

Analyticity of Entropy Rate of Hidden Markov Chains

Guangyue Han, Brian Marcus

Department of Mathematics
University of British Columbia
Vancouver, B.C., V6T 1Z2

e-mail: ghan, marcus@math.ubc.ca

August 19, 2006

Abstract

We prove that under mild positivity assumptions the entropy rate of a hidden Markov chain varies analytically as a function of the underlying Markov chain parameters. A general principle to determine the domain of analyticity is stated. An example is given to estimate the radius of convergence for the entropy rate. We then show that the positivity assumptions can be relaxed, and examples are given for the relaxed conditions. We study a special class of hidden Markov chains in more detail: binary hidden Markov chains with an unambiguous symbol, and we give necessary and sufficient conditions for analyticity of the entropy rate for this case. Finally, we show that under the positivity assumptions the hidden Markov chain *itself* varies analytically, in a strong sense, as a function of the underlying Markov chain parameters.

1 Introduction

For $m, n \in \mathbb{Z}$ with $m \leq n$, we denote a sequence of symbols y_m, y_{m+1}, \dots, y_n by y_m^n . Consider a stationary stochastic process Y with a finite set of states $\{1, 2, \dots, B\}$ and distribution $p(y_m^n)$. Denote the conditional distributions by $p(y_{n+1}|y_m^n)$. The entropy rate of Y is defined as

$$H(Y) = \lim_{n \rightarrow \infty} -E_p(\log(p(y_0|y_{-n}^{-1}))),$$

where E_p denotes expectation with respect to the distribution p .

Let Y be a stationary first order Markov chain with

$$\Delta(i, j) = p(y_1 = j | y_0 = i).$$

It is well known that

$$H(Y) = - \sum_{i,j} p(y_0 = i) \Delta(i, j) \log \Delta(i, j).$$

A *hidden Markov chain* Z (or function of a Markov chain) is a process of the form $Z = \Phi(Y)$, where Φ is a function defined on $\{1, 2, \dots, B\}$ with values $\{1, 2, \dots, A\}$. Often

a hidden Markov chain is defined as a Markov chain observed in noise. It is well known that the two definitions are equivalent (the equivalence is typified by Example 4.1).

For a hidden Markov chain, $H(Z)$ turns out (see Equation (2.4) below) to be the integral of a certain function defined on a simplex with respect to a measure due to Blackwell [4]. However Blackwell's measure is somewhat complicated and the integral formula appears to be difficult to evaluate in most cases.

Recently there has been a rebirth of interest in computing the entropy rate of a hidden Markov chain, and many approaches have been adopted to tackle this problem. For instance, some researchers have used Blackwell's measure to bound the entropy rate [20] and others introduced a variation [8] on bounds due to [3]. An efficient Monte Carlo method for computing the entropy rate of a hidden Markov chain was proposed independently by Arnold and Loeliger [2], Pfister et. al. [25], and Sharma and Singh [31].

In another direction, [20, 12, 35] have studied the variation of the entropy rate as parameters of the underlying Markov chain vary. These works motivated us to consider the general question of whether the entropy rate of a hidden Markov chain is smooth, or even analytic [30, 32], as a function of the underlying parameters. Indeed, this is true under mild positivity assumptions:

Theorem 1.1. *Suppose that the entries of Δ are analytically parameterized by a real variable vector $\vec{\varepsilon}$. If at $\vec{\varepsilon} = \vec{\varepsilon}_0$,*

1. *For all $a \in \{1, 2, \dots, A\}$, there is at least one j with $\Phi(j) = a$ such that the j -th column of Δ is strictly positive – and –*
2. *Every column of Δ is either all zero or strictly positive,*

then $H(Z)$ is a real analytic function of $\vec{\varepsilon}$ at $\vec{\varepsilon}_0$.

Note that this theorem holds if all the entries of Δ are positive. The more general form of our hypotheses is very important (see Example 4.1).

Real analyticity at a point is important because it means that the function can be expressed as a convergent power series in a neighborhood of the point. The power series can be used to approximate or estimate the function. For convenience of the reader, we recall some basic concepts of analyticity in Section 3.

Several authors have observed that the entropy rate of a hidden Markov chain can be viewed as the top Lyapunov exponent of a random matrix product [11, 12, 10]. Results in [1, 22, 23, 27] show that under certain conditions the top Lyapunov exponent of a random matrix product varies analytically as either the underlying Markov process varies analytically or as the matrix entries vary analytically, but not both. However, when regarding the entropy rate as a Lyapunov exponent of a random matrix product, the matrix entries depend on the underlying Markov process. So, the results from Lyapunov theory do not appear to apply directly. Nevertheless, much of the main idea of our proof of Theorem 1.1 is essentially contained in Peres [23]. In contrast to Peres' proof, we do not use the language of Lyapunov exponents and we use only basic complex analysis and no functional analysis. Also the hypotheses in [23] do not carry over to our setting. To the best of our knowledge the statement and proof of Theorem 1.1 has not appeared in the literature. For analyticity

of certain other statistical quantities, see also related work in the area of statistical physics in [7, 5, 15, 6].

After discussing background in Sections 2 and 3, we prove Theorem 1.1 in Section 4. As an example, we show that the entropy rate of a hidden Markov chain obtained by observing a binary Markov chain in binary symmetric noise, with noise parameter ε , is analytic at any $\varepsilon = \varepsilon_0 \geq 0$, provided that the Markov transition probabilities are all positive.

In Section 5, we infer from the proof of Theorem 1.1 a general principle to determine a domain of analyticity for the entropy rate. We apply this to the case of hidden Markov chains obtained from binary Markov chains in binary symmetric noise to find a lower bound on the radius of convergence of a power series in ε at $\varepsilon_0 = 0$. Given the recent results of [36], which compute the derivatives of all orders at $\varepsilon_0 = 0$, this gives an explicit power series for entropy rate near $\varepsilon_0 = 0$.

In Section 6, we show how to relax the conditions of Theorem 1.1 and apply this to give more examples where the entropy rate is analytic.

The entropy rate can fail to be analytic. In Section 7 we give examples and then give a complete set of necessary and sufficient conditions for analyticity in the special case of binary hidden Markov chains with an unambiguous symbol, i.e., a symbol which can be produced by only one symbol of the Markov chain.

Finally in Section 8, we resort to more advanced techniques to prove a stronger version, Theorem 8.1, of Theorem 1.1. This result gives a sense in which the hidden Markov chain *itself* varies analytically with $\vec{\varepsilon}$. The proof of this result requires some measure theory and functional analysis, along with ideas from equilibrium states [26], which are reviewed in Appendix C. Our first proof of Theorem 1.1 was derived as a consequence of Theorem 8.1. It also follows from Theorem 8.1 that, in principle, many statistical properties in addition to entropy rate vary analytically.

Most results of this paper were first announced in [9].

2 Iteration on the Simplex

Let W be the simplex, comprising the vectors

$$\{w = (w_1, w_2, \dots, w_B) \in \mathbb{R}^B : w_i \geq 0, \sum_i w_i = 1\},$$

and let W_a be all $w \in W$ with $w_i = 0$ for $\Phi(i) \neq a$. Let $W^{\mathbb{C}}$ denote the complex version of W , i.e., $W^{\mathbb{C}}$ denotes the complex simplex comprising the vectors

$$\{w = (w_1, w_2, \dots, w_B) \in \mathbb{C}^B : \sum_i w_i = 1\},$$

and let $W_a^{\mathbb{C}}$ denote the complex version of W_a , i.e., $W_a^{\mathbb{C}}$ consists of all $w \in W^{\mathbb{C}}$ with $w_i = 0$ for $\Phi(i) \neq a$. For $a \in A$, let Δ_a denote the $B \times B$ matrix such that $\Delta_a(i, j) = \Delta(i, j)$ for j with $\Phi(j) = a$, and $\Delta_a(i, j) = 0$ otherwise. For $a \in A$, define the scalar-valued and vector-valued functions r_a and f_a on W by

$$r_a(w) = w\Delta_a\mathbf{1},$$

and

$$f_a(w) = w\Delta_a/r_a(w).$$

Note that f_a defines the action of the matrix Δ_a on the simplex W . For any fixed n and z_{-n}^0 , define

$$x_i = x_i(z_{-n}^i) = p(y_i = \cdot | z_i, z_{i-1}, \dots, z_{-n}), \quad (2.1)$$

(here \cdot represent the states of the Markov chain Y), then from Blackwell [4], $\{x_i\}$ satisfies the random dynamical iteration

$$x_{i+1} = f_{z_{i+1}}(x_i), \quad (2.2)$$

starting with

$$x_{-n-1} = p(y_{-n-1} = \cdot). \quad (2.3)$$

We remark that Blackwell showed that

$$H(Z) = - \int \sum_a r_a(w) \log r_a(w) dQ(w), \quad (2.4)$$

where Q , known as *Blackwell's measure*, is the limiting probability distribution, as $n \rightarrow \infty$, of $\{x_0\}$ on W . However, we do not use Blackwell's measure explicitly in this paper.

Next, we consider two metrics on a compact subset S of the interior of a subsimplex W' of W . Without loss of generality, we assume that W' consists of all points from W with the last $B - k$ coordinates equal to 0. The Euclidean metric $d_{\mathbf{E}}$ on S is defined as usual, namely for $u, v \in S$,

$$u = (u_1, u_2, \dots, u_B), v = (v_1, v_2, \dots, v_B) \in S,$$

we have

$$d_{\mathbf{E}}(u, v) = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + \dots + (u_k - v_k)^2}.$$

The Hilbert metric [29] $d_{\mathbf{B}}$ on S is defined as follows:

$$d_{\mathbf{B}}(u, v) = \max_{i \neq j \leq k} \log \left(\frac{u_i/u_j}{v_i/v_j} \right).$$

The following result is well known (for instance, see [1]). For completeness, we give a detailed proof in Appendix A.

Proposition 2.1. *$d_{\mathbf{E}}$ and $d_{\mathbf{B}}$ are equivalent (denoted by $d_{\mathbf{E}} \sim d_{\mathbf{B}}$) on any compact subset S of the interior of a subsimplex W' of W , i.e., there are positive constants $C_1 < C_2$ such that for any two points $u, v \in S$,*

$$C_1 d_{\mathbf{B}}(u, v) < d_{\mathbf{E}}(u, v) < C_2 d_{\mathbf{B}}(u, v).$$

Proposition 2.2. *Assume that at $\vec{\epsilon}_0$, Δ satisfies conditions 1 and 2 of Theorem 1.1. Then for sufficiently large n and all choices of $a_1, \dots, a_n \in \{1, 2, \dots, A\}$ and $b \in \{1, 2, \dots, A\}$ (here a_1, \dots, a_n and b denote hidden Markov symbols), the mapping $f_{a_n} \circ f_{a_{n-1}} \circ \dots \circ f_{a_1}$ is a contraction mapping under the Euclidean metric on W_b .*

Proof. $\hat{W}_b = f_b(W)$ is a compact subset of the interior of some subsimplex W'_b of W_b ; this subsimplex corresponds to column indices j such that $\Phi(j) = b$ and the j -th column is strictly positive. Therefore one can define the Hilbert metric accordingly on \hat{W}_b . Each f_a is a contraction mapping on each \hat{W}_b under the Hilbert metric [29]; namely there exists $0 < \rho < 1$ such that for any a and b , and for any two points $u, v \in \hat{W}_b$,

$$d_{\mathbf{B}}(f_a(u), f_a(v)) < \rho d_{\mathbf{B}}(u, v).$$

Thus, for any choices of a_2, a_3, \dots, a_n , we have

$$d_{\mathbf{B}}(f_{a_n} \circ f_{a_{n-1}} \circ \dots \circ f_{a_2}(u), f_{a_n} \circ f_{a_{n-1}} \circ \dots \circ f_{a_2}(v)) < \rho^{n-1} d_{\mathbf{B}}(u, v).$$

By Proposition 2.1, there exists a positive constant C such that

$$d_{\mathbf{E}}(f_{a_n} \circ f_{a_{n-1}} \circ \dots \circ f_{a_2}(u), f_{a_n} \circ f_{a_{n-1}} \circ \dots \circ f_{a_2}(v)) < C \rho^{n-1} d_{\mathbf{E}}(u, v).$$

Let L be a universal Lipschitz constant for any $f_{a_1} : W_b \rightarrow W'_{a_1}$ with respect to the Euclidean metric. Choose n large enough such that $C \rho^{n-1} < 1/L$. So, for sufficiently large n , any composition of the form $f_{a_n} \circ \dots \circ f_{a_1}$ is a Euclidean contraction on W_b . □

Remark 2.3. Using a slightly modified proof, one can show that for sufficiently large n , any composition of the form $f_{a_n} \circ \dots \circ f_{a_1}$ is a Euclidean contraction on the whole simplex W .

3 Brief background on analyticity

In this section, we briefly review the basics in complex analysis for the purpose of this paper. For more details, we refer to [30, 32].

A real (or complex) function of several variables is analytic at a given point if it admits a convergent Taylor series representation in a real (or complex) neighborhood of the given point. We say that it is real (or complex) analytic in a neighborhood if it is real (or complex) analytic at each point of the neighborhood.

The relationship between real and complex analytic functions is as follows: 1) Any real analytic function can be extended to a complex analytic function on some complex neighborhood; 2) Any real function obtained by restricting a complex analytic function from a complex neighborhood to a real neighborhood is a real analytic function.

The main fact regarding analytic functions used in this paper is that the uniform limit of a sequence of complex analytic functions on a fixed complex neighborhood is complex analytic. The analogous statement does not hold (in fact, fails dramatically!) for real analytic functions.

As an example of a real-valued parametrization of a matrix, consider:

$$\Delta(\varepsilon) = \begin{bmatrix} 2\varepsilon & \varepsilon & 1 - 3\varepsilon \\ \varepsilon & 1 - \varepsilon - \sin(\varepsilon) & \sin(\varepsilon) \\ \varepsilon^2 & \varepsilon^3 & 1 - \varepsilon^2 - \varepsilon^3 \end{bmatrix}.$$

Denote the states of Δ by $\{1, 2, 3\}$ and let $\Phi(1) = \Phi(2) = 0$, $\Phi(3) = 1$. Each entry of Δ is a real analytic function of ε at any given point $\varepsilon = \varepsilon_0$. For $\varepsilon_0 > 0$ and sufficiently small,

Δ is stochastic (i.e., each row sums to 1 and each entry is nonnegative) and in fact strictly positive (i.e., each entry is positive). According to Theorem 1.1, for such values of ε_0 , the entropy rate of the hidden Markov chain defined by $\Delta(\varepsilon)$ and Φ is real analytic as a function of ε at ε_0 .

While we typically think of analytic parametrizations as having the “look” of the preceding example, there is a conceptually simpler parametrization – namely, parameterize an $n \times n$ matrix Δ by its entries themselves; if Δ is required to be stochastic, we choose the parameters to be any set of $n - 1$ entries in each row (so, the real variable vector is an $n(n - 1)$ -tuple). Clearly, for analyticity it does not matter which entries are chosen. We call this the *natural parametrization*.

Suppose that $H(Z)$ is analytic with respect to this parametrization. Then, $H(Z)$ viewed as a function of any other analytic parametrization of the entries of Δ is the composition of two analytic functions and thus must be analytic. We thus have that the following two statements are equivalent.

- $H(Z)$ is analytic with respect to the natural parameterization.
- $H(Z)$ is analytic with respect to any analytic parameterization.

We shall use this implicitly through the paper.

4 Proof of Theorem 1.1

Notation: We rewrite Δ , Z , $f_a(x)$, $p(z_0|z_{-n}^{-1})$ with parameter vector $\vec{\varepsilon}$ as $\Delta^{\vec{\varepsilon}}$, $Z^{\vec{\varepsilon}}$, $f_a^{\vec{\varepsilon}}(x)$ and $p^{\vec{\varepsilon}}(z_0|z_{-n}^{-1})$, respectively. We use the notation \hat{W}_a to mean $f_a^{\varepsilon_0}(W)$. Let $\Omega_{\mathbb{C}} = \Omega_{\mathbb{C}}(r)$ denote the set of points of distance at most r from $\vec{\varepsilon}_0$ in the complex parameter space \mathbb{C}^m . Let $N_b = N_b(R)$ denote the set of all points in $W_b^{\mathbb{C}}$ of distance at most R from \hat{W}_b .

We first prove that for some $r > 0$, $\log p^{\vec{\varepsilon}}(z_0|z_{-n}^{-1})$ can be extended to a complex analytic function of $\vec{\varepsilon} \in \Omega_{\mathbb{C}}(r)$ and for two symbol sequences $z_{-n_1}^0$ and $\hat{z}_{-n_2}^0$, $|\log p^{\vec{\varepsilon}}(z_0|z_{-n_1}^{-1}) - \log p^{\vec{\varepsilon}}(\hat{z}_0|\hat{z}_{-n_2}^{-1})|$ decays exponentially fast in n , when $n \leq n_1, n_2$ and $z_{-n}^0 = \hat{z}_{-n}^0$, uniformly in $\vec{\varepsilon} \in \Omega_{\mathbb{C}}(r)$.

Note that for each a, b , $f_a^{\vec{\varepsilon}}(w)$ is a rational function of the entries of $\Delta^{\vec{\varepsilon}}$ and $w \in \hat{W}_b$. So, by viewing the real vector variables $\vec{\varepsilon}$ and w as complex vector variables, we can naturally extend $f_a^{\vec{\varepsilon}}(w)$ to a complex-valued function of complex vector variables $\vec{\varepsilon}$ and w . Since Δ satisfies conditions 1 and 2 at $\vec{\varepsilon}_0$, for sufficiently small r and R , the denominator of $f_a^{\vec{\varepsilon}}(w)$ is nonzero for $\vec{\varepsilon}$ in $\Omega_{\mathbb{C}}(r)$ and w in $N_b(R)$. Thus, $f_a^{\vec{\varepsilon}}(w)$ is a complex analytic function of $(\vec{\varepsilon}, w)$ in the neighborhood $\Omega_{\mathbb{C}}(r) \times N_b(R)$.

Assuming conditions 1 and 2, we claim that Δ has an isolated (in modulus) maximum eigenvalue 1 at $\vec{\varepsilon}_0$. To see this, we apply Perron-Frobenius theory [29] as follows. By permuting the indices, we can express:

$$\Delta = \begin{bmatrix} U & 0 \\ V & 0 \end{bmatrix}$$

where U is the submatrix corresponding to indices with positive columns. The nonzero eigenvalues of Δ are the same as the eigenvalues of U , which is a positive stochastic matrix. Such a matrix has isolated (in modulus) maximum eigenvalue 1.

The stationary distribution $p^\varepsilon(y = \cdot)$ (the eigenvector corresponding to the maximum eigenvalue 1) is a rational function of the entries of Δ^ε , since it is a solution of the equation $v\Delta^\varepsilon = v$. So, in the same way as for $f_a^\varepsilon(w)$ we can naturally extend $p^\varepsilon(y = \cdot)$ to a complex analytic function $p^\varepsilon(y = \cdot)$ on $\Omega_{\mathbb{C}}$.

Extending (2.1) for each i , we define

$$x_i^\varepsilon = x_i^\varepsilon(z_{-n}^i) = p^\varepsilon(y_i = \cdot | z_{-n}^i), \quad (4.5)$$

by iterating the following complexified random dynamical system (extending (2.2) and (2.3)):

$$x_{i+1}^\varepsilon = f_{z_{i+1}}^\varepsilon(x_i^\varepsilon), \quad (4.6)$$

starting with

$$x_{-n-1}^\varepsilon = p^\varepsilon(y_{-n-1} = \cdot). \quad (4.7)$$

By Proposition 2.2, for sufficiently large n , we can replace the set of mappings $\{f_a^{\varepsilon_0}\}$ with the set $\{f_{a_n}^{\varepsilon_0} \circ f_{a_{n-1}}^{\varepsilon_0} \circ \cdots \circ f_{a_1}^{\varepsilon_0}\}$ and then assume that each $f_a^{\varepsilon_0}$ is a Euclidean contraction on each W_b with contraction coefficient $\rho < 1$. Since \hat{W}_b is compact and the definition of ρ -contraction is given by strict inequality, we can choose r and R sufficiently small such that

$$f_a^\varepsilon \text{ is a Euclidean } \rho\text{-contraction on each } N_b(R), \varepsilon \in \Omega_{\mathbb{C}}(r). \quad (4.8)$$

Further, we claim that by choosing r still smaller, if necessary,

$$x_i^\varepsilon \in \cup_b N_b(R), \text{ for all } i, n \text{ and all choices of } z_{-n}^i, \varepsilon \in \Omega_{\mathbb{C}}(r). \quad (4.9)$$

To see this, fixing ρ and R , choose r so small that

$$|f_a^\varepsilon(x) - f_a^{\varepsilon_0}(x)| \leq R(1 - \rho), \quad x \in \cup_b \hat{W}_b, \varepsilon \in \Omega_{\mathbb{C}}(r) \quad (4.10)$$

and

$$|p^\varepsilon(\cdot) - p^{\varepsilon_0}(\cdot)| \leq R(1 - \rho), \quad \varepsilon \in \Omega_{\mathbb{C}}(r). \quad (4.11)$$

Now consider the difference

$$\begin{aligned} & x_{i+1}^\varepsilon - x_{i+1}^{\varepsilon_0} \\ &= f_{z_{i+1}}^\varepsilon(x_i^\varepsilon) - f_{z_{i+1}}^{\varepsilon_0}(x_i^{\varepsilon_0}) = f_{z_{i+1}}^\varepsilon(x_i^\varepsilon) - f_{z_{i+1}}^\varepsilon(x_i^{\varepsilon_0}) + f_{z_{i+1}}^\varepsilon(x_i^{\varepsilon_0}) - f_{z_{i+1}}^{\varepsilon_0}(x_i^{\varepsilon_0}). \end{aligned} \quad (4.12)$$

Then by (4.8), (4.10) and (4.11), and (4.12), for $i > -n - 1$, we have

$$|x_{i+1}^\varepsilon - x_{i+1}^{\varepsilon_0}| \leq \rho|x_i^\varepsilon - x_i^{\varepsilon_0}| + R(1 - \rho).$$

So,

$$|x_{i+1}^\varepsilon - x_{i+1}^{\varepsilon_0}| \leq R,$$

and thus for all i , we have $x_{i+1}^\varepsilon \in \cup_b N_b(R)$, yielding (4.9). Each x_i^ε is the composition of analytic functions on $\Omega_{\mathbb{C}}(r)$ and so is complex analytic on $\Omega_{\mathbb{C}}(r)$.

For $0 \leq n_1, n_2 \leq \infty$, we say two sequences $\{z_{-n_1}^0\}$ and $\{\hat{z}_{-n_2}^0\}$ have a common tail if there exists $n \geq 0$ with $n \leq n_1, n_2$ such that $z_i = \hat{z}_i, -n \leq i \leq 0$ (denoted by $z_{-n_1}^0 \stackrel{n}{\sim} \hat{z}_{-n_2}^0$).

Let

$$\begin{aligned}x_i^{\vec{\varepsilon}} &= x_i^{\vec{\varepsilon}}(z_{-n_1}^i) = p^{\vec{\varepsilon}}(y_i = \cdot | z_{-n_1}^i), \\ \hat{x}_i^{\vec{\varepsilon}} &= \hat{x}_i^{\vec{\varepsilon}}(\hat{z}_{-n_2}^i) = p^{\vec{\varepsilon}}(y_i = \cdot | \hat{z}_{-n_2}^i).\end{aligned}$$

Then we have

$$x_{i+1}^{\vec{\varepsilon}} = f_{z_{i+1}}^{\vec{\varepsilon}}(x_i^{\vec{\varepsilon}}), \quad \hat{x}_{i+1}^{\vec{\varepsilon}} = f_{\hat{z}_{i+1}}^{\vec{\varepsilon}}(\hat{x}_i^{\vec{\varepsilon}}).$$

From (4.8) and (4.9), if $z_{-n_1}^0 \stackrel{n}{\sim} \hat{z}_{-n_2}^0$, then there exists a positive constant L independent of n_1 and n_2 such that

$$|x_0^{\vec{\varepsilon}} - \hat{x}_0^{\vec{\varepsilon}}| \leq L\rho^n. \quad (4.13)$$

Naturally

$$p^{\vec{\varepsilon}}(z_0 | z_{-n}^{-1}) = \sum_{\{y_0: \Phi(y_0)=z_0\}} \sum_{y_{-1}} \Delta^{\vec{\varepsilon}}(y_{-1}, y_0) p^{\vec{\varepsilon}}(y_{-1} | z_{-n}^{-1}). \quad (4.14)$$

Then, there is a positive constant L' , independent of n_1, n_2 , such that

$$|p^{\vec{\varepsilon}}(z_0 | z_{-n_1}^{-1}) - p^{\vec{\varepsilon}}(\hat{z}_0 | \hat{z}_{-n_2}^{-1})| \leq L'\rho^n. \quad (4.15)$$

Since $\Delta^{\vec{\varepsilon}_0}$ satisfies conditions 1 and 2, $p^{\vec{\varepsilon}}(z_0 | z_{-n}^{-1})$ is bounded away from 0, uniformly in $\vec{\varepsilon} \in \Omega_{\mathbb{C}}$, n and choices of z_{-n}^{-1} ; thus there is a positive constant L'' , independent of n_1, n_2 , such that

$$|\log p^{\vec{\varepsilon}}(z_0 | z_{-n_1}^{-1}) - \log p^{\vec{\varepsilon}}(\hat{z}_0 | \hat{z}_{-n_2}^{-1})| \leq L''\rho^n. \quad (4.16)$$

Since for each $y \in \{1, \dots, B\}$, $p^{\vec{\varepsilon}}(y)$ is analytic, from

$$p^{\vec{\varepsilon}}(z) = \sum_{\Phi(y)=z} p^{\vec{\varepsilon}}(y),$$

we deduce that $p^{\vec{\varepsilon}}(z)$ is analytic. Furthermore since $p^{\vec{\varepsilon}}(z_0 | z_{-n}^{-1})$ is analytic on $\Omega_{\mathbb{C}}$, we conclude $p^{\vec{\varepsilon}}(z_{-n}^0)$ is analytic on $\Omega_{\mathbb{C}}$.

Choose σ so that

$$1 < \sigma < 1/\rho.$$

If r and R are chosen sufficiently small, then

$$\sum_{z_0} |p^{\vec{\varepsilon}}(z_0 | z_{-n}^{-1})| \leq \sigma, \quad \varepsilon \in \Omega_{\mathbb{C}}(r) \text{ and all sequences } z \quad (4.17)$$

and

$$\sum_{z_0} |p^{\vec{\varepsilon}}(z_0)| \leq \sigma, \quad \varepsilon \in \Omega_{\mathbb{C}}(r). \quad (4.18)$$

Then we have

$$\sum_{z_{-n-1}^0} |p^{\vec{\varepsilon}}(z_{-n-1}^0)| = \sum_{z_{-n-1}^0} |p^{\vec{\varepsilon}}(z_{-n-1}^{-1}) p^{\vec{\varepsilon}}(z_0 | z_{-n-1}^{-1})| \leq \sum_{z_{-n-1}^{-1}} |p^{\vec{\varepsilon}}(z_{-n-1}^{-1})| \sum_{z_0} |p^{\vec{\varepsilon}}(z_0 | z_{-n-1}^{-1})| \leq \sigma \sum_{z_{-n}^0} |p^{\vec{\varepsilon}}(z_{-n}^0)|,$$

implying

$$\sum_{z_{-n-1}^0} |p^{\vec{\varepsilon}}(z_{-n-1}^0)| \leq \sigma^{n+2}. \quad (4.19)$$

Let

$$H_n^{\vec{\varepsilon}}(Z) = - \sum_{z_{-n}^0} p^{\vec{\varepsilon}}(z_{-n}^0) \log p^{\vec{\varepsilon}}(z_0 | z_{-n}^{-1})$$

and

$$\rho_1 = \rho\delta < 1,$$

then we have

$$\begin{aligned} |H_{n+1}^{\vec{\varepsilon}}(Z) - H_n^{\vec{\varepsilon}}(Z)| &= \left| \sum_{z_{-n-1}^0} p^{\vec{\varepsilon}}(z_{-n-1}^0) \log p^{\vec{\varepsilon}}(z_0 | z_{-n-1}^{-1}) - \sum_{z_{-n}^0} p^{\vec{\varepsilon}}(z_{-n}^0) \log p^{\vec{\varepsilon}}(z_0 | z_{-n}^{-1}) \right| \\ &= \left| \sum_{z_{-n-1}^0} p^{\vec{\varepsilon}}(z_{-n-1}^0) (\log p^{\vec{\varepsilon}}(z_0 | z_{-n-1}^{-1}) - \log p^{\vec{\varepsilon}}(z_0 | z_{-n}^{-1})) \right| \leq \sigma^2 L'' \rho_1^n; \end{aligned}$$

here the latter inequality follows from (4.16) and (4.19). Thus, for $m > n$,

$$|H_m^{\vec{\varepsilon}}(Z) - H_n^{\vec{\varepsilon}}(Z)| \leq \sigma^2 L'' (\rho_1^n + \dots + \rho_1^{m-1}) \leq \frac{\sigma^2 L'' \rho_1^n}{1 - \rho_1}.$$

This establishes the uniform convergence of $H_n^{\vec{\varepsilon}}(Z)$ to a limit $H_\infty^{\vec{\varepsilon}}(Z)$. By Theorem 2.4.1 of [32], the uniform limit of complex analytic functions on a fixed complex neighborhood is analytic on that neighborhood, and so $H_\infty^{\vec{\varepsilon}}(Z)$ is analytic on $\Omega_{\mathbb{C}}$.

For real $\vec{\varepsilon}$, $H_\infty^{\vec{\varepsilon}}(Z)$ coincides with the entropy rate function $H(Z^{\vec{\varepsilon}})$, and so Theorem 1.1 follows.

Example 4.1. Consider a binary symmetric channel with crossover probability ε . Let $\{Y_n\}$ be the input Markov chain with the transition matrix

$$\Pi = \begin{bmatrix} \pi_{00} & \pi_{01} \\ \pi_{10} & \pi_{11} \end{bmatrix}. \quad (4.20)$$

At time n the channel can be characterized by the following equation

$$Z_n = Y_n \oplus E_n,$$

where \oplus denotes binary addition, E_n denotes the i.i.d. binary noise with $p_E(0) = 1 - \varepsilon$ and $p_E(1) = \varepsilon$, and Z_n denotes the corrupted output. Then (Y_n, E_n) is jointly Markov, so $\{Z_n\}$ is a hidden Markov chain with the corresponding

$$\Delta = \begin{bmatrix} \pi_{00}(1 - \varepsilon) & \pi_{00}\varepsilon & \pi_{01}(1 - \varepsilon) & \pi_{01}\varepsilon \\ \pi_{00}(1 - \varepsilon) & \pi_{00}\varepsilon & \pi_{01}(1 - \varepsilon) & \pi_{01}\varepsilon \\ \pi_{10}(1 - \varepsilon) & \pi_{10}\varepsilon & \pi_{11}(1 - \varepsilon) & \pi_{11}\varepsilon \\ \pi_{10}(1 - \varepsilon) & \pi_{10}\varepsilon & \pi_{11}(1 - \varepsilon) & \pi_{11}\varepsilon \end{bmatrix};$$

here, Φ maps states 1 and 4 to 0 and maps states 2 and 3 to 1. This class of hidden Markov chains has been studied extensively (e.g., [12], [20]).

By Theorem 1.1, when ε and π_{ij} 's are positive, the entropy rate $H(Z)$ is analytic as a function of ε and π_{ij} 's. This still holds when $\varepsilon = 0$ and the π_{ij} 's are positive, because in this case, we have

$$\Delta = \begin{bmatrix} \pi_{00} & 0 & \pi_{01} & 0 \\ \pi_{00} & 0 & \pi_{01} & 0 \\ \pi_{10} & 0 & \pi_{11} & 0 \\ \pi_{10} & 0 & \pi_{11} & 0 \end{bmatrix}.$$

5 Domain of Analyticity

Suppose Δ is analytically parameterized by a vector variable $\vec{\varepsilon}$, and Conditions 1 and 2 in Theorem 1.1 are satisfied at $\vec{\varepsilon} = \vec{\varepsilon}_0$. In principle, the proof of Theorem 1.1 determines a neighborhood $\Omega_{\mathbb{C}}(r)$ of $\vec{\varepsilon}_0$ on which the entropy rate is analytic. Specifically, if one can find ρ, r and R such that all of the following hold, then the entropy rate is analytic on $\Omega_{\mathbb{C}}(r)$.

1. Find ρ such that each $f_a^{\varepsilon_0}$ is a Euclidean ρ -contraction on each W_b . Then choose positive r, R such that for all $\vec{\varepsilon} \in \Omega_{\mathbb{C}}(r)$, each $f_a^{\vec{\varepsilon}}$ is a Euclidean ρ -contraction on each $N_b(R)$ (see (4.8)).
2. Next find r smaller (if necessary) such that for all $\vec{\varepsilon} \in \Omega_{\mathbb{C}}(r)$, the image of the stationary vector of $\Delta^{\vec{\varepsilon}}$, under any composition of the mappings $\{f_a^{\vec{\varepsilon}}\}$, stays within $\cup_b N_b(R)$ (see (4.9)). Note that the argument in the proof shows that this holds if (4.10) and (4.11) hold.
3. Finally, find r, R such that the sum of the absolute values of the complexified conditional probabilities, conditioned on any given past symbol sequence, is $< 1/\rho$, and similarly for the sum of the absolute values of the complexified stationary probabilities (see (4.17) and (4.18)).

In fact, the proof shows that one can always find such ρ, r, R , but in condition 1 above one may need to replace f_a 's by all n -fold compositions of the f_a 's, for some n .

Recall from Example 4.1 the family of hidden Markov chains Z^ε determined by passing a binary Markov chain through a binary symmetric channel with cross-over probability ε . Recall that $H(Z^\varepsilon)$ is an analytic function of ε at $\varepsilon = 0$ when the Markov transition probabilities are all positive. We shall determine a complex neighborhood of 0 such that the entropy rate, as a function of ε , is analytic on this neighborhood.

Let $u_n = p(y_n = 0 | z_1^n)$ and $v_n = p(y_n = 1 | z_1^n)$. For $z_{n+1} = 1$ we have

$$u_{n+1} = \frac{\varepsilon(\pi_{00}u_n + \pi_{10}v_n)}{\varepsilon(\pi_{00}u_n + \pi_{10}v_n) + (1 - \varepsilon)(\pi_{01}u_n + \pi_{11}v_n)},$$

$$v_{n+1} = \frac{(1 - \varepsilon)(\pi_{01}u_n + \pi_{11}v_n)}{\varepsilon(\pi_{00}u_n + \pi_{10}v_n) + (1 - \varepsilon)(\pi_{01}u_n + \pi_{11}v_n)}.$$

Since $u_n + v_n = 1$, u_{n+1} is a function of u_n ; let g_1 denote this function.

For $z_{n+1} = 0$ we have

$$u_{n+1} = \frac{(1 - \varepsilon)(\pi_{00}u_n + \pi_{10}v_n)}{(1 - \varepsilon)(\pi_{00}u_n + \pi_{10}v_n) + \varepsilon(\pi_{01}u_n + \pi_{11}v_n)},$$

$$v_{n+1} = \frac{\varepsilon(\pi_{01}u_n + \pi_{11}v_n)}{(1 - \varepsilon)(\pi_{00}u_n + \pi_{10}v_n) + \varepsilon(\pi_{01}u_n + \pi_{11}v_n)}.$$

Again, u_{n+1} is a function of u_n ; let g_0 denote this function.

And for the conditional probability, we have

$$p(z_n = 0 | z_1^{n-1}) = ((1 - \varepsilon)\pi_{00} + \varepsilon\pi_{01})u_n + ((1 - \varepsilon)\pi_{10} + \varepsilon\pi_{11})v_n.$$

Since $u_n + v_n = 1$, $p(z_n = 0|z_1^{n-1})$ is a function of u_n ; let r_0 denote this function. And

$$p(z_n = 1|z_1^{n-1}) = (\varepsilon\pi_{00} + (1 - \varepsilon)\pi_{01})u_n + (\varepsilon\pi_{10} + (1 - \varepsilon)\pi_{11})v_n.$$

Again, $p(z_n = 1|z_1^{n-1})$ is a function of u_n ; let r_1 denote this function.

Note that g_0, g_1, r_0, r_1 are all implicitly parameterized by ε . The stationary vector (π_0, π_1) of Y , which doesn't depend on ε , is equal to $(\pi_{10}/(\pi_{10} + \pi_{01}), \pi_{01}/(\pi_{10} + \pi_{01}))$.

We shall choose ρ with $0 < \rho < 1$, $r > 0$ and $R > 0$ such that for all ε with $|\varepsilon| < r$

1. g_0 and g_1 are ρ -contraction mappings on R -neighborhoods of 0 and 1 in the complex plane,
2. the set of all $\{g_{a_n} \circ g_{a_{n-1}} \circ \cdots \circ g_{a_1}(\pi_0)\}$ are within the R -neighborhoods of 0 and 1,
3. and $|r_0(u)| + |r_1(u)| < 1/\rho$ for u in R -neighborhoods of 0 and 1 in the complex plane.

By the general principle above, the entropy rate should be analytic on $|\varepsilon| < r$.

More concretely, condition 1, 2 and 3 translate to (here $\rho < 1$):

1. $|g'_0(u)| < \rho$, $|g'_1(u)| < \rho$ on $(|\varepsilon| < r$ and $|u| < R)$ and $(|\varepsilon| < r$ and $|1 - u| < R)$,
2. $\max\{|g_0(0) - 1|, |g_0(1) - 1|, |g_1(0)|, |g_1(1)|\} < R(1 - \rho)$ on $|\varepsilon| < r$ (this follows from (4.10); (4.11) is trivial since the stationary vector of Y doesn't depend on ε),
3. $|r_0(u)| + |r_1(u)| < 1/\rho$ on $(|\varepsilon| < r$ and $|u| < R)$ and $(|\varepsilon| < r$ and $|1 - u| < R)$.

A straightforward computation shows that the following conditions guarantee conditions 1, 2, 3:

$$\begin{aligned} 0 &\leq \frac{\sqrt{r(r+1)}|-\pi_{00}\pi_{11} + \pi_{10}\pi_{01}|}{\pi_{11} - |\pi_{10} - \pi_{11}|r - (|\pi_{00} - \pi_{10} - \pi_{01} + \pi_{11}|r + |\pi_{01} - \pi_{11}|)R} < \sqrt{\rho}, \\ 0 &\leq \frac{\sqrt{r(r+1)}|-\pi_{00}\pi_{11} + \pi_{10}\pi_{01}|}{\pi_{01} - |\pi_{00} - \pi_{01}|r - (|\pi_{00} - \pi_{10} - \pi_{01} + \pi_{11}|r + |\pi_{01} - \pi_{11}|)R} < \sqrt{\rho}, \\ 0 &\leq \frac{\sqrt{r(r+1)}|-\pi_{11}\pi_{00} + \pi_{01}\pi_{10}|}{\pi_{00} - |\pi_{01} - \pi_{00}|r - (|\pi_{00} - \pi_{10} + \pi_{11} - \pi_{01}|r + |\pi_{10} - \pi_{00}|)R} < \sqrt{\rho}, \\ 0 &\leq \frac{\sqrt{r(r+1)}|-\pi_{11}\pi_{00} + \pi_{01}\pi_{10}|}{\pi_{10} - |\pi_{11} - \pi_{10}|r - (|\pi_{00} - \pi_{10} + \pi_{11} - \pi_{01}|r + |\pi_{10} - \pi_{00}|)R} < \sqrt{\rho}, \\ 0 &\leq \frac{r\pi_{00}}{\pi_{01} - |\pi_{00} - \pi_{01}|r} < R(1 - \rho), \quad 0 \leq \frac{r\pi_{10}}{\pi_{11} - |\pi_{10} - \pi_{11}|r} < R(1 - \rho), \\ 0 &\leq \frac{r\pi_{11}}{\pi_{10} - |\pi_{11} - \pi_{10}|r} < R(1 - \rho), \quad 0 \leq \frac{r\pi_{01}}{\pi_{00} - |\pi_{01} - \pi_{00}|r} < R(1 - \rho), \\ 2(|\pi_{00} - \pi_{01} - \pi_{10} + \pi_{11}|r + |\pi_{01} - \pi_{11}|)R + 2|\pi_{10} - \pi_{11}|r + 1 &< 1/\rho, \end{aligned}$$

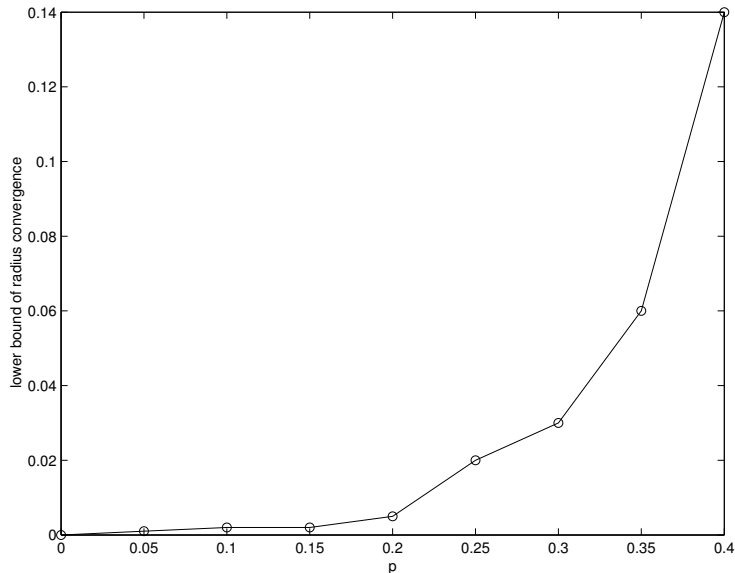


Figure 1: lower bound on radius of convergence as a function of p

$$2(|\pi_{10} - \pi_{11} - \pi_{00} + \pi_{01}|r + |\pi_{11} - \pi_{01}|)R + 2|\pi_{00} - \pi_{01}|r + 1 < 1/\rho.$$

In other words, for given ρ with $0 < \rho < 1$, choose r and R to satisfy all the constraints above. Then the entropy rate is an analytic function of ε on $|\varepsilon| < r$.

Let $\pi_{00} = \pi_{11} = p$ and $\pi_{01} = \pi_{10} = 1 - p$. We plot lower bounds on radius of convergence of $H(Z)$ (as a function of ε) against p in Figure 1. For a fixed p , the lower bound is obtained by randomly generating many 3-tuples (r, R, ρ) and taking the maximal r from the 3-tuples which satisfy the inequality conditions above. One can see in the plot that as p goes to 0.5, the lower bound is rapidly increasing. This is not surprising, since when $p = 0.5$, the corresponding entropy rate is a constant function of ε , and thus the radius of convergence is ∞ .

6 Relaxed Conditions

We do not know a complete set of necessary and sufficient conditions on Δ and Φ that guarantee analyticity of entropy rate. However, in this section, we show how the hypotheses in Theorem 1.1 can be relaxed and still guarantee analyticity. We then give several examples. In Section 7, we do give a complete set of necessary and sufficient conditions for a very special class of hidden Markov chains.

In this section, we assume that Δ has a simple maximum eigenvalue 1; this implies that Δ has a unique stationary vector \vec{s} .

For a mapping f from W_b to W and $w \in W_b$. Let f' denote the first derivative of f at w restricted to the subspace spanned by directions parallel to the simplex W_b and let $\|\cdot\|$ denote the Euclidean norm of a linear mapping. We say that $\{f_a : a \in \{1, 2, \dots, A\}\}$ is *eventually contracting* at $w \in W_b$ if there exists n such that for any $a_0, a_1, \dots, a_n \in \{1, 2, \dots, A\}$,

$\|(f_{a_n} \circ f_{a_{n-1}} \circ \cdots \circ f_{a_0})'(w)\|$ is strictly less than 1. We say that $\{f_a : a \in \{1, 2, \dots, A\}\}$ is *contracting* at $w \in W_b$ if it is *eventually contracting* at w with $n = 0$. Using the mean value theorem, one can show that if $\{f_a : a \in \{1, 2, \dots, A\}\}$ is *contracting* at each w in a compact convex subset K of W_b then each f_a is a contraction mapping on K .

Let L denote the limit set of

$$\{(f_{a_n} \circ f_{a_{n-1}} \circ \cdots \circ f_{a_0})(\vec{s}) : a_1, a_2, \dots, a_n \in \{1, 2, \dots, A\}, n \geq 0\}.$$

Theorem 6.1. *If at $\Delta = \hat{\Delta}$,*

1. *1 is a simple eigenvalue for $\hat{\Delta}$,*
2. *For every a and all w in L , $r_a(w) > 0$,*
3. *For every b , $\{f_a : a \in \{1, 2, \dots, A\}\}$ is eventually contracting at all w in the convex hull of the intersection of L and W_b ,*

then $H(Z)$ is analytic at $\Delta = \hat{\Delta}$.

Proof. Let \mathcal{X} denote the right infinite shift space $\{a_0^\infty : a_i \in \{1, 2, \dots, A\}\}$. Let L_b^δ be the set of all points in W_b of distance at most δ from $L \cap W_b$, and let $L^\delta = \cup_b L_b^\delta$. Choose δ so small that

- For every $a \in \{1, 2, \dots, A\}$ and w in L^δ , $r_a(w) > 0$ – and –
- For every b , $\{f_a : a \in \{1, 2, \dots, A\}\}$ is eventually contracting at all w in the convex hull K_b^δ of L_b^δ .

Since K_b^δ is compact, there exists n such that for any $a_0, a_1, \dots, a_n \in \{1, 2, \dots, A\}$ and any $w \in K_b^\delta$, $\|(f_{a_n} \circ f_{a_{n-1}} \circ \cdots \circ f_{a_0})'(w)\|$ is strictly less than 1. For simplicity, we may assume that $\{f_a\}$ is contracting on K_b^δ , and so each f_a is a contraction mapping on K_b^δ . Since $L_b^\delta \subseteq K_b^\delta$, it follows that $f_a(L^\delta) \subseteq L^\delta$.

For any $c_0^\infty \in \mathcal{X}$, there exists n such that $\{(f_{c_n} \circ f_{c_{n-1}} \circ \cdots \circ f_{c_0})(\vec{s})\} \in L^\delta$. Let $\mathcal{X}_{c_0^\infty}^n$ denote the cylinder set $\{a_0^\infty : a_0 = c_0, a_1 = c_1, \dots, a_n = c_n\}$. Since $f_a(L^\delta) \subset L^\delta$, we conclude that for any $a_0^\infty \in \mathcal{X}_{c_0^\infty}^n$ and all $m \geq n$, $\{(f_{a_m} \circ f_{a_{m-1}} \circ \cdots \circ f_{a_0})(\vec{s})\} \in L^\delta$. By the compactness of \mathcal{X} , we can find finitely many such cylinder sets to cover \mathcal{X} . Consequently we can find n such that for any $a_0^\infty \in \mathcal{X}$ and any $m \geq n$, we have $\{(f_{a_m} \circ f_{a_{m-1}} \circ \cdots \circ f_{a_0})(\vec{s})\} \in L^\delta$. We can now apply the proof of Theorem 1.1 – namely, we can use the contraction (along any symbolic sequence z_{-n}^0) to extend $H_n(Z) = H(Z_0|Z_{-n}^{-1})$ from real to complex and prove the uniform convergence of $H_n(Z)$ to $H(Z)$ in complex parameter space. \square

Remark 6.2.

(1) If $\hat{\Delta}$ has a strictly positive column (or more generally, there is a j such that for all i , there exists n such that $\hat{\Delta}_{ij}^n > 0$), then condition 1 of Theorem 6.1 holds by Perron-Frobenius theory.

(2) If for each symbol a , $\hat{\Delta}_a$ is row allowable (i.e., no row is all zero), then $r_a(w) > 0$ for all $w \in W$ and so condition 2 of Theorem 6.1 holds.

Theorem 6.1 relaxes the positivity assumptions of Theorem 1.1. Indeed given conditions 1 and 2 of Theorem 1.1, by Remark 6.2, conditions 1 and 2 of Theorem 6.1 hold. For condition 3 of Theorem 6.1, first observe that L is contained in $\cup_b f_b(W)$. Using the equivalence of the Euclidean metric and the Hilbert metric, Proposition 2.2 shows that for every b , $\{f_a : a \in \{1, 2, \dots, A\}\}$ is eventually contracting on $f_b(W)$, which is a convex set containing the intersection of L and W_b .

Theorem 6.1 also applies to many cases not covered by Theorem 1.1. Suppose that some column of $\hat{\Delta}$ is strictly positive and each $\hat{\Delta}_a$ is row allowable. By Remark 6.2, Theorem 6.1 applies whenever we can guarantee condition 3. For this, it is sufficient to check that for each a, b , f_a is a contraction, with respect to the Euclidean metric, on the convex hull of the intersection of L with each W_b . This can be done by explicitly computing derivatives. This is illustrated by the following example.

Example 6.3. Consider a hidden Markov chain Z defined by :

$$\hat{\Delta} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix},$$

with $\Phi(1) = \Phi(2) = 0$ and $\Phi(3) = \Phi(4) = 1$. We assume that some column of $\hat{\Delta}$ is strictly positive and both $\hat{\Delta}_0$ and $\hat{\Delta}_1$ are row allowable.

Parameterize W_0 by $(y, 1-y, 0, 0)$ and parameterize W_1 by $(0, 0, y, 1-y)$ (with $y \in [0, 1]$). We can explicitly compute the derivatives of f_0 and f_1 with respect to y :

$$f'_0|_{(y,1-y,0,0)} = \frac{a_{11}a_{22} - a_{12}a_{21}}{((a_{11} + a_{12} - a_{21} - a_{22})y + a_{21} + a_{22})^2},$$

$$f'_0|_{(0,0,y,1-y)} = \frac{a_{31}a_{42} - a_{32}a_{41}}{((a_{31} + a_{32} - a_{41} - a_{42})y + a_{41} + a_{42})^2},$$

$$f'_1|_{(y,1-y,0,0)} = \frac{a_{13}a_{24} - a_{14}a_{23}}{((a_{13} + a_{14} - a_{23} - a_{24})y + a_{23} + a_{24})^2},$$

$$f'_1|_{(0,0,y,1-y)} = \frac{a_{33}a_{44} - a_{34}a_{43}}{((a_{33} + a_{34} - a_{43} - a_{44})y + a_{43} + a_{44})^2},$$

Note that the row allowability condition guarantees that the denominators in these expressions never vanish.

Choose a_{ij} 's such that each of these derivatives is less than 1; then we conclude that the entropy rate is analytic at $\hat{\Delta}$. One way to do this is to make each of the 2×2 upper/lower left/right matrices singular.

Or choose the a_{ij} 's such that

$$\hat{\Delta} = \begin{bmatrix} \alpha_1 & * & \beta_1 & 0 \\ 0 & \alpha_2 & 0 & \beta_2 \\ \lambda_1 & * & \eta_1 & 0 \\ 0 & \lambda_2 & 0 & \eta_2 \end{bmatrix}$$

where $0 < \alpha_1 < \alpha_2$, $0 < \beta_1 < \beta_2$, $0 < \lambda_1 < \lambda_2$, $0 < \eta_1 < \eta_2$ and $*$ denote a real positive number (note that Theorem 1.1 doesn't apply for this special case). Let (s_2, s_4) be the Perron eigenvector of the stochastic matrix:

$$\begin{bmatrix} \alpha_2 & \beta_2 \\ \lambda_2 & \eta_2 \end{bmatrix}.$$

Then $\vec{s} = (0, s_2, 0, s_4)$ is the stationary vector of Δ corresponding to the simple eigenvalue 1. Let $w_0 = (0, 1, 0, 0)$ and $w_1 = (0, 0, 0, 1)$. One checks that for $n \geq 0$, $f_{a_n} \circ f_{a_{n-1}} \circ \dots \circ f_{a_0}(\vec{s}) = w_{a_n}$. Therefore L consists of $\{w_0, w_1\}$. Using the expressions above, we see that

$$\begin{aligned} f'_0|_{w_0} &= \alpha_1/\alpha_2 < 1, f'_0|_{w_1} = \lambda_1/\lambda_2 < 1, \\ f'_1|_{w_0} &= \beta_1/\beta_2 < 1, f'_1|_{w_1} = \eta_1/\eta_2 < 1. \end{aligned}$$

So, f_0 and f_1 are contraction mappings at $\{w_0, w_1\}$, and so condition 3 holds. Thus, the entropy rate $H(Z)$ is analytic at $\hat{\Delta}$.

7 Hidden Markov Chains with Unambiguous Symbol

Definition 7.1. A symbol a is called *unambiguous* if $\Phi^{-1}(a)$ contains only one element.

Remark 7.2. Note that unambiguous symbol is referred to as ‘‘singleton clump’’ in some ergodic theory work, such as [24].

When an unambiguous symbol is present, the entropy rate can be expressed in a simple way: letting a_1 be an unambiguous symbol,

$$H(Z) = \sum_{a_i \neq a_1} p(a_{i_n} a_{i_{n-1}} \dots a_{i_2} a_1) H(z | a_{i_n} a_{i_{n-1}} \dots a_{i_2} a_1). \quad (7.21)$$

In this section, we focus on the case of a binary hidden Markov chain, in which 0 is unambiguous. Then, we can rewrite (7.21) as

$$H(Z^\varepsilon) = p^\varepsilon(0)H^\varepsilon(z|0) + p^\varepsilon(10)H^\varepsilon(z|10) + \dots + p^\varepsilon(1^{(n)}0)H^\varepsilon(z|1^{(n)}0) + \dots, \quad (7.22)$$

where $1^{(n)}$ denotes the sequence of n 1's and

$$H^\varepsilon(z|1^{(n)}0) = -p^\varepsilon(0|1^{(n)}0) \log p^\varepsilon(0|1^{(n)}0) - p^\varepsilon(1|1^{(n)}0) \log p^\varepsilon(1|1^{(n)}0).$$

Example 7.3. Fix $a, b, \dots, h > 0$ and for $\varepsilon \geq 0$ let

$$\Delta(\varepsilon) = \begin{bmatrix} \varepsilon & a - \varepsilon & b \\ g & c & d \\ h & e & f \end{bmatrix}.$$

Assume $a, b, \dots, h > 0$ are chosen such that $\Delta(\varepsilon)$ is stochastic. The symbols of the Markov chain are the matrix indices $\{1, 2, 3\}$. Let Z^ε be the binary hidden Markov chain defined by: $\Phi(1) = 0$ and $\Phi(2) = \Phi(3) = 1$. We claim that $H(Z^\varepsilon)$ is not analytic at $\varepsilon = 0$.

Let $\pi(\varepsilon)$ be the stationary vector of $\Delta(\varepsilon)$ (which is unique since $\Delta(\varepsilon)$ is irreducible). Observe that

$$p^\varepsilon(0) = \pi_1(\varepsilon), \quad p^\varepsilon(00) = \pi_1(\varepsilon)\varepsilon,$$

and for $n \geq 1$.

$$p^\varepsilon(1^{(n)}0) = \pi_1(\varepsilon)(a - \varepsilon, b) \begin{bmatrix} c & d \\ e & f \end{bmatrix}^{n-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Since $\Delta(\varepsilon)$ is irreducible, $\pi(\varepsilon)$ is analytic in ε and positive. Now,

$$p^\varepsilon(0)H^\varepsilon(z|0) = -p^\varepsilon(00) \log p^\varepsilon(0|0) - p^\varepsilon(10) \log p^\varepsilon(1|0). \quad (7.23)$$

The first term in (7.23) is

$$-p^\varepsilon(00) \log p^\varepsilon(0|0) = -\pi_1(\varepsilon)\varepsilon \log \varepsilon,$$

which is not analytic (or even differentiable at $\varepsilon = 0$). The second term in (7.23) is

$$-p^\varepsilon(10) \log p^\varepsilon(1|0) = -\pi_1(\varepsilon)(a - \varepsilon + b) \log(\pi_1(\varepsilon)(a - \varepsilon + b)),$$

which is analytic at $\varepsilon = 0$. Thus, $H^\varepsilon(z|0)$ is not analytic at $\varepsilon = 0$. Similarly it can be shown that all of the terms of (7.22), other than $H^\varepsilon(z|0)$, are analytic at $\varepsilon = 0$. Since the matrix

$$\begin{bmatrix} c & d \\ e & f \end{bmatrix}$$

has spectral radius < 1 , the terms of (7.22) decay exponentially; it follows that the infinite sum of these terms is analytic. Thus, $H(Z^\varepsilon)$ is the sum of two functions of ε , one of which is analytic and the other is not analytic at $\varepsilon = 0$. Thus, $H(Z^\varepsilon)$ is not analytic at $\varepsilon = 0$.

Example 7.4. Fix $a, b, \dots, g > 0$ and consider the stochastic matrix

$$\Delta(\varepsilon) = \begin{bmatrix} e & a & b \\ f - \varepsilon & c & \varepsilon \\ g & 0 & d \end{bmatrix}.$$

The symbols of the Markov chain are the matrix indices $\{1, 2, 3\}$. Again let Z^ε be the binary hidden Markov chain defined by $\Phi(1) = 0$ and $\Phi(2) = \Phi(3) = 1$. We show that $H(Z^\varepsilon)$ is analytic at $\varepsilon = 0$ when $c \neq d$, and not analytic when $c = d$. Note that

$$p^\varepsilon(0) = \pi_1(\varepsilon),$$

and for $n \geq 1$.

$$p^\varepsilon(1^{(n)}0) = \pi_1(\varepsilon)(a, b) \begin{bmatrix} c & \varepsilon \\ 0 & d \end{bmatrix}^{n-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

When $c \neq d$, we assume $c > d$, then

$$\begin{bmatrix} c & \varepsilon \\ 0 & d \end{bmatrix}^n = \begin{bmatrix} c^n & \varepsilon c^{n-1} \frac{1-(d/c)^n}{1-d/c} \\ 0 & d^n \end{bmatrix}.$$

Since $\Delta(\varepsilon)$ is irreducible, $\pi(\varepsilon)$ is analytic in ε and positive. Simple computation leads to:

$$\begin{aligned} p^\varepsilon(1|1^{(n)}0) &= (ac^n + a\varepsilon c^{n-1} \frac{1 - (d/c)^n}{1 - d/c} + bd^n) / (ac^{n-1} + a\varepsilon c^{n-2} \frac{1 - (d/c)^{n-1}}{1 - d/c} + bd^{n-1}) \\ &= (ac^2 + a\varepsilon c \frac{1 - (d/c)^n}{1 - d/c} + bd^2(d/c)^{n-2}) / (ac + \varepsilon \frac{1 - (d/c)^{n-1}}{1 - d/c} + bd(d/c)^{n-2}), \end{aligned}$$

and

$$\begin{aligned} p^\varepsilon(0|1^{(n)}0) &= ((f - \varepsilon)ac^{n-1} + g(a\varepsilon c^{n-2} \frac{1 - (d/c)^{n-1}}{1 - d/c} + bd^{n-1})) / (ac^{n-1} + a\varepsilon c^{n-2} \frac{1 - (d/c)^{n-1}}{1 - d/c} + bd^{n-1}) \\ &= ((f - \varepsilon)ac + g(a\varepsilon \frac{1 - (d/c)^{n-1}}{1 - d/c} + bd(d/c)^{n-2})) / (ac + \varepsilon \frac{1 - (d/c)^{n-1}}{1 - d/c} + bd(d/c)^{n-2}). \end{aligned}$$

In this case all terms are analytic. Again since

$$\begin{bmatrix} c & \varepsilon \\ 0 & d \end{bmatrix}$$

has spectral radius < 1 , the term $p^\varepsilon(1^{(n)}0)H^\varepsilon(z|1^{(n)}0)$ is exponentially decaying with respect to n . Therefore the infinite sum of these terms is also analytic, and so the entropy rate is a real analytic function of ε .

When $c = d$, we have

$$\begin{aligned} p^\varepsilon(1|1^{(n)}0) &= (ac^{n+1} + a\varepsilon(n+1)c^n + bc^{n+1}) / (ac^n + a\varepsilon nc^{n-1} + bc^n) \\ &= (ac^2 + a\varepsilon(n+1)c + bc^2) / (ac + a\varepsilon n + bc), \end{aligned}$$

and

$$\begin{aligned} p^\varepsilon(0|1^{(n)}0) &= ((f - \varepsilon)ac^n + ga\varepsilon nc^{n-1} + gbc^n) / (ac^n + a\varepsilon nc^{n-1} + bc^n) \\ &= ((f - \varepsilon)ac + ga\varepsilon n + gbc) / (ac + a\varepsilon n + bc). \end{aligned}$$

For any n , consider a small neighborhood N_n of $-(a+b)c/an$ in \mathbb{C} such that $-(a+b)c/aj \in N_n$ only holds for $j = n$. When $\varepsilon \rightarrow -(a+b)c/an$, the complexified term $p^\varepsilon(1^{(n)}0)H^\varepsilon(z|1^{(n)}0) \rightarrow \infty$. Meanwhile, the sum of all the other terms can be analytically extended to N_n (from any path I from a positive ε to $-(a+b)c/an$ with $-(a+b)c/aj \notin I$ for $j \neq n$). Thus, by the uniqueness of analytic continuation of $H(Z^\varepsilon)$, we conclude that $H(Z^\varepsilon)$ blows up when one approaches $-(a+b)c/an$ and therefore is not analytic at $\varepsilon = 0$ (although it is smooth from the right at $\varepsilon = 0$).

The two examples above show that under certain conditions the entropy rate of a binary hidden Markov chain with unambiguous symbol can fail to be analytic at the boundary. We now show that these examples typify all the types of failures of analyticity at the boundary (in the case of a binary hidden Markov chains with an unambiguous symbol).

We will need the following result.

Lemma 7.5. *Let $A(\vec{\varepsilon})$ be an analytic parameterization of complex matrices. Let λ be the spectral radius of $A(\vec{\varepsilon}_0)$. Then for any $\eta > 0$, there exists a complex neighborhood Ω of $\vec{\varepsilon}_0$ and positive constant C such that for all $\vec{\varepsilon} \in \Omega$ and all i, j, k*

$$|A_{ij}^k(\vec{\varepsilon})| \leq C(\lambda + \eta)^k.$$

Proof. Following [29], we consider

$$(I - zA)^{-1} = I + zA + z^2A^2 + \dots.$$

And

$$(I - zA)^{-1} = \frac{\text{Adj}(I - zA)}{\det(I - zA)} = \frac{\text{Adj}(I - zA)}{(1 - \lambda_1 z)(1 - \lambda_2 z) \cdots (1 - \lambda_n z)},$$

where $\lambda_1, \dots, \lambda_n$ are the eigenvalues of A . So every entry of $(I - zA)^{-1}$ takes the form:

$$\begin{aligned} & (p_0 + p_1 z + \cdots + p_m z^m) \prod_{j=1}^n \sum_{i=0}^{\infty} \lambda_j^i z^i \\ &= \sum_{k=0}^{\infty} \sum_{u=0}^m p_u \sum_{i_1+i_2+\cdots+i_n=k-u} \lambda_1^{i_1} \lambda_2^{i_2} \cdots \lambda_n^{i_n} z^k. \end{aligned}$$

Since the eigenvalues of a complex matrix vary continuously with entries, the lemma follows. \square

Now let $S(n)$ denote the set of all the $n \times n$ complex matrices with isolated (in modulus) maximum eigenvalue.

Lemma 7.6. *$S(n)$ is connected.*

Proof. let $A, B \in S(n)$, then we consider their Jordan forms:

$$A = U \text{diag}(\lambda_1, C) U^{-1}, \quad B = V \text{diag}(\eta_1, D) V^{-1},$$

here λ_1, η_1 are maximum eigenvalues for A, B , respectively, C, D correspond to other Jordan blocks, and $U, V \in GL(n, \mathbb{C})$ (here $GL(n, \mathbb{C})$ denotes the set of all the $n \times n$ nonsingular complex matrices). Since $GL(n, \mathbb{C})$ is connected [19], it suffices to prove that there is a path in $S(n)$ from $\text{diag}(\lambda_1, C)$ to $\text{diag}(\eta_1, D)$. This is straightforward: first connect $\text{diag}(\lambda_1, C)$ to $\text{diag}(\eta_1, \eta_1/\lambda_1 C)$ by a continuous rescaling; then connect $\eta_1/\lambda_1 C$ to D by the path $t\eta_1/\lambda_1 C + (1-t)D$ (the path $\text{diag}(\eta_1, t\eta_1/\lambda_1 C + (1-t)D)$ stays within $S(n)$ since the matrices along this path are upper triangular with all diagonal entries, except η_1 , of modulus less than $|\eta_1|$). \square

For a complex analytic function $f(z_1, z_2, \dots, z_n)$, let $V(f)$ denote the ‘‘hypersurface’’ defined by f , namely

$$V(f) = \{(z_1, z_2, \dots, z_n) \in \mathbb{C}^n : f(z_1, z_2, \dots, z_n) = 0\}.$$

Now let $\Omega_{\mathbb{C}}$ denote a connected open set in \mathbb{C}^n . It is well known that the following Lemma holds (for completeness, we include a brief proof).

Lemma 7.7. $\Omega_{\mathbb{C}} \setminus V(f)$ is connected.

Proof. For simplicity, we first assume $\Omega_{\mathbb{C}}$ is a ball $B_r(z_0)$ (here $z_0 \in \mathbb{C}^n$ is the center of the ball and r is the radius, i.e., $B_r(z_0) = \{z \in \mathbb{C}^n : |z - z_0| < r\}$) in \mathbb{C}^n . For any two distinct point $P, Q \in \Omega_{\mathbb{C}} \setminus V(f)$, consider the “complex line”

$$L_{\mathbb{C}}^{PQ} = \{zP + (1 - z)Q : z \in \mathbb{C}\}.$$

$L_{\mathbb{C}}^{PQ} \cap V(f) \cap \Omega_{\mathbb{C}}$ consists of only isolated points (A non-constant one variable complex analytic function must have isolated zeros in the complex plane [30]). It then follows that for the compact real line segment:

$$L_{\mathbb{R}}^{PQ} = \{tP + (1 - t)Q : t \in [0, 1]\},$$

$L_{\mathbb{R}}^{PQ} \cap V(f) \cap \Omega_{\mathbb{C}}$ consists of only finitely many points. Certainly one can choose an arc in $L_{\mathbb{C}}^{PQ} \cap \Omega_{\mathbb{C}}$ to avoid these points and connect P and Q . This implies that $\Omega_{\mathbb{C}} \setminus V(f)$ is connected.

In the general case, $\Omega_{\mathbb{C}}$ is a connected open set in \mathbb{C}^n . Let I be an arc in $\Omega_{\mathbb{C}}$ connecting P and Q , and let $\{B_{r_j}(z_j)\}$ be a collection of balls covering I such that each $B_{r_j}(z_j) \cap B_{r_{j+1}}(z_{j+1}) \neq \emptyset$. Pick a point P_j in $B_{r_j}(z_j) \cap B_{r_{j+1}}(z_{j+1})$ such that $P_j \in \Omega_{\mathbb{C}} \setminus V(f)$. Applying the same argument as above to every ball $B_{r_j}(z_j)$, we see that P is connected to Q in $\Omega_{\mathbb{C}} \setminus V(f)$ through the points P_j 's. Thus we prove the lemma. \square

Theorem 7.8. Let Δ be an irreducible stochastic $d \times d$ matrix. Write Δ in the form:

$$\Delta = \begin{bmatrix} a & r \\ c & B \end{bmatrix} \quad (7.24)$$

where a is a scalar and B is a $(d - 1) \times (d - 1)$ matrix. Let Φ be the function defined by $\Phi(1) = 0$, and $\Phi(2) = \dots = \Phi(d) = 1$. Then for any parametrization $\Delta(\vec{\varepsilon})$ such that $\Delta(\vec{\varepsilon}_0) = \Delta$, letting $Z^{\vec{\varepsilon}}$ denote the hidden Markov chain defined by $\Delta(\vec{\varepsilon})$ and Φ , $H(Z^{\vec{\varepsilon}})$ is analytic at $\vec{\varepsilon}_0$ if and only if

1. $a > 0$, and $rB^j c > 0$ for $j = 0, 1, \dots$.
2. The maximum eigenvalue of B is simple and strictly greater in absolute value than the other eigenvalues of B .

Proof. Proof of sufficiency.

We write

$$\Delta(\vec{\varepsilon}) = \begin{bmatrix} a(\vec{\varepsilon}) & r(\vec{\varepsilon}) \\ c(\vec{\varepsilon}) & B(\vec{\varepsilon}) \end{bmatrix}, \quad (7.25)$$

where $a(\vec{\varepsilon})$ is a scalar and $B(\vec{\varepsilon})$ is a $(d - 1) \times (d - 1)$ matrix.

Since $\Delta(\vec{\varepsilon}_0)$ is stochastic and irreducible, its spectral radius is 1, and 1 is a simple eigenvalue of Δ . Thus, if $\Omega_{\mathbb{C}}$ is sufficiently small, for all $\vec{\varepsilon} \in \Omega_{\mathbb{C}}$, any fixed row $\pi(\vec{\varepsilon}) = (\pi_1(\vec{\varepsilon}), \pi_2(\vec{\varepsilon}), \dots, \pi_d(\vec{\varepsilon}))$ of $\text{Adj}(I - \Delta(\vec{\varepsilon}))$ is a left eigenvector of $\Delta(\vec{\varepsilon})$ associated with eigenvalue 1 and is an analytic function of $\vec{\varepsilon}$. Normalizing, we can assume that $\pi(\vec{\varepsilon})\mathbf{1} = 1$, $\pi(\vec{\varepsilon})$ is analytic in $\vec{\varepsilon}$, and $\pi_j(\vec{\varepsilon}_0) > 0$ for $j = 1, 2, \dots, d$.

The entries of $r(\vec{\varepsilon})$, $B(\vec{\varepsilon})$, and $c(\vec{\varepsilon})$ are real analytic in $\vec{\varepsilon}$ and can be extended to complex analytic functions in a complex neighborhood $\Omega_{\mathbb{C}}$ of $\vec{\varepsilon}_0$. Thus, for all n , $\pi_1(\vec{\varepsilon})r(\vec{\varepsilon})B(\vec{\varepsilon})^{n-1}\mathbf{1}$ and $\pi_1(\vec{\varepsilon})r(\vec{\varepsilon})B(\vec{\varepsilon})^{n-1}c(\vec{\varepsilon})$ can be extended to complex analytic functions on $\Omega_{\mathbb{C}}$ (in fact, each of these functions is a polynomial in $\vec{\varepsilon}$).

Since $B(\vec{\varepsilon}_0)$ is a proper sub-matrix of the irreducible stochastic matrix $\Delta(\vec{\varepsilon}_0)$, its spectral radius is strictly less than 1. Thus, by Lemma 7.5, there exists $0 < \lambda^* < 1$ and a constant $C_1 > 0$, such that for some complex neighborhood $\Omega_{\mathbb{C}}$ of $\vec{\varepsilon}_0$, all $\vec{\varepsilon} \in \Omega_{\mathbb{C}}$, and all n ,

$$|B_{ij}^n(\vec{\varepsilon})| < C_1(\lambda^*)^n.$$

Since $\pi_1(\vec{\varepsilon})$, $r(\vec{\varepsilon})$ and $c(\vec{\varepsilon})$ are continuous in $\vec{\varepsilon}$, there is a constant $C_2 > 0$ such that for all $\vec{\varepsilon} \in \Omega_{\mathbb{C}}$ and all n :

$$|\pi_1(\vec{\varepsilon})r(\vec{\varepsilon})B(\vec{\varepsilon})^n\mathbf{1}| < C_2(\lambda^*)^n. \quad (7.26)$$

We will need the following result, proven in Appendix B.

Lemma 7.9. *Let*

$$a(\vec{\varepsilon}, n) \equiv \frac{\pi_1(\vec{\varepsilon})r(\vec{\varepsilon})B(\vec{\varepsilon})^n\mathbf{1}}{\pi_1(\vec{\varepsilon})r(\vec{\varepsilon})B(\vec{\varepsilon})^{n-1}\mathbf{1}}$$

and

$$b(\vec{\varepsilon}, n) \equiv \frac{\pi_1(\vec{\varepsilon})r(\vec{\varepsilon})B(\vec{\varepsilon})^{n-1}c(\vec{\varepsilon})}{\pi_1(\vec{\varepsilon})r(\vec{\varepsilon})B(\vec{\varepsilon})^{n-1}\mathbf{1}}.$$

For a sufficiently small neighborhood $\Omega_{\mathbb{C}}$ of $\vec{\varepsilon}_0$, both $|a(\vec{\varepsilon}, n)|$ and $|b(\vec{\varepsilon}, n)|$ are bounded from above and away from zero, uniformly in $\vec{\varepsilon} \in \Omega_{\mathbb{C}}$ and n .

Define

$$H_n^{\vec{\varepsilon}} = -a(\vec{\varepsilon}, n) \log a(\vec{\varepsilon}, n) - b(\vec{\varepsilon}, n) \log b(\vec{\varepsilon}, n),$$

where $a(\vec{\varepsilon}, n)$ and $b(\vec{\varepsilon}, n)$ are as in Lemma 7.9. Choosing $\Omega_{\mathbb{C}}$ to be a smaller neighborhood of $\vec{\varepsilon}_0$, if necessary, $a(\vec{\varepsilon}, n)$ and $b(\vec{\varepsilon}, n)$ are constrained to lie in a closed disk not containing 0. Thus for all n , $H_n^{\vec{\varepsilon}}$ is an analytic function of $\vec{\varepsilon}$, with $|H_n^{\vec{\varepsilon}}|$ bounded uniformly in $\vec{\varepsilon} \in \Omega_{\mathbb{C}}$ and n . Since $\pi_1(\vec{\varepsilon})r(\vec{\varepsilon})B(\vec{\varepsilon})^{n-1}\mathbf{1}$ is analytic on $\Omega_{\mathbb{C}}$ and exponentially decaying (by (7.26)), the infinite series

$$H^{\vec{\varepsilon}}(Z) = \pi_1(\vec{\varepsilon})H_0^{\vec{\varepsilon}} + \pi_1(\vec{\varepsilon})r(\vec{\varepsilon})\mathbf{1}H_1^{\vec{\varepsilon}} + \cdots + \pi_1(\vec{\varepsilon})r(\vec{\varepsilon})B(\vec{\varepsilon})^{n-1}\mathbf{1}H_n^{\vec{\varepsilon}} + \cdots \quad (7.27)$$

converges uniformly on $\Omega_{\mathbb{C}}$ and thus defines an analytic function on $\Omega_{\mathbb{C}}$.

Note that for $\vec{\varepsilon} \geq 0$,

$$p^{\vec{\varepsilon}}(1^{(n)}0) = \pi_1(\vec{\varepsilon})r(\vec{\varepsilon})B(\vec{\varepsilon})^{n-1}\mathbf{1} \quad (7.28)$$

and

$$p^{\vec{\varepsilon}}(01^{(n)}0) = \pi_1(\vec{\varepsilon})r(\vec{\varepsilon})B(\vec{\varepsilon})^{n-1}c(\vec{\varepsilon}). \quad (7.29)$$

By (7.28), (7.29) and (7.22), $H^{\vec{\varepsilon}}(Z)$ agrees with the entropy rate when $\Delta(\vec{\varepsilon}) \geq 0$, as desired.

Remark 7.10. We show how sufficiency relates to Theorem 6.1. Namely, the assumptions in Theorem 7.8 imply those of Theorem 6.1. Condition 1 of Theorem 6.1 follows from the fact that Δ is assumed irreducible. For conditions 2 and 3 of Theorem 6.1, one first notes that the image of f_0 is a single point W_0 , and the f_1 -orbit of W_0 and f_1 -orbit of \vec{s} converge to a point p_1 . It follows that L is the union of W_0 , the f_1 -orbit of W_0 and p_1 . The assumptions in Theorem 7.8. imply that $r_a > 0$ on L (i.e., condition 2 of Theorem 6.1 holds) and that for sufficiently large n , the n -fold composition of f_1 is contracting on the convex hull of the intersection of L and W_1 (so condition 3 of Theorem 6.1 holds). To see the latter, one uses the ideas in the proof of sufficiency.

Proof of necessity

We first consider condition 2. We shall use the natural parameterization and view $H(Z)$ as a function of Δ , or more precisely of (B, r) . Note that there is a one-to-one correspondence between Δ and (B, r) ; we shall use this correspondence throughout the proof.

Suppose Δ doesn't satisfy condition 2, however $H(Z)$ is analytic at Δ with respect to the natural parameterization. In other words, suppose there exists a complex neighborhood N_Δ of Δ (here N_Δ corresponds to $N_B \times N_r$ where N_B is neighborhood of B and N_r is neighborhood of r) such that $H(Z)$ can be analytically extended to N_Δ , while the corresponding B doesn't have isolated (in modulus) maximum eigenvalue.

We first claim there exists $\tilde{\Delta} \in N_\Delta$ with $\tilde{r}\tilde{B}^k\mathbf{1} = 0$, here \tilde{r} and \tilde{B} correspond to $\tilde{\Delta}$ and \tilde{B} has distinct eigenvalues (in modulus). Indeed we can first (for simplicity) perturb Δ to $\tilde{\Delta}$ such that the corresponding \tilde{B} has distinct eigenvalues in modulus. Then

$$\begin{aligned} \tilde{B} &= \tilde{U} \text{diag}(\tilde{\lambda}_1, \tilde{\lambda}_2, \dots, \tilde{\lambda}_{d-1}) \tilde{U}^{-1} \\ &= (\tilde{v}_1, \tilde{v}_2, \dots, \tilde{v}_{d-1}) \text{diag}(\tilde{\lambda}_1, \tilde{\lambda}_2, \dots, \tilde{\lambda}_{d-1}) (\tilde{w}_1^t, \tilde{w}_2^t, \dots, \tilde{w}_{d-1}^t)^t \end{aligned}$$

where $|\tilde{\lambda}_1| > |\tilde{\lambda}_2| > \dots > |\tilde{\lambda}_{d-1}|$, and \tilde{v}_i, \tilde{w}_i 's are appropriately scaled right and left eigenvectors of \tilde{B} , respectively. Then we have

$$r\tilde{B}^k\mathbf{1} = r\tilde{v}_1\tilde{w}_1\mathbf{1}\tilde{\lambda}_1^k + r\tilde{v}_2\tilde{w}_2\mathbf{1}\tilde{\lambda}_2^k + \dots + r\tilde{v}_{d-1}\tilde{w}_{d-1}\mathbf{1}\tilde{\lambda}_{d-1}^k.$$

Further consider a perturbation of B from

$$\tilde{B} = \tilde{U} \text{diag}(\tilde{\lambda}_1, \tilde{\lambda}_2, \dots, \tilde{\lambda}_{d-1}) \tilde{U}^{-1}$$

to

$$\tilde{B} = V\tilde{U} \text{diag}(\tilde{\lambda}_1, \tilde{\lambda}_2, \dots, \tilde{\lambda}_{d-1}) \tilde{U}^{-1}V^{-1},$$

where V is a complex matrix close to the $(d-1) \times (d-1)$ identity matrix I_{d-1} . So we can pick V such that $\tilde{v}_1\tilde{w}_1V^{-1}\mathbf{1} \neq 0$, $\tilde{v}_1\tilde{w}_1V^{-1}\tilde{c} \neq 0$, $\tilde{v}_2\tilde{w}_2V^{-1}\mathbf{1} \neq 0$. Clearly $\tilde{v}_1\tilde{w}_1V^{-1}\mathbf{1}$ is not proportional to $\tilde{v}_2\tilde{w}_2V^{-1}\mathbf{1}$. Then by a further perturbation of r to \tilde{r} , we can simultaneously require that $\tilde{r}\tilde{v}_1\tilde{w}_1\mathbf{1} \neq 0$, $\tilde{r}\tilde{v}_1\tilde{w}_1\tilde{c} \neq 0$, $\tilde{r}\tilde{v}_2\tilde{w}_2\mathbf{1} \neq 0$, $|\tilde{r}\tilde{v}_1\tilde{w}_1\mathbf{1}| \neq |\tilde{r}\tilde{v}_2\tilde{w}_2\mathbf{1}|$, where we redefine $\tilde{v}_i = V\tilde{v}_i$ and $\tilde{w}_i = \tilde{w}_iV^{-1}$. For any θ and $\eta > 0$, it can be checked that

$$\bigcup_{k=0}^{\infty} \{z^k : |z - e^{i\theta}| < \eta\} = \mathbb{C} \setminus \{0\}.$$

Since $\tilde{\lambda}_2$ is a perturbation of $\tilde{\lambda}_1$, it follows that for large enough k , one can perturb $\tilde{\lambda}_2$ to satisfy the equation

$$\left(\tilde{\lambda}_2/\tilde{\lambda}_1\right)^k = \frac{-\tilde{r}\tilde{v}_1\tilde{w}_1\mathbf{1} - \tilde{r}\tilde{v}_3\tilde{w}_3\mathbf{1}(\tilde{\lambda}_3/\tilde{\lambda}_1)^k - \cdots - \tilde{r}\tilde{v}_{d-1}\tilde{w}_{d-1}\mathbf{1}(\tilde{\lambda}_{d-1}/\tilde{\lambda}_1)^k}{\tilde{r}\tilde{v}_2\tilde{w}_2\mathbf{1}},$$

with $|\tilde{\lambda}_2| \neq |\tilde{\lambda}_1|$ and $|\tilde{\lambda}_2|$ strictly greater than $|\tilde{\lambda}_j|$ for $j \geq 3$. Thus we prove the claim.

We now pick a positive matrix $\hat{\Delta} \in N_\Delta$ with corresponding \hat{r} and \hat{B} . We then pick $\tilde{\Delta} \in N_\Delta$ with corresponding \tilde{r} and \tilde{B} (with distinct eigenvalues in modulus) such that $\tilde{r}\tilde{B}^{k_1}\mathbf{1} = 0$ for some k_1 , and we can further require that $\tilde{r}\tilde{v}_1\tilde{w}_1\mathbf{1} \neq 0$, $\tilde{r}\tilde{v}_1\tilde{w}_1\tilde{c} \neq 0$ (see the proof for the previous claim), where as before, \tilde{v}_1, \tilde{w}_1 are eigenvectors corresponding to the largest eigenvalue of \tilde{B} . According to Lemma 7.6, there is an arc $I_1 \subset S(d-1)$ connecting \hat{B} to \tilde{B} ; we then connect \hat{r} and \tilde{r} using an arc I_2 in \mathbb{C}^{d-1} . According to Lemma 7.7, we can choose the arc $I = (I_1, I_2)$ to avoid the hypersurface $V((rv_1w_1\mathbf{1})(rv_1w_1c)) \subset \mathbb{C}^{(d-1)^2} \times \mathbb{C}^{d-1}$; in other words, we can assume that along the path I , $rv_1w_1\mathbf{1} \neq 0$ and $rv_1w_1c \neq 0$; here v_1, w_1, c are determined by the variable matrix B along the path I_1 and r is the variable point along path I_2 (we remind the reader that the coordinates of v_1 and w_1 are all analytic functions of the entries of B). We then claim that there is a neighborhood N_I of I such that $V_k \cap N_I \neq \emptyset$ and $W_k \cap N_I \neq \emptyset$ hold for only finitely many k , where $V_k = \{(B, r) : rB^k\mathbf{1} = 0\}$ and $W_k = \{(B, r) : rB^k c = 0\}$. Indeed for any $\Delta \in I$ with corresponding $B \in S(d-1)$, by the Jordan form we have

$$rB^k\mathbf{1} = rv_1w_1\mathbf{1}\lambda_1^k + o(\lambda_1^k),$$

where λ_1 is the isolated maximum eigenvalue and v_1, w_1 are appropriately scaled right and left eigenvectors of B , respectively. Since $rv_1w_1\mathbf{1} \neq 0$ on I , there exists a complex connected neighborhood N_I of I such that $rv_1w_1\mathbf{1} \neq 0$ on N_I and $rv_1w_1\mathbf{1}\lambda_1^k$ dominates uniformly on N_I (see Lemma 7.5). Consequently, $|rB^k\mathbf{1}| > 0$ on N_I for large enough k . In other words, $V_k \cap N_I \neq \emptyset$ holds for only finitely many k . Similarly since $rv_1w_1c \neq 0$ on I , there exists a complex neighborhood N_I of I (here we use the same notation for a possibly different neighborhood) such that $W_k \cap N_I \neq \emptyset$ holds only for finitely many k . From now on, we assume such k 's are less than some K , which depends on N_I .

We claim that we can further choose I and find a new neighborhood N_I in $\mathbb{C}^{d-1} \times S(d-1)$ of I such that $V_k \cap N_I \neq \emptyset$ holds only for $k = k_1$ and $W_k \cap N_I = \emptyset$ for all k . Consider $\hat{\Delta}$ with corresponding \hat{B} , let $F_i = F_i(\hat{B}) = \{r : r\hat{B}^i\mathbf{1} = 0\}$, which is a hyperplane orthogonal to the vector $\hat{B}^i\mathbf{1}$ in \mathbb{C}^{d-1} . Similarly we define $G_i = G_i(\hat{B}) = \{r : r\hat{B}^i\tilde{c} = 0\}$. Recall that $\hat{B} = \tilde{U}\text{diag}(\lambda_1, \lambda_2, \dots, \lambda_{d-1})\tilde{U}^{-1}$; we can require that $\tilde{U}^{-1}\mathbf{1}$ has no zero coordinates by a small perturbation of \tilde{U} if necessary. We then show that F_i 's and G_j 's define different hyperplanes in \mathbb{C}^{d-1} . Indeed suppose $F_i = F_j$. It follows that $\tilde{U}\text{diag}(\tilde{\lambda}_1^i, \tilde{\lambda}_2^i, \dots, \tilde{\lambda}_{d-1}^i)\tilde{U}^{-1}\mathbf{1}$ is proportional to $\tilde{U}\text{diag}(\tilde{\lambda}_1^j, \tilde{\lambda}_2^j, \dots, \tilde{\lambda}_{d-1}^j)\tilde{U}^{-1}\mathbf{1}$. It then follows that $(\tilde{\lambda}_1^i, \tilde{\lambda}_2^i, \dots, \tilde{\lambda}_{d-1}^i)$ is proportional to $(\tilde{\lambda}_1^j, \tilde{\lambda}_2^j, \dots, \tilde{\lambda}_{d-1}^j)$. However since not all eigenvalues have the same modulus, this implies that $i = j$. With a perturbation of \tilde{c} (equivalently a perturbation of row sums of \hat{B}), if necessary, we conclude that the F_i 's and G_i 's determine different hyperplanes, i.e., $F_i \neq F_j$, $G_i \neq G_j$ for $i \neq j \leq K$, and $F_i \neq G_j$ for all i, j . Thus, with a perturbation of \tilde{r} if necessary, we can choose a new $\hat{\Delta}$ contained in V_{k_1} , but not contained in any V_k with $k \neq k_1$ or W_k for all k . Again by Lemma 7.7, one can choose a new I inside original N_I , connecting

$\hat{\Delta}$ and $\tilde{\Delta}$, to avoid all V_k 's and W_k 's except V_{k_1} , then choose a smaller new neighborhood N_I of the new I to make sure that $V_k \cap N_I \neq \phi$ only holds for $k = k_1$ and $W_k \cap N_I = \phi$ for all k .

Since the perturbed complex matrix B still has spectral radius strictly less than 1, all the complexified terms in the entropy rate formula (see (7.27)) with $k \neq k_1$ are exponentially decaying and thus sum up to an analytic function on N_I (i.e., the sum of these terms can be analytically continued to N_I), while the unique analytic extension of the k_1 -th term on N_I blows up as one approaches $V_{k_1} \cap N_I$ from $\hat{\Delta}$. Again by the uniqueness of analytic extension of $H(Z)$ on N_I , this would be a contradiction to the assumption that $H(Z)$ is analytic at Δ (here we are applying the uniqueness theorem of analytic continuation of a function of several complex variables, see page 21 in [30]). Thus we prove the necessity of condition 2.

We now consider condition 1. Suppose Δ doesn't satisfies condition 1, namely $a = 0$ or $rB^k c = 0$ for some k , however $H(Z)$ is analytic at Δ . With the proof above for the necessity of condition 2, we can now assume the corresponding $B \in S(d-1)$.

If $a = 0$, consider any perturbation of Δ to Δ_1 such that $\tilde{B} \in S(d-1)$, $\tilde{r}\tilde{v}_1\tilde{w}_1\mathbf{1} \neq 0$, $\tilde{r}\tilde{v}_1\tilde{w}_1\tilde{c} \neq 0$, $\tilde{r}\tilde{B}^k\mathbf{1} \neq 0$ and $\tilde{r}\tilde{B}^k\tilde{c} \neq 0$ for all k (here we follow the notation as in the proof of necessity of condition 2). Then using similar arguments, we can prove the sum of all the terms except the first term in the entropy rate formula (see (7.27)) can be analytically extended to $\tilde{\Delta}$. However this implies that $a \log a$ is a well-defined analytic function on some neighborhood of 0 in \mathbb{C} , which is a contradiction. Similar arguments can be applied to the case that $rB^k c = 0$ for some k 's. Thus we prove the necessity of condition 1. \square

8 Analyticity of a Hidden Markov Chain in a Strong Sense

In this section, we show that if Δ is analytically parameterized by a real variable vector $\vec{\varepsilon}$, and at $\vec{\varepsilon}_0$, Δ satisfies conditions 1 and 2 of Theorem 1.1, then the hidden Markov chain *itself* is a real analytic function of $\vec{\varepsilon}$ at $\vec{\varepsilon}_0$ in a strong sense. We assume (for this section only) that the reader is familiar with the basics of measure theory and functional analysis [17, 34, 18]. Our approach uses a connection between the entropy rate of a hidden Markov chain and symbolic dynamics explored in [16].

Let \mathcal{X} denote the set of left infinite sequences with finite alphabet. A cylinder set is a set of the form: $(\{x_{-\infty}^0 : x_0 = z_0, \dots, x_{-n} = z_{-n}\})$. The Borel sigma-algebra is the smallest sigma-algebra containing the cylinder sets. A Borel probability measure (BPM) ν on \mathcal{X} is a measure on the Borel measurable sets of \mathcal{X} such that $\nu(\mathcal{X}) = 1$. Such a measure is uniquely determined by its values on the cylinder sets.

For real $\vec{\varepsilon}$, consider the measure $\nu^{\vec{\varepsilon}}$ on \mathcal{X} defined by:

$$\nu^{\vec{\varepsilon}}(\{x_{-\infty}^0 : x_0 = z_0, \dots, x_{-n} = z_{-n}\}) = p^{\vec{\varepsilon}}(z_{-n}^0). \quad (8.30)$$

Note that $H(Z)$ can be rewritten as

$$H^{\vec{\varepsilon}}(Z) = \int -\log p^{\vec{\varepsilon}}(z_0 | z_{-\infty}^{-1}) d\nu^{\vec{\varepsilon}}. \quad (8.31)$$

Usually, the Borel sigma-algebra is defined to be the smallest sigma-algebra containing the open sets; in this case, the open sets are defined by the metric: for any two elements ξ

and η in \mathcal{X} , define $d(\xi, \eta) = 2^{-k}$ where $k = \inf\{|i| : \xi_i \neq \eta_i\}$. The metric space (\mathcal{X}, d) is compact.

Let $C(\mathcal{X})$ be the space of real-valued continuous functions on \mathcal{X} . Then $C(\mathcal{X})$ is a Banach space (i.e., complete normed linear space) with the sup norm $\|f\|_\infty = \sup\{|f(x)| : x \in \mathcal{X}\}$. Then any BPM ν acts as a bounded linear functional on $C(\mathcal{X})$, namely $\nu(f) = \int f d\nu$. As such, the set of BPM's is a subset of the dual space, $C(\mathcal{X})^*$, which is itself a Banach space; the norm of a BPM ν is defined: $\|\nu\| = \sup_{\{f \in C(\mathcal{X}) : \|f\|_\infty = 1\}} \int f d\nu$. In fact, since \mathcal{X} is compact, $C(\mathcal{X})^*$ is the linear span of the BPM's.

It makes sense to ask if $\vec{\varepsilon} \mapsto \nu^{\vec{\varepsilon}}$ is analytic as a mapping from the parameter space to $C(\mathcal{X})^*$; by definition, this would mean that $\nu^{\vec{\varepsilon}}$ can be expressed as a power series in the coordinates of $\vec{\varepsilon}$. However, as the following example shows, this mapping is not even continuous.

Let \mathcal{X} be the set of binary left infinite sequences. Let ν_p denote the i.i.d. $(p, 1-p)$ measure, with $0 < p < 1$. We claim that, for fixed $p > q$ and ε with $0 < \varepsilon < (p-q)/2$, by application of the law of large numbers to ν_p and ν_q , one can find a finite union C of cylinder sets such that

$$\nu_p(C) > 1 - \varepsilon$$

and

$$\nu_q(C) < \varepsilon.$$

To see this, first say that a word $w = w_1 \dots w_n$ is (p, ε, n) -typical if the frequency of 0's in w is in $[p - \varepsilon, p + \varepsilon]$. Let $C_{p, \varepsilon, n}$ be the union of the cylinder sets corresponding to the (p, ε, n) -typical words. The law of large numbers asserts that for n sufficiently large

$$\nu_p(C_{p, \varepsilon, n}) > 1 - \varepsilon \text{ and } \nu_q(C_{q, \varepsilon, n}) > 1 - \varepsilon$$

Since $p > q + 2\varepsilon$, then $C_{p, \varepsilon, n}$ and $C_{q, \varepsilon, n}$ are disjoint. Thus, we have

$$\nu_q(C_{p, \varepsilon, n}) < 1 - \nu_q(C_{q, \varepsilon, n}) < \varepsilon.$$

Set $C = C_{p, \varepsilon, n}$.

Now, if f is the characteristic function of C , we have

$$\|\nu_p - \nu_q\| \geq \int f d\nu_p - \int f d\nu_q = \nu_p(C) - \nu_q(C) > 1 - 2\varepsilon.$$

So, $\|\nu_p - \nu_q\| \geq 1$. It follows that ν_q cannot converge in norm to ν_p as $q \rightarrow p$, and so the map $p \mapsto \nu_p$ from R to $C(\mathcal{X})^*$ is discontinuous.

On the other hand, using the work of Ruelle [26], we now show that $\vec{\varepsilon} \mapsto \nu^{\vec{\varepsilon}}$ is analytic as a mapping from the parameter space to another natural space.

For $f \in C(\mathcal{X})$, define $var_n(f) = \sup\{|f(\xi) - f(\xi')| : \xi_{-i} = \xi'_{-i} \text{ for } i \leq n\}$. We denote by F^θ the subset of $f \in C(\mathcal{X})$ such that

$$\|f\|_\theta \equiv \sup_{n \geq 0} (\theta^{-n} var_n(f)) < +\infty.$$

F^θ is a Banach space with the norm $\|f\| = \max(\|f\|_\infty, \|f\|_\theta)$. Let $(F^\theta)^*$ denote the dual space (i.e., the set of bounded linear functionals) on F^θ . For any $\nu \in (F^\theta)^*$, the norm of ν

is naturally defined as $\|\nu\| = \sup_{\{f \in F^\theta: \|f\|=1\}} \nu(f)$. Using complex functions instead of real functions, one defines $F_{\mathbb{C}}^\theta$ and $(F_{\mathbb{C}}^\theta)^*$ similarly.

In the following theorem, we prove the analyticity of a hidden Markov chain in a strong sense.

Theorem 8.1. *Suppose that the entries of Δ are analytically parameterized by a real variable vector $\vec{\varepsilon}$. If at $\vec{\varepsilon} = \vec{\varepsilon}_0$, Δ satisfies conditions 1 and 2 in Theorem 1.1, then the mapping $\vec{\varepsilon} \mapsto \log p^{\vec{\varepsilon}}(z_0|z_{-\infty}^{-1})$ is analytic at $\vec{\varepsilon}_0$ from the real parameter space to F^ρ (here ρ is the contraction constant in the proof of Theorem 1.1). Moreover the mapping $\vec{\varepsilon} \mapsto \nu^{\vec{\varepsilon}}$ is analytic at $\vec{\varepsilon}_0$ from the real parameter space to $(F^\rho)^*$.*

Proof. For complex $\vec{\varepsilon}$, by (4.16), one shows that $\log p^{\vec{\varepsilon}}(z_0|z_{-\infty}^{-1})$ can be defined on $\Omega_{\mathbb{C}}$ as the uniform (in $\vec{\varepsilon}$ and $z \in \mathcal{X}$) limit of $\log p^{\vec{\varepsilon}}(z_0|z_{-n}^{-1})$ as $n \rightarrow \infty$, and $\log p^{\vec{\varepsilon}}(z_0|z_{-\infty}^{-1})$ belongs to $F_{\mathbb{C}}^\rho$. By (4.5), (4.6), (4.7) and (4.14) it follows that $p^{\vec{\varepsilon}}(z_0|z_{-n}^{-1})$ is analytic on $\Omega_{\mathbb{C}}$. As a result of (4.16), if Δ satisfies conditions 1 and 2, for fixed $z \in \mathcal{X}$, $\log p^{\vec{\varepsilon}}(z_0|z_{-\infty}^{-1})$ is the uniform limit of analytic functions and hence is analytic on $\Omega_{\mathbb{C}}$ (see Theorem 2.4.1 of [32]).

For a given sequence $z = z_{-\infty}^0$, let $f(\vec{\varepsilon}; z) = \log p^{\vec{\varepsilon}}(z_0|z_{-\infty}^{-1})$. Let $D_{\vec{\varepsilon}}f|_{\vec{\varepsilon}_0}(z)$ denote the vector of partial derivatives of $f(\vec{\varepsilon}; z)$ with respect to $\vec{\varepsilon}$ at $\vec{\varepsilon} = \vec{\varepsilon}_0$. Using (4.16) and the Cauchy integral formula in several variables [32] (which expresses the derivative of an analytic function at a point as an integral on a closed surface around the point), we obtain the following. There is a positive constant C' such that whenever $z \stackrel{n}{\sim} \hat{z}$, for all $\vec{\varepsilon}_0 \in \Omega_{\mathbb{C}}$

$$|D_{\vec{\varepsilon}}f|_{\vec{\varepsilon}_0}(z) - D_{\vec{\varepsilon}}f|_{\vec{\varepsilon}_0}(\hat{z})| \leq C' \rho^n. \quad (8.32)$$

For a direction \vec{h} in the parameter space, let $D_{\vec{\varepsilon}}f|_{\vec{\varepsilon}_0}(\vec{h}; z)$ denote the directional derivative of $f(\vec{\varepsilon}; z)$ at $\vec{\varepsilon} = \vec{\varepsilon}_0$ in direction \vec{h} . Let $D_{\vec{\varepsilon}}f|_{\vec{\varepsilon}_0}(\vec{h}; \cdot)$ denote the function on \mathcal{X} , whose value on $z = z_{-\infty}^0 \in \mathcal{X}$ is given by $D_{\vec{\varepsilon}}f|_{\vec{\varepsilon}_0}(\vec{h}; z)$. By (8.32), $D_{\vec{\varepsilon}}f|_{\vec{\varepsilon}_0}(\vec{h}; \cdot)$ belongs to $F_{\mathbb{C}}^\rho$.

Now, we must prove that the mapping $\vec{\varepsilon} \mapsto \log p^{\vec{\varepsilon}}(z_0|z_{-\infty}^{-1})$ is complex differentiable (therefore analytic) from $\Omega_{\mathbb{C}}$ to $F_{\mathbb{C}}^\rho$. For this, it suffices to prove that for all $\vec{\varepsilon}_0 \in \Omega_{\mathbb{C}}$,

$$\|f(\vec{\varepsilon}_0 + \vec{h}; \cdot) - f(\vec{\varepsilon}_0; \cdot) - D_{\vec{\varepsilon}}f|_{\vec{\varepsilon}_0}(\vec{h}; \cdot)\|_\infty \leq o(|\vec{h}|). \quad (8.33)$$

and

$$\|f(\vec{\varepsilon}_0 + \vec{h}; \cdot) - f(\vec{\varepsilon}_0; \cdot) - D_{\vec{\varepsilon}}f|_{\vec{\varepsilon}_0}(\vec{h}; \cdot)\|_\rho \leq o(|\vec{h}|). \quad (8.34)$$

Let $D_{\vec{\varepsilon}}^2f|_{\vec{\varepsilon}_0}(\vec{h}, \vec{h}; z)$ denote the second directional derivative in direction (\vec{h}, \vec{h}) of $f(\vec{\varepsilon}; z)$ at $\vec{\varepsilon} = \vec{\varepsilon}_0$. Again applying the Cauchy integral formula in several variables, it follows that there exists a positive constant C'' such that for all $\vec{\varepsilon}_0 \in \Omega_{\mathbb{C}}$ we have

$$|D_{\vec{\varepsilon}}^2f|_{\vec{\varepsilon}_0}(\vec{h}, \vec{h}; z)| \leq C'' |\vec{h}|^2 \quad (8.35)$$

and whenever $z \stackrel{n}{\sim} \hat{z}$,

$$\int_0^1 (1-t) |(D_{\vec{\varepsilon}}^2f|_{\vec{\varepsilon}_0}(\vec{h}, \vec{h}; z) - D_{\vec{\varepsilon}}^2f|_{\vec{\varepsilon}_0}(\vec{h}, \vec{h}; \hat{z}))| dt \leq C'' |\vec{h}|^2 \rho^n, \quad (8.36)$$

From the Taylor formula with integral remainder, we have:

$$f(\vec{\varepsilon}_0 + \vec{h}; z) - f(\vec{\varepsilon}_0; z) - D_{\vec{\varepsilon}}f|_{\vec{\varepsilon}_0}(\vec{h}; z) = \int_0^1 (1-t) D_{\vec{\varepsilon}}^2f|_{\vec{\varepsilon}_0+t\vec{h}}(\vec{h}, \vec{h}; z) dt. \quad (8.37)$$

To prove (8.33), use (8.35) and (8.37). To prove (8.34), use (8.36) and (8.37). Therefore $\vec{\varepsilon} \mapsto \log p^{\vec{\varepsilon}}(\cdot)$ is analytic as a mapping from $\Omega_{\mathbb{C}}$ to $F_{\mathbb{C}}^{\rho}$. Restricting the mapping $\vec{\varepsilon} \mapsto \log p^{\vec{\varepsilon}}(z_0 | z_{-\infty}^{-1})$ to the real parameter space, we conclude that it is real analytic (as a mapping into F^{ρ}). Using this and the theory of equilibrium states [26], the ‘‘Moreover’’ is proven in Appendix C. \square

Corollary 8.2. *Suppose that at $\vec{\varepsilon}_0$, Δ satisfies conditions 1 and 2 in Theorem 1.1, and $\vec{\varepsilon} \mapsto f^{\vec{\varepsilon}} \in F^{\rho}$ is analytic at $\vec{\varepsilon}_0$, then $\vec{\varepsilon} \mapsto \nu^{\vec{\varepsilon}}(f^{\vec{\varepsilon}})$ is analytic at $\vec{\varepsilon}_0$. In particular, we recover Theorem 1.1: $\vec{\varepsilon} \mapsto H^{\vec{\varepsilon}}(Z)$ is analytic at $\vec{\varepsilon}_0$.*

Proof. The map

$$\begin{aligned} \Omega &\rightarrow F^{\rho} \times (F^{\rho})^* \rightarrow \mathbb{R} \\ \vec{\varepsilon} &\mapsto (f^{\vec{\varepsilon}}, \nu^{\vec{\varepsilon}}) \mapsto \nu^{\vec{\varepsilon}}(f^{\vec{\varepsilon}}) \end{aligned}$$

is analytic at $\vec{\varepsilon}_0$, as desired. \square

Acknowledgements: We are grateful to Wael Bahsoun, Joel Feldman, Robert Israel, Izabella Laba, Erik Ordentlich, Yuval Peres, Gadiel Seroussi, Wojciech Szpankowski and Tsachy Weissman for helpful discussions.

Appendices

A Proof of Proposition 2.1

Proof. Without loss of generality, we assume S is convex (otherwise consider the convex hull of S). It follows from standard arguments that max norm and sum norm are equivalent. More specifically, for another metric d_1 defined by

$$d_1(u, v) = \sqrt{\sum_{i \neq j \leq k} \log^2 \left(\frac{u_i/u_j}{v_i/v_j} \right)},$$

we have $d_{\mathbf{B}} \sim d_1$. For metric d_2 defined by

$$d_2(u, v) = \sqrt{\sum_{i \neq j \leq k} (u_i/u_j - v_i/v_j)^2}.$$

Applying mean value theorem to log function, one concludes that $d_1 \sim d_2$. Note that

$$\begin{aligned} u_i - v_i &= \frac{u_i}{u_1 + u_2 + \cdots + u_k} - \frac{v_i}{v_1 + v_2 + \cdots + v_k} \\ &= \frac{1}{u_1/u_i + u_2/u_i + \cdots + u_k/u_i} - \frac{1}{v_1/v_i + v_2/v_i + \cdots + v_k/v_i} \end{aligned}$$

Applying the mean value theorem to function f , defined as

$$f(x_1, x_2, \dots, x_B) = \frac{1}{x_1 + x_2 + \dots + x_k},$$

we conclude that there exists $\xi \in S$ such that

$$u_i - v_i = \nabla f|_{\xi} \cdot (u_1/u_i - v_1/v_i, \dots, u_k/u_i - v_k/v_i).$$

It follows from Cauchy inequality that there exists a positive constant D_1 such that

$$d_{\mathbf{E}}(u, v) < D_1 d_2(u, v).$$

Similarly consider $u_i/u_j - v_i/v_j$, and apply mean value theorem to function g , defined as $g(x, y) = x/y$, we show that there exists a positive constant D_2 such that

$$d_2(u, v) < D_2 d_{\mathbf{E}}(u, v).$$

Namely $d_2 \sim d_{\mathbf{E}}$. Thus the claim in this Proposition follows, namely there exist two positive constant $C_1 < C_2$ such that for any two points $u, v \in S$,

$$C_1 d_{\mathbf{B}}(u, v) < d_{\mathbf{E}}(u, v) < C_2 d_{\mathbf{B}}(u, v).$$

□

B Proof of Lemma 7.9:

Recall that for a non-negative matrix B , the *canonical form* of B is:

$$B = \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1n} \\ 0 & B_{22} & \cdots & B_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & B_{nn} \end{bmatrix},$$

where B_{ii} is either an irreducible matrix (called *irreducible components*) or a 1×1 zero matrix.

Condition 2 in Theorem 7.8 is equivalent to the statement that $B = B(\vec{\varepsilon}_0)$ has a unique irreducible component of maximal spectral radius and that this component is primitive. Let C denote the square matrix obtained by restricting B to this component and let S_C denote the set of indices corresponding to this component. Let λ_1 denote the spectral radius of B , equivalently the spectral radius of C .

Let $\lambda_1(\vec{\varepsilon})$ denote the largest, in modulus, eigenvalue of $B(\vec{\varepsilon})$. Since the entries of $B(\vec{\varepsilon})$ are analytic in $\vec{\varepsilon}$ and λ_1 is simple, it follows that if the complex neighborhood Ω is chosen sufficiently small, then $\lambda_1(\vec{\varepsilon})$ is analytic function of $\vec{\varepsilon} \in \Omega$.

The columns (resp., rows) of $Adj(\lambda_1(\vec{\varepsilon})I - B(\vec{\varepsilon}))$ are right (resp., left) eigenvectors of $B(\vec{\varepsilon})$ corresponding to $\lambda_1(\vec{\varepsilon})$. By choosing $x(\vec{\varepsilon})$ (resp. $y(\vec{\varepsilon})$) to be a fixed column (resp. row) of $Adj(\lambda_1(\vec{\varepsilon})I - B(\vec{\varepsilon}))$ and then replacing $x(\vec{\varepsilon})$ and $y(\vec{\varepsilon})$ by appropriately rescaled versions, we may assume that:

- $x(\vec{\varepsilon}_0), y(\vec{\varepsilon}_0) \geq 0$, and they are positive on S_C
- $y(\vec{\varepsilon}) \cdot x(\vec{\varepsilon}) = 1$
- $x(\vec{\varepsilon})$ and $y(\vec{\varepsilon})$ are analytic in $\vec{\varepsilon} \in \Omega$

Let

$$V(\vec{\varepsilon}) = \lambda_1(\vec{\varepsilon})x(\vec{\varepsilon}) \cdot y(\vec{\varepsilon})$$

and

$$U(\vec{\varepsilon}) = B(\vec{\varepsilon}) - V(\vec{\varepsilon}).$$

Then

$$U(\vec{\varepsilon})V(\vec{\varepsilon}) = B(\vec{\varepsilon})V(\vec{\varepsilon}) - V^2(\vec{\varepsilon}) = \lambda_1^2(\vec{\varepsilon})x(\vec{\varepsilon}) \cdot y(\vec{\varepsilon}) - \lambda_1^2(\vec{\varepsilon})x(\vec{\varepsilon}) \cdot y(\vec{\varepsilon}) = \mathbf{0}.$$

And similarly

$$V(\vec{\varepsilon})U(\vec{\varepsilon}) = \mathbf{0}.$$

Let $\mu(\vec{\varepsilon})$ denote the spectral radius of $U(\vec{\varepsilon})$. By condition 2, $\mu(\vec{\varepsilon}_0) < \lambda_1(\vec{\varepsilon}_0)$. Thus, there is a constant $\nu > 0$ such that if the neighbourhood Ω is sufficiently small, then for all $\vec{\varepsilon} \in \Omega$

$$\mu(\vec{\varepsilon}) < \nu < |\lambda_1(\vec{\varepsilon})|.$$

Thus, by Lemma 7.5, and making still Ω smaller if necessary, there is a constant $K_1 > 0$ such that for all i, j , all n and all $\vec{\varepsilon} \in \Omega$,

$$|U_{ij}^n(\vec{\varepsilon})| < K_1 \nu^n. \tag{B.38}$$

Let $r = r(\vec{\varepsilon}_0)$, $c = c(\vec{\varepsilon}_0)$, $x = x(\vec{\varepsilon}_0)$ and $y = y(\vec{\varepsilon}_0)$.

Let $s_0 \in S_C$. Since $\Delta(\vec{\varepsilon}_0)$ is irreducible, for some j_0 , $(rB^{j_0})_{s_0} > 0$. Similarly, there exist a state s_1 of the underlying Markov chain and j_1 such that $B_{s_0 s_1}^{j_1} c_{s_1} > 0$. Now

$$rB^n c \geq (rB^{j_0})_{s_0} C_{s_0 s_0}^{n-j_0-j_1} B_{s_0 s_1}^{j_1} c_{s_1}.$$

Since C is primitive, by Perron-Frobenius theory, $C_{s_0 s_0}^{n-j_0-j_1}$ grows like $\lambda_1^{n-j_0-j_1}$ (up to a scalar) as n goes to infinity; it then follows that there is a constant K_2 such that for sufficiently large n ,

$$rx \cdot yc \lambda_1^n + rU^n c = rV^n c + rU^n c = rB^n c > K_2 \lambda_1^n,$$

which by (B.38) implies that $rx \cdot yc > 0$. Therefore if Ω is sufficiently small, there exists a positive constant K_4 such that

$$|r(\vec{\varepsilon})x(\vec{\varepsilon}) \cdot y(\vec{\varepsilon})c(\vec{\varepsilon})| > K_4,$$

for $\vec{\varepsilon} \in \Omega$.

Let K_3 be an upper bound on the entries of $|x(\vec{\varepsilon})|$, $|y(\vec{\varepsilon})|$, $|r(\vec{\varepsilon})|$ and $|c(\vec{\varepsilon})|$.

Thus, for all n and all $\vec{\varepsilon} \in \Omega$, we have

$$|r(\vec{\varepsilon})B^n(\vec{\varepsilon})c(\vec{\varepsilon})| \leq |r(\vec{\varepsilon})U^n(\vec{\varepsilon})c(\vec{\varepsilon})| + |r(\vec{\varepsilon})V^n(\vec{\varepsilon})c(\vec{\varepsilon})| \leq |\mathcal{B}|^2 K_3^2 K_1 \nu^n + |\mathcal{B}|^2 K_3^4 |\lambda_1(\vec{\varepsilon})|^n$$

and

$$|r(\vec{\varepsilon})B^n(\vec{\varepsilon})c(\vec{\varepsilon})| \geq |r(\vec{\varepsilon})V^n(\vec{\varepsilon})c(\vec{\varepsilon})| - |r(\vec{\varepsilon})U^n(\vec{\varepsilon})c(\vec{\varepsilon})| \geq K_4|\lambda_1(\vec{\varepsilon})|^n - |\mathcal{B}|^2K_3^2K_1\nu^n.$$

With similar upper and lower bounds for $|r(\vec{\varepsilon})B^n(\vec{\varepsilon})\mathbf{1}|$, it follows that for sufficiently large n and all $\vec{\varepsilon} \in \Omega$,

$$\frac{\pi_1(\vec{\varepsilon})r(\vec{\varepsilon})B(\vec{\varepsilon})^n\mathbf{1}}{\pi_1(\vec{\varepsilon})r(\vec{\varepsilon})B(\vec{\varepsilon})^{n-1}\mathbf{1}}$$

and

$$\frac{\pi_1(\vec{\varepsilon})r(\vec{\varepsilon})B(\vec{\varepsilon})^{n-1}c(\vec{\varepsilon})}{\pi_1(\vec{\varepsilon})r(\vec{\varepsilon})B(\vec{\varepsilon})^{n-1}\mathbf{1}}$$

are uniformly bounded from above and away from zero. By condition 1, for any finite collection of n , there is a (possibly smaller) neighborhood Ω of $\vec{\varepsilon}_0$, such that for all $\vec{\varepsilon} \in \Omega$, these quantities are uniformly bounded from above and away from zero. This completes the proof of Lemma 7.9 (and therefore the proof of sufficiency for Theorem 7.8.)

C $\vec{\varepsilon} \mapsto \nu^{\vec{\varepsilon}}$ is analytic

In this appendix, we follow the notation in Section 8. Let $\tau : \mathcal{X} \rightarrow \mathcal{X}$ be the right shift operator, which is a continuous mapping on \mathcal{X} under the topology induced by the metric d . For $f \in C(\mathcal{X})$, one defines the *pressure* via a variational principle [26]:

$$P(f) = \sup_{\mu \in M(\mathcal{X}, \tau)} \left(H_\mu(\tau) + \int f d\mu \right),$$

where $M(\mathcal{X}, \tau)$ denotes the set of τ -invariant probability measures on \mathcal{X} and $H_\mu(\tau)$ denotes measure-theoretic entropy. A member μ of $M(\mathcal{X}, \tau)$ is called an *equilibrium state* for f if $P(f) = H_\mu(\tau) + \int f d\mu$.

For $f \in C(\mathcal{X})$ the Ruelle operator $\mathcal{L}_f : C(\mathcal{X}) \rightarrow C(\mathcal{X})$ is defined [26] by

$$(\mathcal{L}_f h)(x) = \sum_{y \in \tau^{-1}(x)} e^{f(y)} h(y).$$

The connection between pressure and the Ruelle operator is as follows [26, 28]. When $f \in F^\theta$, $P(f)$ is $\log \lambda$, where λ is the spectral radius of \mathcal{L}_f . The restriction of \mathcal{L}_f to F^θ still has spectral radius λ , and λ is isolated from all other eigenvalues of the restricted operator. Using this, Ruelle applied standard perturbation theory for linear operators [13] to conclude that pressure $P(f)$ is real analytic on F^θ . Moreover, he showed that each $f \in F^\theta$ has a unique equilibrium state μ_f and the first order derivative of $f \mapsto P(f)$ on F^θ is μ_f , viewed as a linear functional on F^θ . So, the analyticity of $P(f)$ implies that the equilibrium state μ_f is also analytic in $f \in F^\theta$.

We first claim that for $f(\vec{\varepsilon}, z) = \log p^{\vec{\varepsilon}}(z_0 | z_{-\infty}^{-1})$, we have $\mu_{f(\vec{\varepsilon}, \cdot)} = \nu^{\vec{\varepsilon}}$ as in (8.30).

To see this, first observe that the spectral radius λ of $\mathcal{L} = \mathcal{L}_{f(\vec{\varepsilon}, \cdot)}$ is 1; this follows from the observations:

- the function $\bar{1}$ which is identically 1 on \mathcal{X} is a fixed point of \mathcal{L} – and –

- (see Proposition 5.16 of [26]) $\mathcal{L}^n(\bar{1})/\lambda^n$ converges to a strictly positive function.

Thus $P(f(\vec{\varepsilon}, \cdot)) = 0$. So, for $\mu^{\vec{\varepsilon}} = \mu_{f(\vec{\varepsilon}, \cdot)}$, we have

$$h_{\mu^{\vec{\varepsilon}}}(\tau) + \int f(\vec{\varepsilon}, \cdot) d\mu^{\vec{\varepsilon}} = 0.$$

But from (8.31), we have

$$h_{\nu}(\tau) + \int f(\vec{\varepsilon}, \cdot) d\nu^{\vec{\varepsilon}} = 0.$$

By uniqueness of the equilibrium state, we thus obtain $\mu_{f(\vec{\varepsilon}, \cdot)} = \nu^{\vec{\varepsilon}}$ as claimed.

Since $\vec{\varepsilon} \mapsto f(\vec{\varepsilon}, \cdot)$ is analytic, it then follows that $\vec{\varepsilon} \mapsto \nu^{\vec{\varepsilon}}$ is analytic, thereby completing the proof of Theorem 8.1.

References

- [1] L. Arnold, V. M. Gundlach and L. Demetrius. Evolutionary formalism for products of positive random matrices. *Annals of Applied Probability*, 4:859–901, 1994.
- [2] D. Arnold and H. Loeliger. The information rate of binary-input channels with memory. Proc. 2001 IEEE Int. Conf. on Communications, (Helsinki, Finland), pp. 2692–2695, June 11-14 2001.
- [3] J. J. Birch. Approximations for the entropy for functions of Markov chains. *Ann. Math. Statist.*, 33:930–938, 1962.
- [4] D. Blackwell. The entropy of functions of finite-state Markov chains. *Trans. First Prague Conf. Information Theory, Statistical Decision Functions, Random Processes*, pages 13–20, 1957.
- [5] M. Cassandro and E. Olivieri. Renormalization group and analyticity in one dimension: A proof of Dobrushin’s theorem *Commun. Math. Phys.*, 80, 255-269, 1981.
- [6] J. R. Chazottes and E. Ugaldé. Projection of Markov measures may be Gibbsian. *J. Statist. Phys.*, Volume 111, Numbers 5-6, 1245-1272.
- [7] R. L. Dobrushin. Analyticity of correlation functions in one-dimensional classical systems with slowly decreasing potentials. *Commun. Math. Phys.* 32, 269-289, 1973.
- [8] S. Egner, V. Balakirsky, L. Tolhuizen, S. Baggen and H. Hollmann. On the entropy rate of a hidden Markov model. In *Proceedings of the 2004 IEEE International Symposium on Information Theory*, page 12, Chicago, U.S.A., 2004.
- [9] G. Han and B. Marcus. Analyticity of entropy rate of a hidden Markov chain In Proc. of IEEE International Symposium on Information Theory, Adelaide, Australia, September 4-September 9 2005, pages 2193-2197.

- [10] R. Gharavi and V. Anantharam. An upper bound for the largest Lyapunov exponent of a Markovian product of nonnegative matrices. Preprint, January 1995.
- [11] T. Holliday, A. Goldsmith and P. Glynn. On entropy and Lyapunov exponents for finite state channels. 2003. Available at <http://wsl.stanford.edu/Publications/THolliday/Lyapunov.pdf>.
- [12] P. Jacquet, G. seroussi and W. Szpankowski. On the entropy of a hidden Markov process. In *Proceedings of the 2004 IEEE International Symposium on Information Theory*, page 10, Chicago, U.S.A., 2004.
- [13] T. Kato. *Perturbation Theory for Linear Operators*. Springer Verlag, Berlin-Heidelberg-New York, 1976.
- [14] D. Lind and B. Marcus. *An Introduction to Symbolic Dynamics and Coding*. Cambridge University Press, 1995.
- [15] J. Lőrinczi, C. Maes and K. V. Velde. Transformations of Gibbs measures. *Probab. Theory Relat. Fields*, Volume 112, 121-147, 1998.
- [16] B. Marcus, K. Petersen and S. Williams. Transmission rates and factors of Markov chains. *Contemporary Mathematics*, 26:279–294, 1984.
- [17] A. Mukherjea and K. Pothoven. *Real and functional analysis*. Plenum Press, New York, 1978.
- [18] L. Nachbin. *Introduction to functional analysis : Banach spaces and differential calculus*. New York : M. Dekker, 1981.
- [19] A. Onishchik. *Lie groups and Lie algebra I*. Encyclopaedia of mathematical sciences ; v. 20. Springer-Verlag, 1993.
- [20] E. Ordentlich and T. Weissman. On the optimality of symbol by symbol filtering and denoising. *Information Theory, IEEE Transactions*, Volume 52, Issue 1, Jan. 2006 Page(s):19 - 40.
- [21] E. Ordentlich and T. Weissman. New bounds on the entropy rate of hidden Markov process. *Information Theory Workshop*, 2004. IEEE 24-29 Oct. 2004 Page(s):117 - 122
- [22] Y. Peres. *Analytic dependence of Lyapunov exponents on transition probabilities*, volume 1486 of *Lecture Notes in Mathematics, Lyapunov's exponents, Proceedings of a Workshop*. Springer Verlag, 1990.
- [23] Y. Peres. Domains of analytic continuation for the top Lyapunov exponent. *Ann. Inst. H. Poincaré Probab. Statist.*, 28(1):131–148, 1992.
- [24] K. Petersen, A. Quas and S. Shin. Measures of maximal relative entropy. *Ergod. Th. and Dynam. Sys.*, 23, 207-223, 2003

- [25] H. Pfister, J. Soriaga and P. Siegel. The achievable information rates of finite-state ISI channels. Proc. IEEE GLOBECOM, (San Antonio, TX), pp. 2992–2996, Nov. 2001.
- [26] D. Ruelle. *Thermodynamic formalism : the mathematical structures of classical equilibrium statistical mechanics*. Addison-Wesley Pub. Co., Advanced Book Program, Reading, Mass, 1978.
- [27] D. Ruelle. Analyticity properties of the characteristic exponents of random matrix products. *Adv. Math.*, 32:68–80, 1979.
- [28] D. Ruelle. Differentiation of SRB states. *Comm. Math. Phys.*, 187(1):227–241, 1997.
- [29] E. Seneta. *Springer Series in Statistics. Non-negative Matrices and Markov Chains*. Springer-Verlag, New York Heidelberg Berlin, 1980.
- [30] B. V. Shabat. *Introduction to complex analysis*. Translations of mathematical monographs ; v. 110. American Mathematical Society, Providence, R.I., 1992.
- [31] V. Sharma and S. Singh. Entropy and channel capacity in the regenerative setup with applications to Markov channels. Proc. IEEE Intern. Symp. on Inform. Theory, (Washington, D.C.), p. 283, June 24-29 2001.
- [32] J. L. Taylor. *Several complex variables with connections to algebraic geometry and Lie groups*. American Mathematical Society, Providence, R.I., 2002.
- [33] P. Walters. *An introduction to ergodic theory*. volume 79 of *Graduate texts in mathematics*. Springer-Verlag, New York, 1982.
- [34] K. Yosida. *Functional analysis*, 4th edition. Springer-Verlag, Berlin, 1974.
- [35] O. Zuk, I. Kanter and E. Domany. The entropy of a binary hidden Markov process. *J. Stat. Phys.*, 121(3-4): 343-360 (2005)
- [36] O. Zuk, E. Domany, I. Kanter, and M. Aizenman. Taylor series expansions for the entropy rate of hidden Markov Processes. ICC 2006, Istanbul.