Large deviations for general Markov chains on discrete state spaces

Léo Daures[1]

[1]Université Paris Cité and Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, F-75013 Paris, France
email: daures@lpsm.paris

Abstract

We study the large deviations of Markov chains under the sole assumption that the state space is discrete. In particular, we do not require any of the usual irreducibility and exponential tightness assumptions. Using subadditive arguments, we provide an elementary and self-contained proof of the level-2 and level-3 large deviation principles. Due to the possible reducibility of the Markov chain, the rate functions may be nonconvex and may differ, outside a specific set, from the Donsker-Varadhan entropy and other classical rate functions.

Contents

1 Introduction

Let $X = (X_n)_{n\geqslant 1}$ be a Markov chain on a discrete countable state space S associated with a stochastic kernel p and an initial distribution β, and let $\mathbb{P}$ denote the law of X. From the seminal work of Donsker and Varadhan [16, 17, 18, 19] to the recent publication of a comprehensive monograph by de Acosta [10], large deviations of Markov chains have become a standard subject in probability theory. A *Large Deviation Principle* (LDP) describes an exponential rate of decay of probabilities. This work primarily focuses on *weak LDPs*. We recall that weak LDPs differ from standard LDPs in that the upper bound is only required to hold for all compact (instead of all closed) sets; see Definition 1.1. Following standard terminology, we will occasionally refer to the standard LDP as the *full LDP* to emphasize the contrast with the weak LDP.

For Markov chains, it is usual to study the large deviations of the time average of a given function $f : S \to \mathbb{R}^d$, of the occupation times, and of the empirical process, respectively defined as

$$A_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i), \qquad L_n^{(1)} = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}, \qquad L_n^{(\infty)} = \frac{1}{n}\sum_{i=1}^{n} \delta_{T^i(X)}, \qquad n \geqslant 1, \tag{1.1}$$

where T is the shift operator acting on $S^{\mathbb{N}}$ by $T(x_1, x_2, \dots) = (x_2, x_3, \dots)$. Here, $A_n f$ is a d-dimensional vector, $L^{(1)}n$ is a probability measure on S, and $L_n^{(\infty)}$ is a probability measure on $S^{\mathbb{N}}$. Following the terminology of [20], we refer to the (weak) LDPs of the three sequences in (1.1) as *level-1*, *level-2*, and *level-3* (weak) LDPs respectively.

The level-1, level-2, and level-3 full LDPs for Markov chains have been extensively studied in the literature. Most existing results are stated under assumptions that ensure two properties: one of *irreducibility* and one of *exponential tightness* (see Section 1.1 for definitions). The irreducibility property can be a consequence of the uniformity assumption **(U)** as in [13, 11], or assumed directly, while requiring the kernel p to be *matrix-irreducible* or *φ-irreducible*,[1] as in [10]. Typically, irreducibility yields the convexity of the rate functions; simple examples of Markov chains which are not matrix-irreducible, and which satisfy LDPs with nonconvex rate functions were already exhibited in [14]. The property of exponential tightness can be implied by various hypothesis; one is **(H*)**, originally introduced by Donsker and Varadhan in [18], but assumption **(U)** of [13, 11] is also sufficient. Broadly speaking, exponential tightness guarantees that the LDPs are full and that the rate functions are good. In the absence of exponential tightness, the validity of the full level-2 and level-3 LDPs, as well as that of the (weak) level-1 LDP, seems to depend on the fine details of the Markov chain. We recall in Appendix D some well-known[2] examples of Markov chains (on countable spaces) which lack exponential tightness or goodness of the rate function, and for which only the level-2 and level-3 *weak* LDPs hold, and which even fail to satisfy the level-1 weak LDP.

In this paper, we prove that (X_n) always satisfies the level-2 and level-3 weak LDP; see Theorems 1.4 and 1.5 below. Besides discreteness of the space S, no assumption is made on the kernel p or on the initial distribution β, and in particular, the results hold without exponential tightness or irreducibility. In view of the above discussion, the stated weak LDPs are the best achievable results in this situation. These LDPs will be obtained as consequences of an auxiliary result, which is interesting in its own right, namely the weak LDP for the *pair empirical measures*; see Theorem 1.3.

In the derivation of the above weak LDPs, we will introduce the sets $\mathcal{A}^{(k)}$ of *admissible measures*; see Definition 1.2. These sets contain the measures that are relevant for the LDPs, and

[1] Most authors simply call *irreducibility* the notion of φ-irreducibility. Even if it may be improper we will keep the name φ-irreducibility to avoid any ambiguity.

[2] See for instance Section 3 of [3], Example 1 of [5], Exercises 13.14 and 13.15 of [28] and Example 10.3 of [10].

the rate functions are trivially infinite on their complement. Since LDPs describe exponential decay of probabilities, we stress that measures that give positive probability to transient states may be admissible and have a finite rate function. Examples are provided in Appendix D.

The conclusions of Theorems 1.3, 1.4, and 1.5 are not new in themselves. In 2002, de La Fortelle and Fayolle [21] claimed the weak LDP for the pair empirical measure with modified version of the relative entropy as rate function; unfortunately their proof appears to contain some small gaps. In 2005, Jiang and Wu [22] proved the level-2 and level-3 weak LDPs with modified versions of the Donsker-Varadhan entropy as rate functions. This proof was based on some functional analysis arguments from a previous rich article by Wu [30]. Eventually, in 2015, the level-2 weak LDP made its way to the book of Rassoul-Agha and Seppäläinen [28], where a proof is outlined as an adaptation of the proof for the irreducible case. All these works define a notion of admissible measure. We believe that the novelty of the present article lies in two major features. First, we provide a proof based on subadditivity, which is self-contained and purely probabilistic. Second, we provide an exhaustive identification of the LDP rate functions.

More precisely, the first feature consists in the use of the *subadditive method*, which requires significant adaptations in order to handle reducible Markov chains. In large deviations theory, the subadditive method consists in applying a version of the subadditive lemma (see for example Lemma 6.1.11 in [11]), in order to prove the existence of the *Ruelle-Lanford function* (RL function), which, in turn, implies the desired weak LDP; see Section 2.2. Subadditive techniques have been employed extensively in the study of large deviations of irreducible Markov chains, leading to concise and elegant proofs of weak LDPs. For instance, under assumption **(U)**, subadditive arguments are used to derive the level-2 and level-3 LDPs in Chapter 4 of [13] and Chapter 6 of [11]. The method also appears in [6] under some *decoupling* assumptions which cover irreducible Markov chains on finite state spaces. Some subadditive arguments have also been used outside of a full application of the method, as in Chapter 2 of [10] (building on ideas from [15] and [9]), where subadditive arguments yield a LDP lower bound. The classical subadditive method, and the variations used in all the above references, inherently relies on irreducibility and always produces convex rate functions. The improvements to the method that we provide involve carefully *slicing* and *stitching* finite-length trajectories, in a way that accomodates (possible) reducibility. To the best of the author's knowledge, this work provides the first instance of a subadditive argument beyond irreducibility and convexity. In addition to offering an elegant and self-contained proof of the level-2 and level-3 weak LDPs, we hope that these developments will pave the way for future extensions and applications of the subadditive method.

The second feature of the present article is the comprehensive identification of the rate functions. We identify the rate functions with all known standard expressions on the set of admissible measures $\mathcal{A}^{(k)}$, and show that they are infinite outside $\mathcal{A}^{(k)}$. The latter property is at the root of the above-mentioned nonconvexity of the rate functions in non-irreducible cases. In addition to the possible lack of convexity, the rate functions that we obtain may not be good, and the LDPs may not be full. The lack of these properties makes the identification of the rate functions more delicate than usual.

This article proceeds as follows. In the remainder of the current section, we provide some notation and definitions, and state the main results of this article. In Section 2, we present the modifications of the subadditive method required to accommodate the generality of our setup and derive the weak LDP for the *pair empirical measure*, defined in (1.5). Section 3 is dedicated to the computation of the rate function of this weak LDP. Sections 4 and 5 use this weak LDP to derive respectively the weak LDP for the occupation times, and the weak LDP for the empirical process.

1.1 Definitions and main results

Let us recall the definition of exponential tightness, goodness, and of the weak and full LDP.

Definition 1.1. *Let $(Y_n) = (Y_n)_{n\in\mathbb{N}}$ be a random process on a Hausdorff space $\mathcal{X}$, endowed with its Borel σ-algebra.*

1. *We say that (Y_n) is exponentially tight if for all $\alpha < \infty$, there exists a compact set $K \subseteq \mathcal{X}$ such that*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(Y_n \notin K) < -\alpha.$$

2. *We say that (Y_n) satisfies the weak (respectively, full) LDP if there exists a lower semicontinuous function $I : \mathcal{X} \to [0, +\infty]$ such that*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(Y_n \in K) \leqslant - \inf_{x\in K} I(x), \tag{1.2}$$

for every compact (respectively, closed) set K in $\mathcal{X}$ and

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(Y_n \in U) \geqslant - \inf_{x\in U} I(x), \tag{1.3}$$

for every open set U of $\mathcal{X}$. If so, I is called the rate function *of the LDP. We say that the rate function is* good *if its level sets are compact, i.e. if $\{x \in \mathcal{X} : I(x) \leqslant C\}$ is compact for every $C \in \mathbb{R}$. We refer to* (1.3) *as the* lower bound *and to* (1.2) *as the* upper bound *of the (weak) LDP.*

When a random process satisfies a full LDP or a weak LDP on a Polish space, its rate function is unique: see Lemma 4.1.4 and Exercise 4.1.30 of [11].

Let $k \in \mathbb{N} := \{1, 2, \ldots\}$. Let $\mathcal{P}(S^k)$ denote the set of probability measures on S^k, embedded in the vector space $\mathcal{M}(S^k)$ of finite signed measures on S^k. Let $\mathcal{B}(S^k)$ denote the set of bounded real-valued functions on S^k.[3] We define the dual pairing

$$(\mu, V) \mapsto \langle \mu, V\rangle = \sum_{u\in S^k} V(u)\mu(u), \qquad V \in \mathcal{B}(S^k),\ \mu \in \mathcal{M}(S^k), \tag{1.4}$$

and we call *weak topology* the coarsest topology on $\mathcal{M}(S^k)$ making $\langle \cdot, V\rangle$ continuous for all $V \in \mathcal{B}(S^k)$.[4] We equip $\mathcal{M}(S^k)$ with the weak topology. A convenient metric to work with is the total variation (TV) distance, denoted $|\cdot|_{\mathrm{TV}}$. Since S is discrete, $|\cdot|_{\mathrm{TV}}$ metrizes the weak topology on $\mathcal{P}(S^k)$; this is because $|\cdot|_{\mathrm{TV}}$ can be expressed as half the ℓ^1 norm on $\mathcal{P}(S^k)$. In $\mathcal{P}(S^k)$, the open $|\cdot|_{\mathrm{TV}}$-ball for of radius ρ centered at μ is denoted by $\mathcal{B}(\mu, \rho)$. When S_1 is a subset of S, we identify $\mathcal{P}(S_1^k)$ with the set $\{\mu \in \mathcal{P}(S^k) \mid \operatorname{supp}\mu \subseteq S_1^k\}$, where $\operatorname{supp}\mu = \{u \in S^k \mid \mu(u) > 0\}$ is the support of μ. The space $\mathcal{P}(S^k)$ is complete; see for instance Theorem 6.5 of [25].

Most of the work in this paper is done in $\mathcal{P}(S^2)$, with the pair empirical measure L_n defined as

$$L_n = L_n^{(2)} = \frac{1}{n} \sum_{k=1}^{n} \delta_{(X_k, X_{k+1})} \in \mathcal{P}(S^2), \qquad n \geqslant 1. \tag{1.5}$$

The measure L_n, which counts the transitions between each pair of states, carries more useful information than the occupation times $L_n^{(1)}$. Following [21], we say that a finite measure μ on S^2 is *balanced* if its first and second marginal are equal *i.e* if

$$\mu(A \times S) = \mu(S \times A), \quad A \subseteq S. \tag{1.6}$$

We call $\mu^{(1)}$ the measure on S for which $\mu^{(1)}(A)$ is given by the left-hand side of (1.6) (or either side of (1.6) when μ is balanced). The set of all balanced probability measures is closed and we denote it by $\mathcal{P}_{\mathrm{bal}}(S^2)$. It will play a central role in the weak LDP of Theorem 1.3 below because, by definition, L_n is $\frac{2}{n}$-close to a balanced measure.

As we consider Markov chains that may be reducible, the decomposition of S into irreducible classes plays a central role in the following. The stochastic kernel p defines a relation $\rightsquigarrow$ on S with $x \rightsquigarrow y$ if there exists a finite sequence $x = x_1, x_2, \ldots, x_{n+1} = y$ in S satisfying $p(x_i, x_{i+1}) > 0$ for all $1 \leqslant i \leqslant n$.[5] We then say that y is *reachable* from x. An irreducible class is a maximal subset C of S such that $x \rightsquigarrow y$ for all $(x, y) \in C^2$. We denote by B the set of all points $x \in S$ that do not belong to any irreducible class; these are the points such that $x \not\rightsquigarrow x$, meaning that they can be visited by X at most once. The state space S can be partitioned as

$$S = B \cup \left(\bigcup_{j\in\mathcal{J}} C_j \right), \tag{1.7}$$

[3]We do not mention measurability nor continuity since S^k is discrete!

[4]The *dual* pairing of (1.4) can also be used to define a weak topology on $\mathcal{B}(S^k)$; see Appendix B.

[5]Note that, unlike other definitions of communication, this one requires the sequence to have at least two elements. In particular, we do not have $x \rightsquigarrow x$ in general.

where the C_j are the irreducible classes and $\mathcal{J}$ is countable.[6] We say that the class $C_{j'}$ is reachable from C_j and write $C_j \rightsquigarrow C_{j'}$ if some $y \in C_{j'}$ is reachable from some $x \in C_j$.[7] The relation $\rightsquigarrow$ is a partial order on $(C_j)_{j\in\mathcal{J}}$. We also define reachability from β by writing $\beta \rightsquigarrow y$ if $y \in \operatorname{supp}\beta$ or if there exists $x \in \operatorname{supp}\beta$ such that $x \rightsquigarrow y$, and $\beta \rightsquigarrow C_j$ if some $y \in C_j$ is reachable from β.[8] If there is a j such that $x \rightsquigarrow C_j$ for all $x \in S$, the Markov chain is *φ-irreducible*. If there is only one class and B is empty, the Markov chain is *matrix-irreducible*.

Using this decomposition into irreducible classes, we can now define admissible measures. In this definition and in the following, if μ is a measure and A is a set, $\mu|_A$ denotes the restriction of μ to A, which is the measure defined by $\mu|_A(\cdot) = \mu(A \cap \cdot)$.

Definition 1.2 (Admissibility)**.** *Let $k \in \mathbb{N}$. Let $\mu \in \mathcal{P}(S^k)$. We say that μ is* pre-admissible *if there exists a set of indices $\mathcal{J}_\mu \subseteq \mathcal{J}$ such that*

$$\mu = \sum_{j\in\mathcal{J}_\mu} \mu|_{C_j^k},$$

in which case we impose that $\mathcal{J}_\mu$ is minimal. Moreover, if μ is pre-admissible, we say that μ is admissible *if the order $\rightsquigarrow$ is total on $(C_j)_{j\in\mathcal{J}_\mu}$ and $\beta \rightsquigarrow C_j$ for all $j \in \mathcal{J}_\mu$. We denote by $\mathcal{A}^{(k)}$ the set of all admissible measures, and we set $\mathcal{A}^{(2)}_{\mathrm{bal}} = \mathcal{A}^{(2)} \cap \mathcal{P}_{\mathrm{bal}}(S^2)$.*

Those definitions are general, but we will mostly work with $k = 1$ or $k = 2$. When $k = 1$, (pre-)admissible measures vanish on B. When $k = 2$, (pre-)admissible measures are supported by pairs of points that belong to the same irreducible class. Properties of $\mathcal{A}^{(k)}$ are further explored in appendix A.2. In particular, setting

$$\mathcal{D}^{(2)} = \{\mu \in \mathcal{M}(S^2) \mid \forall (x, y) \in S^2,\ p(x, y) = 0 \Rightarrow \mu(x, y) = 0\}, \tag{1.8}$$

Proposition A.7 shows that $\mathcal{A}^{(2)}_{\mathrm{bal}} \cap \mathcal{D}^{(2)}$ consists of precisely those measures which can be asymptotically approximated by $L_n^{(2)}$ with positive probability, and hence which are relevant for the corresponding LDP; see Remark A.3.

From a technical point of view, the central result of this paper is the following theorem.

Theorem 1.3 (Weak LDP for the pair empirical measures)**.** *The sequence $(L_n^{(2)})$ satisfies the weak LDP in $\mathcal{P}(S^2)$ with rate function $I^{(2)}$ given by*

$$I^{(2)}(\mu) = \begin{cases} (\Lambda^{(2)})^*(\mu) = J^{(2)}(\mu) = R^{(2)}(\mu) & \text{if } \mu \in \mathcal{A}^{(2)}_{\mathrm{bal}}, \\ \infty & \text{otherwise}, \end{cases} \tag{1.9}$$

where the functions $\Lambda^{(2)}$, $J^{(2)}$, and $R^{(2)}$ are defined by equations (3.1), (3.2) and (3.5) respectively and $\cdot^*$ denotes the convex conjugate as in Definition B.1.

This theorem embodies two statements: first, $(L_n^{(2)})$ satisfies the weak LDP, and second, the rate function of this weak LDP can be determined explicitly.

The first point is developed in Section 2, where a self-contained proof of the weak LDP for $(L_n^{(2)})$ is provided (Theorem 2.5) through the use of RL functions (Section 2.2). The second point, the identification of the rate function, is investigated in details in Section 3. The expressions in the first line of (1.9) are standard in large deviations theory:

- The function $(\Lambda^{(2)})^*$ is the convex conjugate of the scaled cumulant generating function (SCGF) of the sequence $(L_n^{(2)})$, defined in (3.1).[9] Since they provide information on the exponential scale, SCGFs are common objects in large deviations theory. Their convex conjugates are often identified as rate functions of LDPs, for instance via the Gärtner-Ellis Theorem (see V.2 in [12] or Chapter 2.3 of [11]) or Varadhan's Lemma (see III.3 in [12], or Chapters 4.3 and 4.5 of [11]). SCGFs are also relevant in the specific context of Markov chains, as in Theorem 6.3.8 of [11] and in Chapters 2 and 3 of [10].

[6] Note that (1.7) is *not* the Doeblin decomposition. The family $(C_j)_{j\in\mathcal{J}}$ contains both transient and essential classes. In particular, the sets C_j are not the absorbing classes of the Markov chain, and may not be absorbing at all.

[7] Or equivalently, if all $y \in C_{j'}$ are reachable from all $x \in C_j$.

[8] Or equivalently, if all $y \in C_j$ are reachable from β.

[9] Λ is also often called pressure, although this terminology is not physically accurate in general.

- The function $J^{(2)}$ is the Donsker-Varadhan entropy (DV entropy), defined in (3.2) and first introduced by Donsker and Varadhan in [16].[10] Since then, it has been presented as the standard rate function of LDPs in Markovian setups (see for instance Theorem IV.7 of [12], or Chapter 6.5 of [11]). In the literature, it is common to find that the DV entropy satisfies the LDP lower bound, under very mild assumptions; see Chapter 3 of [10]. See also Chapter 13.2 of [28].

- The function $R^{(2)}$ is defined in (3.5), and is expressed in terms of relative entropy. Going back to Sanov's theorem, relative entropy is always expected to play a role in level-2 large deviations. This function appears for instance in the LDPs of Theorem IV.3 of [12], of Chapter 6.5.2 of [11], and of Chapter 13.2 of [28].

We further notice that the functions $(\Lambda^{(2)})^*$, $J^{(2)}$, and $R^{(2)}$ do not depend on β, so the rate function $I^{(2)}$ only depends on β through the set $\mathcal{A}^{(2)}_{\text{bal}}$. The second line of (1.9) is also interesting by itself, and is specific to the reducible case. It says that outside of $\mathcal{A}^{(2)}_{\text{bal}}$, the rate function is infinite. This is only a sufficient condition and $I^{(2)}$ can be infinite inside $\mathcal{A}^{(2)}$; see Remark 3.2. A brief analysis of the geometry of $\mathcal{A}^{(2)}_{\text{bal}}$ (see Appendix A.2) reveals that this property prevents the rate function from being convex in many cases, with the consequences discussed in the introduction.

The consequences of Theorem 1.3 are the level-2 and level-3 weak LDPs:

Theorem 1.4 (Level-2 weak LDP)**.** *The sequence $(L_n^{(1)})_{n\geqslant 1}$ satisfies the weak LDP in $\mathcal{P}(S)$ with rate function $I^{(1)}$ given by*

$$I^{(1)}(\mu) = \begin{cases} (\Lambda^{(1)})^*(\mu) = J^{(1)}(\mu) = R^{(1)}(\mu) & \text{if } \mu \in \mathcal{A}^{(1)}, \\ \infty & \text{otherwise}, \end{cases} \tag{1.10}$$

where the functions $\Lambda^{(1)}$, $J^{(1)}$, and $R^{(1)}$ are defined by equations (4.2), (4.4), and (4.5), respectively and $\cdot^*$ denotes the convex conjugate as in Definition B.1.

Theorem 1.5 (Level-3 weak LDP)**.** *Let $\mathcal{P}(S^{\mathbb{N}})$ be equipped with the weak topology; see Section 5.2. The sequence $(L_n^{(\infty)})_{n\geqslant 1}$ satisfies the weak LDP in $\mathcal{P}(S^{\mathbb{N}})$, with rate function $I^{(\infty)}$ given by*

$$I^{(\infty)}(\mu) = \begin{cases} (\Lambda^{(\infty)})^*(\mu) = J^{(\infty)}(\mu) = R^{(\infty)}(\mu) & \text{if } \mu \in \mathcal{A}^{(\infty)}_{\text{bal}}, \\ \infty & \text{otherwise}, \end{cases} \tag{1.11}$$

where the functions $(\Lambda^{(\infty)})^*$, $J^{(\infty)}$, and $R^{(\infty)}$ are defined by equations (5.18), (5.19), and (5.20) respectively and the set $\mathcal{A}^{(\infty)}_{\text{bal}}$ is introduced in Definition 5.4. Moreover, assuming that $\mu \in \mathcal{A}^{(\infty)}_{\text{bal}}$ and $H(\mu^{(1)}|\beta) < \infty$, we have the additional expression

$$I^{(\infty)}(\mu) = \lim_{k\to\infty} \frac{1}{k} H(\mu^{(k)}|\mathbb{P}_k), \tag{1.12}$$

where $\mu^{(k)}$ is a marginal of μ that will be defined in Section 5.2 and $\mathbb{P}_k$ is the marginal of $\mathbb{P}$ on the first k coordinates; see (2.2). The relative entropy $H(\cdot|\cdot)$ is defined in (3.4).

The same comments as for Theorem 1.3 can be made: the sets $\mathcal{A}^{(1)}$ and $\mathcal{A}^{(\infty)}_{\text{bal}}$ which appear in (1.10) and (1.11) may cause the rate functions to be nonconvex for reducible Markov chains.

The path we follow to derive Theorems 1.4 and 1.5 from Theorem 1.3 will come as no surprise to the experienced reader:

- The weak LDP for $(L_n^{(1)})$ is obtained using the contraction principle (see III.5 in [12] or Theorem 4.2.1 of [11]), which has to be adapted, since we only have the weak LDP and the rate function may not be good.

- The weak LDP for $(L_n^{(\infty)})$ is obtained using the Dawson-Gärtner Theorem (see for instance Theorem 4.6.1 of [11]). We follow the tracks of Sections 6.5.2, 6.5.3 of [11] or Section 4.4 of [13], where the level-3 LDP for Markov chains is proved under the uniformity assumption **(U)**. Once again, in the absence of this assumption, some technical adaptations are required.

[10] The function $J^{(2)}$ is actually the DV entropy for the pair empirical measures; the traditional DV entropy is the function $J^{(1)}$ defined in (4.4).

We briefly comment on the fact that if the state space S is finite, then all LDPs are full and all rate functions are good. In addition, in this case many technical aspects of the proofs can be considerably simplified; see Appendix C for a discussion. However, the main conceptual modifications of the subadditive method that we develop to handle reducible Markov chains, namely the *slicing* and *stitching* procedure below, remain necessary.

2 Weak LDP by subadditivity

In this section, we prove the first part of Theorem 1.3, *i.e.* that the sequence (L_n) defined in (1.5) satisfies the weak LDP. We use a subadditive method.

2.1 Notations

Trajectories of $(X_n)_{n\geqslant 1}$ are sequences of elements of S. In the following, we will often refer to the elements of S as *letters*, and to finite sequences of elements of S, *i.e.* finite pieces of trajectories, as *words*. The set of all words is denoted by S_{fin}, the set of words of length t by S^t and the length of a finite word u by $|u|$. We will use the symbol e for the empty word (which has length 0). We extend the definition of p to finite words by setting

$$p(u) = p(u_1, u_2) \times \ldots \times p(u_{|u|-1}, u_{|u|}), \qquad u \in S_{\mathrm{fin}},\ |u| \geqslant 2, \tag{2.1}$$

and $p(u) = 1$ if $|u| \leqslant 1$. The function p, seen as a function over S_{fin}, satisfies the following property: for $u, v \in S_{\mathrm{fin}}$, $p(uv) = p(u)p(v)p(u_{|u|}, v_1)$, where the quantity $p(u_{|u|}, v_1)$ depends only on the last letter of u and the first letter of v. The law $\mathbb{P}$ induces a probability measure $\mathbb{P}_t$ on S^t by

$$\mathbb{P}_t(u) = \mathbb{P}(X_1 = u_1, \ldots, X_t = u_t) = \beta(u_1)p(u). \tag{2.2}$$

An empirical measure can be associated with each word. We define

$$M[u] = \sum_{i=1}^{t} \delta_{(u_i, u_{i+1})} \in \mathcal{M}(S^2), \qquad L[u] = \frac{1}{t} M[u] \in \mathcal{P}(S^2), \qquad u \in S^{t+1}. \tag{2.3}$$

If u has length 1 or 0, we set $L[u]$ to be the null measure. We will extensively use the fact that $|M[u]|_{\mathrm{TV}} = |u| - 1$ if $u \neq e$.[11] Given $t \in \mathbb{N}$, $S_1 \subseteq S$, $\mu \in \mathcal{P}(S^2)$ and $\rho > 0$, we define the following sets of words:

- $S_{\mathrm{fin},+}$ is the set of words u such that $p(u) > 0$.
- $W_t(\mu, \rho)$ is the set of words $u \in S^{t+1}$ such that $L[u] \in \mathcal{B}(\mu, \rho)$ and $\mathbb{P}_{t+1}(u) > 0$,
- $W_{t,S_1}(\mu, \rho)$ is the intersection of $W_t(\mu, \rho)$ and S_1^{t+1}.

Observe also that S_1^t is exactly the set of words $u \in S^t$ such that $L[u] \in \mathcal{P}(S_1^2)$.

Remark 2.1. One could apply the argument of the current section to $L_n^{(1)}$ instead of $L_n^{(2)}$, and thus obtain the level-2 weak LDP in a straightforward manner. However, from there, identifying the rate function and deriving the level-3 weak LDP would be more complicated, and in fact the pair empirical measures $(L_n^{(2)})$ would come into play in the process anyway. We choose to study the weak LDP for $(L_n^{(2)})$ first, as it carries enough information on both the level-2 and level-3 weak LDPs, as made clear in Sections 4 and 5.

2.2 Ruelle-Lanford functions and weak LDP for the pair empirical measures

Definition 2.2 (Ruelle-Lanford functions)**.** *For all Borel sets $A \subseteq \mathcal{P}(S^2)$, let*

$$\underline{s}(A) = \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(L_n \in A),$$
$$\overline{s}(A) = \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(L_n \in A).$$

[11] We also have $|M[e]|_{\mathrm{TV}} = 0 = |e|$.

For all $\nu \in \mathcal{P}(S^2)$, we define $\underline{s}(\nu)$ and $\overline{s}(\nu)$ as the monotone limits $\underline{s}(\nu) = \lim_{\delta\to 0} \underline{s}(\mathcal{B}(\nu,\delta))$ and $\overline{s}(\nu) = \lim_{\delta\to 0} \overline{s}(\mathcal{B}(\nu,\delta))$. We say that (L_n) has a Ruelle-Lanford (RL) function if $\underline{s} = \overline{s}$ on $\mathcal{P}(S^2)$, in which case the function $\underline{s} = \overline{s}$ is called the RL function of (L_n).

The notion of RL function was first introduced by Ruelle in [29] as an expression for the entropy. Later, in [23], Lanford revealed how efficient it can be in the theory of large deviations. This gave rise to the subadditive methods, originally developed for independent and identically distributed random variables in [23], [2] and [1]. It has later been subjected to some adaptations to derive results for a larger class of random variables; see the introduction for some examples of use of subadditivity for large deviations for Markov chains. See [24] for a more detailed historical account.

The interest in considering RL functions in our case resides in the following standard result; for a proof, see, for example, Theorem 4.1.1 in [11], Proposition 3.5 of [26], Theorem 3.1 of [24] and, more recently, Section 3.2 of [6].

Lemma 2.3. *Suppose that (L_n) has a RL function s. Then, s is upper semicontinuous and (L_n) satisfies the weak LDP with rate function $-s$.*

Thanks to this lemma, the goal of this section is now simply to prove that (L_n) has a RL function. The proof of the existence of the RL function and the derivation of its basic properties are achieved via the following technical proposition, whose proof will be covered by the next sections and concluded in Section 2.5.1. It states a key *decoupled inequality*, which is the core of our subadditive method.

In the following proposition, for the purpose of the sole existence of the RL function, one can consider $\mu_1 = \mu_2$[12] and $S_2 = S$ (and thus overlook the existence of S_1).

Proposition 2.4. *Let $0 < \delta \leqslant 1/12$, and let, for any integers n, t,*

$$N = N(n,t) = \left\lceil \frac{t+1}{n(1-4\delta)} \right\rceil, \qquad N_1 = \left\lfloor \frac{N}{2} \right\rfloor, \qquad N_2 = N - N_1. \tag{2.4}$$

Let $\mu_1, \mu_2 \in \mathcal{A}^{(2)}$ such that $\mu := \frac{1}{2}\mu_1 + \frac{1}{2}\mu_2 \in \mathcal{A}^{(2)}$. Then, there exist an integer n_0, a sequence of integers (t_n), and a finite set $S_1 \subseteq S$ such that for all $n \geqslant n_0$, all $t \geqslant t_n$, and all set $S_2 \supseteq S_1$,[13]

$$\mathbb{P}_{t+1}(W_{t,S_2}(\mu, 20\delta)) \geqslant C_{n,t}\mathbb{P}_{n+1}(W_{n,S_2}(\mu_1,\delta))^{N_1}\mathbb{P}_{n+1}(W_{n,S_2}(\mu_2,\delta))^{N_2}, \tag{2.5}$$

for some constants $C_{n,t}$ satisfying

$$\lim_{n\to\infty}\lim_{t\to\infty}\frac{1}{t}\log C_{n,t} = 0. \tag{2.6}$$

Taking Proposition 2.4 for granted, we can derive the existence of the RL function of (L_n) and the weak LDP for (L_n).

Theorem 2.5. *The sequence (L_n) has a RL function $s := \underline{s} = \overline{s}$ and satisfies the weak LDP with rate function $I^{(2)} := -s$.*

Proof. Let $\mu \in \mathcal{A}^{(2)}$. We set $\mu_1 = \mu_2 = \mu$ and $S_2 = S$ in Proposition 2.4. Then, (2.5) yields

$$\mathbb{P}(L_t \in \mathcal{B}(\mu, 20\delta)) \geqslant C_{n,t}\mathbb{P}(L_n \in \mathcal{B}(\mu,\delta))^N.$$

By taking the logarithm on both sides, dividing by t and letting $t \to \infty$, we obtain

$$\underline{s}(\mathcal{B}(\mu, 20\delta)) \geqslant \lim_{t\to\infty}\frac{1}{t}\log C_{n,t} + \frac{1}{n(1-4\delta)}\log \mathbb{P}_{n+1}(L_n \in \mathcal{B}(\mu,\delta)).$$

By taking the limit superior as $n \to \infty$ and using (2.6), we have

$$\underline{s}(\mathcal{B}(\mu, 20\delta)) \geqslant \frac{1}{(1-4\delta)}\overline{s}(\mathcal{B}(\mu,\delta)).$$

Further taking the limit as $\delta \to 0$ yields $\underline{s}(\mu) \geqslant \overline{s}(\mu)$. Since, obviously, we also have $\underline{s}(\mu) \leqslant \overline{s}(\mu)$, the sequence (L_n) has a RL function s and the weak LDP with the rate function $-s$ follows from Lemma 2.3. □

[12] Taking two different measures in the construction is a common technique in subadditive methods, as it can often be used to later derive the convexity of the rate function; see [6]. Even though our rate function is not convex, this technique still delivers useful properties; see Property 1 of Proposition 2.7.

[13] The radius 20δ in (2.5) is far from being optimal, but any multiple of δ acceptable for the purpose of the subadditive method; see how the factor 20 eventually does not matter in the proof of Theorem 2.5.

By Theorem 2.5, the function $I^{(2)}$ is the unique rate function[14] associated with the weak LDP for (L_n). Theorem 2.5 does not provide much information to compute $I^{(2)}$ in practice. However, we can access a few useful properties of $I^{(2)}$ as direct consequences of Proposition 2.4, and of Proposition 2.6, which we state here and will prove in Section 2.5.2. Inequality (2.7) can be seen as a reciprocal inequality of (2.5) in the case of disjoint supports and will indeed be obtained by reversing the construction used to derive Proposition 2.4.

Proposition 2.6. *Let $\lambda_1 \in (0,1)$ and let $\lambda_2 = 1-\lambda_1$. Let $0 < \delta < \min(\lambda_1/12, \lambda_2/12)$ and let, for any integer n,*

$$t_i = t_i(n) = \lfloor n(\lambda_i - 2\delta)\rfloor - 1, \qquad i \in \{1,2\}.$$

Let $\mu_1, \mu_2 \in \mathcal{A}^{(2)}$ be such that $\mu := \lambda_1\mu_1 + \lambda_2\mu_2 \in \mathcal{A}^{(2)}$ and $\mathcal{J}_{\mu_1} \cap \mathcal{J}_{\mu_2} = \emptyset$. Then, there exists an integer n_0 such that for all $n \geqslant n_0$,[15]

$$\mathbb{P}_{t_1+1}\Big(W_{t_1}\Big(\mu_1, \frac{125}{\lambda_1^2}\delta\Big)\Big)\mathbb{P}_{t_2+1}\Big(W_{t_2}\Big(\mu_2, \frac{125}{\lambda_2^2}\delta\Big)\Big) \geqslant \widetilde{C}_n \mathbb{P}_{n+1}(W_n(\mu,\delta)), \tag{2.7}$$

for some constants $\widetilde{C}_n$ satisfying

$$\lim_{n\to\infty} \frac{1}{n} \log \widetilde{C}_n = 0. \tag{2.8}$$

Taking Proposition 2.4 and Prop 2.6 for granted, we deduce the following properties of the rate function $I^{(2)}$.

Proposition 2.7. *The rate function $I^{(2)}$ of Theorem 2.5 satisfies the following properties.*

1. *Let $\mu \in \mathcal{A}^{(2)}$ and let $\mathcal{F} = \{\nu \in \mathcal{A}^{(2)} \mid \mathcal{J}_\nu \subseteq \mathcal{J}_\mu\}$. Then, $I^{(2)}|_{\mathcal{F}}$ is convex. In particular, we have the following. Let $\lambda_1 \in [0,1]$ and let $\lambda_2 = (1-\lambda_1)$. Let $\mu_1, \mu_2 \in \mathcal{A}^{(2)}$ be such that $\mu := \lambda_1\mu_1 + \lambda_2\mu_2 \in \mathcal{A}^{(2)}$. Then,*

$$I^{(2)}(\mu) \leqslant \lambda_1 I^{(2)}(\mu_1) + \lambda_2 I^{(2)}(\mu_2). \tag{2.9}$$

2. *Let $\lambda_1 \in [0,1]$ and let $\lambda_2 = (1-\lambda_1)$. Let $\mu_1, \mu_2 \in \mathcal{A}^{(2)}$ be such that $\mu := \lambda_1\mu_1 + \lambda_2\mu_2 \in \mathcal{A}^{(2)}$ and $\mathcal{J}_{\mu_1} \cap \mathcal{J}_{\mu_1} = \emptyset$. Then,*

$$I^{(2)}(\mu) = \lambda_1 I^{(2)}(\mu_1) + \lambda_2 I^{(2)}(\mu_2).$$

3. *Let $\mu \in \mathcal{P}(S^2) \setminus \mathcal{A}^{(2)}_{\mathrm{bal}}$. Then,*

$$I^{(2)}(\mu) = \infty.$$

4. *For all Borel set $A \subseteq \mathcal{P}(S^2)$, let $\underline{s}_\infty(A)$ be the supremum of $\underline{s}(A \cap \mathcal{P}(K^2))$ over all finite $K \subseteq S$. Let $\mu \in \mathcal{P}(S^2)$. Then, the function $r \mapsto \underline{s}_\infty(\mathcal{B}(\mu,r))$ is nondecreasing and*

$$I^{(2)}(\mu) = -\lim_{r\to 0} \underline{s}_\infty(\mathcal{B}(\mu,r)).$$

One should read these properties with extra care. Property 1 does not say that $I^{(2)}$ is convex; see Example D.1. The affinity stated in Property 2 does not hold between every two measures; any two-states matrix-irreducible Markov chain is already a convincing counter-example. Also notice that Property 3 only states a sufficient condition for $I^{(2)}$ to be infinite; see Remark 3.2 and Example D.4.

Proof of Proposition 2.7. 1. We begin with Property 1. Let $\nu_1, \nu_2 \in \mathcal{F}$ and fix $\nu = \frac{1}{2}(\nu_1+\nu_2)$. We first prove that

$$I^{(2)}(\nu) \leqslant \frac{1}{2}(I^{(2)}(\nu_1) + I^{(2)}(\nu_2)). \tag{2.10}$$

To do so, we use Proposition 2.4. Considering (2.5) for ν, ν_1, ν_2 with $S_2 = S$, we have

$$\mathbb{P}(L_t \in \mathcal{B}(\nu, 20\delta)) \geqslant C_{n,t} \log \mathbb{P}(L_n \in \mathcal{B}(\nu_1,\delta))^{N_1} \mathbb{P}(L_n \in \mathcal{B}(\nu_2,\delta))^{N_2}.$$

[14] In particular, $I^{(2)}$ is lower semicontinuous.

[15] As for the factor 20 in (2.5), the factor $125/\lambda_i^2$ surely is not optimal, but this does not matter for the sequel; see the proof of Property 2 of Proposition 2.7.

By taking the logarithm of both sides, dividing by t and letting $t \to \infty$, we get

$$\underline{s}(\mathcal{B}(\nu, 20\delta)) \geqslant \lim_{t\to\infty} \frac{1}{t} \log C_{n,t} + \frac{1}{2n(1-4\delta)} (\log \mathbb{P}(L_n \in \mathcal{B}(\nu_1, \delta)) + \log \mathbb{P}(L_n \in \mathcal{B}(\nu_2, \delta))).$$

Taking the limit inferior of this expression as $n \to \infty$ yields

$$\underline{s}(\mathcal{B}(\nu, 20\delta)) \geqslant \frac{1}{2(1-4\delta)} (\underline{s}(\mathcal{B}(\nu_1, \delta)) + \underline{s}(\mathcal{B}(\nu_2, \delta))).$$

Since $I^{(2)} = -\underline{s}$, finally taking the limit as $\delta \to 0$ yields (2.10). Let $\mu_1, \mu_2 \in \mathcal{F}$. Then, $\lambda\mu_1 + (1-\lambda)\mu_2 \in \mathcal{F}$ for all $\lambda \in [0, 1]$. Using a bisection argument based on (2.10), together with the lower semicontinuity of I, we obtain, for all $\lambda \in [0, 1]$,

$$I(\lambda\mu_1 + (1-\lambda)\mu_2) \leqslant \lambda I(\mu_1) + (1-\lambda)I(\mu_2),$$

which shows that $I^{(2)}|_{\mathcal{F}}$ is convex. In particular, we have the following. For all $\lambda_1 \in [0, 1]$ and $\lambda_2 = (1-\lambda_1)$, $\mu_1, \mu_2 \in \mathcal{A}^{(2)}$ such that $\mu := \lambda_1\mu_1 + \lambda_2\mu_2 \in \mathcal{A}^{(2)}$, both μ_1 and μ_2 are absolutely continuous with respect to μ, hence they belong to $\mathcal{F}$ and satisfy (2.9).

2. We now prove Property 2. Let μ_1, μ_2 and λ_1, λ_2 be as in the statement of Property 2. By Property 1, we only have to prove that

$$\lambda_1 I^{(2)}(\mu_1) + \lambda_2 I^{(2)}(\mu_2) \leqslant I^{(2)}(\mu). \tag{2.11}$$

To do so, we use Proposition 2.6. For convenience, we set $\delta_i = 125\delta/\lambda_i^2$ for $i \in \{1, 2\}$. Taking the logarithm on both sides in (2.7), dividing by n, and then letting $n \to \infty$ yields

$$(\lambda_1 - 2\delta)\overline{s}(\mathcal{B}(\mu_1, \delta_1)) + (\lambda_2 - 2\delta)\overline{s}(\mathcal{B}(\mu_2, \delta_2)) \geqslant 0 + \underline{s}(\mathcal{B}(\mu, \delta)).$$

Since $I^{(2)} = -s$, taking the limit as $\delta \to 0$ yields (2.11).

3. Property 3 is equivalent to $s(\mu) = -\infty$ for all $\mu \in \mathcal{P}(S^2) \setminus \mathcal{A}^{(2)}_{\text{bal}}$. Let $\mu \in \mathcal{P}(S^2) \setminus \mathcal{A}^{(2)}_{\text{bal}}$. by contrapositive of ($1 \Rightarrow 2$) in Proposition A.7, there exists $\ell \in \mathbb{N}$ and $\delta > 0$ such that all words w satisfying $|w| \geqslant \ell$ and $L[w] \in \mathcal{B}(\mu, \delta)$ must satisfy $\mathbb{P}_{|w|}(w) = 0$. Hence, for all $n \geqslant \ell$, we have $\mathbb{P}(L_n \in \mathcal{B}(\mu, \delta)) = 0$. Therefore, $\overline{s}(\mathcal{B}(\mu, \delta)) = -\infty$, implying $s(\mu) = -\infty$.

4. We now prove Property 4. Let $\mu \in \mathcal{P}(S^2)$. The monotonicity of $r \mapsto \underline{s}_\infty(\mathcal{B}(\mu, r))$ is immediate by definition, hence Property 4 reformulates as

$$\lim_{r\to 0} \underline{s}_\infty(\mathcal{B}(\mu, r)) = \lim_{r\to 0} \underline{s}(\mathcal{B}(\mu, r)).$$

The bound $\underline{s}_\infty(A) \leqslant \underline{s}(A)$ for all Borel sets $A \subseteq \mathcal{P}(S^2)$ is immediate, thus we have to show that

$$\lim_{r\to 0} \underline{s}_\infty(\mathcal{B}(\mu, r)) \geqslant \lim_{r\to 0} \underline{s}(\mathcal{B}(\mu, r)). \tag{2.12}$$

When $\mu \notin \mathcal{A}^{(2)}_{\text{bal}}$, there is nothing to prove because the right-hand side limit is $s(\mu) = -\infty$ by Property 3. Assume that $\mu \in \mathcal{A}^{(2)}_{\text{bal}}$. Let $\delta > 0$, and take S_1 as given by Proposition 2.4. Fix n and t as in Proposition 2.4. There exists a finite set $S_2 \subseteq S$, possibly depending on n, such that $S_1 \subseteq S_2$ and

$$\mathbb{P}_{n+1}(W_{n,S_2}(\mu, \delta)) \geqslant \frac{1}{2}\mathbb{P}_{n+1}(W_n(\mu, \delta)).$$

Thus, by (2.5),

$$\begin{aligned}\mathbb{P}_{t+1}(W_{t,S_2}(\mu, 20\delta)) &\geqslant C_{n,t}\mathbb{P}_{n+1}\left(W_{n,S_2}(\mu, \delta)\right)^N \\ &\geqslant \frac{C_{n,t}}{2^N}\mathbb{P}_{n+1}(W_n(\mu, \delta))^N.\end{aligned}$$

Taking the logarithm of both sides yields, dividing by t and taking the limit inferior as $t \to \infty$ yields

$$\begin{aligned}\underline{s}_\infty(\mathcal{B}(\mu, 20\delta)) &\geqslant \underline{s}(\mathcal{B}(\mu, 20\delta) \cap \mathcal{P}(S_2^2)) \\ &\geqslant \lim_{t\to\infty} \frac{1}{t} \log C_{n,t} - \frac{\log 2}{n(1-4\delta)} + \frac{1}{n(1-4\delta)} \log \mathbb{P}_n(W_n(\mu, \delta)).\end{aligned}$$

By further taking the limit inferior as $n \to \infty$ (both ends of the chain of inequalities are independent of S_2, hence no problem ensues from the dependence of S_2 on n), we get

$$\underline{s}_\infty(\mathcal{B}(\mu, 20\delta)) \geqslant \underline{s}(\mathcal{B}(\mu, \delta)).$$

We obtain (2.12) by taking the limit as $\delta \to 0$.

□

From now on, our goal is to prove Propositions 2.4 and 2.6. This will be achieved in Sections 2.5.1 and 2.5.2 respectively.

Remark 2.8. Note that, while Proposition 2.4 is used in both the proofs the existence and properties of the RL function (Theorem 2.5 and Proposition 2.7), Proposition 2.6 is only used to derive one property of the RL function (Property 2 in Proposition 2.7). The reader solely interested in the proof of the weak LDP can skip Section 2.5.2.

2.3 Two preliminary examples

Section 2.2 made clear that obtaining an inequality of the form (2.5) is the pivotal argument of our subadditive method. Before embarking in the proof of Proposition 2.4, let us illustrate our strategy on simple examples.

Example 2.9. Let S be finite, and assume that (X_n) is matrix-irreducible and aperiodic. In other words, we assume that there exists $\tau > 0$ such that, for all $x, y \in S$, there exists $\xi_{x,y} \in S^\tau$ satisfying $p(x\xi_{x,y}y) > 0$. Let $\mu \in \mathcal{P}(S^2)$ and $\delta > 0$. Then, for all $u, v \in S_{\text{fin}}$, there exists $\xi \in S^\tau$ such that[16]

$$\mathbb{P}_{|u|+|v|+\tau}(u\xi v) \geqslant c\mathbb{P}_{|u|}(u)\mathbb{P}_{|v|}(v), \qquad c := \inf_{(x,y)\in S^2} p(x\xi_{x,y}y) > 0. \tag{2.13}$$

Notice that if $u, v \in W_n(\mu, \delta)$ for some $n \geqslant (2\tau + 1)/2\delta$, then $u\xi v \in W_{2n+\tau}(\mu, 2\delta)$. Building on the above inequality, one can derive that for any family of words $(u^i) \in (S^n)^N$, there exists a family of words $(\xi^i) \in (S^\tau)^{N-1}$ such that

$$\mathbb{P}_{Nn+(N-1)\tau}(u^1\xi^1u^2\xi^2\dots u^N) \geqslant c^{N-1}\prod_{i=1}^{N}\mathbb{P}_n(u^i).$$

By summing this inequality over $W_n(\mu, \delta)^N$, one can find

$$\mathbb{P}_{Nn+(N-1)\tau}(W_{Nn+(N-1)\tau}(\mu, 2\delta)) \geqslant C_{n,N}\mathbb{P}_n(W_n(\mu, \delta))^N,$$

where $C_{n,N}$ satisfies

$$\lim_{n\to\infty}\lim_{N\to\infty}\frac{1}{Nn + (N-1)\tau}\log C_{n,N} = 0.$$

This inequality resembles (2.5) and is sufficient for using the machinery of Setion 2.2. Therefore, (L_n) satisfies the weak LDP.

This example is well-known in the literature. A general construction for *decoupled systems*, which include the case of Example 2.9, is provided in [6]. The specific case of finite matrix-irreducible (but not necessarily aperiodic) Markov chains is mentioned in Example 2.20 as an application of the general construction. See also Examples 3 and 4 of [26].

The central inequality in the construction of Example 2.9 is (2.13). However, it is clear that the assumption of matrix-irreducibility was crucial in deriving such an inequality. In absence of irreducibility, how can we obtain an inequality that resembles (2.13) enough to enable Proposition 2.4? The purpose of the rest of the section is to adapt this construction to reducible setups. As an introduction, let us illustrate the method with another simple, reducible, example.

Example 2.10. Let $S = \{1, 2, 3, 4, 5, 6, 7\}$ and consider a Markov chain represented on Figure 1, the initial measure β being the uniform law on S. The irreducible classes of the Markov chain are $(\{1\}, \{2\}, \{3\}, \{5, 6\}, \{7\})$ and $B = \{4\}$. The chain is neither matrix-irreducible nor φ-irreducible

[16] Note that we used $\beta(v_1) \leqslant 1$.

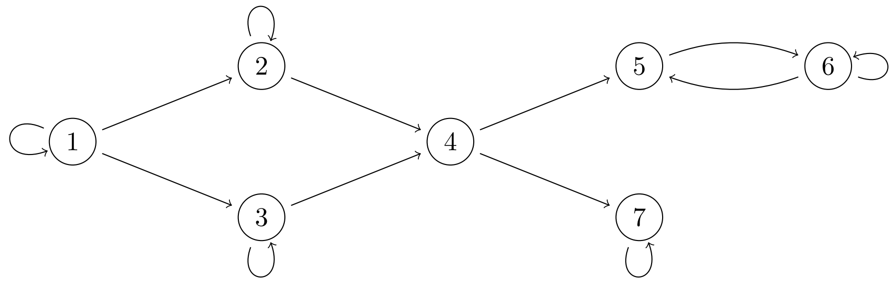


Figure 1: The 7-letter Markov chain of Example 2.10. The transition probabilities of each arrow do not matter for this example, as long as they are positive on each edge of the graph and zero otherwise.

and thus, the existence of linking words ξ as in Example (2.9) is not guaranteed. The two challenges brought by reducibility are the following. First, there is a notion of irreversibility. Since it is impossible to 'come back' to 1, the words $u = 1112$ and $v = 1245$ cannot possibly satisfy (2.13). Second, some letters are incompatible with each other. Since neither $2 \rightsquigarrow 3$ nor $3 \rightsquigarrow 2$, no word of positive probability can contain simultaneously the letters 2 and 3. Thus we will not be able to derive an inequality of the form (2.13) with the words $u = 2222$ and $v = 3333$.

Let us consider a measure $\mu \in \mathcal{A}^{(2)}_{\mathrm{bal}}$ and $\delta > 0$. For instance, we take

$$\mu = \frac{1}{4}\delta_{1,1} + \frac{1}{8}\delta_{2,2} + \frac{1}{4}\delta_{5,6} + \frac{1}{4}\delta_{6,5} + \frac{1}{8}\delta_{6,6} \in \mathcal{A}^{(2)}_{\mathrm{bal}}.$$

It is immediate that μ is balanced, preadmissible with $(C_j)_{j\in\mathcal{J}_\mu} = (\{1\},\{2\},\{5,6\})$, and admissible because $\beta \rightsquigarrow \{1\} \rightsquigarrow \{2\} \rightsquigarrow \{5,6\}$.

Let $u = 1112224565665665$ and $v = 1111334565656566$. Both words have their empirical measure close to μ. Let us try to find a word w such that $L[w]$ is almost as close to μ, and $\mathbb{P}_{|w|}(w)$ is comparable to the product $\mathbb{P}_{|u|}(u)\mathbb{P}_{|v|}(v)$. The word w will play the role of $u\xi v$ in (2.13). The key idea is to *slice* words u and v into subwords that are easy to reassemble: we write

$$u = (111)(222)(4)(565665665), \qquad v = (1111)(334)(565656566).$$

We can drop the subwords 4 of u and 334 of v, as they consist of letters not appearing in the support of μ (in addition, the letter 3 in v is incompatible with the letter 2 of u. Also notice that neither u nor v has the letter 7, as it is incompatible with their letters 5 and 6). We then *stitch* together the remaining subwords 111, 222, 565665665, 1111 and 565656566 while respecting the total order on $(\{1\},\{2\},\{5,6\})$, and adding transition letters when needed. We set[17]

$$w = (111)(1111)(222)(4)(565665665)(6)(565656566).$$

The transition letters 4 and 6 were added in the word w because $p(2,5) = p(5,5) = 0$. The measure $L[w]$ is reasonably close to μ. We also have

$$\mathbb{P}_{|w|}(w) = kp(111)p(1111)p(222)p(565665665)p(565656566),$$

where k is a product of a few missing factors of the form $p(x,y)$ and $\beta(1)$. Since

$$\mathbb{P}_{|u|}(u) \leqslant p(111)p(222)p(565665665), \qquad \mathbb{P}_{|v|}(v) \leqslant p(1111)p(565656566),$$

we find $\mathbb{P}_{|w|}(w) \geqslant k\mathbb{P}_{|u|}(u)\mathbb{P}_{|v|}(v)$. This maneuver can be reproduced for any $u, v \in S_{\mathrm{fin}}$ whose empirical measures are close to μ. By bounding k away from 0, we obtain an inequality that is similar to (2.13).

In Example 2.10, the construction of a longer word w satisfying the desired properties resulted from two operations. First, the operation of *slicing* the words u and v into smaller subwords, and second the operation of rearranging and *stitching* those smaller subwords together. These operations were carried out in such a way that $p(w)$ is close to $p(u)p(v)$ and $L[w]$ is close to $\frac{1}{2}L[u] + \frac{1}{2}L[v]$. The next section provides a general construction for these slicing and stitching operations.

[17] One can notice that w is shorter than $|u| + |v|$, in contrast with more standard subadditive arguments.

2.4 The slicing map and the stitching map

We begin by introducing a *slicing map* and a *stitching map*, whose properties will be used to prove Proposition 2.4 in Section 2.5. The slicing map is used to turn a word u into a collection of subwords u^j, which contains most of the letters of u belonging to the class C_j. The stitching map is used to turn a collection of words into a longer word containing all of them as subwords.

In this section, K is a finite subset of S and we denote by $\mathcal{J}_K$ the set $\{j \in \mathcal{J} \mid C_j \cap K \neq \emptyset\}$. Also in this section, we let $\mathcal{J}_0 \subseteq \mathcal{J}_K$ be such that the order $\rightsquigarrow$ is total on $(C_j)_{j\in\mathcal{J}_0}$. Notice that $\mathcal{J}_K$ and $\mathcal{J}_0$ are necessarily finite. We also assume that they are not empty. For convenience, we rename the elements of $\mathcal{J}_0$ so that

$$\begin{cases} \mathcal{J}_0 = \{1, \dots, r\}, \\ C_j \rightsquigarrow C_{j+1} \ \forall j \in \{1, \dots, r-1\}. \end{cases} \tag{2.14}$$

We denote $C_j \cap K$ by K_j for all $j \in \mathcal{J}_0$. For later purposes, if $\beta \rightsquigarrow C_1$, we also denote by K_0 the singleton $\{z_0\}$ where z_0 is an arbitrary reference letter in $\operatorname{supp}\beta$ such that $z_0 \rightsquigarrow C_1$.

2.4.1 The slicing map

Definition 2.11 (The slicing map)**.** *The slicing map is the map $F_{\mathcal{J}_0} : S_{\text{fin},+} \to S_{\text{fin}}^{\mathcal{J}_0}$ defined as follows. Let $u \in S_{\text{fin},+}$. For all $j \in \mathcal{J}_0$, if u does not have any letter in K_j, we define $u^j = e$. Otherwise, we define u^j as the largest subword of u whose first and last letters are in K_j. We set $F_{\mathcal{J}_0}(u) = (u^j)_{j\in\mathcal{J}_0}$.*

In the following, except in Section 2.5.2, the set $\mathcal{J}_0$ will be fixed, so we will simply write F instead of $F_{\mathcal{J}_0}$.

Remark 2.12. Let $u \in S_{\text{fin},+}$ and $(u^j)_{j\in\mathcal{J}_0} = F(u)$. It follows from the definition that when $u^j \neq e$, the first and last letters of u^j belong to $K_j \subseteq C_j$, but the letters in between may not belong to K_j. However, since $u \in S_{\text{fin},+}$ implies that $u_i \rightsquigarrow u_{i+1}$ for all $1 \leqslant i \leqslant |u| - 1$, we observe that all letters of u^j belong to C_j, and that the only letters of u that belong to K_j are those contained in u^j. In particular, the subwords u^j are non-overlapping. The word u can be decomposed as follows:

$$u = \zeta^1 u^1 \zeta^2 u^2 \dots \zeta^r u^r \zeta^{r+1}, \tag{2.15}$$

where the ζ^j are the remaining subwords of u, in between the subwords u^j. In order to prevent any ambiguity in (2.15), we set $\zeta^{j+1} = e$ when $u^j = e$.

A visual representation of the application of F to two words u^1 and u^2, decomposed respectively into $F(u^1) = (u^{1,1}, u^{1,2})$ and $F(u^2) = (u^{2,1}, u^{2,2})$, is given in Figure 3(a). Two useful properties of F are provided in Lemmas 2.13 and 2.14 below. In the following, we let

$$\Gamma = \bigcup_{j\in\mathcal{J}_0} K_j^2. \tag{2.16}$$

Lemma 2.13. *Let $u \in S_{\text{fin},+}$ and let $(u^j)_{j\in\mathcal{J}_0} = F(u)$. Then, for all $j \in \mathcal{J}_0$,*

$$\left| M[u]|_{C_j^2} - M[u^j] \right|_{\text{TV}} \leqslant M[u](C_j^2 \backslash K_j^2), \tag{2.17}$$

and

$$\left| M[u] - \sum_{j\in\mathcal{J}_0} M[u^j] \right|_{\text{TV}} \leqslant M[u](S^2 \backslash \Gamma). \tag{2.18}$$

Proof. Let $j \in \mathcal{J}_0$. By Remark 2.12, we have $M[u^j] \leqslant M[u]|_{C_j^2}$ in $\mathcal{M}(S^2)$, because every term $\delta_{(u_k,u_{k+1})}$ of the sum $M[u^j]$ also appears in the sum $M[u]|_{C_j^2}$. Thus, the left-hand side of (2.17) is equal to $M[u](C_j^2) - M[u^j](C_j^2)$. Since we also have

$$M[u^j](C_j^2) \geqslant M[u^j](K_j^2) = M[u](K_j^2),$$

we obtain (2.17).

In the same way, $\sum_{j\in\mathcal{J}_0} M[u^j] \leqslant M[u]$, so the left-hand side of (2.18) is equal to $M[u](S^2) - \sum_{j\in\mathcal{J}_0} M[u^j](S^2)$. Since

$$M[u^j](S^2) \geqslant M[u^j](K_j^2) = M[u](K_j^2),$$

we obtain (2.18). □

In the following, if $\underline{u} = (u^1, \dots, u^k) \in S_{\text{fin}}^k$ is a list of words, we let $k_{\underline{u}} = k$ be the number of entries in $\underline{u}$ and

$$|\underline{u}| = \sum_{i=1}^{k} |u^i| \tag{2.19}$$

be the total length of words of $\underline{u}$.

Lemma 2.14. *Let F_n be the restriction of the map F to $S^{n+1} \cap S_{\text{fin},+}$. Then, for all $\underline{v} = (v^j)_{j\in\mathcal{J}_0}$,*

$$\mathbb{P}_{n+1}(F_n^{-1}(\underline{v})) \leqslant c_{\text{sl}} \prod_{j=1}^{r} p(v^j), \tag{2.20}$$

where $c_{\text{sl}} = (n+1)^{r+1}$.

Proof. Inequality (2.20) is trivial if $\underline{v}$ is not in the range of F_n. Assume $\underline{v}$ is in the range of F_n and let $u \in S_{\text{fin},+}$ be such that $F_n(u) = \underline{v}$, *i.e.* such that $u^j = v^j$ for all $1 \leqslant j \leqslant r$, where we let u^j be the subword $F(u)^j$. We begin by comparing $\mathbb{P}_{n+1}(u)$ and the product of every $p(u^j)$ by identifying their common factors, and bounding the remaining ones. Consider the decomposition (2.15) of u. The quantity $p(u)$ is a product of factors $p(u^j)$, $p(\zeta^j)$, and $p(x,y)$, x and y being the first or last letters of some u^j or ζ^j. In this product, we bound by 1 each factor $p(x,y)$ except the ones for which y is the first letter of some ζ^j. We have

$$\mathbb{P}_{n+1}(u) \leqslant \left(\prod_{j=1}^{r} p(u^j)\right) Q(u), \qquad Q(u) = \prod_{1\leqslant j\leqslant r} q_u^j(\zeta^j), \tag{2.21}$$

where, for any $1 \leqslant j \leqslant r$, the function q_u^j is defined on S_{fin} as follow:

- if $\zeta^j \neq e$ and $j > 1$, then $q_u^j(\xi) = p(x\xi)$, where x is the letter preceding ζ^j in u (such a letter necessarily exists, since by convention $\zeta^j = e$ when $u^{j-1} = e$);
- if $\zeta^j \neq e$ and $j = 1$, then $q_u^j(\xi) = \mathbb{P}_{|\xi|}(\xi)$;[18]
- if $\zeta^j = e$, then $q_u^j(\xi) = 1$ for all $\xi \in S_{\text{fin}}$.

Notice that, for all $1 \leqslant j \leqslant r$,[19]

$$\sum_{\xi\in S^{|\zeta^j|}} q_u^j(\xi) = 1. \tag{2.22}$$

Inequality (2.21) already resembles the conclusion (2.20). However, the map F_n is far from injective, because of the complete loss of data carried by the ζ^i, which prevents us from deriving (2.20) from (2.21) by simply bounding $Q(u)$ by 1. We will prove that

$$\sum_{u\in F_n^{-1}(\underline{v})} Q(u) \leqslant (n+1)^{r+1}. \tag{2.23}$$

Taking (2.23) for granted, we obtain (2.20) by summing (2.21) over all $u \in F_n^{-1}(\underline{v})$.

Thus, it only remains to prove (2.23). From now on, since $u \in F_n^{-1}(\underline{v})$ may vary, we will write $\zeta^j(u)$ instead of simply ζ^j to prevent any confusion. To compute the sum on the left-hand

[18] Notice that the initial distribution β only appears in $Q(u)$ if $\zeta^1 \neq e$. If $\zeta^1 = e$, the inequality (2.21) bounds $\beta(u_1)$ by 1, while if $\zeta^1 \neq e$ we include $\beta(u_1)$ in $Q(u)$ in anticipation of some summation below.

[19] In other words, for any j, the function q_u^j defines a probability measure on $S^{|\zeta^j|}$. Also notice that ζ^j is involved in (2.22) only through its length.

side of (2.23), we partition $F_n^{-1}(\underline{v})$ according to the length of the $\zeta^j(u)$. Let us define integers $\ell_1, \ldots, \ell_{r+1} \geqslant 0$ such that

$$\ell_1 + \ldots + \ell_{r+1} + |\underline{v}| = n + 1,$$

and consider all preimages u of $\underline{v}$ such that the length of each $\zeta^j(u)$ is exactly ℓ_j. Notice that all u satisfying this constraint have the same functions $q_u^j(\cdot)$, which we simply call $q^j(\cdot)$. Under this constraint, identifying a specific datum u amounts simply to identifying its subwords $\zeta^j(u)$. The sum of $Q(u)$ over this specific set of preimages is, by (2.22),

$$\sum_{\zeta^1 \in S^{\ell_1}} \sum_{\zeta^2 \in S^{\ell_2}} \cdots \sum_{\zeta^{r+1} \in S^{\ell_{r+1}}} \prod_{j=1}^{r+1} q^j(\zeta^j) = \prod_{j=1}^{r+1} \sum_{\zeta^j \in S^{\ell_j}} q^j(\zeta^j) = \prod_{j=1}^{r+1} 1 = 1.$$

This means that the left-hand side of (2.23) is bounded above by the number of possible ways to choose $(\ell_1, \ldots, \ell_{r+1})$. A crude upper bound on this number is obtained as follows. There are at most $n+1$ possibilities for each ℓ_j, thus there are at most $(n+1)^{r+1}$ possibilities for $(\ell_1, \ldots, \ell_{r+1})$. The bound (2.23) is proved, and the proof is complete. □

2.4.2 The stitching map

Recall that z_0 was defined at the beginning of the section to be a letter in $\operatorname{supp}\beta$ such that $z_0 \rightsquigarrow C_1$ and $K_0 = \{z_0\}$. Also recall that if $\underline{v}$ is a list of words, $k_{\underline{v}}$ denotes the number of entries in $\underline{v}$ and $|\underline{v}|$ denotes the total length of $\underline{v}$, as in (2.19).

Lemma 2.15 (Transition words)**.** *Assume that $\beta \rightsquigarrow C_1$. Let Δ be the set of pairs (x, y) where $x \in K_j$, $y \in K_{j'}$, for some $0 \leqslant j \leqslant r$ and $1 \vee j \leqslant j' \leqslant r$. Then, for all $(x, y) \in \Delta$, there exists a finite word $\xi_{x,y}$ such that*

$$\eta := \inf_{(x,y)\in\Delta} p(x\xi_{x,y}y) > 0, \qquad \tau := \sup_{(x,y)\in\Delta} |\xi_{x,y}| + 1 < \infty. \tag{2.24}$$

Proof. For all $(x, y) \in \Delta$, we have $x \rightsquigarrow y$, so the definition of the relation $\rightsquigarrow$ yields a word $\xi_{x,y}$ satisfying $p(x\xi_{x,y}y) > 0$. Since Δ is finite, the conclusion is immediate. □

Definition 2.16 (The stitching map)**.** *Let $t \in \mathbb{N}$. We say that a finite sequence of non-empty words $\underline{v} = (v^1, \ldots, v^k) \in S_{\mathrm{fin}}^k$ is* stitchable *if there exists $1 \leqslant j(1) \leqslant j(2) \leqslant \ldots \leqslant j(k)$ such that for each $1 \leqslant i \leqslant k$, the first and last letter of v^i belongs to $K_{j(i)}$. We let*

$$A^{(t)} = \left\{ \underline{v} \in \bigcup_{k\in\mathbb{N}} S_{\mathrm{fin}}^k \;\middle|\; \underline{v} \textit{ is stitchable}, \; |\underline{v}| \geqslant t+1 \right\}$$

denote the set of stitchable sequences of total length greater than $t+1$. We define the stitching map $G_t : A^{(t)} \to S^{t+1}$ in the following way. When $\underline{v} \in A^{(t)}$, for $2 \leqslant i \leqslant k_{\underline{v}}$, let ξ^i be the word $\xi_{x,y}$ from Lemma 2.15 where x is the last letter of v^{i-1} and y is the first letter of v^i. Denote by ξ^1 the word $\xi_{z_0,y}$, where y is the first letter of v^1. Let

$$G(\underline{v}) = z_0 \xi^1 v^1 \xi^2 v^2 \ldots \xi^{k_{\underline{v}}} v^{k_{\underline{v}}},$$

and let $G_t(\underline{v})$ be the prefix of the word $G(\underline{v})$ of length $t+1$.

Note that $G_t(\underline{v})$ is well defined when $\underline{v} \in A^{(t)}$ because in that case, $|G(\underline{v})| \geqslant t+1$. A visual representation of the stitching map is provided in Figure 3(b).

Lemma 2.17. *Let $k \in \mathbb{N}$ and $t \in \mathbb{N}$. For all $\underline{v} \in A^{(t)}$,*

$$\left| M[G_t(\underline{v})] - \sum_{i=1}^{k_{\underline{v}}} M[v^i] \right|_{\mathrm{TV}} \leqslant |\underline{v}| - (t+1) + 2k_{\underline{v}}\tau, \tag{2.25}$$

where τ is as in (2.24).

Proof. We first compute the distance between $M[G(\underline{v})]$ and $\sum_i M[v^i]$. As they are both sums of Dirac measures, we only have to count the number of terms that differ between the two. They share every Dirac term $\delta_{(v^i_j, v^i_{j+1})}$, thus, their difference consists of the sum of terms involving at least one letter of some ξ^i or z_0. Since the words $z_0\xi^1, \xi^2, \dots, \xi^{k_{\underline{v}}}$ each contribute to at most τ terms in the sum defining $M[G_t(\underline{v})]$, we find

$$\left| M[G(\underline{v})] - \sum_{i=1}^{k_{\underline{v}}} M[v^i] \right|_{\mathrm{TV}} \leqslant k_{\underline{v}}\tau. \tag{2.26}$$

Since $G_t(\underline{v})$ is a prefix of $G(\underline{v})$, the TV distance between $M[G_t(\underline{v})]$ and $M[G(\underline{v})]$ is the difference between the length of $G(\underline{v})$ and $t+1$. By definition of the stitching map, we have $|G(\underline{v})| \leqslant |\underline{v}| + k_{\underline{v}}\tau$, hence

$$\left| M[G_t(\underline{v})] - M[G(\underline{v})] \right|_{\mathrm{TV}} \leqslant |\underline{v}| - (t+1) + k_{\underline{v}}\tau. \tag{2.27}$$

Combining (2.27) with (2.26) yields (2.25), as claimed. □

Lemma 2.18. *Let $t, k_*, l_* \in \mathbb{N}$. Let*

$$B^{(t)}_{k_*,l_*} = \{\underline{v} \in A^{(t)} \mid k_{\underline{v}} \leqslant k_*,\ |\underline{v}| \leqslant l_*\}. \tag{2.28}$$

Then, for all $w \in S^{t+1}$,

$$\mathbb{P}_{t+1}(w) \geqslant c_{\mathrm{st}} \sum_{\substack{\underline{v} \in B^{(t)}_{k_*,l_*} \\ G_t(\underline{v}) = w}} \prod_{i=1}^{k_{\underline{v}}} p(v^i), \tag{2.29}$$

where

$$c_{\mathrm{st}} = \beta(z_0)\eta^{k_*} \frac{1}{(l_* + \tau k_*)^2} \left(\frac{2k_*}{e(l_* + (\tau+2)k_*)} \right)^{2k_*}.$$

Proof. First, notice that for all $\underline{v} \in B^{(t)}_{k_*,l_*}$, we have $\mathbb{P}_{|G(\underline{v})|}(G(\underline{v})) = \beta(z_0)p(G(\underline{v}))$ and

$$p(G(\underline{v})) \geqslant \eta^{k_{\underline{v}}} \prod_{i=1}^{k_{\underline{v}}} p(v^i) \geqslant \eta^{k_*} \prod_{i=1}^{k_{\underline{v}}} p(v^i). \tag{2.30}$$

Let $w \in S^{t+1}$. The remainder of the proof consists in summing (2.30) over all $\underline{v} \in B^{(t)}_{k_*,l_*}$ such that $G_t(\underline{v}) = w$, in order to obtain (2.29). Assume $w \in G_t(B^{(t)}_{k_*,l_*})$, and consider a word $\kappa \in S_{\mathrm{fin}}$ satisfying $|\kappa| + t + 1 \leqslant l_* + \tau k_*$. Let us show that

$$c\mathbb{P}_{|w\kappa|}(w\kappa) \geqslant \sum_{\substack{\underline{v} \in B^{(t)}_{k_*,l_*} \\ G(\underline{v}) = w\kappa}} \prod_{i=1}^{k_{\underline{v}}} p(v^i), \qquad c = k_*\beta(z_0)^{-1}\eta^{-k_*} \left(\frac{e(l_* + (\tau+2)k_*)}{2k_*} \right)^{2k_*}. \tag{2.31}$$

Indeed, by summing (2.30) over all $\underline{v} \in G^{-1}(w\kappa) \cap B^{(t)}_{k_*,l_*}$, we get

$$\#\big(G^{-1}(w\kappa) \cap B^{(t)}_{k_*,l_*}\big) p(w\kappa) \geqslant \eta^{k_*} \sum_{\substack{\underline{v} \in B^{(t)}_{k_*,l_*} \\ G(\underline{v}) = w\kappa}} \prod_{i=1}^{k_{\underline{v}}} p(v^i). \tag{2.32}$$

To prove (2.31), it suffices to provide an estimate of the cardinality of $G^{-1}(w\kappa) \cap B^{(t)}_{k_*,l_*}$. Let $k \leqslant k_*$. To determine a preimage $\underline{v}$ of $w\kappa$ by G such that $k_{\underline{v}} = k$, it suffices to specify the indices where the words v^i and ξ^i start in the word $G(\underline{v}) = w\kappa = z_0\xi^1 v^1 \dots \xi^k v^k$. This piece of information is given by the datum of a nondecreasing sequence of $2k$ elements in $\{0, 1, \dots, |w\kappa| - 1\}$. Thus,

$$\#\{\underline{v} \in G^{-1}(w\kappa) \mid k_{\underline{v}} = k\} \leqslant \binom{|w\kappa| + 2k - 1}{2k} \leqslant \left(\frac{e(|w\kappa| + 2k_*)}{2k_*} \right)^{2k_*}.$$

In the bound above, we have used that $\binom{x+y}{y} \leqslant (e(x+y)/y)^y$ for any integers $y \leqslant x$, and that the function defined in the right-hand side of the bound is nondecreasing in x and y. Hence, we have

$$\#\{\underline{v} \in G^{-1}(w\kappa) \mid k_{\underline{v}} \leqslant k_*\} \leqslant k_* \Big(\frac{e(|w\kappa| + 2k_*)}{2k_*}\Big)^{2k_*}.$$

By inclusion, this is also an upper bound on the cardinality of $G^{-1}(w\kappa) \cap B^{(t)}_{k_*,l_*}$. Hence, since $|w\kappa| = |\kappa| + t + 1 \leqslant l_* + \tau k_*$, we have

$$\#\big(G^{-1}(w\kappa) \cap B^{(t)}_{k_*,l_*}\big) \leqslant k_* \Big(\frac{e(l_* + (\tau + 2)k_*)}{2k_*}\Big)^{2k_*}.$$

Applying this bound to (2.32) and multiplying the inequality by $\beta(z_0)$, we obtain (2.31) with the announced constant c. Now, we can use the estimations on $\mathbb{P}_{|w\kappa|}(w\kappa)$ provided by (2.31) to bound the value of $\mathbb{P}_{t+1}(w)$. Observe that for all $m \geqslant 0$,

$$\sum_{\kappa \in S^m} \mathbb{P}_{t+1+m}(w\kappa) = \mathbb{P}_{t+1}(w) \sum_{\kappa \in S^m} p(w_{t+1}\kappa) = \mathbb{P}_{t+1}(w).$$

Hence, by summing both sides in (2.31) over all possible suffixes κ, we have:

$$\begin{aligned}
(l_* + \tau k_* - (t+1))c\mathbb{P}_{t+1}(w) &= \sum_{\substack{\kappa \in S_{\text{fin}} \\ |w\kappa| \leqslant l_* + \tau k_*}} c\mathbb{P}_{|w\kappa|}(w\kappa) \\
&\geqslant \sum_{\substack{\kappa \in S_{\text{fin}} \\ |w\kappa| \leqslant l_* + \tau k_*}} \sum_{\substack{\underline{v} \in B^{(t)}_{k_*,l_*} \\ G(\underline{v}) = w\kappa}} \prod_{i=1}^{k_{\underline{v}}} p(v^i) \\
&= \sum_{\substack{\underline{v} \in B^{(t)}_{k_*,l_*} \\ G_t(\underline{v}) = w}} \prod_{i=1}^{k_{\underline{v}}} p(v^i).
\end{aligned}$$

The last equality holds because all $\underline{v} \in B^{(t)}_{k_*,l_*}$ satisfy $|G(\underline{v})| \leqslant l_* + \tau k_*$. A rough upper bound on the obtained constant yields (2.29) with constant c_{st}. □

2.5 Coupling trajectories

Using the slicing and stitching maps of the previous section, we build here the *coupling* and *decoupling maps*. In Section 2.5.1, the properties of the coupling map are used to finally prove Proposition 2.4, thus allowing the machinery of Section 2.2 to provide the existence of the RL function and the weak LDP for (L_n). In Section 2.5.2, the properties of the decoupling map yield Proposition 2.6, which is used in the derivation of the additional properties of the rate function, in Proposition 2.7.

2.5.1 The coupling map and proof of Proposition 2.4

The goal of this section is to prove Proposition 2.4. Let us fix δ, μ_1, μ_2, μ, N, N_1, and N_2 as in Proposition 2.4. Note that this choice of parameters yields the two convenient bounds[20]

$$\frac{1}{1-4\delta} \leqslant 1 + 6\delta, \qquad t \geqslant \frac{n}{\delta} \;\Rightarrow\; 1 \leqslant \frac{Nn}{t} \leqslant 1 + 8\delta. \tag{2.33}$$

We also fix a finite set $K \subseteq S$ such that

$$\mu_1(K^2) \geqslant 1 - \delta, \qquad \mu_2(K^2) \geqslant 1 - \delta, \tag{2.34}$$

and we set $\mathcal{J}_0 = \mathcal{J}_\mu \cap \mathcal{J}_K$, where $\mathcal{J}_K = \{j \in j \mid K \cap C_j \neq \emptyset\}$. The measure μ being admissible, $\mathcal{J}_0$ satisfies the same assumptions as in Section 2.4, and we keep the notations of (2.14).

[20] The first one is the reason why we chose $\delta < 1/12$.

Lemma 2.19. *Let $k \in \mathbb{N}$. There exists a map $\sigma : F(S_{\mathrm{fin},+})^k \to \bigcup_{i=0}^{rk} S_{\mathrm{fin},+}^i$ such that, for all $\underline{w} \in F(S_{\mathrm{fin},+})^k$, the list $\sigma(\underline{w})$ is stitchable and the elements of $\sigma(\underline{w})$ are exactly the non-empty elements of $\underline{w}$.*

Proof. Let $\underline{u} = (u^1, \ldots, u^k) \in S_{\mathrm{fin},+}^k$. Applying F to each word u^i yields a $k \times r$ matrix $(u^{i,j})$ whose lines are $(u^{i,j})_{1 \leqslant j \leqslant r} = F(u^i)$. If $u^{i,j} \neq e$, then the first and last letters of $u^{i,j}$ belong to K_j. We set $\sigma(F(u^1), \ldots, F(u^k))$ to the be the list obtained by reading the entries of the matrix $(u^{i,j})$ column by column, while skipping empty words. □

We introduce the *coupling map* in Definition 2.20 below. The coupling map takes N words of length $n+1$ and turns them into one word of length $t+1$, using the slicing and stitching maps.

Definition 2.20 (The coupling map)**.** *Let $t, n \in \mathbb{N}$. Let $\underline{u} = (u^1, \ldots, u^N) \in (S_{\mathrm{fin},+})^N$ satisfy $|u^i| = n+1$ for all $1 \leqslant i \leqslant N$. Let $\underline{v} = \sigma(F(u^1), \ldots, F(u^N))$, where σ is the reordering map of Lemma 2.19. If $|\underline{v}| \geqslant t+1$, we set $\Psi_{n,t}(\underline{u}) = G_t(\underline{v})$, where G_t is the stitching map from Definition 2.16. We call*

$$\Psi_{n,t} : (S_{\mathrm{fin},+} \cap S^{n+1})^N \to S_{\mathrm{fin}}$$

the coupling map.

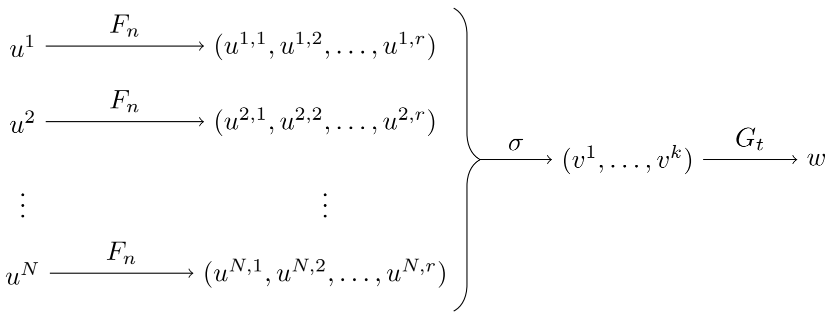


Figure 2: Definition of the coupling map.

The definition of the coupling map as the composition of several maps is sketched in Figure 2. Another useful illustration of the way the coupling map acts on trajectories is provided in Figure 3.

Lemma 2.21. *There exists a finite set $S_1 \subseteq S$ such that for all integers n, t and $S_2 \supseteq S_1$, for all $\underline{u} = (u^1, \ldots, u^N) \in (S_2^{n+1})^N$, the word $\Psi_{n,t}(u)$ belongs to S_2^{t+1} if it exists.*

Proof. We fix S_1 to be the union of K, $\{z_0\}$, and the set of all letters involved in the transition words $\xi_{x,y}$ with $x, y \in K$. Then, S_1 is a finite set. Let $S_2 \supseteq S_1$ and let $\underline{u} = (u^1, \ldots, u^N) \in (S_2^{n+1})^N$. By construction, any letter in the word $\Psi_{n,t}(u)$ belongs either to a transition word $\xi_{x,y}$ or to some u^i. In both cases, it is an element of S_2. □

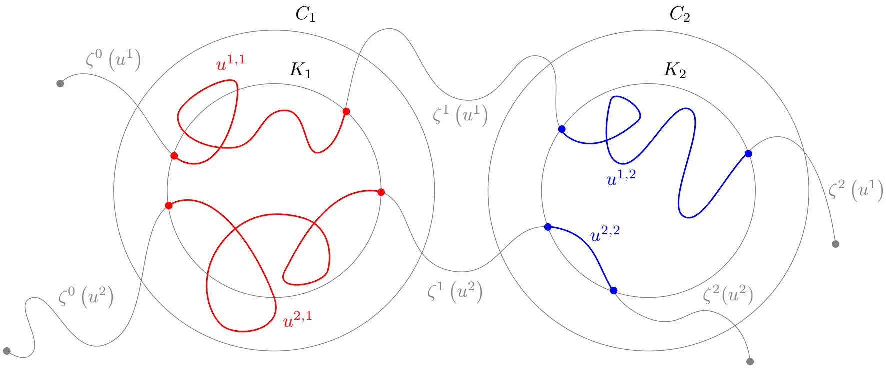
Lemma 2.22. *Let n, t be integers. Let $\underline{u} = (u^1, \ldots, u^N) \in (S^{n+1})^N$ satisfy $\sum_{i=1}^N M[u^i](\Gamma) \geqslant t+1$, where Γ was defined in (2.16). Then, $\Psi_{n,t}(\underline{u})$ is well defined and*

$$\left| M[\Psi_{n,t}(\underline{u})] - \sum_{i=1}^N M[u^i] \right|_{\mathrm{TV}} \leqslant 2\big(Nn - (t+1)\big) + 2N(r\tau + 1). \tag{2.35}$$

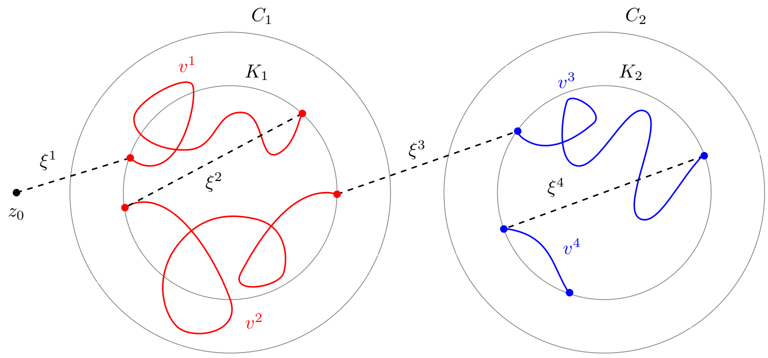
Proof. Let $\underline{v} = (v^1, \ldots, v^k)$ be the stitchable list $\sigma(F(u^1), \ldots, F(u^N))$ as in Definition 2.20.[21] We first argue that $\underline{v} \in A^{(t)}$, so that $\Psi_{n,t}(\underline{u})$ is well defined. By its definition, $\underline{v}$ is stitchable, hence it suffices to show that $|\underline{v}| \geqslant t+1$, *i.e.* that

$$\sum_{i=1}^k |v^i| = \sum_{i=1}^N \sum_{j=1}^r \left| M[F_n(u^i)^j] \right|_{\mathrm{TV}} + k \geqslant t+1. \tag{2.36}$$

[21]Since the number of empty words in the list may vary, k is not fixed here. The number k depends on the data $(u^1, \ldots, u^N)$.



(a) The action of the slicing map on two words u^1 and u^2.



(b) The action of the stitching map on a stitchable list (v^1, v^2, v^3, v^4).

Figure 3: Example of use of the coupling map, with two words u^1 and u^2. The slicing map produces two subwords $u^{1,1}$, $u^{1,2}$ of u^1 and two subwords $u^{2,1}$, $u^{2,2}$ of u^2 in (3(a)). These are then reordered as the stitchable sequence $(v^1, \dots, v^4)$ and reassembled by the stitching map in (3(b)). Here, we assume that the length of the obtained word is $t+1$ so the operation of taking the prefix of length $t+1$ is not represented. Trajectories are depicted as smooth curves for easier readability, even though they are discrete.

By Lemma 2.13 and the assumption on $\underline{u}$,

$$\sum_{i=1}^{N}\sum_{j=1}^{r}\left|M[F_n(u^i)^j]\right|_{\mathrm{TV}} \geqslant \sum_{i=1}^{N}\left(\left|M[u^i]\right|_{\mathrm{TV}} - \left|M[u^i] - \sum_{j=1}^{r} M[F_n(u^i)^j]\right|_{\mathrm{TV}}\right)$$
$$\geqslant \sum_{i=1}^{N}\big(n - (n - M[u^i](\Gamma))\big) \geqslant t+1.$$

This shows (2.36), thus $\Psi_{n,t}(\underline{u})$ is well defined. We now turn to the proof of (2.35). By definition,

$$\sum_{i=1}^{N}\sum_{j=1}^{r} M[F(u^i)^j] = \sum_{i=1}^{k} M[v^i], \qquad \Psi_{n,t}(\underline{u}) = G_t(\underline{v}).$$

On the one hand, by Lemma 2.17,

$$\left|M[\Psi_{n,t}(\underline{u})] - \sum_{i=1}^{k} M[v^i]\right|_{\mathrm{TV}} \leqslant |\underline{v}| - (t+1) + 2k\tau \leqslant N(n+1) - (t+1) + 2Nr\tau.$$

On the other hand, by Lemma 2.13,

$$\begin{aligned}\Big|\sum_{i=1}^{N}\sum_{j=1}^{r}M[F_n(u^i)^j]-\sum_{i=1}^{N}M[u^i]\Big|_{\mathrm{TV}} &\leqslant \sum_{i=1}^{N}\Big|\sum_{j=1}^{r}M[F(u^i)^j]-M[u^i]\Big|_{\mathrm{TV}}\\ &\leqslant \sum_{i=1}^{N}M[u^i](S^2\setminus\Gamma)\\ &\leqslant Nn-(t+1).\end{aligned}$$

Combining these two bounds yields (2.35) and completes the proof. □

Lemma 2.23. *For all $w\in S^{t+1}$ in the range of $\Psi_{n,t}$,*

$$\mathbb{P}_{t+1}(w)\geqslant C_{n,t}\mathbb{P}_{n+1}^{\otimes N}\left(\Psi_{n,t}^{-1}(w)\right), \tag{2.37}$$

for some constants $C_{n,t}$ satisfying (2.6).

Proof. To prove (2.37), we compute the probability of the preimage $\Psi_{n,t}^{-1}(w)$ of a fixed word w in the range of $\Psi_{n,t}$. Let $\underline{u}=(u^1,\dots,u^N)$ be a preimage of w by $\Psi_{n,t}$, that is to say $\underline{u}$ satisfies $G(\sigma(F_n(u^1),\dots,F_n(u^N)))=w$. Since we only consider words u^i of length $n+1$, that are each sliced by F_n into at most r subwords, the stitchable sequence $\sigma(F_n(u^1),\dots,F_n(u^N))$ is of total length at most $N(n+1)$ and has less than Nr entries. In other words, every preimage $\underline{u}$ of w by $\Psi_{n,t}$ satisfies

$$\sigma(F_n(u^1),\dots,F_n(u^N))\in B^{(t)}_{Nr,N(n+1)}, \tag{2.38}$$

where $B^{(t)}_{k_*,l_*}$ is defined by (2.28). Then, by (2.38),

$$\begin{aligned}\mathbb{P}_n^{\otimes N}(\Psi_{n,t}^{-1}(w)) &= \sum_{\substack{\underline{v}\in A^{(t)}\\ G_t(\underline{v})=w}}\ \sum_{u\in\sigma^{-1}(\underline{v})}\prod_{i=1}^{N}\mathbb{P}_n(F_n^{-1}(u^i))\\ &= \sum_{\substack{\underline{v}\in B^{(t)}_{Nr,N(n+1)}\\ G_t(\underline{v})=w}}\ \sum_{u\in\sigma^{-1}(\underline{v})}\prod_{i=1}^{N}\mathbb{P}_n(F_n^{-1}(u^i)).\end{aligned} \tag{2.39}$$

Let $\underline{v}\in B^{(t)}_{Nr,N(n+1)}$ be such that $G_t(\underline{v})=w$ and let $u=(u^{i,j})\in\sigma^{-1}(\underline{v})$. By Lemma 2.14,

$$\mathbb{P}_n(F_n^{-1}(u^i))\leqslant c_{\mathrm{sl}}\prod_{j=1}^{r}p(u^{i,j})$$

for all $1\leqslant i\leqslant N$. There is a one-to-one correspondence between the family of non-empty subwords $u^{i,j}$ and $(v^1,\dots,v^{k_{\underline{v}}})$, hence,

$$\prod_{i=1}^{N}\mathbb{P}_n(F^{-1}(u^i))\leqslant c_{\mathrm{sl}}^N\prod_{i=1}^{N}\prod_{j=1}^{r}p(u^{i,j})=c_{\mathrm{sl}}^N\prod_{j=1}^{k_{\underline{v}}}p(v^j).$$

Therefore, by injecting this bound into (2.39), we obtain

$$\mathbb{P}_n^{\otimes N}(\Psi_{n,t}^{-1}(w))\leqslant\sum_{\substack{\underline{v}\in B^{(t)}_{Nr,N(n+1)}\\ G_t(\underline{v})=w}}\#\sigma^{-1}(\underline{v})c_{\mathrm{sl}}^N\left(\prod_{j=1}^{k_{\underline{v}}}p(v^j)\right). \tag{2.40}$$

We need a combinatorial argument to bound the factor $\#\sigma^{-1}(\underline{v})$, *i.e.* the number of $N\times r$ matrices of words $U=(u^{i,j})$ such that $\sigma(U)=\underline{v}$. If we know which of the words $u^{i,j}$ are empty, there is only one possibility, that is filling the matrix of words with entries v^l in the right order while leaving blank the specified entries. Hence $\#\sigma^{-1}(\underline{v})$ is simply the number of choices of which of the Nr words are empty, knowing that there must be $Nr - k_{\underline{v}}$ of them. Using the crude bound

$$\#\sigma^{-1}(\underline{v}) = \binom{Nr}{k_{\underline{v}}} \leqslant 2^{Nr}$$

in (2.40), we obtain

$$\mathbb{P}_n^{\otimes N}(\Psi_{n,t}^{-1}(w)) \leqslant 2^{Nr} c_{\mathrm{sl}}^N \sum_{\substack{\underline{v} \in B^{(t)}_{Nr, N(n+1)} \\ G_t(\underline{v}) = w}} \prod_{j=1}^{k_{\underline{v}}} p(v^j).$$

Hence, applying Lemma 2.18 with $k_* = Nr$ and $l_* = N(n+1)$, we obtain

$$\mathbb{P}_n^{\otimes N}(\Psi_{n,t}^{-1}(w)) \leqslant \frac{2^{Nr} c_{\mathrm{sl}}^N}{c_{\mathrm{st}}} \mathbb{P}_{t+1}(w).$$

The bound (2.37) is proved with

$$C_{n,t} = \frac{c_{\mathrm{st}}}{2^{Nr} c_{\mathrm{sl}}^N} = 2^{-Nr}(n+1)^{-N(r+1)}(N(n+1+\tau r))^{-2}\beta(z_0)\eta^{Nr}\left(\frac{2r}{e(n+1+(\tau+2)r)}\right)^{2Nr}.$$

It remains to show that $C_{n,t}$ satisfies condition (2.6). Notice that $C_{n,t}$ is actually a function of variables (n, N) that only depends on t through the definition of N. By (2.4), at fixed n, we have

$$\begin{aligned}\lim_{t\to\infty} \frac{1}{t} \log C_{n,t} &= \frac{1}{n(1-4\delta)} \lim_{N\to\infty} \frac{1}{N} \log C_{n,t} \\ &= \frac{1}{n(1-4\delta)}\left(-r\log 2 - (r+1)\log(n+1) + r\log\eta + 2r\log\frac{2r}{e(n+1+(\tau+2)r)}\right).\end{aligned}$$

Further taking the limit as $n \to \infty$, we get (2.6). □

We can finally prove Proposition 2.4, using the properties of the coupling map established above.

Proof of Proposition 2.4. Set n_0 to be an integer such that $4(r\tau+1)/n_0 \leqslant \delta$. For all $n \in \mathbb{N}$, set t_n to be an integer such that $n/t_n \leqslant \delta$. Let S_1 be given by Lemma 2.21. Let $n \geqslant n_0$, $t \geqslant t_n$, and $S_2 \supseteq S_1$. Denote by $\mathcal{W}$ the set $W_{n,S_2}(\mu_1,\delta)^{N_1} \times W_{n,S_2}(\mu_2,\delta)^{N_2}$. Proposition 2.4 relies on the following statement:

$$\Psi_{n,t}(\mathcal{W}) \subseteq W_{S_2,t}(\mu, 20\delta), \tag{2.41}$$

which we will prove below. Once (2.41) is proved, we have

$$\mathbb{P}_{t+1}\left(W_{t,S_2}(\mu,20\delta)\right) \geqslant \mathbb{P}_{t+1}\left(\Psi_{n,t}(\mathcal{W})\right),$$

hence by Lemma 2.23,

$$\begin{aligned}\mathbb{P}_{t+1}\left(W_{t,S_2}(\mu,20\delta)\right) &\geqslant C_{n,t}\mathbb{P}_{n+1}^{\otimes N}\left(\Psi_{n,t}^{-1}\left(\Psi_{n,t}(\mathcal{W})\right)\right) \\ &\geqslant C_{n,t}\mathbb{P}_{n+1}^{\otimes N}\left(\mathcal{W}\right).\end{aligned}$$

This yields (2.5), with the constant $C_{n,t}$ from Lemma 2.23, which satisfies (2.6). It only remains to prove (2.41), which we do now. Let $\underline{u} = (u^1, \ldots, u^N) \in \mathcal{W}$. Equation (2.41) encapsulates three properties:

1. The word $\Psi_{n,t}(\underline{u})$ is well defined.
2. $L[\Psi_{n,t}(\underline{u})]$ is at most 20δ-close to μ.
3. Every letter of $\Psi_{n,t}(\underline{u})$ belongs to S_2.

Once the existence of $\Psi_{n,t}(\underline{u})$ is proved, Property 3 will be an immediate consequence of Lemma 2.21. We now prove Properties 1 and 2 by showing that $\underline{u}$ satisfies the assumptions of Lemma 2.22, namely that $\sum_{i=1}^{N} M[u^i](\Gamma) \geqslant t+1$. Let $i \leqslant N$ and let ν_i denote μ_1 or μ_2 depending on whether $i \leqslant N_1$ or $i > N_1$. Since ν_i is admissible, and by (2.34), we have $\nu_i(\Gamma) = \nu_i(K^2) \geqslant 1-\delta$. Thus,

$$L[u^i]\,(\Gamma) \geqslant \nu_i\,(\Gamma) - |L[u^i] - \nu_i|_{\mathrm{TV}} \geqslant 1-\delta-\delta.$$

Hence $M[u^i](\Gamma) \geqslant n(1-2\delta)$, and

$$\sum_{i=1}^{N} M[u^i](\Gamma) \geqslant Nn(1-2\delta) \geqslant t+1.$$

By Lemma 2.22, Property 1 holds. Another consequence of Lemma 2.22 is that

$$|L[\Psi_{n,t}(\underline{u})] - \mu|_{\mathrm{TV}} \leqslant \frac{1}{t}(Nn-(t+1)+2N(r\tau+1)) + \left|\frac{1}{t}\sum_{i=1}^{N} M[u^i] - \mu\right|_{\mathrm{TV}}. \tag{2.42}$$

Using (2.33) and $n \geqslant n_0$, the first term on the right-hand side of (2.42) above is no larger than $1+8\delta-1+\delta$. Since $\underline{u} \in \mathcal{W}$, the second term satisfies

$$\begin{aligned}
\left|\frac{1}{t}\sum_{i=1}^{N} M[u^i] - \mu\right|_{\mathrm{TV}} &\leqslant \frac{1}{t}\sum_{i=1}^{N_1} \left|M[u^i] - n\mu_1\right|_{\mathrm{TV}} + \frac{1}{t}\sum_{i=N_1+1}^{N_2} \left|M[u^i] - n\mu_2\right|_{\mathrm{TV}} \\
&\qquad + \left|\frac{N_1}{t}n\mu_1 + \frac{N_2}{t}n\mu_2 - \mu\right|_{\mathrm{TV}} \\
&\leqslant \frac{N_1}{t}n\delta + \frac{N_2}{t}n\delta + \left|\frac{N_1}{t}n - \frac{1}{2}\right| + \left|\frac{N_2}{t}n - \frac{1}{2}\right| \\
&\leqslant (1+8\delta)\delta + \frac{1}{2}8\delta + \frac{1}{2}8\delta + \delta.
\end{aligned}$$

Combining the bounds on the two terms the right-hand side of (2.42), we obtain

$$|L[\Psi_{n,t}(\underline{u})] - \mu|_{\mathrm{TV}} \leqslant 20\delta,$$

and we have finally shown Property 2. This completes the proof of (2.41) and concludes the proof of the proposition. □

Remark 2.24. In irreducible cases, the construction of the coupling map is considerably simplified, yet not trivial. Assume here that there is only one irreducible class C_1 in (1.7) (in particular, this assumption is satisfied when the Markov chain is φ-irreducible). Then, the slicing map is reduced to simply removing a prefix and a suffix from the argument. The use of the slicing map on $\underline{u} := (u^1, \dots, u^N)$ results in a stitchable sequence of at most N words and the coupled word $\Psi_{n,t}(\underline{u})$ is simply the prefix of length t of the word

$$z_0\xi^1 v^1 \xi^2 \dots \xi^v v^k,$$

where $(v_1, \dots, v^k)$ is the list $(F(u^1), \dots, F(u^N))$ with empty words removed.

If S is finite and the Markov chain is matrix-irreducible, as in Example 2.20 of [6], we recover the map $\psi_{n,t}(u)$ from [6]. Indeed, one can take $C_1 = K_1 = K = S$. Then, the slicing map is the identity, and we have

$$\Psi_{n,t}(\underline{u}) = z_0\xi(z_0, u_1^1)\psi_{n,t}(\underline{u}), \qquad \underline{u} \in (S^n)^N.$$

2.5.2 The decoupling map and proof of Proposition 2.6

The construction of the previous section achieved to prove Proposition 2.4, which is already enough to derive Theorem 2.5 in Section 2.2. The goal of the present section is to prove Proposition 2.6, which is used in the proof of Property 2 of Proposition 2.7. To do so, we present another construction that uses slicing and stitching maps.

Let λ_1, λ_2, δ, t_1, t_2, μ_1, μ_2 and μ be as in Proposition 2.6. For convenience, we also set

$$\delta_i = \frac{125}{\lambda_i^2}\delta, \qquad i \in \{1,2\}.$$

Note that, by the choice of δ and the definition of t_i, if $n \geqslant 1/\delta$, we have the following two bounds for $i \in \{1,2\}$:[22]

$$\frac{1}{\lambda_i} \leqslant \frac{n}{t_i} \leqslant \frac{1}{\lambda_i}\Big(1 + \frac{4}{\lambda_i}\delta\Big), \qquad 0 \leqslant n - t_1 - t_2 \leqslant 8n\delta. \tag{2.43}$$

Let $K \subseteq S$ be a finite set such that $\mu(K^2) > 1 - \delta$. We set

$$\mathcal{J}_i = \mathcal{J}_K \cap \mathcal{J}_{\mu_i}, \qquad \Gamma_i = \bigcup_{j \in \mathcal{J}_i} K_j^2, \qquad i \in \{1,2\},$$

where $\mathcal{J}_K = \{j \in j \mid K \cap C_j \neq \emptyset\}$. We set $\mathcal{J}_0 := \mathcal{J}_1 \cup \mathcal{J}_2$ and $r := |\mathcal{J}_1 \cup \mathcal{J}_2| < \infty$. For all $j \in \mathcal{J}_0$, let $K_j = K \cap C_j$. In the following, when $\mathcal{J}' \subseteq \mathcal{J}$, and ν is a measure, we denote

$$\nu|_{\mathcal{J}'} = \nu|_{\bigcup_{j\in\mathcal{J}'} C_j^2}.$$

In order to prove Proposition 2.6, we introduce the *decoupling map*. It acts in the opposite way of the coupling map: while the coupling map takes words approximating μ_1 and μ_2 and creates a word approximating μ, the decoupling map takes a word approximating μ and creates two words approximating μ_1 and μ_2 respectively (given that $\mu = \lambda_1\mu_1 + \lambda_2\mu_2$ and μ_1, μ_2 are supported on disjoint classes).

Definition 2.25 (The decoupling map)**.** *Let $u \in S^{n+1} \cap S_{\mathrm{fin},+}$. Applying the slicing maps $F_{\mathcal{J}_1}$ and $F_{\mathcal{J}_2}$ from Definition 2.11 to u defines two sequences $\underline{v}_1$ and $\underline{v}_2$ of subwords of u indexed by $\mathcal{J}_1$ and $\mathcal{J}_2$, respectively. Both lists $\sigma(\underline{v}_1)$ and $\sigma(\underline{v}_2)$ are stitchable, where σ is as in Lemma 2.19.*[23] *If $|\underline{v}_1| \geqslant t_1 + 1$ and $|\underline{v}_2| \geqslant t_2 + 1$, we define the words*

$$\widetilde{\Psi}_{1,n}(u) = G_{t_1}(\sigma(\underline{v}_1)), \qquad \widetilde{\Psi}_{2,n}(u) = G_{t_2}(\sigma(\underline{v}_2)),$$

where G_{t_i} is the stitching map from Definition 2.16, and we call $\widetilde{\Psi}_n : u \mapsto \Big(\widetilde{\Psi}_{1,n}(u), \widetilde{\Psi}_{2,n}(u)\Big)$ the decoupling map.

Beware that the decoupling map is by no means the inverse of the coupling map. Figure 4 provides a sketch of the definition of the decoupling map.

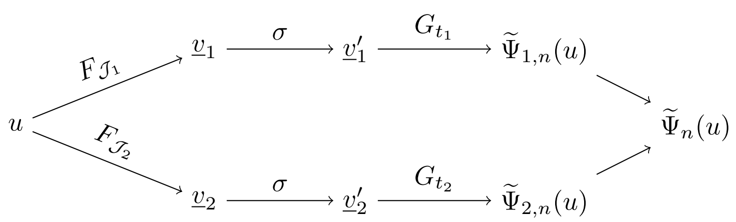


Figure 4: The definition of the decoupling map as composition of slicing and stitching maps.

Lemma 2.26. *Let $u \in S_{\mathrm{fin}}$ be such that $M[u](\Gamma_1) \geqslant t_1 + 1$ and $M[u](\Gamma_2) \geqslant t_2 + 1$. Then, for $i \in \{1,2\}$, $\widetilde{\Psi}_{i,n}(u)$ is well defined and*

$$\Big| M\Big[\widetilde{\Psi}_{i,n}(u)\Big] - M[u]|_{\mathcal{J}_i} \Big|_{\mathrm{TV}} \leqslant 2\big(|u| - 1 - t_1 - t_2\big) + 2r\tau. \tag{2.44}$$

Proof. Let $i \in \{1,2\}$. Let $\underline{v}'_i = \sigma(F_{\mathcal{J}_i}(u))$. Since $\underline{v}'_i$ is stitchable, we only have to verify that $|\underline{v}'_i| \geqslant t_i + 1$ in order for $\widetilde{\Psi}_{i,n}(u)$ to be well defined. Using the reverse triangular inequality, the bound (2.18) in Lemma 2.13 applied to $F_{\mathcal{J}_i}$ yields

$$|M[u]|_{\mathrm{TV}} - \sum_{j \in \mathcal{J}_i} \big|M[u^j]\big|_{\mathrm{TV}} \leqslant M[u](S^2 \backslash \Gamma_i) \leqslant n - (t_i + 1),$$

[22] The first bound is the reason why we required δ to be no larger than $\lambda_i/12$.
[23] Here the action of σ is simply to remove empty words, since the lists are already in the right order.

so that indeed, we have in particular $|\underline{v}_i'| \geqslant \sum_{j\in\mathcal{J}_i} \big|M[u^j]\big|_{\mathrm{TV}} \geqslant t_i+1$. We now prove (2.44). Since $M[u]|_{\mathcal{J}_i}$ is the sum of $M[u]|_{C_j^2}$ over $j\in\mathcal{J}_i$, the bound (2.17) of Lemma 2.13 yields

$$\left|M[u]|_{\mathcal{J}_i} - \sum_{j\in\mathcal{J}_i} M[u^j]\right|_{\mathrm{TV}} = \sum_{j\in\mathcal{J}_i} \big|M[u]|_{C_j^2} - M[u^j]\big|_{\mathrm{TV}} \leqslant \sum_{j\in\mathcal{J}_i} M[u](C_j^2) - M[u](\Gamma_i). \quad (2.45)$$

The elements of S^2 that are in a C_j^2 for some $j\in\mathcal{J}_i$ are not in any $K_{j'}^2$ with $j'\in\mathcal{J}_{3-i}$. Hence we have a bound on the first term of the right-hand side of (2.45):

$$\sum_{j\in\mathcal{J}_i} M[u](C_j^2) \leqslant (|u|-1) - \sum_{j'\in\mathcal{J}_i} M[u](K_{j'}^2) \leqslant |u| - t_{3-i}.$$

Also bounding the second term of the right-hand side of (2.45) by $-(t_i+1)$, we have

$$\left|M[u]|_{\mathcal{J}_i} - \sum_{j\in\mathcal{J}_i} M[u^j]\right|_{\mathrm{TV}} \leqslant |u| - t_1 - t_2 - 1. \quad (2.46)$$

Moreover, by Lemma 2.17,

$$\begin{aligned}\left|M[\widetilde{\Psi}_{i,n}(u)] - \sum_{j\in\mathcal{J}_i} M[u^j]\right|_{\mathrm{TV}} &\leqslant |\underline{v}_i'| - (t_i+1) + 2r\tau \\ &\leqslant |u| - (t_1+1) - (t_2+1) + 2r\tau. \quad (2.47)\end{aligned}$$

The last inequality holds because $|\underline{v}_1'| + |\underline{v}_2'| \leqslant |u|$. Combining (2.46) and (2.47), we obtain (2.44). □

Lemma 2.27. *For all $(w^1, w^2) \in S^{t_1+1}\times S^{t_2+1}$,*

$$\mathbb{P}_{t_1+1}(w^1)\mathbb{P}_{t_2+1}(w^2) \geqslant \widetilde{C}_n \mathbb{P}_{n+1}(\widetilde{\Psi}_n^{-1}(w^1,w^2)), \quad (2.48)$$

for some constants $\widetilde{C}_n$ satisfying (2.8).

Proof. If (w^1,w^2) is not in the range of $\widetilde{\Psi}_n$, (2.48) is trivial, so we now assume that (w^1,w^2) is in the range of $\widetilde{\Psi}_n$. For convenience, we define, for $i\in\{1,2\}$,

$$\begin{aligned}\mathcal{V}_i &= \{\underline{v}\in S_{\mathrm{fin},+}^{\mathcal{J}_i} \mid G_{t_i}(\sigma(\underline{v})) = w^i,\ \forall j\in\mathcal{J}_i,\ v^j\in C_j^{|v^j|}\}, \\ \mathcal{V}_i' &= \{\underline{v}'\in S_{\mathrm{fin},+}^{\mathcal{J}'} \mid \mathcal{J}'\subseteq\mathcal{J}_i,\ G_{t_i}(\underline{v}') = w^i,\ \forall j\in\mathcal{J}',\ e\neq v^j\in C_j^{|v^j|}\}.\end{aligned}$$

Notice that σ is bijective between $\mathcal{V}_i$ and $\mathcal{V}_i'$ since its action only consists in removing empty words from the list. Let $u\in\widetilde{\Psi}_n^{-1}(w^1,w^2)$. There are $\underline{v}_1\in\mathcal{V}_1$ and $\underline{v}_2\in\mathcal{V}_2$ such that $F_{\mathcal{J}_1}(u) = \underline{v}_1$ and $F_{\mathcal{J}_2}(u) = \underline{v}_2$. Notice that then, $F_{\mathcal{J}_0}(u)$ is the list $\underline{v} = (v^j)_{j\in\mathcal{J}_0}$ obtained by combining and reordering the (possibly empty) words of $\underline{v}_1$ and $\underline{v}_2$. Since $\mathcal{J}_1\cap\mathcal{J}_2 = \emptyset$, the map $h : (\underline{v}_1,\underline{v}_2)\mapsto\underline{v}$ is a bijection. It follows that

$$\mathbb{P}_{n+1}(\Psi_n^{-1}(w^1,w^2)) = \sum_{\substack{\underline{v}_1\in\mathcal{V}_1\\ \underline{v}_2\in\mathcal{V}_2}} \mathbb{P}_{n+1}(F_n^{-1}(h(\underline{v}_1,\underline{v}_2))),$$

where F_n is the restriction of $F_{\mathcal{J}_0}$ to $S_{\mathrm{fin},+}\cap S^{n+1}$ as in Lemma 2.14. By Lemma 2.14 applied to each term of the sum, with $(v^j)_{j\in\mathcal{J}_0} := h(\underline{v}_1,\underline{v}_2)$, we have

$$\begin{aligned}\mathbb{P}_n(\widetilde{\Psi}_n^{-1}(w^1,w^2)) &\leqslant c_{\mathrm{sl}} \sum_{\substack{\underline{v}_1\in\mathcal{V}_1\\ \underline{v}_2\in\mathcal{V}_2}} \prod_{j\in\mathcal{J}_0} p(v^j) \\ &= c_{\mathrm{sl}}\left(\sum_{\underline{v}_1\in\mathcal{V}_1}\prod_{j\in\mathcal{J}_1} p((\underline{v}_1)^j)\right)\left(\sum_{\underline{v}_2\in\mathcal{V}_2}\prod_{j\in\mathcal{J}_2} p((\underline{v}_2)^j)\right) \\ &= c_{\mathrm{sl}}\left(\sum_{\underline{v}_1'\in\mathcal{V}_1'}\prod_{j=1}^{k_{\underline{v}_1'}} p((\underline{v}_1')^j)\right)\left(\sum_{\underline{v}_2'\in\mathcal{V}_2'}\prod_{j=1}^{k_{\underline{v}_2'}} p((\underline{v}_2')^j)\right).\end{aligned}$$

We have $\mathcal{V}_i' \subseteq G_{t_i}^{-1}(w^i)$ for $i \in \{1,2\}$, so we can crudely bound the two last sums using Lemma 2.18. By Lemma 2.18 with $k_* = r$ and $l_* = n+1$. We get

$$\mathbb{P}_n(\Psi_n^{-1}(w_1, w_2)) \leqslant \frac{c_{\mathrm{sl}}}{c_{\mathrm{st}}^2} \mathbb{P}_{t_1}(w^1)\mathbb{P}_{t_2}(w^2).$$

The bound (2.48) is proved with $\widetilde{C}_n = c_{\mathrm{st}}^2/c_{\mathrm{sl}}$. It remains to show (2.8). The constant $\widetilde{C}_n$ satisfies

$$\begin{aligned}\frac{1}{n}\log \widetilde{C}_n =& \frac{r+1}{n}\log(n+1) - \frac{4}{n}\log(n+1+\tau r)\\ &+ \frac{2}{n}\log\beta(z_0) + \frac{2r}{n}\log(\eta) + \frac{4r}{n}\log\frac{2r}{e(n+1+(\tau+2)r)}.\end{aligned}$$

This quantity vanishes in the limit as $n \to \infty$. □

Proof of Proposition 2.6. Let n_0 be an integer such that $(2r\tau/t_1) \vee (2r\tau/t_2) \leqslant \delta$, with τ as in Lemma 2.15, and let $n \geqslant n_0$. We will show that

$$\widetilde{\Psi}_n(W_n(\mu,\delta)) \subseteq W_{t_1}(\mu_1,\delta_1) \times W_{t_2}(\mu_2,\delta_2). \tag{2.49}$$

Once (2.49) has been proved, we can apply Lemma 2.27 to obtain

$$\begin{aligned}\mathbb{P}_{t_1+1}(W_{t_1}(\mu_1,\delta_1))\mathbb{P}_{t_2+1}(W_{t_2}(\mu_2,\delta_2)) &\geqslant \mathbb{P}_{t_1+1}\otimes\mathbb{P}_{t_2+1}(\widetilde{\Psi}_n(W_n(\mu,\delta)))\\ &\geqslant \widetilde{C}_n\mathbb{P}_{n+1}(W_n(\mu,\delta)).\end{aligned}$$

This gives us (2.7) where the constant $\widetilde{C}_n$ satisfies (2.8). To check that (2.49) holds, we first need to make sure that $\widetilde{\Psi}_n(u)$ is well defined for all $u \in W_n(\mu,\delta)$: since

$$L[u](\Gamma_i) \geqslant \mu(\Gamma_i) - |L[u]-\mu|_{\mathrm{TV}} \geqslant \lambda_i - \delta - \delta, \tag{2.50}$$

we have $M[u](\Gamma_i) \geqslant n(\lambda_i - 2\delta) \geqslant t_i + 1$, thus by Lemma 2.26, $\widetilde{\Psi}_n(u)$ is well defined. We now prove (2.49). Let $u \in W_n(\mu,\delta)$ and $i \in \{1,2\}$. We have to compare $L\left[\widetilde{\Psi}_{i,n}(u)\right]$ and μ_i. We shall first establish a few preliminary bounds. By (2.50), we have

$$L[u]\bigg(\bigcup_{j\in\mathcal{J}_\mu} C_j^2\bigg) \geqslant L[u](\Gamma_1\cup\Gamma_2) \geqslant 1-4\delta.$$

Hence, since $\mathcal{J}_{\mu_1}$ and $\mathcal{J}_{\mu_2}$ are disjoint,

$$\begin{aligned}\left|\lambda_1\mu_1 - L[u]|_{\mathcal{J}_{\mu_1}}\right|_{\mathrm{TV}} + \left|\lambda_2\mu_2 - L[u]|_{\mathcal{J}_{\mu_2}}\right|_{\mathrm{TV}} &= \left|\mu - L[u]|_{\mathcal{J}_{\mu_1}\cup\mathcal{J}_{\mu_2}}\right|_{\mathrm{TV}}\\ &\leqslant \delta + 4\delta,\end{aligned}$$

thus both terms are no larger than 5δ. In particular, we have $L[u]|_{\mathcal{J}_{\mu_i}}(A) \leqslant \lambda_i\mu_i(A) + 5\delta$ for all $A \subseteq S^2$. Together with (2.50), this implies

$$\begin{aligned}\left|L[u]|_{\mathcal{J}_{\mu_i}} - L[u]|_{\mathcal{J}_i}\right|_{\mathrm{TV}} &= L[u]\bigg(\bigcup_{j\in\mathcal{J}_{\mu_i}} C_j^2\bigg) - L[u]\bigg(\bigcup_{j\in\mathcal{J}_i} C_j^2\bigg)\\ &\leqslant \lambda_i + 5\delta - (\lambda_i - 2\delta)\\ &= 7\delta.\end{aligned}$$

We can now compare $L\left[\widetilde{\Psi}_{i,n}(u)\right]$ and μ_i. On the one hand, by Lemma 2.26,

$$\begin{aligned}\left|L\left[\widetilde{\Psi}_{i,n}(u)\right] - \frac{n}{t_i}L[u]|_{\mathcal{J}_i}\right|_{\mathrm{TV}} &\leqslant \frac{2}{t_i}(n-t_1-t_2) + \frac{2r\tau}{t_i}\\ &\leqslant \frac{n}{t_i}16\delta + \delta.\end{aligned}$$

The bound on $n - t_1 - t_2$ is obtained thanks to the choice of δ and n that enabled (2.43). The choice of $n \geqslant n_0$ also enabled the bound on $2r\tau/t_i$. On the other hand,

$$\begin{aligned}\left|\frac{n}{t_i}L[u]|_{\mathcal{J}_i} - \mu_i\right|_{\mathrm{TV}} &\leqslant \frac{n}{t_i}\left|L[u]|_{\mathcal{J}_i} - L[u]|_{\mathcal{J}_{\mu_i}}\right|_{\mathrm{TV}} + \left|\left(\frac{1}{\lambda_i} - \frac{n}{t_i}\right)L[u]|_{\mathcal{J}_{\mu_i}}\right|_{\mathrm{TV}} \\ &\qquad + \frac{1}{\lambda_i}\left|L[u]|_{\mathcal{J}_{\mu_i}} - \lambda_i\mu_i\right|_{\mathrm{TV}} \\ &\leqslant \frac{n}{t_i}7\delta + \frac{4}{\lambda_i^2}\delta + \frac{1}{\lambda_i}5\delta.\end{aligned}$$

The term $4\delta/\lambda_i^2$ in the middle is obtained thanks to (2.43). The two other therms are obtained thanks to the preliminary bounds stated above in the proof. Combining these two bounds yields

$$\begin{aligned}\left|L\left[\widetilde{\Psi}_{i,n}(u)\right] - \mu_i\right|_{\mathrm{TV}} &\leqslant \frac{n}{t_i}23\delta + \frac{4}{\lambda_i^2}\delta + \frac{1}{\lambda_i}5\delta + \delta \\ &\leqslant \frac{96}{\lambda_i^2}\delta + \frac{1}{\lambda_i}28\delta + \delta.\end{aligned}$$

The bound on n/t_i is obtained thanks to (2.43). Roughly bounding both $1/\lambda_i$ and 1 by $1/\lambda_i^2$, we get the bound $125\delta/\lambda_i^2$, which proves (2.49) and completes the proof. □

3 Identification of the rate function

Theorem 1.3 was only partially proven in the previous section; Theorem 2.5 states that $(L_n^{(2)})$ satisfies the weak LDP with rate function $I^{(2)}$, but it remains to prove (1.9). This is the goal of the current section. For convenience, we set $I := I^{(2)}$ in this section.

3.1 Usual rate functions

Let us properly define the functions involved in (1.9). In the following, $\mathbb{E}[\cdot]$ denotes the expectation associated to the law of (X_n). Fix $\mu \in \mathcal{P}(S^2)$. Recall that $\mathcal{M}(S^2)$ denotes the space of finite signed measures on S^2 and that $\mathcal{B}(S^2)$ denotes the set of bounded functions on S^2. In addition, $\mathcal{E}(S^2) := \exp(\mathcal{B}(S^2))$ denotes the set of positive functions on S^2 that are bounded away from 0 and ∞. The supremum norm on $\mathcal{B}(S^2)$ is denoted $\|\cdot\|$.[24] The SCGF $\Lambda = \Lambda^{(2)}$ of $(L_n^{(2)})$ is defined by

$$\Lambda(V) = \limsup_{n\to\infty} \frac{1}{n}\log \mathbb{E}\left[e^{n\langle L_n, V\rangle}\right], \qquad V \in \mathcal{B}(S^2). \tag{3.1}$$

We also define the truncated SCGF $\Lambda_\infty = \Lambda_\infty^{(2)}$ of $(L_n^{(2)})$ by

$$\Lambda_K(V) = \limsup_{n\to\infty} \frac{1}{n}\log \mathbb{E}\left[e^{n\langle L_n, V\rangle}\mathbb{1}_{\mathcal{P}(K^2)}(L_n)\right], \qquad V \in \mathcal{B}(S^2),\ K \subseteq S \text{ finite},$$

and

$$\Lambda_\infty(V) = \sup\{\Lambda_K(V) \mid K \subseteq S,\ |K| < \infty\}.$$

Clearly, $\Lambda_K \leqslant \Lambda_\infty \leqslant \Lambda$. See Example D.2 for a situation in which Λ_∞ and Λ are not equal. Their convex conjugates Λ^* and Λ_∞^* are defined by Definition B.1 and are lower semicontinuous.

As defined in [16], the DV entropy $J = J^{(2)}$ is given by[25]

$$J(\mu) = \sup_{f\in\mathcal{E}(S^2)} \left\langle \mu, \log\frac{f}{Pf}\right\rangle, \qquad \mu \in \mathcal{P}_{\mathrm{bal}}(S^2), \tag{3.2}$$

where[26]

$$Pf(x,y) := P^{(2)}f(x,y) = \sum_{z\in S} p(y,z)f(y,z), \qquad (x,y) \in S^2,\ f \in \mathcal{E}(S^2). \tag{3.3}$$

[24] We do not equip $\mathcal{B}(S^2)$ with the topology induced by $\|\cdot\|$. See Appendix B.

[25] The supremum in (3.2) is often taken over $\{f \in \mathcal{E}(S^2) \mid f \geqslant 1\}$ instead of $\mathcal{E}(S^2)$. In fact, when $f \in \mathcal{E}(S^2)$, replacing f by $f/\inf f$ does not change the value of the bracket in (3.2), hence the two definitions agree.

[26] Any stochastic kernel r on a space $\mathcal{X}$ defines an operator on $\mathcal{E}(\mathcal{X})$ by $rf(x) = \sum_{y\in\mathcal{X}} r(x,y)f(y)$. The quantity $Pf(x,y)$ in (3.3) has the same definition with $\mathcal{X} = S^2$ and r the stochastic kernel of the Markov chain $((X_n, X_{n+1}))_{n\geqslant 1}$, defined by $r((x_1,y_1),(x_2,y_2)) = \mathbb{1}_{\{y_1=x_2\}}p(x_2,y_2)$. Notice that $Pf(x,y)$ does not depend on x.

The function J is the standard (see for instance [16], [17], [11], [28], [10]) DV entropy for the Markov chain $((X_n, X_{n+1}))_{n\geqslant 1}$. The function J is lower semicontinuous as a pointwise supremum of a family of continuous functions.

If $\mu, \nu \in \mathcal{P}(S^k)$ for some $k \in \mathbb{N}$, we denote by $H(\mu|\nu)$ the relative entropy of μ with respect to ν, which is defined by

$$H(\mu|\nu) = \begin{cases} \sum_{u\in S^k} \mu(u) \log \frac{\mu(u)}{\nu(u)}, & \text{if } \mu \ll \nu, \\ \infty, & \text{otherwise.} \end{cases} \tag{3.4}$$

In this definition and throughout the rest of this paper, we adopt the conventions that $0 \times \log 0 = 0$. The fact that $H(\mu|\nu)$ is well defined in $[0, +\infty]$ comes from properties of the real function $s \mapsto s \log s$, namely that it is bounded below and convex. Additional properties of the relative entropy can be found in Chapter 5 of [28]. We use $H(\cdot|\cdot)$ to define a function $R = R^{(2)}$ in the following way. Let $\mu \in \mathcal{P}_{\text{bal}}(S^2)$. When r is a probability kernel on S, we define the probability measure $\mu^{(1)} \otimes r$ on S^2 by $\mu^{(1)} \otimes r(x, y) = \mu^{(1)}(x)r(x, y)$. We define

$$R(\mu) = H(\mu|\mu^{(1)} \otimes p), \quad \mu \in \mathcal{P}_{\text{bal}}(S^2). \tag{3.5}$$

By the variational formula for the relative entropy (see for instance Theorem 5.4 of [28]), we also have, for all $\mu \in \mathcal{P}(S^2)$,

$$R(\mu) = \sup_{V\in\mathcal{B}(S^2)} \big(\langle \mu, V\rangle - \log\langle \mu^{(1)} \otimes p, e^V\rangle\big).$$

This implies in particular that R is lower semicontinuous, since both the functions $\mu \mapsto \langle \mu, V\rangle$ and $\mu \mapsto \langle \mu^{(1)} \otimes p, e^V\rangle$ are continuous for a fixed V. Let q be the stochastic kernel defined by $q(x, y) = \mu(x, y)/\mu^{(1)}(x)$ if $\mu^{(1)}(x) > 0$, and arbitrarily otherwise. It follows that $\mu^{(1)} \otimes q = \mu$. If $\mu \in \mathcal{D}^{(2)}$, then $\mu \ll \mu^{(1)} \otimes p$ and we have the convenient formula

$$R(\mu) = \sum_{x,y\in S} \mu(x, y) \log \frac{q(x, y)}{p(x, y)} = \sum_{x\in S} \mu^{(1)}(x)H(q(x, \cdot)|p(x, \cdot)). \tag{3.6}$$

As mentioned earlier, the convex conjugate of the SCGF, the DV entropy, and the function $R^{(2)}$ are standard in large deviations theory of Markov chains. For convenience, until the end of Section 3, we will omit the exponent and simply write Λ, Λ_∞, J and R instead of $\Lambda^{(2)}$, $\Lambda^{(2)}_\infty$, $J^{(2)}$ and $R^{(2)}$ respectively. We propose a complete identification of I.

Proposition 3.1. *The function I satisfies*

$$I(\mu) = \begin{cases} \Lambda^*_\infty(\mu) = \Lambda^*(\mu) = J(\mu) = R(\mu), & \text{if } \mu \in \mathcal{A}^{(2)}_{\text{bal}}, \\ \infty, & \text{otherwise.} \end{cases} \tag{3.7}$$

Proof. Property 3 of Lemma 2.7 already showed that I is infinite on the complement of $\mathcal{A}^{(2)}_{\text{bal}}$. The description of I over $\mathcal{A}^{(2)}_{\text{bal}}$ requires a more involved proof. The conclusion comes as consequence of several propositions.

- By Proposition 3.4, the function I coincides with its own convex biconjugate I^{**} on $\mathcal{A}^{(2)}$.
- In Proposition 3.8, we show that $I^* = \Lambda_\infty$ on $\mathcal{B}(S^2)$. Hence, by the previous point, we have $I = I^{**} = \Lambda^*_\infty$ on $\mathcal{A}^{(2)}_{\text{bal}}$.
- Finally, Proposition 3.11, establishes that

$$\Lambda^*_\infty(\mu) = \Lambda^*(\mu) = J(\mu) = R(\mu), \qquad \mu \in \mathcal{A}^{(2)}_{\text{bal}}. \tag{3.8}$$

Put together, these equalities yield the conclusion. We prove Propositions 3.4, 3.8, and 3.11 in Sections 3.2, 3.3, and 3.4 below, respectively. ☐

Remark 3.2. Note that Proposition 3.1 does not say that I is finite on $\mathcal{A}^{(2)}_{\text{bal}}$. For instance, if a measure $\mu \in \mathcal{A}^{(2)}_{\text{bal}}$ does not belong to $\mathcal{A}^{(2)}_{\text{bal}} \cap \mathcal{D}^{(2)}$, one can easily check that Equation (3.7) holds with

$$I(\mu) = \Lambda^*_\infty(\mu) = \Lambda^*(\mu) = J(\mu) = R(\mu) = \infty.$$

See Example D.4 for a non-trivial example of a measure that belongs to $\mathcal{A}^{(2)}_{\text{bal}} \cap \mathcal{D}^{(2)}$ while having infinite rate function.

Corollary 3.3. *Let $\mu \in \mathcal{A}^{(2)}_{\mathrm{bal}}$, and let $\tilde{\mu}_j = \mu(C^2_j)^{-1}\mu|_{C^2_j}$ for all $j \in \mathcal{J}_\mu$. We have*

$$I(\mu) = \sum_{j\in\mathcal{J}_\mu} \mu(C^2_j) I(\tilde{\mu}_j). \tag{3.9}$$

Proof. Since the function R satisfies (3.9), this is a direct consequence of Proposition 3.1. $\square$

3.2 The convex biconjugate of I

Let I^{**} denote the convex biconjugate of I, defined on $\mathcal{P}(S^2)$ as in Appendix B. In this section, we prove the following property of the rate function I.

Proposition 3.4. *Let $\mu \in \mathcal{A}^{(2)}$. Then,*

$$I(\mu) = I^{**}(\mu).$$

The identification $I = I^{**}$ does not hold on the full space $\mathcal{P}(S^2)$, but rather on the smaller set of admissible measures. In fact, I^{**} is convex, whereas we do not expect I to be, by Property 3 of Proposition 2.7. Let $\operatorname{co} I$ denote the convex envelope of I, which is the largest convex function under I and satisfies

$$\begin{aligned}\operatorname{co} I(\mu) = \inf \Bigg\{ \sum_{i=1}^k \lambda_i I(\mu_i) \,\Bigg|\, & k \in \mathbb{N},\ (\mu_i) \in \mathcal{P}(S^2)^k,\ (\lambda_i) \in (0,1)^k, \\ & \sum_{i=1}^k \lambda_i = 1,\ \sum_{i=1}^k \lambda_i \mu_i = \mu \Bigg\}, \qquad \mu \in \mathcal{P}(S^2).\end{aligned} \tag{3.10}$$

In the proof of Proposition 3.4, we need properties of I and $\operatorname{co} I$, stated in Lemmas 3.5 and 3.6 below.

Proof. Let $\mu \in \mathcal{A}^{(2)}$. By Corollary B.2, I^{**} is the lower semicontinuous envelope of $\operatorname{co} I$. By Lemma 3.6, $\operatorname{co} I$ agrees with its lower semicontinuous envelope on $\mathcal{A}^{(2)}$, thus $I^{**}(\mu) = \operatorname{co} I(\mu)$. Lemma 3.5 yields the conclusion. $\square$

Lemma 3.5. *Let $\mu \in \mathcal{A}^{(2)}$. Then, $I(\mu) = \operatorname{co} I(\mu)$.*

Lemma 3.6. *Let $\mu \in \mathcal{A}^{(2)}$. Then, the function $\operatorname{co} I$ is lower semicontinuous at μ.*

Proof of Lemma 3.5. By definition, $\operatorname{co} I(\mu) \leqslant I(\mu)$. Let us prove the reverse bound. Let $\delta > 0$. By (3.10), there exist finite sequences $(\lambda_i) \in (0,1)^k$ and $(\mu_i) \in \mathcal{P}(S^2)^k$ such that

$$\sum_{i=1}^k \lambda_i = 1, \quad \sum_{i=1}^k \lambda_i \mu_i = \mu, \quad \operatorname{co} I(\mu) + \delta \geqslant \sum_{i=1}^k \lambda_i I(\mu_i).$$

Each μ_i must be absolutely continuous with respect to μ. Hence, each μ_i is admissible and satisfies $\mathcal{J}_{\mu_i} \subseteq \mathcal{J}_\mu$. By Property 1 of Proposition 2.7 and Jensen's inequality, we have

$$\operatorname{co} I(\mu) + \delta \geqslant I\left(\sum_{i=1}^k \lambda_i \mu_i\right) = I(\mu).$$

This holds for every $\delta > 0$, thus $\operatorname{co} I(\mu) \geqslant I(\mu)$. $\square$

Proof of Lemma 3.6. Let (μ_n) be a sequence of probability measures such that $\mu_n \to \mu$. We have to prove that

$$\liminf_{n\to\infty} \operatorname{co} I(\mu_n) \geqslant \operatorname{co} I(\mu).$$

This is immediate when the left-hand side is infinite. Otherwise, by taking a subsequence if necessary, we can assume that $\operatorname{co} I(\mu_n) < \infty$ for all $n \in \mathbb{N}$ without loss of generality. By (3.10), for all $n \in \mathbb{N}$, there exist finite sequences $(\lambda_{n,i}) \in (0,1)^{k_n}$ and $(\mu_{n,i}) \in \mathcal{P}(S^2)^{k_n}$ such that

$$\sum_{i=1}^{k_n} \lambda_{n,i} = 1, \qquad \sum_{i=1}^{k_n} \lambda_{n,i}\mu_{n,i} = \mu_n, \qquad \operatorname{co} I(\mu_n) + \frac{1}{n} \geqslant \sum_{i=1}^{k_n} \lambda_{n,i} I(\mu_{n,i}).$$

Since $\operatorname{co} I(\mu_n) < \infty$, we must have $I(\mu_{n,i}) < \infty$ for all $i \leqslant k_n$. Thus, by Property 3 of Proposition 2.7, the measure $\mu_{n,i}$ is admissible for all $i \leqslant k_n$. Let D denote the support of μ. For all $i \leqslant k_n$, let $q_{n,i} = \mu_{n,i}(D)$ and

$$\mu'_{n,i} = \begin{cases} \frac{1}{q_{n,i}}\mu_{n,i}|_D, & \text{if } q_{n,i} \neq 0, \\ \mu, & \text{otherwise,} \end{cases} \qquad \mu''_{n,i} = \begin{cases} \frac{1}{1-q_{n,i}}\mu_{n,i}|_{S^2\setminus D}, & \text{if } q_{n,i} \neq 1, \\ \mu, & \text{otherwise.} \end{cases}$$

The measures $\mu'_{n,i}$ and $\mu''_{n,i}$ are admissible. We have $\mu_{n,i} = q_{n,i}\mu'_{n,i} + (1-q_{n,i})\mu''_{n,i}$ and

$$I(\mu_{n,i}) = q_{n,i} I(\mu'_{n,i}) + (1-q_{n,i}) I(\mu''_{n,i}).$$

Indeed, this is immediate if $q_{n,i} \in \{0,1\}$. Otherwise, $q_{n,i} \in (0,1)$ and this is a consequence of Property 2 of Proposition 2.7 with the partition $\mathcal{J}_{\mu_{n,i}} = \big(\mathcal{J}_{\mu_{n,i}} \cap \mathcal{J}_\mu\big) \cup \big(\mathcal{J}_{\mu_{n,i}} \cap \mathcal{J}_\mu^c\big)$, where $\mathcal{J}_\mu^c$ denotes the complement of $\mathcal{J}_\mu$ in $\mathcal{J}$. Let

$$p_n = \sum_{i=1}^{k_n} \lambda_{n,i} q_{n,i} = \mu_n(D).$$

Since $p_n \to 1$ as $n \to \infty$, we can set

$$\lambda'_{n,i} = \frac{\lambda_{n,i} q_{n,i}}{p_n},$$

for n large enough. Then, we have

$$\mu_n = p_n \sum_{i=1}^{k_n} \lambda'_{n,i}\mu'_{n,i} + \sum_{i=1}^{k_n} \lambda_{n,i}(1-q_{n,i})\mu''_{n,i},$$

$$\operatorname{co} I(\mu_n) + \frac{1}{n} \geqslant p_n \sum_{i=1}^{k_n} \lambda'_{n,i} I(\mu'_{n,i}) + \sum_{i=1}^{k_n} \lambda_{n,i}(1-q_{n,i}) I(\mu''_{n,i}) \geqslant p_n \sum_{i=1}^{k_n} \lambda'_{n,i} I(\mu'_{n,i}).$$

For all $i \leqslant k_n$, the measure $\mu'_{n,i}$ is an element of $\{\nu \in \mathcal{A}^{(2)} \mid \mathcal{J}_\nu \subseteq \mathcal{J}_\mu\}$. By Property 1 of Proposition 2.7, since $\sum_{i=1}^{k_n} \lambda'_{n,i} = 1$, we have

$$\operatorname{co} I(\mu_n) + \frac{1}{n} \geqslant p_n I\bigg(\sum_{i=1}^{k_n} \lambda'_{n,i}\mu'_{n,i}\bigg).$$

Moreover,

$$\begin{aligned}
\bigg|\sum_{i=1}^{k_n} \lambda'_{n,i}\mu'_{n,i} - \mu\bigg|_{\mathrm{TV}} &\leqslant \bigg|\sum_{i=1}^{k_n} \lambda'_{n,i}\mu'_{n,i} - \mu_n\bigg|_{\mathrm{TV}} + |\mu - \mu_n|_{\mathrm{TV}} \\
&= \bigg|\sum_{i=1}^{k_n} (\lambda'_{n,i} - \lambda_{n,i}q_{n,i})\mu'_{n,i} - \sum_{i=1}^{k_n} \lambda_{n,i}(1-q_{n,i})\mu''_{n,i}\bigg|_{\mathrm{TV}} + |\mu - \mu_n|_{\mathrm{TV}} \\
&\leqslant \sum_{i=1}^{k_n} (\lambda'_{n,i} - \lambda_{n,i}q_{n,i}) + \sum_{i=1}^{k_n} \lambda_{n,i}(1-q_{n,i}) + |\mu - \mu_n|_{\mathrm{TV}} \\
&= 2(1-p_n) + |\mu - \mu_n|_{\mathrm{TV}}.
\end{aligned}$$

Since $p_n \to 1$ as $n \to \infty$, this proves that $\sum_{i=1}^{k_n} \lambda'_{n,i}\mu'_{n,i} \to \mu$ as $n \to \infty$. By the lower semicontinuity of I and Lemma 3.5, we have

$$\liminf_{n\to\infty} \operatorname{co} I(\mu_n) \geqslant \liminf_{n\to\infty} p_n I\bigg(\sum_{i=1}^{k_n} \lambda'_{n,i}\mu'_{n,i}\bigg) \geqslant I(\mu) = \operatorname{co} I(\mu).$$

□

Remark 3.7. We have described I^{**} over the set $\mathcal{A}^{(2)}$, which is enough for the purpose of the rest of this paper. However, it is fair to wonder whether the identification $I^{**} = \operatorname{co} I$ also holds on the whole space $\mathcal{P}(S^2)$. In full generality, it does not, as illustrated by Example D.2. It turns out that $\operatorname{co} I$ is not always lower semicontinuous on $\mathcal{P}(S^2)$, preventing it to agree with I^{**}. One can actually show that $I^{**} = \sigma\text{-co}\, I$ on $\mathcal{P}(S^2)$, where $\sigma\text{-co}\, I$ is the σ-convex envelope of I. In other words,

$$I^{**}(\mu) = \sigma\text{-co}\, I(\mu) = \inf\left\{ \sum_{m=1}^{\infty} \lambda_m I(\mu_m) \;\middle|\; (\mu_m) \in \mathcal{P}(S^2)^{\mathbb{N}},\, (\lambda_m) \in [0,1]^{\mathbb{N}}, \right.$$
$$\left. \sum_{m=1}^{\infty} \lambda_m = 1,\; \sum_{m=1}^{\infty} \lambda_m \mu_m = \mu \right\}.$$

For all $\mu \in \mathcal{A}^{(2)}$, we have $\sigma\text{-co}\, I(\mu) = \operatorname{co} I(\mu)$.

3.3 A variation of Varadhan's lemma

The functional Λ plays an important role in large deviations theory. In fact, one may expect that $I^* = \Lambda$ by Varadhan's Lemma. However, standard versions of Varadhan's Lemma require the LDP to be full with good rate function, whereas Theorem 2.5 only provides the weak LDP. The purpose of this section is to derive a version of Varadhan's lemma that fits the context of our weak LDP. This requires using the function Λ_∞ instead of Λ.

Proposition 3.8. *For all $V \in \mathcal{B}(S^2)$, $I^*(V) = \Lambda_\infty(V)$.*

Proof. This is a direct consequence of Lemmas 3.9 and 3.10 below. □

Lemma 3.9. *For all $V \in \mathcal{B}(S^2)$, $I^*(V) \geqslant \Lambda_\infty(V)$.*

The proof of Lemma 3.9 is a simple adaptation of Varadhan's upper bound to the weak LDP; see for example Lemma 4.3.6 in [11]. The compactness properties usually obtained from the goodness of the rate function in Varadhan's upper bound are here obtained through the definition of Λ_∞.

Proof. Let $K \subseteq S$ be a finite set and $V \in \mathcal{B}(S^2)$. Let $\delta > 0$. As I is lower semicontinuous and $\langle \cdot, V\rangle$ is continuous, for every $\mu \in \mathcal{P}(S^2)$, there exists an open neighborhood A_μ of μ such that

$$\inf_{\nu \in \overline{A}_\mu} I(\nu) \geqslant I(\mu) - \delta, \qquad \sup_{\nu \in \overline{A}_\mu} \langle \nu, V\rangle \leqslant \langle \mu, V\rangle + \delta.$$

Let $\alpha \in \mathbb{R}$. The set $\mathcal{L} := \{\mu \in \mathcal{P}(K^2) \mid I(\mu) \leqslant \alpha\}$ is a compact subset of $\mathcal{P}(S^2)$, hence there exists a finite sequence $\mu_1, \ldots, \mu_k \in \mathcal{L}$ such that $\mathcal{L} \subseteq \bigcup_{i=1}^k A_{\mu_i}$. Denote by C the complement in $\mathcal{P}(K^2)$ of this union, which is compact. We have

$$\mathcal{P}(K^2) = \left(\bigcup_{i=1}^k \mathcal{P}(K^2) \cap \overline{A}_{\mu_i} \right) \cup C.$$

Therefore we have

$$\begin{aligned} \mathbb{E}\left[e^{n\langle L_n, V\rangle} \mathbb{1}_{\mathcal{P}(K^2)}(L_n)\right] &\leqslant \sum_{i=1}^k \mathbb{E}\left[e^{n\langle L_n, V\rangle} \mathbb{1}_{\mathcal{P}(K^2)\cap \overline{A}_{\mu_i}}(L_n)\right] + \mathbb{E}\left[e^{n\langle L_n, V\rangle} \mathbb{1}_C(L_n)\right] \\ &\leqslant \sum_{i=1}^k e^{n(\langle \mu_i, V\rangle + \delta)} \mathbb{P}(L_n \in \mathcal{P}(K^2) \cap \overline{A}_{\mu_i}) \\ &\qquad + e^{n\|V\|} \mathbb{P}(L_n \in C). \end{aligned} \tag{3.11}$$

We use Theorem 2.5. For all $i \in \{1, \ldots, k\}$, by the weak LDP upper bound applied on the compact set $\mathcal{P}(K^2) \cap \overline{A}_{\mu_i}$,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(L_n \in \mathcal{P}(K^2) \cap \overline{A}_{\mu_i}) = - \inf_{\mu \in \mathcal{P}(K^2)\cap \overline{A}_{\mu_i}} I(\mu) \leqslant -I(\mu_i) + \delta.$$

Therefore,

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n} \log\left(\sum_{i=1}^{k} e^{n(\langle \mu_i, V\rangle + \delta)} \mathbb{P}(L_n \in \mathcal{P}(K^2) \cap A_{\mu_i})\right) &\leqslant \max_{1\leqslant i\leqslant k} (\langle \mu_i, V\rangle + 2\delta - I(\mu_i)) \\
&\leqslant \sup_{\mu\in\mathcal{P}(K^2)} (\langle \mu, V\rangle - I(\mu)) + 2\delta \\
&\leqslant \sup_{\mu\in\mathcal{P}(S^2)} (\langle \mu, V\rangle - I(\mu)) + 2\delta.
\end{aligned}$$

We also handle the last term on the right-hand side of (3.11) with the weak LDP of Theorem 2.5. By definition, $I(\mu) > \alpha$ for all $\mu \in C$, thus the weak LDP upper bound applied on the compact set C yields

$$\limsup_{n\to\infty} \frac{1}{n} \log e^{n\|V\|} \mathbb{P}(L_n \in C) \leqslant \|V\| - \inf_{\mu\in C} I(\mu) \leqslant \|V\| - \alpha.$$

Combining these bounds and (3.11), we get

$$\begin{aligned}
\Lambda_K(V) &= \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[e^{n\langle L_n, V\rangle} \mathbb{1}_{\mathcal{P}(K^2)}(L_n)\right] \\
&\leqslant (\|V\| - \alpha) \vee \left(\sup_{\mu\in\mathcal{P}(S^2)} (\langle \mu, V\rangle - I(\mu)) + 2\delta\right) \\
&= \sup_{\mu\in\mathcal{P}(S^2)} (\langle \mu, V\rangle - I(\mu)) + 2\delta,
\end{aligned}$$

where the last inequality holds for a choice of $\alpha < \infty$ large enough. We get $\Lambda_K(V) \leqslant I^*(V)$ by taking the limit as $\delta \to 0$. Since this holds for any finite set K, we have $\Lambda_\infty(V) \leqslant I^*(V)$. ☐

Lemma 3.10. *For all $V \in \mathcal{B}(S^2)$, $I^*(V) \leqslant \Lambda_\infty(V)$.*

The lower bound $I^* \leqslant \Lambda$ of Varadhan's lemma holds even in the context of the weak LDP because it only relies on the LDP lower bound (see Lemma 4.3.4 of [11]). In order to obtain the stronger bound $I^* \leqslant \Lambda_\infty$, we follow the same approach. For this bound, we shall use the formula of Property 4 in Proposition 2.7.

Proof. Let $V \in \mathcal{B}(S^2)$, and let $\mu \in \mathcal{P}(S^2)$. Fix $\delta > 0$ and let K be a finite subset of S. There exists $r > 0$, independent of K, such that the open ball $\mathcal{B}(\mu, r)$ satisfies

$$\inf_{\mathcal{B}(\mu,r)\cap\mathcal{P}(K^2)} \langle \cdot, V\rangle \geqslant \inf_{\mathcal{B}(\mu,r)} \langle \cdot, V\rangle \geqslant \langle \mu, V\rangle - \delta.$$

Therefore, by the definition of $\underline{s}$ (see Definition 2.2),

$$\begin{aligned}
\Lambda_K(V) &\geqslant \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[e^{n\langle L_n, V\rangle} \mathbb{1}_{\mathcal{B}(\mu,r)\cap\mathcal{P}(K^2)}(L_n)\right] \\
&\geqslant \inf_{\mathcal{B}(\mu,r)\cap\mathcal{P}(K^2)} \langle \cdot, V\rangle + \underline{s}(\mathcal{B}(\mu, r) \cap \mathcal{P}(K^2)) \\
&\geqslant \langle \mu, V\rangle - \delta + \underline{s}(\mathcal{B}(\mu, r) \cap \mathcal{P}(K^2)).
\end{aligned} \tag{3.12}$$

Notice that, since K is independent of r, the supremum of the right-hand side of (3.12) over all finite $K \subseteq S$ is $\langle \mu, V\rangle - \delta + \underline{s}_\infty(\mathcal{B}(\mu, r))$, where $\underline{s}_\infty$ is as in Property 4 of Proposition 2.7. Therefore, by Property 4 of Proposition 2.7,

$$\Lambda_\infty(V) \geqslant \langle \mu, V\rangle - \delta + \underline{s}_\infty(\mathcal{B}(\mu, r)) \geqslant \langle \mu, V\rangle - \delta - I(\mu).$$

The proof is completed by taking the limit as $\delta \to 0$ and the supremum over all $\mu \in \mathcal{P}(S^2)$. ☐

3.4 Alternative expressions of the rate function

The goal of this section is to prove the chain of identities (3.8). We will need some properties of balanced measures stated in Appendix A.1. We will also need the following expressions for Λ and

Λ_∞. Fix a bounded function $V \in \mathcal{B}(S^2)$ and a finite set $K \subseteq S$. Define two operators P^V and P_K^V by

$$P^V f(x,y) = e^{V(x,y)} Pf(x,y), \qquad P_K^V(x,y) = \mathbb{1}_{K^2}(x,y) e^{V(x,y)} Pf(x,y), \qquad f \in \mathcal{E}(S^2),$$

where Pf is as in (3.3). These operators satisfy

$$\langle \beta^{(2)}, (P^V)^n 1\rangle = \mathbb{E}\left[e^{n\langle L_n, V\rangle}\right], \qquad \langle \beta^{(2)}, (P_K^V)^n 1\rangle = \mathbb{E}\left[e^{n\langle L_n, V\rangle} \mathbb{1}_{\mathcal{P}(K^2)}(L_n)\right],$$

where $\beta^{(2)} := \beta \otimes p$. It follows that

$$\Lambda(V) = \limsup_{n\to\infty} \frac{1}{n} \log\langle \beta^{(2)}, (P^V)^n 1\rangle, \qquad \Lambda_K(V) = \limsup_{n\to\infty} \frac{1}{n} \log\langle \beta^{(2)}, (P_K^V)^n 1\rangle. \tag{3.13}$$

Proposition 3.11. *Equality (3.8) holds. In other words, $\Lambda_\infty^*(\mu) = \Lambda^*(\mu) = J(\mu) = R(\mu)$ for all $\mu \in \mathcal{A}_{\text{bal}}^{(2)}$.*

Proof. The identity $J(\mu) = R(\mu)$ holds in all generality and is proved in Theorem 13.1 of [28], where I is our J, $\nu_{\mathcal{Y}}$ denotes the first marginal of a probability measure ν, and κ is the second marginal of ν. In this case, the condition $\nu_{\mathcal{Y}} = \kappa$ corresponds to $\nu \in \mathcal{P}_{\text{bal}}(S^2)$, and $H(\nu|Q)$ becomes our R. We do not repeat the proof here.

Another inequality that does not need further development is $\Lambda_\infty \geqslant \Lambda$, which is immediate since $\Lambda \geqslant \Lambda_\infty$ by definition. As a consequence, in order to complete the proof, it suffices to prove that, for all $\mu \in \mathcal{A}_{\text{bal}}^{(2)}$,

$$\Lambda^*(\mu) \geqslant J(\mu) \geqslant \Lambda_\infty^*(\mu). \tag{3.14}$$

We prove this chain of inequalities in three steps. The first step is to show $\Lambda^*(\mu) \geqslant J(\mu)$, the second step is to show $J(\mu) \geqslant \Lambda_\infty^*(\mu)$ under assumption that μ has a finite support, and the third step consists in approximating μ by finitely supported measures in the general case to obtain $J(\mu) \geqslant \Lambda_\infty^*(\mu)$ for arbitrary $\mu \in \mathcal{A}_{\text{bal}}^{(2)}$.

Step 1. We first show that $\Lambda^*(\mu) \geqslant J(\mu)$.[27] Let $f \in \mathcal{E}(S^2)$. The function $V = \log \frac{f}{Pf}$ is bounded and satisfies $(P^V)^n f = f$. Hence, as f is bounded above and below, Equation (3.13) yields

$$\begin{aligned}
\Lambda(V) &= \limsup_{n\to\infty} \frac{1}{n} \log\langle \beta^{(2)}, (P^V)^n 1\rangle \\
&= \limsup_{n\to\infty} \frac{1}{n} \log\langle \beta^{(2)}, (P^V)^n f\rangle \\
&= \limsup_{n\to\infty} \frac{1}{n} \log\langle \beta^{(2)}, f\rangle = 0.
\end{aligned}$$

Therefore,

$$\Lambda^*(\mu) \geqslant \langle \mu, V\rangle = \left\langle \mu, \log\frac{f}{Pf}\right\rangle.$$

This holds true for any $f \in \mathcal{E}(S^2)$, hence we get $\Lambda^*(\mu) \geqslant J(\mu)$.

Step 2. We prove $J(\mu) \geqslant \Lambda_\infty^*(\mu)$ under the assumption that μ has finite support.[28] In this case, there exists a finite $K \subseteq S$ satisfying $\mu(K^2) = 1$. Without loss of generality, since $\mu \in \mathcal{A}^{(2)}$, we can assume that K^2 is contained in the union of C_j^2 with $j \in \mathcal{J}_\mu$, hence any state of K is reachable form β. For any $g \in \mathcal{E}(S^2)$, we have $\langle \beta^{(2)}, P^n g\rangle = \langle \beta^{(2)} P^n, g\rangle$, where νP is defined on $\mathcal{P}(S^2)$ by $\nu P(x,y) = \sum_{x'\in S} \nu(x', x) p(x,y)$. The measure $\beta^{(2)} P^n$ is supported by pairs of states (x,y) such that x (resp. y) is reachable from β in n steps (resp. $n+1$ steps). Fix $V \in \mathcal{B}(S^2)$, and let $\lambda > \Lambda_\infty(V) \geqslant \Lambda_K(V)$. We define the function

$$f = \sum_{n=0}^{\infty} e^{-\lambda n} (P_K^V)^n 1,$$

[27] The argument of Step 1 is standard. See for instance Lemma 4.1.36 of [13] or Theorem 4.1 of [10]

[28] The method used in Step 2 is a variation of a common argument. See for instance Lemma 4.1.36 of [13], Lemma 3.3 of [22], or Theorem 4.1 of [10]. The big difference is the definition of the SCGF involved in the inequality with J, since Λ_∞ is not mentioned by these references. See also our proof of Proposition 4.4.

and we show that $f \in \mathcal{E}(S^2)$. Since f is bounded below by 1 and coincides with 1 outside K^2, it suffices to show that f is finite on K^2. By (3.13), we have

$$\limsup_{n\to\infty} \frac{1}{n} \log\langle \beta^{(2)}, e^{-n\lambda}(P_K^V)^n 1\rangle = \Lambda_K(V) - \lambda < 0,$$

thus $\langle \beta^{(2)}, f\rangle < \infty$. It follows that f is finite on the support of $\beta^{(2)}$. Moreover, for all $(x, y) \in K^2$,

$$Pf(x,y) = e^{-V(x,y)} P_K^V f(x,y) = e^{\lambda - V(x,y)}(f(x,y) - 1). \tag{3.15}$$

It follows that f is also finite on the support of $\beta^{(2)}P$ because

$$\begin{aligned}\langle \beta^{(2)}P, f\rangle &= \beta^{(2)}((K^2)^c) + \langle \beta^{(2)}P, f\mathbb{1}_{K^2}\rangle \\ &= \beta^{(2)}((K^2)^c) + \langle \beta^{(2)}, Pf\mathbb{1}_{K^2}\rangle \\ &\leqslant 1 + e^{\lambda + \|V\|}\langle \beta^{(2)}, f\rangle < \infty.\end{aligned}$$

By induction, f is finite on the support of $\beta^{(2)}P^n$ for all n, thus f is finite on the set of all pairs reachable from β. Therefore f is finite on K^2 and we can conclude that $f \in \mathcal{E}(S^2)$. Therefore, by definition of $J(\mu)$ and (3.15),

$$\begin{aligned} J(\mu) \geqslant \left\langle \mu, \log\frac{f}{Pf}\right\rangle &= \sum_{(x,y)\in K^2} \mu(x,y) \log\frac{f(x,y)}{Pf(x,y)} \\ &= \sum_{(x,y)\in K^2} \mu(x,y)(V(x,y) - \lambda) + \sum_{(x,y)\in K^2} \mu(x,y)\log\frac{f(x,y)}{f(x,y)-1} \\ &\geqslant \langle \mu, V\rangle - \lambda + 0. \end{aligned}$$

The inequality $J(\mu) \geqslant \langle \mu, V\rangle - \lambda$ holds true for all $\lambda > \Lambda_\infty(V)$, hence $J(\mu) \geqslant \langle \mu, V\rangle - \Lambda_\infty(V)$. Therefore, $J(\mu) \geqslant \Lambda_\infty^*(\mu)$.

At this point, we have proven (3.14) – and thus (3.8) – on the set of measures of $\mathcal{A}_{\text{bal}}^{(2)}$ that have finite support.

Step 3. We turn to the general case, where the support of μ is no longer assumed to be finite. Consider the sequence (μ_n) given by Lemma A.2 in Appendix A.1. By Property 1 of Lemma A.2, each μ_n has finite support and belongs to $\mathcal{A}_{\text{bal}}^{(2)}$ (the admissibility of μ_n follows from the admissibility of μ, since μ_n is absolutely continuous with respect to μ). Hence, by the previous step, we have $R(\mu_n) = J(\mu_n) = \Lambda^*(\mu_n)$. By Property 2 of Lemma A.2 and the lower semicontinuity of Λ^*, we have

$$J(\mu) = R(\mu) = \lim_{n\to\infty} R(\mu_n) = \lim_{n\to\infty} \Lambda^*(\mu_n) \geqslant \Lambda^*(\mu).$$

This completes the proof. □

Remark 3.12. Step 3 of the proof of Proposition 3.11 is one reason why we had to work with pair measures from the beginning. Indeed, Lemma A.2 only works with pair measures, and has no counterpart with the standard occupation measures $(L_n^{(1)})$.

4 Level-2 weak LDP

4.1 A weak contraction principle

In order to deduce the weak LDP for $(L_n^{(1)})$ from that of $(L_n^{(2)})$, we will use a variation of the contraction principle, adapted from the usual one and stated in Lemma 4.1 below. It is comparable to Theorem 3.3 of [24].[29]

[29]Theorem 3.3 of [24], uses the assumption that $\mathcal{X}$ and $\mathcal{Y}$ are locally compact to prove that the contracted rate function is lower semicontinuous. In our Lemma 4.1, we drop this assumption and replace it with the hypothesis that they are metric.

Lemma 4.1 (Weak contraction principle). *Let $\mathcal{X}$ and $\mathcal{Y}$ be metric spaces, and (A_n) a sequence of random variables taking values in $\mathcal{X}$ and satisfying the weak LDP with rate function $I_{\mathcal{X}}$. Let $F : \mathcal{X} \to \mathcal{Y}$ be a continuous function such that $F^{-1}(K)$ is compact for every compact set K of $\mathcal{Y}$. Then, $(F(A_n))$ satisfies the weak LDP in $\mathcal{Y}$ with rate function $I_{\mathcal{Y}}$ defined by*

$$I_{\mathcal{Y}}(y) = \inf_{x \in F^{-1}(y)} I_{\mathcal{X}}(x).$$

Proof. We only have to ensure that $I_{\mathcal{Y}}$ is lower semicontinuous. Once this is proved, it suffices to reproduce any proof of a full contraction principle while replacing the word "closed" with "compact". See for instance Theorem III.20 of [12] or Theorem 4.1.2 of [11].

Let (y_n) be a sequence of elements of $\mathcal{Y}$ converging to some $y \in \mathcal{Y}$. We must show that

$$I_{\mathcal{Y}}(y) \leqslant \liminf_{n\to\infty} I_{\mathcal{Y}}(y_n). \tag{4.1}$$

If the right-hand side of (4.1) is infinite, this is clear. Otherwise, consider a subsequence $(y_{\varphi(n)})$ satisfying $I_{\mathcal{Y}}(y_{\varphi(n)}) < \infty$ for all n and such that $\liminf_{n\to\infty} I_{\mathcal{Y}}(y_n) = \lim_{n\to\infty} I_{\mathcal{Y}}(y_{\varphi(n)})$. Then $y_{\varphi(n)} \to y$ as $n \to \infty$. By definition of $I_{\mathcal{Y}}$, there exists a sequence (x_n) of elements of $\mathcal{X}$ such that, for all n, $F(x_n) = y_{\varphi(n)}$ and

$$I_{\mathcal{X}}(x_n) \leqslant I_{\mathcal{Y}}(y_{\varphi(n)}) + \frac{1}{n}.$$

The sequence (x_n) takes values in the compact set $F^{-1}\left(\{y_{\varphi(n)} \mid n \in \mathbb{N}\} \cup \{y\}\right)$. Hence it has a subsequence $(x_{\psi(n)})$ that converges to a certain $x \in \mathcal{X}$ as $n \to \infty$. Since F is continuous, $F(x) = y$. The function $I_{\mathcal{X}}$ being lower semicontinuous, we get

$$\begin{aligned} I_{\mathcal{Y}}(y) &\leqslant I_{\mathcal{X}}(x) \\ &\leqslant \liminf_{n\to\infty} I_{\mathcal{X}}(x_{\psi(n)}) \\ &\leqslant \liminf_{n\to\infty} \left(I_{\mathcal{Y}}(y_{\psi\circ\varphi(n)}) + \frac{1}{n}\right) \\ &= \lim_{n\to\infty} I_{\mathcal{Y}}(y_{\varphi(n)}) \\ &= \liminf_{n\to\infty} I_{\mathcal{Y}}(y_n). \end{aligned}$$

□

Remark 4.2. It is common to derive the full level-1 LDP as a corollary of the full level-2 LDP. For this, the proof would consist in applying Lemma 4.1, or another variant of the contraction principle to the continuous function $F : \mu \mapsto \langle \mu^{(1)}, f\rangle$, where $f : S \to R^d$ is any function. However, this is not possible in general. For instance, in Example D.5, the level-1 weak LDP fails even though the level-2 weak LDP holds.

In fact, the route based on Lemma 4.1 is successful only in the case of a finite S,[30] which is detailed in Appendix C.

4.2 Weak LDP for the occupation times

In this section, we aim to prove Theorem 1.4, namely, to show that $(L_n^{(1)})$ satisfies the weak LDP with rate function $I^{(1)}$ satisfying (1.10). Let us define the functions involved in (1.10). The function $\Lambda^{(1)}$ is the SCGF of $(L_n^{(1)})$, defined as

$$\Lambda^{(1)}(V) = \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[e^{n\langle L_n^{(1)}, V\rangle}\right], \qquad V \in \mathcal{B}(S). \tag{4.2}$$

Its convex conjugate is defined by Definition B.1. The kernel p defines an operator on $\mathcal{B}(S)$ by $pf(x) = \sum_{y\in S} p(x,y) f(y)$. For any function $V \in \mathcal{B}(S)$, let p^V denote the operator defined by $p^V f(x) = e^{V(x)} pf(x)$. Like in (3.13),

$$\langle \beta, (p^V)^n 1\rangle = \mathbb{E}_\beta\left[e^{n\langle L_n^{(1)}, V\rangle}\right],$$

[30]One can show that the critical assumption that $F^{-1}(K)$ is compact for every compact set of $\mathcal{Y}$ in Lemma 4.1 is satisfied if and only if S is finite.

and $\Lambda^{(1)}$ satisfies the equation

$$\Lambda^{(1)}(V) = \limsup_{n\to\infty} \frac{1}{n} \log\langle \beta, (p^V)^n 1\rangle. \tag{4.3}$$

Let $\mathcal{E}(S)$ denote the set of positive functions on S that are bounded away from 0 and ∞. The DV entropy $J^{(1)}$ is defined by

$$J^{(1)}(\mu) = \sup_{f\in\mathcal{E}(S)} \left\langle \mu, \log\frac{f}{pf}\right\rangle = \sup_{f\in\mathcal{E}(S)} \sum_{x\in S} \mu(x) \log\frac{f(x)}{pf(x)}, \qquad \mu \in \mathcal{P}(S). \tag{4.4}$$

The functional $R^{(1)}$ is defined on $\mathcal{P}(S)$ as the contraction of the function $R^{(2)}$ defined in (3.5):

$$R^{(1)}(\mu) = \inf\{R^{(2)}(\nu) \mid \nu \in \mathcal{P}_{\mathrm{bal}}(S^2),\ \nu^{(1)} = \mu\}, \qquad \mu \in \mathcal{P}(S). \tag{4.5}$$

All these functions coincide on $\mathcal{A}^{(1)}$, as proved in Proposition 4.4 below.

Let $\pi : \mathcal{P}_{\mathrm{bal}}(S^2) \to \mathcal{P}(S)$ be the map defined by $\pi(\mu) = \mu^{(1)}$. It is continuous with respect to the TV metric. Our strategy to prove Theorem 1.4 consists in applying the contraction principle of Lemma 4.1 to π, and use Proposition 4.4 to make the rate function explicit. Lemma 4.3 and Proposition 4.4 are stated and proved below.

Proof of Theorem 1.4. By Lemma 4.3, $(L_n^{(1)})$ satsifies a weak LDP with rate function given by (4.6). It remains to show (1.10). Let $\mu \in \mathcal{P}(S)$. If $\mu \notin \mathcal{A}^{(1)}$, by Proposition A.6, there are no $\nu \in \mathcal{A}^{(2)}_{\mathrm{bal}}$ such that $\mu = \nu^{(1)}$, therefore by (4.6), we have $I^{(1)}(\mu) = \infty$. If $\mu \in \mathcal{A}^{(1)}$, by (4.6) and the equality $I^{(2)} = R^{(2)}$ of Theorem 1.3, we have $I^{(1)}(\mu) = R^{(1)}(\mu)$. The remaining equalities follow from Proposition 4.4. ☐

Lemma 4.3. *The sequence $(L_n^{(1)})$ satisfies a weak LDP in $\mathcal{P}(S)$ with rate function $I^{(1)}$ given by*

$$I^{(1)}(\mu) = \inf\{I^{(2)}(\nu) \mid \nu \in \mathcal{P}_{\mathrm{bal}}(S^2),\ \nu^{(1)} = \mu\}. \tag{4.6}$$

Proof. We only have to apply Lemma 4.1 to π. Since π is defined on $\mathcal{P}_{\mathrm{bal}}(S^2)$, we shall work only with balanced measures. In order to do so, we set

$$\tilde{L}_n^{(2)} = L_n^{(2)} - \frac{1}{n}\delta_{(X_n,X_{n+1})} + \frac{1}{n}\delta_{(X_n,X_1)} \in \mathcal{P}_{\mathrm{bal}}(S^2).$$

We have $L_n^{(1)} = \pi(\tilde{L}_n^{(2)})$. Moreover, for all $\mu \in \mathcal{P}(S^2)$, all $\delta > 0$ and all $n \geqslant 2/\delta$, since the TV distance between $L_n^{(2)}$ and $\tilde{L}_n^{(2)}$ is at most $\frac{2}{n}$, we have

$$\mathbb{P}(L_n^{(2)} \in \mathcal{B}(\mu,\delta)) \leqslant \mathbb{P}(\tilde{L}_n^{(2)} \in \mathcal{B}(\mu,2\delta)) \leqslant \mathbb{P}(L_n^{(2)} \in \mathcal{B}(\mu,3\delta)).$$

Thus,

$$\underline{s}(\mu) \leqslant \lim_{\delta\to 0}\liminf_{n\to\infty} \frac{1}{n}\log\mathbb{P}(\tilde{L}_n^{(2)} \in \mathcal{B}(\mu,2\delta)) \leqslant \lim_{\delta\to 0}\limsup_{n\to\infty} \frac{1}{n}\log\mathbb{P}(\tilde{L}_n^{(2)} \in \mathcal{B}(\mu,2\delta)) \leqslant \overline{s}(\mu),$$

where $\underline{s}$ and $\overline{s}$ are as in Definition 2.2. Since $\underline{s} = \overline{s}$ by Theorem 2.5, $(\tilde{L}_n^{(2)})$ has a RL function that coincides with $s = \underline{s} = \overline{s}$. Therefore, the sequences $(L_n^{(2)})$ and $(\tilde{L}_n^{(2)})$ satisfy the same weak LDP by Lemma 2.3. It only remains to check that π satisfies the assumptions of Lemma 4.1. By Prokhorov's theorem any set $K \subseteq \mathcal{P}(S)$ is compact if and only if it is tight. Let $\varepsilon > 0$ and let K be compact. Then, by tightness, there exists a finite set $S_1 \subseteq S$ such that every $\mu \in K$ satisfies $\mu(S_1^c) \leqslant \varepsilon$. Let $\nu \in \pi^{-1}(K) \subseteq \mathcal{P}_{\mathrm{bal}}(S^2)$. Then,

$$\nu((S_1 \times S_1)^c) = \nu((S_1^c \times S) \cup (S \times S_1^c)) \leqslant \nu^{(1)}(S_1^c) + \nu^{(1)}(S_1^c) \leqslant 2\varepsilon.$$

Therefore $\pi^{-1}(K)$ is tight, and thus compact. Thus, by Theorem 1.3 and Lemma 4.1, $(L_n^{(1)})$ satisfies the weak LDP with rate function $I^{(1)}$ defined by (4.6). ☐

Proposition 4.4. *Let $\mu \in \mathcal{A}^{(1)}$. Then,*

$$(\Lambda^{(1)})^*(\mu) = J^{(1)}(\mu) = R^{(1)}(\mu).$$

Proof. The equality $J^{(1)}(\mu) = R^{(1)}(\mu)$ holds in all generality in $\mathcal{P}(S)$. For the proof of this equality, we refer the reader to Theorem 13.1 of [28]. To apply Equation (13.5) of [28] in our context, take $I_{\mathcal{X}}$ as our $J^{(1)}$ and I as our $J = J^{(2)}$. Then, Equation (13.5) yields

$$J^{(1)}(\mu) = \inf_{\nu \in \pi^{-1}(\mu)} J^{(2)}(\nu) = \inf_{\nu \in \pi^{-1}(\mu)} R^{(2)}(\nu) = R^{(1)}(\mu).$$

Let us show that $(\Lambda^{(1)})^*(\mu) = J^{(1)}(\mu)$. The first inequality $(\Lambda^{(1)})^*(\mu) \geqslant J^{(1)}(\mu)$ is handled as usual:[31] for any $f \in \mathcal{E}(S)$, we set $V = \log \frac{f}{pf} \in \mathcal{B}(S)$, which satisfies $p^V f = f$. Thus, as in the first step of the proof of Proposition 3.11, by (4.3), we have

$$\begin{aligned}\Lambda^{(1)}(V) &= \limsup_{n\to\infty} \frac{1}{n} \log\langle \beta, (p^V)^n 1\rangle \\ &= \limsup_{n\to\infty} \frac{1}{n} \log\langle \beta, (p^V)^n f\rangle \\ &= \limsup_{n\to\infty} \frac{1}{n} \log\langle \beta, f\rangle = 0.\end{aligned}$$

Therefore,

$$(\Lambda^{(1)})^*(\mu) \geqslant \langle \mu, V\rangle = \Big\langle \mu, \log \frac{f}{pf} \Big\rangle,$$

hence $(\Lambda^{(1)})^*(\mu) \geqslant J^{(1)}(\mu)$. We turn to the converse inequality.[32] Let S_β be the set of elements $x \in S$ such that $\beta \rightsquigarrow x$. Since $\mu \in \mathcal{A}^{(1)}$, we have $\operatorname{supp}\mu \subseteq S_\beta$. For every $V \in \mathcal{B}(S)$, we define

$$\widetilde{\Lambda}^{(1)}(V) = \sup_{x\in S_\beta} \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{E}_x \left[e^{n\langle L_n^{(1)}, V\rangle} \right],$$

where $\mathbb{E}_x[\cdot]$ denotes the expectation with respect to the law of the Markov chain with kernel p and initial measure δ_x. We will show that

$$J^{(1)}(\mu) \geqslant (\widetilde{\Lambda}^{(1)})^*(\mu) \geqslant (\Lambda^{(1)})^*(\mu). \tag{4.7}$$

For the second inequality, it suffices to show that $\Lambda^{(1)}(V) \geqslant \widetilde{\Lambda}^{(1)}(V)$ for all $V \in \mathcal{B}(S)$. Let $V \in \mathcal{B}(S)$ and $\delta > 0$. There exists $x \in S_\beta$ such that

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{E}_x \left[e^{\langle L_n^{(1)}, V\rangle} \right] \geqslant \widetilde{\Lambda}^{(1)}(V) - \delta.$$

Since $\beta \rightsquigarrow x$, there exists a finite sequence $\operatorname{supp}\beta \ni x_1, x_2, \dots, x_k = x$ satisfying $\mathbb{P}_k(x_1 \dots x_k) > 0$. By the Markov property, we get

$$\begin{aligned}\mathbb{E}\left[e^{\langle L_n^{(1)}, V\rangle} \right] &\geqslant \mathbb{E}\left[e^{\langle L_n^{(1)}, V\rangle} \mathbb{1}_{\{X_1 = x_1, \dots, X_k = x_k\}} \right] \\ &= \mathbb{E}_x \left[e^{\langle L_{n-k+1}^{(1)}, V\rangle} \right] e^{V(x_1) + \ldots + V(x_{k-1})} \mathbb{P}_k(x_1 \dots x_k).\end{aligned}$$

Hence,

$$\Lambda^{(1)}(V) \geqslant \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{E}_x \left[e^{\langle L_{n-k+1}^{(1)}, V\rangle} \right] + 0 \geqslant \widetilde{\Lambda}^{(1)}(V) - \delta.$$

This holds for any $\delta > 0$, thus $\Lambda^{(1)}(V) \geqslant \tilde{\Lambda}^{(1)}(V)$.[33] We turn to the first inequality in (4.7). We want to optimize the bracket in the definition of $J^{(1)}$ in (4.4). Since $\operatorname{supp}\mu \subseteq S_\beta$, we can consider functions of $\mathcal{E}(S_\beta)$ instead of functions of $\mathcal{E}(S)$ in (4.4). Let $V \in \mathcal{B}(S)$ and $\lambda > \widetilde{\Lambda}^{(1)}(V)$. Let

$$f_n(x) = \sum_{k=0}^{n} e^{-\lambda k} (p^V)^k 1(x), \qquad x \in S_\beta.$$

[31] See the first step of the proof of Proposition 3.11, or Lemma 4.1.36 of [13] and Theorem 4.1 of [10]

[32] To prove the converse inequality, we wish to reproduce the proof of Proposition 3.11, *i.e.* to find a quantity α such that $(\Lambda^{(1)})^*(\mu) \leqslant \alpha \leqslant J^{(1)}(\mu)$. In the proof of Proposition 3.11, this quantity was $(\Lambda_\infty^{(2)})^*(\mu)$, satisfying $(\Lambda^{(2)})^*(\mu) \leqslant (\Lambda_\infty^{(2)})^*(\mu) \leqslant J^{(2)}(\mu)$. However, due to the lack of counterpart of Lemma A.2 in $\mathcal{P}(S)$, the argument here to handle measures with infinite support is slightly different. This argument is more standard; see [22]. The same argument could be used to prove that $(\Lambda^{(2)})^*(\mu) \leqslant J^{(2)}(\mu)$ in Proposition 3.11, but it would not provide any equality implying $(\Lambda_\infty^{(2)})^*(\mu)$.

[33] Lemma 3.3 of [22] provides an alternative proof of this inequality. This inequality can be strict; see Example D.2.

Since the sum is finite, we have $f_n \in \mathcal{E}(S_\beta)$, with the bound $f_n \geqslant f_0 = 1$. The sequence (f_n) converges pointwise over S_β because, for all $x \in S_\beta$,

$$\begin{aligned}\limsup_{k\to\infty} \frac{1}{k} \log\left(e^{-k\lambda}(p^V)^k 1(x)\right) &= \limsup_{k\to\infty} \frac{1}{k} \log\left(e^{-k\lambda}\mathbb{E}_x\left[e^{n\langle L_n^{(1)}, V\rangle}\right]\right) \\ &\leqslant \widetilde{\Lambda}^{(1)}(V) - \lambda \\ &< 0.\end{aligned}$$

We call f the pointwise limit of (f_n) on S_β. It is not necessarily bounded so we cannot use it in the optimization of the bracket in (4.4). However, for every $n \geqslant 1$, we have

$$J^{(1)}(\mu) \geqslant \left\langle \mu, \log \frac{f_n}{pf_n} \right\rangle = \sum_{x\in S_\beta} \mu(x) \log \frac{f_n(x)}{pf_n(x)}.$$

Moreover, we have, for all $n \in \mathbb{N}$,

$$pf_n = e^{-V} p^V f_n = e^{V-\lambda}(f_{n+1} - 1).$$

Furthermore, for all $n \in \mathbb{N}$,

$$f_{n+1} - f_n = e^{-\lambda(n+1)}(p^V)^{n+1} 1 \leqslant e^{-\lambda + \|V\|} e^{-\lambda n}(p^V)^n 1 \leqslant e^{-\lambda+\|V\|} f_n,$$

thus

$$\frac{f_n}{pf_n} = e^{\lambda - V} \frac{f_n}{f_{n+1} - 1} \geqslant e^{\lambda - V} \frac{f_n}{(1 + e^{-\lambda + \|V\|}) f_n - 1} \geqslant 1.$$

Thus, by Fatou's lemma,

$$\begin{aligned}J^{(1)}(\mu) &\geqslant \liminf_{n\to\infty} \sum_{x\in S_\beta} \mu(x) \log \frac{f_n(x)}{pf_n(x)} \\ &\geqslant \sum_{x\in S_\beta} \mu(x) \liminf_{n\to\infty} \log \frac{f_n(x)}{pf_n(x)} \\ &= \langle \mu, V\rangle - \lambda + \sum_{x\in S_\beta} \mu(x) \lim_{n\to\infty} \log \frac{f_n(x)}{f_{n+1}(x) - 1} \\ &= \langle \mu, V\rangle - \lambda + \sum_{x\in S_\beta} \mu(x) \log \frac{f(x)}{f(x) - 1} \geqslant \langle \mu, V\rangle - \lambda.\end{aligned}$$

This holds[34] for every $\lambda > \widetilde{\Lambda}^{(1)}$, thus $J^{(1)}(\mu) \geqslant \langle \mu, V\rangle - \widetilde{\Lambda}^{(1)}$ and ultimately $J^{(1)}(\mu) \geqslant (\widetilde{\Lambda}^{(1)})^*(\mu)$. The proof is complete. □

Before leaving this section, let us add a corollary of Lemma 4.3 as a remark about $I^{(1)}$. Like $I^{(2)}$, the function $I^{(1)}$ can be decomposed over irreducible classes.

Corollary 4.5. *Let $\mu \in \mathcal{A}^{(1)}$ and let $\tilde{\mu}_j = \mu(C_j)^{-1}\mu|_{C_j}$ for all $j \in \mathcal{J}_\mu$. We have*

$$I^{(1)}(\mu) = \sum_{j\in\mathcal{J}_\mu} \mu(C_j) I^{(1)}(\tilde{\mu}_j).$$

Proof. We use the expression of $I^{(1)}(\mu)$ provided by (4.6) in Lemma 4.3. The set $\pi^{-1}(\mu)$ is exactly the set of all $\nu \in \mathcal{A}^{(2)}_{\text{bal}}$ such that $\mathcal{J}_\nu = \mathcal{J}_\mu$ and for all $j \in \mathcal{J}_\mu$, $\nu(C_j^2) = \mu(C_j)$ and

[34] The equality line holds because f is never infinite. If there exists x in supp mu such that $f(x) = 1$, the sum is infinite and the final inequality still holds.

$\pi(\mu(C_j)^{-1}\nu|_{C_j^2}) = \tilde{\mu}_j$. Therefore, by Corollary 3.3,

$$\begin{aligned} I^{(1)}(\mu) &= \inf\left\{ \sum_{j\in\mathcal{J}_\mu} \mu(C_j) I^{(2)}\left(\frac{1}{\mu(C_j)}\nu|_{C_j^2}\right) \;\middle|\; \nu \in \pi^{-1}(\mu)\right\} \\ &= \inf\left\{ \sum_{j\in\mathcal{J}_\mu} \mu(C_j) I^{(2)}(\nu_j) \;\middle|\; \forall j \in \mathcal{J}_\mu,\ \nu_j \in \pi^{-1}(\tilde{\mu}_j)\right\} \\ &= \sum_{j\in\mathcal{J}_\mu} \mu(C_j) \inf\{I^{(2)}(\nu_j) \mid \nu_j \in \pi^{-1}(\tilde{\mu}_j)\} \\ &= \sum_{j\in\mathcal{J}_\mu} \mu(C_j) I^{(1)}(\tilde{\mu}_j). \end{aligned}$$

□

5 Level-3 weak LDP

In this section, we prove Theorem 1.5. We follow the projective limit approach of Section 4.4 of [13], and of Sections 6.5.2 and 6.5.3 of [11]. Since we work with weak LDPs and without the uniformity assumption **(U)** used in these books, we need to slightly adapt their propositions. In Section 5.1, we derive the weak LDPs for the k-th empirical measures $(L_n^{(k)})$, and derive convenient expressions for the associated rate functions. In Section 5.2, we turn these weak LDPs into the weak LDP for the sequence $(L_n^{(\infty)})$.

5.1 Weak LDP for the k-th empirical measures

In this section, we work in $\mathcal{P}(S^k)$ with $k \geqslant 3$. We use the notations of Section 2.1. For all $n \in \mathbb{N}$, the k-th empirical measure of a word $w \in S^{n+k-1}$ is defined as

$$L^{(k)}[w] = \frac{1}{n}\sum_{i=1}^{n} \delta_{(w_i,\dots,w_{i+k-1})}.$$

The k-th empirical measure of (X_n) at time n is denoted $L_n^{(k)}$ and is defined as the k-th empirical measure of the word $(X_1,\dots,X_{n+k-1})$. The sequence $(L_n^{(k)})$ is the object of interest in this section.

Our goal is to translate the weak LDP for $(L_n^{(2)})$ of Theorem 1.3 to the weak LDP for $(L_n^{(k)})$. To achieve this, we consider the Markov chain (Y_n) on S^{k-1} defined by $Y_n = (X_n,\dots,X_{n+k-2})$, whose pair empirical measures carry the same information as $(L_n^{(k)})$, as we will see. Using the notation of (2.1), the initial measure of the Markov chain (Y_n) is given by $\beta_Y(u) = \beta(u_1)p(u)$ and its transition kernel by

$$p_Y(u,v) = p(u_{k-1},v_{k-1})\prod_{i=1}^{k-2}\mathbb{1}_{\{u_{i+1}=v_i\}}, \qquad u,v\in S^{k-1}.$$

Notice that a word $u \in S^{k-1}$ is reachable from $v \in S^{k-1}$ for p_Y if and only if there exists a word $w \in S_{\text{fin},+}$ of which u is prefix and v is suffix (u and v may overlap in w). The pair empirical measure associated to this Markov chain is

$$L_{Y,n}^{(2)} = \frac{1}{n}\sum_{i=1}^{n}\delta_{(Y_n,Y_{n+1})} = \frac{1}{n}\sum_{i=1}^{n}\delta_{((X_n,\dots,X_{n+k-2}),(X_{n+1},\dots,X_{n+k-1}))}.$$

According to Theorem 1.3, $(L_{Y,n}^{(2)})$ satisfies the weak LDP with an explicitly known rate function. In Proposition 5.2 below, we will prove that this weak LDP is actually the weak LDP for $(L_n^{(k)})$. Before stating Proposition 5.2, let us introduce some notations. As in (1.8), let

$$\mathcal{D}^{(k)} = \{\mu \in S^k \mid \forall u \in S^k,\, p(u) = 0 \Rightarrow \mu(u) = 0\}. \tag{5.1}$$

Let $\mu \in \mathcal{P}(S^k)$. For any $k' \leqslant k$, we define a measure $\mu^{(k')} \in \mathcal{P}(S^{k'})$ by setting

$$\mu^{(k')}(A) = \mu(A \times S^{k-k'}), \qquad A \subseteq S^{k'}.$$

We denote by $\pi_{k,k'} : \mathcal{P}(S^k) \to \mathcal{P}(S^{k'})$ the continuous map $\nu \mapsto \nu^{(k')}$. As in (1.6), we say that the measure μ is balanced, and we write $\mu \in \mathcal{P}_{\mathrm{bal}}(S^k)$ if

$$\mu(A \times S) = \mu(S \times A), \qquad A \subseteq S^{k-1}. \tag{5.2}$$

Like previously, we denote by $\mathcal{A}_{\mathrm{bal}}^{(k)}$ the set of measures of $\mathcal{P}(S^k)$ that are both admissible and balanced. Note that $\mathcal{P}_{\mathrm{bal}}(S^k)$ and $\mathcal{A}_{\mathrm{bal}}^{(k)}$ are both closed. If μ is balanced, by induction on (5.2), one can show that $\mu^{(k')}$ satisfies

$$\mu^{(k')}(A) = \mu(S^{k_1} \times A \times S^{k_2}), \qquad A \subseteq S^{k'},$$

for all $k', k_1, k_2 \geqslant 0$ such that $k_1 + k_2 = k - k'$. It follows that $\mu^{(k')}$ is balanced in $\mathcal{P}(S^{k'})$ when μ is balanced in $\mathcal{P}(S^k)$.

Before stating Proposition 5.2, we also need to define some rate functions. Like previously, we define the k-th SCGF by

$$\Lambda^{(k)}(V) = \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[e^{n\langle L_n^{(k)}, V\rangle}\right], \qquad V \in \mathcal{B}(S^k).$$

Its convex conjugate $(\Lambda^{(k)})^*$ is given by Definition B.1. Let $\mathcal{E}(S^k)$ denote the set of positive functions on S^k that are bounded away from 0 and ∞. We define, for all $\mu \in \mathcal{P}(S^k)$,[35]

$$J^{(k)}(\mu) = \sup_{f\in\mathcal{E}(S^k)} \left\langle \mu, \log \frac{f}{P^{(k)}f} \right\rangle, \qquad P^{(k)}f : u \mapsto \sum_{y\in S} p(u_k, y) f(u_2, \dots, u_k, y),$$

and

$$R^{(k)}(\mu) = H(\mu | \mu^{(k-1)} \otimes p).$$

Note that when $\mu \in \mathcal{D}^{(k)}$, we have $\mu \ll \mu^{(k-1)} \otimes p$ and

$$R^{(k)}(\mu) = \sum_{u\in S^{k-1}} \sum_{y\in S} \mu^{(k)}(uy) \log \frac{\mu^{(k)}(uy)}{\mu^{(k-1)}(u) p(u_{k-1}, y)}. \tag{5.3}$$

In the case $k = 2$, the rate functions $(\Lambda^{(2)})^*$, $J^{(2)}$ and $R^{(2)}$ are necessarily infinite outside $\mathcal{D}^{(2)}$; see Remark 3.2. As one can see in Remark 5.1, in the general case, the set $\mathcal{D}^{(k)}$ plays the same role for $(\Lambda^{(k)})^*$, $J^{(k)}$ and $R^{(k)}$.

Remark 5.1. As for the $k = 2$ case, it is easy to check that for all $\mu \in \mathcal{P}(S^k)$ that does not belong to $\mathcal{D}^{(k)}$, we have

$$(\Lambda^{(k)})^*(\mu) = J^{(k)}(\mu) = R^{(k)}(\mu) = \infty.$$

We can now get to the main proposition of this section.

Proposition 5.2. *The sequence $(L_n^{(k)})$ satisfies the weak LDP with rate function $I^{(k)}$ given by*

$$I^{(k)}(\mu) = \begin{cases} (\Lambda^{(k)})^*(\mu) = J^{(k)}(\mu) = R^{(k)}(\mu), & \text{if } \mu \in \mathcal{A}_{\mathrm{bal}}^{(k)}, \\ \infty, & \text{otherwise.} \end{cases} \tag{5.4}$$

The proof of this proposition is rather technical. It consists in applying a contraction principle to a map Φ that turns $L_{n,Y}^{(2)}$ into $L_n^{(k)}$. Then, it must be checked that the obtained rate function matches the claim (5.4).[36]

[35] The operator $P^{(k)}$ corresponds to the operator P of (3.3), but for the Markov chain $((X_n, \dots, X_{n+k-1}))$. Hence $J^{(k)}$ is just the DV entropy relative to this Markov chain. Also notice that the quantity $P^{(k)}f(u)$ does not depend on u_1.

[36] A brief analysis of the reducibility structure of the Markov chain (Y_n) reveals that its irreducible classes are slightly more intricate than C_j^{k-1}, which complicates the definition of admissible measures for (Y_n).

Proof. Let

$$\varphi:\begin{cases} U & \to S^k, \\ (u,v) & \mapsto u_1 v, \end{cases} \qquad U=\{(u,v)\in(S^{k-1})^2 \mid u_{i+1}=v_i\ 1\leqslant i\leqslant k-2\}.$$

The application φ is a one-to-one correspondence between U and S^k. For all $\nu\in\mathcal{P}(U)$, let $\Phi(\nu):=\nu\circ\varphi^{-1}\in\mathcal{P}(S^k)$. The application Φ is a one-to-one correspondence between $\mathcal{P}(U)$ and $\mathcal{P}(S^k)$ (the inverse application is defined by $\Phi^{-1}(\mu)=\mu\circ\varphi\in\mathcal{P}(U)$ for all $\mu\in\mathcal{P}(S^k)$). It is even a homeomorphism since both the maps Φ and Φ^{-1} preserve the TV distance. Moreover, if $v\in S^k$, we have $\Phi(\delta_{\varphi^{-1}(v)})=\delta_v$. It follows that

$$L_n^{(k)}=\frac{1}{n}\sum_{i=1}^{n}\Phi(\delta_{\varphi^{-1}(X_i\ldots X_{i+k-1})})=\Phi(L_{Y,n}^{(2)}). \tag{5.5}$$

By Theorem 1.3, $(L_{Y,n}^{(2)})$ satisfies the weak LDP on $\mathcal{P}((S^{k-1})^2)$. Since the measure $L_{Y,n}^{(2)}$ always belongs to the closed set $\mathcal{P}(U)$, this weak LDP is actually a weak LDP on $\mathcal{P}(U)$ equipped by the induced weak topology. We denote by I_Y the associated rate function. Therefore, by (5.5) and since Φ is a homeomorphism, $(L_n^{(k)})$ satisfies the weak LDP with rate function $I^{(k)}=I_Y\circ\Phi^{-1}$.[37]

It remains to identify $I^{(k)}$, *i.e.* to establish (5.4). Let $\mathcal{A}_{Y,\mathrm{bal}}^{(2)}$ denote the set of admissible, balanced measures of $\mathcal{P}((S^{k-1})^2)$ relative to (Y_n), which naturally appears in the expression of I_Y given by (3.7). We must first precisely describe this set of measures. We now prove that

$$\Phi(\mathcal{A}_{\mathrm{bal},Y}^{(2)}\cap\mathcal{D}_Y^{(2)})=\mathcal{A}_{\mathrm{bal}}^{(k)}\cap\mathcal{D}^{(k)}, \tag{5.6}$$
$$\mathcal{D}_Y^{(2)}:=\{\nu\in\mathcal{P}(U)\mid\forall(u,v)\in U,\ p_Y(u,v)=0\Rightarrow\nu(u,v)=0\}.$$

For convenience, we set $T_{k-1}:=S_{\mathrm{fin},+}\cap S^{k-1}=\{u\in S^{k-1}\mid p(u)>0\}$. Let $\nu\in\mathcal{P}(U)$. The following holds.

- $\nu\in\mathcal{D}_Y^{(2)}$ if and only if $\Phi(\nu)\in\mathcal{D}^{(k)}$. Indeed, for all $(u,v)\in U$,

$$\nu(u,v)=0\Leftrightarrow\Phi(\nu)(\varphi(u,v))=0,\qquad p(u)p(v)=0\Leftrightarrow p(\varphi(u,v))=0.$$

- ν is balanced in $\mathcal{P}((S^{k-1})^2)$ if and only if $\Phi(\nu)$ is balanced in $\mathcal{P}(S^k)$.[38] Indeed, for all $A\subseteq S^{k-1}$,

$$\Phi(\nu)(A\times S)=\nu(\varphi^{-1}(A\times S))=\nu((A\times S^{k-1})\cap U)=\nu(A\times S^{k-1}),$$
$$\Phi(\nu)(S\times A)=\nu(\varphi^{-1}(S\times A))=\nu((S^{k-1}\times A)\cap U)=\nu(S^{k-1}\times A).$$

  Hence, ν satisfies (1.6) if and only if $\Phi(\nu)$ satisfies (5.2).

- Let $(C_{Y,j})_{j\in\mathcal{J}_Y}$ be the sequence of irreducible classes of the Markov chain (Y_n), as in (1.7). We have $\mathcal{J}_Y=\mathcal{J}$ and

$$C_{Y,j}=C_j^{k-1}\cap T_{k-1},\qquad j\in\mathcal{J}_Y. \tag{5.7}$$

  Indeed, let $u,v\in C_j^{k-1}\cap T_{k-1}$. Then, $u_{k-1}\rightsquigarrow v_1$, thus there exists a word ξ such that $p(u_{k-1}\xi v_1)>0$. Since $p(u)>0$ and $p(v)>0$, we have $w:=u\xi v\in S_{\mathrm{fin},+}$, implying that v is reachable from u for p_Y. In the same way, u is reachable from v for p_Y, thus u and v are in the same irreducible class for p_Y. Now let $u,v\in S^{k-1}$ be in the same class for p_Y. There exist $w,w'\in S_{\mathrm{fin},+}$ such that u is a prefix of w and a suffix of w' and v is a prefix of w' and a suffix of w. Using these two words, we can construct a word $w''\in S_{\mathrm{fin},+}$ of the form $w''=u\xi v\zeta u\xi v$. The fact that $p(w'')>0$ implies that $u,v\in T_{k-1}$ and that any letter of u or v is reachable form any letter of u or v. Therefore, all the letters of u and v are in the same irreducibility class. This proves (5.7). Therefore, ν is admissible if and only if

$$\sum_{j\in\mathcal{J}_\nu}\nu((C_j^{k-1})^2\cap T_{k-1}^2)=1, \tag{5.8}$$

  where $\mathcal{J}_\nu$ satisfies the admissibility condition of Definition 1.2.

[37] This can be seen as a consequence of Lemma 4.1.

[38] The question of whether ν is balanced requires seeing ν as a measure in $\mathcal{P}((S^{k-1})^2)$ with support in U instead of a measure of $\mathcal{P}(U)$.

Let $\mu \in \mathcal{P}(S^k)$ and let $\nu = \Phi^{-1}(\mu) \in \mathcal{P}(U)$. If $\mu \in \mathcal{A}^{(k)}_{\mathrm{bal}} \cap \mathcal{D}^{(k)}$, then $\nu \in \mathcal{P}_{\mathrm{bal}}((S^{k-1})^2) \cap \mathcal{D}^{(2)}_Y$ and $\sum_{j\in\mathcal{J}_\mu} \mu(C^k_j) = 1$. Since $\nu \in \mathcal{D}^{(2)}_Y$, we have $\operatorname{supp}\nu \subseteq T^2_{k-1}$. It follows that

$$\sum_{j\in\mathcal{J}_\mu} \nu((C^{k-1}_j)^2 \cap T^2_{k-1}) = \sum_{j\in\mathcal{J}_\mu} \nu((C^{k-1}_j)^2) = \sum_{j\in\mathcal{J}_\mu} \mu(C^k_j) = 1.$$

Since $\mathcal{J}_\mu$ satisfies the admissibility conditions of Definition 1.2, the measure ν is admissible. Reciprocally, if $\nu \in \mathcal{A}^{(2)}_{\mathrm{bal},Y} \cap \mathcal{D}^{(2)}_Y$, then $\mu \in \mathcal{P}_{\mathrm{bal}}(S^k) \cap \mathcal{D}^{(k)}$ and (5.8) holds. Since $\nu \in \mathcal{D}^{(2)}_Y$, we have $\operatorname{supp}\nu \subseteq T^2_{k-1}$. It follows that

$$\sum_{j\in\mathcal{J}_\nu} \mu(C^k_j) = \sum_{j\in\mathcal{J}_\nu} \nu((C^{k-1}_j)^2) = \sum_{j\in\mathcal{J}_\nu} \nu((C^{k-1}_j)^2 \cap T^2_{k-1}) = 1.$$

Since $\mathcal{J}_\nu$ satisfies the admissibility conditions of Definition 1.2, the measure μ is admissible. This achieves to prove (5.6).

Now we can prove (5.4). Let $\mu \in \mathcal{P}(S^k)$ and let $\nu = \Phi^{-1}(\mu)$. If $\mu \notin \mathcal{A}^{(k)}_{\mathrm{bal}} \cap \mathcal{D}^{(k)}$, then by (5.6), $\nu \notin \mathcal{A}^{(2)}_{\mathrm{bal},Y} \cap \mathcal{D}^{(2)}_Y$, thus $I^{(k)}(\mu) = I_Y(\nu) = \infty$. Indeed, by Proposition 3.1 and Remark 3.2, I_Y is infinite outside of $\mathcal{A}^{(2)}_{\mathrm{bal},Y} \cap \mathcal{D}^{(2)}_Y$. By Remark 5.1, this proves (5.4) for $\mu \notin \mathcal{A}^{(2)}_{\mathrm{bal}} \cap \mathcal{D}^{(k)}$. Assume now that $\mu \in \mathcal{A}^{(k)}_{\mathrm{bal}} \cap \mathcal{D}^{(k)}$. Applying Proposition 3.1 to the chain (Y_n) shows that

$$I_Y(\nu) = (\Lambda^{(2)}_{\infty,Y})^*(\nu) = (\Lambda^{(2)}_Y)^*(\nu) = J^{(2)}_Y(\nu) = R^{(2)}_Y(\nu),$$

where the subscript Y was added to distinguish the rate functions corresponding to (Y_n) instead of (X_n). Thus, in order to complete the proof of (5.4), we need to identify $(\Lambda^{(2)}_{\infty,Y})^*(\nu) = (\Lambda^{(k)}_\infty)^*(\mu)$, $(\Lambda^{(2)}_Y)^*(\nu) = (\Lambda^{(k)})^*(\mu)$, $J^{(2)}_Y(\nu) = J^{(k)}(\mu)$ and $R^{(2)}_Y(\nu) = R^{(k)}(\mu)$.

1. We begin with $(\Lambda^{(2)}_Y)^*(\nu) = (\Lambda^{(k)})^*(\mu)$. For all $V \in \mathcal{B}((S^{k-1})^2)$ and all $\gamma \in \mathcal{P}(U)$, we have

$$\langle \gamma, V\rangle = \langle \Phi(\gamma), V \circ \varphi^{-1}\rangle. \tag{5.9}$$

By definition,

$$(\Lambda^{(2)}_Y)^*(\nu) = \sup_{V\in\mathcal{B}((S^{k-1})^2)} \left(\langle \nu, V\rangle - \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[e^{n\langle L^{(2)}_{n,Y}, V\rangle} \right] \right). \tag{5.10}$$

Since both $L^{(2)}_{n,Y}$ and ν belong to $\mathcal{P}(U)$, the quantity being optimized in (5.10) does not depend on the values of V outside U. Hence, the supremum can be taken over $V \in \mathcal{B}(U)$. Thus, since the application $V \mapsto V\circ\varphi^{-1}$ is a bijection between $\mathcal{B}(U)$ and $\mathcal{B}(S^k)$, Identity (5.9) shows that the expressions of $(\Lambda^{(2)}_Y)^*(\nu)$ and of $(\Lambda^{(k)})^*(\mu)$ coincide.

2. We now show $J^{(2)}_Y(\nu) = J^{(k)}(\mu)$. By definition,

$$J^{(2)}_Y(\nu) = \sup_{f\in\mathcal{E}((S^{k-1})^2)} \Big\langle \nu, \log\frac{f}{P_Y f}\Big\rangle, \tag{5.11}$$

where[39]

$$P_Y f : (u,v) \mapsto \sum_{y\in S} p(v_{k-1}, y) f(v, v_2 \dots v_{k-1} y).$$

Since ν belongs to $\mathcal{P}(U)$, the quantity being optimized in (5.11) does not depend on the values of f outside U. Hence, the supremum can be taken over $f \in \mathcal{E}(U)$. Since $f \mapsto f\circ\varphi^{-1}$ is a bijection between $\mathcal{E}(U)$ and $\mathcal{E}(S^k)$, and after noticing that

$$(P_Y f)\circ\varphi^{-1}(w) = \sum_{y\in S} p(w_k, x) f(w_2 \dots w_k, w_3 \dots w_k y) = P^{(k)}(f\circ\varphi^{-1})(w), \qquad w \in S^k,$$

Identity (5.9) shows again that the expressions of $J^{(2)}_Y(\nu)$ and of $J^{(k)}(\mu)$ coincide.

[39]The function $P_Y f$ does not depend on its first variable.

3. We now show $R_Y^{(2)}(\nu) = R^{(k)}(\mu)$. The measure ν belongs to $\mathcal{D}_Y^{(2)}$, thus by (5.3),

$$R_Y^{(2)}(\nu) = \sum_{(u,v)\in U} \nu(u,v) \log \frac{\nu(u,v)}{\nu^{(1)}(u) p_Y(u,v)}.$$

For all $(u,v) \in U$, we have $p_Y(u,v) = p(u_{k-1}, v_{k-1})$. Moreover, for all $u \in S^{k-1}$,

$$\nu^{(1)}(u) = \sum_{\substack{v\in S^{k-1}\\(u,v)\in U}} \nu(u,v) = \sum_{y\in S} \nu(u, u_2 \dots u_{k-1} y) = \sum_{y\in S} \mu(uy) = \mu^{(k-1)}(u).$$

By replacing $p_Y(u,v)$ and $\nu^{(1)}(u)$ by these values in the expression of $R_Y^{(2)}(\nu)$, since $(u,v) \mapsto (u, v_k)$ is a bijection between U and $S^{k-1} \times S$, we find that the expression of $R_Y^{(2)}(\nu)$ and the expression (3.6) of $R^{(k)}(\mu)$ coincide.

□

Let us also mention a relation between $R^{(k)}$ and relative entropy, which will be useful in the proof of Proposition 5.6 below. Notice that both sides in (5.12) can be ∞.

Lemma 5.3. *Let $\mu \in \mathcal{P}_{\mathrm{bal}}(S^k)$ be absolutely continuous with respect to $\mathbb{P}_k$. Then,*

$$H(\mu|\mathbb{P}_k) = H(\mu^{(k-1)}|\mathbb{P}_{k-1}) + R^{(k)}(\mu). \tag{5.12}$$

Proof. For all $u \in \operatorname{supp} \mu^{(k-1)}$, let $q(u,\cdot)$ be the probability measure on S defined by the relation $\mu(uy) = \mu^{(k-1)}(u) q(u,y)$ for all $y \in S$. Since $\mu \in \mathcal{D}^{(k)}$, Equation (5.3) holds. Equation (5.12) is the result of the following computations, which we justify below.

$$\begin{aligned}
H(\mu|\mathbb{P}_k) &= \sum_u \mu^{(k-1)}(u) \sum_y q(u,y) \log \frac{\mu(uy)}{\mathbb{P}_k(uy)} \\
&= \sum_u \mu^{(k-1)}(u) \sum_y q(u,y) \left(\log \frac{\mu^{(k-1)}(u)}{\mathbb{P}_{k-1}(u)} + \log \frac{q(u,y)}{p(u_{k-1},y)} \right) \qquad (5.13)\\
&= \sum_u \mu^{(k-1)}(u) \left(\log \frac{\mu^{(k-1)}(u)}{\mathbb{P}_{k-1}(u)} + \sum_y q(u,y) \log \frac{q(u,y)}{p(u_{k-1},y)} \right) \qquad (5.14)\\
&= \sum_u \mu^{(k-1)}(u) \log \frac{\mu^{(k-1)}(u)}{\mathbb{P}_{k-1}(u)} + \sum_u \mu^{(k-1)}(u) \sum_y q(u,y) \log \frac{q(u,y)}{p(u_{k-1},y)} \qquad (5.15)\\
&= H(\mu^{(k-1)}(u)|\mathbb{P}_{k-1}) + R^{(k)}(\mu).
\end{aligned}$$

In these computations, the sums are taken for $u \in \operatorname{supp} \mu^{(k-1)}$ and $y \in \operatorname{supp} q(u,\cdot)$. The transitions from (5.13) to (5.14) and from (5.14) to (5.15) must be justified. Let $u \in \operatorname{supp} \mu^{(k-1)}$. Then, the series

$$\sum_{y\in \operatorname{supp} q(u,\cdot)} q(u,y) \log \frac{\mu^{(k-1)}(u)}{\mathbb{P}_{k-1}(u)} = \log \frac{\mu^{(k-1)}(u)}{\mathbb{P}_{k-1}(u)}$$

is absolutely convergent in $(-\infty, +\infty)$, since $\mu^{(k-1)}$ is absolutely continuous with respect to $\mathbb{P}_{k-1}$. Moreover, the sum

$$\sum_{y\in \operatorname{supp} q(u,\cdot)} q(u,y) \log \frac{q(u,y)}{p(u_{k-1},y)} = H(q(u,\cdot)|p(u_{k-1},\cdot))$$

is properly defined in $[0, +\infty]$ as a relative entropy, because $q(u,\cdot)$ is absolutely continuous with respect to $p(u_{k-1},\cdot)$. This justifies that the expression within brackets in (5.14) is properly defined in $(-\infty, +\infty]$ for every u, which justifies the transition from (5.13) to (5.14). In (5.15), notice that both

$$u \mapsto \frac{\mu^{(k-1)}(u)}{\mathbb{P}_{k-1}(u)} \log \frac{\mu^{(k-1)}(u)}{\mathbb{P}_{k-1}(u)}, \qquad u \mapsto \sum_{y\in \operatorname{supp} q(u,\cdot)} q(u,y) \log \frac{q(u,y)}{p(u_{k-1},y)},$$

are bounded below; the former by $\inf_{s>0} s \log s = -e^{-1}$, and the latter by 0 as a relative entropy. Hence the two sums in (5.15) are well defined in $(-\infty, +\infty]$, and the transition from (5.14) to (5.15) is justified. □

5.2 Weak LDP for the empirical process

In this section, we prove Theorem 1.5. We follow the projective limit approach of Sections 6.5.2 and 6.5.3 of [11], Section 4.4 of [13], or originally Theorem 3.3 of [7]. We equip $S^{\mathbb{N}}$ with the product topology, which is metrized by the distance $d^{(\infty)}$ given by $d^{(\infty)}(\omega,\omega') = 2^{-n}$, where $n = \inf\{k \in \mathbb{N} \mid \omega_k \neq \omega'_k\}$. Since S is separable, the Borel σ-algebra of $S^{\mathbb{N}}$ is generated by the product of Borel σ-algebras of S; see Theorem 1.10 of [25]. Let $\mathcal{P}(S^{\mathbb{N}})$ denote the space of probability measures on $S^{\mathbb{N}}$. Let $\mathcal{C}(S^{\mathbb{N}})$ and $\mathcal{C}_{\mathrm{u}}(S^{\mathbb{N}})$ respectively denote the space of bounded continuous functions and of bounded, uniformly continuous, functions on $S^{\mathbb{N}}$. We define the dual pairing

$$(V,\mu) \mapsto \langle \mu, V\rangle = \int_{S^{\mathbb{N}}} V \mathrm{d}\mu, \qquad V \in \mathcal{V},\ \mu \in \mathcal{P}(S^{\mathbb{N}}), \tag{5.16}$$

where $\mathcal{V} = \mathcal{C}(S^{\mathbb{N}})$ or $\mathcal{V} = \mathcal{C}_{\mathrm{u}}(S^{\mathbb{N}})$. Functions $V \in \mathcal{B}(S^k)$ for some $k \in \mathbb{N}$ can be seen as functions of $\mathcal{C}_{\mathrm{u}}(S^k)$ that depend only on their first k variables, satisfying $\langle \mu, V\rangle = \langle \mu^{(k)}, V\rangle$ for $\mu \in \mathcal{P}(S^{\mathbb{N}})$. We have

$$\bigcup_{k\geqslant 2} \mathcal{B}(S^k) \subseteq \mathcal{C}_{\mathrm{u}}(S^{\mathbb{N}}) \subseteq \mathcal{C}(S^{\mathbb{N}}). \tag{5.17}$$

The weak topology on $\mathcal{P}(S^{\mathbb{N}})$ is the coarsest topology that makes each $\langle \cdot, V\rangle$ continuous, for all $V \in \mathcal{V}$ (by the Portmanteau theorem, both $\mathcal{V} = \mathcal{C}(S^{\mathbb{N}})$ and $\mathcal{V} = \mathcal{C}_{\mathrm{u}}(S^{\mathbb{N}})$ define the same weak topology; see Theorem D.10 of [11] or Theorem 6.1 of [25]). We equip $\mathcal{P}(S^{\mathbb{N}})$ with this topology. By Lemma 6.5.14 of [11], the topological space $\mathcal{P}(S^{\mathbb{N}})$ is the projective limit of $(\mathcal{P}(S^k))_{k\geqslant 2}$. This means that the weak topology is generated by sets

$$\left\{\mu \in \mathcal{P}(S^{\mathbb{N}}),\ \left|\langle \mu^{(k)}, V\rangle - x\right| < \varepsilon\right\}, \qquad k \geqslant 2,\ V \in \mathcal{B}(S^k),\ x \in \mathbb{R},\ \varepsilon > 0.$$

Let T denote the shift map on $S^{\mathbb{N}}$ defined by $T(x_1, x_2, \ldots) = (x_2, x_3, \ldots)$, and let $\mathcal{P}_{\mathrm{bal}}(S^{\mathbb{N}})$ denote the set of shift-invariant measures of $\mathcal{P}(S^{\mathbb{N}})$. For $\mu \in \mathcal{P}(S^{\mathbb{N}})$ and $k \geqslant 1$, we define $\mu^{(k)} \in \mathcal{P}(S^k)$ by setting $\mu^{(k)}(A) = \mu(A \times S^{\mathbb{N}})$ for all $A \subseteq S^k$. When μ is shift-invariant, $\mu^{(k)}$ is balanced. Conversely, if every $\mu^{(k)}$ is balanced, then μ is shift-invariant because, for all $A \subseteq S^k$, by (5.2),

$$\mu(T^{-1}(A \times S^{\mathbb{N}})) = \mu(S \times A \times S^{\mathbb{N}}) = \mu^{(k+1)}(S \times A) = \mu^{(k+1)}(A \times S) = \mu(A \times S^{\mathbb{N}}).$$

We denote by $\pi_k : \mathcal{P}(S^{\mathbb{N}}) \to \mathcal{P}(S^k)$ the continuous map $\mu \mapsto \mu^{(k)}$. Notice that $\pi_{k'} = \pi_{k,k'} \circ \pi_k$ for all $1 \leqslant k' \leqslant k$, where $\pi_{k,k'}$ is a continuous map defined in Section 5.1. Notice that two measures μ, ν are equal in $\mathcal{P}_{\mathrm{bal}}(S^{\mathbb{N}})$ if and only if $\pi_k(\mu) = \pi_k(\nu)$ for all $k \geqslant 2$. Recall that the empirical process $(L_n^{(\infty)})$ is defined as an element of $\mathcal{P}(S^{\mathbb{N}})$ by (1.1). Clearly, $\pi_k(L_n^{(\infty)}) = L_n^{(k)}$. Our goal is to prove Theorem 1.5, *i.e.* to derive the weak LDP in $\mathcal{P}(S^{\mathbb{N}})$ for $(L_n^{(\infty)})$ and prove Equation (1.11).

We shall first define the rate functions appearing in (1.11). Let $\mu \in \mathcal{P}(S^{\mathbb{N}})$. We define the level-3 SCGF by

$$\Lambda^{(\infty)}(V) = \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[e^{n\langle L_n^{(\infty)}, V\rangle}\right], \qquad V \in \mathcal{C}(S^{\mathbb{N}}).$$

We let[40]

$$(\Lambda^{(\infty)})^*(\mu) = \sup_{V\in\mathcal{C}(S^{\mathbb{N}})} \left(\langle \mu, V\rangle - \Lambda^{(\infty)}(V)\right), \qquad \mu \in \mathcal{P}(S^{\mathbb{N}}). \tag{5.18}$$

Let $S^{-\mathbb{N}_0} = \{\ldots, \omega_{-2}, \omega_{-1}, \omega_0 \mid \omega_i \in S,\ i \in \mathbb{N}_0\}$, equipped with the product topology and the distance $d_-^{(\infty)}$ given by $d_-^{(\infty)}(\omega,\omega') = 2^n$ where $n = \sup\{k \in -\mathbb{N}_0 \mid \omega_k \neq \omega'_k\}$. Since μ is shift invariant, Kolmogorov's extension theorem ensures that we can define a unique measure $\mu^* \in \mathcal{P}(S^{-\mathbb{N}_0})$ satisfying

$$\mu^*(S^{-\mathbb{N}} \times \{u\}) = \mu^{(k)}(u), \qquad u \in S^k,\ k \geqslant 1.$$

[40] If we equip $\mathcal{C}(S^{\mathbb{N}})$ with the weak topology induced by $\langle \cdot, \cdot\rangle$ and $\mathcal{P}(S^{\mathbb{N}})$, the function $(\Lambda^{(\infty)})^*$ is the convex conjugate of $\Lambda^{(\infty)}$.

Let $\mathcal{E}(S^{-\mathbb{N}_0}) := \exp(\mathcal{C}(S^{-\mathbb{N}_0}))$ denote the set of continuous positive functions on $S^{-\mathbb{N}_0}$ that are bounded away from 0 and ∞. Let

$$J^{(\infty)}(\mu) = \sup_{f\in\mathcal{E}(S^{-\mathbb{N}_0})} \Big\langle \mu^*, \log \frac{f}{P^{(\infty)}f}\Big\rangle, \tag{5.19}$$

where

$$P^{(\infty)}g(\omega) = \sum_{y\in S} p(\omega_0, y)g(\omega y), \qquad g \in \mathcal{E}(S^{-\mathbb{N}_0}).$$

In the definition of $P^{(\infty)}$ above, ωy denotes the word indexed by $-\mathbb{N}_0$ whose letters are $(\omega y)_i = y$ if $i = 0$ and $(\omega y)_i = \omega_{i+1}$ if $i < 0$. The symbol $\langle\cdot,\cdot\rangle$ in (5.19) denotes the same integral as in (5.16) but indexed over $S^{-\mathbb{N}_0}$. The function $J^{(\infty)}$ is the level-3 DV entropy.[41] We also define $R^{(\infty)}$ like in in Sections 4.4 of [13] and 6.5.3 of [11]. Using again Kolmogorov's extension Theorem, we define a measure $\nu^* \in \mathcal{P}(S^{-\mathbb{N}})$ by

$$\nu^*(S^{-\mathbb{N}_0} \times \{u\}) = \mu^{(k-1)}(u_1 \dots u_{k-1})p(u_{k-1}, u_k), \qquad u \in S^k,\ k \geqslant 2.$$

If μ^* is absolutely continuous with respect to ν^*, let $\frac{d\mu^*}{d\nu^*}$ denote the Radon-Nykodym derivative of μ^* with respect to ν^*. We set $R^{(\infty)}(\mu)$ as the relative entropy of μ^* with respect to ν^*, that is

$$R^{(\infty)}(\mu) = \begin{cases} \int_{S^{-\mathbb{N}_0}} \log \frac{d\mu^*}{d\nu^*} d\mu^*, & \text{if } \mu^* \ll \nu^*,\\ \infty, & \text{otherwise.}\end{cases} \tag{5.20}$$

These three definitions of rate functions are natural extensions of (3.1), (3.2), and (3.5). The last definition needed in the statement of Theorem 1.5 is the definition of $\mathcal{A}^{(\infty)}$ and $\mathcal{A}^{(\infty)}_{\mathrm{bal}}$. This definition is also a natural extension of Definition 1.2. In fact, $\mathcal{A}^{(\infty)}_{\mathrm{bal}}$ is the projective limit of the sequence $(\mathcal{A}^{(k)}_{\mathrm{bal}})$; see Appendix A.3.

Definition 5.4 (Admissibility)**.** *Let $\mu \in \mathcal{P}(S^{\mathbb{N}})$. We say that μ is* pre-admissible *if there exists a set of indices $\mathcal{J}_\mu \subseteq \mathcal{J}$ such that*

$$\mu = \sum_{j\in\mathcal{J}_\mu} \mu|_{C_j^{\mathbb{N}}},$$

in which case we impose that $\mathcal{J}_\mu$ is minimal. If μ is pre-admissible, we say that μ is admissible *if the order $\rightsquigarrow$ is total on $(C_j)_{j\in\mathcal{J}_\mu}$ and $\beta \rightsquigarrow C_j$ for all $j \in \mathcal{J}_\mu$. We denote by $\mathcal{A}^{(\infty)}$ the set of all admissible measures of $\mathcal{P}(S^{\mathbb{N}})$, and we set $\mathcal{A}^{(\infty)}_{\mathrm{bal}} = \mathcal{A}^{(\infty)} \cap \mathcal{P}_{\mathrm{bal}}(S^{\mathbb{N}})$.*

Now we can turn to the proof of Theorem 1.5. We first use a variant of the Dawson-Gärtner Theorem to derive the weak LDP with an abstract rate function and second compute this rate function with Lemma 5.5 and Proposition 5.6, which are stated and proved below.

Proof of Theorem 1.5. Let $I^{(\infty)} : \mathcal{P}(S^{\mathbb{N}}) \to [0, +\infty]$ be defined by

$$I^{(\infty)}(\mu) = \sup_{k\geqslant 2} I^{(k)}(\mu^{(k)}), \qquad \mu \in \mathcal{P}(S^{\mathbb{N}}),$$

where $I^{(k)}$ is the rate function of Proposition 5.2. Since it is the supremum of a family of lower semicontinuous functions, this function is lower semicontinuous. We begin by proving that $(L_n^{(\infty)})$ satisfies the weak LDP with rate function $I^{(\infty)}$, by a variant of the Dawson-Gärtner Theorem. The proof of the LDP lower bound with $I^{(\infty)}$ is the classical proof for the lower bound of the Dawson-Gärtner Theorem, as in Theorem 4.6.1 of [11] and Exercise 2.1.21 of [13], and we do not repeat it here. It remains to show that the weak LDP upper bound holds with $I^{(\infty)}$. The proof of the upper bound is similar to the one of [11] or [13], up to some adaptations required by the fact that we only have weak LDPs and no good rate function.[42] Let K be a compact subset of $\mathcal{P}(S^{\mathbb{N}})$, and $\alpha \in \mathbb{R}$. For $k \in \mathbb{N}$, we set $\mathcal{L}_k = \{\mu \in \mathcal{P}(S^k) \mid I^{(k)}(\mu) \leqslant \alpha\}$. We also set $\mathcal{L}_\infty = \{\mu \in \mathcal{P}(S^{\mathbb{N}}) \mid I^{(\infty)}(\mu) \leqslant \alpha\}$. By definition of $I^{(\infty)}$, a measure $\mu \in \mathcal{P}(S^{\mathbb{N}})$ satisfies

[41] This function is not common in the literature.

[42] By definition of the weak LDP upper bound, we do not need the compactness usually brought by the goodness of the rate function.

$I^{(\infty)}(\mu) \leqslant \alpha$ if and only if it satisfies $I^{(k)}(\mu^{(k)}) \leqslant \alpha$ for all k. Therefore, $K \cap \mathcal{L}_\infty$ is the projective limit of closed sets $\pi_k(K) \cap \mathcal{L}_k$, *i.e.*

$$K \cap \mathcal{L}_\infty = \bigcap_{k \geqslant 2} \pi_k^{-1}(\pi_k(K) \cap \mathcal{L}_k).$$

Assume that $\alpha < \inf_{\mu \in K} I^{(\infty)}(\mu)$. It follows that $K \cap \mathcal{L}_\infty$ is empty. Each $\pi_k(K) \cap \mathcal{L}_k$ is compact, thus by property of the projective limit of compact sets, one of them must be empty; see Theorem B.4 in [11] or TG I.6.8 in [4]. There exists $k \geqslant 2$ such that $\pi_k(K) \cap \mathcal{L}_k$ is empty, implying that $I^{(k)}(\mu) > \alpha$ for all $\mu \in \pi_k(K)$. By Proposition 5.2, the weak LDP upper bound for $(L_n^{(k)})$ holds with $I^{(k)}$. Thus,

$$\limsup_{n \to \infty} \frac{1}{n} \log \mathbb{P}(L_n^{(\infty)} \in K) \leqslant \limsup_{n \to \infty} \frac{1}{n} \log \mathbb{P}(L_n^{(k)} \in \pi_k(K)) \leqslant -\alpha.$$

This holds true as long as $\alpha < \inf_{\mu \in K} I^{(\infty)}$, thus

$$\limsup_{n \to \infty} \frac{1}{n} \log \mathbb{P}(L_n^{(\infty)} \in K) \leqslant - \inf_{\mu \in K} I^{(\infty)}.$$

Both the lower and upper bound hold with $I^{(\infty)}$, hence $(L_n^{(\infty)})$ satisfies the weak LDP with rate function $I^{(\infty)}$. We now prove that $I^{(\infty)}$ satisfies (1.11) and (1.12). If $\mu \notin \mathcal{A}_{\mathrm{bal}}^{(\infty)}$, then by Proposition A.9, there exists k such that $\mu^{(k)} \notin \mathcal{A}_{\mathrm{bal}}^{(k)}$, therefore $I^{(\infty)}(\mu^{(\infty)}) \geqslant I^{(k)}(\mu^{(k)}) = \infty$. Assume now that $\mu \in \mathcal{A}_{\mathrm{bal}}^{(\infty)}$. By Proposition 5.2 and Lemma 5.5, we have

$$I^{(\infty)}(\mu) = \sup_{k \geqslant 2} \sup_{V_k \in \mathcal{B}(S^k)} (\langle \mu, V_k \rangle - \Lambda^{(k)}(V_k)) = (\Lambda^{(\infty)})^*(\mu),$$

$$I^{(\infty)}(\mu) = \sup_{k \geqslant 2} \sup_{f_k \in \mathcal{E}(S^k)} \left\langle \mu^{(k)}, \log \frac{f_k}{P^{(k)} f_k} \right\rangle = J^{(\infty)}(\mu),$$

$$I^{(\infty)}(\mu) = \sup_{k \geqslant 2} R^{(k)}(\mu^{(k)}) = R^{(\infty)}(\mu),$$

which establishes (1.11). Finally, the identity (1.12) is proved in Proposition 5.6 below. □

Lemma 5.5. *Let $\mu \in \mathcal{P}(S^{\mathbb{N}})$.*

1. *$\Lambda^{(\infty)}$ satisfies*

$$(\Lambda^{(\infty)})^*(\mu) = \sup_{V \in \mathcal{C}_{\mathrm{u}}(S^{\mathbb{N}})} \Big(\langle \mu, V \rangle - \Lambda^{(\infty)}(V) \Big) = \sup_{k \geqslant 2} \sup_{V_k \in \mathcal{B}(S^k)} \Big(\langle \mu^{(k)}, V_k \rangle - \Lambda^{(k)}(V_k) \Big). \tag{5.21}$$

2. *Let $\mathcal{E}_{\mathrm{u}}(S^{-\mathbb{N}_0}) := \exp(\mathcal{C}_{\mathrm{u}}(S^{-\mathbb{N}_0}))$ denote the set of uniformly continuous positive functions on $S^{-\mathbb{N}_0}$ that are bounded away from 0 and ∞. Then, $J^{(\infty)}$ satisfies*

$$J^{(\infty)}(\mu) = \sup_{f \in \mathcal{E}_{\mathrm{u}}(S^{-\mathbb{N}_0})} \left\langle \mu^*, \log \frac{f}{P^{(\infty)} f} \right\rangle = \sup_{k \geqslant 2} \sup_{f_k \in \mathcal{E}(S^k)} \left\langle \mu^{(k)}, \log \frac{f_k}{P^{(k)} f_k} \right\rangle. \tag{5.22}$$

3. *The sequence $(R^{(k)}(\mu^{(k)}))$ is nondecreasing and $R^{(\infty)}$ satisfies*

$$R^{(\infty)}(\mu) = \lim_{k \to \infty} R^{(k)}(\mu^{(k)}).$$

Proof. 1. By (5.17), since $\Lambda^{(\infty)}(V) = \Lambda^{(k)}(V)$ for all $k \geqslant 2$ and $V \in \mathcal{B}(S^k)$, each quantity in (5.21) is no larger that the one on its left. We show show that the rightmost quantity in (5.21) is as large as the leftmost one. Let $V \in \mathcal{C}(S^{\mathbb{N}})$, and set, for all $x \in S^k$,

$$V_k(x) = \inf\{V(\omega) \mid \omega_1 = x_1, \dots, \omega_k = x_k\}, \quad k \geqslant 2.$$

The functions V_k belongs to $\mathcal{B}(S^k)$. When seen as an element of $\mathcal{C}(S^{\mathbb{N}})$, it satisfies $V_k \leqslant V$, thus $\Lambda^{(k)}(V_k) \leqslant \Lambda^{(\infty)}(V)$. Moreover, the sequence (V_k) is nondecreasing, thus by monotone convergence $\langle \mu, V_k \rangle \to \langle \mu, V \rangle$ as $k \to \infty$. Therefore, the rightmost term in (5.21) is bounded below by

$$\limsup_{k \to \infty} \big(\langle \mu^{(k)}, V_k \rangle - \Lambda^{(k)}(V_k) \big) \geqslant \langle \mu, V \rangle - \Lambda^{(\infty)}(V).$$

Taking the supremum over $V \in \mathcal{C}(S^{\mathbb{N}})$ yields the desired bound.

2. For all $k \geqslant 2$ and all $f \in \mathcal{E}(S^k)$, define $f^{*,k}$ as a function on $S^{-\mathbb{N}_0}$ by $f^{*,k}(\omega) = f(\omega_{1-k}, \omega_{-k}, \ldots, \omega_0)$. This is an uniformly continuous function that is bounded from above and below, hence we have

$$\bigcup_{k \geqslant 2} \mathcal{E}(S^k) \subseteq \mathcal{E}_{\mathrm{u}}(S^{-\mathbb{N}_0}) \subseteq \mathcal{E}(S^{-\mathbb{N}_0}).$$

Also notice that

$$\langle \mu^*, \log f^{*,k} \rangle = \langle \mu^{(k)}, \log f \rangle, \qquad \langle \mu^*, \log P^{(\infty)} f^{*,k} \rangle = \langle \mu^{(k)} \log P^{(k)} f \rangle.$$

This proves that each quantity in (5.22) is no larger than the one on its left. We now show that the rightmost quantity in (5.22) is as large as the leftmost one. Let $f \in \mathcal{E}(S^{-\mathbb{N}_0})$, and set, for all $x \in S^k$,

$$f_k(x) = \inf\{f(\omega) \mid \omega_{1-k} = x_1, \omega_{2-k} = x_2, \ldots, \omega_0 = x_k, \}.$$

The function f_k belongs to $\mathcal{E}(S^k)$ and satisfies $f_k^{*,k} \leqslant f$, thus

$$\langle \mu^*, \log P^{(\infty)} f \rangle \geqslant \langle \mu^*, \log P^{(\infty)} f_k^{*,k} \rangle = \langle \mu^{(k)}, \log P^{(k)} f_k \rangle.$$

Moreover, the sequence $(\log f_k^{*,k})$ is nondecreasing, thus by monotone convergence,

$$\langle \mu^{(k)}, \log f_k \rangle = \langle \mu^*, \log f_k^{*,k} \rangle \to \langle \mu^*, \log f \rangle, \qquad k \to \infty.$$

Therefore, the rightmost term in (5.22) is bounded below by

$$\limsup_{k \to \infty} \left(\langle \mu^{(k)}, \log f_k \rangle - \langle \mu^{(k)}, \log P^{(k)} f_k \rangle \right) \geqslant \langle \mu^*, \log f \rangle - \langle \mu^*, \log P^{(\infty)} f \rangle.$$

Taking the supremum over $f \in \mathcal{E}(S^{-\mathbb{N}_0})$ yields the desired bound.

3. Property 3 is known as Pinsker's Lemma. We refer the reader to Lemma 6.5.13 in [11] or to [27] for a proof.

□

Proposition 5.6. *Let $\mu \in \mathcal{P}_{\mathrm{bal}}(S^{\mathbb{N}})$, and assume that $H(\mu^{(1)}|\beta) < \infty$. Then,*

$$R^{(\infty)}(\mu) = \lim_{k \to \infty} \frac{1}{k} H(\mu^{(k)}|\mathbb{P}_k), \tag{5.23}$$

where $H(\mu^{(k)}|\mathbb{P}_k)$ denotes the relative entropy of $\mu^{(k)}$ with respect to $\mathbb{P}_k$.

Proof. Assume first that $R^{(k)}(\mu^{(k)}) = \infty$ for some k. Then, Property 3 of Lemma 5.5, $R^{(l)}(\mu^{(l)}) = R^{(\infty)}(\mu) = \infty$ for all $l \geqslant k$. Since $H(\mu^{(k)}|\mathbb{P}_k) \geqslant 0$, by Lemma 5.3, $H(\mu^{(l)}|\mathbb{P}_l) = \infty$ for all $l \geqslant k$, and (5.23) holds. Assume now that $R^{(k)}(\mu^{(k)})$ is finite for all k. Then, by Lemma 5.3, the relative entropy $H(\mu^{(k)}|\mathbb{P}_k)$ is either finite for all $k \geqslant 1$, or infinite for all $k \geqslant 1$. As we assumed that $H(\mu^{(1)}|\mathbb{P}_1) = H(\mu^{(1)}|\beta) < \infty$, the second case is ruled out. Therefore $R^{(k)}(\mu^{(k)})$ can always be expressed as the telescopic difference between $H(\mu^{(k)}|\mathbb{P}_k)$ and $H(\mu^{(k-1)}|\mathbb{P}_{k-1})$. Hence $\lim_{k \to \infty} \frac{1}{k} H(\mu^{(k)}|\mathbb{P}_k)$ is the Cesaro limit of the sequence $R^{(k)}(\mu^{(k)})$, which is also its regular limit $R^{(\infty)}(\mu)$ by Property 3 of Lemma 5.5. □

A Admissible measures

A.1 Balanced measures

We call *minimal cycle* any word of the form $u = vv_1$ with $v \in S_{\mathrm{fin}}$, such that every letter of v appears exactly once in v. Denote by $\mathcal{C}$ the set of all minimal cycles u satisfying $p(u) > 0$.[43] In this section, we present a construction that was mentioned by de La Fortelle and Fayolle in Propositions 2 and 6 of [21] under the name of balanced measure decomposition. We provide here a detailed proof, including the following crucial lemma, which seems to have been overlooked in [21]. Notations are defined in Section 2.1.

[43] Or equivalently, the set of all minimal cycles u satisfying $L[u] \in \mathcal{D}^{(2)}$.

Lemma A.1. *Let $\nu \in \mathcal{M}(S^2)$ be a balanced non-negative measure. If $\nu \neq 0$, then there exists a minimal cycle v and $\alpha > 0$ such that $\nu \geqslant \alpha M[v]$ in $\mathcal{M}(S^2)$.*

Proof. Since ν is finite and non-zero, we can assume that $\nu \in \mathcal{P}_{\text{bal}}(S^2)$ without loss of generality. Let q be a probability kernel defined by $q(x,y) = \nu(x,y)/\nu^{(1)}(x)$ if $x \in \operatorname{supp} \nu^{(1)}$ and arbitrarily otherwise. We have $\nu^{(1)} \otimes q = \nu$.[44] Let (Z_n) be the Markov chain defined on S by the initial measure $\nu^{(1)}$ and the kernel q. Since ν is balanced, $\nu^{(1)}$ is stationary for the kernel q. Consider a state $z \in \operatorname{supp} \nu^{(1)} \neq \emptyset$. Denoting by $\mathbb{E}_\nu[\cdot]$ the expectation associated to the Markov chain (Z_n), we have

$$\mathbb{E}_\nu\left[\sum_{n=1}^{\infty} \mathbb{1}_{\{Z_n = z\}}\right] = \sum_{n=1}^{\infty} \nu^{(1)}(z) = \infty.$$

Therefore, there is a positive probability that (Z_n) visits z several times. This means that there exists a word u, whose first and last letter are z, satisfying $q(u_i, u_{i+1}) > 0$ for all $i \leqslant |u| - 1$. In particular, there exists a minimal cycle v satisfying $q(v_i, v_{i+1}) > 0$ for all $i \leqslant |v| - 1$. Hence, by definition of $\nu^{(1)}$ and by induction, for all $i \leqslant |v| - 1$,

$$\nu^{(1)}(v_i) \geqslant \nu^{(1)}(v_{i-1}) q(v_{i-1}, v_i) \geqslant \ldots \geqslant \nu^{(1)}(v_1) q(v_1, v_2) \ldots q(v_{i-1}, v_i) > 0.$$

Therefore,

$$\alpha := \min_{1 \leqslant i \leqslant |v|-1} \nu^{(1)}(v_i) q(v_i, v_{i+1}) > 0,$$

and we have $\nu \geqslant \alpha M[v]$. □

Lemma A.2 provides a useful way to approximate certain balanced measures.[45] Recall that $\mathcal{D}^{(2)}$ was defined in (1.8) and $R^{(2)}$ in (3.5). Notice that, by Property 1, each μ_n is (pre-)admissible whenever μ is (pre-)admissible.

Lemma A.2. *Let $\mu \in \mathcal{P}_{\text{bal}}(S^2) \cap \mathcal{D}^{(2)}$. There exists a sequence $(\mu_n) \in \mathcal{P}(S^2)^{\mathbb{N}}$ satisfying the following properties:*

1. *For all $n \in \mathbb{N}$, the measure μ_n is balanced, absolutely continuous with respect to μ, and of finite support.*
2. *$\mu_n \to \mu$ and $R^{(2)}(\mu_n) \to R^{(2)}(\mu)$ as $n \to \infty$.*
3. *For all $n \in \mathbb{N}$, μ_n is of the form*

$$\mu_n = \sum_{k=1}^{n} \alpha_{n,k} M[u^k], \tag{A.1}$$

where $u^k \in \mathcal{C}$ and $\alpha_{k,n} \in [0,1]$ for all $1 \leqslant k \leqslant n$.

Proof. We first construct the words u^k and coefficients $\alpha_{n,k}$ of Property 3. Since $\mathcal{C}$ is countable, we can index its elements and write $\mathcal{C}$ as $\{u^n \mid n \in \mathbb{N}\}$. We define by induction a sequence of nonnegative numbers (a_n) and a sequence of balanced measures[46] $(\nu_n)_{n \geqslant 0}$ by $\nu_0 = 0$ and

$$\nu_n = \sum_{k=1}^{n} a_k M[u^k], \qquad n \geqslant 1$$
$$a_{n+1} = \min\left\{\mu(u_i^n, u_{i+1}^n) - \nu_n(u_i^n, u_{i+1}^n) \,\middle|\, 1 \leqslant i \leqslant |u^n| - 1\right\}, \qquad n \geqslant 0.$$

Since minimal cycles go through each of their letters exactly once, $M[u^{n+1}](x,y) \leqslant 1$ for all $n \in \mathbb{N}$ and all $x, y \in S$. Therefore, $\nu_{n+1}(x,y)$ is equal either to $\nu_n(x,y)$ or $\nu_n(x,y) + a_{n+1}$, implying by induction that $0 \leqslant \nu_n \leqslant \mu$ in $\mathcal{M}(S^2)$ for all $n \geqslant 0$. Let $\nu \in \mathcal{M}(S^2)$ be the limit of (ν_n). It must satisfy $0 \leqslant \nu \leqslant \mu$. If $\nu \neq \mu$, by Lemma A.1, there exists a minimal cycle v and $\alpha > 0$ such that $\mu - \nu \geqslant \alpha M[v]$. Since then $\mu \geqslant \alpha M[v]$ and $\mu \in \mathcal{D}^{(2)}$, the minimal cycle v must be

[44] $\nu^{(1)} \otimes q$ denotes the measure on S^2 defined by $\nu^{(1)} \otimes q(x,y) = \nu^{(1)}(x) q(x,y)$, as in Section 3.1.

[45] As mentioned by de La Fortelle and Fayolle, since measures of the form $L[v]$, where $v \in \mathcal{C}$, are extreme points of the convex set $\mathcal{P}_{\text{bal}}(S)$, this result can also be seen as an extreme point decomposition. When S is finite, $\mathcal{P}_{\text{bal}}(S)$ is compact and the decomposition is simply an application of the Krein-Milman theorem.

[46] Not necessarily probability measures.

an element of $\mathcal{C}$, thus $v = u^n$ for some $n \in \mathbb{N}$. Therefore, there exists an index i such that $a_{n+1} = \mu(u_i^n, u_{i+1}^n) - \nu_n(u_i^n, u_{i+1}^n)$, thus

$$\begin{aligned}\mu(u_i^n, u_{i+1}^n) - \nu(u_i^n, u_{i+1}^n) &\geqslant \mu(u_i^n, u_{i+1}^n) - \nu_{n+1}(u_i^n, u_{i+1}^n)\\ &= \mu(u_i^n, u_{i+1}^n) - \nu_n(u_i^n, u_{i+1}^n) - a_{n+1}\\ &= 0.\end{aligned}$$

Since $M[u^n](u_i^n, u_{i+1}^n) = 1$, this contradicts $\mu - \nu \geqslant \alpha M[u^n]$. Thus $\nu = \mu$.

We now define the measures μ_n of the form (A.1). For technical reasons, they are obtained from ν_n not by simple normalization, but by adding the "missing probability" $1 - \nu_n(S^2)$ to the first cycle u^1 instead. Up to reindexing of (u^n), we can assume $a_1 > 0$ without loss of generality, which ensures that $M[u^1]$ is absolutely continuous with respect to μ. Let, for all $n \in \mathbb{N}$,

$$\alpha_{n,1} = a_1 + \frac{1 - \nu_n(S^2)}{|u^1| - 1}, \qquad \alpha_{n,k} = a_k, \qquad 2 \leqslant k \leqslant n,$$

and let μ_n be the measure defined by (A.1) with these coefficients. By construction, the measure μ_n satisfies Properties 1 and 3.

We now prove Property 2. The first part of Property 2 is also a direct consequence of the construction of (μ_n), because

$$\mu_n = \nu_n + \frac{1 - \nu_n(S^2)}{|u^1| - 1} M[u^1] \to \mu + 0, \qquad n \to \infty.$$

For later purposes, note that if U denotes the set of letters in the word u^1, then for all $n \in \mathbb{N}$ and all $x, y \in S$,

$$\begin{aligned}(x, y) \notin U^2 &\implies \mu_n(x, y) = \nu_n(x, y) \leqslant \mu(x, y),\\ x \notin U &\implies \mu_n^{(1)}(x) = \nu_n^{(1)}(x) \leqslant \mu^{(1)}(x).\end{aligned} \tag{A.2}$$

It remains to prove the convergence of $R^{(2)}(\mu_n)$ to $R^{(2)}(\mu)$. If $R^{(2)}(\mu) = \infty$, by lower semicontinuity of $R^{(2)}$, it follows from $\mu_n \to \mu$ that $R^{(2)}(\mu_n) \to \infty$ as $n \to \infty$. Now, assume that $R^{(2)}(\mu)$ is finite. In particular, $\mu \in \mathcal{D}^{(2)}$ and (3.6) holds. Equation (3.6) also holds for μ_n for all $n \in \mathbb{N}$ because $\mu_n \ll \mu$. For all $n \in \mathbb{N}$, we have

$$\begin{aligned}&R^{(2)}(\mu_n) = R_1(\mu_n) - R_2(\mu_n),\\ &R_1(\mu_n) := H(\mu_n | \mu^{(1)} \otimes p) = \sum_{(x,y) \in S^2} \mu_n(x, y) \log \frac{\mu_n(x, y)}{\mu^{(1)}(x) p(x, y)},\\ &R_2(\mu_n) := H(\mu_n^{(1)} | \mu^{(1)}) = \sum_{x \in S} \mu_n^{(1)}(x) \log \frac{\mu_n^{(1)}(x)}{\mu^{(1)}(x)}.\end{aligned}$$

Since μ_n is finitely supported, both $R_1(\mu_n)$ and $R_2(\mu_n)$ are finite sums, thus they are finite. Let us next compute the limit of $R_1(\mu_n)$. We have

$$\begin{aligned}R_1(\mu_n) &= \sum_{(x,y) \in U^2} k_n(x, y) + \sum_{(x,y) \in (U^2)^c} k_n(x, y),\\ k_n(x, y) &= \mu_n(x, y) \log \frac{\mu_n(x, y)}{\mu^{(1)}(x) p(x, y)}.\end{aligned}$$

Since it involves a finite number of terms, the first sum converges as $n \to \infty$. Let $f : s \mapsto s \log s$ and $g : s \mapsto \max(e^{-1}, |f(s)|)$, defined on $[0, +\infty)$.[47] We have $|f| \leqslant g$, and the function g is nondecreasing. Hence, for all $(x, y) \in (U^2)^c$,

$$\begin{aligned}|k_n(x, y)| &= \mu^{(1)}(x) p(x, y) \Big| f\Big(\frac{\mu_n(x, y)}{\mu^{(1)}(x) p(x, y)}\Big) \Big|\\ &\leqslant \mu^{(1)}(x) p(x, y) g\Big(\frac{\mu_n(x, y)}{\mu^{(1)}(x) p(x, y)}\Big)\\ &\leqslant \mu^{(1)}(x) p(x, y) g\Big(\frac{\mu(x, y)}{\mu^{(1)}(x) p(x, y)}\Big).\end{aligned}$$

[47] We recall the convention that $0 \log 0 = 0$.

The last line used (A.2) and the monotonicity of g. Since g coincides with f on $[1,+\infty)$ and with e^{-1} on $[0,1]$, we get, for all $(x,y) \in (U^c)^2$,

$$|k_n(x,y)| \leqslant k(x,y) := \begin{cases} \mu(x,y)\log\frac{\mu(x,y)}{\mu^{(1)}(x)p(x,y)}, & \text{if } \mu(x,y) > \mu^{(1)}(x,y)p(x,y), \\ e^{-1}\mu^{(1)}(x)p(x,y), & \text{otherwise.} \end{cases}$$

Since $\mu^{(1)} \otimes p$ is a probability measure, we have

$$\sum_{\substack{(x,y)\in(U^2)^c \\ \mu(x,y)\leqslant\mu^{(1)}(x)p(x,y)}} e^{-1}\mu^{(1)}(x)p(x,y) \leqslant e^{-1} < \infty,$$

and since the negative terms in the first sum of the expression of $R^{(2)}$ given in (3.6) cannot get smaller than $-e^{-1}\mu^{(1)}(x)p(x,y)$, we have

$$\sum_{\substack{(x,y)\in(U^2)^c \\ \mu(x,y)>\mu^{(1)}(x)p(x,y)}} \mu(x,y)\log\frac{\mu(x,y)}{\mu^{(1)}(x)p(x,y)} \leqslant R^{(2)}(\mu) + e^{-1} < \infty.$$

Hence the sum of $k(x,y)$ over $(x,y) \in (U^c)^2$ is finite. Therefore, by dominated convergence, we get $R_1(\mu_n) \to R^{(2)}(\mu)$ as $n \to \infty$. Let us compute the limit of $R_2(\mu_n)$. We have

$$R_2(\mu_n) = \sum_{x\in U} \mu_n^{(1)}(x)\log\frac{\mu_n^{(1)}(x)}{\mu^{(1)}(x)} + \sum_{x\in U^c} \mu_n^{(1)}(x)\log\frac{\mu_n^{(1)}(x)}{\mu^{(1)}(x)}.$$

Since it involves a finite number of terms which vanish when $n \to \infty$, the first sum vanishes as $n \to \infty$. By (A.2) and since f is bounded below by $-e^{-1}$, we have, for all $x \in U^c$,

$$-e^{-1}\mu^{(1)}(x) \leqslant \mu_n^{(1)}(x)\log\frac{\mu_n^{(1)}(x)}{\mu^{(1)}(x)} \leqslant 0.$$

Thus, by dominated convergence, the second sum vanishes too as $n \to \infty$. Therefore, $R_2(\mu_n) \to 0$ as $n \to \infty$, which completes the proof of Property 2. ☐

A.2 Admissible measures for finite k

This appendix is dedicated to the study of the set $\mathcal{A}^{(k)}$ of Definition 1.2. As mentioned in the introduction, the cases $k = 1$ and $k = 2$ are the most important ones for the main part of this paper. Larger values of k are only considered to obtain the level-3 weak LDP in Section 5.

Properties of pre-admissibility and admissibility of a measure $\mu \in \mathcal{P}(S^k)$ only depend on the support of μ. In particular, μ is pre-admissible if and only if

$$\sum_{j\in\mathcal{J}} \mu(C_j^k) = 1.$$

If so, μ is admissible if and only if for any given $j_1, j_2 \in \mathcal{J}_\mu$, we have $\beta \rightsquigarrow C_{j_1} \rightsquigarrow C_{j_2}$ or $\beta \rightsquigarrow C_{j_2} \rightsquigarrow C_{j_1}$. For all $k \geqslant 2$, we denote by $\mathcal{A}_{\text{bal}}^{(k)}$ the set of measures that are admissible and balanced; see Section 5.1. We will also consider the set $\mathcal{D}^{(k)}$ defined in (1.8) and (5.1).

Remark A.3. In [21], [22] and Corollary 13.6 of [28], admissible measures are defined as elements of $\mathcal{A}^{(2)} \cap \mathcal{D}^{(2)}$ rather than simply elements of $\mathcal{A}^{(2)}$. Elements of $\mathcal{A}^{(2)} \cap \mathcal{D}^{(2)}$ and of $\mathcal{A}^{(2)}$ differ in the fact that the latter may have transitions that are not ‘allowed’ under p. This apparent conflict between definitions of admissibility is without consequences, since $R(\mu) = J(\mu) = \Lambda^*(\mu) = \infty$ anyway for all $\mu \in \mathcal{A}^{(2)} \setminus \mathcal{D}^{(2)}$ (see Remark 3.2). In particular, the value of the right-hand side of (1.9) and (3.7) remains unchanged when $\mathcal{A}_{\text{bal}}^{(2)}$ is replaced by $\mathcal{A}_{\text{bal}}^{(2)} \cap \mathcal{D}^{(2)}$.

Proposition A.4. *$\mathcal{A}^{(k)}$ and $\mathcal{A}_{\text{bal}}^{(k)}$ are closed in $\mathcal{P}(S^k)$.*

Proof. Since $\mathcal{P}_{\text{bal}}(S^k)$ is closed in $\mathcal{P}(S^k)$, it suffices to show that $\mathcal{A}^{(k)}$ is closed. Let (μ_n) be a sequence in $\mathcal{A}^{(k)}$ that converges to some $\mu \in \mathcal{P}(S^k)$ as $n \to \infty$. For all $u \in S^k$ that does not belong to C_j^k for any $j \in \mathcal{J}$, we have $\mu_n(u) = 0$ for all n, thus $\mu(u) = 0$. Therefore, μ is pre-admissible. Let $j_1, j_2 \in \mathcal{J}_\mu$. There exists $n \in \mathbb{N}$, such that $\mu_n(C_{j_1}^k) > 0$ and $\mu_n(C_{j_2}^k) > 0$. Since μ_n is admissible, $\beta \rightsquigarrow C_{j_1} \rightsquigarrow C_{j_2}$ or $\beta \rightsquigarrow C_{j_2} \rightsquigarrow C_{j_1}$. Therefore, μ is admissible. ☐

Remark A.5. Sets of the form $\{\nu \in \mathcal{A}^{(k)} \mid \mathcal{J}_\nu \subseteq \mathcal{J}_\mu\}$ for $\mu \in \mathcal{A}^{(k)}$ are convex faces of the (possibly infinite dimensional) simplex $\mathcal{P}(S^k)$. The set $\mathcal{A}^{(k)}$, which may be nonconvex, can always be written as a union of faces:

$$\mathcal{A}^{(k)} = \bigcup_{\mu \in \mathcal{A}^{(k)}} \{\nu \in \mathcal{A}^{(k)} \mid \mathcal{J}_\nu \subseteq \mathcal{J}_\mu\}.$$

A visual representation of this decomposition is provided in Figure 5.

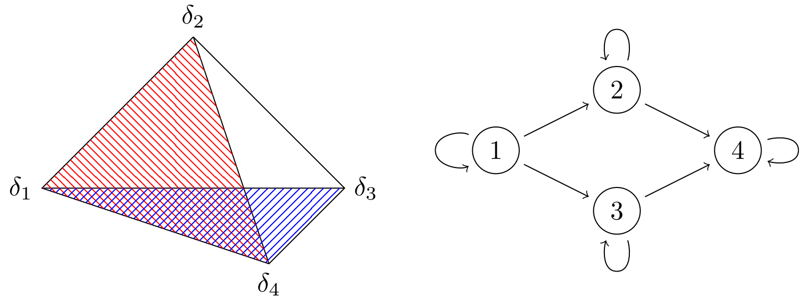


Figure 5: Let (X_n) be the Markov chain on $S = \{1, 2, 3, 4\}$ whose allowed transitions are given by the graph on the right-hand side of the figure, and whose initial measure is $\beta = \delta_1$. The values of transition probabilities do not matter for this example as long as they are positive on each edge of the graph and null otherwise. The set $\mathcal{P}(S)$ is a (finite-dimensional) simplex with extreme points $\{\delta_1, \delta_2, \delta_3, \delta_4\}$. The set $\mathcal{A}^{(1)}$ is the hatched domain, that is the union of the faces $\{\nu \in \mathcal{A}^{(1)} \mid \mathcal{J}_\nu \subseteq \mathcal{J}_{\frac{1}{3}(\delta_1+\delta_2+\delta_4)}\}$ and $\{\nu \in \mathcal{A}^{(1)} \mid \mathcal{J}_\nu \subseteq \mathcal{J}_{\frac{1}{3}(\delta_1+\delta_3+\delta_4)}\}$.

The following two propositions are characterizations of $\mathcal{A}^{(1)}$ and of $\mathcal{A}^{(2)}_{\text{bal}} \cap \mathcal{D}^{(2)}$. Recall that, if $\mu \in \mathcal{A}^{(2)}_{\text{bal}}$, the measure $\mu^{(1)} \in \mathcal{A}^{(1)}$ is defined by (1.6).

Proposition A.6. *We have* $\mathcal{A}^{(1)} = \{\mu^{(1)} \mid \mu \in \mathcal{A}^{(2)}_{\text{bal}}\}$.

Proof. If $\nu \in \mathcal{A}^{(1)}$, consider the measure $\mu \in \mathcal{P}_{\text{bal}}(S^2)$ defined by $\mu(x, y) = \mathbb{1}_{\{x=y\}}\nu(x)$. The measure μ satisfies $\mu^{(1)} = \nu$. Moreover, $\mu(C_j^2) = \nu(C_j)$ hence μ is pre-admissible with $\mathcal{J}_\mu = \mathcal{J}_\nu$, and therefore admissible.[48] Conversely, let $\mu \in \mathcal{A}^{(2)}_{\text{bal}}$. Since $\mu^{(1)}(C_j) = \mu(C_j^2)$, the measure $\mu^{(1)}$ is pre-admissible with $\mathcal{J}_{\mu^{(1)}} = \mathcal{J}_\mu$, thus it is admissible. □

Proposition A.7 characterizes $\mathcal{A}^{(2)}_{\text{bal}}$ in an intuitive way. Although we prove here the full equivalence for the sake of completeness, we will only use the simpler implication (1 ⇒ 2).

Proposition A.7. *Let* $\mu \in \mathcal{P}(S^2)$*. The following properties are equivalent.*

1. *There exists a sequence of words* (w^m) *with* $|w^m| \to \infty$ *such that* $\mathbb{P}_{|w^m|}(w^m) > 0$ *for all* m *and* $L[w^m] \to \mu$ *as* $m \to \infty$.[49]
2. $\mu \in \mathcal{A}^{(2)}_{\text{bal}} \cap \mathcal{D}^{(2)}$.

Proof. Let $\mu \in \mathcal{P}(S^2)$ and assume Property 1. Let us show that $\mu \in \mathcal{D}^{(2)}$. Let $(x, y) \in S^2$ be such that $p(x, y) = 0$. Since $\mathbb{P}_{|w^m|}(w^m) > 0$, the letters x, y cannot be consecutive letters in w^m. Therefore, $L[w^m](x, y) = 0$, thus $\mu(x, y) = 0$. We now prove that μ is balanced. First, observe that

$$\inf_{\nu \in \mathcal{P}_{\text{bal}}(S^2)} |L[w^m] - \nu|_{\text{TV}} \leqslant |L[w^m] - L[w^m w_1^m]|_{\text{TV}} \leqslant \frac{2}{|w^m - 1|} \to 0, \qquad m \to \infty.$$

By the closedness of $\mathcal{P}_{\text{bal}}(S^2)$, μ is balanced. Let us show that μ is admissible. Let $xy \in S^2$ be a two-letter word that does not belong to C_j^2 for any $j \in \mathcal{J}$. Since $\mathbb{P}_{|w^m|}(w^m) > 0$, the subword

[48]Recall that the definition of admissibility of μ does not require $\mu \in \mathcal{D}^{(2)}$, which is why checking admissibility is so easy. See Remark A.3.

[49]The notation $L[w^m]$ is introduced in (2.3)

xy appears at most once in w^m. As a result, $\mu(x,y) = \lim_{n\to\infty} L[w^m](x,y) = 0$. It follows that μ is pre-admissible. Let $j_1, j_2 \in \mathcal{J}_\mu$. There exists some $m \geqslant 1$ such that $L[w^m](C_{j_1}^2) > 0$ and $L[w^m](C_{j_2}^2) > 0$. Hence there exists a letter $x_1 \in C_{j_1}$ and a letter $x_2 \in C_{j_2}$ in w^m, and since w^m has positive probability under $\mathbb{P}$, we have $\beta \rightsquigarrow x_1 \rightsquigarrow x_2$ or $\beta \rightsquigarrow x_2 \rightsquigarrow x_1$. It follows that $\beta \rightsquigarrow C_{j_1} \rightsquigarrow C_{j_2}$ or $\beta \rightsquigarrow C_{j_2} \rightsquigarrow C_{j_1}$, hence μ is admissible, so Property 2 holds.

Conversely, assume now Property 2. Consider the sequence (μ_n), the words u^k and the coefficients $\alpha_{n,k}$ given by Lemma A.2, and let $m \in \mathbb{N}$. We define the word w^m as follows. By Properties 1 and 2 of Lemma A.2, $\mu_n \in \mathcal{A}_{\mathrm{bal}}^{(2)}$, and there exists a certain n such that $|\mu_n - \mu|_{\mathrm{TV}} \leqslant 1/m$. By the admissibility condition, up to reordering words u^k, we can assume that the first letter of each word u^{k+1} is reachable from the last letter of its predecessor u^k, and we fix a word ξ^{k+1} such that $p(u^k\xi^{k+1}u^{k+1}) > 0$. Additionally, u_1^1 is reachable from β, and we can choose a word ξ^1 such that $\beta(\xi^1)p(\xi^1 u^1) > 0$. Let $N \in \mathbb{N}$ be such that

$$\lambda_{N,k} := \lfloor \alpha_{n,k} N \rfloor \geqslant 1, \qquad 1 \leqslant k \leqslant n.$$

Since each u^k is a minimal cycle, we can set $u^k = \tilde{u}^k \tilde{u}_1^k$. We set

$$v^N = \xi^1 \left(\tilde{u}^1\right)^{\lambda_{N,1}} u_1^1 \xi^2 \left(\tilde{u}^2\right)^{\lambda_{N,2}} u_1^2 \cdots \xi^k \left(\tilde{u}^n\right)^{\lambda_{N,n}} u_1^n, \qquad \ell_N = |v^N| - 1.$$

Let τ be the maximal length of words ξ^k. We have

$$\begin{aligned}
\left|L[v^N] - \mu\right|_{\mathrm{TV}} &\leqslant \frac{\tau n}{\ell_N} + \frac{n}{\ell_N} + \left|\sum_{k=1}^n \frac{\lambda_{N,k}}{\ell_N} M[u^k] - \mu\right|_{\mathrm{TV}} \\
&\leqslant \frac{(\tau+1)n}{\ell_N} + |\mu_n - \mu|_{\mathrm{TV}} + \left|\sum_{k=1}^n \left(\frac{\lambda_{N,k}}{\ell_N} - \alpha_{n,k}\right) M[u^k]\right|_{\mathrm{TV}} \\
&\leqslant \frac{(\tau+1)n}{\ell_N} + \frac{1}{m} + \sum_{k=1}^n \left|\frac{\lambda_{N,k}}{\ell_N} - \alpha_{n,k}\right| (|u^k| - 1).
\end{aligned}$$

When $N \to \infty$, the first term vanishes. Moreover, we have $|v^N| = N + o(N)$, thus for all $1 \leqslant k \leqslant n$, the coefficient $\lambda_{N,k}/\ell_N$ converges to $\alpha_{n,k}$, thus each term of the sum vanishes. It follows that the sum vanishes because n is fixed. Consequently, there exists $N(m) \in \mathbb{N}$ such that for all $N \geqslant N(m)$,

$$\left|L[v^N] - \mu\right|_{\mathrm{TV}} \leqslant \frac{1}{m} + \frac{1}{m} + \frac{1}{m}.$$

The word $w^m := v^{N(m)}$ satisfies $\mathbb{P}_{|w^m|}(v^m) > 0$ and $|L[w^m] - \mu|_{\mathrm{TV}} \leqslant 3/m$. Since it is always possible to choose $N(m+1) \geqslant N(m) + 1$, the length of w^m is at least m, thus $|w^m| \to \infty$ as $m \to \infty$. ☐

Remark A.8. Proposition A.7 is adapted from Proposition 1 of [21], which is, unfortunately, false as formulated. Indeed, the set described in this proposition actually coincides with $\mathcal{A}_{\mathrm{bal}}^{(2)} \cap \mathcal{D}^{(2)}$, which sometimes differs from the set $\mathcal{M}$ defined in [21]. See Example D.2 for instance. This overlook seems to have no consequence in the proofs in [21], provided we replace the definition of $\mathcal{M}$ with that of $\mathcal{A}_{\mathrm{bal}}^{(2)} \cap \mathcal{D}^{(2)}$.

A.3 Admissible measures for infinite k

In this appendix, we study the set $\mathcal{A}_{\mathrm{bal}}^{(\infty)}$ introduced in Definition 5.4. We show that $\mathcal{A}_{\mathrm{bal}}^{(\infty)}$ is the projective limit of $(\mathcal{A}_{\mathrm{bal}}^{(k)})_{k\geqslant 1}$, defined in Definition 1.2. The objects of this appendix, for instance measures $\mu^{(k)}$, where $\mu \in \mathcal{P}(S^{\mathbb{N}})$, and the applications $\pi_k : \mu \mapsto \mu^{(k)}$, were introduced at the beginning of Section 5.2.

Lemma A.9. *The set $\mathcal{A}_{\mathrm{bal}}^{(\infty)}$ is the projective limit of $(\mathcal{A}_{\mathrm{bal}}^{(k)})$. In other words, $\mathcal{A}_{\mathrm{bal}}^{(\infty)}$ is closed and satisfies*

$$\mathcal{A}_{\mathrm{bal}}^{(\infty)} = \bigcap_{k=1}^{\infty} \pi_k^{-1}(\mathcal{A}_{\mathrm{bal}}^{(k)}). \tag{A.3}$$

Proof. Since every $\mathcal{A}^{(k)}_{\text{bal}}$ is closed and every π_k is continuous, it suffices to prove (A.3). Let $\mu \in \mathcal{A}^{(\infty)}_{\text{bal}}$ and $k \geqslant 1$. The measure $\mu^{(k)}$ is balanced, and satisfies

$$\mu^{(k)} = \pi_k \left(\sum_{j \in \mathcal{J}_\mu} \mu|_{C_j^{\mathbb{N}}} \right) = \sum_{j \in \mathcal{J}_\mu} \pi_k \left(\mu|_{C_j^{\mathbb{N}}} \right) = \sum_{j \in \mathcal{J}_\mu} \mu^{(k)}|_{C_j^k}. \tag{A.4}$$

Thus $\mu^{(k)}$ is pre-admissible and $\mathcal{J}_{\mu^{(k)}} \subseteq \mathcal{J}_\mu$. Moreover, for all $j \in \mathcal{J}$,

$$\mu^{(k)}(C_j^k) = \mu(C_j^k \times S^{\mathbb{N}}) = \mu(C_j^{\mathbb{N}}),$$

thus $\mathcal{J}_\mu$ is minimal in (A.4) and we have $\mathcal{J}_{\mu^{(k)}} = \mathcal{J}_\mu$. Since the order $\rightsquigarrow$ is total on $(C_j)_{j \in \mathcal{J}_\mu}$ and $\beta \rightsquigarrow C_j$ for all $j \in \mathcal{J}_\mu$, it follows that $\mu^{(k)} \in \mathcal{A}^{(k)}_{\text{bal}}$. We now prove the converse inclusion. Let $\mu \in \bigcap_{k=1}^\infty \pi_k^{-1}(\mathcal{A}^{(k)}_{\text{bal}})$. Then, μ is shift-invariant. It remains to prove that $\mu \in \mathcal{A}^{(\infty)}$. More precisely, we will show that μ is pre-admissible with $\mathcal{J}_\mu = \mathcal{J}_{\mu^{(1)}}$, and therefore admissible. It suffices to show that

$$\sum_{j \in \mathcal{J}_{\mu^{(1)}}} \mu(C_j^{\mathbb{N}}) = 1. \tag{A.5}$$

Let $k \in \mathbb{N}$. For all $j \in \mathcal{J}$, we have

$$\mu^{(1)}(C_j) = \mu(C_j \times S^{\mathbb{N}}) = \mu^{(k)}(C_j \times S^{k-1}) = \mu^{(k)}(C_j^k),$$

where the last equality uses that $\mu^{(k)}$ is pre-admissible. Thus, the sequence $(\mu^{(k)}(C_j^k))_{k \geqslant 1}$ is constant for every $j \in \mathcal{J}$. Therefore,

$$\mu(C_j^{\mathbb{N}}) = \mu\left(\bigcap_{k \in \mathbb{N}} C_j^k \times S^{\mathbb{N}} \right) = \lim_{k \to \infty} \mu^{(k)}(C_j^k) = \mu^{(1)}(C_j).$$

By definition of $\mathcal{J}_{\mu^{(1)}}$, (A.5) is satisfied. □

B Convex conjugates and duality

Let $k \in \mathbb{N}$. We recall that $\mathcal{B}(S^k)$ denotes the set of all bounded functions on S^k and $\mathcal{M}(S^k)$ denotes the set of all finite signed measures on S^k. In addition to the weak topology on $\mathcal{M}(S^k)$ defined in the introduction, we define the *weak topology* on $\mathcal{B}(S^k)$ as the coarsest topology such that $\langle \mu, \cdot \rangle$ is continuous for all $\mu \in \mathcal{M}(S^k)$. The sets $\mathcal{M}(S^k)$ and $\mathcal{B}(S^k)$ are equipped with their respective weak topologies.[50] We recall that $\mathcal{P}(S^k)$ denotes the set of probability measures on S^k.

Definition B.1. *Let $\Lambda : \mathcal{B}(S^k) \to (-\infty, +\infty]$ be a function. The* convex conjugate *of Λ is defined over $\mathcal{M}(S^k)$ by*

$$\Lambda^*(\mu) = \sup_{V \in \mathcal{B}(S^k)} \left(\langle \mu, V \rangle - \Lambda(V) \right).$$

Let $J : \mathcal{M}(S^k) \to (-\infty, +\infty]$ be a function. The convex conjugate of J is defined over $\mathcal{B}(S^k)$ by

$$J^*(V) = \sup_{\mu \in \mathcal{M}(S^k)} \left(\langle \mu, V \rangle - J(\mu) \right).$$

Let $I : \mathcal{P}(S^k) \to (-\infty, +\infty]$ be a function. Let J be the extension of I to $\mathcal{M}(S^k)$ defined by setting $J(\mu) = \infty$ for all $\mu \notin \mathcal{P}(S^k)$. The convex conjugate of I is defined over $\mathcal{B}(S^k)$ as the convex conjugate of J and is still denoted I^*.

The convex conjugate of a function on $\mathcal{B}(S^k)$, $\mathcal{M}(S^k)$, or $\mathcal{P}(S^k)$ is always a lower semicontinuous function because it is the supremum of a family of continuous functions. The notion of convex conjugation is designed for vector spaces, which is the reason why we need to extend the function

[50] We follow the same philosophy as in Chapter 2.3 of [31]: rather than taking a Banach space and considering its dual equipped with the operator norm, we define simultaneously two topologies on two sets that make them the dual of each other. Beware that this weak topology on $\mathcal{B}(S^k)$ is not the topology of uniform convergence!

$I : \mathcal{P}(S^k) \to (-\infty, +\infty]$ to $\mathcal{M}(S^k)$ before defining its convex conjugate. Nevertheless, we still have

$$I^*(V) = \sup_{\mu \in \mathcal{P}(S^k)} \left(\langle \mu, V \rangle - I(\mu) \right), \qquad V \in \mathcal{B}(S^k),$$

as a consequence of Definition B.1.

Let $I : \mathcal{P}(S^k) \to (-\infty, +\infty]$ be a function. If I is finite at one point (at least), I^* is a function from $\mathcal{B}(S^k)$ to $(-\infty, +\infty]$. The convex biconjugate of I is the function $I^{**} := (I^*)^*$ defined on $\mathcal{M}(S^k)$. Otherwise, I^{**} is set to be the infinite function. The function I^{**} is described by the Fenchel-Moreau Theorem; see for instance Theorems 2.3.3 and 2.3.4 of [31]. In this paper, we only use the following corollary. For a function $I : \mathcal{P}(S^k) \to (-\infty, +\infty]$, we denote by $\operatorname{cl} I$ the lower semicontinuous envelope of I and by $\operatorname{co} I$ the convex envelope of I.

Corollary B.2. *Let $I : \mathcal{P}(S^k) \to (-\infty, +\infty]$ be a function. Then, for all $\mu \in \mathcal{P}(S^k)$,*

$$I^{**}(\mu) = \operatorname{cl}(\operatorname{co} I)(\mu).$$

Proof. If I is infinite everywhere, we have $I^{**} = \infty = \operatorname{co} I$ everywhere. Otherwise, let J be the extension of I to $\mathcal{M}(S^k)$ defined as in Definition B.1. By the Fenchel-Moreau Theorem (see Theorem 2.3.4 of [31]), and by definition of I^{**}, we have

$$I^{**}(\mu) = J^{**}(\mu) = \operatorname{cl}_{\mathcal{M}(S^k)}(\operatorname{co}_{\mathcal{M}(S^k)} J)(\mu), \qquad \mu \in \mathcal{P}(S^k),$$

where $\operatorname{cl}_{\mathcal{M}(S^k)}(\cdot)$ denotes the lower semicontinuous envelope taken on $\mathcal{M}(S^k)$ and $\operatorname{co}_{\mathcal{M}(S^k)}(\cdot)$ denotes the convex envelope taken on $\mathcal{M}(S^k)$. Since $\mathcal{P}(S^k)$ is convex in $\mathcal{M}(S^k)$, by definition of J, we have, for all $\mu \in \mathcal{M}(S^k)$,

$$\operatorname{co}_{\mathcal{M}(S^k)} J(\mu) = \begin{cases} \operatorname{co} I(\mu), & \text{if } \mu \in \mathcal{P}(S^k), \\ \infty, & \text{otherwise.} \end{cases}$$

Since $\mathcal{P}(S^k)$ is closed in $\mathcal{M}(S^k)$, it follows that for all $\mu \in \mathcal{P}(S^k)$,

$$\operatorname{cl}_{\mathcal{M}(S^k)}(\operatorname{co}_{\mathcal{M}(S^k)} J)(\mu) = \operatorname{cl}(\operatorname{co} I)(\mu),$$

which concludes the proof. □

C The finite case

In this section, we consider the case of a finite state space S. This assumption, as it implies that $\mathcal{P}(S)$ is compact, provides tightness and goodness of rate functions. By doing so, it greatly simplifies many proofs in the paper. Nevertheless, we stress that finiteness does not imply any kind of irreducibility. In the following theorem, functions $(\Lambda^{(1)})^*$, $J^{(1)}$, $R^{(1)}$, $(\Lambda^{(\infty)})^*$, $J^{(\infty)}$ and $R^{(\infty)}$ and sets $\mathcal{A}^{(1)}$ and $\mathcal{A}^{(\infty)}_{\mathrm{bal}}$ are as in Theorems 1.4 and 1.5.

Theorem C.1. *Assume that S is finite. Then, the following hold:*

1. *(Level-1 full LDP) For all $f : S \to \mathbb{R}^d$, the sequence $(A_n f)$ satisfies the full LDP with good rate function I_f, which satisfies*

$$I_f(a) = \inf\{I^{(1)}(\mu) \mid \mu \in \mathcal{P}(S),\ \langle \mu, f \rangle = a\}, \qquad a \in \mathbb{R}^d,$$

where $I^{(1)}$ is the function of Property 2;

2. *(Level-2 full LDP) The sequence $(L_n^{(1)})$ satisfies the full LDP with good rate function $I^{(1)}$, which satisfies, for all $\mu \in \mathcal{P}(S)$,*

$$I^{(1)}(\mu) = \begin{cases} (\Lambda^{(1)})^*(\mu) = J^{(1)}(\mu) = R^{(1)}(\mu), & \text{if } \mu \in \mathcal{A}^{(1)}, \\ \infty, & \text{otherwise;} \end{cases}$$

3. *(Level-3 full LDP) The sequence* $(L_n^{(\infty)})$ *satisfies the full LDP with good rate function* $I^{(\infty)}$*, which satisfies, for all* $\mu \in \mathcal{P}(\mathcal{S}^{\mathbb{N}})$*,*

$$I^{(\infty)}(\mu) = \begin{cases} (\Lambda^{(\infty)})^*(\mu) = J^{(\infty)}(\mu) = R^{(\infty)}(\mu), & \text{if } \mu \in \mathcal{A}_{\mathrm{bal}}^{(\infty)}, \\ \infty, & \text{otherwise.} \end{cases}$$

Moreover, assuming that $\mu \in \mathcal{A}_{\mathrm{bal}}^{(\infty)}$ *and* $H(\mu^{(1)}|\beta) < \infty$*, we have the additional expression*

$$I^{(\infty)}(\mu) = \lim_{k\to\infty} \frac{1}{k} H(\mu^{(k)}|\mathbb{P}_k).$$

Theorem C.1 could be obtained as a corollary of Theorems 1.4 and 1.5. However, in order to highlight how the finiteness assumption simplifies many proofs, we propose a sketch of a direct proof by going through the paper again with this assumption in mind.

Sketch of proof. Here are the main steps of the proof, and how much simpler they are.

- The slicing, stitching, coupling and decoupling maps defined in Sections 2.4 and 2.5 have the same definition as in the infinite case. However, their properties are simpler to prove. Indeed, since S is now finite, we do not need to introduce a finite set $K \subseteq S$ to be able to bound probabilities and lengths of words (that is, we can simply take $K = S_1 = S$ throughout). Moreover, the sums that appear, for example in (2.23), can be bounded by simply counting the number of terms and crudely replacing the probabilities being summed by 1; see Example 2.20 of [6].

- We then use the RL method of Section 2.2 to show that $(L_n^{(2)})$ has a RL function. By Lemma 2.3, $(L_n^{(2)})$ satisfies the weak LDP with rate function $I^{(2)}$. But since $\mathcal{P}(S^2)$ is compact, the LDP is full and the rate function good.

- We derive Proposition 3.4 using convexity properties of $I^{(2)}$ as in Section 3.2.

- By Varadhan's Lemma (the standard version for full LDPs; see Section III.3 in [12] or Theorem 4.3.1 of [11]), we have $(I^{(2)})^* = \Lambda^{(2)} = \Lambda_\infty^{(2)}$ everywhere. Section 3.3 can be skipped entirely.

- The content of Section 3.4 is standard in the finite case. To derive $J^{(2)} = R^{(2)}$, one can reproduce the proof of theorem 13.1 of [28], as we did in Section 3.4, but an alternative proof, that gets simpler in finite dimension, consists in explicitly optimizing (3.2).[51] We then have to prove that $(\Lambda^{(2)})^* = J^{(2)}$ on $\mathcal{A}_{\mathrm{bal}}^{(2)}$. We outline here a proof of this standard equality, following the same approach as in the proof of Proposition 4.4.[52] Let $\mu \in \mathcal{A}_{\mathrm{bal}}^{(2)}$. We deduce the inequality $J^{(2)}(\mu) \leqslant (\Lambda^{(2)})^*(\mu)$ like in step 1 of the proof of Proposition 3.11. The proof of the converse inequality is simplified in the following way. Since the set $S_\beta := \{x \in S \mid \beta \rightsquigarrow x\}$ is finite, for all $V \in \mathcal{B}(S^2)$,

$$\Lambda^{(2)}(V) = \max_{x\in S_\beta} \left(\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{E}_x \left[e^{n\langle L_n^{(2)}, V\rangle} \right] \right) =: \widetilde{\Lambda}^{(2)}(V).$$

  Here, $\mathbb{E}_x[\cdot]$ denotes the expectation associated with the Markov chain conditioned to $(X_1 = x)$. For all $\lambda > \widetilde{\Lambda}^{(2)}(V)$, the expression

$$u_n(x) = \sum_{k=0}^{n} e^{-\lambda k} (P^V)^k 1(x), \qquad x \in S,$$

  defines a function $u_n \in \mathcal{E}(S^2)$. After observing that $Pu_n = e^{V-\lambda}(u_{n+1} - 1)$ and that (u_n) converges pointwise to a function u satisfying $u(x) > 1$ for all $x \in S_\beta$, we obtain

$$J^{(2)}(\mu) \geqslant \langle \mu, V\rangle - \lambda + \sum_{x\in S_\beta} \mu(x) \liminf_{n\to\infty} \log \frac{u_n(x)}{u_{n+1} - 1} \geqslant \langle \mu, V\rangle - \lambda + 0.$$

[51] The supremizer is the function $(x, y) \mapsto \mathbb{1}_{p(x,y)>0}\mu(x, y)/\mu^{(1)}(x)p(x, y)$, though it does not belong to $\mathcal{E}(S^2)$, and should be approximated to get $J^{(2)}$.

[52] Since $\Lambda^{(2)}$ and $\Lambda_\infty^{(2)}$ are equal, there is no need to reproduce the full proof of Proposition 3.11 for comparing $J^{(2)}$ with $(\Lambda_\infty^{(2)})^*$.

This shows $J^{(2)}(\mu) \geqslant (\Lambda^{(2)})^*(\mu)$. Notice that we did not use Lemma A.2 in this reasoning.

- Since the LDP is full, the content of Section 4 can be entirely replaced by the simple use of the standard contraction principle (see Section III.5 in [12] or Theorem 4.2.1 of [11]) to obtain the full LDP for $(L_n^{(1)})$. Recall also that, in the infinite case, Remark 4.2 prevented us from using Lemma 4.1 to obtain the LDP for $(A_n f)$. In the finite case, the standard contraction principle yields the full LDP for $(A_n f)$.
- We use the same proofs as in Section 5 to get the level-3 full LDP. In the proof of Theorem 1.5, we can use the Dawson-Gärtner Theorem.

□

Remark C.2. An even shorter route to the level-2 full LDP of Theorem C.1 is to perform the same steps while replacing $L_n^{(2)}$ by $L_n^{(1)}$ everywhere. This makes the formulation of certain statements simpler and skips the use of a contraction principle. This approach would not have worked in the infinite S case, because we needed measures on S^2 in Section 3.4. More precisely, Step 3 of the proof of Proposition 3.11 used the decomposition of balanced measures given in Lemma A.2.

D Examples and counterexamples

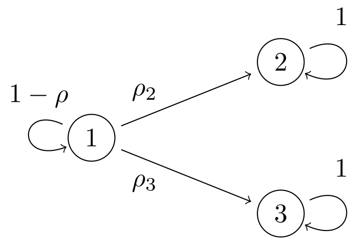


Figure 6: The 3-state Markov chain of Example D.1.

Example D.1. Let $S = \{1, 2, 3\}$, $\beta = \delta_1$, and

$$p = \begin{pmatrix} 1-\rho & \rho_2 & \rho_3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

where $\rho_2, \rho_3 \in [0, 1]$ are such that $\rho = \rho_2 + \rho_3 \leqslant 1$. Although very simple and easily solvable by hand, the problem in this example does not fit into the irreducibility framework. We get

$$\mathcal{A}^{(1)} = \{\lambda\delta_1 + (1-\lambda)\delta_2 \mid \lambda \in [0, 1]\} \cup \{\lambda\delta_1 + (1-\lambda)\delta_3 \mid \lambda \in [0, 1]\},$$

and $L_n^{(1)}$ satisfies the full LDP with rate function I, where

$$I(\mu) = \begin{cases} -\mu(1)\log(1-\rho), & \text{if } \mu \in \mathcal{A}^{(1)}, \\ \infty, & \text{otherwise.} \end{cases}$$

$I^{(1)}$ is not convex, since $I(\frac{1}{2}\delta_2 + \frac{1}{2}\delta_3) = \infty$ and $I(\delta_2) = I(\delta_3) = 0$. The same argument holds for $I^{(2)}$.

Example D.2 (The right-only one-dimensional random walk)**.** Let $S = \mathbb{Z}$ be equipped with the stochastic kernel defined by $p(x, \cdot) = \rho\delta_{x+1} + (1-\rho)\delta_x$, where $\rho \in [0, 1]$ is a parameter. This kernel is shown in Figure 7. Let $\beta \in \mathcal{P}(S)$ satisfy $\inf \operatorname{supp} \beta = -\infty$. If $\rho = 1$, there are no irreducible classes. In this case $\mathcal{A}^{(1)} = \emptyset$ and $L_n^{(1)}$ satisfies the weak LDP with rate function $I_1 = \infty$. If $\rho \in (0, 1)$, all singletons are irreducible classes, and $\mathcal{A}^{(1)} = \mathcal{P}(S)$. Then, $L_n^{(1)}$ satisfies the weak LDP with rate function I_ρ defined as the constant function $-\log(1-\rho)$. If $\rho = 0$, all singletons

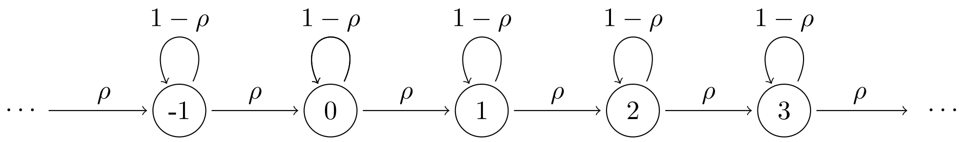


Figure 7: The right-only one-dimensional random walk on $\mathbb{Z}$.

are irreducible classes, but none is reachable from another. Thus $\mathcal{A}^{(1)} = \{\delta_x \mid x \in \operatorname{supp}\beta\}$ and $(L_n^{(1)})$ satsfies the weak LDP with rate function I_0 defined by

$$I_0(\mu) = \begin{cases} 0, & \text{if } \exists x \in \operatorname{supp}\beta,\ \mu = \delta_x\ , \\ \infty, & \text{otherwise.} \end{cases}$$

This example illustrates several interesting facts:

1. When $\rho = 1$, the functions $\Lambda^{(1)}$ and $\widetilde{\Lambda}^{(1)}$ of the proof of Proposition 4.4 may not coincide. Let $V(x) = \mathbb{1}_{x \geqslant 0}$ and let $\beta = \sum_{k \geqslant 1} 2^{-k}\delta_{-k}$. Then, $\Lambda^{(1)}(V) = 1 - \log 2 > 0 = \widetilde{\Lambda}^{(1)}(V)$.

2. When $\rho \in (0, 1)$, no state is recurrent, yet the rate function is not everywhere infinite.

3. When $\rho \in (0, 1)$, $\mathcal{A}^{(1)} = \mathcal{P}(S)$ and I_ρ is convex, but the Markov chain is not irreducible. Although the convexity of the rate function is closely related to the irreducibility of the Markov chain, it is not a sufficient condition.

4. When $\rho \in (0, 1)$, the set $\mathcal{M}$ defined in [21] and mentioned in Remark A.8 is the set of balanced measures whose support is bounded from below. It does not coincide with $\mathcal{A}_{\text{bal}}^{(2)} \cap \mathcal{D}^{(2)}$. As a consequence, the statement provided in Theorem 7 of [21] is false. For instance, the measure $\mu = \sum_{x<0} 2^x \delta_{(x,x)}$ is admissible, thus $I_\rho(\mu) = -\log(1-\rho) < \infty$, but $\mu \notin \mathcal{M}$. As mentioned in Remark A.8, Theorem 7 of [21] holds if the definition of $\mathcal{M}$ is replaced by that of $\mathcal{A}_{\text{bal}}^{(2)} \cap \mathcal{D}^{(2)}$.

5. When $\rho \in (0, 1)$, one can compute that $\Lambda_\infty(0) = \log\rho < 0 = \Lambda(0)$, showing that the inequality $\Lambda_\infty \leqslant \Lambda$ is not an equality in general.

6. When $\rho \in (0, 1)$, Λ is not lower semicontinuous in the weak topology on $\mathcal{B}(S^2)$.[53] If it were, since it is also convex, the Fenchel-Moreau Theorem would cause $\Lambda^{**} = \Lambda$ on $\mathcal{B}(S^2)$ (see Theorem 2.3.3 of [31] for instance). Yet this equality is false because $\Lambda^{**}(0) = I_\rho^*(0) = \log(1-\rho) < 0 = \Lambda(0)$.

7. When $\rho = 0$, let $\operatorname{co} I_0$ denote the convex envelope of I_0. One can compute that $\operatorname{co} I_0(\mu)$ is zero if the support of μ is finite and ∞ otherwise. This function is not lower semicontinuous. Moreover, both I_0^{**} and the σ-convex envelope of I_0 are the zero function. Hence, $\operatorname{co} I_0$ does agree with I_0^{**} on $\mathcal{A}^{(1)}$ but not on $\mathcal{P}(S)$.

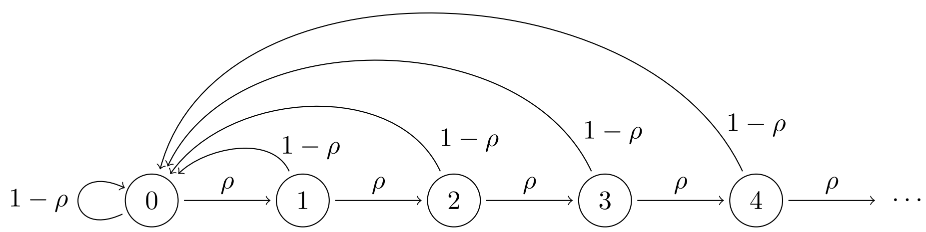


Figure 8: The Markov chain of Example D.3. The state 0 is reachable from every other state in one step and the state x is reachable from 0 in x steps, thus the Markov chain is matrix-irreducible.

[53] Yet it is 1-Lipschitz, thus continuous with respect to the uniform convergence topology. The weak topology is defined in Appendix B.

Example D.3. Let $S = \mathbb{N}_0 = \{0, 1, 2, \dots\}$, and $p(x, \cdot) = \rho\delta_{x+1} + (1-\rho)\delta_0$ for all x, where ρ is a parameter in $(0, 1)$. The kernel p is represented in Figure 8. This Markov chain is matrix-irreducible. As mentioned in the introduction, this example shows that, in absence of exponential tightness, it is a very subtle question to determine whether the LDP is full or not. For every initial measure β, the empirical measure satisfies the weak LDP. If $\beta = \delta_0$, one can show that this LDP is actually full. If $\beta \neq \delta_0$, the empirical measure fails to satisfy the full LDP, as proved in detail in Example 10.3 of [10] and in [3].

Example D.4. Let $S = \mathbb{N}$ and p be defined by $p(n, \cdot) = e^{-n}\delta_n + (1 - e^{-n})\delta_{n+1}$ for all $n \in \mathbb{N}$. Let $\beta = \delta_1$. Then, we have

$$\mathcal{A}^{(2)}_{\text{bal}} \cap \mathcal{D}^{(2)} = \mathcal{A}^{(2)}_{\text{bal}} = \mathcal{P}(\{(n, n) \mid n \in \mathbb{N}\}).$$

Let

$$\mu = \sum_{n=1}^{\infty} \frac{6}{\pi^2 n^2} \delta_{(n,n)} \in \mathcal{A}^{(2)}_{\text{bal}} \cap \mathcal{D}^{(2)}.$$

By Theorem 1.3, μ satisfies

$$I^{(2)}(\mu) = R^{(2)}(\mu) = \sum_{n=1}^{\infty} \frac{6}{\pi^2 n^2} n = \infty.$$

In this example, the inclusion of the domain of $I^{(2)}$ – that is the set $\{\mu \in \mathcal{P}(S^2) \mid I^{(2)}(\mu) < \infty\}$ – in $\mathcal{A}^{(2)}_{\text{bal}} \cap \mathcal{D}^{(2)}$ is strict.

In the following example, from [5], the Markov chain satisfies a level-2 weak LDP but fails to satisfy a level-1 weak LDP. By doing so, it disproves any general statement of a contraction principle from a weak level-2 LDP to a weak level-1 LDP.

Example D.5. Let $n_k = 3^k$ if $k \geqslant 1$ and $n_0 = 1/2$, and let $S = \{(k, j) \in \mathbb{N}_0^2 \mid 1 \leqslant j \leqslant 2n_k\}$. Let p be defined by Figure 9, with the following parameters: $\rho \in (0, 1)$ such that $\alpha := 1 - 4\log(1-\rho) > 0$, and $q_k = C_1 e^{-\alpha n_k}$, where C_1 is such that $q_1 + q_2 + \ldots = 1 - \rho$. Let $\beta(k, j) = C_2(1-\rho)^{j-1} q_k$

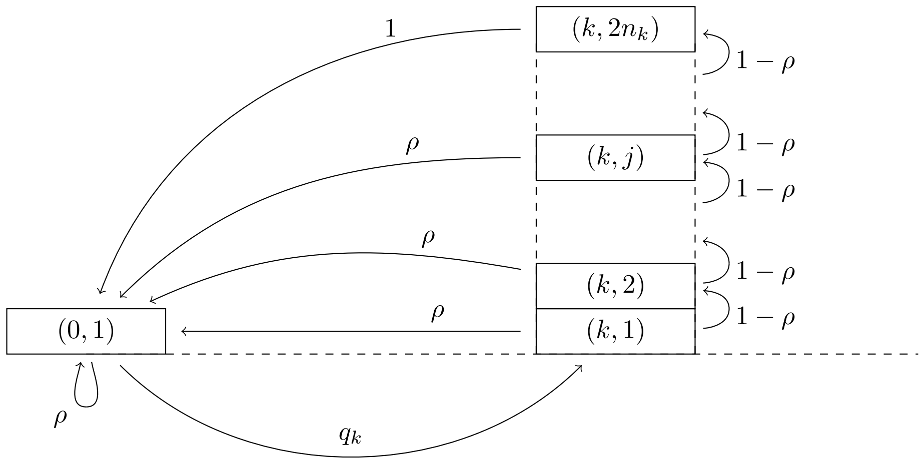


Figure 9: Visual representation of the Markov chain of example D.5. There is one 'tower' as represented on the right of the picture per value of $k \in \mathbb{N}$.

where C_2 is a normalizing constant. This is a matrix-irreducible Markov chain. We consider the following observable:

$$f(k, j) = \begin{cases} +1, & j \geqslant n_k + 1, \\ -1, & j \leqslant n_k. \end{cases}$$

Suppose that the Markov chain satisfies a level-1 weak LDP. Denote $A_n = A_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$. Since (A_n) takes its values in the compact set $[-1, 1]$, it must satisfy the full LDP. Therefore, by Varadhan's lemma, the limit

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[e^{3\alpha A_n}\right]$$

exists. But [5] provided a proof that it actually does not converge. Eventually, this shows by contradiction that the level-1 weak LDP is not valid. The level-2 weak LDP remains valid by Theorem 1.4.

E Glossary

Markov chain

- S, countable state space, 2
- (X_n), Markov chain, 2
- β, initial measure, 2
- p, transition kernel, 2, 7
- $\mathbb{P}$, law of the Markov chain, 2
- $\mathbb{P}_t$, induced law on S^t, 7
- $\mathbb{E}$, expectation, 26
- $\rightsquigarrow$, reachability relation, 4
- (C_j), irreducible classes, 4
- $\mathcal{J}$, 4
- L_n, short for $L_n^{(2)}$, 4
- $L_n^{(k)}$, empirical measure, 2, 4, 38

Sets

- $\mathcal{B}(\cdot)$, bounded measurable functions, 4
- $\mathcal{E}(\cdot)$, 26, 35, 39, 44
- $\mathcal{C}(\cdot)$, continuous functions, 43
- $\mathcal{C}_\mathrm{u}(\cdot)$, uniformly continuous functions, 43
- $\mathcal{E}_\mathrm{u}(\cdot)$, 45
- $\mathcal{M}(\cdot)$, finite signed measures, 4
- $\mathcal{P}(\cdot)$, probability measures, 4, 43
- $\mathcal{P}_\mathrm{bal}(\cdot)$, balanced probability measures, 4, 43
- $\mathcal{A}^{(k)}$, admissible measures, 5, 44
- $\mathcal{A}^{(k)}_\mathrm{bal}$, admissible balanced measures, 5, 44
- $\mathcal{D}^{(k)}$, 5, 38

Functions and rate functions

- I, short for $I^{(2)}$, 8
- $I^{(k)}$, rate function, 8, 35, 40, 44
- $\underline{s}, \overline{s}$, 7
- s, RL function, 8
- $\langle\cdot,\cdot\rangle$, dual pairing, 4, 43
- $\|\cdot\|$, supremum norm, 26
- $\cdot^*$, convex conjugate, 43, 52
- $H(\cdot|\cdot)$, relative entropy, 27
- J, short for $J^{(2)}$, 26
- $J^{(k)}$, DV entropy, 26, 35, 39, 44
- R, short for $R^{(2)}$, 27
- $R^{(k)}$, 27, 35, 39, 44
- Λ, short for $\Lambda^{(2)}$, 26
- $\Lambda^{(k)}$, SCGF, 26, 34, 39, 43
- $(\Lambda^{(k)})^*$, 26, 34, 39, 43, 52
- Λ_∞, short for $\Lambda_\infty^{(2)}$, 26
- $(\Lambda_\infty^{(2)})^*$, 26, 52

Measures

- $|\mu|_\mathrm{TV}$, total variation norm of μ, 4
- $\mathcal{B}(\mu,\rho)$, ball around μ, 4
- $\operatorname{supp}\mu$, support of μ, 4
- $\mu^{(k)}$, marginal of μ, 4, 39, 43
- $\mu\otimes q$, 27
- $\mu|_A$, 5
- π_k, 43
- $\pi_{k,k'}$, 35, 39

Words and operations on words, slicing and stitching

- e, empty word, 7
- $p(u)$, 7
- $M[u]$, 7
- $L[u]$, 7
- $|u|$, length of u, 7
- $|\underline{u}|$, total length of $\underline{u}$, 14
- $k_{\underline{u}}$, number of items in $\underline{u}$, 14
- S_fin, set of words, 7
- $S_{\mathrm{fin},+}$, set of words, 7
- W_t, set of words, 7
- $W_{t,S'}$, set of words, 7
- $\mathcal{J}_0$, finite set of indices, 13
- r, number of classes, 13
- K, finite subset of S, 13
- K_j, finite subset of a class, 13
- Γ, 13
- $\xi_{x,y}$, ξ^i, transition words, 15
- η, minimal transition probability, 15
- τ, maximal transition length, 15
- $A^{(t)}$, stitchable lists, 15
- σ, reordering map, 18
- $F_{\mathcal{J}'}$, F, F_n, slicing map, 13, 14
- G_t, stitching map, 15
- $\Psi_{n,t}$, coupling map, 18
- $\widetilde{\Psi}_n$, decoupling map, 23

Acknowledgments

I would like to thank Noé Cuneo for many pivotal conversations on this topic. I am also grateful to Orphée Collin, Ali Ellouze, Armen Shirikyan and Ofer Zeitouni for helpful discussions and

advice.

References

[1] Azencott, R. (1980). Grandes déviations et applications. In *Eighth Saint Flour Probability Summer School—1978 (Saint Flour, 1978)*, Volume 774 of *Lecture Notes in Math.*, pp. 1–176. Springer, Berlin.

[2] Bahadur, R. R. and S. L. Zabell (1979). Large deviations of the sample mean in general vector spaces. *Ann. Probab. 7*(4), 587–621.

[3] Baxter, J. R., N. C. Jain, and S. R. S. Varadhan (1991). Some familiar examples for which the large deviation principle does not hold. *Comm. Pure Appl. Math. 44*(8-9), 911–923.

[4] Bourbaki, N. (1971). *Éléments de mathématique. Topologie générale. Chapitres 1 à 4.* Hermann, Paris.

[5] Bryc, W. and A. Dembo (1996). Large deviations and strong mixing. *Ann. Inst. H. Poincaré Probab. Statist. 32*(4), 549–569.

[6] Cuneo, N., V. Jakšić, C.-A. Pillet, and A. Shirikyan (2019). Large deviations and fluctuation theorem for selectively decoupled measures on shift spaces. *Rev. Math. Phys. 31*(10), 1950036, 54.

[7] Dawson, D. A. and J. Gärtner (1987). Large deviations from the McKean-Vlasov limit for weakly interacting diffusions. *Stochastics 20*(4), 247–308.

[8] de Acosta, A. (1990). Large deviations for empirical measures of Markov chains. *J. Theoret. Probab. 3*(3), 395–431.

[9] de Acosta, A. and P. Ney (1998). Large deviation lower bounds for arbitrary additive functionals of a Markov chain. *Ann. Probab. 26*(4), 1660–1682.

[10] de Acosta, A. D. (2022). *Large deviations for Markov chains*, Volume 229 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge.

[11] Dembo, A. and O. Zeitouni (2010). *Large deviations techniques and applications*, Volume 38 of *Stochastic Modelling and Applied Probability*. Springer-Verlag, Berlin. Corrected reprint of the second (1998) edition.

[12] den Hollander, F. (2000). *Large deviations*, Volume 14 of *Fields Institute Monographs*. American Mathematical Society, Providence, RI.

[13] Deuschel, J.-D. and D. W. Stroock (1989). *Large deviations*, Volume 137 of *Pure and Applied Mathematics*. Academic Press, Inc., Boston, MA.

[14] Dinwoodie, I. H. (1993). Identifying a large deviation rate function. *Ann. Probab. 21*(1), 216–231.

[15] Dinwoodie, I. H. and P. Ney (1995). Occupation measures for Markov chains. *J. Theoret. Probab. 8*(3), 679–691.

[16] Donsker, M. D. and S. R. S. Varadhan (1975a). Asymptotic evaluation of certain Markov process expectations for large time. I. *Comm. Pure Appl. Math. 28*, 1–47.

[17] Donsker, M. D. and S. R. S. Varadhan (1975b). Asymptotic evaluation of certain Markov process expectations for large time. II. *Comm. Pure Appl. Math. 28*, 279–301.

[18] Donsker, M. D. and S. R. S. Varadhan (1976). Asymptotic evaluation of certain Markov process expectations for large time. III. *Comm. Pure Appl. Math. 29*(4), 389–461.

[19] Donsker, M. D. and S. R. S. Varadhan (1983). Asymptotic evaluation of certain Markov process expectations for large time. IV. *Comm. Pure Appl. Math. 36*(2), 183–212.

[20] Ellis, R. S. (2006). *Entropy, large deviations, and statistical mechanics.* Classics in Mathematics. Springer-Verlag, Berlin. Reprint of the 1985 original.

[21] Fayolle, G. and A. de La Fortelle (2002). Entropy and the principle of large deviations for discrete-time Markov chains. *Problemy Peredachi Informatsii 38*(4), 121–135.

[22] Jiang, Y. W. and L. M. Wu (2005). Large deviations for empirical measures of not necessarily irreducible countable Markov chains with arbitrary initial measures. *Acta Math. Sin. (Engl. Ser.) 21*(6), 1377–1390.

[23] Lanford, O. E. (1973). Entropy and equilibrium states in classical statistical mechanics. pp. 1–113.

[24] Lewis, J. and C.-E. Pfister (1995). Thermodynamic probability theory: Some aspects of large deviations. *Russian Mathematical Surveys 50.*

[25] Parthasarathy, K. R. (2005). *Probability measures on metric spaces / K. R. Parthasarathy.* Providence (R.I.): AMS Chelsea Publishing, American Mathematical Society.

[26] Pfister, C.-E. (2002). Thermodynamical aspects of classical lattice systems. In *In and out of equilibrium (Mambucaba, 2000)*, Volume 51 of *Progr. Probab.*, pp. 393–472. Birkhäuser Boston, Boston, MA.

[27] Pinsker, M. S. (1964). *Information and information stability of random variables and processes.* Holden-Day, Inc., San Francisco, Calif.-London-Amsterdam.

[28] Rassoul-Agha, F. and T. Seppäläinen (2015). *A course on large deviations with an introduction to Gibbs measures*, Volume 162 of *Graduate Studies in Mathematics.* American Mathematical Society, Providence, RI.

[29] Ruelle, D. (1965). Correlation functionals. *J. Mathematical Phys. 6*, 201–220.

[30] Wu, L. (2000). Uniformly integrable operators and large deviations for Markov processes. *J. Funct. Anal. 172*(2), 301–376.

[31] Zălinescu, C. (2002). *Convex analysis in general vector spaces.* World Scientific Publishing Co., Inc., River Edge, NJ.