On the Posterior Distribution of a Random Process Conditioned on Empirical Frequencies of a Finite Path: the i.i.d and finite Markov chain case

Wenqing Hu* and Hong Qian†

Abstract

We obtain the posterior distribution of a random process conditioned on observing the empirical frequencies of a finite sample path. We find under a rather broad assumption on the “dependence structure” of the process, c.f. independence or Markovian, the posterior marginal distribution of the process at a given time index can be identified as certain empirical distribution computed from the observed empirical frequencies of the sample path. We show that in both cases of discrete-valued i.i.d. sequence and finite Markov chain, a certain “conditional symmetry” given by the observation of the empirical frequencies leads to the desired result on the posterior distribution. Results for both finite-time observations and its asymptotic infinite-time limit are connected via the idea of Gibbs conditioning. Finally, since our results demonstrate a central role of the empirical frequency in understanding the information content of data, we use the Large Deviations Principle (LDP) to construct a general notion of “data-driven entropy”, from which one can apply a formalism from the recent study of statistical thermodynamics to data.

Keywords: Posterior Distribution, Empirical Frequency, Gibbs Conditioning, Large Deviations Principle, Statistical Thermodynamics.

2020 Mathematics Subject Classification Numbers: 60J10, 60F10, 62C10, 62C12.

1 Introduction

Ever since the axiomatic construction of A. N. Kolmogorov [15], Probability Theory always starts with a probability space \((\Omega, \mathcal{F}, P)\), where \(\Omega\) is a sample space of elementary events, \(\mathcal{F}\) is a \(\sigma\)-algebra on the sample space, and \(P\) is a probability measure given a priori. A major concern when applying this paradigm to real-world problems, identified

*Department of Mathematics and Statistics, Missouri University of Science and Technology (formerly University of Missouri, Rolla). Email: huwen@mst.edu
†Department of Applied Mathematics, University of Washington. Email: hqian@uw.edu
as the “second-half of probability theory” by E. T. Jaynes \cite{Jaynes}, the pioneer of Maximum Entropy Principle, lies in that one usually does not know the probability measure in any realistic way. As a matter of fact the $P$ is not a scientific observable which should be a function on $\Omega$. This has not prevented some researchers relying on the “frequentist” point of view and use observed empirical distribution as a substitution for the real probability. However, when one “fits” data to a statistical model, the form of the unknown probability distribution necessarily enters as assumptions, hidden or explicitly. This is most succinctly pointed out by P. J. Huber in \cite{Huber}:

\begin{quote}
Statistical inferences are based only in part upon the observations. An equally important base is formed by prior assumptions about the underlying situation. Even in the simplest cases, there are explicit or implicit assumptions about randomness and independence, about distributional models, perhaps prior distributions for some unknown parameters, and so on.
\end{quote}

The route of building statistical models from probability, as an integral part of data science, thus, should always start with the “basic assumption of a probability space including a prior probability measure”. This realization led to the rise of Bayesian logic in which posterior conditional probability becomes a central object, while in many science and engineering, empirical frequency is still a legitimate surrogate for the probability. Inspired by the upcoming “big data” in the near future, the main concern of the present work is to secure a probability measure from empirical frequency data. We are particularly motivated by the following question:

\begin{quote}
Given an observation of the empirical frequencies of a random process, to what extent can we recover the probability structure of the original random process via conditioning?
\end{quote}

Our main rationale for addressing the question lies as follows: Under a rather broad assumption of the dependence structure on the process, such as “independence” or “Markovian”, we can identify the posterior marginal distribution of the process at a given time index as certain empirical distribution that is computable from the empirical frequencies of observed outcomes of the process, e.g. a finite sample path. In the simplest case for an i.i.d sequence, the distribution from the computation is the empirical frequencies of the outcomes. In general one considers “sample path frequencies” admitted by the observed empirical frequencies of the outcomes, through an analysis of sample path trajectories. A summarizing conclusion in short: “posterior probability distribution is given by some kind of empirical frequency based on observations”. There is a quite related idea under the notion of de Finetti’s exchangeability \cite{Diaconis}. In particular Diaconis and Freedman \cite{DiaconisFreedman} have constructed a symmetry based definition for the very notion of Markov dependence in a sequence. Our work, however, studies the posterior
conditional probability within the framework of a Markov process by exploring such a symmetry. Indeed, the results we demonstrate below can be easily extended to the exchangeable case since all we need is the symmetry provided by exchangeability.

In this work, we carry out the analysis for the two cases: (1) a sequence of i.i.d random variables taking discrete values, and (2) a sequence of random variables that follow a finite-state Markov chain. The former is included since some of the subtleties for the latter are more clearly exhibited. In both cases, the fact that the posterior marginal distribution is given by the empirical distribution can be roughly understood as a result of “conditional symmetry” for the joint distribution of the process under interest. To illustrate this simply: Suppose we have observed the empirical frequencies of a sequence of i.i.d. random variables $X_1, ..., X_n$ taking positive integer values. This implies that the number of $X$’s taking value $i \in \mathbb{N}$ is a known number of count $\nu_i \in \mathbb{Z}_+$, such that $\sum_{i \in \mathbb{N}} \nu_i = n$. When conditioned on the observation of only the quantities $\nu_i, i \in \mathbb{N}$, the random variable $X_\ell$ at a fixed index $1 \leq \ell \leq n$ may take the value of any number in the unordered list (it may contain repeated elements) $[i_1, ..., i_n]$ such that $\# \{\ell : i_\ell = i\} = \nu_i$ for any $i \in \mathbb{N}$. However, using the i.i.d. (exchangeable) property, it is intuitively true that $X_\ell$ may take each of the possible $i_\ell$’s with the same conditioned probability. This is a result of the fact that any sample trajectory $X_1, ..., X_n$ of an i.i.d. sequence conditioned on the observed frequencies has the same joint distribution. Such a conditional symmetry leads to the fact that the observed empirical frequencies give the posterior marginal of $X_\ell$.

While simple to understand, the situation is more involved when the i.i.d. sequence is replaced by a finite-state Markov chain. In this case, upon the observation of $n+1$ steps $Y_1, Y_2, ..., Y_n, Y_{n+1}$ in the chain, we count the empirical frequencies of the occurrences of one-step transitions $X_1 = (Y_1, Y_2), ..., X_n = (Y_n, Y_{n+1})$. Given the frequencies of these transitions and fix the starting point of the Markov chain $Y$, we observe that the joint probability distribution of $X_1, ..., X_n$ remains the same regardless of how we do admissible permutations of the occurrences of the one-step transitions. Here an admissible permutation means that the resulting sequence of the $X$’s will still form a trajectory of a Markov chain. Thus we obtained “conditional symmetry” at the level of sample path trajectories. As a result, the posterior marginal distribution of $X_1$ upon observing the empirical frequencies of $X_1, ..., X_n$, is given by the empirical distribution of admissible trajectories that match the observed empirical frequencies. This means that the posterior marginal distribution of $X_1$ is proportional to the number of all admissible trajectories of $X_1, ..., X_n$ that match the observed empirical frequencies. In this case, we can still understand the posterior distribution as some kind of “empirical frequency”, but it is at the level of sample path trajectories rather than the above direct observation of the $X$’s.

Problems of similar nature, namely the calculation of posterior distributions of
random process conditioned on given observations, have been considered in previous works, see [3] and the references cited within. The key difference is that we consider finite-time sample path while [3] considered conditioning under large deviation principle (LDP), e.g. with infinite-time limit. To connect these two types of results, we investigate the infinite-time limit of our finite-time results. By using combinatorial enumeration results known as Whittle’s formula [24,25], we are able to show that for the finite Markov chain under additional ergodicity condition, in the asymptotic limit as the number of observations tends to infinity, the numbers of admissible trajectories starting from different one-step transitions are evenly distributed. This yields the fact that in the infinite-time limit, the “conditional symmetry” at the level of sample paths for ergodic finite Markov chains is reduced to the “conditional symmetry” at the level of the observations of one-step transitions. Therefore for an ergodic finite Markov chain, as the number of observations tends to infinity, the posterior distribution of the one-step transition tends to be given asymptotically by the empirical frequency of the observed one-step transitions, a result that aligns nicely with the i.i.d. case. Notice that under the ergodicity assumption, the latter empirical frequency also converges to the actual transition probabilities multiplied by the stationary measure. In fact, such an asymptotic result can also be obtained from the general principle of Gibbs conditioning (see Sec. 3.3 of [6] and Sec. 7.3 of [5]). However, our approach is more systematic and reveals more clearly how different levels of symmetry at play in obtaining the limit of the posterior distribution, and it is expected to be generalizable to broader classes of random processes.

Our result can be interpreted more conceptually from a measure-theoretic point of view since the conditional symmetry that we revealed here is simply a result of the product structure of the underlying probability measure defining the process. We expect that a general principle should hold beyond the cases we can rigorously demonstrate here, that a “conditional symmetry” leads to the procedure of using empirical distributions calculated from observed frequencies to stand for posterior marginal distributions. This principle is expected to be also applicable in understanding the posterior marginal distribution of continuous-time Markov processes, for which we leave the exact formulation of the result and its proof to a future time.

As was pointed out in [4], there have been two different paradigms for doing scientific research: the Keplerian paradigm, which is the data-driven approach, and the Newtonian paradigm, which is considered the first-principle-based approach. The natural question that arises with the rapid advancement of current data science is to understand how much information one can extract from large collection of data. Essentially, the largest amount of information we can extract from data is the maximum capability of doing science under the Keplerian paradigm. Our work thus can be regarded as providing an understanding to this question at least from one perspective. The underlying fact
is: Given only the frequency observations of a random process, a conditional symmetry occurs to the posterior distribution, and no more additional information can be extracted from frequency observations. This illustrates the importance of empirical frequency observations when working under the Keplerian paradigm. Based on this, one can further apply the Large Deviations Principle \[6,10\] to the empirical frequencies either at the level of random variable or at the level of sample paths. Following this approach, as we shall illustrate, one can achieve a natural integration of the frequentist school and the Bayesian school, with the concept of entropy emerging at the center stage. The entropy concept here, however, is much more general and broader than that of Gibbs, which appears in statistical thermodynamics, and Shannon’s which defines current information theory. It can be regarded as an abstract “data-driven” entropy from which we can apply the formalism of thermodynamics to data sciences. For asymptotic infinitely large samples and in the context of LDP, this is the essential idea of Gibbs conditioning.

The paper is organized as follows: In Section 2 we derive the results when the process \(X_1, \ldots, X_n\) consists of an i.i.d. sequence; in Section 3 we consider the finite Markov chain case; in Section 4 we study the infinite-time limit of the results in Sec. 3; in Section 5 we discuss a novel understanding of the mathematical content of statistical thermodynamics, a branch of theoretical physics, in terms of the large deviations theory and Koopman-Damois exponential family of models. This gives the paradigm of Maximum Entropy Principle and a deep connection to an overlooked theory developed by L. Szilard [23] and B. B. Mandelbrot [19]. We further propose some conjectures and provide an outlook for future research.

2 Conditioned on sample frequency: The i.i.d case

Let \(X_1, ..., X_n, \ldots\) be an i.i.d sequence defined on the probability space \((\Omega, \mathcal{F}, \mathbf{P})\) with common distribution as a random variable \(X\) taking values in \(\mathbb{N}\). Given a sequence of sample frequencies \(\nu_k \in \mathbb{N}_+\) satisfying

\[
\sum_{k \in \mathbb{N}} \nu_k = n,
\]

we consider the event

\[
\mathcal{E}_{\{\nu_k\}} = \left\{ \sum_{\ell=1}^n \mathbb{1}_k(X_\ell) = \nu_k, k \in \mathbb{N} \right\},
\]

where \(\mathbb{1}_k(X_\ell) = \begin{cases} 1, & \text{if } X_\ell = k, \\ 0, & \text{otherwise} \end{cases}\). Thus \(\mathcal{E}_{\{\nu_k\}}\) stands for the event that the trajectory \(X_\ell, \ell = 1, 2, ..., n\) takes on value \(k\) with frequency \(\nu_k, k \in \mathbb{N}\), respectively. The event \(\mathcal{E}_{\{\nu_k\}}\) can be viewed as the “observation” of the trajectory up to time \(\ell = n\). Conditioned on this event, we claim the following
Theorem 1 (posterior distribution for the i.i.d. case). Given \( m \in \mathbb{N} \) and any \( 1 \leq \ell \leq n \), we have
\[
P(X_{\ell} = m \mid \mathcal{E}_{\{\nu_k\}}) = \frac{\nu_m}{n}.
\] (2)

Proof. Given \( \mathcal{E}_{\{\nu_k\}} \), for any \( k \in \mathbb{N} \), we know that among each \( X_{\ell} \) in the sequence \( X_1, \ldots, X_n \), \( \ell = 1, \ldots, n \), the number \( k \) occurs at a multiple of \( \nu_k \) times. Since each \( \nu_k \geq 0 \) is a non-negative integer, and \( \sum_{k \in \mathbb{N}} \nu_k = n \), we see that there are only finitely many \( k \)'s such that \( \nu_k \geq 1 \), and for all the rest of \( k \in \mathbb{N} \) we have \( \nu_k = 0 \). We order the \( k \)'s such that \( \nu_k \geq 1 \) in an increasing sequence as \( 1 \leq k_1 < k_2 < \ldots < k_I < \infty \) for some finite \( I \in \mathbb{N} \). Consider the set of \( n \) distinguished elements
\[
\mathcal{X}_{\text{distinguished}} = \bigcup_{i=1}^{I} \{ (k_i, 1), \ldots, (k_i, \nu_{k_i}) \}.
\] (3)

Let us randomly pick each element in \( \mathcal{X}_{\text{distinguished}} \) once and without replacement, so that we can establish \( n! \) sequences of length \( n \) with each sequence \( S \) consists of ordered elements of the form \( (k, a) \) for some \( k \in \{k_1, \ldots, k_I\} \) and some \( 1 \leq a \leq \nu_k \). An element-wise projection \( \mathcal{Y} \) with \( (k, a) \xrightarrow{\mathcal{Y}} k \) applied to each of the above sequence results in an outcome of the sequence \( X_1, \ldots, X_n \) conditioned on \( \mathcal{E}_{\{\nu_k\}} \). We construct a probability measure \( \mathcal{P} \) on the space of all such sequences \( \mathcal{S} \), such that for each sequence we have
\[
\mathcal{P}(S) = C \prod_{i=1}^{I} (\mathcal{P}(X = k_i))^{\nu_{k_i}}
\] (4)
for an undetermined normalizing constant \( C > 0 \). For a given outcome \( (i_1, \ldots, i_n) \) of the sequence \( X_1, \ldots, X_n \), such that when counting frequencies, the sequence \( (i_1, \ldots, i_n) \) satisfies \( \mathcal{E}_{\{\nu_k\}} \), we must have \( \{X_1 = i_1, \ldots, X_n = i_n\} \subset \mathcal{E}_{\{\nu_k\}} \). Therefore for such sequences we have
\[
\mathcal{P}(X_1 = i_1, \ldots, X_n = i_n, \mathcal{E}_{\{\nu_k\}}) = \mathcal{P}(X_1 = i_1, \ldots, X_n = i_n) = \prod_{i=1}^{I} (\mathcal{P}(X = k_i))^{\nu_{k_i}}.
\] Thus we have
\[
\mathcal{P}(X_1 = i_1, \ldots, X_n = i_n \mid \mathcal{E}_{\{\nu_k\}}) = D \prod_{i=1}^{I} (\mathcal{P}(X = k_i))^{\nu_{k_i}},
\] (5)
where \( D > 0 \) is an undetermined normalizing constant for the conditional probability measure \( \mathcal{P}(\cdot \mid \mathcal{E}_{\{\nu_k\}}) \). Comparing (4) and (5), we see that for any given outcome \( (i_1, \ldots, i_n) \) of the sequence \( X_1, \ldots, X_n \) satisfying \( \mathcal{E}_{\{\nu_k\}} \) we must have
\[
\mathcal{P}(X_1 = i_1, \ldots, X_n = i_n \mid \mathcal{E}_{\{\nu_k\}}) = K \mathcal{P}(\mathcal{S}),
\] (6)
for some constant \( K > 0 \).

Combining (4), (5) and (6), we see the following two facts
Fact 1: For each of the \( n! \) different sequences \( \mathcal{S} \), \( \mathcal{P}(\mathcal{S}) \) has the same value.

Fact 2: The sequence \( \mathcal{S} \) in the RHS of (6) can be an arbitrary length-\( n \) sequence picked from the \( n! \) possible choices.

For each sequence \((i_1, \ldots, i_n)\), we collect all possible sequences \( \mathcal{S} \) such that \( \mathcal{Y}(\mathcal{S}) = (i_1, \ldots, i_n) \). We claim that we have

\[
\mathcal{P}(X_1 = i_1, \ldots, X_n = i_n | \mathcal{E}_{(\nu_k)}) = \mathcal{P} \text{ (all possible } \mathcal{S} \text{ such that } \mathcal{Y}(\mathcal{S}) = (i_1, \ldots, i_n)) \ . \quad (7)
\]

This is because due to the above Fact 1, we have

\[
\mathcal{P} \text{ (all possible } \mathcal{S} \text{ such that } \mathcal{Y}(\mathcal{S}) = (i_1, \ldots, i_n)) = \text{ (Number of all possible } \mathcal{S} \text{ such that } \mathcal{Y}(\mathcal{S}) = (i_1, \ldots, i_n)) \cdot \mathcal{P}(\mathcal{S}) .
\]

Now we note that by simple combinatorics we have

Fact 3: For each realization \((i_1, \ldots, i_n)\) of \( X_1, \ldots, X_n \),

\[
\text{ (Number of all possible } \mathcal{S} \text{ such that } \mathcal{Y}(\mathcal{S}) = (i_1, \ldots, i_n) \text{) is independent of the choice of } (i_1, \ldots, i_n).
\]

Therefore by Fact 3 we see that (7) is equivalent to

\[
\mathcal{P}(X_1 = i_1, \ldots, X_n = i_n | \mathcal{E}_{(\nu_k)}) = K_1 \mathcal{P}(\mathcal{S}) .
\]

It is then easy to see that \( K_1 = K \) just by normalization of the probability measure \( \mathcal{P}(\bullet) \) and the conditional probability measure \( \mathcal{P}(\bullet | \mathcal{E}_{(\nu_k)}) \), as well as the above Fact 2.

So we proved that (7) is valid.

From (7) we know that for an element \( m \in \{k_1, \ldots, k_I\} \) we have

\[
\mathcal{P}(X_\ell = m | \mathcal{E}_{(\nu_k)}) = \sum_{a=1}^{\nu_m} \mathcal{P} \text{ (all possible } \mathcal{S} \text{ whose } \ell\text{-th element is } (m, a)) . \quad (8)
\]

We claim that for each element \((m, a) \in \mathcal{X}_{\text{distinguished}}\) we also have

Fact 4: For each \((m, a) \in \mathcal{X}_{\text{distinguished}}\),

\[
\text{ (Number of all possible } \mathcal{S} \text{ whose } \ell\text{-th element is } (m, a) \text{) is independent of } (m, a) .
\]

1. This number is actually \( \nu_{k_1} \cdot \nu_{k_2} \cdot \ldots \cdot \nu_{k_I} \).
2. This number is actually \( (n-1)! \).
Fact 5: $\mathcal{P}$ (all possible $\mathcal{S}$ whose $\ell$-th element is $(m, a)$) def $= p$ is independent of $(m, a)$.

The above Fact 4 is a simple combinatorial observation, and Fact 5 is a consequence of Facts 1 and 4. Since $\mathcal{X}^{\text{distinguished}}$ contains $n$ elements, by Fact 5 we know that $np = 1$, i.e., $p = \frac{1}{n}$. This fact combined with [3] give us [2]. Note that when $m \not\in \{k_1, ..., k_I\}$ we have $\nu_m = 0$ and [2] is trivial. So we have proved the whole statement.

Remark 1 (Alternate proof of Theorem 1). One can establish a less intuitive but more direct combinatorial proof of Theorem 1 as follows.

Alternate Proof of Theorem 1. Given $E\{\nu_k\}$, the occurrences of $X_\ell = m$ happen on an arbitrary length-$\nu_m$ sub-index sequence $1 \leq \ell_1 < \ldots < \ell_{\nu_m} \leq n$ such that $\ell = \ell_1, ..., \ell_{\nu_m}$.

Consider the event $E^m_{(\ell_1, ..., \ell_{\nu_m})} \equiv \{X_{\ell_1} = \ldots = X_{\ell_{\nu_m}} = m \text{ for } 1 \leq \ell_1 < \ldots < \ell_{\nu_m} \leq n \text{ and } X_\ell \neq m \text{ for } \ell \neq \ell_1, ..., \ell_{\nu_m}\}$.

Then for two distinct sequences $(\ell_1, ..., \ell_{\nu_m}) \neq (\tilde{\ell}_1, ..., \tilde{\ell}_{\nu_m})$ we must have $E^m_{(\ell_1, ..., \ell_{\nu_m})} \cap E^m_{(\tilde{\ell}_1, ..., \tilde{\ell}_{\nu_m})} = \emptyset$.

Let us also define the event $E^m_{\nu_m, \nu_m} = \left\{ \sum_{\ell=1}^{n-\nu_m} 1_k(X_\ell) = \nu_k, k \in \mathbb{N} \backslash \{m\} \right\}$.

Define $P(E^m_{\nu_m}) = p$, then it is easy from the above that

$$P(E_{\nu_k}) = \binom{n}{\nu_m} p \cdot P(E_{\nu_k}^{m, \nu_m}),$$

$$P(X_\ell = m, E_{\nu_k}) = \binom{n-1}{\nu_m-1} p \cdot P(E_{\nu_k}^{m, \nu_m}),$$

Thus this implies that

$$P(X_\ell = m | E_{\nu_k}) = \frac{(\nu_m-1)}{n} = \frac{\nu_m}{n},$$

which is [2].

Remark 2 (Conditional Symmetry). In the above proof, we have extended the conditional probability $P(\bullet | E_{\nu_k})$ on the space of all outcome sequences of $X_1, ..., X_n$ that match the frequency outcome $E_{\nu_k}$, to the probability $\mathcal{P}(\bullet)$ on an “lifted” probability space of sequences $\mathcal{S}$. In this correspondence, the Facts 1,2,3 lead to [3], which states that the conditional probability under our interest is equal to an absolute probability given by $\mathcal{P}$. This further helps us to understand that the “conditional symmetry” is represented by an “absolute symmetry”, exactly stated as the Fact 1 in the above proof.
Actually, even without lifting the conditional probability measure $P(\bullet|E_{\nu_k})$ to the absolute probability measure $\mathcal{P}(\bullet)$, the conditional symmetry can still be easily seen from (5), which is a result of i.i.d property. By (5), we know that the joint probability of $X_1, ..., X_n$ conditioned on $E_{\nu_k}$ will remain the same regardless of how we place the outcomes $i_1, ..., i_n$.

The reason why we would like to lift the original conditional probability to the absolute probability $\mathcal{P}$ lies behind our stated Facts 4 and 5 in the above proof. Here we would like to explore another type of symmetry: the “combinatorial symmetry” that comes from direct counting. By Fact 4, such kind of symmetry asserts that the number of certain combinatorial configurations are the same regardless of specific constraints. This enables us to directly obtain the fact that the empirical frequency of observed outcomes is the posterior marginal distribution. We will see that such “combinatorial symmetry” may be broken in the finite Markov chain case.

Our strategy of proof, although a bit more complicated than a direct combinatorial proof (see Remark 3), reveals the more general symmetric structure of the problem.

**Remark 3.** It can be easily seen using the same argument as in the proof of Theorem 1 that the joint conditional distribution

$$P(X_{\ell_1} = m_1, X_{\ell_2} = m_2|E_{\nu_k}) = \frac{\nu_{m_1}\nu_{m_2}}{n(n-1)}.$$  

This together with Theorem 1 imply that $X_{\ell_1}$ and $X_{\ell_2}$ are not conditionally independent. In a heuristic explanation, the conditioning puts a “nonlinear constraint” on the joint distributions which makes them dependent.

### 3 Conditioned on sample frequency: The finite Markov chain case

Let $Y_1, ..., Y_n, ...$ be a time-homogeneous Markov chain with finite state space $\Sigma = \{1, ..., N\}$, $|\Sigma| = N$. Let the transition probability matrix of the process $\{Y_\ell\}_{\ell \geq 1}$ be given by $P = (p_{ij})_{1 \leq i,j \leq N}$. Assume the process starts from an initial probability distribution $\pi^0 = (\pi_1^0, ..., \pi_N^0)$, $0 \leq \pi_i^0 \leq 1$, $\sum_{i=1}^N \pi_i^0 = 1$, such that $P(Y_1 = i) = \pi_i^0$.

Define the “consecutive pair” process $X_\ell = (Y_\ell, Y_{\ell+1})$, $\ell \geq 1$. It is easy to see that $\{X_\ell\}_{\ell \geq 1}$ is also a Markov process with transition probability matrix

$$P^{(2)} = (p_{(i,j),(k,l)}^{(2)})_{1 \leq i,j,k,l \leq N},$$

and the matrix elements are computed from $p_{ij}$ via the formula (see [6, Section 3.1.13])

$$p_{(i,j),(k,l)}^{(2)} = 1_{j=k} \cdot p_{kl}.$$
Let us first look more carefully at the procedure by which we transfer a trajectory of \( \{Y_t\}_{t \geq 1} \) to \( \{X_t\}_{t \geq 1} \). The process \( \{Y_t\}_{t \geq 1} \) has induced a measure \( \mu \) on \( \Sigma^{n+1} \), where each length-\( (n+1) \) string \( i_1, \ldots, i_n, i_{n+1} \) is measured as

\[
\mu(i_1, \ldots, i_n, i_{n+1}) = \pi_i^0 \prod_{\ell=1}^{n} p_{i_{\ell} i_{\ell+1}} .
\]

The above measure \( \mu \) induces a new measure \( \mu^{(2)} \) on \( (\Sigma \times \Sigma)^n \) such that for each length-\( n \) string of the pairs \( (i_1, i_2), (i_2, i_3), \ldots, (i_n, i_{n+1}) \) we have

\[
\mu^{(2)}((i_1, i_2), (i_2, i_3), \ldots, (i_n, i_{n+1})) = \mu(i_1, \ldots, i_n, i_{n+1}) = \pi_i^0 \prod_{\ell=1}^{n} p_{i_{\ell} i_{\ell+1}} .
\]

Moreover, for any length-\( n \) string of the pairs \( (i_{11}, i_{12}), (i_{21}, i_{22}), \ldots, (i_{n1}, i_{n2}) \) such that there exist some \( i_{\ell2} \neq i_{(\ell+1)1} \) for some \( 1 \leq \ell \leq n-1 \), we have

\[
\mu^{(2)}((i_{11}, i_{12}), (i_{21}, i_{22}), \ldots, (i_{n1}, i_{n2})) = 0 .
\]

For future presentation, we introduce the following definition.

**Definition 1** (string of chain type). We denote a length-\( n \) string of the pairs

\[
((i_{11}, i_{12}), (i_{21}, i_{22}), \ldots, (i_{n1}, i_{n2}))
\]

such that \( i_{12} = i_{21}, \ldots, i_{(n-1)2} = i_{n1} \) as a “string of chain type”. For such a string of chain type, we denote by \( i_{11} \) its “head”.

Thus (10) and (11) are saying that \( \mu^{(2)} \) only charges on strings of chain type. Let the sample space for the trajectory \( X_1, \ldots, X_n \) be given by \( (\Omega = (\Sigma \times \Sigma)^n, \mathcal{F}, \mathbb{P}) \), then it is easy to see that \( \mu^{(2)} \) is the probability measure for the trajectory \( X_1, \ldots, X_n \), i.e., we have

\[
\mathbb{P}(X_1 = (i_{11}, i_{12}), X_2 = (i_{21}, i_{22}), \ldots, X_n = (i_{n1}, i_{n2})) = \mu^{(2)}((i_{11}, i_{12}), (i_{21}, i_{22}), \ldots, (i_{n1}, i_{n2})) .
\]

Given a sequence of sample frequencies \( \nu_{(i,j)} \in \mathbb{N}_+ \) satisfying

\[
\sum_{i=1}^{N} \sum_{j=1}^{N} \nu_{(i,j)} = n ,
\]

we consider the event

\[
\mathcal{E}_{\nu_{(i,j)}} = \left\{ \sum_{\ell=1}^{n} \mathbbm{1}_{(i,j)}(X_\ell) = \nu_{(i,j)}, 1 \leq i, j \leq N \right\} ,
\]

where \( \mathbbm{1}_{(i,j)}(X_\ell) = \begin{cases} 1, & \text{if } X_\ell = (i, j) , \\ 0, & \text{otherwise} . \end{cases} \) Thus \( \mathcal{E}_{\nu_{(i,j)}} \) stands for the event that the trajectory \( X_\ell (\ell = 1, \ldots, n) \) takes on value \( (i, j) \) with frequency \( \nu_{(i,j)}, 1 \leq i, j \leq N, \)
respectively. The event $E_{\nu(i,j)}$ can be viewed as the “observation” of the trajectory of $X_\ell$ up to time $\ell = n$.

Let us suppose that we have observed an outcome of the event $E_{\nu(i,j)}$. Then we claim that the Markov chain $Y_\ell$ for a fixed index $1 \leq \ell \leq n$ cannot be an arbitrary element chosen from the state space $\Sigma = \{1, \ldots, N\}$. This can be seen from the following example.

**Example 1.** Suppose $\{Y_\ell\}_{\ell \geq 1}$ is a stationary Markov chain with a 3-element state space $\{1, 2, 3\}$ and stationary measure $\pi = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Set $n = 2$ and suppose we have observed $E_{\nu(i,j)} = \{\nu(1,2) = \nu(2,3) = 1, \nu(i,j) = 0 \text{ for all other pairs of } (i, j)\}$.

Then it is easy to see that $P(Y_1 = 1|E_{\nu(i,j)}) = 1$ while $P(Y_1 = 2|E_{\nu(i,j)}) = P(Y_2 = 3|E_{\nu(i,j)}) = 0$. This indicates that conditioning on the observed frequencies $E_{\nu(i,j)}$ may break stationarity and pick specific possible choices of $Y_\ell$ ($1 \leq \ell \leq n$) based on the observed frequencies $E_{\nu(i,j)}$.

Due to the above example, we introduce “conditional admissible states” for $Y_\ell$ as the following

**Definition 2** (conditional admissible states). Given the stationary Markov chain $\{Y_\ell\}_{\ell = 1}^{n+1}$ with state space $\Sigma = \{1, 2, \ldots, N\}$ and the observed sample frequencies $E_{\nu(i,j)}$ defined in (13), we define the “conditional admissible states” $\Sigma^\nu(\ell|E_{\nu(i,j)})$ for $Y_\ell$ ($1 \leq \ell \leq n$) as the set of all possible choices of $1 \leq i \leq N$ such that $P(Y_\ell = i|E_{\nu(i,j)}) > 0$. Thus

$$\Sigma^\nu(\ell|E_{\nu(i,j)}) = \left\{ i : 1 \leq i \leq N \ , \ P(Y_\ell = i|E_{\nu(i,j)}) > 0 \right\} , \ 1 \leq \ell \leq n .$$

Given a state $i \in \{1, 2, \ldots, N\}$ and some $1 \leq \ell \leq n$, we further denote the indicator function

$$1^\nu_i = 1_{\Sigma^\nu(\ell|E_{\nu(i,j)})}(i) ,$$

which indicates that state $i$ is conditionally admissible at $Y_\ell$ given the observed frequencies $E_{\nu(i,j)}$.

Given an outcome of the event $E_{\nu(i,j)}$, the sequence $X_1, \ldots, X_n$ must take the form of a string of chain type $X_\ell = (i_1, i_2, \ldots, X_n = (i_{n-1}, i_n)$, such that when counting frequencies, the elements in the string satisfy $E_{\nu(i,j)}$. We introduce the following

**Definition 3** (number of strings of chain type with given term). Given an $(i, j)$ such that $\nu(i,j) \geq 1$ on the event $E_{\nu(i,j)}$, we define by $\#_{1}^{(i,j)}(E_{\nu(i,j)})$ to be the number of different strings of chain type $X_\ell = (i_1, i_2, \ldots, X_n = (i_{n-1}, i_n)$ with the $\ell$-th element being $X_\ell = (i, j)$, and satisfying $E_{\nu(i,j)}$.

It is easy to provide the following example showing that for $j_1 \neq j_2$ we may have $\#_{1}^{(i,j_1)}(E_{\nu(i,j)}) \neq \#_{1}^{(i,j_2)}(E_{\nu(i,j)})$. So that Definition 3 is non-trivial.
Example 2. Suppose \( \{ Y_\ell \}_{\ell \geq 1} \) is a Markov chain with a 3-element state space \( \{ 1, 2, 3 \} \). Set \( n = 3 \) and suppose we have observed

\[
E_{\nu_{(i,j)}} = \{ \nu_{(1,2)} = \nu_{(1,3)} = 0, \nu_{(i,j)} = 1 \text{ for all other pairs of } (i,j) \}.
\]

Then it is easy to see that \( \#_1^{(1,2)}(E_{\nu_{(i,j)}}) = 1 \) and \( \#_1^{(1,3)}(E_{\nu_{(i,j)}}) = 0 \).

With the above definitions, we can compute the posterior marginal probability of \( X_1 \) conditioned upon observed frequencies \( E_{\nu_{(i,j)}} \) as the following

Proposition 1. Given \( 1 \leq i, j \leq N \), then we have

\[
P(X_1 = (i,j) \mid E_{\nu_{(i,j)}}, Y_1 = i) = 1_{1^{\chi_{\nu_{(i,j)}}}} \cdot \frac{\#_2^{(1)}(E_{\nu_{(i,j)}})}{\sum_{k=1}^N \#_2^{(1,2)}(E_{\nu_{(i,j)}})}.
\]

Here we follow the convention that if the events \( \{ Y_1 = i \} \) and \( E_{\nu_{(i,j)}} \) are disjoint, then

\[
P(X_1 = (i,j) \mid E_{\nu_{(i,j)}}, Y_1 = i) = 0.
\]

Proof. The proof follows the same scheme proposed in the i.i.d case (Theorem 1), but with delicate and interesting differences. Given the two events \( E_{\nu_{(i,j)}} \) and \( \{ Y_1 = i \} \) and suppose these two events are not disjoint, then for any pair \( (i,j) \in \Sigma \times \Sigma \), we know that among each \( X_\ell \) in the sequence \( X_1, ..., X_n, \ell = 1, 2, ..., n \), the pair \( (i,j) \) occurs at a multiple of \( \nu_{(i,j)} \) times. Since each \( \nu_{(i,j)} \) is a non-negative integer, and \( \sum_{i=1}^N \sum_{j=1}^N \nu_{(i,j)} = n \), we see that there are only finitely many pairs of \( (i,j) \)’s such that \( \nu_{(i,j)} \geq 1 \), and for all the rest of \( (i,j) \in \Sigma \times \Sigma \) we have \( \nu_{(i,j)} = 0 \). For those pairs of \( (i,j) \)’s such that \( \nu_{(i,j)} \geq 1 \), we order them in an alphabetical order as \( (i_1,j_1) < (i_2,j_2) < ... < (i_I,j_I) \), where \( (i_1,j_1) < (i_2,j_2) \) if and only if \( i_1 < i_2 \) or \( i_1 = i_2, j_1 < j_2 \), and \( I \in \mathbb{N} \) is finite. Consider the set of \( n \) distinguished elements

\[
\chi_{\text{distinguished}} = \bigcup_{i=1}^I \{ ((i_1,j_1), 1), ..., ((i_I,j_I), \nu_{(i_I,j_I)}) \}.
\]

Consider all possible length-\( n \) ordered sequences \( \mathcal{I} \) consisting of distinguished elements of the form \( (((i,j), a) \in \chi_{\text{distinguished}} \) for some \( (i,j) \in \{ (i_1,j_1), ..., (i_I,j_I) \} \) and some \( 1 \leq a \leq \nu_{(i,j)} \), such that the element-wise projection \( \mathcal{Y} \) with \( ((i,j), a) \rightarrow (i,j) \) applied to each of the above sequence \( \mathcal{I} \) results in a string of chain type \( \mathcal{Y}(\mathcal{I}) \) with head \( i \). Such a string of chain type is an outcome of the sequence \( X_1, ..., X_n \) conditioned on \( E_{\nu_{(i,j)}} \) and \( \{ Y_1 = i \} \). We construct a probability measure \( \mathcal{P} \) on the space of all such sequences \( \mathcal{I} \), such that for each sequence we have

\[
\mathcal{P}(\mathcal{I}) = C(i) \prod_{i=1}^I p_{\nu_{(i,j)}}.
\]
for an undetermined normalizing constant $C(i) > 0$, that may depend on $i$. For a given outcome $(i_1, i_2), (i_2, i_3), \ldots, (i_{n-1}, i_n)$ of the sequence $X_1, \ldots, X_n$ such that $i_1 = i$ and when counting frequencies, the sequence $(i_1, i_2), (i_2, i_3), \ldots, (i_{n-1}, i_n)$ satisfies $E_{\nu(i,j)}$, we must have $\left\{ X_1 = (i, i_2), \ldots, X_n = (i_{n-1}, i_n) \right\} \subset E_{\nu(i,j)} \cap \{ Y_1 = i \}$. Therefore for such sequences we must have $P(X_1 = (i, i_2), \ldots, X_n = (i_{n-1}, i_n), E_{\nu(i,j)}, Y_1 = i) = P(X_1 = (i, i_2), \ldots, X_n = (i_{n-1}, i_n)) = \pi_0^i \prod_{l=1}^{n} \nu(i,j)$. Since $P(Y_1 = i) = \pi_0^i$ is the initial probability distribution, we further have

$$
P(X_1 = (i, i_2), \ldots, X_n = (i_{n-1}, i_n) | E_{\nu(i,j)}, Y_1 = i) = \frac{1}{P(E_{\nu(i,j)}, Y_1 = i)} P(X_1 = (i, i_2), \ldots, X_n = (i_{n-1}, i_n), E_{\nu(i,j)}, Y_1 = i)
$$

$$
= \frac{1}{P(E_{\nu(i,j)} | Y_1 = i) P(Y_1 = i)} P(X_1 = (i, i_2), \ldots, X_n = (i_{n-1}, i_n), E_{\nu(i,j)}, Y_1 = i)
$$

$$
= \frac{1}{P(E_{\nu(i,j)} | Y_1 = i)} \prod_{l=1}^{n} \nu(i,j)
$$

$$
= D(i) \prod_{l=1}^{n} \nu(i,j),
$$

(17)

where the constant $D(i) = \frac{1}{P(E_{\nu(i,j)} | Y_1 = i)} > 0$ is an undetermined normalizing constant for the conditional measure $P(E_{\nu(i,j)} | Y_1 = i)$. Comparing (16) and (17), we see that for any given outcome $X_1 = (i, i_2), \ldots, X_n = (i_{n-1}, i_n)$ of the sequence $X_1, \ldots, X_n$ satisfying $E_{\nu(i,j)}$ and $\{ Y_1 = i \}$ we must have

$$
P(X_1 = (i, i_2), \ldots, X_n = (i_{n-1}, i_n) | E_{\nu(i,j)}, Y_1 = i) = K(i) \mathcal{P} \left( \mathcal{S} \right),
$$

(18)

for some constant $K(i) > 0$.

Combining (16), (17) and (18), we see the following two facts

Fact 1': For each of the different sequences $\mathcal{S}$ such that $\mathcal{Y}(\mathcal{S})$ is a string of chain type with head $i$, $\mathcal{P}(\mathcal{S})$ has the same value.

Fact 2': The sequence $\mathcal{S}$ in the RHS of (18) can be an arbitrary length-$n$ sequence picked from all the possible choices of $\mathcal{S}$ such that $\mathcal{Y}(\mathcal{S})$ is a string of chain type with head $i$.

For each string of chain type $((i, i_2), \ldots, (i_{n-1}, i_n))$ with head $i$, we collect all possible sequences $\mathcal{S}$ such that $\mathcal{Y}(\mathcal{S}) = ((i, i_2), \ldots, (i_{n-1}, i_n))$. We claim that we have

$$
P(X_1 = (i, i_2), \ldots, X_n = (i_{n-1}, i_n) | E_{\nu(i,j)}, Y_1 = i) = \mathcal{P} \left( \text{all possible } \mathcal{S} \text{ such that } \mathcal{Y}(\mathcal{S}) = ((i, i_2), \ldots, (i_{n-1}, i_n)) \right).
$$

(19)
This is because due to the above Fact 1’, we have
\[
\mathcal{P} \left( \text{all possible } \mathcal{I} \text{ such that } \mathcal{Y}(\mathcal{I}) = ((i, i_2), \ldots, (i_{n-1}, i_n)) \right) = (\text{Number of all possible } \mathcal{I} \text{ such that } \mathcal{Y}(\mathcal{I}) = ((i, i_2), \ldots, (i_{n-1}, i_n))) \cdot \mathcal{P}(\mathcal{I}).
\]
Now we note that by simple combinatorics we have
\[
\text{Fact 3’: For each realization } (i, i_2), \ldots, (i_{n-1}, i_n) \text{ of } X_1, \ldots, X_n,
\]
\[
\text{(Number of all possible } \mathcal{I} \text{ such that } \mathcal{Y}(\mathcal{I}) = ((i, i_2), \ldots, (i_{n-1}, i_n)))
\]
is independent of the choice of \((i, i_2), \ldots, (i_{n-1}, i_n)\).

Therefore by Fact 3’ we see that (19) is equivalent to
\[
P(X_1 = (i, i_2), \ldots, X_n = (i_{n-1}, i_n)|\mathcal{E}_{\nu(i,j)}, Y_1 = i) = K_1(i)\mathcal{P}(\mathcal{I}).
\]
It is then easy to see that \(K_1(i) = K(i)\) just by normalization of the probability measure \(\mathcal{P}(\bullet)\) and the conditional probability measure \(P(\bullet|\mathcal{E}_{\nu(i,j)}, Y_1 = i)\), as well as the above Fact 2’. So we proved that (19) is valid.

From (19) we know that for an element \((i, j) \in \{i_1, j_1\}, \ldots, (i_I, j_I)\) with the first component \(i\) being fixed, we have
\[
P(X_1 = (i, j)|\mathcal{E}_{\nu(i,j)}, Y_1 = i) = 1_j^2 \cdot \sum_{a=1}^{\nu(i,j)} \mathcal{P} \text{ (all possible } \mathcal{I} \text{ whose 1-st element is } ((i, j), a)) .
\]
We claim that for each element \((i, j), a) \in \mathcal{X}^{\text{distinguished}}\) with \(i\) being fixed we also have

\[
\text{Fact 4’: For each element } ((i, j), a) \in \mathcal{X}^{\text{distinguished}} \text{ with } i \text{ being fixed,
}
\]
\[
\text{(Number of all possible } \mathcal{I} \text{ whose 1-st element is } ((i, j), a))
\]
is independent of \(a\), and is actually equal to
\[
\frac{\nu(i, j_1)! \ldots \nu(i, j_I)!}{\nu(i, j)} \cdot 1_j^2 \cdot \#_1^{((i, j))}(\mathcal{E}_{\nu(i,j)}).
\]
\[
\text{Fact 5’: } \mathcal{P} \text{ (all possible } \mathcal{I} \text{ whose 1-st element is } ((i, j), a)) = \frac{\nu(i, j_1)! \ldots \nu(i, j_I)!}{\nu(i, j)} \cdot 1_j^2 \cdot \#_1^{((i, j))}(\mathcal{E}_{\nu(i,j)}),
\]
p, where \(p\) is independent of \((i, j), a)\).

The above Fact 4’ is a simple combinatorial observation, and Fact 5’ is a consequence of Facts 1’ and 4’. Since we have
\[
\sum_{((i, j), a) \in \mathcal{X}^{\text{distinguished}}, i \text{ fixed}} \mathcal{P} \text{ (all possible } \mathcal{I} \text{ whose 1-st element is } ((i, j), a)) = 1,
\]
\[3\text{This number is actually } \nu(i_1, j_2)! \nu(i_2, j_2)! \ldots \nu(i_{n-1}, j_{n-1})! .
\]
by Facts 4’ and 5’ we see that
\[
1 = \sum_{k_2=1}^{N} 1_{k_2}^{x} \cdot \nu(i,k_2) \cdot \frac{\nu(i_1,j_1)! \ldots \nu(i_r,j_r)!}{\nu(i,k_2)} \cdot \#_1^{(i,k_2)}(\mathcal{E}_{\nu(i,k_2)}) \cdot p ,
\]
so
\[
p = \frac{1}{\sum_{k_2=1}^{N} 1_{k_2}^{x} \cdot \nu(i_1,j_1)! \ldots \nu(i_r,j_r)! \cdot \#_1^{(i,k_2)}(\mathcal{E}_{\nu(i,k_2)})} .
\]
By the above equation, (20) and Fact 5’ we see that (14) holds when \(\mathcal{E}_{\nu(i,j)}\) is not disjoint with \(\{Y_1 = i\}\).

Finally we see that if \(\mathcal{E}_{\nu(i,j)}\) is disjoint with \(\{Y_1 = i\}\), then \(1_{i}^{x} = 0\) and (14) is automatically true. So we have proved the whole statement.

Remark 4 (Conditional Symmetry). The above argument is parallel to the proof of Theorem 1. Here, as is the same in Remark 2, we used the idea of extending the probability space and our obtained identity (19) is again relating the conditional probability to an absolute probability. Fact 1’ is parallel to Fact 1, and is a manifestation of the conditional symmetry in the Markov chain case.

However, in the Markov chain case, we do not have the combinatorial symmetry in exactly the same way as Fact 4 proposed in the proof of Theorem 1. Rather, this is replaced by our new Fact 4’, which leads to Fact 5’. This is the main reason why we cannot simply use the frequency formula
\[
P(X_1 = (i,j)|\mathcal{E}_{\nu(i,j)}, Y_1 = i) = \frac{\nu(i,j)}{\sum_{k_2=1}^{N} \nu(i,k_2)}
\]
(21)
to replace (14). However, we claim that asymptotically it is still true to use (21) to replace our obtained formula (14). This issue will be further investigated in the next Section.

Again, in a same fashion as Remark 1, our result for the Markov chain case can also be proved directly using the fact that the joint distribution of \(X_1, \ldots, X_n\) remains the same conditioned on \(\mathcal{E}_{\nu(i,j)}\) and \(Y_1 = i\) regardless of how we place the outcomes \((i, i_1), \ldots, (i_{n-1}, i_n)\). In our proof, the main purpose of introducing an extended probability space is to show that the combinatorial symmetry will be broken into Fact 5’, so that we cannot simply use the observed frequencies to calculate the posterior marginal distribution.

Proposition 2. Given \(1 \leq i \leq N\), then we have
\[
P(Y_1 = i|\mathcal{E}_{\nu(i,j)}) = \frac{1_{i}^{x} \pi_{0}^{i}}{\sum_{k_1=1}^{N} 1_{k_1}^{x} \pi_{0}^{k_1}} .
\]
(22)
Proof. We first consider the case when the events \( \{ Y_1 = i \} \) and \( \mathcal{E}_{\nu(i,j)} \) are not disjoint. In this case, there must be a pair \((i, j)\) with the first component \(i\) being fixed, such that \(\nu(i,j) \geq 1\). Thus \(\mathcal{E}_{\nu(i,j)} \subseteq \{ Y_1 = i \} \) and we have

\[
P(Y_1 = i | \mathcal{E}_{\nu(i,j)}) = \frac{1}{P(\mathcal{E}_{\nu(i,j)})} P(Y_1 = i, \mathcal{E}_{\nu(i,j)}) = C \pi_i^0 ,
\]

where \(C > 0\) is a normalizing constant.

It is easy to see that \(1_i^0 = 0\) if and only if the events \(\{ Y_1 = i \} \) and \(\mathcal{E}_{\nu(i,j)}\) are disjoint. Thus in general we have

\[
P(Y_1 = i | \mathcal{E}_{\nu(i,j)}) = C1_i^0 \pi_i^0 .
\]

By normalization of the conditional probability \(P(\bullet | \mathcal{E}_{\nu(i,j)})\) we know that the normalization constant \(C = \frac{1}{\sum_{k_1=1}^N 1_k^0 \pi_{k_1}^0}\), so we proved the statement of the Proposition.

Combining Propositions 1 and 2 we easily have

**Theorem 2** (posterior distribution for the finite Markov chain case). Given \(1 \leq i, j \leq N\), then we have

\[
P(X_1 = (i, j) | \mathcal{E}_{\nu(i,j)}) = \frac{1_i^0 \pi_i^0}{\sum_{k_1=1}^N 1_k^0 \pi_{k_1}^0} \cdot \frac{1_j^0}{\sum_{k_2=1}^N 1_k^0 \pi_{k_1}^0} \cdot \frac{\#_{\nu(i,j)}(\mathcal{E}_{\nu(i,j)})}{\#_{\nu(i,j)}(\mathcal{E}_{\nu(i,j)})} . \tag{23}
\]

Proof. This is an easy consequence of the simple conditional probability formula

\[
P(X_1 = (i, j) | \mathcal{E}_{\nu(i,j)}) = P(Y_1 = i | \mathcal{E}_{\nu(i,j)}) \cdot P(X_1 = (i, j) | \mathcal{E}_{\nu(i,j)}, Y_1 = i) , \tag{24}
\]

as well as the fact that \(X_1 = (i, j)\) implies \(Y_1 = i\).

For the rest of this section, we would like to focus more on the combinatorial calculation of the number of strings of chain type \(\#_{\nu(i,j)}(\mathcal{E}_{\nu(i,j)})\) that we have introduced in Definition 3. Actually, this number is calculated based on the so-called Whittle’s formula (see [25], also [2, Theorem 2.1]). For the reader’s convenience, we shall first describe Whittle’s result below. Our formulation of this result is based on [2, Theorem 2.1], but the mathematical terms and symbols we use will follow those in our present paper.

Given a frequency event \(\mathcal{E}_{\nu(i,j)}\) and a string of chain type \(X_1 = (i_1, i_2), ..., X_n = (i_n, i_{n+1})\) that satisfies \(\mathcal{E}_{\nu(i,j)}\), it is easy to observe that

\[
\sum_{j=1}^N \nu(i,j) - \sum_{j=1}^N \nu(j,i) = 1_{\{i=i_1\}} - 1_{\{i=i_{n+1}\}} . \tag{25}
\]
It is easy to see that once $E_{\{\nu_{(i,j)}\}}$ is given, then $i_1$ is fixed if we fix $i_{n+1}$, and $i_{n+1}$ is fixed if we fix $i_1$. However, if we do not fix either $i_1$ or $i_{n+1}$, then we may have different choices of both of them. As an example, the strings of chain type $(1,2),(2,1)$ and $(2,1),(1,2)$ correspond to the same frequency event $E_{\{\nu_{(i,j)}\}}$, but the choices of $i_1$ and $i_{n+1}$ can be different.

Recall we have assumed that the state space of the chain $\{Y_\ell\}_{\ell \geq 1}$ is a finite set $\Sigma = \{1,2,\ldots,N\}$. Let us fix some $u, v \in \{1,2,\ldots,N\}$ and consider all possible strings of chain type $X_1 = (i_1,i_2), \ldots, X_n = (i_n,i_{n+1})$ that satisfy the given frequency event $E_{\{\nu_{(i,j)}\}}$, such that $i_1 = u, i_{n+1} = v$. The total number of such strings of chain type is denoted by $N_{uv}^{(n)}(E_{\{\nu_{(i,j)}\}})$. We shall first form a matrix $F^*$ of size $N \times N$ with elements $F^* = (\nu^*_ij)_{1 \leq i,j \leq N}$, where

$$
\nu^*_ij = \begin{cases} 
1_{i=j} - \frac{\nu(i,j)}{\sum_{j=1}^{N} \nu(i,j)}, & \text{if } \sum_{j=1}^{N} \nu(i,j) > 0, \\
1_{i=j}, & \text{if } \sum_{j=1}^{N} \nu(i,j) = 0.
\end{cases}
$$

The following result is due to Whittle in 1955 (see [25]) and its proof is also presented in [2, Theorem 2.1].

**Theorem 3** (Whittle’s formula, 1955). We have

$$
N_{uv}^{(n)}(E_{\{\nu_{(i,j)}\}}) = \frac{\prod_{i=1}^{N} \left( \sum_{j=1}^{N} \nu(i,j) \right)!}{\prod_{i=1}^{N} \prod_{j=1}^{N} \nu(i,j)!} F^*_{vu},
$$

where $F^*_{vu}$ is the $(v,u)$-th cofactor of the matrix $F^*$ and $0! = 1$.

Very simply, the Whittle’s formula provides us with an easy calculation of the quantity $#_1^{(i,j)}(E_{\{\nu_{(i,j)}\}})$ that we have introduced in Definition 3. To this end, for given
\( i, j \in \{1, 2, ..., N\} \) we define the \( N \times N \) matrix \( \tilde{F}^* = (\tilde{\nu}_{ij}^*)_{1 \leq i, j \leq N} \), where

\[
\tilde{\nu}_{ij}^* = \begin{cases} 
1_{\{i=j\}} - \frac{\nu_{\{i,j\}} - 1}{\sum_{k=1}^{N} \nu_{\{i,k\}} - 1} , & \text{if } \tilde{i} = \tilde{j} = j \text{ and } \sum_{k=1}^{N} \nu_{\{i,k\}} > 1 , \nu_{\{i,j\}} \geq 1 , \\
1_{\{i=j\}} - \frac{\nu_{\{i,j\}} - 1}{\sum_{k=1}^{N} \nu_{\{i,k\}} - 1} , & \text{if } \tilde{i} = \tilde{j} \neq j \text{ and } \sum_{k=1}^{N} \nu_{\{i,k\}} > 1 , \nu_{\{i,j\}} \geq 1 , \\
1_{\{i=j\}} , & \text{if } \tilde{i} = i \text{ and } \sum_{k=1}^{N} \nu_{\{i,k\}} = 1 , \nu_{\{i,j\}} = 1 , \\
\frac{\nu_{\{i,j\}}}{\sum_{k=1}^{N} \nu_{\{i,k\}}} , & \text{if } \tilde{i} \neq i \text{ and } \sum_{k=1}^{N} \nu_{\{i,k\}} > 0 , \\
1_{\{i=j\}} , & \text{if } \tilde{i} \neq i \text{ and } \sum_{k=1}^{N} \nu_{\{i,k\}} = 0 .
\end{cases}
\]

(28)

Then we have

Corollary 1 (The exact calculation of the number of strings of chain type with given term). Given the frequency event \( \mathcal{E}_{\{(i,j)\}} \) that satisfies (26) with fixed \( i_1 = i \) and \( i_{n+1} = v \), and suppose that \( \nu_{\{i,j\}} \geq 1 \) for fixed \( j \). Then the quantity \( \#_1^{(i,j)}(\mathcal{E}_{\{(i,j)\}}) \) that we have introduced in Definition 3 can be calculated as

\[
\#_1^{(i,j)}(\mathcal{E}_{\{(i,j)\}}) = \begin{cases} 
\frac{\nu_{\{i,j\}}}{\sum_{j=1}^{N} \nu_{\{i,j\}}} \cdot N_{iv}^{(n)}(\mathcal{E}_{\{(i,j)\}}) \cdot \frac{\tilde{F}_{vj}^*}{\tilde{F}_{vi}^*} , & \text{if } N_{iv}^{(n)}(\mathcal{E}_{\{(i,j)\}}) > 0 ; \\
0 , & \text{if } N_{iv}^{(n)}(\mathcal{E}_{\{(i,j)\}}) = 0 ,
\end{cases}
\]

(29)

where \( N_{iv}^{(n)}(\mathcal{E}_{\{(i,j)\}}) \) is the quantity defined in the Whittle’s formula (27) with \( u = i, v = v; F_{vj}^* \) is the \((v, i)\)-th cofactor of the matrix \( F^* = (\nu_{ij})_{1 \leq i, j \leq N} \) with elements defined in (26), and \( \tilde{F}_{vj}^* \) is the \((v, j)\)-th cofactor of the matrix \( \tilde{F}^* = (\tilde{\nu}_{ij})_{1 \leq i, j \leq N} \) with elements defined in (25).

Proof. Given \( \mathcal{E}_{\{(i,j)\}} \) and some fixed \( i, j \in \{1, 2, ..., N\} \), let us first suppose that at least one string of chain type \( X_1 = (i, j), X_2 = (j, i_3), ..., X_n = (i_n, i_{n+1}) \) that satisfies \( \mathcal{E}_{\{(i,j)\}} \) exists, which implies that \( \#_1^{(i,j)}(\mathcal{E}_{\{(i,j)\}}) \geq 1 \). It is easy to observe that the sequence \( X_2 = (j, i_3), ..., X_n = (i_n, i_{n+1}) \) forms a string of chain type that satisfies the frequency event \( \mathcal{E}_{\{(i,j)\}} \) with

\[
\tilde{\nu}_{\{i,j\}} = \begin{cases} 
\nu_{\{i,j\}} , & \text{if } (\tilde{i}, \tilde{j}) \neq (i, j) , \\
\nu_{\{i,j\}} - 1 , & \text{if } (\tilde{i}, \tilde{j}) = (i, j) .
\end{cases}
\]

(30)

We note that, when \( \mathcal{E}_{\{(i,j)\}} \) and the starting state \( i \) are fixed, then the final state \( i_{n+1} = v \in \{1, 2, ..., N\} \) will be fixed, and thus \( N_{iv}^{(n)}(\mathcal{E}_{\{(i,j)\}}) \geq 1 \). Let us also note that
\( \mathcal{E}(\tilde{\nu}_{(i,j)}) \) must satisfy (25) in the form of

\[
N \sum_{j=1}^{N} \tilde{\nu}(i,j) - N \sum_{j=1}^{N} \tilde{\nu}(j,i) = 1_{\{i=j\}} - 1_{\{i=v\}}.
\]

This means that given \( \mathcal{E}(\tilde{\nu}_{(i,j)}) \) and the starting state \( j \), the final state must be \( v \). Thus it is easy to see that

\[
\#^{(i,j)}_{1}(\mathcal{E}_{\nu_{(i,j)}}) = N^{(n-1)}_{j,v}(\mathcal{E}_{\tilde{\nu}_{(i,j)}}).
\]

According to Theorem 3, the RHS of (31) is equal to

\[
N^{(n-1)}_{j,v}(\mathcal{E}_{\tilde{\nu}_{(i,j)}}) = \frac{\prod_{i=1}^{N} \left( \sum_{j=1}^{N} \tilde{\nu}_{(i,j)} \right)!}{\prod_{i=1}^{N} \prod_{j=1}^{N} \tilde{\nu}_{(i,j)}!} \tilde{F}_{vj},
\]

where \( \tilde{F}_{vj} \) is the \( (v, j) \)-th cofactor of the matrix \( \tilde{F}^* = (\tilde{\nu}_{ij})_{1 \leq i,j \leq N} \), with

\[
\tilde{\nu}^*_{ij} = \begin{cases} 
1_{\{i=j\}} - \frac{\sum_{j=1}^{N} \tilde{\nu}(i,j)}{\sum_{j=1}^{N} \tilde{\nu}(i,j)}, & \text{if } \sum_{j=1}^{N} \tilde{\nu}(i,j) > 0, \\
1_{\{i=j\}}, & \text{if } \sum_{j=1}^{N} \tilde{\nu}(i,j) = 0.
\end{cases}
\]

It is then easy to see that (33) is the same as (28) due to (30).

By (30), we also know that

\[
\prod_{i=1}^{N} \left( \sum_{j=1}^{N} \tilde{\nu}_{(i,j)} \right)! = \prod_{i=1}^{N} \left( \sum_{j=1}^{N} \nu_{(i,j)} \right)!^{\nu_{(i,j)}}.
\]

Therefore

\[
\prod_{i=1}^{N} \left( \sum_{j=1}^{N} \tilde{\nu}_{(i,j)} \right)! = \frac{\prod_{i=1}^{N} \left( \sum_{j=1}^{N} \nu_{(i,j)} \right)!^{\nu_{(i,j)}}}{\prod_{j=1}^{N} \nu_{(i,j)}!}.
\]

Moreover, by (30) we know that

\[
\prod_{i=1}^{N} \prod_{j=1}^{N} \tilde{\nu}_{(i,j)}! = \prod_{i=1}^{N} \prod_{j=1}^{N} \nu_{(i,j)}!.
\]
we apply the second part of (29) and we know that it is still correct.

Using (38), we arrive at the matrix

\[
F = 1,
\]

Then \( F \) is the matrix defined in (33); and \( N_{iv} \) with \( i, v \leq 1 \) that satisfies \( \tilde{e} \text{ is the } (v, u) \)-th cofactor of the matrix \( F \) \( \text{matrix} \), \( tiv \) \( \text{defined in (26)}. Under our assumption we find that for each \( 1 \leq j \leq N \text{, the argument is the same. Thus the Lemma is proved.} \]

Proof. Recall that the matrix \( F^* = (\nu_{ij})_{1 \leq i, j \leq N} \) defined in (26). A version of the same Lemma can be found in [2] Lemma 4. For comprehensiveness, we will also provide a short proof of this Lemma.

**Lemma 1.** Suppose that for each \( 1 \leq v \leq N \) we have \( \sum_{k=1}^{N} \nu_{(v,k)} > 0 \). Then for each \( 1 \leq v \leq N \) and each pair \( 1 \leq i, j \leq N \) we have \( F^*_{vi} = F^*_{vj} \).

**Proof.** Recall that the matrix \( F^* = (\nu_{ij})_{1 \leq i, j \leq N} \) has the matrix elements \( \nu_{ij} \) that we defined in (26). Under our assumption we find that for each \( 1 \leq v \leq N \) we have

\[
\sum_{i=1}^{N} \nu_{v,i} = 0.
\]

To prove the statement of this lemma, without loss of generality we can assume that \( v = 1, i = 1 \) and \( j = 2 \). Set the matrices

\[
F_{11}^* = \begin{pmatrix}
\nu_{21}^* & \nu_{23}^* & \cdots & \nu_{2N}^*
\nu_{31}^* & \nu_{33}^* & \cdots & \nu_{3N}^*
\vdots & \vdots & \ddots & \vdots
\nu_{N1}^* & \nu_{N3}^* & \cdots & \nu_{NN}^*
\end{pmatrix},
F_{12}^* = \begin{pmatrix}
\nu_{21}^* & \nu_{23}^* & \cdots & \nu_{2N}^*
\nu_{31}^* & \nu_{33}^* & \cdots & \nu_{3N}^*
\vdots & \vdots & \ddots & \vdots
\nu_{N1}^* & \nu_{N3}^* & \cdots & \nu_{NN}^*
\end{pmatrix}.
\]

Then \( F_{11}^* = \det F_{11}^* \) and \( F_{12}^* = -\det F_{12}^* \). It is then easy to see that by adding each of the second, third, ..., until the last columns of the matrix \( F_{11}^* \) to its first column, and using (35), we arrive at the matrix \( -F_{12}^* \). So we conclude that \( F_{11}^* = F_{12}^* \). For any other \( 1 \leq i, j \leq N \), the argument is the same. Thus the Lemma is proved.

\[
20
\]
4 The infinite-sample limit of the Markov chain case

In this section, by making use of Theorem 2 we tend to investigate the limit of the posterior probability $P(X_1 = (i, j) | \mathcal{E}_{\{\nu_{(i,j)}\}})$ as the number of observations $n$ tends to infinity. Recall that we have introduced in Section 3 the Markov chain $\{Y_\ell\}_{\ell \geq 1}$ with finite state space $\Sigma = \{1, \ldots, N\}$, $|\Sigma| = N$. Let the transition probability matrix of the process $\{Y_\ell\}_{\ell \geq 1}$ be given by $P = (p_{ij})_{1 \leq i, j \leq N}$. Assume the process starts from an initial probability distribution $\pi^0 = (\pi^0_1, \ldots, \pi^0_N)$, $0 \leq \pi^0_i \leq 1$, $\sum_{i=1}^N \pi^0_i = 1$, such that $P(Y_1 = i) = \pi^0_i$. Within this section, we will work under the following

Assumption 1 (Positive Transition Probabilities). The Markov chain $\{Y_\ell\}_{\ell \geq 1}$ has positive transition probabilities, i.e. each transition probability $p_{ij} > 0$ for any $1 \leq i, j \leq N$. Thus the Markov chain $\{Y_\ell\}_{\ell \geq 1}$ is ergodic, and its invariant measure is given by $\pi_i > 0$, $i = 1, 2, \ldots, N$ with $\sum_{i=1}^N \pi_i = 1$.

Let us fix the number of states $N$ and the number of observations $n$, so that we consider the Markov chain $Y_1, \ldots, Y_n, Y_{n+1}$ and the “consecutive pair” process $X_\ell = (Y_\ell, Y_{\ell+1})$ for $1 \leq \ell \leq n+1$ as in Section 3. Under this assumption, the number of trajectories $(Y_1, \ldots, Y_{n+1})$ is finite and is equal to $N^{n+1}$. Since each trajectory corresponds to a frequency event $\mathcal{E}_{\{\nu_{(i,j)}\}}$ as defined in (13), and two trajectories may correspond to the same frequency event, we see that the number of different frequency events $\mathcal{E}_{\{\nu_{(i,j)}\}}$ will not exceed $N^{n+1}$ and thus will be finite. Furthermore, it is easy to observe that two different frequency events must be disjoint. Thus we label all possible frequency events by the index $\lambda \in \Lambda$ where $\Lambda$ is a finite set, and we denote each frequency event by $\mathcal{E}_{\{\nu_{(i,j)}\}}^\lambda$. We claim that in the limit we have

Theorem 4 (Asymptotic of the posterior probability). For any $\varepsilon > 0$ small enough, there exist some $M \geq 1$ and some $n_0 = n_0(\varepsilon, M) \in \mathbb{N}$ such that for any $n \geq n_0$, there exists a family of frequency events $\mathcal{E}_{\{\nu_{(i,j)}\}}^\lambda$, $\lambda \in \Lambda \subseteq \Lambda$ such that

$$P \left( \bigcup_{\lambda \in \Lambda} \mathcal{E}_{\{\nu_{(i,j)}\}}^\lambda \right) \geq 1 - \frac{\varepsilon}{M} \quad , \quad (39)$$

and for each frequency event $\mathcal{E}_{\{\nu_{(i,j)}\}}^\lambda$, $\lambda \in \Lambda$, the posterior probability of $X_1$ conditioned on $\mathcal{E}_{\{\nu_{(i,j)}\}}^\lambda$ is close to the unconditioned probability of $X_1$, i.e. for any $1 \leq i, j \leq N$ we have

$$\left| P(X_1 = (i, j) | \mathcal{E}_{\{\nu_{(i,j)}\}}^\lambda) - \frac{\prod_{k=1}^N \pi^0_{k1}^{k_1} \cdot p_{ij}}{\sum_{k=1}^N \prod_{k=1}^N \pi^0_{k1}^{k_1}} \cdot p_{ij} \right| < \varepsilon \quad . \quad (40)$$

For example: when the chain $\{Y_\ell\}_{\ell \geq 1}$ is given by $1 \rightarrow 1 \rightarrow 1 \rightarrow 2 \rightarrow 1$ or $1 \rightarrow 2 \rightarrow 1 \rightarrow 1 \rightarrow 1$, both cases will correspond to $\mathcal{E}_{\{\nu_{(i,j)}\}}^\lambda = \{\nu_{(1,2)} = 1, \nu_{(1,1)} = 2, \nu_{(i,j)} = 0\}$ for all other pairs of $(i, j)$.\footnote{For example: when the chain $\{Y_\ell\}_{\ell \geq 1}$ is given by $1 \rightarrow 1 \rightarrow 1 \rightarrow 2 \rightarrow 1$ or $1 \rightarrow 2 \rightarrow 1 \rightarrow 1 \rightarrow 1$, both cases will correspond to $\mathcal{E}_{\{\nu_{(i,j)}\}}^\lambda = \{\nu_{(1,2)} = 1, \nu_{(1,1)} = 2, \nu_{(i,j)} = 0\}$ for all other pairs of $(i, j)$.}
Proof. By using the conditional probability formula (24) we have

\[
P(X_1 = (i, j) | E_{\nu(i,j)}^\lambda) - \frac{1^{i,j} \pi^0_{1}}{\sum_{k=1}^{N} 1^{i,j} \pi^0_{k1}} \cdot p_{ij}
\]

\[
= P(Y_1 = i | E_{\nu(i,j)}^\lambda) \cdot P(X_1 = (i, j) | E_{\nu(i,j)}^\lambda, Y_1 = i) - \frac{1^{i,j} \pi^0_{i}}{\sum_{k=1}^{N} 1^{i,j} \pi^0_{k1}} \cdot p_{ij}
\]

\[
\leq P(Y_1 = i | E_{\nu(i,j)}^\lambda) \cdot P(X_1 = (i, j) | E_{\nu(i,j)}^\lambda, Y_1 = i) - P(Y_1 = i | E_{\nu(i,j)}^\lambda) \cdot p_{ij}
\]

\[
\leq P(X_1 = (i, j) | E_{\nu(i,j)}^\lambda, Y_1 = i) - p_{ij}
\]

(41)

where in (a) we have used (22) in Proposition 2 and in (b) we have used the fact that

\[
P(Y_1 = i | E_{\nu(i,j)}^\lambda) \leq 1.
\]

Given \( \varepsilon > 0 \), we will prove in Lemma 2 that there exists some \( M_1 \geq 1 \) and some \( n_0^{(1)} = n_0^{(1)}(\varepsilon, M_1) \in \mathbb{N} \), so that for any \( n \geq n_0^{(1)} \) and any frequency event \( E_{\nu(i,j)}^\lambda, \lambda \in \tilde{\Lambda} \) chosen as in Lemma 3 with \( \mu = \frac{\varepsilon}{M_1} \), we have

\[
P(X_1 = (i, j) | E_{\nu(i,j)}^\lambda, Y_1 = i) - \frac{\nu(i,j)}{\sum_{k=1}^{N} \nu(i,k)} \leq \frac{\varepsilon}{2}.
\]

(42)

Here we identify each frequency event \( E_{\nu(i,j)}^\lambda, \lambda \in \tilde{\Lambda} \) as such that each of the ratios

\[
\frac{\nu(i,j)}{\sum_{k=1}^{N} \nu(i,k)}
\]

is close to \( \pi_i p_{ij} \), where \( \{\pi_i\}_{i=1}^{N} \) is the invariant measure of \( \{Y^\epsilon, \nu(i,j)\}_{i,j,k=1}^{N} \) introduced in Assumption 1 (see Lemma 3 for details).

Notice that by the simple ergodic theorem for Markov chains (see [29, Section 1.10]), there exists some \( n_0^{(2)} = n_0^{(2)}(\varepsilon) \in \mathbb{N} \) such that for any \( n \geq n_0^{(2)} \) we have

\[
\left| \frac{\nu(i,j)}{\sum_{k=1}^{N} \nu(i,k)} - p_{ij} \right| < \frac{\varepsilon}{2},
\]

(43)

for sufficiently small \( \varepsilon > 0 \).

Using (11), (12) and (13) we know that if we pick \( M = \max(M_1, M_2) \) and \( n_0 = \max(n_0^{(1)}, n_0^{(2)}) \), then we can form the set of frequency events \( E_{\nu(i,j)}^\lambda, \lambda \in \tilde{\Lambda} \) satisfying (39) and for any \( n \geq n_0 \), any such frequency event \( E_{\nu(i,j)}^\lambda \), we must have

\[
P(X_1 = (i, j) | E_{\nu(i,j)}^\lambda) - \frac{1^{i,j} \pi^0_{1}}{\sum_{k=1}^{N} 1^{i,j} \pi^0_{k1}} \cdot p_{ij}
\]

\[
\leq P(X_1 = (i, j) | E_{\nu(i,j)}^\lambda, Y_1 = i) - p_{ij}
\]

\[
\leq P(X_1 = (i, j) | E_{\nu(i,j)}^\lambda, Y_1 = i) - \frac{\nu(i,j)}{\sum_{k=1}^{N} \nu(i,k)} + \frac{\nu(i,j)}{\sum_{k=1}^{N} \nu(i,k)} - p_{ij}
\]

\[
\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,
\]

(44)
where in (a) we have used (41), in (b) we have used the triangle inequality, and in (c) we have used (42) and (43). This implies (40).

**Remark 5** \((Y_1\) cannot be any state). It is not true that when the number of samples \(n\) is large we must have \(1_i^T \pi_0 = 1\) for any \(i \in \{1, 2, \ldots, N\}\). That is to say, not all states are admissible at the first position. This is simply because of (25), since when the frequency event \(E^\lambda\{\nu_{(i,j)}\}\) is given, the initial state \(i_1\) and the final state \(i_{n+1}\) must satisfy (25). Because of this reason, the term \(1_i^T \pi_0 \sum_{k=1}^{N} \nu_{(k)} \cdot p_{ij}\) in (40) cannot be replaced by \(\pi_0^T p_{ij}\).

**Lemma 2** (Asymptotic of Proposition 1). For any \(\varepsilon > 0\) small enough, there exists some \(M \geq 1\) and some \(n_0 = n_0(\varepsilon, M) \in \mathbb{N}\), so that for any \(n \geq n_0\) and any of the frequency events \(E^\lambda\{\nu_{(i,j)}\}\) chosen from \(\lambda \in \Lambda\) as in Lemma 3 with \(\mu = \frac{\varepsilon}{M}\), for any \(1 \leq i, j \leq N\) we have

\[
\left| P \left( X_1 = (i, j) \mid E^\lambda\{\nu_{(i,j)}\}, Y_1 = i \right) \right| \leq \frac{\varepsilon}{2}.
\]  

(44)

**Proof.** Suppose we have been given one of any frequency events \(E^\lambda\{\nu_{(i,j)}\}\) with \(\lambda \in \Lambda\) and also \(Y_1 = i\). Under this assumption, recall that in our proof of Proposition 1 we have introduced the set \(\mathcal{X}^{\text{distinguished}}\) in (15) and all possible length-\(n\) ordered sequences \(\mathcal{S}\) consisting of distinguished elements of the form \(((i, j), a) \in \mathcal{X}^{\text{distinguished}}\), such that the element-wise projection \(\mathcal{Y}\) with \(((i, j), a) \mathcal{Y} (i, j)\) applied to each of the above sequence \(\mathcal{S}\) results in a string of chain type \(\mathcal{Y}(\mathcal{S})\) with head \(i\). By Fact 1’ in the proof of Proposition 1 we know that

\[
\mathcal{P} (\text{all possible } \mathcal{S} \text{ whose 1-st element is } ((i, j), a)) = (\text{Number of all possible } \mathcal{S} \text{ whose 1-st element is } ((i, j), a)) \cdot p
\]

where \(p \in (0, 1)\) is independent of \(((i, j), a)\). Since we have

\[
\sum_{((i, j), a) \in \mathcal{X}^{\text{distinguished}}, i \text{ fixed}} \mathcal{P} (\text{all possible } \mathcal{S} \text{ whose 1-st element is } ((i, j), a)) = 1,
\]

and by Fact 4’ stated in the proof of Proposition 1 indicating that

\[
(\text{Number of all possible } \mathcal{S} \text{ whose 1-st element is } ((i, k), a))
\]

is independent of \(a\), we obtain that for any choices of \(1 \leq a \leq \nu_{(i,k)}\),

\[
1 = \sum_{k=1}^{N} \nu_{(i,k)} \cdot (\text{Number of all possible } \mathcal{S} \text{ whose 1-st element is } ((i, k), a)) \cdot p,
\]

23
which gives
\[
p = \frac{1}{\sum_{k=1}^{N} \nu(i,k) \cdot (\text{Number of all possible } S \text{ whose 1-st element is } ((i, k), a))}.
\] (45)

We recall the formula (20) in the proof of Proposition 1 and we combine it with the above (45) to obtain
\[
P(X_1 = (i, j) | E_{\lambda}^{\nu(i,j)}, Y_1 = i) = \frac{\nu(i,j)}{\sum_{a=1}^{\nu(i,j)} \gamma(k,j)}.
\] (46)

where we denote
\[
\gamma(k,j) = \frac{\text{Number of all possible } S \text{ whose 1-st element is } ((i, k), a)}{\text{Number of all possible } S \text{ whose 1-st element is } ((i, j), a')}
\]

Notice that \(\gamma(k,j)\) is independent of \(a\) and \(a'\). The finite existence of \(\gamma(k,j)\) is guaranteed by part (a) of Lemma 5.

Given any \(\varepsilon > 0\), by part (b) of Lemma 5 we know that there exists some \(M_1 \geq 1\) and some \(n_0^{(1)} = n_0^{(1)}(\varepsilon, M_1) \in \mathbb{N}\) such that for any \(n \geq n_0^{(1)}\), for any of our given frequency events \(E_{\nu(i,j)}^{\varepsilon}\), \(\lambda \in \tilde{\Lambda}\) chosen as in Lemma 1 with \(\mu = \frac{\varepsilon}{M_1}\) and any \(1 \leq k, j \leq N\), that
\[
|\gamma(k,j) - 1| < \frac{\varepsilon}{4}.
\]

This combined with (46) enable us to estimate
\[
\begin{align*}
& \left| P(X_1 = (i, j) | E_{\nu(i,j)}^{\varepsilon}, Y_1 = i) - \frac{\nu(i,j)}{\sum_{k=1}^{N} \nu(i,k)} \right| \\
\leq & \frac{\nu(i,j)}{\sum_{k=1}^{N} \nu(i,k)} \left(1 - \frac{1}{\sum_{k=1}^{N} \nu(i,k)} \right) - \frac{1}{\sum_{k=1}^{N} \nu(i,k)} = \frac{\nu(i,j)}{\sum_{k=1}^{N} \nu(i,k)} \cdot \frac{1}{\sum_{k=1}^{N} \nu(i,k)} \\
< & \frac{\varepsilon}{4} \cdot \frac{\nu(i,j)}{\sum_{k=1}^{N} \nu(i,k)} \cdot \frac{1}{\sum_{k=1}^{N} \nu(i,k)} = \frac{\varepsilon}{4} \cdot \frac{\nu(i,j)}{\sum_{k=1}^{N} \nu(i,k)} \\
< & \frac{\varepsilon}{1 - \frac{\varepsilon}{4}} - \frac{\varepsilon}{4} \leq \frac{\varepsilon}{1 - \frac{\varepsilon}{4}}.
\end{align*}
\] (47)
By Lemma 4 we know that there exists some $M_2 \geq 1$ and some $n_0^{(2)} = n_0^{(2)}(\varepsilon, M_2) \in \mathbb{N}$ and when $n \geq n_0^{(2)}$, for any of our frequency events $E_{\nu(i,j)}^\lambda$, $\lambda \in \tilde{\Lambda}$ chosen as in Lemma 3 with $\mu = \varepsilon M_2$, we have
\[ |1^{1/\varepsilon} - 1| = 0 . \]

We pick $M = \max(M_1, M_2)$ and $n_0 = \max(n_0^{(1)}, n_0^{(2)})$. By using (47), we can form the set of frequency events $E_{\nu(i,j)}^\lambda$, $\lambda \in \tilde{\Lambda}$ satisfying Lemma 3 with $\mu = \varepsilon M_2$ and for any $n \geq n_0$, any such frequency event $E_{\nu(i,j)}^\lambda$, we must have
\[
\left| \mathbb{P}(X_1 = (i,j)|E_{\nu(i,j)}^\lambda, Y_1 = i) - \frac{\nu(i,j)}{\sum_{k=1}^{N} \nu(i,k)} \right| + \left| 1^{1/\varepsilon} - 1 \right| \cdot \frac{\nu(i,j)}{\sum_{k=1}^{N} \nu(i,k)} < \varepsilon \leq \frac{1}{2} ,
\]
when $0 < \varepsilon \leq 2$. This proves (44).

**Remark 6.** As we have explained in Remark 1, an alternate and seemingly simpler proof of the above Lemma 2 can be obtained by directly using Proposition 1 and Corollary 1. The argument is parallel to the one used in the proof of Lemma 5 below. We omit details here. However, the way of proof we adopt here reveals more of the underlying symmetric structure of the problem (i.e. “conditional symmetry” at different levels, see Remark 8). We expect that such arguments based on symmetry are more fundamental and should be extended to more general classes of processes.

**Lemma 3 (Asymptotic of frequency events).** For any $\mu > 0$ there exists some $n_0 = n_0(\mu) \in \mathbb{N}$ such that for any $n \geq n_0$, there exists a family of frequency events $E_{\nu(i,j)}^\lambda$, $\lambda \in \tilde{\Lambda} \subseteq \Lambda$ with
\[
\mathbb{P}\left( \bigcup_{\lambda \in \tilde{\Lambda}} E_{\nu(i,j)}^\lambda \right) \geq 1 - \mu ,
\]
and for each frequency event $E_{\nu(i,j)}^\lambda$, $\lambda \in \tilde{\Lambda}$, its corresponding frequencies $\nu_{(i,j)}$ satisfy that for any $1 \leq i, j \leq N$,
\[
\left| \frac{\nu(i,j)}{n} - \pi_i p_{ij} \right| < \mu ,
\]
where $\pi_i, i = 1, 2, ..., N$ is the invariant measure of the Markov chain $\{Y_\ell\}_{\ell \geq 1}$ and $p_{ij}$ are the transition probabilities.

**Proof.** By our Assumption 1 from the weak Law of Large Numbers for ergodic Markov chain (see [17, Section 5.4]) we know that for any $\mu > 0$, there exists some $n_0 = n_0(\mu) \in \mathbb{N}$
and for all \( n \geq n_0 \) we have

\[
P \left( \left| \frac{\nu_{(i,j)}}{n} - \pi_ip_{ij} \right| \geq \mu \right) \leq \mu ,
\]

which implies the statement of the Lemma.

\( \square \)

**Remark 7** (Desired frequency events cannot be of full probability). It is not appropriate to conclude here that \( P \left( \bigcup_{\lambda \in \tilde{\Lambda}} E_{\bar{\nu}_{(i,j)}}^{\lambda} \right) = 1 \) by using the strong Law of Large Numbers for ergodic Markov chain (see [20 Section 1.10]), since in that case the threshold \( n_0 \) may depend on the element \( \omega \) in the probability space \( \Omega \), i.e. the assertion that

\[
P \left( \lim_{n \to \infty} \frac{\nu_{(i,j)}}{n} = \pi_ip_{ij} \right) = 1
\]

implies that for some \( \tilde{\Omega} \subseteq \Omega \) with \( P(\tilde{\Omega}) = 1 \), for any \( \mu > 0 \) and any \( \omega \in \tilde{\Omega} \), there exists some \( n_0 \in \mathbb{N} \) that may depend on \( \omega \), such that \( \left| \frac{\nu_{(i,j)}}{n} - \pi_ip_{ij} \right| < \mu \).

Recall that the indicator function \( 1_{i,j}^{\ell} \) is defined as in Definition 2, which indicates that state \( i \) is conditionally admissible at \( Y_\ell \) given the observed frequencies \( E_{\nu_{(i,j)}} \). As the number of observations \( n \) tends to infinity, we have

**Lemma 4** (All states are asymptotically conditionally admissible at \( Y_2 \)). For any \( j \in \{1, 2, ..., N\} \) and any \( 1 \leq \ell \leq n \), for any \( \varepsilon > 0 \) and any \( M \geq 1 \), there exist some \( n_0 \in \mathbb{N} \) such that when \( n \geq n_0 \), for any frequency event \( E_{\tilde{\nu}_{(i,j)}}^{\lambda} \) chosen from \( \lambda \in \tilde{\Lambda} \) as in Lemma 3 with \( \mu = \varepsilon/M \) and admitting \( Y_1 = i \), we have

\[
1_{j}^{\ell} = 1.
\]

That is to say, in the \( n \to \infty \) asymptotic, all states are conditionally admissible at \( Y_2 \).

**Proof.** Without loss of generality we shall suppress the upper-index \( \lambda \) in \( E_{\bar{\nu}_{(i,j)}}^{\lambda} \). Since \( N_{iw}^{(n)}(E_{\nu_{(i,j)}}) \geq 1 \), by (50) in the proof of Corollary 1 we know that

\[
N_{jv}^{(n-1)}(E_{\bar{\nu}_{(i,j)}}) = \frac{\nu_{(i,j)}}{N} \cdot \sum_{j=1}^{N} \nu_{(i,j)} \cdot \frac{1}{\sum_{j=1}^{N} \nu_{(i,j)}} \cdot \frac{\tilde{F}_{v}^{s}}{F_{v}^{s}},
\]

where we use the same notations in as Corollary 1 and we refer to the reader for more details there. We then see that in order to prove (50) it suffices to show that \( N_{jv}^{(n-1)}(E_{\bar{\nu}_{(i,j)}}) \geq 1 \). As \( n \) is large, by Lemma 3 we know that for any \( 1 \leq i, j \leq N \) we must have \( \nu_{(i,j)} \geq 1 \). By Lemma 4 we know that for any \( 1 \leq j \leq N \) we have \( \tilde{F}_{v}^{s} \to F_{v}^{s} \) as \( n \to \infty \). By (49) we know that the assumption of Lemma 1 is satisfied, and thus we know that \( \frac{\tilde{F}_{v}^{s}}{F_{v}^{s}} = 1 \). These facts imply that \( N_{jv}^{(n-1)}(E_{\bar{\nu}_{(i,j)}}) \geq 1 \) when \( n \) is large, and thus (50) is proved.

\( \square \)
Recall in Remark 4 we have explained that the Fact 4’ in the proof of Proposition 1 is different from the “conditional symmetry” as the Fact 4 of the proof of Theorem 1. Here we show that as \( n \to \infty \), Fact 4’ will asymptotically become “conditionally symmetric” and thus it becomes a version of Fact 4 stated in the proof of Theorem 1. Using the same notations as in the proof of Proposition 1, we have

**Lemma 5** (Asymptotic conditional symmetry). For any \( \varepsilon > 0 \), there exist some \( M \geq 1 \) and some \( n_0 \in \mathbb{N} \) such that when \( n \geq n_0 \), for any of the frequency events \( \mathcal{E}_{(\nu_{(i,j)})}^\lambda \) chosen from \( \lambda \in \Lambda \) as in Lemma 3 with \( \mu = \frac{\varepsilon}{M} \) and admitting \( Y_1 = i \in \{1, 2, ..., N\} \), we have

(a) For any \( j \in \{1, 2, ..., N\} \) and any \( 1 \leq a \leq \nu_{(i,j)} \), the number of all possible \( \mathcal{F} \) whose 1-st element is \((i, j), a \geq 1\);

(b) For any two elements \((i, j_1), a_1\), \((i, j_2), a_2\) \( \in \mathcal{F}_{\text{distinguished}} \) with \( j_1 \neq j_2, j_1, j_2 \in \{1, 2, ..., N\} \) and \( 1 \leq a_1 \leq \nu_{(i,j_1)} \), \( 1 \leq a_2 \leq \nu_{(i,j_2)} \), we have

\[
\begin{align*}
\left| \text{Number of all possible } \mathcal{F} \text{ whose 1-st element is } ((i, j_1), a_1) \right| & < \frac{\varepsilon}{4}.
\end{align*}
\] (51)

**Proof.** We make use of Fact 4’ in the proof of Proposition 1, so that for each element \((i, j), a) \in \mathcal{F}_{\text{distinguished}} \) with \( i \) being fixed,

\[
\begin{align*}
\text{(Number of all possible } \mathcal{F} \text{ whose 1-st element is } ((i, j), a))
&= \frac{\nu_{(i,j)}! \cdots \nu_{(i,j)}!}{\nu_{(i,j)}!} \cdot 1^j \cdot \#_{1}^{(i,j)}(\mathcal{E}_{\nu_{(i,j)}}^\lambda).
\end{align*}
\]

Since \( Y_1 = i \in \{1, 2, ..., N\} \) is admitted by the frequency event \( \mathcal{E}_{\nu_{(i,j)}}^\lambda \), we know that \( \#_{1}^{(i,j)}(\mathcal{E}_{\nu_{(i,j)}}^\lambda) \geq 1 \). By Lemma 4 we know that when \( n \) is large, \( 1^j = 1 \). Thus part (a) is proved.

For part (b), by (29) in Corollary 1 we know that

\[
\begin{align*}
\#_{1}^{(i,j)}(\mathcal{E}_{\nu_{(i,j)}}^\lambda) &= \frac{\nu_{(i,j)}}{\nu_{(i,j)}} \cdot \frac{N_{(i,j)}}{N} \cdot \frac{\tilde{F}_{v_i}}{F_{v_i}}.
\end{align*}
\]

Therefore

\[
\begin{align*}
\text{(Number of all possible } \mathcal{F} \text{ whose 1-st element is } ((i, j), a))
&= \frac{\nu_{(i,j)}! \cdots \nu_{(i,j)}!}{\nu_{(i,j)}!} \cdot 1^j \cdot \frac{\nu_{(i,j)}}{\nu_{(i,j)}} \cdot \frac{N_{(i,j)}^{(i,j)}}{N_{(i,j)}} \cdot \frac{\tilde{F}_{v_i}}{F_{v_i}}
\ &= \frac{\nu_{(i,j)}! \cdots \nu_{(i,j)}!}{\sum_{j=1}^{N} \nu_{(i,j)}} \cdot \frac{N_{(i,j)}^{(i,j)}}{N_{(i,j)}} \cdot \frac{\tilde{F}_{v_i}}{F_{v_i}}.
\end{align*}
\]
Here the matrices $F^* = (\nu^*_{ij})_{1 \leq i,j \leq N}$ and $\tilde{F}^* = (\tilde{\nu}^*_{ij})_{1 \leq i,j \leq N}$ are defined in (26) and (28); and $\tilde{F}_{vj}^*$, $F_{vi}^*$ are the $(v,j)$-th and $(v,i)$-th cofactor of these two matrices, respectively.

Thus we have

\[
\frac{\text{Number of all possible } S \text{ whose 1-st element is } ((i, j), a_1)}{\text{Number of all possible } S \text{ whose 1-st element is } ((i, j), a_2)} = 1 \cdot \frac{\tilde{F}_{vj}^*}{F_{vj}^*} \cdot \frac{\tilde{F}_{ij}^*}{F_{ij}^*}.
\]

By Lemma 6 we know that for any $1 \leq j \leq N$ we have $\tilde{F}_{vj}^* \rightarrow F_{vj}^*$ as $n \rightarrow \infty$. By (49) we know that the assumption of Lemma 1 is satisfied, and thus we know that $\frac{\tilde{F}_{vj}}{F_{vj}} = 1$. These facts together with (52) and Lemma 1 imply (51). So we have also proved part (b).

Remark 8 (Asymptotic Conditional Symmetry at the level of observations of one-step transitions). From Lemma 5 we see that, as the number of observations $n$ tends to infinity, a higher level of symmetry is manifested at the fact that the numbers of admissible trajectories starting from different initial one-step transitions tend to be evenly distributed. From here, the “conditional symmetry” at the level of sample path trajectories as shown in Fact 4’ of the proof of Proposition 4 is reduced to the “conditional symmetry” at the level of observations of one-step transitions. The latter level of symmetry is essentially the same as Fact 4 in the proof of Theorem 1 for the i.i.d. case. By this reason, when $n \rightarrow \infty$, the ergodic finite Markov chain case should have the same kind of posterior distribution as the i.i.d. case, with the only difference of replacing the frequencies of state occurrences by the frequencies of state transitions. This is exactly what Lemma 2 indicates.

Recall that we have defined the matrices $F^* = (\nu^*_{ij})_{1 \leq i,j \leq N}$ and $\tilde{F}^* = (\tilde{\nu}^*_{ij})_{1 \leq i,j \leq N}$ in (26) and (28), respectively. The following lemma shows that under Assumption 1 these two matrices have asymptotically the same elements.

Lemma 6 (Asymptotic of the matrices $F^*$ and $\tilde{F}^*$). For the given $1 \leq i,j \leq N$ that are used to define $\tilde{F}^* = (\tilde{\nu}^*_{ij})_{1 \leq i,j \leq N}$ in (28), assume that $\nu(i,j) \geq 1$ in $\mathcal{E}_{\{i,j\}}$ and the Assumption 1 holds. Then for any $\varepsilon > 0$ there exists some $n_0 \geq 1$ such that when $n \geq n_0$ and for any pair $(k,l), 1 \leq k,l \leq N$ we have

\[
|\tilde{\nu}^*_{kl} - \nu^*_{kl}| < \varepsilon. \tag{53}
\]

Proof. According to (28), if $k \neq i$, then we actually have $\tilde{\nu}^*_{kl} = \nu^*_{kl}$, so (53) is automatically true. Now suppose $k = i$, then since $\nu(i,j) \geq 1$, we must have $\sum_{l=1}^{N} \nu^*_{(i,l)} > 0$. Thus
according to (26), for each $1 \leq l \leq N$ we have
\[
\nu^*_n = 1_{\{i = l\}} - \frac{\nu_{(i,l)}}{\sum_{i=1}^{N} \nu_{(i,l)}}.
\] (54)

Without loss of generality we can assume that $n$ is large, so $\sum_{l=1}^{N} \nu_{(i,l)} = 1$, $\nu_{(i,j)} = 1$ will not happen, because otherwise we must have $n = 1$. Thus we know from (28) that
\[
\tilde{\nu}^*_n = \begin{cases} 
1_{\{i = j\}} - \frac{\nu_{(i,j)} - 1}{\sum_{i=1}^{N} \nu_{(i,l)} - 1}, & \text{if } l = j, \\
1_{\{i = l\}} - \frac{\nu_{(i,l)} - 1}{\sum_{i=1}^{N} \nu_{(i,l)} - 1}, & \text{if } l \neq j.
\end{cases}
\] (55)

From (55) and (54) we can calculate that when $l = j$, we have
\[
|\nu^*_n - \tilde{\nu}^*_n| = \left| \frac{\nu_{(i,j)} - 1}{\sum_{i=1}^{N} \nu_{(i,l)} - 1} \right| < \frac{1}{\sum_{i=1}^{N} \nu_{(i,l)} - 1},
\]
and when $l \neq j$ we have
\[
|\nu^*_n - \tilde{\nu}^*_n| = \left| \frac{\nu_{(i,l)} - 1}{\sum_{i=1}^{N} \nu_{(i,l)} - 1} \right| < \frac{1}{\sum_{i=1}^{N} \nu_{(i,l)} - 1}.
\]

Thus in order that (53) holds, it suffices to have
\[
\sum_{i=1}^{N} \nu_{(i,l)} - 1 > \frac{1}{\varepsilon},
\]
which is a result of (49) in Lemma 3.

5 Discussion, Remarks, and Generalization

5.1 Koopman-Damois exponential family of models and maximum entropy principle

Recent studies on applying the probability theory of large deviations to nanothermodynamics [4, 11, 18, 22] have shed considerable new light on the nature of Gibbs’ equilibrium theory of statistical mechanics and statistical chemistry: The former introduces the notion of statistical ensemble for the probability of energy and system’s
volume, and the latter generalizes the method to counting the number of atoms and molecules. It is clear now that Gibbs’ theory, in fact the entire theory of thermodynamics proper, is an approach that combines Bayesian statistical inference and limit theorems by formulating a posterior probability for a representative member in a large system that is conditioned on a limit law. In fact, each specific limit law also implies a family of probabilistic models with parameters (conjugate variables) that can be determined by the empirical observation as sufficient statistics. The last aspect of thermodynamics is precisely the phenomenological thermodynamics that accounts for fluctuations first proposed by L. Szilard in 1925 [23] and further developed by B. Mandelbrot [19], in terms of the Koopman-Damois (KD), also known as exponential, family of models [14, 16]. The sufficient statistics also has a deep logic relation to the Maximum Entropy Principle (MEP) which uniquely determines the posterior distribution based on idealized “date”, as elucidated in [24].

The KD family is a consequence of the large-deviation posterior inference, via exponential tilting [1, 3]. The present work further shows that such a factorized form is valid even for posterior conditioned on finite observations. More specifically, one has the measure-theoretic formulation

\[
P(dx|\hat{\theta}) = \exp \left[ \eta(\hat{\theta})g(x) - \Lambda(\hat{\theta}) \right] \mu(dx) \quad (56a)
\]

\[
\times \exp \left[ I(g(x)) \right] \mu(dx), \quad (56b)
\]

\[
\Lambda(\theta) = \sup_g \left\{ \eta(\theta)g - I(g) \right\}, \quad (56c)
\]

\[
\hat{\theta}(g^*) = \arg \sup_{\theta} \left\{ \eta(\theta)g^* - \Lambda(\theta) \right\}. \quad (56d)
\]

w.r.t. a reference measure \( \mu \). In [56], \( \eta \) is the conjugate variable to the random variable \( g(x) \), its cumulant generating function \( \Lambda(g) \) is the Legendre-Fenchel transform of the LDRF \( I(g) \), and \( g^* \) is the empirical arithmetic mean value of observing \( g \). In classical thermodynamics, \( I(g) \) and \( \Lambda(\hat{\theta}) \) are identified with the Gibbs entropy and the Helmholtz free energy, respectively, when \( \eta \) is temperature and \( g \) is internal energy. Gibbs’ statistical mechanics approach with explicit consideration of atoms and molecules and Szilard-Mandelbrot’s statistical inference approach to phenomenological thermodynamics including fluctuations, thus, are now unified under the Probability Theory.

Indeed, contraction principle in the theory of large deviations consists of three parts: It provides a mathematical justification for a new, lower level LDP and it supplies a method for computing the new, corresponding LDRF in terms of a constrained optimization of the entropy function at hand. The very contraction principle also predicts a posterior probability conditioned on the constrain; this is the central idea of Gibbs conditioning [6]. The last aspect is the statistical foundation of MEP. Alternatively stated, maximizing the entropy function under an asymptotic constraint defines a conditional law on a set of infinitely large samples, independent or correlated. The
Gibbs conditioning problem associated with Sanov’s LDP for i.i.d. sample frequencies was carefully studied by van Campenhout and Cover [24]. This version of the MEP has been applied to a wide range of problems in recent years [21].

5.2 A conjecture for continuous-time Markov Chains

Let \( \{Y_t\}_{t \geq 0} \) be a continuous-time Markov process with finite state space \( \Sigma = \{1, 2, ..., N\} \) and Q-matrix \( Q = (q_{ij})_{1 \leq i, j \leq N} \) and where \( q_{ii} = -\sum_{j \in \{1, 2, ..., N\} \setminus \{i\}} q_{ij} \). Suppose we observed the process \( Y_t \) during time \( t \in [0, T] \), \( T > 0 \). For each pairs of states \( i, j \in \{1, 2, ..., N\} \) we record the number of jumps from \( i \) to \( j \) up to time \( T \), and we denote it by \( \nu_{(i,j)}^{[0,T]} \). Let the integer \( N_i^{[0,T]} \geq 0 \) be the number of visits of the process \( \{Y_t\}_{t \geq 0} \) to state \( i \in \{1, 2, ..., N\} \) up to time \( T > 0 \). We record the empirical occupation times

\[
m_i^{[0,T]} = \frac{1}{N_i^{[0,T]}} \int_0^T \mathbf{1}_{\{Y_t = i\}} dt, \quad i = 1, 2, ..., N.
\]

Note that if \( N_i^{[0,T]} = 0 \), then we must have \( \int_0^T \mathbf{1}_{\{Y_t = i\}} dt = 0 \) and for any \( j \in \{1, 2, ..., N\} \) we have \( \nu_{(i,j)}^{[0,T]} = \nu_{(j,i)}^{[0,T]} = 0 \). In this case, by convention, we will set \( m_i^{[0,T]} = +\infty \).

Given the frequency event

\[
\mathcal{E}_{\nu_{(i,j)}^{[0,T]}} \equiv \{\nu_{(i,j)}^{[0,T]} : 1 \leq i, j \leq N\},
\]

such that

\[
\sum_{j=1}^N \nu_{(i,j)}^{[0,T]} - \sum_{j=1}^N \nu_{(j,i)}^{[0,T]} = \mathbf{1}_{\{i = i_1\}} - \mathbf{1}_{\{i = i_{n+1}\}},
\]

we consider all length-\( n \) strings of chain type

\[
X_1 = (i_1, i_2), X_2 = (i_2, i_3), ..., X_n = (i_n, i_{n+1}),
\]

satisfying \( \mathcal{E}_{\nu_{(i,j)}^{[0,T]}} \) and such that \( n = \sum_{i,j=1}^N \nu_{(i,j)}^{[0,T]} \).

Given an \( (i, j) \) such that \( \nu_{(i,j)}^{[0,T]} \geq 1 \) on the event \( \mathcal{E}_{\nu_{(i,j)}^{[0,T]}} \), we define by \( \#_{k}^{(i,j)}(\mathcal{E}_{\nu_{(i,j)}^{[0,T]}}) \) to be the number of different strings of chain type \( X_1 = (i_1, i_2), ..., X_n = (i_{n-1}, i_n) \) with the \( \ell \)-th element being \( X_\ell = (i, j) \), and satisfying \( \mathcal{E}_{\nu_{(i,j)}^{[0,T]}} \). Let \( \mathbf{1}_{(i,j)}^{k} \) be the indicator function which is 1 if there is at least one string of chain type \( X_1 = (i_1, i_2), ..., X_n = (i_{n-1}, i_n) \) satisfying \( \mathcal{E}_{\nu_{(i,j)}^{[0,T]}} \) with \( X_\ell = (i, *) \), and it is 0 otherwise. With all these, we conjecture that
**Conjecture.** Conditioned on $\mathcal{E}_{\nu_{i,j}(0,T)}$ and $m_{i}^{[0,T]} \ (i = 1, 2, \ldots, N)$, we have

$$
\lim_{\Delta t \to 0} \frac{1}{\Delta t} P \left( Y_{\Delta t} = j \mid \mathcal{E}_{\nu_{i,j}(0,T)}, m_{i}^{[0,T]}, Y_{0} = i \right) = \begin{cases} 
1_{i}^{[0,T]} \cdot q_{ij}^{[0,T]}, & \text{if } j \neq i; \\
-1_{i}^{[0,T]} \cdot \frac{1}{m_{i}^{[0,T]}}, & \text{if } j = i.
\end{cases}
$$

Here for $j \neq i$ we have

$$
q_{ij}^{[0,T]} = \frac{1}{m_{i}^{[0,T]}} \cdot \frac{1_{2}^{[0,T]} \cdot \#_{1}^{(i,j)}(\mathcal{E}_{\nu_{i,j}(0,T)})}{\sum_{k_{2} = 1}^{N_{2}} 1_{2}^{[0,T]} \cdot \#_{1}^{(i,k_{2})}(\mathcal{E}_{\nu_{i,k_{2}}(0,T)})}.
$$

As time $T \to \infty$, we further expect that

$$
\lim_{T \to \infty} \left| \frac{\hat{m}_{i}^{[0,T]} - m_{i}^{[0,T]}}{m_{i}^{[0,T]}} \right| = 0, \quad \lim_{T \to \infty} m_{i}^{[0,T]} = m_{i}, \quad \lim_{T \to \infty} q_{ij}^{[0,T]} = q_{ij}.
$$

We expect to prove the above conjecture using similar arguments as we did in Sections 3-4 of this paper.

**Acknowledgements.** WH would like to thank Louis Waitong Fan, Hao Ge, Yong Liu, Hui He and Yin Ouyang for reading an earlier version of the manuscript and helpful comments, and financial support from the Simons Foundation. HQ acknowledges Chris Burdzy and Jian-Sheng Xie for discussions and the Olga Jung Wan Endowed Professorship.

**References**

[1] A. C. Barato and R. Chetrite. A formal view on level 2.5 large deviations and fluctuation relations. *Journal of Statistical Physics*, 160(5):1154–1172, 2015.

[2] P. Billingsley. Statistical methods in Markov chains. *Ann. Math. Statist.*, 32(1):12–40, 1961.

[3] R. Chetrite and H. Touchette. Nonequilibrium Markov processes conditioned on large deviations. *Annales Henri Poincaré*, 16(9):2005–2057, 2015.

[4] J. Commons, Y.-J. Yang, and H. Qian. Duality symmetry, two entropy functions, and an eigenvalue problem in Gibbs’ theory. *arXiv:2108.08948*, 2021.

[5] I. Csiszár, T. Cover, and B.-S. Choi. Conditional limit theorems under Markov conditioning. *IEEE Transactions on Information Theory*, IT-33(6):788–801, 1987.
[6] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications, Second Edition*. Springer, 1998.

[7] P. Diaconis and D. Freedman. de Finetti’s theorem for Markov chains. *The Annals of Probability*, 8:115–130, 1980.

[8] P. Diaconis and B. Skyrms. *Ten Great Ideas About Chance*. Princeton Univ. Press, New Jersey, 2018.

[9] W. E. The dawning of a new era in applied mathematics. *Notices of the American Mathematical Society*, 68(4):565–571, 2021.

[10] M. I. Freidlin and A. D. Wentzell. *Random Perturbations of Dynamical Systems, third edition*. Springer, 2012.

[11] T. L. Hill. *Thermodynamics of Small Systems*. Dover, New York, 1963.

[12] P. J. Huber. *Robust Statistics*. John Wiley & Sons, 1981.

[13] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[14] H. Jeffreys. An extension of the Pitman–Koopman theorem. *Mathematical Proceedings of the Cambridge Philosophical Society*, 56(4):393–395, 1960.

[15] A. N. Kolmogoroff. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, 1933.

[16] B. O. Koopman. On distributions admitting a sufficient statistic. *Trans. Amer. Math. Soc.*, 39(3):399–409, 1936.

[17] L. B. Koralov and Y. G. Sinai. *Theory of Probability and Random Processes (Second Edition)*. Springer, 2007.

[18] Z. Lu and H. Qian. Emergence and breaking of duality symmetry in thermodynamic behavior: repeated measurements and macroscopic limit. *arXiv:2009.12644*, 2020.

[19] B. Mandelbrot. The role of sufficiency and of estimation in thermodynamics. *The Annals of Mathematical Statistics*, 33(3):1021–1038, 1962.

[20] J. R. Norris. *Markov Chains*. Cambridge University Press, 1997.

[21] S. Pressé, K. Ghosh, J. Lee, and K. A. Dill. Principles of maximum entropy and maximum caliber in statistical physics. *Rev. Mod. Phys.*, 85(3):1115–1141, 2013.

[22] H. Qian. Thermodynamic behavior of statistical event counting in time: Independent and correlated measurements. *arXiv:2109.12806*, 2021.
[23] L. Szilard. Über die ausdehnung der phänomenologischen thermodynamik auf die schwankungserscheinungen. *Z. Physik.*, 32:753–7888, 1925.

[24] J. van Campenhout and T. Cover. Maximum entropy and conditional probability. *IEEE Transactions on Information Theory*, 27(4):483–489, 1981.

[25] P. Whittle. Some distribution and moment formulae for the Markov Chain. *Journal of the Royal Statistical Society. Series B (Methodological)*, 17(2):235–242, 1955.