The Annals of Applied Probability
2017, Vol. 27, No. 2, 811–845
DOI: 10.1214/16-AAP1216

FINITE-LENGTH ANALYSIS ON TAIL PROBABILITY FOR MARKOV CHAIN AND APPLICATION TO SIMPLE HYPOTHESIS TESTING

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

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

Using terminologies of information geometry, we derive upper and lower bounds of the tail probability of the sample mean for the Markov chain with finite state space. Employing these bounds, we obtain upper and lower bounds of the minimum error probability of the type-2 error under the exponential constraint for the error probability of the type-1 error in a simple hypothesis testing for a finite-length Markov chain, which yields the Hoeffding-type bound. For these derivations, we derive upper and lower bounds of cumulant generating function for Markov chain with finite state space. As a byproduct, we obtain another simple proof of central limit theorem for Markov chain with finite state space.

1. Introduction. Since the notion of a Markov chain provides a natural model for (joint) probability distributions with stochastic correlation, we focus on the Markov chain with finite state space. Under this model, we often focus on the sample mean of n samples, and discuss the cumulant generating function and the tail probability. Many existing studies investigated their asymptotic behaviors [7, 9–11, 14–16, 28, 31, 33, 34, 40]. For example, the papers [16, 28, 31, 34, 40, 53] showed the central limit theorem, that is, they proved that the difference between the sample mean and the expectation asymptotically obeys the Gaussian distribution. Donsker and Varadhan [11] did some pioneering works on the large deviation theory of the Markov chain, which influenced later works in this field. Nowadays, it is known that the exponential decaying rate of the Markov chain is characterized by the Legendre transform of the asymptotic cumulant generating function (see Dembo and Zeitouni [10]). Further, other existing studies [45, 46] investigated the simple hypothesis testing for Markov chains. They derived the Hoeffding bound [26] for two Markov chains, that is, the exponentially decreasing rate of the type-2

Received May 2015; revised May 2016.

[1]Supported in part by JSPS Postdoctoral Fellowships for Research Abroad.

[2]Supported in part by a MEXT Grant-in-Aid for Scientific Research (A) No. 23246071 and supported in part by the National Institute of Information and Communication Technology (NICT), Japan.

MSC2010 subject classifications. 62M02, 62F03.

Key words and phrases. Simple hypothesis testing, tail probability, finite-length Markov chain, information geometry, relative entropy, relative Rényi entropy.

error probability under the exponential constraint for the type-1 error probability. In the independently and identically distributed (i.i.d.) case, by extending Stein's lemma, Strassen [52] derived the asymptotic expansion of the exponential decreasing rate of the type-2 error probability up to the order $\sqrt{n}$, under the constant constraint for the type-1 error probability, whose quantum extension was recently done by the papers in [38, 54].

Indeed, it is not difficult to give a bound when it is not so tight or its computation is not so easy. Here, we should mention a proper requirement for a better finite-length bound as follows:

(1) Asymptotic tightness. For example, in the case of the tail probability, the bound can recover one of the following in the limit $n \to \infty$;

  (T1) Central limit theorem [28, 34, 40];
  (T2) Moderate deviation [9, 34];
  (T3) Large deviation [10, 11, 33, 34].

(2) Computability. The bound should have less computational complexity, for example, $O(1)$, $O(n)$ or $O(n \log n)$. For example, we call the bound $O(1)$-*computable* when its computation complexity is $O(1)$.

In the i.i.d. case, it is known that the Markov inequality derives an upper bound of the tail probability that attains the asymptotic tightness in the sense of (T2) and (T3) and is called Chernoff bound [10, 39]. Also, the paper [33] derived a finite-length upper bound of the tail probability in the sense of (T3) for the Markov chain case by another method. However, even in the i.i.d. case, there is no $O(1)$-computable finite-length lower bound that attains the asymptotic tightness in the sense of (T2) nor (T3).

The Berry–Esseen theorem gives upper and lower $O(1)$-computable bounds of the tail probability that attain the asymptotic tightness in the sense of (T1) in the i.i.d. case (see, e.g., [13]). The paper [25], Theorem 2, extended the Berry–Esseen theorem to the Markov chain case, and gave similar upper and lower $O(1)$-computable bounds for Markov chains. Also, the paper [53] generalized the Berry–Esseen theorem to a general setting including the Markov chain case.

In the case of simple hypothesis testing, the three kinds of asymptotic tightness are characterized as follows:

(H1) Constant constraint for the type-1 error probability $\varepsilon = \text{const.}$

(H2) Moderate deviation type constraint for the type-1 error probability $\varepsilon = e^{-n^{1-2t}r}$ with $t \in (0, \frac{1}{2})$.

(H3) Large deviation type constraint for the type-1 error probability $\varepsilon = e^{-nr}$ (Hoeffding bound [45, 46]).

In the i.i.d. case (including the quantum case), the paper [54] derived lower and upper $O(1)$-computable finite-length bounds for the type-2 error probability that

attain the asymptotic tightness in the sense of (H1). Also, it is not difficult to derive an upper $O(1)$-computable finite-length bound for the type-2 error probability that attains the asymptotic tightness in the sense of (H2) nor (H3). However, no study addressed a lower $O(1)$-computable finite-length bound for the type-2 error probability that attains the asymptotic tightness in the sense of (H2) nor (H3) even in the i.i.d. case.

This paper derives the finite-length bounds for the above topics satisfying the above requirement. First, we derive upper and lower bounds of the cumulant generating function when n observations are given. We show that these limits recover the asymptotic cumulant generating function [10]. Using our evaluation of the cumulant generating function, we also derive upper and lower $O(1)$-computable bounds of the tail probability that attains the asymptotic tightness in the sense of (T2) and (T3) in the Markov chain case as well as in the i.i.d. case. Our analysis covers the sample mean of two-input functions like $g(X_{k+1}, X_k)$ as well as the simple sample mean $\sum_{i=1}^n \frac{X_i}{n}$. As a byproduct, employing the evaluation of the cumulant generating function, we simply reproduce the central limit theorem [28, 34, 40]. Indeed, since we address a general function $g(X_{k+1}, X_k)$, our evaluations can be applied to the sample mean of the hidden Markov random variables.

For simple hypothesis testing, this paper derives the lower and upper $O(1)$-computable bounds of the type-2 error probability under the same constraint with finite observations whose limits recover the asymptotic bound (H3) [45, 46] and the asymptotic bound (H2). For describing these finite-length bounds, we employ the notation given by the transition matrix version of information geometry, that is, the relative entropy (Kullback–Leibler divergence), the relative Rényi entropy, exponential family, natural parameter and expectation parameter [24, 44, 45]. Further, employing the Markov version of the Berry–Esseen theorem [25], Theorem 2, we also obtain another type $O(1)$-computable finite-length bound, which derives the asymptotic bound (H1) as a generalization of the result by Strassen [52].

Indeed, there are two ways to define a transition matrix version of exponential family. We employ the definition by [24, 44, 45], which is different from the definition by [5, 6, 12, 27, 35, 50, 51]. The exponential family to be used plays an essential role in our derivation. That is, the exponential family enables us to discuss simple hypothesis testing and the parameter estimation [24] in a unified manner. The obtained bounds are used for the evaluations of several information theoretical problems [21].

As another significance of the obtained result, we point out an interesting application of the simple hypothesis testing to topics in information theory, channel coding, data compression and secure random number generation, etc. For example, the optimal performance of channel coding is evaluated by using the combination of the first and second kinds of error probabilities of simple hypothesis testing [19, 43, 48], which yields the second-order analysis [18, 48]. Its history is reviewed in the recent review paper [17]. Similar evaluations are available for other topics in information theory. Hence, applying the obtained evaluation for Markovian chain,

we can discuss channel coding, data compression and secure random number generation for the Markovian case [20–23]. This kind of application yields the finite block-length evaluation, which gives the evaluation of the optimal performance of real finite block-length codes.

The remainder of this paper is organized as follows. Section 2 gives the brief summary of obtained results. In Section 3, we review an exponential family of transition matrices [24, 44, 45] in the one-parameter case. In Section 4, we characterize Legendre transform of the potential function. In Section 6, we give useful upper and lower bounds of the cumulant generating function. In Section 7, we give a simple alternative proof of the central limit theorem for the Markov chain case. In Section 8, we also give useful upper and lower bounds of the tail probability with finite observation, which produces the large deviation bound of the tail probability. In Section 9, using these bounds, we derive upper and lower bounds of the type-2 error probability of simple hypothesis testing, which yields the Hoeffding type bounds.

2. Summary of results. Here, we prepare notation and definitions. For a given transition matrix W over $\mathcal{X}$, we define $W^{\times n}(x_n, x_{n-1}, \ldots, x_1|\bar{x}) := W(x_n|x_{n-1})W(x_{n-1}|x_{n-2})\cdots W(x_1|\bar{x})$ and $W^n(x|\bar{x}) = \sum_{x_{n-1},\ldots,x_1} W^{\times n}(x, x_{n-1}, \ldots, x_1|\bar{x})$. 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 $x, \bar{x} \in \mathcal{X}$, there exists a natural number n such that $W^n(x|\bar{x}) > 0$ [41]. An irreducible matrix W is called *ergodic* when there are no input $\bar{x}$ and no integer n' such that $W^n(\bar{x}|\bar{x}) = 0$ unless n is divisible by n' [41]. It is known that the output distribution of $W^n P$ converges to the stationary distribution of W for a given ergodic transition matrix W [30, 41].

2.1. *Cumulant generating function.* Assume that $n+1$ random variables $X_1, \ldots, X_{n+1}$ obey the Markov process with the transition matrix $W(x|\bar{x})$. In this paper, for a two-input function $g(x, \bar{x})$ and the Markovian sequence $X^{n+1} := (X_{n+1}, \ldots, X_1)$, we focus on the random variable $g^n(X^{n+1}) := \sum_{i=1}^n g(X_{i+1}, X_i)$. This is because a two-input function $g(x, \bar{x})$ is closely related to an exponential family of transition matrices. Indeed, the simple sample mean can be treated in the formulation by choosing $g(x, \bar{x})$ as x or $\bar{x}$. Here, when we choose a general function $g(x)$, $g^n(X^{n+1}) = \sum_{i=1}^n g(X_{i+1})$ is the sample mean of the hidden Markov random variable. So, our results can be applied to the hidden Markov random case.

We denote the Perron–Frobenius eigenvalue of $W(x|\bar{x})e^{\theta g(x,\bar{x})}$ by λ_θ and define the potential function $\phi(\theta) := \log \lambda_\theta$. Then we focus on the cumulant generating function $\phi_n(\theta) := \log \mathsf{E}[e^{\theta g^n(X^{n+1})}]$, where E denotes the expectation. We will define functions $\underline{\delta}(\theta)$ and $\overline{\delta}(\theta)$ in Section 6 so that $\underline{\delta}(\theta) \to 0$ and $\overline{\delta}(\theta) \to 0$ as $\theta \to 0$. Then we will evaluate $\phi_n(\theta)$ as

$$n\phi(\theta) + \underline{\delta}(\theta) \le \phi_n(\theta) \le n\phi(\theta) + \overline{\delta}(\theta). \tag{2.1}$$

2.2. *Tail probability.* Given an irreducible and ergodic transition matrix W, we will evaluate the tail probability of the random variable $g^n(X^{n+1})$ by using the one-parameter exponential family W_θ given in [24], Section 3, and the relative entropies $D(W_\theta\|W_{\bar\theta})$ and $D_{1+s}(W_\theta\|W_{\bar\theta})$ defined in Section 3 as follows. Now, we focus on the asymptotic expectation of the sample mean $\mathsf{E}[g] := \lim_{n\to\infty}\frac{1}{n}\mathsf{E}[g^n(X^{n+1})]$. For any $a > \mathsf{E}[g]$, we will show

$$-\log P\{g^n(X^{n+1}) \ge na\} \ge nD(W_{\phi'^{-1}(a)}\|W_0) - \overline{\delta}(\theta), \tag{2.2}$$

where $\phi'^{-1}(a)$ is the inverse function of $\phi'(\theta) = \frac{d\phi}{d\theta}(\theta)$, that is, $\frac{d\phi}{d\theta}(\phi'^{-1}(a)) = a$. Conversely, we will show

$$\begin{aligned} &-\log P\{g^n(X^{n+1}) \ge na\} \\ &\le \inf_{\substack{s>0\\ \theta>\phi'^{-1}(a)}} nD_{1+s}(W_\theta\|W_0) + \frac{1}{s}\big[\overline{\delta}\big((1+s)\theta\big) - \underline{\delta}(\theta)\big] \\ &\quad - \frac{1+s}{s}\log\big(1 - e^{-nD(W_{\phi'^{-1}(a)}\|W_\theta)+\overline{\delta}(\phi'^{-1}(a))-\underline{\delta}(\theta)}\big). \end{aligned} \tag{2.3}$$

Similarly, for $a < \mathsf{E}[g]$, we will show

$$-\log P\{g^n(X^{n+1}) \le na\} \ge nD(W_{\phi'^{-1}(a)}\|W_0) - \overline{\delta}(\theta). \tag{2.4}$$

Conversely, we will show

$$\begin{aligned} &-\log P\{g^n(X^{n+1}) \le na\} \\ &\le \inf_{\substack{s>0\\ \theta<\phi'^{-1}(a)}} nD_{1+s}(W_\theta\|W_0) + \frac{1}{s}\big[\overline{\delta}\big((1+s)\theta\big) - \underline{\delta}(\theta)\big] \\ &\quad - \frac{1+s}{s}\log\big(1 - e^{-nD(W_{\phi'^{-1}(a)}\|W_\theta)+\overline{\delta}(\phi'^{-1}(a))-\underline{\delta}(\theta)}\big). \end{aligned} \tag{2.5}$$

2.3. *Simple hypothesis testing.* Now, we consider the hypothesis testing with the two hypotheses $W_0^{\times n}\times P_0$ and $W_1^{\times n}\times P_1$. Usually, the null hypothesis is written with the parameter 0, and the alternative one is with 1. However, this paper employs the opposite parameterization because of the following reason. Although the main results of this paper are upper and lower bounds of type-2 error probability by using several functions, the constructions of these functions are closely related to the parameterization of the null and alternative hypotheses. If we parametrize them in the conventional way, the forms of these functions become more complicated so that many important formulas could not be written in one line. To avoid these kinds of troubles, we employ the parametrization opposite to the conventional case.

Then we consider the minimum type-2 error probability

$$\begin{aligned} &\beta_\varepsilon\big(W_1^{\times n}\times P_1\|W_0^{\times n}\times P_0\big) \\ &\quad := \min_{S\subset\mathcal{X}^{n+1}}\big\{1 - W_0^{\times n}\times P_0(S)|W_1^{\times n}\times P_1(S) \le \varepsilon\big\}. \end{aligned} \tag{2.6}$$

Using the one-parameter exponential family W_θ of transition matrices with the generator $g(x,\bar{x}) := \log \frac{W_1(x|\bar{x})}{W_0(x|\bar{x})}$ and the cumulant generating function $\phi(\theta)$ defined by $g(x,\bar{x})$, we will show that

$$\begin{aligned}
&\sup_{0\le\theta\le 1} \frac{n(-\theta r - \phi(\theta)) - \underline{\delta}(\theta)}{1-\theta} \\
&\quad \le -\log \beta_{e^{-nr}}\big(W_1^{\times n} \times P_1 \| W_0^{\times n} \times P_0\big) \\
&\quad \le \inf_{s>0,\theta\in(\hat{\theta}(r),1)} nD_{1+s}(W_\theta\|W_0) + \frac{1}{s}\big(\overline{\delta}\big((1+s)\theta\big) - (1+s)\underline{\delta}(\theta)\big) \\
&\qquad - \frac{1+s}{s}\log\big(1 - 2e^{-nD(W_{\hat{\theta}(r)}\|W_\theta) - \underline{\delta}(\theta) + \frac{(1-\theta)\overline{\delta}(\hat{\theta}(r))}{1-\hat{\theta}(r)}}\big),
\end{aligned} \tag{2.7}$$

where the functions $\hat{\theta}(r)$ is given in Section 4. We will also asymptotically characterize $\beta_\varepsilon(W_1^{\times n} \times P_1 \| W_0^{\times n} \times P_0)$ with a fixed ε.

3. Geometric structure for transition matrices. In this section, we review the definition and the properties of the one-parameter exponential family of transition matrices [44, 45] by following the logical order of [24], Section 4, although a large part of results for exponential family of transition matrices were obtained by Nagaoka [44] and Nakagawa and Kanaya [45]. This is because the logical order of [24], Section 4, is more suitable for the context of this paper. These relations are explained in [24], Remarks 3.5, 4.12, and 4.14. Note that the definition of exponential family in this paper is different from that by the papers [5, 6, 12, 27, 35, 50, 51] as explained in [24], Remark 4.13.

3.1. *Preparations*. For the definition and the properties of the one-parameter exponential family of transition matrices, we prepare the following things.

LEMMA 3.1 ([24], Lemma 3.1). *Consider an irreducible and ergodic transition matrix W over $\mathcal{X}$ and a real-valued function g on $\mathcal{X}\times\mathcal{X}$. Then we define the support $\mathcal{X}_W^2 := \{(x,\bar{x}) \in \mathcal{X}^2 | W(x|\bar{x}) > 0\}$. Define $\phi(\theta)$ as the logarithm of the Perron–Frobenius eigenvalue of the matrix*:

$$\tilde{W}_\theta(x|\bar{x}) := W(x|\bar{x}) e^{\theta g(x,\bar{x})}. \tag{3.1}$$

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

(1) *No real-valued function f on $\mathcal{X}$ satisfies that $g(x,\bar{x}) = f(x) - f(\bar{x}) + c$ for any $(x,\bar{x}) \in \mathcal{X}_W^2$ 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 ergodic transition matrices W and V with the same support, we define the relative entropy and the relative Rényi entropies. For this purpose, we denote the logarithm of the Perron–Frobenius eigenvalue of the matrix $W(x|\bar{x})^{1+s}V(x|\bar{x})^{-s}$ by $\varphi(1+s)$. Then we define

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

Note that the limit $\lim_{s\to 0} D_{1+s}(W\|V)$ equals $D(W\|V)$. Since W and V are distinct, the function $\log \frac{W(x|\bar{x})}{V(x|\bar{x})}$ satisfies the condition for the function g in Lemma 3.1. Hence, the function $s \mapsto sD_{1+s}(W\|V)$ is strictly convex, which implies that $sD_{1+s}(W\|V) < (1-\frac{s}{\bar{s}})0 + \frac{s}{\bar{s}}\bar{s}D_{1+\bar{s}}(W\|V)$ for $0 < s < \bar{s}$. Since a similar relation holds for $0 > s > \bar{s}$, the relative Rényi entropy $D_{1+s}(W\|V)$ is strictly monotone increasing with respect to s.

3.2. *Exponential family.* Now, we focus on a transition matrix $W(x|\bar{x})$ from $\mathcal{X}$ to $\mathcal{X}$ and a real-valued function g on $\mathcal{X}\times\mathcal{X}$ satisfying the condition in Lemma 3.1. In the following, we assume that the function g satisfies condition in Lemma 3.1. Then we will define the matrix $W_\theta(x|\bar{x})$ from $\mathcal{X}$ to $\mathcal{X}$ for θ by following steps below. For this purpose, we define the matrix $\tilde{W}_\theta(x|\bar{x})$ from $\mathcal{X}$ to $\mathcal{X}$ by

$$\tilde{W}_\theta(x|\bar{x}) := W(x|\bar{x})e^{\theta g(x,\bar{x})}. \tag{3.3}$$

Using the Perron–Frobenius eigenvalue λ_θ of $\tilde{W}_\theta$, we define the potential function

$$\phi(\theta) := \log \lambda_\theta. \tag{3.4}$$

Due to Lemma 3.1, the second derivative $\frac{d^2\phi}{d\theta^2}$ is strictly positive. Hence, the potential function $\phi(\theta)$ is strictly convex. In the following, using the strictly convex function $\phi(\theta)$, we define a one-parameter exponential family for transition matrices.

Note that, since the value $\sum_x \tilde{W}_\theta(x|\bar{x})$ generally depends on $\bar{x}$, we cannot make a transition matrix by simply multiplying a constant with the matrix $\tilde{W}_\theta$. To make a transition matrix from the matrix $\tilde{W}_\theta$, 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 $\tilde{W}_\theta$. We denote the Perron–Frobenius eigenvectors of $\tilde{W}_\theta$ and its transpose $\tilde{W}_\theta^T$ by $\tilde{P}_\theta$ and $\hat{P}_\theta$. Since the irreducibility of W guarantees the irreducibility of $\tilde{P}_\theta^T$, the relation $\hat{P}_\theta(x) > 0$ holds. According to [24, 34, 44, 45],[3] we define the matrix $W_\theta(x|\bar{x})$ as

$$W_\theta(x|\bar{x}) := \lambda_\theta^{-1}\hat{P}_\theta(x)\tilde{W}_\theta(x|\bar{x})\hat{P}_\theta(\bar{x})^{-1}. \tag{3.5}$$

[3]The Appendix of [24] explains detailed relation the papers [24, 34, 44, 45] for an exponential family of transition matrices.

The matrix $W_\theta(x|\bar{x})$ is a transition matrix because vector $(1,\dots,1)^T$ is an eigenvector of the transpose W_θ^T. In the following, we call the family of transition matrices $\mathcal{E} := \{W_\theta\}$ an *exponential family* of transition matrices with the generator g.

Using the potential function $\phi(\theta)$, we explain several concepts for transition matrices based on Lemma 3.1, formally. We call the parameter θ the natural parameter, and the parameter $\eta(\theta) := \phi'(\theta) = \frac{d\phi}{d\theta}(\theta)$ the expectation parameter. For η, we define the inverse function $\phi'^{-1}(\eta)$ of ϕ' as

$$\phi'\big(\phi'^{-1}(\eta)\big) = \eta. \tag{3.6}$$

Then we define the Fisher information for the natural parameter by the second derivative $\frac{d^2\phi}{d\theta^2}(\theta)$. The Fisher information for the expectation parameter is given as $\frac{d^2\phi}{d\theta^2}(\theta)^{-1}$.

LEMMA 3.2 ([24], Lemma 4.4). *The relative entropy and the relative Rényi entropies between two transition matrices W_θ and $W_{\bar{\theta}}$ are characterized as*

$$D(W_\theta \| W_{\bar{\theta}}) = (\theta - \bar{\theta})\frac{d\phi}{d\theta}(\theta) - \phi(\theta) + \phi(\bar{\theta}), \tag{3.7}$$

$$D_{1+s}(W_\theta \| W_{\bar{\theta}}) = \frac{\phi((1+s)\theta - s\bar{\theta}) - (1+s)\phi(\theta) + s\phi(\bar{\theta})}{s}. \tag{3.8}$$

In the following, E_W denotes the expectation with respect to the joint distribution when the conditional distribution is given by the transition matrix W and the input distribution is given by the stationary distribution of W. Then, for a generator g and a real number a, we define the mixture subfamily $\mathcal{M}_{g,a}$ as

$$\mathcal{M}_{g,a} := \big\{W | \mathsf{E}_W g(X, X') = a\big\}. \tag{3.9}$$

A transition matrix version of the Pythagorean theorem [3] holds as follows.

THEOREM 3.3 ([45], Lemma 5, [24], Corollary 4.8). *For a transition matrix V, a generator g, and a real number a, we define*

$$V^* := \operatorname*{argmin}_{W \in \mathcal{M}_{g,a}} D(W\|V). \tag{3.10}$$

(1) *Any transition matrix $W \in \mathcal{M}_{g,a}$ satisfies*

$$D(W\|V) = D\big(W\|V^*\big) + D\big(V^*\|V\big). \tag{3.11}$$

(2) *The transition matrix V^* is the intersection of the set $\mathcal{M}_{g,a}$ and the exponential family generated by g containing V.*

Due to Lemma 3.2, the Fisher information $\frac{d^2\phi}{d\theta^2}(\theta_0)$ can be characterized by the limits of the relative entropy and relative Rényi entropy as follows.

LEMMA 3.4. *Under the limit* $\delta \to 0$, *we have*

$$\lim_{\delta\to 0}\frac{1}{\delta^2}D(W_{\theta_0+\delta}\|W_{\theta_0}) = \lim_{\delta\to 0}\frac{1}{\delta^2}D(W_{\theta_0}\|W_{\theta_0+\delta}) = \frac{1}{2}\frac{d^2\phi}{d\theta^2}(\theta_0), \tag{3.12}$$

$$\begin{aligned}\lim_{\delta\to 0}\frac{1}{\delta^2}D_{1+s}(W_{\theta_0+\delta}\|W_{\theta_0}) &= \lim_{\delta\to 0}\frac{1}{\delta^2}D_{1+s}(W_{\theta_0}\|W_{\theta_0+\delta})\\ &= \frac{1+s}{2}\frac{d^2\phi}{d\theta^2}(\theta_0).\end{aligned} \tag{3.13}$$

4. Relation with Legendre transform. Given two irreducible and ergodic transition matrices W and V, we choose the exponential family W_θ with the generator $g(x,\bar{x}) := \log V(x|\bar{x}) - \log W(x|\bar{x})$ so that $W_0 = W$ and $W_1 = V$. In fact, an arbitrary exponential family W_θ can be written as the above form by choosing two irreducible and ergodic transition matrices as $W := W_0$ and $V := W_1$. The Legendre transform $\sup_{\theta\ge 0}[\theta a - \phi(\theta)]$ of the convex function ϕ can be characterized as follows.

LEMMA 4.1. *When* $a > -D(W\|V)$,

$$\begin{aligned}\inf_{\substack{s>0\\ \theta>\phi'^{-1}(a)}} D_{1+s}(W_\theta\|W_0) &= \inf_{\substack{s>0\\ \theta>\phi'^{-1}(a)}} \frac{\phi((1+s)\theta)-(1+s)\phi(\theta)}{s}\\ &= \phi'^{-1}(a)a - \phi\big(\phi'^{-1}(a)\big)\\ &= \sup_{\theta\ge 0}\big[\theta a - \phi(\theta)\big] = D(W_{\phi'^{-1}(a)}\|W_0).\end{aligned} \tag{4.1}$$

Similarly, when $a < -D(W\|V)$,

$$\begin{aligned}\inf_{\substack{s>0\\ \theta<\phi'^{-1}(a)}} \frac{\phi((1+s)\theta)-(1+s)\phi(\theta)}{s} &= \phi'^{-1}(a)a - \phi\big(\phi'^{-1}(a)\big)\\ &= \sup_{\theta\le 0}\big[\theta a - \phi(\theta)\big] = D(W_{\phi'^{-1}(a)}\|W_0).\end{aligned}$$

PROOF. Since the function $\theta \mapsto \phi(\theta)$ is convex, the function $s \mapsto \frac{\phi((1+s)\theta)-(1+s)\phi(\theta)}{s}$ is monotone increasing due to Lemma C.1. Hence, $\inf_{s>0}\frac{\phi((1+s)\theta)-(1+s)\phi(\theta)}{s} = \theta\frac{d\phi}{d\theta}(\theta) - \phi(\theta)$ for $\theta > \phi'^{-1}(a)$. Thus,

$$\inf_{\substack{s>0\\ \theta>\phi'^{-1}(a)}} \frac{\phi((1+s)\theta)-(1+s)\phi(\theta)}{s} = \inf_{\theta>\phi'^{-1}(a)} \theta\frac{d\phi}{d\theta}(\theta) - \phi(\theta).$$

Since the function $\theta \mapsto \phi(\theta)$ is convex, the function $\theta \mapsto \theta\frac{d\phi}{d\theta}(\theta) - \phi(\theta)$ is monotone increasing for $\theta \ge 0$. Therefore,

$$\inf_{\theta>\phi'^{-1}(a)} \theta\frac{d\phi}{d\theta}(\theta) - \phi(\theta) = \phi'^{-1}(a)a - \phi\big(\phi'^{-1}(a)\big) = \sup_{\theta\ge 0}\big[\theta a - \phi(\theta)\big],$$

where the second equation follows from the convexity of $\phi(\theta)$. The final equation in (4.1) is shown by $D(W_{\phi'^{-1}(a)}\|W_0) = \phi'^{-1}(a)a - \phi(\phi'^{-1}(a))$. $\square$

Now, for an arbitrary convex function ϕ and $r > 0$, we define the function $\hat{\theta}(r) = \hat{\theta}[\phi](r)$ as the smaller solution of the equation

$$(\theta - 1)\frac{d\phi}{d\theta}(\theta) - \phi(\theta) = D(W_\theta\|W_1) = r \tag{4.2}$$

with respect to θ. Hence, due to the convexity of ϕ, we have

$$\begin{aligned}
&\inf_{s>0,\theta\in(0,\hat{\theta}(r))} D_{1+s}(W_\theta\|W_0) \\
&\quad= \inf_{s>0,\theta\in(0,\hat{\theta}(r))} \frac{1}{s}\big[\phi\big((1+s)\theta\big) - (1+s)\phi(\theta)\big] \\
&\quad= \inf_{s>0} \frac{1}{s}\big[\phi\big((1+s)\hat{\theta}(r)\big) - (1+s)\phi\big(\hat{\theta}(r)\big)\big] = \hat{\theta}(r)\frac{d\phi}{d\theta}\big(\hat{\theta}(r)\big) - \phi\big(\hat{\theta}(r)\big) \qquad (4.3)\\
&\quad= -\hat{\theta}(r)\frac{r + \phi(\hat{\theta}(r))}{1-\hat{\theta}(r)} - \phi\big(\hat{\theta}(r)\big) = \frac{-\hat{\theta}(r)r - \phi(\hat{\theta}(r))}{1-\hat{\theta}(r)} \\
&\quad\overset{(a)}{=} \sup_{0\le\theta<1} \frac{-\theta r - \phi(\theta)}{1-\theta} = \sup_{0\le\theta<1} \frac{\theta(-r + D_{1-\theta}(W_0\|W_1))}{1-\theta},
\end{aligned}$$

which implies the following lemma. Here, (a) can be derived as follows. Due to Lemma C.1, the maximum can be attained when (4.2) holds, that is, $\theta = \hat{\theta}(r)$. Hence, we have (a).

LEMMA 4.2. *When* $0 \le r \le D(W_0\|W_1)$,

$$\begin{aligned}
\sup_{0\le\theta\le1} \frac{-\theta r - \phi(\theta)}{1-\theta} &= \sup_{0\le\theta\le1} \frac{\theta(-r + D_{1-\theta}(W_0\|W_1))}{1-\theta} \\
&= \inf_{s>0,\theta\in(0,\hat{\theta}(r))} D_{1+s}(W_\theta\|W_0) = D(W_{\hat{\theta}(r)}\|W_0) \\
&= \min_{W:D(W\|W_1)\le r} D(W\|W_0).
\end{aligned}$$

Here, when $\frac{\phi(\theta)}{1-\theta}$ is regarded as a function of $\delta := \frac{-\theta}{1-\theta}$, that is, is described as $f(\delta)$, $\sup_{0\le\theta\le1} \frac{-\theta r - \phi(\theta)}{1-\theta}$ is given as the Legendre transform of f, i.e., $\sup_{0\le\theta\le1} \delta r - f(\delta)$.

PROOF OF LEMMA 4.2. The first and second equations follow from (4.3). The third equation follows from (4.3) and the relation $D(W_{\hat{\theta}(r)}\|W_0) = \hat{\theta}(r)\frac{d\phi}{d\theta}(\hat{\theta}(r)) - \phi(\hat{\theta}(r))$. Now, we show the final equation. We choose W satisfying that $D(W\|$

$W_1) \le r$. We also choose a such that $W \in \mathcal{M}_{g,a}$, which is defined in Theorem 3.3. Then we denote the intersection of the set $\mathcal{M}_{g,a}$ and the exponential family $\{W_\theta\}$ by $W_{\bar\theta}$. Theorem 3.3 implies that $D(W_{\bar\theta}\|W_1) \le r$ and $D(W_{\bar\theta}\|W_0) \le D(W\|W_0)$. Thus, we obtain

$$\min_{W:D(W\|W_1)\le r} D(W\|W_0) = \min_{\theta:D(W_\theta\|W_1)\le r} D(W_\theta\|W_0). \tag{4.4}$$

Due to the condition $0 \le r \le D(W_1\|W_0)$, the above value equals $D(W_{\hat\theta(r)}\|W_0)$. □

5. Information processing inequality. Now, we introduce a condition for a transition matrix as follows. 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 $y' \in \mathcal{Y}$. When the Markov chain for X and Y generated by W and W satisfies the above condition, the sequence for X is also a Markov chain, not a hidden Markov chain. This is the reason of the name of "nonhidden." As a transition matrix version of information processing inequality, we have the following theorem.

THEOREM 5.1. *When two transition matrices W and V on $\mathcal{X}\times\mathcal{Y}$ are nonhidden for $\mathcal{X}$, the following hold for $s \in (-1,0)\cup(0,\infty)$:*

$$D(W\|V) \ge D(W_X\|V_X), \qquad D_{1+s}(W\|V) \ge D_{1+s}(W_X\|V_X). \tag{5.1}$$

PROOF. For $s > 0$, let λ and $b = (b_{x,y})$ be the Perron–Frobenius eigenvalue and eigenvector of the matrix $W(x,y|x',y')^{1+s}V(x,y|x',y')^{-s}$. Since the reverse Hölder inequality implies

$$\begin{aligned}
&\sum_y W(x,y|x',y')^{1+s}V(x,y|x',y')^{-s} \\
&\quad\ge \Big(\sum_y W(x,y|x',y')^{(1+s)/(1+s)}\Big)^{1+s}\Big(\sum_y V(x,y|x',y')^{-s/(-s)}\Big)^{-s} \\
&\quad= W_X(x|x')^{1+s}V_X(x|x')^{-s},
\end{aligned}$$

the number $c_x = \sum_y b_{x,y}$ satisfies

$$\begin{aligned}
\lambda c_x &= \sum_y \lambda b_{x,y} = \sum_y \sum_{x',y'} W(x,y|x',y')^{1+s}V(x,y|x',y')^{-s} b_{x',y'} \\
&\ge \sum_{x',y'} W_X(x|x')^{1+s}V_X(x|x')^{-s} b_{x',y'} = \sum_{x'} W_X(x|x')^{1+s}V_X(x|x')^{-s} c_{x'},
\end{aligned}$$

which implies that $\max_x \frac{\sum_{x'} W_X(x|x')^{1+s}V_X(x|x')^{-s}c_{x'}}{c_x} \le \lambda$. Due to the Collatz–Wielandt formula, λ is larger than the Perron–Frobenius eigenvalue of the matrix

$W_X(x|x')^{1+s}V_X(x|x')^{-s}$. Hence, we obtain the second inequality of (5.1) with $s > 0$.

When $s \in (-1, 0)$, replacing the role of the reverse Hölder inequality by the Hölder inequality, we can show that $\min_x \frac{\sum_{x'} W_X(x|x')^{1+s}V_X(x|x')^{-s}c_{x'}}{c_x} \geq \lambda$. Due to the Perron–Frobenius theorem, λ is smaller than the Perron–Frobenius eigenvalue of the matrix $W_X(x|x')^{1+s}V_X(x|x')^{-s}$. Hence, we obtain the second inequality of (5.1) with $s \in (-1, 0)$. Taking the limit $s \to 0$, we obtain the first inequality of (5.1). □

Theorem 5.1 can be regarded as a part of information processing inequality as follows. In the case of the information processing inequality between two distributions P and P', we compare the relative entropy between P and P' and the relative entropy between VP and VP' for a given transition matrix V. Since the relative entropy between P and P' equals the relative entropy between $V \times P$ and $V \times P'$, it is enough to compare the relative entropy between $V \times P$ and $V \times P'$ and the relative entropy between VP and VP'. The difference between these relative entropies can be characterized as existence or nonexistence of the marginalization for the input system. Therefore, the information processing inequality can be reduced to the information processing inequality with respect to the marginalization. As the inequalities in Theorem 5.1 give the relations among the relative entropies before/after the marginalization, they can be regarded as an information processing inequality. Therefore, it can be expected that Theorem 5.1 will play roles of information processing inequality in information theory.

6. Cumulant generating function. In the following, we consider the Markov chain $X^{n+1} = (X_1, \ldots, X_n, X_{n+1})$ generated by the transition matrix W_0 and an arbitrary initial distribution P_0. That is, the random variable X^{n+1} is subject to the distribution $W_0^{\times n} \times P_0$. We consider the random variable $\tilde{g}^n(X^{n+1}) := \sum_{i=1}^n g(X_{i+1}, X_i) + h(X_1)$ for a function h on $\mathbb{R}$. Then we define the cumulant generating function

$$\phi_n(\theta) := \log \mathsf{E}_0\big[e^{\theta \tilde{g}^n(X^{n+1})}\big], \tag{6.1}$$

where E_0 denotes the expectation under the distribution $W_0^{\times n} \times P_0$.

LEMMA 6.1. *Let v_θ be the eigenvector of $\tilde{W}_\theta^T$ with respect to the Perron–Frobenius eigenvalue λ_θ such that $\min_x v_\theta(x) = 1$. Let $w_\theta(x) := P_0(x)e^{\theta h(x)}$. Then we have*

$$n\phi(\theta) + \underline{\delta}(\theta) \leq \phi_n(\theta) \leq n\phi(\theta) + \overline{\delta}(\theta), \tag{6.2}$$

where

$$\overline{\delta}(\theta) := \log\langle v_\theta | w_\theta\rangle, \qquad \underline{\delta}(\theta) := \log\langle v_\theta | w_\theta\rangle - \log \max_x v_\theta(x). \tag{6.3}$$

REMARK 6.2. The recent paper [42], Lemma 24,[4] showed a related evaluation as

$$n\phi(\theta) - (1+2\theta)\log\eta \le \phi_n(\theta) \le n\phi(\theta) + (1+2\theta)\log\eta, \tag{6.4}$$

where η is a constant, which is independent of θ. In fact, to apply this kind evaluation to our proof of our Lemma 7.1, the error term needs to go to zero when θ goes to zero. However, their evaluation (6.4) does not satisfy this requirement because the error term $(1+2\theta)\log\eta$ does not go to zero while our evaluation satisfies this requirement as Lemma 6.4. This comparison shows that in our evaluation, Lemma 6.1 is tighter than their evaluation (6.4) in this sense.

PROOF OF LEMMA 6.1. Let u be the vector such that $u(x) = 1$ for every $x \in \mathcal{X}$. From the definition of $\phi_n(\theta)$, we have the following sequence of calculations:

$$\begin{aligned} e^{\phi_n(\theta)} &= \sum_{x_n,\dots,x_1} P(x_1) \prod_{i=1}^{n} W(x_{i+1}|x_i) e^{\theta \sum_{i=1}^n g(x_{i+1},x_i)+h(x_1)} \\ &= \langle u|\tilde{W}_\theta^n w_\theta\rangle \le \langle v_\theta|\tilde{W}_\theta^n w_\theta\rangle = \langle (\tilde{W}_\theta^T)^n v_\theta|w_\theta\rangle = \lambda_\theta^n \langle v_\theta|w_\theta\rangle = e^{n\phi(\theta)}\langle v_\theta|w_\theta\rangle, \end{aligned}$$

which implies the right-hand side inequality of (6.2). On the other hand, we have the following sequence of calculations:

$$\begin{aligned} e^{\phi_n(\theta)} &= \langle u|\tilde{W}_\theta^n w_\theta\rangle \ge \frac{1}{\max_x v_\theta(x)}\langle v_\theta|\tilde{W}_\theta^n w_\theta\rangle \\ &= \frac{1}{\max_x v_\theta(x)}\langle (\tilde{W}_\theta^T)^n v_\theta|w_\theta\rangle = \lambda_\theta^n \frac{\langle v_\theta|w_\theta\rangle}{\max_x v_\theta(x)} = e^{n\phi(\theta)}\frac{\langle v_\theta|w_\theta\rangle}{\max_x v_\theta(x)}, \end{aligned}$$

which implies the left-hand side inequality of (6.2). □

By taking the limit in (6.2) of Lemma 6.1, we have the following.

COROLLARY 6.3 ([10], Theorem 3.1.1). *For $\theta \in \mathbb{R}$, we have*

$$\lim_{n\to\infty} \frac{1}{n}\phi_n(\theta) = \phi(\theta). \tag{6.5}$$

LEMMA 6.4.

$$\lim_{\theta\to 0} \overline{\delta}(\theta) = 0, \qquad \lim_{\theta\to 0} \underline{\delta}(\theta) = 0. \tag{6.6}$$

[4] After the submission of the preliminary conference version [57] of this paper in April 2014, a related paper was posted in arXiv in September 2014 [42].

PROOF. From the construction of v_θ and w_θ, the vectors v_θ and w_θ are continuous for θ. Hence,

$$\lim_{\theta\to 0}\langle v_\theta|w_\theta\rangle = \langle u|w_0\rangle = \sum_x P(x) = 1, \tag{6.7}$$

which implies the first equation of (6.6). Similarly,

$$\lim_{\theta\to 0}\max_x v_\theta(x) = \max_x u(x) = 1. \tag{6.8}$$

Combining (6.7) and (6.8), we obtain the second equation of (6.6). □

Using these relations, we can show the following lemma.

LEMMA 6.5. *For any initial distributions P_0 and P_1, we have*

$$\lim_{n\to\infty}\frac{1}{n}D\big(W_0^{\times n}\times P_0\|W_1^{\times n}\times P_1\big) = D(W_0\|W_1), \tag{6.9}$$

$$\lim_{n\to\infty}\frac{1}{n}D_{1+s}\big(W_0^{\times n}\times P_0\|W_1^{\times n}\times P_1\big) = D_{1+s}(W_0\|W_1). \tag{6.10}$$

PROOF. Now, we choose the functions $g(x,\bar{x}) := \log\frac{W_1(x|\bar{x})}{W_0(x|\bar{x})}$ and $h(\bar{x}) := \log\frac{P_1(\bar{x})}{P_0(\bar{x})}$. Under these choices,

$$D_{1+s}\big(W_0^{\times n}\times P_0\|W_1^{\times n}\times P_1\big) = \frac{\phi_n(-s)}{s}, \qquad D_{1+s}(W_0\|W_1) = \frac{\phi(-s)}{s}. \tag{6.11}$$

Hence, combining (6.2) ad (6.11), we obtain (6.10).

Since the relative Rényi entropy $D_{1+s}(W_0^{\times n}\times P_0\|W_1^{\times n}\times P_1)$ is monotone increasing with respect to s and $\lim_{s\to 0} D_{1+s}(W_0^{\times n}\times P_0\|W_1^{\times n}\times P_1) = D(W_0^{\times n}\times P_0\|W_1^{\times n}\times P_1)$, we have

$$\begin{aligned} D_{1-\delta}(W_0\|W_1) &= \lim_{n\to\infty}\frac{1}{n}D_{1-\delta}\big(W_0^{\times n}\times P_0\|W_1^{\times n}\times P_1\big) \\ &\le \lim_{n\to\infty}\frac{1}{n}D\big(W_0^{\times n}\times P_0\|W_1^{\times n}\times P_1\big) \\ &\le \lim_{n\to\infty}\frac{1}{n}D_{1+\delta}\big(W_0^{\times n}\times P_0\|W_1^{\times n}\times P_1\big) \\ &= D_{1+\delta}(W_0\|W_1) \end{aligned}$$

for $\delta > 0$. Since $\lim_{s\to 0} D_{1+s}(W_0\|W_1) = D(W_0\|W_1)$, we obtain (6.9). □

7. Asymptotic variance. First, we prepare the following lemma.

LEMMA 7.1 ([36]). *The cumulant generating function of the random variable* $\sqrt{n}(\frac{\tilde{g}^n(X^{n+1})}{n} - \eta(0))$ *converges as follows*:

$$\begin{aligned} &\log \mathsf{E}_0\Big[\exp\Big[\delta\sqrt{n}\Big(\frac{\tilde{g}^n(X^{n+1})}{n} - \eta(0)\Big)\Big]\Big] \\ &\quad = \phi_n\Big(\frac{\delta}{\sqrt{n}}\Big) - \delta\sqrt{n}\eta(0) \to \delta^2 \frac{1}{2}\frac{d^2\phi}{d\theta^2}(0). \end{aligned} \tag{7.1}$$

PROOF. Using (6.2) and (6.6), we have

$$\begin{aligned} \lim_{n\to\infty} \phi_n\Big(\frac{\delta}{\sqrt{n}}\Big) - \delta\sqrt{n}\eta(0) &\le \lim_{n\to\infty} n\phi\Big(\frac{\delta}{\sqrt{n}}\Big) - \delta\sqrt{n}\eta(0) + \overline{\delta}\Big(\frac{\delta}{\sqrt{n}}\Big) \\ &= \lim_{n\to\infty} \delta^2 \frac{\phi(\frac{\delta}{\sqrt{n}}) - (\frac{\delta}{\sqrt{n}})\frac{d\phi}{d\theta}(0)}{(\frac{\delta}{\sqrt{n}})^2} = \frac{\delta^2}{2}\frac{d^2\phi}{d\theta^2}(0). \end{aligned}$$

Similarly, the opposite inequality can be shown by (6.2) and (6.6). Hence, we obtain the desired relation. □

The right-hand side of (7.1) is the cumulant generating function of Gaussian distribution with the variance $\frac{d^2\phi}{d\theta^2}(0)$ and average 0. Since the limit of cumulant generating function uniquely decides the limit of the distribution function [49], Lemma 7.1 reproduces the central limit theorem as a corollary.

COROLLARY 7.2 ([4, 28, 34, 36, 40]). *The limiting distribution of* $\sqrt{n} \times (\frac{\tilde{g}^n(X^{n+1})}{n} - \eta(0))$ *is characterized as*

$$\lim_{n\to\infty} W_0^{\times n} \times P_0\{\tilde{g}^n(X^{n+1}) - n\eta(0) \le \sqrt{n}\delta\} = \Phi\Big(\frac{\delta}{\sqrt{\frac{d^2\phi}{d\theta^2}(0)}}\Big), \tag{7.2}$$

where $\Phi(y) := \int_{-\infty}^{y} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}\,dx$.

The above corollary can be regarded as the Markov version of the central limit theorem. As the refinement of the above argument, the paper [25], Theorem 2, showed the Markov version of the Berry–Esseen theorem as follows.

REMARK 7.3. Our derivation of Corollary 7.2 is much simpler than existing derivations [4, 28, 34, 40] because it employs only our evaluation of the cumulant generating function. For example, the paper [4], Theorem 4, showed the Markov version of the central limit theorem by using a martingale stopping technique.

Only Lalley [36] employ the same method as our paper for Corollary 7.2. Lalley [36] also showed the same statement as Lemma 7.1 in his proof in Theorem 1 of his paper. To show this statement, he showed a statement for an expansion of the Perron eigenvalue $\phi_n(\frac{\delta}{\sqrt{n}})$. He employed regular perturbation theory of operators on the infinite dimensional space [29], Chapter 7, #1, Chapter 4, #3, and Chapter 3, #5. Our proof is based only on more elementary mathematics. Hence, our proof is more helpful for readers who are not familiar to such an advanced mathematics.

PROPOSITION 7.4 ([25], Theorem 2). *For a given constant $\delta > 0$, there exists a constant C such that*

$$\left| W_0^{\times n} \times P_0\{\tilde{g}^n(X^{n+1}) - n\eta(0) \le \sqrt{n}\delta\} - \Phi\left(\frac{\delta}{\sqrt{V}}\right)\right| \le \frac{C}{\sqrt{n}}, \tag{7.3}$$

where V is the asymptotic variance.

For the calculation of C, see [25], Theorem 2. Since Corollary 7.2 shows that the asymptotic variance is $\frac{d^2\phi}{d\theta^2}(0)$, we can replace V by $\frac{d^2\phi}{d\theta^2}(0)$ in the above proposition. The paper [24], Lemma 5.3, also gives another expression of $\frac{d^2\phi}{d\theta^2}(0)$ as follows.

PROPOSITION 7.5. *The second derivative $\frac{d^2\phi}{d\theta^2}(0)$ is calculated as*

$$\frac{d^2\phi}{d\theta^2}(0) = \mathsf{V}_0[g(X, X')] + 2\sum_{x,\bar{x}} W(x|\bar{x})g(x,\bar{x})\frac{d\tilde{P}_\theta(\bar{x})}{d\theta}\bigg|_{\theta=0}, \tag{7.4}$$

where V_0 denotes the variance when X, X' obeys the joint distribution $W_0 \times \tilde{P}_0$.

In this paper, we give another expression of $\frac{d^2\phi}{d\theta^2}(0)$, which is easier to compute in some case than the second derivative of $\phi(\theta)$ at $\theta = 0$ and (7.4). To describe it, we define the matrices $A_{x,\bar{x}} := \tilde{P}_0(x)$, $W_{x,\bar{x}} := W(x|\bar{x})$, and the fundamental matrix $Z := (I - (W - A))^{-1}$ [30], whose existence is guaranteed by the following lemma.

PROPOSITION 7.6 ([30], Theorem 4.3.1). *For a transition matrix W, the matrix $Z = (I - (W - A))^{-1}$ exists and*

$$Z = I + \sum_{n=1}^{\infty}(W^n - A) = \sum_{n=0}^{\infty}(W - A)^n. \tag{7.5}$$

We also have $(W - A)^n = W^n - A$ for every n.

This proposition can be shown by the fact that $\lim_{n\to\infty} W^n = A$. Then we give another expression of $\frac{d^2\phi}{d\theta^2}(0)$ as follows.

THEOREM 7.7.

$$\begin{aligned}\frac{d^2\phi}{d\theta^2}(0) &= \mathsf{V}_0\big[g(X,X')\big] \\ &\quad + 2\sum_{x,\bar{x}}\Big(\sum_{\bar{x}_o} g(x,\bar{x}_o)W(x|\bar{x}_o)\Big)(Z-A)_{x,\bar{x}} \\ &\quad \times \Big(\sum_{x_o} g(\bar{x},x_o)\tilde{P}_0(x_o)W(\bar{x}|x_o)\Big).\end{aligned} \tag{7.6}$$

The proof of Theorem 7.7 is given in Appendix B. Combining (7.4), we obtain

$$\frac{d\tilde{P}_\theta(\bar{x})}{d\theta}\Big|_{\theta=0} = \sum_x\Big(\sum_{\bar{x}_o} g(x,\bar{x}_o)\tilde{P}_0(\bar{x}_o)W(x|\bar{x}_o)\Big)(Z-A)_{x,\bar{x}}. \tag{7.7}$$

REMARK 7.8. Many existing papers considered the case when $g(x,\bar{x})$ is x or $\bar{x}$. In the above case, the literatures [16, 28, 31, 40] showed the central limit theorem by using the asymptotic variance. The literature [28, 40] did not give any expression of the asymptotic variance without the infinite sum. The papers [16, 31] showed another expression by using the spectral measure under a more general setting while it is hard to calculate the spectral measure in general even in the finite state case. The literature [32], Lemma 1.5 of Chapter 1, also derived another closed form for asymptotic variance by using spectral measure. The paper [16] also showed that the variance can be expressed by the sum of covariance between X_1 and X_i, which is also not computable even in the finite state case. In the above case, the paper [34] showed the central limit theorem and the asymptotic variance equals the second derivative of the limit $\lim_{n\to\infty}\frac{\phi_n(\theta)}{n}$. However, it did not give a concrete form of the limit. In this limited case, the literature [30, 47, 56], second equation on page 199, showed that the asymptotic variance is given as the right-hand side of (7.6), and the paper [55] gave another expression for the asymptotic variance. When we apply the result by [30, 47, 56], second equation on page 199, to the transition matrix $P(g(X_{n+1},X_n)=x|g(X_n,X_{n-1})=\bar{x})$, we can derive a formula for the asymptotic variance as follows. First, we define two matrices $\hat{A}_{(x_1,x_2),(\bar{x}_3,\bar{x}_4)} := W(\bar{x}_3|\bar{x}_4)\tilde{P}_0(\bar{x}_4)$, $\hat{W}_{(x_1,x_2),(\bar{x}_3,\bar{x}_4)} := W(x_1|x_2)W(\bar{x}_3|\bar{x}_4)\delta_{x_2,\bar{x}_3}$ from $\mathcal{X}^2$ to $\mathcal{X}^2$. Then we define another matrix $\hat{Z} := (I-(\hat{W}-\hat{A}))^{-1}$. Applying their formula to our case, we have

$$\begin{aligned}\frac{d^2\phi}{d\theta^2}(0) &= \mathsf{V}_0\big[g(X,X')\big] \\ &\quad + 2\sum_{x_1,x_2,\bar{x}_3,\bar{x}_4}\Big(\sum_{\bar{x}_1,\bar{x}_2} g(x_1,x_2)W(x_1|x_2)W(\bar{x}_1|\bar{x}_2)\Big)\delta_{x_2,\bar{x}_1} \\ &\quad \times (\hat{Z}-\hat{A})_{(x_1,x_2),(\bar{x}_3,\bar{x}_4)}g(\bar{x}_3,\bar{x}_4)W(\bar{x}_3|\bar{x}_4)\tilde{P}_0(\bar{x}_4).\end{aligned} \tag{7.8}$$

The number of terms of our formula (7.6) is smaller than that of the above formula. Hence, our formula (7.6) is useful for practical calculation.

8. Tail probability. Combining Proposition A.1, Lemma 4.1 and (6.2), we can derive the following lower bound on the exponent by using the function $\phi'^{-1}(a)$, $\underline{\delta}(\theta)$, $\overline{\delta}(\theta)$ and $\phi(\theta)$ defined in (3.6), (6.3), (6.3) and (3.4).

THEOREM 8.1. *For any $a > \eta(0) = \mathsf{E}_0[g]$, we have*

$$\begin{aligned} &-\log W_0^{\times n} \times P_0\{\tilde{g}^n(X^{n+1}) \ge na\} \\ &\quad \ge \sup_{\theta \ge 0}[n\theta a - n\phi(\theta) - \overline{\delta}(\theta)] \\ &\quad = n\phi'^{-1}(a)a - n\phi(\phi'^{-1}(a)) - \overline{\delta}(\phi'^{-1}(a)) \\ &\quad = nD(W_{\phi'^{-1}(a)} \| W_0) - \overline{\delta}(\phi'^{-1}(a)). \end{aligned} \tag{8.1}$$

Similarly, for $a < \eta(0) = \mathsf{E}_0[g]$, we have

$$\begin{aligned} -\log W_0^{\times n} \times P_0\{\tilde{g}^n(X^{n+1}) \le na\} &\ge \sup_{\theta \le 0}[n\theta a - n\phi(\theta) - \overline{\delta}(\theta)] \\ &= n\phi'^{-1}(a)a - n\phi(\phi'^{-1}(a)) - \overline{\delta}(\phi'^{-1}(a)) \\ &= nD(W_{\phi'^{-1}(a)} \| W_0) - \overline{\delta}(\phi'^{-1}(a)). \end{aligned}$$

Combining Theorem A.2 and (6.2) of Lemma 6.1, we can derive the following converse bound.

THEOREM 8.2. *For any $a > \eta(0) = \mathsf{E}_0[g]$, we have*

$$\begin{aligned} &-\log W_0^{\times n} \times P_0\{\tilde{g}^n(X^{n+1}) \ge na\} \\ &\overset{(a)}{\le} \inf_{\substack{s>0 \\ \theta \in \mathbb{R}, \bar{\theta} \le 0}} \big[n\phi((1+s)\theta) - n(1+s)\phi(\theta) + \overline{\delta}((1+s)\theta) - \underline{\delta}(\theta) \\ &\qquad - (1+s)\log\big(1 - e^{-n[\bar{\theta}a - \phi(\theta+\bar{\theta}) + \phi(\theta) + \overline{\delta}(\theta+\bar{\theta}) - \underline{\delta}(\theta)]}\big)\big]/s \\ &\overset{(b)}{\le} \inf_{\substack{s>0 \\ \theta > \phi'^{-1}(a)}} \big[n\phi((1+s)\theta) - n(1+s)\phi(\theta) + \overline{\delta}((1+s)\theta) - \underline{\delta}(\theta) \\ &\qquad - (1+s)\log\big(1 - e^{n[(\theta - \phi'^{-1}(a))a + \phi(\phi'^{-1}(a)) - \phi(\theta) + \overline{\delta}(\phi'^{-1}(a)) - \underline{\delta}(\theta)]}\big)\big]/s \\ &\overset{(c)}{=} \inf_{\substack{s>0 \\ \theta > \phi'^{-1}(a)}} nD_{1+s}(W_\theta \| W_0) + \frac{1}{s}\big[\overline{\delta}((1+s)\theta) - \underline{\delta}(\theta)\big] \\ &\qquad - \frac{1+s}{s}\log\big(1 - e^{-nD(W_{\phi'^{-1}(a)} \| W_\theta) + \overline{\delta}(\phi'^{-1}(a)) - \underline{\delta}(\theta)}\big). \end{aligned} \tag{8.2}$$

Similarly, for any $a < \eta(0) = \mathsf{E}_0[g]$, *we have*

$$
\begin{aligned}
&-\log W_0^{\times n} \times P_0\{\tilde{g}^n(X^{n+1}) \le na\} \\
&\quad \le \inf_{\substack{s>0 \\ \theta\in\mathbb{R},\bar{\theta}\ge 0}} \big[n\phi\big((1+s)\theta\big) - n(1+s)\phi(\theta) + \overline{\delta}\big((1+s)\theta\big) - \underline{\delta}(\theta) \\
&\qquad - (1+s)\log\big(1 - e^{-n[\bar{\theta}a - \phi(\theta+\bar{\theta}) + \phi(\theta) + \overline{\delta}(\theta+\bar{\theta}) - \underline{\delta}(\theta)]}\big)\big]/s \\
&\quad \le \inf_{\substack{s>0 \\ \theta<\phi'^{-1}(a)}} \big[n\phi\big((1+s)\theta\big) - (n-1)(1+s)\phi(\theta) + \overline{\delta}\big((1+s)\theta\big) - \underline{\delta}(\theta) \\
&\qquad - (1+s)\log\big(1 - e^{n[(\theta-\phi'^{-1}(a))a + \phi(\phi'^{-1}(a)) - \phi(\theta) + \overline{\delta}(\phi'^{-1}(a)) - \underline{\delta}(\theta)]}\big)\big]/s \\
&\quad = \inf_{\substack{s>0 \\ \theta<\phi'^{-1}(a)}} nD_{1+s}(W_\theta \| W_0) + \frac{1}{s}\big[\overline{\delta}\big((1+s)\theta\big) - \underline{\delta}(\theta)\big] \\
&\qquad - \frac{1+s}{s}\log\big(1 - e^{-nD(W_{\phi'^{-1}(a)}\|W_\theta) + \overline{\delta}(\phi'^{-1}(a)) - \underline{\delta}(\theta)}\big).
\end{aligned}
$$

PROOF. (a) Follows from the combination of (a) of Theorem A.2 and (6.2) of Lemma 6.1. (b) and (c) can be shown by the same way as (b) and (c) of Theorem A.2. □

Due to the expressions in Theorems 8.1 and 8.2, the above upper and lower bounds are $O(1)$-computable. These also attain the asymptotic tightness in the sense of (T2) and (T3) as shown in the following corollaries. Indeed, the paper [33] also derived the lower bound of $-\log W_0^{\times n} \times P_0\{\tilde{g}^n(X^{n+1}) \ge na\}$ in a more general setting including infinite state spaces and the continuous case. Since the finite-length bound in the paper [33] contains so many parameters, it is difficult to characterize the difference between their lower bound and the leading term of the true value, that is, $nD(W_{\phi'^{-1}(a)}\|W_0)$. They showed only that the difference is sublinear for n. In contrast, the difference in our lower bound is clearly shown to be the constant $\overline{\delta}(\phi'^{-1}(a))$ in (8.1) of Theorem 8.1. So, our bound can be applied to the moderate deviation as in Corollary 8.4. However, it is not clear whether their bound can be applied to it. Hence, our lower bound is better than theirs.

From Lemma 4.1 and Theorems 8.1 and 8.2, we can derive the large deviation evaluation.

COROLLARY 8.3 ([11], [10], Theorem 3.1.2). *For arbitrary* $\delta > 0$, *we have*

$$
\begin{aligned}
&\lim_{n\to\infty} -\frac{1}{n}\log W_0^{\times n} \times P_0\{\tilde{g}^n(X^{n+1}) - n\eta(0) \ge \delta n\} \\
&\quad = \sup_{\theta\ge 0}\big[\phi'^{-1}\big(\eta(0)+\delta\big) - \phi(\theta)\big],
\end{aligned}
$$

$$\lim_{n\to\infty} -\frac{1}{n}\log W_0^{\times n}\times P_0\{\tilde{g}^n(X^{n+1}) - n\eta(0) \le -\delta n\} = \sup_{\theta\le 0}[\phi'^{-1}(\eta(0)-\delta) - \phi(\theta)].$$

From Theorems 8.1 and 8.2, we can derive the moderate deviation evaluation.

COROLLARY 8.4. *For arbitrary $t \in (0, 1/2)$ and $\delta > 0$, we have*

$$\lim_{n\to\infty} -\frac{1}{n^{1-2t}}\log W_0^{\times n}\times P_0\{\tilde{g}^n(X^{n+1}) - n\eta(0) \ge n^{1-t}\delta\} = \frac{\delta^2}{2\frac{d^2\phi}{d\theta^2}(0)}, \tag{8.3}$$

$$\lim_{n\to\infty} -\frac{1}{n^{1-2t}}\log W_0^{\times n}\times P_0\{\tilde{g}^n(X^{n+1}) - n\eta(0) \le -n^{1-t}\delta\} = \frac{\delta^2}{2\frac{d^2\phi}{d\theta^2}(0)}. \tag{8.4}$$

PROOF. We only prove (8.3). To show the inequality $\ge$ in (8.3), we employ (8.1). That is, we substitute $a_n := \eta(0) + \frac{\delta}{n^t}$ into a in (8.3). Since $\frac{d\phi'^{-1}(\eta)}{d\eta} = \frac{1}{\phi''(\phi'^{-1}(\eta))}$, we have $\phi'^{-1}(a_n) = \frac{\delta}{\frac{d^2\phi}{d\theta^2}(0)n^t} + o(\frac{1}{n^t}) \to 0$. Thus, Relation (6.6) implies $\overline{\delta}(\phi'^{-1}(a_n))) \to 0$. Hence, relation (3.12) yields that

$$\frac{1}{n^{1-2t}}\big(nD(W_{\phi'^{-1}(a_n)}\|W_0) - \overline{\delta}(\phi'^{-1}(a_n))\big) \to \frac{\delta^2}{2\frac{d^2\phi}{d\theta^2}(0)}. \tag{8.5}$$

Applying (8.5) to (8.1), we obtain the part "$\ge$" in (8.3).

To show the inequality $\le$ in (8.3), we employ the final term of (8.2). That is, we substitute $a_n := \eta(0) + \frac{\delta}{n^t}$ and $\theta_n := \phi'^{-1}(a_n) + \frac{\xi}{n^t}\frac{d^2\phi}{d\theta^2}(0)^{-1}$ into a and θ in the final term of (8.2). Then, we have $\theta_n = \frac{\delta+\xi}{\frac{d^2\phi}{d\theta^2}(0)n^t} + o(\frac{1}{n^t}) \to 0$. Thus, relation (6.6) implies that $\frac{1}{s}[\overline{\delta}((1+s)\theta_n) - \underline{\delta}(\theta_n)] \to 0$ and $\overline{\delta}(\phi'^{-1}(a_n)) - \underline{\delta}(\theta_n) \to 0$. We also have $nD(W_{\phi'^{-1}(a_n)}\|W_{\theta_n}) \to \infty$. Hence, (3.13) yields that

$$\begin{aligned}
&\lim_{n\to\infty}\frac{1}{n^{1-2t}}\Bigg[\inf_{\substack{s>0\\ \theta>\phi'^{-1}(a_n)}} nD_{1+s}(W_\theta\|W_0) + \frac{1}{s}\big[\overline{\delta}\big((1+s)\theta\big) - \underline{\delta}(\theta)\big] \\
&\quad - \frac{1+s}{s}\log\big(1 - e^{-nD(W_{\phi'^{-1}(a_n)}\|W_\theta) + \overline{\delta}(\phi'^{-1}(a_n)) - \underline{\delta}(\theta)}\big)\Bigg] \\
&\le \lim_{n\to\infty}\frac{1}{n^{1-2t}}\Bigg[nD_{1+s}(W_{\theta_n}\|W_0) + \frac{1}{s}\big[\overline{\delta}\big((1+s)\theta_n\big) - \underline{\delta}(\theta_n)\big] \\
&\quad - \frac{1+s}{s}\log\big(1 - e^{-nD(W_{\phi'^{-1}(a_n)}\|W_{\theta_n}) + \overline{\delta}(\phi'^{-1}(a_n)) - \underline{\delta}(\theta_n)}\big)\Bigg] \\
&= \lim_{n\to\infty}\frac{n}{n^{1-2t}}D_{1+s}(W_{\theta_n}\|W_0) = \frac{(\delta+\xi)^2}{2\frac{d^2\phi}{d\theta^2}(0)}(1+s).
\end{aligned} \tag{8.6}$$

Finally, we take the limits $\xi \to 0$ and $s \to 0$. Then, applying (8.6) to (8.2), we obtain the part "$\ge$" in (8.3). □

9. Simple hypothesis testing. Next, we consider the binary simple hypothesis testing. To formulate the binary simple hypothesis testing, we consider the case that the null and alternative hypotheses are P_1 and P_0. For theoretical simplicity, we often focus on randomized tests, which is the generalization of the conventional test. Although the conventional test is given as a $\{0,1\}$-valued function of the observed data, a randomized test is given as a $[0,1]$-valued function $T(x)$ of the observed data x. When we observe $T(x) = t$, we support the null hypothesis P_1 with probability t and support the alternative hypothesis P_0 with probability $1-t$. Then the type-1 and type-2 error probabilities are given as $\mathsf{E}_{P_1}[1-T]$ and $\mathsf{E}_{P_0}[T]$, where E_{P_i} denotes the expectation under the distribution P_i. When we choose the random variable T to be the test function with support S, the random variable T realizes the test whose rejection region S.

Then we consider the following value:

$$\begin{aligned}\beta_\varepsilon(P_1\|P_0) &:= \min_T\{\mathsf{E}_{P_0}[T] | \mathsf{E}_{P_1}[1-T] \le \varepsilon\} \\ &= \min_T\{\mathsf{E}_{P_0}[T] | \mathsf{E}_{P_1}[1-T] = \varepsilon\}.\end{aligned} \tag{9.1}$$

Since we allow randomized tests, the optimum test T is realized with the condition $\mathsf{E}_{P_1}[1-T] = \varepsilon$.

Now, we consider the hypothesis testing with the two hypotheses $W_0^{\times n} \times P_0$ and $W_1^{\times n} \times P_1$. Then we choose the functions $g(x,\bar{x}) := \log \frac{W_1(x|\bar{x})}{W_0(x|\bar{x})}$ and $h(\bar{x}) := \log \frac{P_1(\bar{x})}{P_0(\bar{x})}$, which implies that $\phi(1) = 0$. Under these choices, we can evaluate the minimum type-2 error probability in the following way by using the functions $\hat{\theta}(r)$, $\underline{\delta}(\theta)$, $\overline{\delta}(\theta)$ and $\phi(\theta)$ defined in (4.2), (6.3), (6.3) and (3.4) as well as the relative entropies $D(W_{\hat{\theta}(r)}\|W_\theta)$ and $D_{1+s}(W_\theta\|W_0)$.

THEOREM 9.1. *The minimum type-2 error probability* $\beta_{e^{-nr}}(W_1^{\times n} \times P_1 \| W_0^{\times n} \times P_0)$ *defined in* (2.6) *satisfies*

$$\begin{aligned}&\sup_{0\le\theta\le 1} \frac{n(-\theta r - \phi(\theta)) - \underline{\delta}(\theta)}{1-\theta} \\ &\quad\le -\log \beta_{e^{-nr}}\big(W_1^{\times n} \times P_1 \| W_0^{\times n} \times P_0\big) \\ &\quad\overset{(a)}{\le} \inf_{\bar{\theta}\ge 0, s>0, \theta\in(0,1)} \frac{1}{s}\big[n\big(\phi\big((1+s)\theta\big) - (1+s)\phi(\theta)\big) \\ &\qquad + \big(\overline{\delta}\big((1+s)\theta\big) - (1+s)\underline{\delta}(\theta)\big) \\ &\qquad - (1+s)\log\big(1 - 2e^{n\frac{-(1+\bar{\theta})\phi(\theta)+\phi((1+\bar{\theta})\theta-\bar{\theta})-\bar{\theta}r}{1+\bar{\theta}} + \frac{-(1+\bar{\theta})\underline{\delta}(\theta)+\overline{\delta}((1+\bar{\theta})\theta-\bar{\theta})}{1+\bar{\theta}}}\big)\big]\end{aligned}$$

$$\begin{aligned}
&\overset{\text{(b)}}{\le} \inf_{s>0,\theta\in(\hat\theta(r),1)} \frac{1}{s}\big[n\big(\phi\big((1+s)\theta\big)-(1+s)\phi(\theta)\big)\\
&\quad+\big(\overline{\delta}\big((1+s)\theta\big)-(1+s)\underline{\delta}(\theta)\big)\\
&\quad-(1+s)\log\big(1-2e^{n(-\phi(\theta)+\phi(\hat\theta(r))+(\theta-\hat\theta(r))\frac{d\phi}{d\theta}(\hat\theta(r)))-\underline{\delta}(\theta)+\frac{(1-\theta)\overline{\delta}(\hat\theta(r))}{1-\hat\theta(r)}}\big)\big]\\
&\overset{\text{(c)}}{=} \inf_{s>0,\theta\in(\hat\theta(r),1)} nD_{1+s}(W_\theta\|W_0)+\frac{1}{s}\big(\overline{\delta}\big((1+s)\theta\big)-(1+s)\underline{\delta}(\theta)\big)\\
&\quad-\frac{1+s}{s}\log\big(1-2e^{-nD(W_{\hat\theta(r)}\|W_\theta)-\underline{\delta}(\theta)+\frac{(1-\theta)\overline{\delta}(\hat\theta(r))}{1-\hat\theta(r)}}\big).
\end{aligned}$$

PROOF. The inequality (a) can be shown by combining (A.9) and (6.2). To show (b) and (c), we restrict θ in $[\hat\theta(r),1]$ and choose $\bar\theta$ to be $\frac{\theta-\hat\theta(r)}{1-\theta}\ge 0$ similar to the proof of (A.10). Then

$$\begin{aligned}
\frac{-(1+\bar\theta)\underline{\delta}(\theta)+\overline{\delta}((1+\bar\theta)\theta-\bar\theta)}{1+\bar\theta}&=-\underline{\delta}(\theta)+\frac{\overline{\delta}(\hat\theta(r))}{1+\bar\theta}\\
&=-\underline{\delta}(\theta)+\frac{(1-\theta)\overline{\delta}(\hat\theta(r))}{1-\hat\theta(r)}.
\end{aligned}$$

As is shown in the proof of (A.10), we have

$$\begin{aligned}
&\frac{-(1+\bar\theta)\phi(\theta)+\phi((1+\bar\theta)\theta-\bar\theta)-\bar\theta r}{1+\bar\theta}\\
&\quad=-\phi(\theta)+\phi\big(\hat\theta(r)\big)+\big(\theta-\hat\theta(r)\big)\frac{d\phi}{d\theta}\big(\hat\theta(r)\big)\\
&\quad=D(W_{\hat\theta(r)}\|W_\theta).
\end{aligned}$$

Hence, we obtain (b) and (c). □

Due to the expressions in Theorem 9.1, the above upper and lower bounds are $O(1)$-computable. These also attain the asymptotic tightness in the sense of (H2) and (H3) as follows. From Lemma 4.2 and Theorem 9.1, we can recover the Hoeffding-type evaluation as follows.

COROLLARY 9.2 ([46], Theorem 2, [45], Theorem 1).

$$\begin{aligned}
&\lim_{n\to\infty}-\frac{1}{n}\log\beta_{e^{-nr}}\big(W_1^{\times n}\times P_1\|W_0^{\times n}\times P_0\big)\\
&\quad=\sup_{0\le\theta\le1}\frac{-\theta r-\phi(\theta)}{1-\theta}=\sup_{0\le\theta\le1}\frac{\theta(-r+D_{1-\theta}(W_0\|W_1))}{1-\theta}
\end{aligned}\tag{9.2}$$

$$= \inf_{s>0,\theta\in(0,\hat{\theta}(r))} D_{1+s}(W_\theta\|W_0) = D(W_{\hat{\theta}(r)}\|W_0)$$
$$= \min_{W:D(W\|W_1)\le r} D(W\|W_0).$$

REMARK 9.3. Natarajan [46], Theorem 2, showed that the exponent (9.2) equals $\min_{W:D(W\|W_1)\le r} D(W\|W_0)$. Nakagawa and Kanaya [45], Theorem 1, showed that the exponent (9.2) equals $D(W_{\hat{\theta}(r)}\|W_0)$. They did not consider other expressions in (9.2).

From Theorem 9.1, we obtain the following moderate deviation type evaluation.

COROLLARY 9.4. *For $t \in (0, \frac{1}{2})$, we have*

$$\lim_{n\to\infty} -\frac{1}{n^{1-2t}} \log \beta_{e^{-nD(W_0\|W_1)+n^{1-t}\delta}}\bigl(W_1^{\times n}\times P_1\|W_0^{\times n}\times P_0\bigr) = \frac{\delta^2}{2\frac{d^2\phi}{d\theta^2}(0)}. \tag{9.3}$$

That is,

$$\begin{aligned} &-\log \beta_{e^{-n^{1-2t}r}}\bigl(W_0^{\times n}\times P_0\|W_1^{\times n}\times P_1\bigr) \\ &\quad = D(W_0\|W_1)n - \sqrt{2\frac{d^2\phi}{d\theta^2}(0)rn^{1-t}} + o\bigl(n^{1-t}\bigr). \end{aligned} \tag{9.4}$$

PROOF. First, we show (9.3) in the same way as the proof of (8.3). (4.1) implies

$$\sup_{0\le\theta\le 1} \frac{n(-\theta r-\phi(\theta))-\underline{\delta}(\theta)}{1-\theta} \ge \frac{n(-\phi'^{-1}(-r)r-\phi(\phi'^{-1}(-r)))-\underline{\delta}(\phi'^{-1}(-r))}{1-\phi'^{-1}(-r)}$$
$$= \frac{nD(W_{\phi'^{-1}(-r)}\|W_0)-\underline{\delta}(\phi'^{-1}(-r))}{1-\phi'^{-1}(-r)}.$$

Now, we choose $r_n := D(W_0\|W_1)+\delta n^{-t}$. Then we have $\phi'^{-1}(-r_n) = \frac{\delta}{2\frac{d^2\phi}{d\theta^2}(0)n^t} + o(\frac{1}{n^t}) \to 0$. Thus,

$$\frac{nD(W_{\phi'^{-1}(-r_n)}\|W_0)-\underline{\delta}(\phi'^{-1}(-r_n))}{1-\phi'^{-1}(-r_n)} \to \frac{\delta^2}{2\frac{d^2\phi}{d\theta^2}(0)}. \tag{9.5}$$

Applying (9.5) to Theorem 9.1, we obtain the part "$\ge$" in (9.3).

Next, we choose $\theta_n := \phi'^{-1}(a_n) + \frac{\xi}{n^t}\frac{d^2\phi}{d\theta^2}(0)^{-1}$. Then, applying the right-hand side of (c) of Theorem 9.1, we obtain the part "$\le$" in (9.3) as the same way as the proof of the part "$\le$" in (8.3).

Solving $\frac{\delta^2}{2\frac{d^2\phi}{d\theta^2}(0)} = r\delta^2$, we have $\delta = \sqrt{2\frac{d^2\phi}{d\theta^2}(0)r}$. Hence, we have (9.4) from (9.3). □

We also have another type evaluation for the type-2 error probability.

LEMMA 9.5. *When we choose* $g(x,\bar{x}) = \log\frac{W_1(x|\bar{x})}{W_0(x|\bar{x})}$ *and* $\hat{g}(x) = \log\frac{P_1(x)}{P_0(x)}$, *we have*

$$\begin{aligned}
&\sup_a\{a|W_1^{\times n}\times P_1\{g^n(X^{n+1}) < a\} \le \varepsilon\} \\
&\qquad \le -\log\beta_\varepsilon(W_1^{\times n}\times P_1\|W_0^{\times n}\times P_0) \\
&\qquad \le \inf_{\delta>0,a}\{a - \log\delta|W_1^{\times n}\times P_1\{g^n(X^{n+1}) < a\} \ge \varepsilon+\delta\}.
\end{aligned} \tag{9.6}$$

Lemma 9.5 can be shown by substituting $W_i^{\times n}\times P_i$ into P_i $(i=0,1)$ in Lemma A.3 in Appendix A.

Combining (7.3) and (9.6), we can derive lower and upper $O(1)$-computable bounds for $\beta_\varepsilon(W_1^{\times n}\times P_1\|W_0^{\times n}\times P_0)$. Applying Corollary 7.2 to the random variable $\log\frac{W_1^n\times P_1(X^{n+1})}{W_0^n\times P_0(X^{n+1})}$ in Lemma 9.5, we obtain the Stein–Strassen-type evaluation. That is, these bounds attain the asymptotic tightness in the sense of (H1).

THEOREM 9.6.

$$\begin{aligned}
&-\log\beta_\varepsilon(W_1^{\times n}\times P_1\|W_0^{\times n}\times P_0) \\
&\qquad = nD(W_1\|W_0) + \sqrt{n}\sqrt{\frac{d^2\phi}{d\theta^2}(1)}\Phi^{-1}(\varepsilon) + o(\sqrt{n}).
\end{aligned}$$

10. Conclusion. We have derived upper and lower $O(1)$-computable bounds of the cumulant generating function of the Markov chain by using the convex function $\phi(\theta)$. Using these bounds, we have given a simple alternative proof of the central limit theorem of the sample mean in the Markovian chain. Also, using these bounds, we have derived upper and lower $O(1)$-computable bounds of the tail probability of the sample mean, which attains the asymptotic tightness in the sense of (T2) and (T3). Using the above upper and lower bounds, we have derived upper and lower $O(1)$-computable bounds of the minimum error probability of the type-2 error under the constraint for the error probability of the type-1 error, which attains the asymptotic tightness in the sense of (H2) and (H3). These bounds have not been derived even in the independently and identically distributed case. We have also derived other upper and lower $O(1)$-computable bounds that attains the asymptotic tightness in the sense of (H1).

However, in this paper, we have assumed a finite state space. Indeed, the existing papers [1, 2, 37] reported several difficulties to evaluate the tail probability of the sample mean in the continuous probability space even with the discrete time Markov chain. Also, the existing papers [7, 14, 15] reported several examples, in which, the central limit theorem does not hold even on countable state space with the discrete time Markov chain. So, it is remained as a challenging problem to extend the obtained results to such general cases. In this generalization, to avoid such counterexamples, we need to find suitable conditions for such extension. Therefore, we can expect that this extension enables us to handle several Gaussian Markovian chains with discrete time in a simple way. Further, the obtained bounds are useful for several topics in information theory [21].

APPENDIX A: EXPONENTIAL FAMILY OF DISTRIBUTIONS

In this Appendix, we discuss several formulas in an exponential family of distributions $\{P_\theta\}$ with single observation when $P_\theta(x) := P(x)e^{\theta x - \phi(\theta)}$ with cumulant generating function $\phi(\theta) := \log \sum_x P(x)[e^{\theta x}]$.

The exponential families of transition matrices contain exponential families of distributions by considering the family of transition matrices $W_\theta(x|\bar{x}) := P_\theta(x)$ from the family of distributions P_θ. Hence, the definitions and the notation given in Section 3 are applied to the exponential family of distributions $\{P_\theta\}$ in the following.

A.1. Tail probability. First, we define the relative entropy and the relative Rényi entropy between two distributions P and $\bar{P}$ are given as

$$D(P\|\bar{P}) := \sum_x P(x) \log \frac{P(x)}{\bar{P}(x)}, \tag{A.1}$$

$$D_{1+s}(P\|\bar{P}) := \frac{1}{s} \log \sum_x P(x)^{1+s} \bar{P}(x)^{-s}. \tag{A.2}$$

Using the cumulant generating function $\phi(\theta)$, we investigate the lower bound on the tail probability as follows. The following lower bound on the tail probability is nothing but Cramér's theorem in the large deviation theory [10].

PROPOSITION A.1. *For any $a > \mathsf{E}[X]$, we have*

$$\begin{aligned} -\log P_0\{X \ge a\} &\ge \sup_{\theta \ge 0}[\theta a - \phi(\theta)] = \phi'^{-1}(a)a - \phi(\phi'^{-1}(a)) \\ &= D(P_{\phi'^{-1}(a)}\|P_0). \end{aligned}$$

Similarly, for $a < \mathsf{E}[X]$, we have

$$\begin{aligned} -\log P_0\{X \le a\} &\ge \sup_{\theta \le 0}[\theta a - \phi(\theta)] = \phi'^{-1}(a)a - \phi(\phi'^{-1}(a)) \\ &= D(P_{\phi'^{-1}(a)}\|P_0). \end{aligned}$$

By using the monotonicity of the Rényi relative entropy [8], we can derive the following converse bound.

THEOREM A.2. *For any $a > \mathsf{E}[X]$, we have*

$$
\begin{aligned}
&-\log P\{X \geq a\} \\
&\quad \overset{\text{(a)}}{\leq} \inf_{\substack{s>0 \\ \theta\in\mathbb{R},\bar{\theta}\leq 0}} \big[\phi\big((1+s)\theta\big) - (1+s)\phi(\theta) \\
&\qquad - (1+s)\log\big(1 - e^{-[\bar{\theta}a - \phi(\theta+\bar{\theta})+\phi(\theta)]}\big)\big]/s \\
&\quad \overset{\text{(b)}}{\leq} \inf_{\substack{s>0 \\ \theta>\phi'^{-1}(a)}} \big[\phi\big((1+s)\theta\big) - (1+s)\phi(\theta) \\
&\qquad - (1+s)\log\big(1 - e^{(\theta-\phi'^{-1}(a))a+\phi(\phi'^{-1}(a))-\phi(\theta)}\big)\big]/s \\
&\quad \overset{\text{(c)}}{=} \inf_{\substack{s>0 \\ \theta>\phi'^{-1}(a)}} D_{1+s}(P_\theta \| P_0) \\
&\qquad - \frac{1+s}{s}\log\big(1 - e^{-D(P_{\phi'^{-1}(a)}\|P_\theta)}\big).
\end{aligned}
$$

Similarly, for any $a < \mathsf{E}[X]$, we have

$$
\begin{aligned}
&-\log P\{X \leq a\} \\
&\quad \overset{\text{(d)}}{\leq} \inf_{\substack{s>0 \\ \theta\in\mathbb{R},\bar{\theta}\geq 0}} \big[\phi\big((1+s)\theta\big) - (1+s)\phi(\theta) \\
&\qquad - (1+s)\log\big(1 - e^{-[\bar{\theta}a - \phi(\theta+\bar{\theta})+\phi(\theta)]}\big)\big]/s \\
&\quad \overset{\text{(e)}}{\leq} \inf_{\substack{s>0 \\ \theta<\phi'^{-1}(a)}} \big[\phi\big((1+s)\theta\big) \\
&\qquad - (1+s)\phi(\theta) - (1+s)\log\big(1 - e^{(\theta-\phi'^{-1}(a))a+\phi(\phi'^{-1}(a))-\phi(\theta)}\big)\big]/s \\
&\quad \overset{\text{(f)}}{=} \inf_{\substack{s>0 \\ \theta<\phi'^{-1}(a)}} D_{1+s}(P_\theta \| P_0) - \frac{1+s}{s}\log\big(1 - e^{-D(P_{\phi'^{-1}(a)}\|P_\theta)}\big).
\end{aligned}
$$

PROOF. We only show (a)–(c). We can show (d)–(f) almost in a similar manner. For arbitrary $\theta \in \mathbb{R}$, we set $\alpha := P\{X \geq a\}$ and $\beta := P_\theta\{X \geq a\}$. Then, by the monotonicity of the Rényi relative entropy [8], we have

$$
D_{1+s}(P_\theta \| P) \geq \frac{1}{s}\log\big[\beta^{1+s}\alpha^{-s} + (1-\beta)^{1+s}(1-\alpha)^{-s}\big] \geq \frac{1}{s}\log\beta^{1+s}\alpha^{-s}.
$$

Thus, we have

$$-\log\alpha \le \frac{\phi((1+s)\theta)-(1+s)\phi(\theta)-(1+s)\log\beta}{s}.$$

Now, for any $\bar\theta \le 0$, we have

$$\begin{aligned}1-\beta = P_\theta\{X<a\} &\le \sum_x P_\theta(x)e^{\bar\theta(x-a)}\\ &= \sum_x P(x)e^{(\theta+\bar\theta)x-\bar\theta a-\phi(\theta)} = e^{-[\bar\theta a-\phi(\theta+\bar\theta)+\phi(\theta)]}.\end{aligned}$$

Thus, $-\log\alpha \le f(s,\theta,\bar\theta)$, where $f(s,\theta,\bar\theta)$ is the function inside of the right-hand side of (a). Hence, we have (a).

Restricting the range of θ as $\theta > \phi'^{-1}(a)$, we have $\inf_{s>0,\theta\in\mathbb{R},\bar\theta\ge 0} f(s,\theta,\bar\theta) \le \inf_{s>0,\theta>\phi'^{-1}(a),\bar\theta\ge 0} f(s,\theta,\bar\theta)$. This restriction yields

$$\sup_{\bar\theta\le 0}\bigl[\bar\theta a-\phi(\theta+\bar\theta)+\phi(\theta)\bigr] = \bigl(\phi'^{-1}(a)-\theta\bigr)a-\phi\bigl(\phi'^{-1}(a)\bigr)+\phi(\theta),$$

which is achieved by $\bar\theta = \phi'^{-1}(a)-\theta$. Thus, since $\inf_{s>0,\theta>\phi'^{-1}(a),\bar\theta\ge 0} f(s,\theta,\bar\theta)$ equals the right-hand side of (b), we have (b). Furthermore, (c) can be obtained from the relations (A.1) and (A.2). □

A.2. Simple hypothesis testing. For simple hypothesis testing, we have the following lemma for the null and alternative hypotheses are P_0 and P_1. In fact, when two distributions P and Q are given on the probability space $\mathcal{X}$, the one-parametric exponential family P_θ generated by the random variable $Y := \log\frac{Q(X)}{P(X)}$ satisfies that $P_0 = P$ and $P_1 = Q$. Hence, the above case covers the most general setting for the binary hypothesis testing.

LEMMA A.3.

$$\begin{aligned}\sup_a\left\{a\,\middle|\,P_1\left\{\log\frac{P_1(x)}{P_0(x)}<a\right\}\le\varepsilon\right\} &\le -\log\beta_\varepsilon(P_1\|P_0)\\ &\le \inf_{\delta>0,a}\left\{a-\log\delta\,\middle|\,P_1\left\{\log\frac{P_1(x)}{P_0(x)}<a\right\}\ge\varepsilon+\delta\right\}.\end{aligned}$$

PROOF. Let S_a be the set $\{\log\frac{P_1(x)}{P_0(x)}<a\} = \{P_1(x)<e^aP_0(x)\}$ and T_a be the test function with the support S_a. When $\mathsf{E}_{P_1}[T_a]\le\varepsilon$,

$$\begin{aligned}e^{-a} &\ge e^{-a}P_1\left\{\log\frac{P_1(x)}{P_0(x)}\ge a\right\} \ge P_0\left\{\log\frac{P_1(x)}{P_0(x)}\ge a\right\}\\ &= \mathsf{E}_{P_0}[1-T_a]\ge\beta_\varepsilon(P_1\|P_0).\end{aligned}\tag{A.3}$$

Taking the logarithm, we have

$$a \le -\log \beta_\varepsilon(P_1 \| P_0). \tag{A.4}$$

Taking the supremum for a, we obtain the first inequality.

Assume that $P_1\{\log \frac{P_1(x)}{P_0(x)} < a\} \ge \varepsilon + \delta$. We have

$$\begin{aligned} \varepsilon + e^a \mathsf{E}_{P_0}[1 - T] &= \mathsf{E}_{P_1}[T] + e^a \mathsf{E}_{P_0}[1 - T] \\ &\ge \mathsf{E}_{P_1}[T_a] + e^a \mathsf{E}_{P_0}[1 - T_a] \ge \varepsilon + \delta. \end{aligned} \tag{A.5}$$

Thus,

$$\mathsf{E}_{P_0}[1 - T] \ge e^{-a}\delta. \tag{A.6}$$

Taking the minimum for T, we have $\beta_\varepsilon(P_1 \| P_0) \ge e^{-a}\delta$, which implies that

$$-\log \beta_\varepsilon(P_1 \| P_0) \le a - \log \delta. \tag{A.7}$$

Taking the infimum for $a, \delta > 0$, we obtain the second inequality. □

Here, we employ $\hat{\theta}(r) = \hat{\theta}[\phi](r)$ defined at (4.2) for a convex function ϕ. Then, modifying Proposition A.1, we have the following lemma.

LEMMA A.4. *We have*

$$\begin{aligned} &-\log P_1\{Y \le \eta(\hat{\theta}(r))\} \ge D(P_{\hat{\theta}(r)} \| P_1) = r, \\ &-\log P_0\{Y \ge \eta(\hat{\theta}(r))\} \ge D(P_{\hat{\theta}(r)} \| P_0). \end{aligned}$$

Choosing the rejection region $\{Y \le \eta(\hat{\theta}(r))\}$, *we have*

$$-\log \beta_{e^{-r}}(P_1 \| P_0) \ge \sup_{0 \le \theta \le 1} \frac{-\theta r - \phi(\theta)}{1 - \theta}. \tag{A.8}$$

As the opposite inequality, we have the following lemma.

LEMMA A.5.

$$\begin{aligned} &-\log \beta_{e^{-r}}(P_1 \| P_0) \\ &\le \inf_{\bar{\theta} \ge 0, s > 0, \theta \in (0,1)} \frac{1}{s}\Big[\phi\big((1+s)\theta\big) - (1+s)\phi(\theta) \\ &\qquad - (1+s)\log\Big(1 - 2e^{\frac{-(1+\bar{\theta})\phi(\theta) + \phi((1+\bar{\theta})\theta - \bar{\theta}) - \bar{\theta} r}{1+\bar{\theta}}}\Big)\Big] \end{aligned} \tag{A.9}$$

$$\begin{aligned} &\le \inf_{s > 0, \theta \in (\hat{\theta}(r), 1)} \frac{1}{s}\Big[\phi\big((1+s)\theta\big) - (1+s)\phi(\theta) \\ &\qquad - (1+s)\log\Big(1 - 2e^{-\phi(\theta) + \phi(\hat{\theta}(r)) + (\theta - \hat{\theta}(r))\frac{d\phi}{d\theta}(\hat{\theta}(r))}\Big)\Big] \end{aligned} \tag{A.10}$$

$$= \inf_{s > 0, \theta \in (\hat{\theta}(r), 1)} D_{1+s}(P_\theta \| P_0) - \frac{1+s}{s}\log\Big(1 - 2e^{-D(P_{\hat{\theta}(r)} \| P_\theta)}\Big). \tag{A.11}$$

PROOF. We choose the rejection region S as $P_1(S) \le e^{-r}$. The monotonicity of relative Rényi entropy [8] implies that

$$\begin{aligned} D_{1+s}(P_\theta \| P_0) &\ge \frac{1}{s} \log\bigl[P_\theta(S)^{1+s} P_0(S)^{-s} + \bigl(1 - P_\theta(S)\bigr)^{1+s} \bigl(1 - P_0(S)\bigr)^{-s}\bigr] \\ &\ge \frac{1}{s} \log\bigl[\bigl(1 - P_\theta(S)\bigr)^{1+s} \bigl(1 - P_0(S)\bigr)^{-s}\bigr] \\ &= -\log\bigl(1 - P_0(S)\bigr) + \frac{1+s}{s} \log\bigl(1 - P_\theta(S)\bigr) \end{aligned}$$

for $s > 0$. Hence, we have

$$\begin{aligned} -\log\bigl(1 - P_0(S)\bigr) &\le D_{1+s}(P_\theta \| P_0) - \frac{1+s}{s} \log\bigl(1 - P_\theta(S)\bigr) \\ &= \frac{1}{s}\bigl[\phi\bigl((1+s)\theta\bigr) - (1+s)\phi(\theta) - (1+s)\log\bigl(1 - P_\theta(S)\bigr)\bigr]. \end{aligned} \tag{A.12}$$

Next, we focus on the inequality

$$\bigl(1 - P_\theta(S)\bigr) + e^{\gamma} P_1(S) \ge P_\theta\left\{\log \frac{P_1(x)}{P_\theta(x)} \ge -\gamma\right\} + e^{\gamma} P_1\left\{\log \frac{P_1(x)}{P_\theta(x)} < -\gamma\right\},$$

which implies that

$$\bigl(1 - P_\theta(S)\bigr) + e^{\gamma} P_1(S) \ge P_\theta\left\{\log \frac{P_1(x)}{P_\theta(x)} \ge -\gamma\right\}.$$

Hence,

$$P_\theta\left\{\log \frac{P_1(x)}{P_\theta(x)} < -\gamma\right\} + e^{\gamma - r} \ge P_\theta(S).$$

For any $\bar{\theta} \ge 0$, we have

$$\begin{aligned} P_\theta\left\{\log \frac{P_1(x)}{P_\theta(x)} < -\gamma\right\} &\le \sum_x P_\theta(x)^{1+\bar{\theta}} P_1(x)^{-\bar{\theta}} e^{-\bar{\theta}\gamma} \\ &= e^{-(1+\bar{\theta})\phi(\theta) + \phi((1+\bar{\theta})\theta - \bar{\theta}) - \bar{\theta}\gamma}. \end{aligned}$$

Note that $\phi(1) = 0$. Choosing γ so that $-(1+\bar{\theta})\phi(\theta) + \phi((1+\bar{\theta})\theta - \bar{\theta}) - \bar{\theta}\gamma = \gamma - r$, that is, $\gamma = \frac{-(1+\bar{\theta})\phi(\theta) + \phi((1+\bar{\theta})\theta - \bar{\theta}) + r}{1+\bar{\theta}}$, we have

$$P_\theta(S) \le 2e^{\frac{-(1+\bar{\theta})\phi(\theta) + \phi((1+\bar{\theta})\theta - \bar{\theta}) - \bar{\theta} r}{1+\bar{\theta}}}. \tag{A.13}$$

Combining (A.12) and (A.13), we obtain (A.9).

In the following, we restrict θ in $[\hat{\theta}(r), 1]$. Then we can choose $\bar{\theta}$ to be $\frac{\theta - \hat{\theta}(r)}{1-\theta} \ge 0$. Thus, using (4.2) that is, the relation $(\hat{\theta}(r) - 1)\frac{d\phi}{d\theta}(\hat{\theta}(r)) - \phi(\hat{\theta}(r)) = r$,

we have

$$\begin{aligned}
&\frac{-(1+\bar{\theta})\phi(\theta)+\phi((1+\bar{\theta})\theta-\bar{\theta})-\bar{\theta}r}{1+\bar{\theta}} \\
&\quad=-\phi(\theta)+\frac{\phi(\hat{\theta}(r))-\bar{\theta}r}{1+\bar{\theta}} \\
&\quad=-\phi(\theta)+\frac{(1-\theta)\phi(\hat{\theta}(r))-(\theta-\hat{\theta}(r))r}{1-\hat{\theta}(r)} \\
&\quad=-\phi(\theta)+\frac{(1-\theta)\phi(\hat{\theta}(r))-(\theta-\hat{\theta}(r))((\hat{\theta}(r)-1)\frac{d\phi}{d\theta}(\hat{\theta}(r))-\phi(\hat{\theta}(r)))}{1-\hat{\theta}(r)} \\
&\quad=-\phi(\theta)+\phi\big(\hat{\theta}(r)\big)+\big(\theta-\hat{\theta}(r)\big)\frac{d\phi}{d\theta}\big(\hat{\theta}(r)\big)=D(P_{\hat{\theta}(r)}\|P_\theta).
\end{aligned}$$

Hence, we obtain (A.10) and (A.11). □

APPENDIX B: PROOF THEOREM 7.7

First, we prepare the following lemma and corollary, which will be used later.

LEMMA B.1 (Cesáro summability). *Suppose that a sequence of matrices* $\{\beta_n\}_{n=1}^{\infty}$ *satisfies* $\beta_n \to \beta$. *Then we have* $\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\beta_k=\beta$.

COROLLARY B.2. *Suppose that* $\lim_{n\to\infty}\sum_{k=0}^{n-1}\alpha_k=\alpha$. *Then we have* $\lim_{n\to\infty}\sum_{k=0}^{n-1}\frac{n-k}{n}\alpha_k=\alpha$.

PROOF. Apply Lemma B.1 to the sequence $\beta_n=\sum_{k=0}^{n-1}\alpha_k$. □

Now we assume that X^{n+1} obeys the stationary Markov process generated by the transition matrix W_0, and denote the variance by V. As is shown in [24], Lemma 6.2, $g(X^{n+1}):=\sum_{i=1}^{n}g(X_{i+1},X_i)$ satisfies

$$\lim_{n\to\infty}\mathsf{V}\left[\frac{g^n(X^{n+1})}{\sqrt{n}}\right]=\frac{d^2\phi}{d\theta^2}(0).$$

Hence, it is enough for Lemma 7.7 to show that

$$\lim_{n\to\infty}\mathsf{V}\left[\frac{g^n(X^{n+1})}{\sqrt{n}}\right]=\mathsf{V}\big[g(X,X')\big]+2\vec{g}_*^{\,T}(Z-A)\vec{g}^*, \tag{B.1}$$

where $\vec{g}_*:=[\sum_x W(x|\bar{x})g(x,\bar{x})]_{\bar{x}}$, and $\vec{g}^*:=[\sum_{\bar{x}}W(x|\bar{x})\tilde{P}_0(\bar{x})g(x,\bar{x})]_x$.

From Proposition 7.6, $Z = \sum_{n=0}^{\infty}(W - A)^n$ exists. Thus, Corollary B.2 yields that $Z = \lim_{n\to\infty}\sum_{d=0}^{n-1}\frac{n-d}{n}(W - A)^d$, which implies

$$Z - I = \lim_{n\to\infty}\sum_{d=1}^{n-1}\frac{n-d}{n}(W - A)^d = \lim_{n\to\infty}\sum_{d=1}^{n-1}\frac{n-d}{n}(W^d - A), \tag{B.2}$$

where the last equality follows from the relation $(W - A)^n = W^n - A$ given in Proposition 7.6. By elementary calculation, we have

$$\begin{aligned}
&\frac{1}{n-1}\mathsf{V}\left[\sum_{k=2}^{n} g(X_k, X_{k-1})\right] \\
&\quad = \frac{1}{n-1}\mathsf{E}\left[\left(\sum_{k=2}^{n}\bigl(g(X_k, X_{k-1}) - \mathsf{E}[g(X, X')]\bigr)\right)\right. \\
&\quad\quad \left.\times\left(\sum_{\ell=2}^{n}\bigl(g(X_\ell, X_{\ell-1}) - \mathsf{E}[g(X, X')]\bigr)\right)\right] \qquad \text{(B.3)} \\
&\quad = \frac{1}{n-1}\sum_{k=2}^{n}\sum_{\ell=2}^{n}\bigl\{\mathsf{E}[g(X_k, X_{k-1})g(X_\ell, X_{\ell-1})] - \mathsf{E}[g(X, X')]^2\bigr\} \\
&\quad = \left\{\frac{1}{n-1}\sum_{k=2}^{n}\sum_{\ell=2}^{n}\mathsf{E}[g(X_k, X_{k-1})g(X_\ell, X_{\ell-1})]\right\} - (n-1)\mathsf{E}[g(X, X')]^2.
\end{aligned}$$

Since

$$\begin{aligned}
&\Pr\{X_k = x_k, X_{k-1} = x_{k-1}, X_\ell = x_\ell, X_{\ell-1} = x_{\ell-1}\} \\
&\quad = \begin{cases}
W(x_k|x_{k-1})W^{(k-1-\ell)}(x_{k-1}|x_\ell)W(x_\ell|x_{\ell-1})\tilde{P}(x_{\ell-1}), & \text{if } k > \ell+1, \\
W(x_k|x_{k-1})\delta_{x_{k-1}x_\ell}W(x_\ell|x_{\ell-1})\tilde{P}(x_{\ell-1}), & \text{if } k = \ell+1, \\
W(x_k|x_{k-1})\tilde{P}(x_{k-1})\delta_{x_k x_\ell}\delta_{x_{k-1}x_{\ell-1}}, & \text{if } k = \ell, \\
W(x_\ell|x_{\ell-1})\delta_{x_{\ell-1}x_k}W(x_k|x_{k-1})\tilde{P}(x_{k-1}), & \text{if } \ell = k+1, \\
W(x_\ell|x_{\ell-1})W^{(\ell-1-k)}(x_{\ell-1}|x_k)W(x_k|x_{k-1})\tilde{P}(x_{k-1}), & \text{if } \ell > k+1
\end{cases}
\end{aligned}$$

and $\vec{g}_*^T A\vec{g}^* = \mathsf{E}[g(X, X')]^2$, we have

$$\begin{aligned}
&\sum_{k=2}^{n}\sum_{\ell=2}^{n}\mathsf{E}[g(X_k, X_{k-1})g(X_\ell, X_{\ell-1})] \\
&\quad = (n-1)\mathsf{E}[g(X, X')^2] + 2(n-2)\vec{g}_*^T(I - A)\vec{g}^* + 2(n-2)\mathsf{E}[g(X, X')]^2 \\
&\quad\quad + 2\sum_{k>\ell-1}\vec{g}_*^T(W^{k+1-\ell} - A)\vec{g}^* + (n-2)(n-3)\mathsf{E}[g(X, X')]^2.
\end{aligned}$$

Thus, we can rewrite the first term of (B.3) as

$$\begin{aligned}
&\text{the right-hand side of (B.3)}:\\
&= \mathsf{V}\big[g(X,X')\big] + \frac{2}{n-1}\Big\{(n-2)\vec{g}_*^T(I-A)\vec{g}^* \\
&\quad + \sum_{k>\ell+1} \vec{g}_*^T(W^{(k-1-\ell)}-A)\vec{g}^*\Big\} \\
&= \mathsf{V}\big[g(X,X')\big] + \frac{2(n-2)}{n-1}\vec{g}_*^T(I-A)\vec{g}^* \\
&\quad + \frac{2(n-2)}{n-1}\sum_{d=1}^{n-3}\frac{n-2-d}{n-2}\vec{g}_*^T(W^d-A)\vec{g}^* \\
&\to \mathsf{V}\big[g(X,X')\big] + 2\vec{g}_*^T(I-A)\vec{g}^* + 2\vec{g}_*^T(Z-I)\vec{g}^* \\
&= \mathsf{V}\big[g(X,X')\big] + 2\vec{g}_*^T(Z-A)\vec{g}^*,
\end{aligned}$$

where we used (B.2) with replacing n by $n-3$, and took the limit $n \to \infty$. Combining with (B.3), we obtain (B.1).

APPENDIX C: A LEMMA FOR CONVEX FUNCTION

In this Appendix, we give a lemma for a convex function, which is employed in this paper.

LEMMA C.1. *When f is convex the function $x \mapsto \frac{f(x)}{x}$ with $x \in (0,\infty)$ has the minimum when $f'(x)x - f(x) = 0$. In particular, when $f(0) = 0$, the function $x \mapsto \frac{f(x)}{x}$ is monotone increasing for $x \geq 0$.*

PROOF. We have

$$\frac{d}{dx}\frac{f(x)}{x} = \frac{f'(x)x - f(x)}{x^2}.$$

Since

$$\frac{d}{dx}f'(x)x - f(x) = f''(x)x \geq 0,$$

we find that the minimum is realized when $f'(x)x - f(x)$. □

Acknowledgment. The authors are grateful to the referees for helpful comments and informing us about the helpful references [4, 7, 11, 14–16, 31–33, 36, 56].

REFERENCES

[1] ADAMCZAK, R. and BEDNORZ, W. (2012). Orlicz integrability of additive functionals of Harris ergodic Markov chains. Available at arXiv:1201.3567.

[2] ADAMCZAK, R. and BEDNORZ, W. (2015). Exponential concentration inequalities for additive functionals of Markov chains. *ESAIM Probab. Stat.* **19** 440–481. MR3423302

[3] AMARI, S. and NAGAOKA, H. (2000). *Methods of Information Geometry*. *Translations of Mathematical Monographs* **191**. Oxford Univ. Press, Oxford. MR1800071

[4] BEN-ARI, I. and NEUMANN, M. (2012). Probabilistic approach to Perron root, the group inverse, and applications. *Linear Multilinear Algebra* **60** 39–63. MR2869672

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

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

[7] BRADLEY, R. C. JR. (1983). Information regularity and the central limit question. *Rocky Mountain J. Math.* **13** 77–97. MR0692579

[8] CSISZÁR, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.* **2** 299–318. MR0219345

[9] DELYON, B., JUDITSKY, A. and LIPTSER, R. (2006). Moderate deviation principle for ergodic Markov chain. Lipschitz summands. In *From Stochastic Calculus to Mathematical Finance* 189–209. Springer, Berlin. MR2233540

[10] DEMBO, A. and ZEITOUNI, O. (1998). *Large Deviations Techniques and Applications*, 2nd ed. *Applications of Mathematics* (*New York*) **38**. Springer, New York. MR1619036

[11] DONSKER, M. D. and VARADHAN, S. R. S. (1975). Asymptotic evaluation of certain Markov process expectations for large time. I. II. *Comm. Pure Appl. Math.* **28** 1–47; ibid. 28 (1975), 279–301. MR0386024

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

[13] FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications*, 2nd ed. Wiley, New York, New York.

[14] HÄGGSTRÖM, O. (2005). On the central limit theorem for geometrically ergodic Markov chains. *Probab. Theory Related Fields* **132** 74–82. MR2136867

[15] HÄGGSTRÖM, O. (2006). Acknowledgement of priority concerning "On the central limit theorem for geometrically ergodic Markov chains." *Probab. Theory Related Fields* **135** 470. MR2240696

[16] HÄGGSTRÖM, O. and ROSENTHAL, J. S. (2007). On variance conditions for Markov chain CLTs. *Electron. Commun. Probab.* **12** 454–464 (electronic). MR2365647

[17] HAYASHI, M. (2016). Finite-block-length analysis in classical and quantum information theory. Available at arXiv:1605.02821.

[18] HAYASHI, M. (2009). Information spectrum approach to second-order coding rate in channel coding. *IEEE Trans. Inform. Theory* **55** 4947–4966. MR2596952

[19] HAYASHI, M. and NAGAOKA, H. (2003). General formulas for capacity of classical-quantum channels. *IEEE Trans. Inform. Theory* **49** 1753–1768. MR1985576

[20] HAYASHI, M. and WATANABE, S. (2-4, October, (2013)). Non-asymptotic bounds on fixed length source coding for Markov chains. In *Proceedings of* 51*st Annual Allerton Conference on Communication*, *Control*, *and Computing* 875–882. Allerton House, Monticello, IL.

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

[22] HAYASHI, M. and WATANABE, S. (2014). Non-asymptotic and asymptotic analyses on Markov chains in several problems. In *Proceedings of* 2014 *Information Theory and Applications Workshop*, *Catamaran Resort* 1–10. San Diego, CA.

[23] HAYASHI, M. and WATANABE, S. (2016). Uniform random number generation from Markov chains: Non-asymptotic and asymptotic analyses. *IEEE Trans. Inform. Theory* **62** 1795–1822. MR3480083

[24] HAYASHI, M. and WATANABE, S. (2016). Information geometry approach to parameter estimation in Markov chains. *Ann. Statist.* **44** 1495–1535. MR3519931

[25] HERVÉ, L., LEDOUX, J. and PATILEA, V. (2012). A uniform Berry–Esseen theorem on M-estimators for geometrically ergodic Markov chains. *Bernoulli* **18** 703–734. MR2922467

[26] HOEFFDING, W. (1965). Asymptotically optimal tests for multinomial distributions. *Ann. Math. Statist.* **36** 369–408. MR0173322

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

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

[29] KATO, T. (1980). *Perturbation Theory for Linear Operators*. Springer, New York.

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

[31] KIPNIS, C. and VARADHAN, S. R. S. (1986). Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Comm. Math. Phys.* **104** 1–19. MR0834478

[32] KOMOROWSKI, T., LANDIM, C. and OLLA, S. (2012). *Fluctuations in Markov Processes*: *Time Symmetry and Martingale Approximation*. *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **345**. Springer, Heidelberg. MR2952852

[33] KONTOYIANNIS, I., LASTRAS-MONTAÑO, L. A. and MEYN, S. P. (2006). Exponential bounds and stopping rules for MCMC and general Markov chains. In *First International Conference on Performance Evaluation Methodologies and Tools*. ACM, New York.

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

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

[36] LALLEY, S. P. (1986). Ruelle's Perron–Frobenius theorem and the central limit theorem for additive functionals of one-dimensional Gibbs states. In *Adaptive Statistical Procedures and Related Topics* (*Upton, N.Y.*, 1985). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **8** 428–446. IMS, Hayward, CA. MR0898264

[37] ŁATUSZYŃSKI, K., MIASOJEDOW, B. and NIEMIRO, W. (2012). Nonasymptotic bounds on the mean square error for MCMC estimates via renewal techniques. In *Monte Carlo and Quasi-Monte Carlo Methods* 2010. *Springer Proc. Math. Stat.* **23** 539–555. Springer, Heidelberg. MR3173857

[38] LI, K. (2014). Second-order asymptotics for quantum hypothesis testing. *Ann. Statist.* **42** 171–189. MR3178460

[39] MASSART, P. (2007). *Concentration Inequalities and Model Selection*. *Lecture Notes in Math.* **1896**. Springer, Berlin. MR2319879

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

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

[42] MOSONYI, M. and OGAWA, T. (2015). Two approaches to obtain the strong converse exponent of quantum hypothesis testing for general sequences of quantum states. *IEEE Trans. Inform. Theory* **61** 6975–6994. MR3430733

[43] NAGAOKA, H. (2001). Strong converse theorems in quantum information theory. In *Proc. ERATO Conference on Quantum Information Science* (*EQIS*) 2001, 33 (2001) (M. Hayashi, ed.). IEEE, New York. [Also appeared as Chapter 3 of *Asymptotic Theory of Quantum Statistical Inference*, World Scientific, Singapore.]

[44] 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.

[45] 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

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

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

[48] POLYANSKIY, Y., POOR, H. V. and VERDÚ, S. (2010). Channel coding rate in the finite blocklength regime. *IEEE Trans. Inform. Theory* **56** 2307–2359. MR2729787

[49] RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, New York. MR0346957

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

[51] 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

[52] STRASSEN, V. (1964). Asymptotische Abschätzungen in Shannons Informationstheorie. In *Trans. Third Prague Conf. Information Theory, Statist. Decision Functions, Random Processes* (*Liblice*, 1962) 689–723. Publ. House Czech. Acad. Sci., Prague. MR0165997

[53] TIHOMIROV, A. N. (1980). Convergence rate in the central limit theorem for weakly dependent random variables. *Teor. Veroyatn. Primen.* **25** 800–818. MR0595140

[54] TOMAMICHEL, M. and HAYASHI, M. (2013). A hierarchy of information quantities for finite block length analysis of quantum tasks. *IEEE Trans. Inform. Theory* **59** 7693–7710. MR3124668

[55] TREVEZAS, S. and LIMNIOS, N. (2009). Variance estimation in the central limit theorem for Markov chains. *J. Statist. Plann. Inference* **139** 2242–2253. MR2507986

[56] VARADHAN, S. R. S. (2001). *Probability Theory. Courant Lecture Notes in Mathematics* **7**. New York Univ., Courant Institute of Mathematical Sciences, New York. MR1852999

[57] WATANABE, S. and HAYASHI, M. (2014). Finite-length analysis on tail probability and simple hypothesis testing for Markov chain. In *Proceeding of* 2014 *International Symposium on Information Theory and Its Applications* **26–29** 196–200. Melbourne, Australia.

DEPARTMENT OF COMPUTER
AND INFORMATION SCIENCES
TOKYO UNIVERSITY OF AGRICULTURE
AND TECHNOLOGY
KOGANEI TOKYO 184-8588
JAPAN

GRADUATE SCHOOL OF MATHEMATICS
NAGOYA UNIVERSITY
NAGOYA 464-8601
JAPAN
AND
CENTRE FOR QUANTUM TECHNOLOGIES
NATIONAL UNIVERSITY OF SINGAPORE
SINGAPORE 119077
SINGAPORE
E-MAIL: masahito@math.nagoya-u.ac.jp