bound by the Cramér–Rao inequality with the order $\frac{1}{n}$. That is, the estimator $\frac{\tilde{g}^n(X^{n+1})}{n}$ is asymptotically efficient.

PROOF OF LEMMA 7.1.   Similar to (C.2), we have

$$
\begin{aligned}
J_\theta^n = \mathsf{E}_\theta\Bigg[\Bigg(&-n\frac{d\phi}{d\theta}(\theta) + \frac{d}{d\theta}\log \overline{P}_\theta^3(X_{n+1}) \\
&- \frac{d}{d\theta}\log \overline{P}_\theta^3(X_1) + g^n(X^{n+1})\Bigg)^2\Bigg] + J_\theta^1 \\
= \Bigg\|&-n\frac{d\phi}{d\theta}(\theta) + \frac{d}{d\theta}\log \overline{P}_\theta^3(X_{n+1}) \\
&- \frac{d}{d\theta}\log \overline{P}_\theta^3(X_1) + g^n(X^{n+1})\Bigg\|_2^2 + J_\theta^1.
\end{aligned}
$$
(7.10)

Since (7.7) and (7.8) yield that $\frac{1}{n}\|g^n(X^{n+1}) - n\frac{d\phi}{d\theta}(\theta)\|_2^2 \to \frac{d^2\phi}{d\theta^2}(\theta)$, we have

$$
\begin{aligned}
\frac{1}{\sqrt{n}}&\Bigg\|-n\frac{d\phi}{d\theta}(\theta) + \frac{d}{d\theta}\log \overline{P}_\theta^3(X_{n+1}) - \frac{d}{d\theta}\log \overline{P}_\theta^3(X_1) + g^n(X^{n+1})\Bigg\|_2 \\
&\leq \frac{1}{\sqrt{n}}\Bigg(\Bigg\|g^n(X^{n+1}) - n\frac{d\phi}{d\theta}(\theta)\Bigg\|_2 + \Bigg\|\frac{d}{d\theta}\log \overline{P}_\theta^3(X_{n+1})\Bigg\|_2 \\
&\quad + \Bigg\|\frac{d}{d\theta}\log \overline{P}_\theta^3(X_1)\Bigg\|_2\Bigg) \\
&\leq \frac{1}{\sqrt{n}}\Bigg\|g^n(X^{n+1}) - n\frac{d\phi}{d\theta}(\theta)\Bigg\|_2 + \frac{2}{\sqrt{n}}\max_x\Bigg|\frac{d}{d\theta}\log \overline{P}_\theta^3(x)\Bigg| \\
&\to \sqrt{\frac{d^2\phi}{d\theta^2}(\theta)}.
\end{aligned}
$$
(7.11)

The combination of (7.10) and (7.11) yields that $\lim_{n \to \infty} \frac{J_\theta^n}{n} \leq \frac{d^2\phi}{d\theta^2}(\theta)$. Similarly, the opposite inequality can be shown by replacing the role of (7.11) by the follow-

ing inequality:

$$\frac{1}{\sqrt{n}}\left\| -n\frac{d\phi}{d\theta}(\theta) + \frac{d}{d\theta}\log\overline{P}_\theta^3(X_{n+1}) - \frac{d}{d\theta}\log\overline{P}_\theta^3(X_1) + g^n(X^{n+1}) \right\|_2$$

$$\geq \frac{1}{\sqrt{n}}\left\| g^n(X^{n+1}) - n\frac{d\phi}{d\theta}(\theta) \right\|_2 - \frac{2}{\sqrt{n}}\max_x\left| \frac{d}{d\theta}\log\overline{P}_\theta^3(x) \right| \to \sqrt{\frac{d^2\phi}{d\theta^2}(\theta)}.$$

Hence, we obtain (7.9). Since $\frac{d\theta}{d\eta}(\theta) = \frac{d^2\phi}{d\theta^2}(\theta)^{-1}$, (7.9) implies $\lim_{n\to\infty}\frac{\tilde{J}_\theta^n}{n} = \frac{d^2\phi}{d\theta^2}(\theta)^{-1}$. $\square$

8. Estimation with multiparameter case.

8.1. *Estimation with multiparameter exponential family*: *Stationary case.*   Assume that $W_{\vec{\theta}}$ is a multiparameter exponential family of transition matrices with $\vec{\theta} = (\theta^1, \ldots, \theta^d)$ with the generator $\{g_j\}$. Then we assume that the initial distribution is the stationary distribution $\overline{P}_{\vec{\theta}}^1$ on $\mathcal{X}$ of $W_{\vec{\theta}}$ and the random variable $X^{n+1} := (X_{n+1}, \ldots, X_1)$ is subject to $W_{\vec{\theta}}^{\times n} \times \overline{P}_{\vec{\theta}}^1$ with the unknown parameter $\vec{\theta}$. In this subsection, we denote the expectation and the variance under the distribution $W_{\vec{\theta}}^{\times n} \times \overline{P}_{\vec{\theta}}^1$ by $\mathsf{E}_{\vec{\theta}}$ and $\mathsf{V}_{\vec{\theta}}$.

Similar to (6.2), using $\vec{g}^n(X^{n+1}) := [g_j^n(X^{n+1})]_j$, we can show that

$$(8.1) \qquad \mathsf{E}_{\vec{\theta}}\left[\frac{\vec{g}^n(X^{n+1})}{n} \right] = \vec{\eta}(\vec{\theta}),$$

which implies that $\vec{g}^n(X^{n+1})$ is an unbiased estimator of the expectation parameter $\vec{\eta}(\vec{\theta})$. We denote the covariance matrix of $\vec{g}^n(X^{n+1})$ by $\mathsf{Cov}_\theta[\vec{g}^n(X^{n+1})]$. We also denote the covariance matrix of $[\frac{\partial}{\partial\theta^j}\log\overline{P}_\theta^3(X)]_j$ by $\hat{\mathsf{Cov}}_\theta$.

LEMMA 8.1.   *The matrix inequalities*

$$n\mathsf{H}_{\vec{\theta}}[\phi]\left(1 - 2\sqrt{\frac{\|\mathsf{H}_{\vec{\theta}}[\phi]^{-1/2}\hat{\mathsf{Cov}}_\theta\mathsf{H}_{\vec{\theta}}[\phi]^{-1/2}\|}{n}} \right)^2$$

$$(8.2) \qquad \leq \mathsf{Cov}_\theta[\vec{g}^n(X^{n+1})]$$

$$\leq n\mathsf{H}_{\vec{\theta}}[\phi]\left(1 + 2\sqrt{\frac{\|\mathsf{H}_{\vec{\theta}}[\phi]^{-1/2}\hat{\mathsf{Cov}}_\theta\mathsf{H}_{\vec{\theta}}[\phi]^{-1/2}\|}{n}} \right)^2$$

hold, *where the matrix inequality is defined by the positive semi-definiteness.*

PROOF.   First, we fix a real unit vector $\vec{a} = [a_j]_j$. Applying (6.5) to the random variable $\sum_j a_j g_j^n(X^{n+1})$, we obtain

$$n\vec{a}^T \mathsf{H}_{\vec{\theta}}[\phi]\vec{a}\left(1 - 2\sqrt{\frac{\vec{a}^T \hat{\mathsf{Cov}}_{\theta}\vec{a}}{n\vec{a}^T \mathsf{H}_{\vec{\theta}}[\phi]\vec{a}}}\,\right)^2$$

(8.3)
$$\leq \vec{a}^T \mathsf{Cov}_{\theta}[\vec{g}^n(X^{n+1})]\vec{a}$$

$$\leq n\vec{a}^T \mathsf{H}_{\vec{\theta}}[\phi]\vec{a}\left(1 + 2\sqrt{\frac{\vec{a}^T \hat{\mathsf{Cov}}_{\theta}\vec{a}}{n\vec{a}^T \mathsf{H}_{\vec{\theta}}[\phi]\vec{a}}}\,\right)^2.$$

Since $\frac{\vec{a}^T \hat{\mathsf{Cov}}_{\theta}\vec{a}}{\vec{a}^T \mathsf{H}_{\vec{\theta}}[\phi]\vec{a}} \leq \|\mathsf{H}_{\vec{\theta}}[\phi]^{-1/2}\hat{\mathsf{Cov}}_{\theta}\mathsf{H}_{\vec{\theta}}[\phi]^{-1/2}\|$, (8.3) implies (8.2).   □

Lemma 8.1 yields that

(8.4) $\mathsf{Cov}_{\vec{\theta}}\left[\dfrac{\vec{g}^n(X^{n+1})}{n}\right] = \dfrac{\mathsf{Cov}_{\vec{\theta}}[\vec{g}^n(X^{n+1})]}{n^2} = \dfrac{\mathsf{H}_{\vec{\theta}}[\phi]}{n} + o\left(\dfrac{1}{n}\right).$

Now, we denote the Fisher information matrix of the distribution family $\{\overline{P}_{\vec{\theta}}^1\}_{\vec{\theta}}$ by $J_{\vec{\theta}}^1$. The Fisher information matrix $\tilde{J}_{\vec{\eta}(\vec{\theta})}^{n+1}$ for the expectation parameter $\vec{\eta}$ of the distribution family $\{W_{\vec{\theta}}^{\times n} \times \overline{P}_{\vec{\theta}}^1\}_{\vec{\theta}}$ is

$$\tilde{J}_{\vec{\eta}(\vec{\theta})}^{n+1} = \left(\left[\frac{\partial \eta_i(\vec{\theta})}{\partial \theta_j}\right]_{i,j}^T\right)^{-1} J_{\vec{\theta}}^{n+1}\left(\left[\frac{\partial \eta_i(\vec{\theta})}{\partial \theta_j}\right]_{i,j}\right)^{-1}$$

$$= \mathsf{H}_{\vec{\theta}}[\phi]^{-1}\big(n\mathsf{H}_{\vec{\theta}}[\phi] + J_{\vec{\theta}}^1\big)\mathsf{H}_{\vec{\theta}}[\phi]^{-1}$$

$$= \mathsf{H}_{\vec{\theta}}[\phi]^{-1/2}\big(nI + \mathsf{H}_{\vec{\theta}}[\phi]^{-1/2}J_{\vec{\theta}}^1\mathsf{H}_{\vec{\theta}}[\phi]^{-1/2}\big)\mathsf{H}_{\vec{\theta}}[\phi]^{-1/2}.$$

That is, the lower bound of the variance of the unbiased estimator given by Cramér–Rao inequality is $\frac{1}{n}\mathsf{H}_{\vec{\theta}}[\phi]^{1/2}(1I + \frac{1}{n}\mathsf{H}_{\vec{\theta}}[\phi]^{-1/2}J_{\vec{\theta}}^1\mathsf{H}_{\vec{\theta}}[\phi]^{-1/2})^{-1}\mathsf{H}_{\vec{\theta}}[\phi]^{1/2}$, that is, the Cramér–Rao inequality is given as

$$\mathsf{Cov}_{\vec{\theta}}\left[\frac{\vec{g}^n(X^{n+1})}{n}\right]$$

(8.5)
$$\geq \frac{1}{n}\mathsf{H}_{\vec{\theta}}[\phi]^{1/2}\left(I + \frac{1}{n}\mathsf{H}_{\vec{\theta}}[\phi]^{-1/2}J_{\vec{\theta}}^1\mathsf{H}_{\vec{\theta}}[\phi]^{-1/2}\right)^{-1}\mathsf{H}_{\vec{\theta}}[\phi]^{1/2}$$

$$= \frac{1}{n}\mathsf{H}_{\vec{\theta}}[\phi] + O\left(\frac{1}{n^2}\right).$$

The relation (8.4) shows that the unbiased estimator $\frac{\vec{g}^n(X^{n+1})}{n}$ realizes the optimal performance with the order $\frac{1}{n}$.

Therefore, we obtain an asymptotically efficient estimator for the expectation parameter. To estimate the natural parameter, we need to solve the equation

$$(8.6) \qquad \eta_j = \frac{\partial \phi}{\partial \theta^j}(\vec{\theta})$$

for $\vec{\theta}$. Since the function $\phi(\vec{\theta})$ is strictly convex, $\vec{\theta}(\vec{\eta})$ can be derived by the maximization of the concave function as

$$(8.7) \qquad \underset{\vec{\theta}}{\operatorname{argmax}} \ \vec{\eta} \cdot \vec{\theta} - \phi(\vec{\theta}).$$

The calculation complexity does not depend on the number n of data. Hence, when the number d of parameters is not so large, the natural parameter can be estimated efficiently even with a large number n of data.

However, the conventional algorithm for the maximization of the concave function [5] requires the calculation of the derivative. Since the convex function $\phi(\vec{\theta})$ is given as the logarithm of the Perron–Frobenius eigenvalue of the matrix $\overline{W}_\theta$, the calculation of the derivative is not so easy. To overcome this kind of difficulty, we can employ derivative-free optimization algorithms [7, 23] represented by Nelder–Mead method [29]. A derivative-free optimization algorithm maximizes a concave function without calculating the derivative only with calculating the outcomes with several inputs. In particular, it is expected that such an algorithm enables us to numerically derive $\vec{\theta}(\vec{\eta})$ for a given $\vec{\eta}$.

8.2. *Estimation with multiparameter exponential family*: *Nonstationary case.* Next, similar to Section 7, we consider the nonstationary case and assume that the transition matrices $W_{\vec{\theta}}$ are ergodic as well as irreducible. Then we fix an arbitrary initial distributions $P_{\vec{\theta}}$ on $\mathcal{X}$ such that the distribution $P_{\vec{\theta}}$ is smoothly parameterized by the natural parameter $\vec{\theta}$. This assumption contains the special case when the distribution $P_{\vec{\theta}}$ does not depend on the parameter $\vec{\theta}$.

In this subsection, we denote the expectation, the variance, and the covariance matrix under the distribution $W_{\vec{\theta}}^{\times n} \times P_{\vec{\theta}}$ by $\mathsf{E}_{\vec{\theta}}$, $\mathsf{V}_{\vec{\theta}}$ and $\mathsf{Cov}_{\vec{\theta}}$. Then we employ the random variable $\vec{g}^n(X^{n+1}) := (g_j^n(X^{n+1}))$. When we use the random variable $\vec{g}^n(X^{n+1})/n$ as an estimator of the parameter $\vec{\theta}$, the error is measured by the mean square error matrix:

$$\mathsf{MSE}_\theta \left[\frac{\vec{g}^n(X^{n+1})}{n} \right]_{i,j} := \mathsf{E}_\theta \left[\left(\frac{g_i^n(X^{n+1})}{n} - \eta_i(\vec{\theta}) \right) \left(\frac{g_j^n(X^{n+1})}{n} - \eta_j(\vec{\theta}) \right) \right].$$

Similar to (7.7), we can show that

$$(8.8) \qquad \lim_{n \to \infty} \mathsf{E}_{\vec{\theta}} \left[\frac{\vec{g}^n(X^{n+1})}{n} \right] = \vec{\eta}(\vec{\theta}) = \left[\frac{\partial \phi}{\partial \theta^j}(\vec{\theta}) \right]_j.$$

For any vector $\vec{c} = (c_i)$, the application of (7.8) to $\theta = \vec{c} \cdot \vec{\theta}$ implies that

$$
\lim_{n \to \infty} n \vec{c}^T \mathsf{MSE}_{\vec{\theta}}\left[\frac{\vec{g}^n(X^{n+1})}{n}\right]\vec{c} = \lim_{n \to \infty} n \vec{c}^T \mathsf{Cov}_{\vec{\theta}}\left[\frac{\vec{g}^n(X^{n+1})}{n}\right]\vec{c}
$$

(8.9)
$$
= \vec{c}^T \mathsf{H}_{\vec{\theta}}[\phi]\vec{c}.
$$

Here, the convergences in (8.8) and (8.9) are compact uniform with respect to $\vec{\theta}$ because (7.6) can be extended to multiparametric case. Hence, combining Lemma 8.1, we obtain the following theorem in the same way as (7.7) and (7.8).

THEOREM 8.2. *The compact uniform convergences with respect to $\vec{\theta}$*

$$
(8.10) \quad \lim_{n \to \infty} n\mathsf{MSE}_{\vec{\theta}}\left[\frac{\vec{g}^n(X^{n+1})}{n}\right] = \lim_{n \to \infty} n\mathsf{Cov}_{\vec{\theta}}\left[\frac{\vec{g}^n(X^{n+1})}{n}\right] = \mathsf{H}_{\vec{\theta}}[\phi]
$$

hold.

The relation (8.8) shows that the estimator $\frac{\vec{g}^n(X^{n+1})}{n}$ is asymptotically unbiased for the expectation parameter $\vec{\eta}$. The above theorem implies that the mean square error is $\frac{1}{n}\mathsf{H}_{\vec{\theta}}[\phi] + o(\frac{1}{n})$.

Next, for the family $\{W_{\vec{\theta}}^{\times n} \times P_{\vec{\theta}}\}_{\vec{\theta}}$, we consider the Fisher information matrix $J_{\vec{\theta}}^n$ for the natural parameter $\vec{\theta}$ and the Fisher information matrix $\tilde{J}_{\vec{\theta}}^n$ for the expectation parameter $\vec{\eta}$.

LEMMA 8.3. *The limit of the Fisher information matrix $J_{\vec{\theta}}^n$ for the natural parameter $\vec{\theta}$ is characterized as $\lim_{n \to \infty} \frac{J_{\vec{\theta}}^n}{n} = \mathsf{H}_{\vec{\theta}}[\phi]$. Hence, the limit of the Fisher information matrix $\tilde{J}_{\vec{\theta}}^n$ for the expectation parameter $\vec{\eta}$ is characterized as $\lim_{n \to \infty} \frac{\tilde{J}_{\vec{\theta}}^n}{n} = \mathsf{H}_{\vec{\theta}}[\phi]^{-1}$.*

PROOF. We fix a real unit vector $\vec{a} = [a_j]_j$. The application of the relation (7.9) to $\theta = \vec{c} \cdot \vec{\theta}$ yields that $\lim_{n \to \infty} \frac{\vec{a}^T J_{\vec{\theta}}^n \vec{a}}{n} = \vec{a}^T \mathsf{H}_{\vec{\theta}}[\phi]\vec{a}$, which implies $\lim_{n \to \infty} \frac{J_{\vec{\theta}}^n}{n} = \mathsf{H}_{\vec{\theta}}[\phi]$. Since $[\frac{\partial \eta_i(\vec{\theta})}{\partial \theta_j}]_{i,j}$ is $\mathsf{H}_{\vec{\theta}}[\phi]$, we obtain $\lim_{n \to \infty} \frac{\tilde{J}_{\vec{\theta}}^n}{n} = \mathsf{H}_{\vec{\theta}}[\phi]^{-1}$. $\square$

Lemma 8.3 implies that the lower bound of the Cramér–Rao inequality is $\frac{1}{n}\mathsf{H}_{\vec{\theta}}[\phi] + o(\frac{1}{n})$. Therefore, the estimator $\frac{\vec{g}^n(X^{n+1})}{n}$ attains the lower bound by the Cramér–Rao inequality with the order $\frac{1}{n}$. That is, the estimator $\frac{\vec{g}^n(X^{n+1})}{n}$ is asymptotically efficient.

Similar to the one-parameter case, we can show that the random variable $\sqrt{n}(\frac{\vec{g}^n(X^{n+1})}{n} - \vec{\eta}(\vec{\theta}))$ converges to the Gaussian distribution with the covariance matrix $\mathsf{H}_{\vec{\theta}}[\phi]$.

8.3. *Estimation with multiparameter curved exponential family.* Next, we proceed to estimation with a multiparameter curved exponential family. Given a d-dimensional exponential family $\mathcal{E} = \{W_{\vec{\theta}}\}_{\vec{\theta} \in \Theta}$, we choose an injective smooth function $\vec{\theta}_{\mathrm{CRV}}$ from a subset $\Xi \subset \mathbb{R}^{d'}$ to a subset $\Theta \subset \mathbb{R}^d$. Then a d'-parameter subset $\tilde{\mathcal{E}} = \{W_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi})}\}_{\vec{\xi}}$ of transition matrices is called a *curved exponential family* of transition matrices. For example, a mixture family defined in Section 4.2 is also a curved exponential family. As explained in Example 1, the set of all positive transition matrices on a finite-size system forms an exponential family. Hence, any smooth subfamily of transition matrices on a finite-size system forms a curved exponential family. Then we define the Fisher information matrix $\tilde{\mathsf{H}}_{\vec{\xi}}$ as the metric of the submanifold. Assume that the Jacobian matrix $A := [\frac{\partial \eta_i}{\partial \xi_j}|_{\vec{\xi}=\vec{\xi}_o}]_{i,j}$ has the rank d'. When the potential function of the exponential family is ϕ, the Fisher information matrix is written as $\tilde{\mathsf{H}}_{\vec{\xi}_o} = A^T \mathsf{H}_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}_o)}[\phi]^{-1} A$ because the Fisher information matrix for the expectation parameter η at $\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}_o)$ is $\mathsf{H}_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}_o)}[\phi]^{-1}$.

In the following, we assume that the exponential family $\mathcal{E}$ is generated by g_j. Given $n + 1$ observations X^{n+1}, as Figure 1, we define the estimator $\vec{\xi}^n(X^{n+1}) := \mathrm{argmin}_{\vec{\xi}} D(W_{\vec{\theta}(\frac{\vec{g}^n(X^{n+1})}{n})} \| W_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi})})$ for the curved exponential family $\tilde{\mathcal{E}}$. Then, similar to the case of a curved exponential family of probability distributions [2], Section 4.4, we can show that the estimator $\vec{\xi}^n(X^{n+1})$ is asymptotically efficient. That is, the mean square error matrix is asymptotically approximated to $\frac{1}{n}\tilde{\mathsf{H}}_{\vec{\xi}}[\phi]^{-1} + o(\frac{1}{n})$ as follows.

THEOREM 8.4. *The random variable $\vec{\xi}^n(X^{n+1}) - \vec{\xi}_o$ asymptotically obeys the Gaussian distribution with the covariance matrix $\frac{1}{n}\tilde{\mathsf{H}}_{\vec{\xi}_o}[\phi]^{-1}$. Also, when the set Ξ is bounded, the mean square error matrix of our estimator $\vec{\xi}^n(X^{n+1})$ is asymptotically approximated to $\frac{1}{n}\tilde{\mathsf{H}}_{\vec{\xi}}[\phi]^{-1} + o(\frac{1}{n})$.*

PROOF. *One-parameter case*: For simplicity, we first consider the case with $d' = 1$. When ξ is close to ξ_o, $\vec{\eta}_{\mathrm{CRV}}(\xi)$ is $\vec{\eta}_o + \frac{d\vec{\eta}(\vec{\theta}_{\mathrm{CRV}})}{d\xi}|_{\xi=\xi_o}(\xi - \xi_o) + o(\xi - \xi_o)$,
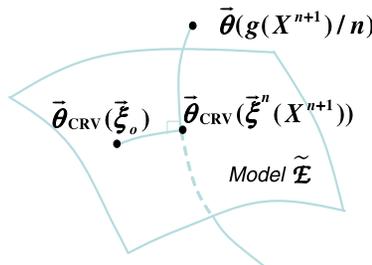

FIG. 1. *Estimator for the curved exponential family.*

where $\vec{\eta}_o := \vec{\eta}(\vec{\theta}_{\text{CRV}}(\xi_o))$ and $\vec{\eta}_{\text{CRV}}(\xi) := \vec{\eta}(\vec{\theta}_{\text{CRV}}(\xi))$. Hence, given an observed value $\frac{\vec{g}^n(X^{n+1})}{n}$ close to the true parameter $\vec{\eta}_o$, we have

$$\xi^n(X^{n+1}) = \xi_o + \frac{\langle \frac{\vec{g}^n(X^{n+1})}{n} - \vec{\eta}_o, \frac{d\vec{\eta}_{\text{CRV}}}{d\xi}|_{\xi=\xi_o} \rangle_{\xi_o}}{\langle \frac{d\vec{\eta}_{\text{CRV}}}{d\xi}|_{\xi=\xi_o}, \frac{d\vec{\eta}_{\text{CRV}}}{d\xi}|_{\xi=\xi_o} \rangle_{\xi_o}}$$

(8.11)
$$+ o\left(\left\| \frac{\vec{g}^n(X^{n+1})}{n} - \vec{\eta}_o \right\|\right),$$

where $\langle v, u \rangle_{\xi_o} := \sum_{i,j} (\mathsf{H}_{\vec{\theta}_{\text{CRV}}(\vec{\xi}_o)}[\phi]^{-1})_{i,j} v_i u_j$ because the minimization $\min_{\xi'} D(W_{\vec{\theta}(\frac{\vec{g}^n(X^{n+1})}{n})} \| W_{\vec{\theta}_{\text{CRV}}(\xi'+\xi_o)})$ can be approximated to

$$\min_{\xi'} \frac{1}{2} \left\langle \frac{\vec{g}^n(X^{n+1})}{n} - \left(\vec{\eta}_o + \frac{d\vec{\eta}_{\text{CRV}}}{d\xi}\bigg|_{\xi=\xi_o} \xi' \right), \right.$$

(8.12)
$$\left. \frac{\vec{g}^n(X^{n+1})}{n} - \left(\vec{\eta}_o + \frac{d\vec{\eta}_{\text{CRV}}}{d\xi}\bigg|_{\xi=\xi_o} \xi' \right) \right\rangle_{\xi_o}.$$

That is, $\xi^n(X^{n+1}) - \xi_o$ can be approximated by $\frac{\langle \frac{\vec{g}^n(X^{n+1})}{n} - \vec{\eta}_o, \frac{d\vec{\eta}_{\text{CRV}}}{d\xi}|_{\xi=\xi_o} \rangle_{\xi_o}}{\langle \frac{d\vec{\eta}_{\text{CRV}}}{d\xi}|_{\xi=\xi_o}, \frac{d\vec{\eta}_{\text{CRV}}}{d\xi}|_{\xi=\xi_o} \rangle_{\xi_o}}$ when $\frac{\vec{g}^n(X^{n+1})}{n}$ is close to the true parameter $\vec{\eta}_o := \vec{\eta}(\vec{\theta}_{\text{CRV}}(\xi_o))$. Since

$$\lim_{n\to\infty} n\mathrm{E}_{\vec{\theta}_{\text{CRV}}(\xi_o)}\left[\left(\left\langle \frac{\vec{g}^n(X^{n+1})}{n} - \vec{\eta}_o, \frac{d\vec{\eta}_{\text{CRV}}}{d\xi}\bigg|_{\xi=\xi_o} \right\rangle_{\xi_o} \right)^2 \right]$$

$$= \left\langle \frac{d\vec{\eta}_{\text{CRV}}}{d\xi}\bigg|_{\xi=\xi_o}, \frac{d\vec{\eta}_{\text{CRV}}}{d\xi}\bigg|_{\xi=\xi_o} \right\rangle_{\xi_o}$$

$$= \tilde{\mathsf{H}}_{\xi_o}[\phi],$$

we have

$$n\left\langle \frac{d\vec{\eta}_{\text{CRV}}}{d\xi}\bigg|_{\xi=\xi_o}, \frac{d\vec{\eta}_{\text{CRV}}}{d\xi}\bigg|_{\xi=\xi_o} \right\rangle_{\xi_o}^{-2}$$

(8.13)
$$\times \mathrm{E}_{\vec{\theta}_{\text{CRV}}(\xi_o)}\left[\left(\left\langle \frac{\vec{g}^n(X^{n+1})}{n} - \vec{\eta}_o, \frac{d\vec{\eta}_{\text{CRV}}}{d\xi}\bigg|_{\xi=\xi_o} \right\rangle_{\xi_o} \right)^2 \right]$$

$$\to \tilde{\mathsf{H}}_{\xi_o}[\phi]^{-1}.$$

Since the probability that this approximation asymptotically does not hold approaches zero, we can apply the central limit theorem for Markovian process [15, 18, 24] to $\langle \frac{\vec{g}^n(X^{n+1})}{n} - \vec{\eta}_o, \frac{d\vec{\eta}_{\text{CRV}}}{d\xi}|_{\xi=\xi_o} \rangle_{\xi_o}$. So, we find that the random variable $\xi^n(X^{n+1}) - \xi_o$ asymptotically obeys the Gaussian distribution with the covariance

$\frac{1}{n}\tilde{\mathsf{H}}_{\xi_o}[\phi]^{-1}$. The above discussion shows only the weak convergence to the Gaussian distribution.

To show the second statement, we need more careful discussion. Rewriting the argument (8.11) in a more precise form, we obtain the following argument. For any small real number $\varepsilon > 0$, there exists $\delta > 0$ satisfying the following condition. When

$$
(8.14) \qquad \left\| \frac{\vec{g}^n(X^{n+1})}{n} - \vec{\eta}_o \right\| < \delta,
$$

we have

$$
(8.15) \qquad \begin{aligned}
\xi^n(X^{n+1}) - \xi_o &= \frac{\langle \frac{\vec{g}^n(X^{n+1})}{n} - \vec{\eta}_o, \frac{d\vec{\eta}_{\mathrm{CRV}}}{d\xi}|_{\xi=\xi_o} \rangle_{\xi_o}}{\langle \frac{d\vec{\eta}_{\mathrm{CRV}}}{d\xi}|_{\xi=\xi_o}, \frac{d\vec{\eta}_{\mathrm{CRV}}}{d\xi}|_{\xi=\xi_o} \rangle_{\xi_o}} \\
&\quad + \varepsilon t(X^{n+1}) \left\| \frac{\vec{g}^n(X^{n+1})}{n} - \vec{\eta}_o \right\|,
\end{aligned}
$$

where $|t(X^{n+1})| \leq 1$.

In the following, we denote the set $\{x^{n+1} \,|\, \|\frac{\vec{g}^n(x^{n+1})}{n} - \vec{\eta}_o\| < \delta\}$ by $\Omega_\delta^{(n)}$. Then, we denote the conditional expectation under the condition (8.14) by $\mathrm{E}_{\vec{\theta}_{\mathrm{CRV}}(\xi_o)|\Omega_\delta^{(n)}}$. Hence, using (8.13) we have

$$
(8.16) \qquad \begin{aligned}
C_{1,\delta}^{(n)} &:= n\mathrm{P}_{\vec{\theta}_{\mathrm{CRV}}(\xi_o)}(\Omega_\delta^{(n)}) \mathrm{E}_{\vec{\theta}_{\mathrm{CRV}}(\xi_o)|\Omega_\delta^{(n)}} \left[\left(\frac{\langle \frac{\vec{g}^n(X^{n+1})}{n} - \vec{\eta}_o, \frac{d\vec{\eta}_{\mathrm{CRV}}}{d\xi}|_{\xi=\xi_o} \rangle_{\xi_o}}{\langle \frac{d\vec{\eta}_{\mathrm{CRV}}}{d\xi}|_{\xi=\xi_o}, \frac{d\vec{\eta}_{\mathrm{CRV}}}{d\xi}|_{\xi=\xi_o} \rangle_{\xi_o}} \right)^2 \right] \\
&\leq n\tilde{\mathsf{H}}_{\vec{\xi}}[\phi]^{-2} \mathrm{E}_{\vec{\theta}_{\mathrm{CRV}}(\xi_o)} \left[\left(\left\langle \frac{\vec{g}^n(X^{n+1})}{n} - \vec{\eta}_o, \frac{d\vec{\eta}_{\mathrm{CRV}}}{d\xi} \Big|_{\xi=\xi_o} \right\rangle_{\xi_o} \right)^2 \right] \\
&\to \tilde{\mathsf{H}}_{\xi_o}[\phi]^{-1},
\end{aligned}
$$

$$
(8.17) \qquad \begin{aligned}
C_{2,\delta}^{(n)} &:= n\mathrm{P}_{\vec{\theta}_{\mathrm{CRV}}(\xi_o)}(\Omega_\delta^{(n)}) \mathrm{E}_{\vec{\theta}_{\mathrm{CRV}}(\xi_o)|\Omega_\delta^{(n)}} \left[t(X^{n+1})^2 \left\| \frac{\vec{g}^n(X^{n+1})}{n} - \vec{\eta}_o \right\|^2 \right] \\
&\leq n\mathrm{P}_{\vec{\theta}_{\mathrm{CRV}}(\xi_o)}(\Omega_\delta^{(n)}) \mathrm{E}_{\vec{\theta}_{\mathrm{CRV}}(\xi_o)|\Omega_\delta^{(n)}} \left[\left\| \frac{\vec{g}^n(X^{n+1})}{n} - \vec{\eta}_o \right\|^2 \right] \\
&\leq n \sum_i \mathrm{E}_{\vec{\theta}_{\mathrm{CRV}}(\xi_o)} \left[\left(\frac{\vec{g}^n(X^{n+1})}{n} - \vec{\eta}_o \right)_i^2 \right] \\
&\overset{(a)}{\to} C_2 := \sum_i \mathsf{H}_{\theta_{\mathrm{CRV}}(\xi_o)}[\phi]_{i,i},
\end{aligned}
$$

where (a) follows from Theorem 8.2. Since $(\alpha + \varepsilon\beta)^2 = \alpha^2 + 2\varepsilon\alpha\beta + \varepsilon^2\beta^2 \leq \alpha^2 + \varepsilon(\alpha^2 + \beta^2) + \varepsilon^2\beta^2 = (1+\varepsilon)\alpha^2 + (\varepsilon + \varepsilon^2)\beta^2$ for two real numbers α and β,

(8.15), (8.16) and (8.17) imply that

$$
(8.18) \quad
\begin{aligned}
n \mathrm{P}_{\vec{\theta}_{\mathrm{CRV}}(\xi_o)}\big(\Omega_\delta^{(n)}\big) \mathrm{E}_{\vec{\theta}_{\mathrm{CRV}}(\xi_o)|\Omega_\delta^{(n)}} \big[\big(\xi^n(X^{n+1}) - \xi_o\big)^2\big] \\
\leq (1 + \varepsilon) C_{1,\delta}^{(n)} + (\varepsilon + \varepsilon^2) C_{2,\delta}^{(n)}.
\end{aligned}
$$

On the other hand, the large deviation theory of Markov chain [8, 12] guarantees that the probability $\mathrm{P}_{\vec{\theta}_{\mathrm{CRV}}(\xi_o)}\big(\Omega_\delta^{(n)^C}\big)$ goes to zero exponentially. So, we have

$$
(8.19) \quad
\begin{aligned}
&\lim_{n \to \infty} n \mathrm{E}_{\vec{\theta}_{\mathrm{CRV}}(\xi_o)} \big[\big(\xi^n(X^{n+1}) - \xi_o\big)^2\big] \\
&= \lim_{n \to \infty} \Big(n \max_\xi (\xi^n(X^{n+1}) - \xi_o)^2 \mathrm{P}_{\vec{\theta}_{\mathrm{CRV}}(\xi_o)}\big(\Omega_\delta^{(n)^C}\big) \\
&\qquad + n \mathrm{P}_{\vec{\theta}_{\mathrm{CRV}}(\xi_o)}\big(\Omega_\delta^{(n)}\big) \mathrm{E}_{\vec{\theta}_{\mathrm{CRV}}(\xi_o)|\Omega_\delta^{(n)}} \big[\big(\xi^n(X^{n+1}) - \xi_o\big)^2\big]\Big) \\
&= \lim_{n \to \infty} n \mathrm{P}_{\vec{\theta}_{\mathrm{CRV}}(\xi_o)}\big(\Omega_\delta^{(n)}\big) \mathrm{E}_{\vec{\theta}_{\mathrm{CRV}}(\xi_o)|\Omega_\delta^{(n)}} \big[\big(\xi^n(X^{n+1}) - \xi_o\big)^2\big] \\
&\overset{(a)}{\leq} (1 + \varepsilon)\tilde{\mathsf{H}}_{\xi_o}[\phi]^{-1} + (\varepsilon + \varepsilon^2) C_{2,\delta},
\end{aligned}
$$

where (a) follows from (8.16), (8.17) and the limit of (8.18). Since ε is arbitrary,

$$
(8.20) \quad \lim_{n \to \infty} n \mathrm{E}_{\vec{\theta}_{\mathrm{CRV}}(\xi_o)} \big[\big(\xi^n(X^{n+1}) - \xi_o\big)^2\big] \leq \tilde{\mathsf{H}}_{\xi_o}[\phi]^{-1},
$$

which implies the second statement.

Multiparameter case: The random variable $\frac{\vec{g}^n(X^{n+1})}{n} - \vec{\eta}_o$ asymptotically obeys the Gaussian distribution with the covariance matrix $\frac{1}{n}\mathsf{H}_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}_o)}[\phi]$, where $\vec{\eta}_o := \vec{\eta}(\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}_o))$. Since the neighborhood of $\vec{\xi}^n(X^{n+1})$ in $\tilde{\mathcal{E}}$ can be approximated to the tangent space at the true point $\vec{\xi}_o$, due to Corollary 4.8, the point $\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}^n(X^{n+1}))$ can be approximately regarded as the projection to the tangent space at $\vec{\xi}_o$ from the observed point $\vec{\theta}(\frac{\vec{g}^n(X^{n+1})}{n})$.

To see the asymptotic covariance matrix of the random variable $\vec{\xi}^n(X^{n+1}) - \vec{\xi}_o$, we choose a $d \times d'$ matrix B_1 and a $d \times (d - d')$ matrix B_2 such that the $d \times d$ matrix $B = (B_1, B_2)$ satisfies that B is invertible and

$$
(8.21) \quad B^T \mathsf{H}_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}_o)}[\phi]^{-1} B = I \quad \text{and} \quad B_2^T \mathsf{H}_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}_o)}[\phi]^{-1} A = 0.
$$

Then $B_1^T \mathsf{H}_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}_o)}[\phi]^{-1} A$ is invertible. So,

$$
\begin{aligned}
&A^T \mathsf{H}_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}_o)}[\phi]^{-1} A \\
&= \big(B_1^T \mathsf{H}_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}_o)}[\phi]^{-1} A\big)^T B_1^T \mathsf{H}_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}_o)}[\phi]^{-1} B_1 \big(B_1^T \mathsf{H}_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}_o)}[\phi]^{-1} A\big) \\
&= \big(B_1^T \mathsf{H}_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}_o)}[\phi]^{-1} A\big)^T \big(B_1^T \mathsf{H}_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}_o)}[\phi]^{-1} A\big).
\end{aligned}
$$

Now, we introduce the new parameter $\vec{\tau}(\vec{\eta}) := B^{-1}\vec{\eta}$ under which, the metric is given as Cartesian inner product. Hence, the covariance matrix of the estimator $B^{-1}(\frac{\vec{g}^n(X^{n+1})}{n})$ for the parameter $\vec{\tau}(\vec{\eta})$ approaches the matrix $\frac{1}{n}I$. More precisely, Theorem 8.2 guarantees that

$$(8.22) \qquad \lim_{n\to\infty} n\mathsf{Cov}_{\vec{\theta}}\left(B^{-1}\left(\frac{\vec{g}^n(X^{n+1})}{n}\right)\right) = I.$$

We denote the vector $(\tau_1, \ldots, \tau_{d'})^T$ by $\vec{\tau}'(\vec{\eta})$. Since the parameter $\vec{\xi}$ is approximately identified with the element of the tangent space, we have $A(\vec{\xi} - \vec{\xi}_o) = \vec{\eta} - \vec{\eta}_o + o(1) = B(\vec{\tau}(\vec{\eta}) - \vec{\tau}(\vec{\eta}_o)) + o(1)$. Hence, (8.21) implies that

$$B_1^T \mathsf{H}_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}_o)}[\phi]^{-1} A(\vec{\xi} - \vec{\xi}_o) = B_1^T \mathsf{H}_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}_o)}[\phi]^{-1} B(\vec{\tau}(\vec{\eta}(\vec{\xi})) - \vec{\tau}(\vec{\eta}_o)) + o(1)$$
$$= B_1^T \mathsf{H}_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}_o)}[\phi]^{-1} B_1(\vec{\tau}'(\vec{\eta}(\vec{\xi})) - \vec{\tau}'(\vec{\eta}_o)) + o(1)$$
$$= \vec{\tau}'(\vec{\eta}(\vec{\xi})) - \vec{\tau}'(\vec{\eta}_o) + o(1).$$

Thus,

$$\vec{\xi} - \vec{\xi}_o = \left(B_1^T \mathsf{H}_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}_o)}[\phi]^{-1} A\right)^{-1}\left(\vec{\tau}'(\vec{\eta}(\vec{\xi})) - \vec{\tau}'(\vec{\eta}_o)\right) + o(1).$$

In this approximation, our estimator $\vec{\xi}^n(X^{n+1})$ for $\vec{\xi}$ is characterized as

$$(8.23) \qquad \begin{aligned} &\vec{\xi}^n(X^{n+1}) - \vec{\xi}_o \\ &= \left(B_1^T \mathsf{H}_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}_o)}[\phi]^{-1} A\right)^{-1} B^{-1}\left(\frac{\vec{g}^n(X^{n+1})}{n} - \vec{\eta}_o\right) + o(1). \end{aligned}$$

Thus, due to (8.22), the covariance matrix of our estimator is calculated as

$$(8.24) \qquad \begin{aligned} &\mathsf{Cov}_{\vec{\theta}}\left(\left(B_1^T \mathsf{H}_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}_o)}[\phi]^{-1} A\right)^{-1} B^{-1}\left(\frac{\vec{g}^n(X^{n+1})}{n}\right)\right) \\ &= \left(B_1^T \mathsf{H}_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}_o)}[\phi]^{-1} A\right)^{-1} \mathsf{Cov}_{\vec{\theta}}\left(B^{-1}\left(\frac{\vec{g}^n(X^{n+1})}{n}\right)\right) \\ &\quad \times \left(\left(B_1^T \mathsf{H}_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}_o)}[\phi]^{-1} A\right)^T\right)^{-1} \\ &= \left(B_1^T \mathsf{H}_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}_o)}[\phi]^{-1} A\right)^{-1} \frac{I}{n}\left(\left(B_1^T \mathsf{H}_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}_o)}[\phi]^{-1} A\right)^T\right)^{-1} \\ &= \frac{1}{n}\left(\left(B_1^T \mathsf{H}_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}_o)}[\phi]^{-1} A\right)^T \left(B_1^T \mathsf{H}_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}_o)}[\phi]^{-1} A\right)\right)^{-1} \\ &= \frac{1}{n}\left(A^T \mathsf{H}_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}_o)}[\phi]^{-1} A\right)^{-1} = \frac{1}{n}\tilde{\mathsf{H}}_{\vec{\xi}_o}[\phi]^{-1}. \end{aligned}$$

Since the probability that this approximation asymptotically does not hold approaches zero, we can apply the central limit theorem for Markovian process [15,

18, 24] to the random variable given in RHS of (8.23). So, we find that the random variable $\vec{\tilde{\xi}}^n(X^{n+1}) - \vec{\tilde{\xi}}_o$ asymptotically obeys the Gaussian distribution with the covariance matrix $\frac{1}{n}\tilde{\mathsf{H}}_{\vec{\tilde{\xi}}_o}[\phi]^{-1}$.

To show the second statement, we apply the same discussion as the latter part of the one-parameter case by replacing (8.11) by (8.23). In this discussion, the role of (8.24) plays the role of (8.13). More precisely, for any vector $\vec{a}$, we apply the same discussion to

$$(8.25) \quad \begin{aligned} &\vec{a}^T(\vec{\tilde{\xi}}^n(X^{n+1}) - \vec{\tilde{\xi}}_o) \\ &= \vec{a}^T(B_1^T \mathsf{H}_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}_o)}[\phi]^{-1}A)^{-1}B^{-1}\left(\frac{\vec{g}^n(X^{n+1})}{n} - \vec{\eta}_o\right) + o(1). \end{aligned}$$

So, we obtain

$$(8.26) \quad \lim_{n\to\infty} n\vec{a}^T \mathsf{MSE}_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}_o)}[\vec{\tilde{\xi}}^n(X^{n+1})]\vec{a} = \vec{a}^T\tilde{\mathsf{H}}_{\vec{\tilde{\xi}}}[\phi]^{-1}\vec{a}.$$

Since $\vec{a}$ is arbitrary, we obtain the second statement. $\square$

REMARK 3. The papers [3, 10, 13, 19, 20, 35, 36] showed that the maximum likelihood estimator (MLE) is asymptotically efficient in the exponential family with their definition (4.4). Since the definition (4.4) is different from ours (4.1), the results in this section are different from theirs. Further, since our asymptotically efficient estimator is given as the sample mean of g, the required calculation amount is smaller than theirs. Even in the case of a curved exponential family, the Pythagorean theorem (4.20) enables us to calculate our asymptotically efficient estimator with small amount of calculation. However, their MLE does not have so simple form because their exponential family does not have such a geometrical structure, for example, expectation parameter and the Pythagorean theorem, etc. Hence, it requires large calculation amounts.

Indeed, when the matrix entries of the transition matrix is to be estimated, the literature [34] showed that the sample mean is the same as the maximum likelihood estimator. However, this fact holds only for such a specific parameter, and cannot be applied to the parameter estimation of our exponential family, in general. Our method can be applied to any parameter of an exponential family in our sense.

8.4. *Implementation of our estimator for curved exponential family.* In this subsection, we consider how to calculate our estimator $\vec{\tilde{\xi}}^n(X^{n+1})$. This calculation depends on the type of parametrization of the transition matrix $W_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi})}$. We can consider two cases as follows:

(1) The entries of the transition matrix $W_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi})}$ are calculated directly from $\vec{\xi}$ with small calculation complexity.

(2) The entries of the transition matrix $W_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi})}$ are calculated by (4.2) via the calculation of $\vec{\theta}_{\mathrm{CRV}}(\vec{\xi})$. In this case, the calculation of these entries has large calculation complexity.

For example, Example 4 belongs to Case (1) because $W_{\vec{\eta}}$ is directly calculated from the parameter $\vec{\eta}$.

In the calculation of the estimator $\vec{\hat{\xi}}^n(X^{n+1})$, first we obtain the estimate $\vec{\eta}'$ of the larger exponential family $\mathcal{E}$ with the expectation parameter. Then we calculate its natural parameter $\vec{\theta}'$ by the method given in the end of Section 8.1. The following steps depend on the above case. In Case (1), we can implement the minimization by employing the final expression in (5.5) with small calculation complexity due to the following reason. The final expression in (5.5) needs only the entries of the transition matrices $W_{\vec{\theta}'}$ and $W_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi})}$ and the Perron–Frobenius eigenvector of $W_{\vec{\theta}'}$. In this case, it is enough to calculate the Perron–Frobenius eigenvalue and the Perron–Frobenius eigenvector of $W_{\vec{\theta}'}$ only at the first step. At each step of the minimization, we do not have any difficult calculation. Therefore, the final expression in (5.5) brings us an easy implementation of the minimization in Case (1).

However, in Case (2), it is better to employ (4.15) instead of the final expression in (5.5) due to the following reason. When the final expression in (5.5) is employed, the calculation of the transition matrix $W_{\vec{\theta}_{\mathrm{CRV}}(\vec{\xi})}$ requires the calculations of the Perron–Frobenius eigenvalue and the Perron–Frobenius eigenvector of the matrix given in (4.1) as in (4.2). To calculate the RHS of (4.15), we need to calculate the partial derivative $\frac{\partial \phi}{\partial \theta^j}(\vec{\theta}')$ and the Perron–Frobenius eigenvalues $\phi(\vec{\theta}')$ and $\phi(\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}))$. Fortunately, the partial derivative $\frac{\partial \phi}{\partial \theta^j}(\vec{\theta}')$ coincides with the expectation parameter $\vec{\eta}'$, which is first obtained. Also, it is enough to calculate the Perron–Frobenius eigenvalue $\phi(\vec{\theta}')$ only once. Hence, at each step of the minimization, we need to calculate only the Perron–Frobenius eigenvalue $\phi(\vec{\theta}_{\mathrm{CRV}}(\vec{\xi}))$, that is, we do not need to calculate the Perron–Frobenius vector. Therefore, (4.15) requires less calculation complexity than the final expression in (5.5) in Case (2).

9. Conclusion. We have revisited the information geometrical structure (the exponential family, the natural parameter, the expectation parameter, relative entropy, relative Rényi entropy, Fisher information matrix and the Pythagorean theorem) of transition matrices by using the convex function $\phi(\theta)$ defined by the Perron–Frobenius eigenvalue of the matrix $\overline{W}_{\vec{\theta}}$ defined by (4.1). Then we have shown that the sample mean of the generating function is an asymptotically efficient estimator for the expectation parameters in the exponential family of transition matrices. Combining this property and the Pythagorean theorem, we have given an asymptotically efficient estimator for a curved exponential family of transition matrices. As a consequence, we have characterized the asymptotic variance of the sample mean in the Markovian chain by using the second derivative of the convex function $\phi(\theta)$.

In this paper, we have assumed that our system consists of finite elements. On the other hand, in Markov chain Monte Carlo (MCMC) methods, it is also important to evaluate the variance of the sample mean, and several approaches to approximately evaluate the variance of the sample mean have been proposed [16,

21, 22, 30, 32, 33]. Our finite-length bound in equation (6.5) in Lemma 6.2 may be an useful alternative approach for the stationary process with the finite state space. However, the most interesting application of MCMC methods is for a nonstationary process and/or a continuous state space; as is reported in the literature [16, 21, 22, 30, 32, 33], there are some difficulties to evaluate the empirical variance of the sample mean for continuous state space even with discrete time Markov chains. Thus, it is desirable to extend our approach for continuous state space. In fact, we do not use the finiteness of the cardinality of state space so explicitly. Therefore, it seems that there is no essential obstacle for extension to the continuous case under a proper regularity condition. Such an extension will bring us an alternative approach to evaluate the variance of the sample mean even for a continuous case. Also, this extension will enable us to handle several Gaussian Markovian chains in a simple way. Also, extending our result evaluation to a nonstationary case is also an interesting problem.

Further, the obtained version of the Pythagorean theorem will be helpful for the hierarchy of exponential families of transition matrices. For an example, a hierarchy of exponential families can be constructed by changing the degree of Markovian chain, it might be interesting to investigate this example.

APPENDIX A: RELATION WITH EXISTING RESULTS

As mentioned in Introduction, some of results in this paper for relative entropy and exponential family have been already stated in [26] (without detailed proof) and we restate those results and give proofs to keep the paper self-contained. For deeper understanding, we summarize the relation with those papers in this Appendix.

Our definition (3.2) for the relative entropy $D(W\|V)$ has the following relation with those by [26–28]. Natarajan [28] and Nakagawa and Kanaya [27] defined the relative entropy $D(W\|V)$ by the final term of (5.5). However, Nagaoka [26] defined the relative entropy $D(W\|V)$ by (4.15) and showed the equivalence with the final term of (5.5). If we consider only the relative entropy $D(W\|V)$, the definition by the final term of (5.5) is natural. However, the relative Rényi entropy $D_{1+s}(W\|V)$ cannot define in the same way. Hence, in order to treat the relative entropy $D(W\|V)$ and the relative Rényi entropy $D_{1+s}(W\|V)$ in a unified way, we adopt the definition (3.2) for the relative entropy $D(W\|V)$ instead of the final term of (5.5). Our definition clarifies the relation between the relative entropy $D(W\|V)$ and the relative Rényi entropy $D_{1+s}(W\|V)$, which is helpful when we apply these quantities to simple hypothesis testing [12], random number generation, data compression and channel coding [11] in Markov chain.

Next, we address the convexity of the function $\phi(\vec{\theta})$. Nakagawa and Kanaya [27], Section III, and Nagaoka [26] showed the convexity $\phi(\vec{\theta})$ in their respective cases. Nagaoka [26] also showed the equivalence between (1) and (5) in Lemma 4.1. However, they did not clearly consider the relation with the other

conditions in Lemma 4.1. In fact, these equivalence relations are essential for the condition of a generator of an exponential family and also for applications to finite-length evaluations of the tail probability, the error probability in simple hypothesis testing [12], source coding, channel coding and random number generation [11] in the Markov chain.

Now, we proceed to the definition of an exponential family for transition matrices. Our logical order of arguments in this definition is different from that by Nagaoka [26] and Nakagawa and Kanaya [27]. We first define the potential function $\phi(\vec{\theta})$ from a given transition matrix W and a given generator $\{g_j\}$ Then we give the parametric transition matrices although their papers [26, 27] gave the parametric transition matrices first. The potential function $\phi(\vec{\theta})$ for a transition matrix W and a generator $\{g_j\}$ produces several information quantities, which play the central roles when we apply the exponential family for transition matrices to finite-length evaluations of the tail probability and the above applications [11, 12] in Markov chain. To characterize these information quantities, we employ an exponential family of transition matrices. So, our logical order adapts such an application. Further, this paper introduces a mixture family while the existing papers [26, 27] did not define a mixture family.

Indeed, Kontoyiannis and Meyn [18], (11), gave a one-parameter family of transition matrices with the same logical order. However, they did not use the terminology "exponential family" and did not show the convexity of the potential function $\phi(\vec{\theta})$. Ito and Amari [14] discussed the geometrical structure of an exponential family of transition matrices only for $\mathcal{W}_\mathcal{X}$ in the same definition as ours. However, they did not treat this set as an exponential family of transition matrices.

Our formula (4.20) in Pythagorean theorem (Proposition 4.6) has the following relation with Nakagawa and Kanaya [27]. Nakagawa and Kanaya [27], Lemma 5, showed (4.20) with $k = 1$. Hence, our relation (4.20) can be regarded as a generalization of Nakagawa and Kanaya [27], Lemma 5. Indeed, the motivation of Nakagawa and Kanaya ([27], Lemma 5) is related to the exponent of simple hypothesis testing. That is, their purpose is to show the relation

$$(A.1) \qquad \min_{W:D(W\|W_1)\leq r} D(W\|W_0) = \min_{\theta:D(W_\theta\|W_1)\leq r} D(W_\theta\|W_0).$$

However, the multiparametric extension (4.20) is essential for estimation in a curved exponential family, which is discussed in Section 8.3.

APPENDIX B: SET OF POSITIVE BISTOCHASTIC MATRICES

To discuss Example 4 in the detail, we investigate the set of bistochastic matrices on $\mathcal{X} = \{0, 1, \ldots, m\}$. First, we divide the linear space of $(m + 1) \times (m + 1)$ matrices into two linear spaces:

$$(B.1) \quad \mathcal{A} := \{(v_x + w_y)_{x,y}|(v_x)_x, (w_x)_x \in \mathbb{R}^{m+1}\},$$

$$(B.2) \quad \mathcal{B} := \left\{(a_{x,y})_{x,y} \,\middle|\, \sum_{x'=0}^{m} a_{x',y} = \sum_{y'=0}^{m} a_{x,y'} = 0 \text{ for } x, y = 0, 1, \ldots, m\right\}.$$

In the following, any two-input function $g(x, x')$ is regarded as an $(m + 1) \times (m + 1)$ matrix. For an arbitrary nonidentical permutation $\sigma \in S_{\mathcal{X}}$, the function $\hat{g}_\sigma$ belongs to $\mathcal{B}$. The function g_j belongs to $\mathcal{A}$. Also, when a function h satisfies $h(x, y) = c + v_x - v_y$ with a constant c and a vector $(v_x) \in \mathbb{R}^{m+1}$, the function h belongs to $\mathcal{A}$. Any nonzero linear combination of $\{g_j\}_{j=1}^m$ cannot be written by the above function h. Thus, to show the linear independence of the set of functions $\{g_j\}_{j=1}^m \cup \{\hat{g}_\sigma\}_{\sigma \in T \cup H}$, it is enough to show the following lemma.

LEMMA B.1. *The set $\{\hat{g}_\sigma\}_{\sigma \in T \cup H}$ is linearly independent in the linear space $\mathcal{B}$.*

The number of elements of the set $\{g_j\}_{j=1}^m \cup \{\hat{g}_\sigma\}_{\sigma \in T \cup H}$ is m^2, which equals the dimension of $\mathcal{B}$. So, the set $\{g_j\}_{j=1}^m \cup \{\hat{g}_\sigma\}_{\sigma \in T \cup H}$ spans the linear space $\mathcal{B}$. For any bistochastic matrix W, we have $W - W_{id} \in \mathcal{B}$. Hence, $W - W_{id}$ can be written as a linear combination of $\{\hat{g}_\sigma\}_{\sigma \in T \cup H}$, that is, $\sum_{\sigma \in T \cup H} \eta_\sigma \hat{g}_\sigma$. Therefore, $W = W_{id} + \sum_{\sigma \in T \cup H} \eta_\sigma \hat{g}_\sigma = W_{\vec{\eta}}$.

PROOF OF LEMMA B.1. Now, we prepare notation. For a two-input function g, we define the symmetric matrix $S[g]_{x,x'} := g(x, x') + g(x', x)$ and the anti-symmetric matrix $A[g]_{x,x'} := g(x, x') - g(x', x)$.

Due to the constraint for $\mathcal{B}$, the diagonal entries of an element of $S(\mathcal{B})$ are determined by other entries. Fixed $0 \le i' < j' \le m$, only the matrix $S[\hat{g}_{(i',j')}]$ has a nonzero (i', j')th entry among the set $\{S[\hat{g}_{(i,j)}]\}_{(i,j) \in T}$. Hence, the set $\{S[\hat{g}_{(i,j)}]\}_{(i,j) \in T}$ is linearly independent in the linear space $S[\mathcal{B}]$.

Due to the constraint for $\mathcal{B}$, the $(0, i)$th entry and $(i, 0)$th entry of an element of $A(\mathcal{B})$ are determined by other entries. Fixed $0 < i' < j' \le m$, only the matrix $A[\hat{g}_{(0,i',j')}]$ has a nonzero (i', j')th entry among the set $\{A[\hat{g}_{(0,i,j)}]\}_{(0,i,j) \in H}$. Hence, the set $\{A[\hat{g}_{(0,i,j)}]\}_{(0,i,j) \in H}$ is linearly independent in the linear space $A[\mathcal{B}]$. Therefore, the set $\{\hat{g}_{(0,i,j)}\}_{(0,i,j) \in H}$ is linearly independent in the linear space $\mathcal{B}$. Since $A[\hat{g}_{(i,j)}] = 0$ for $(i, j) \in T$, the set $\{\hat{g}_\sigma\}_{\sigma \in T \cup H}$ is linearly independent in the linear space $\mathcal{B}$. □

APPENDIX C: PROOFS OF LEMMAS 5.2 AND 5.3

The Fisher information J_θ^2 can be written as

$$
J_\theta^2 = \sum_{x,x'} W_\theta \times \overline{P}_\theta^1(x, x') \left[-\frac{d^2}{d\theta^2} \log W_\theta(x|x') \overline{P}_\theta^1(x') \right]
$$

$$
= \sum_{x,x'} W_\theta \times \overline{P}_\theta^1(x, x') \left[-\frac{d^2}{d\theta^2} \log W_\theta(x|x') - \frac{d^2}{d\theta^2} \log \overline{P}_\theta^1(x') \right]
$$

$$
= \sum_{x,x'} W_\theta \times \overline{P}_\theta^1(x, x') \left[-\frac{d^2}{d\theta^2} \log W_\theta(x|x') \right]
$$

$$+ \sum_{x'} \overline{P}_\theta^1(x') \left[-\frac{d^2}{d\theta^2} \log \overline{P}_\theta^1(x') \right]$$

(C.1)

$$= \sum_{x,x'} W_\theta \times \overline{P}_\theta^1(x, x') \left[-\frac{d^2}{d\theta^2} \log W_\theta(x|x') \right] + J_\theta^1$$

$$= \sum_{x,x'} W_\theta \times \overline{P}_\theta^1(x, x') \left[-\frac{d^2}{d\theta^2} \log \frac{1}{\lambda_\theta} - \frac{d^2}{d\theta^2} \log \frac{\overline{P}_\theta^3(x)}{\overline{P}_\theta^3(x')} \right.$$

$$\left. - \frac{d^2}{d\theta^2} \log W(x|x') - \frac{d^2}{d\theta^2} \theta g(x, x') \right] + J_\theta^1$$

$$\overset{(a)}{=} \sum_{x,x'} W_\theta \times \overline{P}_\theta^1(x, x') \left[-\frac{d^2}{d\theta^2} \log \frac{1}{\lambda_\theta} \right] + J_\theta^1 = \frac{d^2\phi}{d\theta^2}(\theta) + J_\theta^1,$$

where (a) follows from the relation $\sum_{x,x'} W_\theta \times \overline{P}_\theta^1(x, x') \frac{d^2}{d\theta^2} \log \frac{\overline{P}_\theta^3(x)}{\overline{P}_\theta^3(x')} = 0$, which is shown by the following fact: The expectations of $\frac{d^2}{d\theta^2} \log \overline{P}_\theta^3(X)$ and $\frac{d^2}{d\theta^2} \log \overline{P}_\theta^3(X')$ are the same because the marginal distributions of X and X' are the same. Hence, we obtain (5.6). The Fisher information J_θ^2 is also written as

$$J_\theta^2 = \sum_{x,x'} W_\theta \times \overline{P}_\theta^1(x, x') \left(\frac{d}{d\theta} \log W_\theta(x|x') \overline{P}_\theta^1(x') \right)^2$$

$$= \sum_{x,x'} W_\theta \times \overline{P}_\theta^1(x, x') \left[\left(\frac{d}{d\theta} \log W_\theta(x|x') \right)^2 \right.$$

$$\left. + 2 \left(\frac{d}{d\theta} \log W_\theta(x|x') \right) \left(\frac{d}{d\theta} \log \overline{P}_\theta^1(x') \right) + \left(\frac{d}{d\theta} \log \overline{P}_\theta^1(x') \right)^2 \right]$$

$$= \sum_{x,x'} W_\theta \times \overline{P}_\theta^1(x, x') \left[\left(\frac{d}{d\theta} \log W_\theta(x|x') \right)^2 \right.$$

(C.2)

$$\left. + \sum_{x'} \overline{P}_\theta^1(x') \left(\frac{d}{d\theta} \log \overline{P}_\theta^1(x') \right)^2 \right]$$

$$+ 2 \sum_{x,x'} \left(\frac{d}{d\theta} \log W_\theta(x|x') \right) W_\theta(x|x') \left(\frac{d}{d\theta} \log \overline{P}_\theta^1(x') \right) \overline{P}_\theta^1(x')$$

$$= \sum_{x,x'} W_\theta \times \overline{P}_\theta^1(x, x') \left[\frac{d}{d\theta} \log W_\theta(x|x') \right]^2 + J_\theta^1$$

$$= \sum_{x,x'} W_\theta \times \overline{P}_\theta^1(x,x') \left[-\frac{d\phi}{d\theta}(\theta) + \frac{d}{d\theta} \log \overline{P}_\theta^3(x) \right.$$

$$\left. - \frac{d}{d\theta} \log \overline{P}_\theta^3(x') + g(x,x') \right]^2 + J_\theta^1.$$

Combining (5.6) and (C.2), we have

(C.3)
$$\frac{d^2\phi}{d\theta^2}(\theta) = \mathsf{V}_\theta \left[\left(g(x,x') - \frac{d\phi}{d\theta}(\theta) \right) + \frac{d}{d\theta} \log \overline{P}_\theta^3(x) - \frac{d}{d\theta} \log \overline{P}_\theta^3(x') \right]$$
$$> 0,$$

which implies (5.7). Since

$$\left(\frac{d^2\phi}{d\theta^2}(\theta) + \left(\frac{d}{d\theta}\phi(\theta) \right)^2 \right) e^{\phi(\theta)}$$

$$= \frac{d^2}{d\theta^2} e^{\phi(\theta)} = \sum_{x,x'} \frac{d^2}{d\theta^2} W(x|x') e^{\theta g(x,x')} \overline{P}_\theta^2(x')$$

$$= \sum_{x,x'} W(x|x') e^{\theta g(x,x')} \overline{P}_\theta^2(x') g(x,x')^2$$

$$+ 2W(x|x') e^{\theta g(x,x')} \frac{d\overline{P}_\theta^2(x')}{d\theta} g(x,x') + W(x|x') e^{\theta g(x,x')} \frac{d^2\overline{P}_\theta^2(x')}{d\theta^2},$$

we have another expression of $\frac{d^2\phi}{d\theta^2}(\theta)$ as follows:

$$\frac{d^2\phi}{d\theta^2}(\theta) = e^{-\phi(\theta)} \left[\sum_{x,x'} W(x|x') e^{\theta g(x,x')} \overline{P}_\theta^2(x') g(x,x')^2 \right.$$

$$+ 2W(x|x') e^{\theta g(x,x')} \frac{d\overline{P}_\theta^2(x')}{d\theta} g(x,x')$$

$$\left. + W(x|x') e^{\theta g(x,x')} \frac{d^2\overline{P}_\theta^2(x')}{d\theta^2} \right] - \left(\frac{d}{d\theta}\phi(\theta) \right)^2.$$

When $\theta = 0$,

$$\frac{d^2\phi}{d\theta^2}(0) = \left[\sum_{x,x'} W(x|x') \overline{P}_0^2(x') g(x,x')^2 + 2W(x|x') g(x,x') \frac{d\overline{P}_\theta^2(x')}{d\theta} \bigg|_{\theta=0} \right]$$

$$- \eta(0)^2$$

$$= \mathsf{V}_0[g(X,X')] + 2 \sum_{x,x'} W(x|x') g(x,x') \frac{d\overline{P}_\theta^2(x')}{d\theta} \bigg|_{\theta=0}$$

because $\sum_{x,x'} W(x|x') \frac{d^2 \overline{P}_\theta^2(x')}{d\theta^2} = \frac{d^2}{d\theta^2} \sum_{x,x'} W(x|x') \overline{P}_\theta^2(x') = 0$ and $\eta(0) = \mathsf{E}_0[g(X, X')]$. Hence, we obtain (5.8).

APPENDIX D: TWICE DIFFERENTIABILITY

We show the twice differentiability of $\phi(\theta)$, $\overline{P}_\theta^2$ and $\overline{P}_\theta^3$. First, focus on the 1-parameter case. Now, we define the function $F_1(\theta, z) := \det(\overline{W}_\theta - zI)$ with the identity matrix I. Since $\lambda_\theta = e^{\phi(\theta)}$ is the unique solution of $F_1(\theta, z) = 0$ and the function $F_1(\theta, z)$ is twice differentiable, the implicit function theorem guarantees that λ_θ is twice differentiable. Hence, $\phi(\theta)$ is also twice differentiable.

Next, we show that the twice-differentiability of $\overline{P}_\theta^2$ and $\overline{P}_\theta^3$, which are normalized eigenvector with positive entries of $\overline{W}_\theta$ and $\overline{W}_\theta^T$. Now, we define the vector-valued function $F_2(\theta, y) := \overline{W}_\theta y$ and the function $F_3(\theta, y) := \sum_{x \in \mathcal{X}} y_x$. Since $\overline{P}_\theta^3$ is the unique solution of $F_2(\theta, y) = 0$ and $F_3(\theta, y) = 1$ and the functions $F_2(\theta, y)$ and $F_3(\theta, y)$ are twice differentiable, the implicit function theorem guarantees that $\overline{P}_\theta^2$ is twice differentiable. Replacing the role of $\overline{W}_\theta$ by that of $\overline{W}_\theta^T$, we can show the twice differentiability of $\overline{P}_\theta^3$. These discussions can be extended to the case when θ is a d-dimensional parameter.

REFERENCES

[1] AMARI, S.-I. (2009). α-divergence is unique, belonging to both f-divergence and Bregman divergence classes. *IEEE Trans. Inform. Theory* **55** 4925–4931. MR2596950

[2] AMARI, S.-I. and NAGAOKA, H. (2000). *Methods of Information Geometry*. Amer. Math. Soc., Providence, RI. MR1800071

[3] BHAT, B. R. (1988). On exponential and curved exponential families in stochastic processes. *Math. Sci.* **13** 121–134. MR0974067

[4] BHAT, B. R. (2000). *Stochastic Models*: *Analysis and Applications*. New Age International, New Delhi.

[5] BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. MR2061575

[6] BRÈGMAN, L. M. (1967). The relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. Math. Phys.* **7** 200–217.

[7] CONN, A. R., SCHEINBERG, K. and VICENTE, L. N. (2009). *Introduction to Derivative-Free Optimization*. SIAM, Philadelphia, PA. MR2487816

[8] DEMBO, A. and ZEITOUNI, O. (1998). *Large Deviations Techniques and Applications*, 2nd ed. Springer, New York. MR1619036

[9] DOOB, J. L. (1953). *Stochastic Processes*. Wiley, New York. MR0058896

[10] FEIGIN, P. D. (1981). Conditional exponential families and a representation theorem for asymptotic inference. *Ann. Statist.* **9** 597–603. MR0615435

[11] HAYASHI, M. and WATANABE, S. (2013). Non-asymptotic and asymptotic analyses on Markov chains in several problems. Preprint. Available at arXiv:1309.7528.

[12] HAYASHI, M. and WATANABE, S. (2014). Finite-length analysis on simple hypothesis testing for Markov chain. Preprint. Available at arXiv:1401.3801.

[13] HUDSON, I. L. (1982). Large sample inference for Markovian exponential families with application to branching processes with immigration. *Austral. J. Statist.* **24** 98–112. MR0663783

[14] ITO, H. and AMARI, S. (1988). Geometry of information sources. In *Proceedings of the* 11*th Symposium on Information Theory and Its Applications* (*SITA*1988) 57–60. Beppu. (In Japanese.)

[15] JONES, G. L. (2004). On the Markov chain central limit theorem. *Probab. Surv.* **1** 299–320. MR2068475

[16] JOULIN, A. and OLLIVIER, Y. (2010). Curvature, concentration and error estimates for Markov chain Monte Carlo. *Ann. Probab.* **38** 2418–2442. MR2683634

[17] KEMENY, J. G. and SNELL, J. L. (1960). *Finite Markov Chains*. Springer, New York. MR0115196

[18] KONTOYIANNIS, I. and MEYN, S. P. (2003). Spectral theory and limit theorems for geometrically ergodic Markov processes. *Ann. Appl. Probab.* **13** 304–362. MR1952001

[19] KUCHLER, U. and SORENSEN, M. (1989). Exponential families of stochastic processes: A unifying semimartingale approach. *Int. Stat. Rev.* **57** 123–144.

[20] KÜCHLER, U. and SØRENSEN, M. (1998). On exponential families of Markov processes. *J. Statist. Plann. Inference* **66** 3–19. MR1616992

[21] ŁATUSZYŃSKI, K., MIASOJEDOW, B. and NIEMIRO, W. (2013). Nonasymptotic bounds on the estimation error of MCMC algorithms. *Bernoulli* **19** 2033–2066. MR3129043

[22] ŁATUSZYŃSKI, K. and NIEMIRO, W. (2011). Rigorous confidence bounds for MCMC under a geometric drift condition. *J. Complexity* **27** 23–38. MR2745298

[23] MATSUI, K., KUMAGAI, W. and KANAMORI, T. (2014). Parallel distributed block coordinate descent methods based on pairwise comparison oracle. Preprint. Available at arXiv:1409.3912.

[24] MEYN, S. P. and TWEEDIE, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer, London. MR1287609

[25] MITZENMACHER, M. and UPFAL, E. (2005). *Probability and Computing*: *Randomized Algorithms and Probabilistic Analysis*. Cambridge Univ. Press, Cambridge. MR2144605

[26] NAGAOKA, H. (2005). The exponential family of Markov chains and its information geometry. In *Proceedings of the* 28*th Symposium on Information Theory and Its Applications* (*SITA*2005), Okinawa, Japan.

[27] NAKAGAWA, K. and KANAYA, F. (1993). On the converse theorem in statistical hypothesis testing for Markov chains. *IEEE Trans. Inform. Theory* **39** 629–633. MR1224350

[28] NATARAJAN, S. (1985). Large deviations, hypotheses testing, and source coding for finite Markov chains. *IEEE Trans. Inform. Theory* **31** 360–365. MR0794433

[29] NELDER, J. A. and MEAD, R. (1965). A simplex method for function minimization. *The Computer Journal* **7** 308–313.

[30] NIEMIRO, W. and POKAROWSKI, P. (2009). Fixed precision MCMC estimation by median of products of averages. *J. Appl. Probab.* **46** 309–329. MR2535815

[31] PESKUN, P. H. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika* **60** 607–612. MR0362823

[32] RUDOLF, D. (2009). Explicit error bounds for lazy reversible Markov chain Monte Carlo. *J. Complexity* **25** 11–24. MR2475305

[33] RUDOLF, D. (2010). Error bounds for computing the expectation by Markov chain Monte Carlo. *Monte Carlo Methods Appl.* **16** 323–342. MR2747819

[34] SHALIZI, C. (2009). Maximum likelihood estimation for Markov chains. Available at http://www.stat.cmu.edu/~cshalizi/462/lectures/06/markov-mle.pdf.

[35] SØRENSEN, M. (1986). On sequential maximum likelihood estimation for exponential families of stochastic processes. *Internat. Statist. Rev.* **54** 191–210. MR0962935

[36] STEFANOV, V. T. (1995). Explicit limit results for minimal sufficient statistics and maximum likelihood estimators in some Markov processes: Exponential families approach. *Ann. Statist.* **23** 1073–1101. MR1353496

GRADUATE SCHOOL OF MATHEMATICS
NAGOYA UNIVERSITY
FUROCHO, CHIKUSAKU, NAGOYA 464-8602
JAPAN
AND
CENTRE FOR QUANTUM TECHNOLOGIES
NATIONAL UNIVERSITY OF SINGAPORE
3 SCIENCE DRIVE 2, SINGAPORE 117542
SINGAPORE
E-MAIL: masahito@math.nagoya-u.ac.jp

DEPARTMENT OF COMPUTER
  AND INFORMATION SCIENCES
TOKYO UNIVERSITY OF AGRICULTURE
  AND TECHNOLOGY
KOGANEI-SHI, TOKYO 184-8588
JAPAN
E-MAIL: shunwata@cc.tuat.ac.jp