<table>
<thead>
<tr>
<th>Title</th>
<th>Analyticity of entropy rate of hidden Markov chains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Han, G; Marcus, B</td>
</tr>
<tr>
<td>Citation</td>
<td>IEEE Transactions On Information Theory, 2006, v. 52 n. 12, p. 5251-5266</td>
</tr>
<tr>
<td>Issued Date</td>
<td>2006</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://hdl.handle.net/10722/156183">http://hdl.handle.net/10722/156183</a></td>
</tr>
<tr>
<td>Rights</td>
<td>This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.; ©2006 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.</td>
</tr>
</tbody>
</table>
Analyticity of Entropy Rate of Hidden Markov Chains

Guangyue Han and Brian Marcus, Fellow, IEEE

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.

Index Terms—Analyticity, entropy, entropy rate, hidden Markov chain, hidden Markov process.

I. INTRODUCTION

For $m,n \in \mathbb{N}$ with $m \leq n$, we denote a sequence of symbols $y_m, y_{m+1}, \ldots, y_n$ by $y_n^m$. Consider a stationary stochastic process $Y$ with a finite set of states $\{1, 2, \ldots, B\}$ and distribution $p(y_0^n)$. Denote the conditional distributions by $p(y_{n+1}|y_n^m)$. The entropy rate of $Y$ is defined as

$$H(Y) = \lim_{n \to \infty} -E_{p}(\log(p(y_0^n)/y_0^n))$$

where $E_p$ denotes expectation with respect to the distribution $p$.

Let $Y$ be a stationary first-order Markov chain with $\Delta(i,j) = p(y_i = y_j | y_0 = i)$. 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 $\Phi$ is a function defined on $\{1, 2, \ldots, B\}$ with values $\{1, 2, \ldots, A\}$. 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 (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 $\Delta$ are analytically parameterized by a real variable vector $\varepsilon$. If $\varepsilon = \varepsilon_0$

1. for all $a \in \{1, 2, \ldots, A\}$, there is at least one $j$ with $\Phi(j) = a$ such that the $j$th column of $\Delta$ is strictly positive; and
2. every column of $\Delta$ is either all zero or strictly positive; then $H(Z)$ is a real analytic function of $\varepsilon$ at $\varepsilon_0$.

Note that this theorem holds if all the entries of $\Delta$ 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 III.

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

Manuscript received July 12, 2005; revised August 8, 2006. The material in this paper was presented at the IEEE International Symposium on Information Theory, Adelaide, Australia, September 2005.

The authors are with the Department of Mathematics, University of British Columbia, Vancouver, BC V6T 1Z2, Canada (e-mail: ghan@math.ubc.ca; marcus@math.ubc.ca).

Communicated by S. A. Savari, Associate Editor for Source Coding.

Digital Object Identifier 10.1109/TIT.2006.885481
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 II and III, we prove Theorem 1.1 in Section IV. 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 \( \varepsilon \), is analytic at any \( \varepsilon = \varepsilon_0 \geq 0 \), provided that the Markov transition probabilities are all positive.

In Section V, 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 \( \varepsilon \) 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 VI, 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 may fail to be analytic. In Section VII, 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 VIII, 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 \( \varepsilon \). 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].

II. ITERATION ON THE SIMPLEX

Let \( W \) be the simplex, comprising the vectors

\[
\left\{ w = (w_1, w_2, \ldots, w_B) \in \mathbb{R}^B : w_i \geq 0, \sum_i w_i = 1 \right\}
\]

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

\[
\left\{ w = (w_1, w_2, \ldots, w_B) \in \mathbb{C}^B : \sum_i w_i = 1 \right\}
\]

and let \( W^C_i \) denote the complex version of \( W_\alpha \), i.e., \( W^C \) consists of all \( w \in W^C \) with \( w_i = 0 \) for \( \Phi(i) \neq \alpha \). For \( \alpha \in A \), let \( \Delta_\alpha \) denote the \( B \times B \) matrix such that \( \Delta_\alpha(i,j) = \Delta(i,j) \) for \( j \) with \( \Phi(j) = \alpha \), and \( \Delta_\alpha(i,j) = 0 \) otherwise. For \( \alpha \in A \), define the scalar-valued and vector-valued functions \( r_\alpha \) and \( f_\alpha \) on \( W \) by

\[
r_\alpha(w) = w \Delta_\alpha 1
\]

and

\[
f_\alpha(w) = w \Delta_\alpha / r_\alpha(w).
\]

Note that \( f_\alpha \) defines the action of the matrix \( \Delta_\alpha \) on the simplex \( W \). For any fixed \( n \) and \( z^0 \), define

\[
x_i = x_i(z^0_n) = p(y_i = \cdot | z_{i-1}, \ldots, y_{i-n})
\]

(2.1)

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

\[
x_{i+1} = f_{z_{i+1}}(x_i)
\]

(2.2)

starting with

\[
x_{-n+1} = p(y_{-n+1} = \cdot).
\]

(2.3)

We remark that Blackwell showed that

\[
H(Z) = - \int \sum_a r_a(w) \log r_a(w) dQ(w)
\]

(2.4)

where \( Q \), known as Blackwell’s measure, is the limiting probability distribution, as \( n \to \infty \), of \( \{x_i\} \) 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 \( d_E \) on \( S \) is defined as usual, namely, for \( u, v \in S \)

\[
u = (u_1, u_2, \ldots, u_B), v = (v_1, v_2, \ldots, v_B) \in S
\]

we have

\[
d_E(u, v) = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + \cdots + (u_B - v_B)^2}.
\]

The Hilbert metric [29] \( d_B \) on \( S \) is defined as follows:

\[
d_B(u, v) = \max_{\|u_j - v_j\| \leq 1} \log \frac{u_j / v_j}{u_j / v_j}.
\]

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

**Proposition 2.1:** \( d_E \) and \( d_B \) are equivalent (denoted by \( d_E \sim d_B \)) 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 \( u, v \in S \)

\[
C_1 d_B(u, v) < d_E(u, v) < C_2 d_B(u, v).
\]

**Proposition 2.2:** Assume that at \( \varepsilon_0, \Delta \) satisfies Conditions 1 and 2 of Theorem 1.1. Then for sufficiently large \( n \) and all choices of \( a_1, \ldots, a_n \in \{1, 2, \ldots, A\} \) and \( b \in \{1, 2, \ldots, A\} \) (here \( a_1, \ldots, a_n \) and \( b \) denote hidden Markov symbols), the
mapping $f_{a_n} \circ f_{a_{n-1}} \circ \cdots \circ f_{a_1}$ is a contraction mapping under the Euclidean metric on $W_b$.

Proof: $\hat{W}_b := \hat{f}_b(W)$ is a compact subset of the interior of some subsimplex $W^d_b$ 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 $f_{a_j}$ 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 $u, v \in \hat{W}_b$,

$$d_B(u, v) < \rho d_B(u, v).$$

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

$$d_B(f_{a_n} \circ f_{a_{n-1}} \circ \cdots \circ f_{a_2}(u), f_{a_n} \circ f_{a_{n-1}} \circ \cdots \circ f_{a_2}(v)) < \rho^{n-1} d_B(u, v).$$

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

$$d_E(f_{a_n} \circ f_{a_{n-1}} \circ \cdots \circ f_{a_2}(u), f_{a_n} \circ f_{a_{n-1}} \circ \cdots \circ f_{a_2}(v)) < C \rho^{n-1} d_E(u, v).$$

Let $L$ be a universal Lipschitz constant for any $f_{a_i} : W_b \to W^d_b$ 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 $f_{a_n} \circ \cdots \circ f_{a_1}$ 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 $f_{a_n} \circ \cdots \circ f_{a_1}$ is a Euclidean contraction on the whole simplex $W$.

III. 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

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

Denote the states of $\Delta$ by $\{1, 2, 3\}$ and let $\Phi(1) = \Phi(2) = 0, \Phi(3) = 1$. Each entry of $\Delta$ is a real analytic function of $\varepsilon$ at any given point $\varepsilon = \varepsilon_0$. For $\varepsilon_0 > 0$ and sufficiently small, $\Delta$ 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 $\varepsilon_0$, the entropy rate of the hidden Markov chain defined by $\Delta(\varepsilon)$ and $\Phi$ is real analytic as a function of $\varepsilon$ at $\varepsilon_0$.

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 $\Delta$ by its entries themselves; if $\Delta$ 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 $\Delta$ 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 throughout the paper.

IV. PROOF OF THEOREM 1.1

Notation: We rewrite $\Delta, Z, f_0(x), p(z_0|z_{-1})$ with parameter vector $\varepsilon^k$ as $\Delta^\varepsilon, Z^\varepsilon, f_0^\varepsilon(x), \text{ and } p^\varepsilon(z_0|z_{-1})$, respectively. We use the notation $\hat{W}_a$ to mean $\hat{f}_a^0(W)$. Let $\Omega_C = \Omega_C(r)$ denote the set of points of distance at most $r$ from $\varepsilon^0$ in the complex parameter space $C^n$. Let $N_{0} = N_{0}(R)$ denote the set of all points in $W_0^C$ of distance at most $R$ from $\hat{W}_b$.

We first prove that for some $r > 0$, $\log p^\varepsilon(z_0|z_{-1})$ can be extended to a complex analytic function of $\varepsilon^k \in \Omega_C(r)$ and for two symbol sequences $z_{-n}^0$ and $z_{-n}^1$

$$[\log p^\varepsilon(z_0|z_{-1}^0) - \log p^\varepsilon(z_0|z_{-1}^1)]$$

decays exponentially fast in $n$, when $n \leq n_1, n_2$ and $z_{-n}^0 = z_{-n}^1$, uniformly in $\varepsilon^k \in \Omega_C(r)$.

Note that for each $a_1, b, f_0^0(w)$ is a rational function of the entries of $\Delta^\varepsilon$ and $w \in W^d_b$. So, by viewing the real vector variables $\varepsilon$ and $w$ as complex vector variables, we can naturally extend $f_0^\varepsilon(w)$ to a complex-valued function of complex vector variables $\varepsilon$ and $w$. Since $\Delta$ satisfies Conditions 1 and 2 at $\varepsilon_0$, for sufficiently small $r$ and $R$, the denominator of $f_0^\varepsilon(w)$ is nonzero for $\varepsilon$ in $\Omega_C(r)$ and $w$ in $N_{0}(R)$. Thus, $f_0^\varepsilon(w)$ is a complex analytic function of $(\varepsilon, w)$ in the neighborhood $\Omega_C(r) \times N_{0}(R)$.

Assuming Conditions 1 and 2, we claim that $\Delta$ has an isolated (in modulus) maximum eigenvalue 1 at $\varepsilon_0$. 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 \( \Delta \) 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 \( \overline{p}(y = \cdot) \) (the eigenvector corresponding to the maximum eigenvalue 1) is a rational function of the entries of \( \Delta \), since it is a solution of the equation \( v \Delta v = v \). So, in the same way as for \( f_0^\nu(x) \), we can naturally extend \( \overline{p}(y = \cdot) \) to a complex analytic function \( \overline{p}(y = \cdot) \) on \( \Omega_C \).

Extending (2.1) for each \( i \), we define
\[
\overline{x}_i^\nu = \overline{x}_i^\nu(\overline{z}_{-n}^\nu) = \overline{p}(y_i = \overline{z}_{-n}^\nu) \tag{4.5}
\]
by iterating the following complexified random dynamical system (extending (2.2) and (2.3)):
\[
x_{i+1}^\nu = f_{z_{i}^\nu}(x_i^\nu) \tag{4.6}
\]
starting with
\[
x_{-n-1}^\nu = \overline{p}(y_{-n-1} = \cdot). \tag{4.7}
\]

By Proposition 2.2, for sufficiently large \( n \), we can replace the set of mappings \( \{f_0^\nu\} \) with the set \( \{f_{z_1}^0 \circ f_{z_2}^0 \circ \cdots \circ f_{z_n}^0\} \) and then assume that each \( f_{z_0}^\nu \) is a Euclidean contraction on each \( W_k \) with contraction coefficient \( \rho < 1 \). Since \( W_k \) is compact and the definition of \( \rho \)-contraction by strict inequality, we can choose \( r \) and \( R \) sufficiently small such that
\[
f_{z_0}^\nu \text{ is a Euclidean } \rho \text{-contraction on each } N_k(R), \quad \varepsilon \in \Omega_C(r). \tag{4.8}
\]

Further, we claim that by choosing \( r \) still smaller, if necessary \( x_i^\nu \in \cup_k N_k(R), \) for all \( i, n \) and all choices of \( z_{-n}^\nu, \varepsilon \in \Omega_C(r) \),
\[
x_{-n}^\nu \quad \text{To see this, fixing } \rho \text{ and } R, \text{ choose } r \text{ so small that}
\]
\[
|f_{z_0}^\nu(x) - f_{z_0}^\nu(y)| \leq R(1 - \rho), x, y \in \cup_k \tilde{W}_k, \varepsilon \in \Omega_C(r), \tag{4.10}
\]
\[
|\overline{p}(\cdot) - \overline{p}_{z_0}(\cdot)| \leq R(1 - \rho), \varepsilon \in \Omega_C(r). \tag{4.11}
\]

Now consider the difference
\[
x_{i+1}^\nu - x_{i+1}^0 = f_{z_{i+1}}^\nu(x_i) - f_{z_{i+1}}^0(x_i) = f_{z_{i+1}}^\nu(x_i) - f_{z_{i+1}}^\nu(x_i) + f_{z_{i+1}}^\nu(x_i) - f_{z_{i+1}}^0(x_i). \tag{4.12}
\]

Then by (4.8), (4.10) and (4.11), and (4.12), for \( i > -n - 1 \), we have
\[
|\overline{x}_{i+1}^\nu - x_{i+1}^0| \leq \rho |\overline{x}_i^\nu - x_i^0| + R(1 - \rho). \tag{4.13}
\]

So
\[
|\overline{x}_{i+1}^\nu - x_{i+1}^0| \leq R
\]
and thus for all \( i \), we have \( x_{i+1}^\nu \in \cup_k N_k(R) \), yielding (4.9). Each \( x_i^\nu \) is the composition of analytic functions on \( \Omega_C(r) \) and so is complex analytic on \( \Omega_C(r) \).

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

Let
\[
\overline{x}_i^\nu = \overline{x}_i^\nu(z_{-n_1}^0) = \overline{p}(y_i = \overline{z}_{-n_1}^0), \quad \overline{x}_i^\nu = \overline{x}_i^\nu(z_{-n_2}^0) = \overline{p}(y_i = \overline{z}_{-n_2}^0). \tag{4.14}
\]

Then we have
\[
\overline{x}_{i+1}^\nu = f_{z_{i+1}}^\nu(x_i), \quad \overline{x}_{i+1}^\nu = f_{z_{i+1}}^\nu(x_i).
\]

From (4.8) and (4.9), if \( z_{-n_1}^0 \sim z_{-n_2}^0 \), then there exists a positive constant \( L \) independent of \( n_1 \) and \( n_2 \) such that
\[
|\overline{x}_0^\nu - \overline{x}_0^0| \leq L \rho^n. \tag{4.15}
\]

Naturally
\[
\overline{p}(z_0|z_{-n_1}^0) = \sum_{\{y_{-n-1} : (y_{-n-1}, y_0) = (z_{-n_1}^0, z_0)\}} \sum \Delta^\nu(y_{-1}, y_0) \overline{p}(y_{-1}|z_{-n_1}^0). \tag{4.16}
\]

Then, there is a positive constant \( L' \), independent of \( n_1, n_2 \), such that
\[
|\overline{p}(z_0|z_{-n_1}^0) - \overline{p}(z_0|z_{-n_2}^0)| \leq L' \rho^n. \tag{4.17}
\]

Since \( \Delta^\nu \) satisfies Conditions 1 and 2, \( \overline{p}(z_0|z_{-n_1}^0) \) is bounded away from 0, uniformly in \( \nu \in \Omega_C, n \) and choices of \( z_{-n}^0; \) thus, there is a positive constant \( L'' \), independent of \( n_1, n_2 \), such that
\[
|\log \overline{p}(z_0|z_{-n_1}^0) - \log \overline{p}(z_0|z_{-n_2}^0)| \leq L'' \rho^n. \tag{4.18}
\]

Since for each \( y \in \{1, \ldots, B\}, \overline{p}(y) \) is analytic, from
\[
\overline{p}(z) = \sum_{\nu(y) = z} \overline{p}(y),
\]
we deduce that \( \overline{p}(z) \) is analytic. Furthermore, since \( \overline{p}(z_0|z_{-n_1}^0) \) is analytic on \( \Omega_C \), we conclude \( \overline{p}(z_{-n_1}^0) \) is analytic on \( \Omega_C \).

Choose \( \sigma \) so that
\[
1 < \sigma < 1/\rho.
\]

If \( r \) and \( R \) are chosen sufficiently small, then
\[
\sum_{z_0} |\overline{p}(z_0|z_{-n_1}^0)| \leq \sigma, \varepsilon \in \Omega_C(r) \text{ and all sequences } z \tag{4.17}
\]
and
\[
\sum_{z_0} |\overline{p}(z_0)| \leq \sigma, \varepsilon \in \Omega_C(r). \tag{4.18}
\]
Then we have

\[
\sum_{z_{n-1}^0} |p^z(z_{n-1}^0)| \
= \sum_{z_{n-1}^0} |p^z(z_{n-1}^0) p^z(z_0 z_{n-1}^0)| \
\leq \sum_{z_{n-1}^0} |p^z(z_{n-1}^0)| \sum_{z_0} |p^z(z_0 z_{n-1}^0)| \
\leq \sigma \sum_{z_{n-1}^0} |p^\Phi(z_{n-1}^0)|,
\]

implying

\[
\sum_{z_{n-1}^0} |p^\Phi(z_{n-1}^0)| \leq \sigma^{n+2}. \tag{4.19}
\]

Let

\[
H_n^\Phi(Z) = -\sum_{z_{n-1}^0} p^\Phi(z_{n-1}^0) \log p^\Phi(z_0 z_{n-1}^0)
\]

and

\[
\rho_1 = \rho_\delta < 1
\]

then we have

\[
|H_{n+1}^\Phi(Z) - H_n^\Phi(Z)| \
= \left| \sum_{z_{n-1}^0} p^\Phi(z_{n-1}^0) \log p^\Phi(z_0 z_{n-1}^0) \
- \sum_{z_{n-1}^0} p^\Phi(z_{n-1}^0) \log p^\Phi(z_0 z_{n-1}^0) \right| \
= \left| \sum_{z_{n-1}^0} p^\Phi(z_{n-1}^0) \left( \log p^\Phi(z_0 z_{n-1}^0) - \log p^\Phi(z_0 z_{n-1}^0) \right) \right| \
\leq \sigma^2 L'' \rho_1^n;
\]

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

\[
|H_m^\Phi(Z) - H_n^\Phi(Z)| \leq \sigma^2 L'' (\rho_1^n + \cdots + \rho_1^{m-1}) \
\leq \frac{\sigma^2 L'' \rho_1^m}{1 - \rho_1}.
\]

This establishes the uniform convergence of \( H_n^\Phi(Z) \) to a limit \( H_\infty^\Phi(Z) \). 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 \( H_\infty^\Phi(Z) \) is analytic on \( \Omega_C \).

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

**Example 4.1:** Consider a binary-symmetric channel with crossover probability \( \varepsilon \). Let \( \{Y_n\} \) be the input Markov chain with the transition matrix

\[
\Pi = \begin{bmatrix}
\pi_{00} & \pi_{01} \\
\pi_{10} & \pi_{11}
\end{bmatrix}, \tag{4.20}
\]

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 independent and identically distributed (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, \( \Phi \) 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 \( \varepsilon \) and \( \pi_{ij} \)'s are positive, the entropy rate \( H(Z) \) is analytic as a function of \( \varepsilon \) and \( \pi_{ij} \)'s. This still holds when \( \varepsilon = 0 \) and the \( \pi_{ij} \)'s are positive, because in this case, we have

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

**V. Domain of Analyticity**

Suppose \( \Delta \) is analytically parameterized by a vector variable \( \varepsilon \), and Conditions 1 and 2 in Theorem 1.1 are satisfied at \( \varepsilon = \varepsilon_0 \). In principle, the proof of Theorem 1.1 determines a neighborhood \( \Omega_C(r) \) of \( \varepsilon_0 \) on which the entropy rate is analytic. Specifically, if one can find \( \rho, r, R \) such that all of the following hold, then the entropy rate is analytic on \( \Omega_C(r) \).

1. Find \( \rho \) such that each \( f^\Phi_{\delta} \) is a Euclidean \( \rho \)-contraction on each \( W_k \). Then choose positive \( r, R \) such that for all \( \varepsilon \in \Omega_C(r) \), each \( f^\Phi_{\delta} \) is a Euclidean \( \rho \)-contraction on each \( N_k(R) \) (see (4.8)).

2. Next find \( r \) smaller (if necessary) such that for all \( \varepsilon \in \Omega_C(r) \), the image of the stationary vector of \( \Delta^\Phi \), under any composition of the mappings \( \{f^\Phi_{\delta}\} \), stays within \( \cup_k N_k(R) \) (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 \( \rho, r, R \), but in Condition 1 above one may need to replace \( f^\Phi_{\delta} \)'s by all \( n \)-fold compositions of the \( f^\Phi_{\delta} \)'s, for some \( n \).

Recall from Example 4.1 the family of hidden Markov chains \( Z^\varepsilon \) determined by passing a binary Markov chain through a
binary-symmetric channel with crossover probability \( \varepsilon \). Recall that \( H(Z) \) is an analytic function of \( \varepsilon \) 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 \( \varepsilon \), is analytic on this neighborhood. Let \( u_n = p(y_n = 0|z_n^p) \) and \( v_n = p(y_n = 1|z_n^p) \). For \( z_{n+1} = 1 \) we have
\[
\begin{align*}
u_{n+1} &= \frac{\varepsilon(\pi_00u_n + \pi_{10}v_n)}{\varepsilon(\pi_00u_n + \pi_{10}v_n) + (1 - \varepsilon)(\pi_01u_n + \pi_{11}v_n)} \\
v_{n+1} &= \frac{\varepsilon(\pi_01u_n + \pi_{11}v_n)}{\varepsilon(\pi_00u_n + \pi_{10}v_n) + (1 - \varepsilon)(\pi_01u_n + \pi_{11}v_n)}.
\end{align*}
\]
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
\[
\begin{align*}
u_{n+1} &= \frac{1 - \varepsilon}{(1 - \varepsilon)(\pi_00u_n + \pi_{10}v_n) + \varepsilon(\pi_01u_n + \pi_{11}v_n)} \\
v_{n+1} &= \frac{1 - \varepsilon}{\varepsilon(\pi_00u_n + \pi_{10}v_n) + (1 - \varepsilon)(\pi_01u_n + \pi_{11}v_n)}.
\end{align*}
\]
Again, \( u_{n+1} \) is a function of \( u_n \); let \( g_0 \) denote this function. And for the conditional probability, we have
\[
p(z_n = 0|z_n^{n-1}) = ((1 - \varepsilon)\pi_00 + \varepsilon\pi_01)u_n + ((1 - \varepsilon)\pi_{10} + \varepsilon\pi_{11})v_n.
\]
Since \( u_n + v_n = 1 \), \( p(z_n = 0|z_n^{n-1}) \) is a function of \( u_n \); let \( r_0 \) denote this function. And
\[
p(z_n = 1|z_n^{n-1}) = (\varepsilon\pi_00 + (1 - \varepsilon)\pi_01)u_n + (\varepsilon\pi_{10} + (1 - \varepsilon)\pi_{11})v_n.
\]
Again, \( p(z_n = 1|z_n^{n-1}) \) 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 \( \varepsilon \). The stationary vector \( \pi_0, \pi_1 \) of \( Y \), which does not depend on \( \varepsilon \), is equal to \( (\pi_{10}/(\pi_{10} + \pi_{01}), \pi_{01}/(\pi_{10} + \pi_{01})) \).

We shall choose \( \rho \) with \( 0 < \rho < 1, r > 0 \) and \( R > 0 \) such that for all \( \varepsilon \) with \( |\varepsilon| < r \)
\[
1. g_0 \text{ and } g_1 \text{ are } \rho \text{-contraction mappings on } R \text{-neighborhoods of } 0 \text{ and } 1 \text{ in the complex plane;}
2. the set of all \( \{g_{n_0} \circ g_{n_0-1} \circ \cdots \circ g_{n_0}(\pi_0)\} \) are within the \( R \)-neighborhoods of 0 and 1;
3. \( |r_0(u)| + |r_1(u)| < 1/\rho \) 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, Conditions 1, 2, and 3 translate to (here \( \rho < 1 \):)
1. \( |g_0(u)| < \rho, |g_1(u)| < \rho \) on \( (|\varepsilon| < r \text{ and } |u| < R) \) and \( (|\varepsilon| < r \text{ and } |1 - u| < R) \);
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 \( Y \) does not depend on \( \varepsilon \));
3. \( |r_0(u)| + |r_1(u)| < 1/\rho \) on \( (|\varepsilon| < r \text{ and } |u| < R) \) and \( (|\varepsilon| < r \text{ and } |1 - u| < R) \).

A straightforward computation shows that the conditions shown at the bottom of the page guarantee Conditions 1, 2, 3.

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

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

VI. RELAXED CONDITIONS

We do not know a complete set of necessary and sufficient conditions on \( \Delta \) and \( \Phi \) 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 VII, 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 $\Delta$ has a simple maximum eigenvalue $1$; this implies that $\Delta$ 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 $\{f_a : a \in \{1,2,\ldots,A\}\}$ is eventually contracting at $w \in W_b$ if there exists $n$ such that for any $a_0, a_1, \ldots, a_n \in \{1,2,\ldots,A\}$, $\|\langle f_{a_n} \circ f_{a_{n-1}} \circ \cdots \circ f_{a_0}\rangle \|$ is strictly less than $1$. We say that $\{f_a : a \in \{1,2,\ldots,A\}\}$ 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 $\{f_a : a \in \{1,2,\ldots,A\}\}$ 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

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

**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 $L_a$, $r_a(w) > 0$;
3. for every $b$, $\{f_a : a \in \{1,2,\ldots,A\}\}$ 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 $X$ denote the right infinite shift space $\{a^\infty_0 : a_i \in \{1,2,\ldots,A\}\}$. Let $L^\infty_b$ be the set of all points in $W_b$ of distance at most $\delta$ from $L \cap W_b$, and let $L^\infty = \cup_b L^\infty_b$. Choose $\delta$ so small that

- for every $a \in \{1,2,\ldots,A\}$ and $w \in L^\infty$, $r_a(w) > 0$ and
- for every $b$, $\{f_a : a \in \{1,2,\ldots,A\}\}$ is eventually contracting at all $w$ in the convex hull $K^\infty_b$ of $L^\infty_b$.

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

For any $c^\infty_0 \in X$, there exists $n$ such that

$$\{(f_{a_n} \circ f_{a_{n-1}} \circ \cdots \circ f_{a_0}) (\vec{s}) \in I^\infty.$$ 

Let $X^\infty_{c_0}$ denote the cylinder set

$$\{c^\infty_0 : a_0 = a_{i_0}, a_{i_1} = c_{i_1}, \ldots, a_n = c_n\}.$$

Since $f_a(I^\infty) \subseteq I^\infty$, we conclude that for any $a^\infty_0 \in X^\infty_{c_0}$ and all

$$m \geq n, \{(f_{a_m} \circ f_{a_{m-1}} \circ \cdots \circ f_{a_0}) (\vec{s}) \in I^\infty.$$ 

By the compactness of $X$, we can find finitely many such cylinder sets to cover $X$. Consequently, we can find $n$ such that for any $a_0^\infty \in X$ and any $m \geq n$, we have

$$\{(f_{a_m} \circ f_{a_{m-1}} \circ \cdots \circ f_{a_0}) (\vec{s}) \in I^\infty.$$ 

We can now apply the proof of Theorem 1.1—namely, we can use the contraction (along any symbolic sequence $a^\infty_{0_m}$) to extend $H_n(Z) = H(Z_0[Z^\infty_m])$ from real to complex and prove the uniform convergence of $H_n(Z)$ to $H(Z)$ in complex parameter space.
Remark 6.2:
1) If \( \Delta \) has a strictly positive column (or more generally, there is a \( j \) such that for all \( i \), there exists \( \eta_i \) such that \( \Delta_{ij} > 0 \)), then Condition 1 of Theorem 6.1 holds by Perron–Frobenius theory.
2) If for each symbol \( a_i \), \( \Delta_{a_i} \) is row allowable (i.e., no row is all zero), then \( \gamma_{l}(a_i) > 0 \) 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 \( \bigcup_{b} f_b(W) \).

Using the equivalence of the Euclidean metric and the Hilbert metric, Proposition 2.2 shows that for every \( b \), \( \{f_{a_i} : a_i \in \{1, 2, \ldots, A\}\} \) 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 \( \Delta \) is strictly positive and each \( \Delta_{a_i} \) 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 \( a_i \), \( f_{a_i} \) 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

\[
\Delta = \begin{bmatrix}
\alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 \\
\alpha_2 & \alpha_3 & \alpha_4 & 1 \\
\alpha_3 & \alpha_4 & 1 & 1 \\
\alpha_4 & 1 & 1 & 1
\end{bmatrix}
\]

with \( \Phi(1) = \Phi(2) = 0 \) and \( \Phi(3) = \Phi(4) = 1 \). We assume that some column of \( \Delta \) is strictly positive and both \( \Delta_0 \) and \( \Delta_1 \) 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 \( y \):

\[
\frac{df_0}{dy} = \frac{a_{11}a_{22} - a_{12}a_{21}}{(a_{11} + a_{12} - a_{21} - a_{22})y + a_{21} + a_{22})^2}
\]

\[
\frac{df_1}{dy} = \frac{a_{13}a_{44} - a_{14}a_{43}}{(a_{13} + a_{14} - a_{23} - a_{24})y + a_{23} + a_{24})^2}
\]

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

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

Or choose the \( \alpha_{ij} \)'s such that

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

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 does not 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 \( \gamma = (s_2, s_4) \) is the stationary vector of \( \Delta \) corresponding to the simple eigenvalue 1. Let \( u_0 = (0, 1, 0, 0) \) and \( u_1 = (0, 0, 0, 1) \). One checks that for \( n \geq 0 \), \( f_{a_0} \circ f_{a_1} \circ \cdots \circ f_{a_0}(\gamma) = w_{i_n} \). Therefore, \( L \) consists of \( \{u_0, u_1\} \). Using the expressions above, we see that

\[
\begin{align*}
\lambda_0 & = \alpha_1 / \alpha_2 < 1 \\
\lambda_1 & = \beta_1 / \beta_2 < 1
\end{align*}
\]

and so Condition 3 holds. Thus, the entropy rate \( H(Z) \) is analytic at \( \Delta \).

VII. 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 \( a_3 \) be an unambiguous symbol

\[
H(Z) = \sum_{a_3 \neq a_1} p(a_i, a_{i-1}, \ldots, a_2, a_1) 
\]

\( \times H(z \mid a_{i-1}, a_{i-2}, \ldots, a_2, a_1) \). (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

\[
H(Z^\varepsilon) = p(0^\varepsilon)H(0 \mid 0) + p(1^\varepsilon)(0 \mid 1)H(1 \mid 0) + \cdots
\]

\( \cdot \) (7.22)

where \( 1^{(n)} \) denotes the sequence of \( n \) 1’s and

\[
H(1^{(n)}) = -p(1^{(n)} | 0) \log p(1^{(n)} | 0) + \cdots
\]

\( \cdot \) (7.22)

Example 7.3: Fix \( \alpha, \beta, \ldots, h > 0 \) and for \( \varepsilon \geq 0 \) let

\[
\Delta(\varepsilon) = \begin{bmatrix}
\varepsilon & \alpha - \varepsilon & b \\
g & c & d \\
h & e & f
\end{bmatrix}
\]

Assume \( \alpha, \beta, \ldots, h > 0 \) are chosen such that \( \Delta(\varepsilon) \) is stochastic. The symbols of the Markov chain are the matrix indices \( \{1, 2, 3\} \). Let \( Z^\varepsilon \) 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
\[
\vec{p}(0) = \pi_1(\varepsilon), \pi(0) = \pi_1(\varepsilon) \varepsilon
\]
and for \( n \geq 1 \)
\[
\vec{p}(1^n)(0) = \pi_1(\varepsilon)(a - \varepsilon, b) \begin{bmatrix} c \\ \varepsilon \\ d \\ f \end{bmatrix}^{n-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix},
\]
Since \( \Delta(\varepsilon) \) is irreducible, \( \pi(\varepsilon) \) is analytic in \( \varepsilon \) and positive. Now
\[
\vec{p}(0)H^\varepsilon(z|0) = -\vec{p}(0)(0)\log \vec{p}(0)(0) - \vec{p}(10)\log \vec{p}(10),
\]
(7.23)
The first term in (7.23) is
\[
-\vec{p}(0)(0)\log \vec{p}(0)(0) = -\pi_1(\varepsilon) \varepsilon \log \varepsilon
\]
which is not analytic (or even differentiable at \( \varepsilon = 0 \)). The second term in (7.23) is
\[
-\vec{p}(10)\log \vec{p}(10) = -\pi_1(\varepsilon)(a - \varepsilon, b) \log (\pi_1(\varepsilon)(a - \varepsilon, b))
\]
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
\[
\begin{bmatrix} c & d \\ \varepsilon & f \end{bmatrix}
\]
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 \( \varepsilon \), 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 \( a, b, \ldots, g > 0 \) and consider the stochastic matrix
\[
\Delta(\varepsilon) = \begin{bmatrix} \varepsilon & a & b \\ f - \varepsilon & \varepsilon & c \\ g & 0 & d \end{bmatrix}.
\]
The symbols of the Markov chain are the matrix indices \( \{1, 2, 3\} \). Again let \( Z^\varepsilon \) 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 \( c = d \), and not analytic when \( c \neq d \). Note that
\[
\vec{p}(0) = \pi_1(\varepsilon)
\]
and for \( n \geq 1 \)
\[
\vec{p}(1^n)(0) = \pi_1(\varepsilon)(a, b) \begin{bmatrix} c & \varepsilon \\ \varepsilon & 0 \\ d & \varepsilon \end{bmatrix}^{n-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix}.
\]
When \( c \neq d \), we assume \( c > d \), then
\[
\begin{bmatrix} c & \varepsilon \\ \varepsilon & 0 \\ d & \varepsilon \end{bmatrix} = \begin{bmatrix} c^n & \varepsilon^n & (d/c)^n \\ 0 & 1 & 0 \end{bmatrix}.
\]
Since \( \Delta(\varepsilon) \) is irreducible, \( \pi(\varepsilon) \) is analytic in \( \varepsilon \) and positive. Simple computation leads to
\[
\vec{p}(1^n)(0) = \left( ac^2 + \frac{1}{1 - d/c} \left( a + \varepsilon c \right) + bd(\varepsilon)^2 \right) \left(\frac{1}{1 - d/c} + bd(\varepsilon)^2 \right) / \left( a + \varepsilon c \right),
\]
(7.24)
and
\[
\vec{p}(0)(1^n)(0) = \left( f - \varepsilon \right) a c + g \left( \frac{1}{1 - d/c} \left( a + \varepsilon c \right) + bd(\varepsilon)^2 \right) / \left( a + \varepsilon c \right)
\]
which has spectral radius \( < 1 \), the term \( \vec{p}(1^n)(0)H^\varepsilon(z|1^n)(0) \) 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 \( \varepsilon \).

When \( c = d \), we have
\[
\vec{p}(1^n)(0) = \left( a c^{n+1} + \varepsilon c^{n+1} + \varepsilon c^{n+1} \right) / \left( a c^n + \varepsilon c^{n+1} + \varepsilon c^{n+1} \right),
\]
and
\[
\vec{p}(0)(1^n)(0) = \left( (f - \varepsilon) c^n + g \varepsilon c^{n+1} + g c^{n+1} \right) / \left( a c^n + \varepsilon c^{n+1} + \varepsilon c^{n+1} \right)
\]
For any \( n \), consider a small neighborhood \( N_n \) of \( -(a + b)c/an \) in \( \mathbb{C} \) such that \( \Delta(a + b)c/an \in N_n \) only holds for \( j = n \). When \( \varepsilon \rightarrow -(a + b)c/an \), the complexified term \( \vec{p}^\varepsilon(1^n)(0)H^\varepsilon(z|1^n)(0) \rightarrow \infty \). Meanwhile, the sum of all the other terms can be analytically extended to \( N_n \) (from any path \( I \) to a positive \( \varepsilon \) to \( -(a + b)c/an \) with \( -(a + b)c/an \in N_j \) 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 \( A(\varepsilon) \) be an analytic parameterization of complex matrices. Let \( \lambda \) be the spectral radius of \( A(\varepsilon_0) \). Then for any \( \eta > 0 \), there exists a complex neighborhood \( \Omega \) of \( \varepsilon_0 \) and positive constant \( C \) such that for all \( \varepsilon \in \Omega \) and all \( i, j, k \)

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

**Proof:** Following [29], we consider

\[
(I - zA)^{-1} = I + zA + z^2A^2 + \cdots.
\]

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

\[
(p_0 + p_1 z + \cdots + p_m z^m) \prod_{j=1}^{\infty} \sum_{i=0}^{\infty} \lambda_j^{i_1} \lambda_j^{i_2} \cdots \lambda_j^{i_n} z^{k_i}.
\]

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

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

where \( \lambda_1, \eta_1 \) 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 \( \text{diag}(\eta_1, C) \) by a continuous rescaling; then connect \( \eta_1/C \) to \( D \) by the path \( \eta_1/C + (1 - t)D \) (the path \( \text{diag}(\eta_1/C + (1 - t)D) \) stays within \( S(n) \) since the matrices along this path are upper triangular with all diagonal entries, except \( \eta_1, \) of modulus less than \( |\eta_1| \)). \( \square \)

For a complex analytic function \( f(z_1, z_2, \ldots, z_n) \), let \( V(f) \) denote the “hypersurface” defined by \( f \), namely \( V(f) = \{(z_1, z_2, \ldots, z_n) \in \mathbb{C}^n : f(z_1, z_2, \ldots, z_n) = 0\} \).

Now let \( \Omega_c \) 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:** \( \Omega_c \setminus V(f) \) is connected.

**Proof:** For simplicity, we first assume \( \Omega_c \) 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 points \( P, Q \in \Omega_c \setminus V(f) \), consider the “complex line”

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

\( L_{PQ} \cap V(f) \cap \Omega_c \) consists of only isolated points (A nonconstant one variable complex analytic function must have isolated zeros in the complex plane [30]). It then follows that for the compact real line segment

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

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

In the general case, \( \Omega_c \) is a connected open set in \( \mathbb{C}^n \). Let \( f \) be an arc in \( \Omega_c \) 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 \( P_j \in \Omega_c \setminus V(f) \). Applying the same argument as above to every ball \( B_{r_j}(z_j) \), we see that \( P \) is connected to \( Q \in \Omega_c \setminus V(f) \) through the points \( P_j \)’s. Thus, we prove the lemma.

**Theorem 7.8:** Let \( \Delta \) be an irreducible stochastic \( d \times d \) matrix. Write \( \Delta \) in the form

\[
\Delta = \begin{bmatrix} a & r \\ c & B \end{bmatrix}
\]

where \( a \) is a scalar and \( B \) is a \((d-1) \times (d-1)\) matrix. Let \( \Phi \) be the function defined by \( \Phi(1) = 0 \), \( \Phi(2) = \cdots = \Phi(d) = 1 \). Then for any parametrization \( \Delta(\varepsilon) \) such that \( \Delta(\varepsilon_0) = \Delta \), letting \( Z^\varepsilon \) denote the hidden Markov chain defined by \( \Delta(\varepsilon) \) and \( \Phi, H(Z^\varepsilon) \) is analytic at \( \varepsilon_0 \) if and only if

1. \( a > 0 \), and \( rBc > 0 \) for \( j = 0, 1, \ldots; \)
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

\[
\Delta(\varepsilon) = \begin{bmatrix} a(\varepsilon) & r(\varepsilon) \\ c(\varepsilon) & B(\varepsilon) \end{bmatrix}
\]

where \( a(\varepsilon) \) is a scalar and \( B(\varepsilon) \) is a \((d-1) \times (d-1)\) matrix. Since \( \Delta(\varepsilon_0) \) is stochastic and irreducible, its spectral radius is 1, and 1 is a simple eigenvalue of \( \Delta \). Thus, if \( \Omega_c \) is sufficiently small, for all \( \varepsilon \in \Omega_c \), any fixed row \( \pi(\varepsilon) = (\pi_1(\varepsilon), \pi_2(\varepsilon), \ldots, \pi_d(\varepsilon)) \) of \( \text{Adj}(I - \Delta(\varepsilon)) \) is a left eigenvector of \( \Delta(\varepsilon) \) associated with eigenvalue 1 and is an analytic function of \( \varepsilon \). Normalizing, we can assume that \( \pi(\varepsilon_1) = 1, \pi(\varepsilon) \) is analytic in \( \varepsilon \), and \( \pi_j(\varepsilon_0) > 0 \) for \( j = 1, 2, \ldots, d \).
The entries of \( r(z), B(z), \) and \( c(z) \) are real analytic in \( z \) and can be extended to complex analytic functions in a complex neighborhood \( \Omega_C \) of \( z_0 \). Thus, for all \( n \)

\[
\pi_1(z)r(z)B(z)^{n-1} \quad \text{and} \quad \pi_1(z)r(z)B(z)^{n-1}c(z)
\]
can be extended to complex analytic functions on \( \Omega_C \) (in fact, each of these functions is a polynomial in \( z \)).

Since \( B(z_1) \) is a proper submatrix of the irreducible stochastic matrix \( \Delta(z_0) \), 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 \( \Omega_C \) of \( z_0 \), all \( z \in \Omega_C \), and all \( n \)

\[
|B_{ij}^n(z_0)| < C_1(\lambda^*)^n.
\]

Since \( \pi_1(z), r(z) \) and \( c(z) \) are continuous in \( z \), there is a constant \( C_2 > 0 \) such that for all \( z \in \Omega_C \) and all \( n \)

\[
|\pi_1(z)r(z)B(z)^{n+1}| < C_2(\lambda^*)^n.
\]

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

**Lemma 7.9:** Let

\[
a(z,n) \equiv \frac{\pi_1(z)r(z)B(z)^{n+1}}{\pi_1(z)r(z)B(z)^{n+1}1} \quad \text{and} \quad b(z,n) \equiv \frac{\pi_1(z)r(z)B(z)^{n+1}c(z)}{\pi_1(z)r(z)B(z)^{n+1}},
\]

For a sufficiently small neighborhood \( \Omega_C \) of \( z_0 \), both \( |a(z,n)| \) and \( |b(z,n)| \) are bounded from above and from zero, uniformly in \( z \in \Omega_C \) and \( n \).

Define

\[
H^z_n = -a(z,n) \log a(z,n) - b(z,n) \log b(z,n)
\]

where \( a(z,n) \) and \( b(z,n) \) are as in Lemma 7.9. Choosing \( \Omega_C \) to be a smaller neighborhood of \( z_0 \), if necessary, \( a(z,n) \) and \( b(z,n) \) are constrained to lie in a closed disk not containing 0. Thus, for all \( n \), \( H^z_n \) is an analytic function of \( z \) with \( |H^z_n| \) bounded uniformly in \( z \in \Omega_C \) and \( n \). Since \( \pi_1(z)r(z)B(z)^{n+1}1 \) is analytic on \( \Omega_C \) and exponentially decaying (by (7.26)), the infinite series

\[
H^z(Z) = \sum \pi_1(z)H^z_n + \pi_1(z)r(z)1H^1_n + \cdots
\]

\[
+ \pi_1(z)r(z)B(z)^{n+1}1H^1_n + \cdots
\]

(7.27)

converges uniformly on \( \Omega_C \) and thus defines an analytic function on \( \Omega_C \).

Note that for \( z \geq 0 \)

\[
\pi_1(\Omega) = \pi_1(z)r(z)B(z)^{n+1}1
\]

and

\[
\pi_1(\Omega) = \pi_1(z)r(z)B(z)^{n+1}c(z)
\]

By (7.28), (7.29), and (7.22), \( H^z(Z) \) agrees with the entropy rate when \( \Delta^+(z) \geq 0 \), 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 \( \Delta \) 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 \( f_2 \)-orbit of \( z \) converge to a point \( p_1 \). It follows that \( L \) is the union of \( W_0 \) and \( W_1 \); the \( f_1 \)-orbit of \( W_0 \) and \( W_1 \). The assumptions in Theorem 7.8. imply that \( r_n > 0 \) 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:**

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

Suppose \( \Delta \) does not satisfy Condition 2, however, \( H(Z) \) is analytic at \( \Delta \) with respect to the natural parameterization. In other words, suppose there exists a complex neighborhood \( N_\Delta \) of \( \Delta \) (here \( N_\Delta \) 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 \( N_\Delta \), while the corresponding \( B \) does not have isolated (in modulus) maximum eigenvalue.

We first claim there exists \( \Delta \in N_\Delta \) with \( \hat{B} \neq 1 \). Here \( \hat{r} \) and \( \hat{B} \) correspond to \( \Delta \) and \( \hat{B} \) has distinct eigenvalues (in modulus). Indeed, we can first (for simplicity) perturb \( \Delta \) to \( \Delta' \) such that the corresponding \( \hat{B} \) has distinct eigenvalues in modulus. Then

\[
\hat{B} = \hat{U} \hat{W} \hat{B} \hat{W}^{-1} \quad \text{for} \quad \hat{U} \hat{W}^{-1} = \hat{U} \quad \text{and} \quad \hat{U}^{-1} \hat{W} \quad \text{for} \quad \hat{U} \hat{W}^{-1} = \hat{U} \quad \text{and} \quad \hat{U}^{-1} \hat{W}.
\]

Further, consider a perturbation of \( B \) from

\[
\hat{B} = \hat{U} \hat{B} \hat{W} \hat{W}^{-1} \quad \text{to} \quad \hat{B} = \hat{U} \hat{W} \hat{B} \hat{W}^{-1} \hat{V}^{-1}
\]

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 \( \hat{V}_1 \hat{W}_1 \hat{V}_1^{-1} \neq 0 \), \( \hat{V}_2 \hat{W}_2 \hat{V}_2^{-1} \neq 0 \), \( \hat{V}_2 \hat{W}_2 \hat{V}_2^{-1} \neq 0 \). Clearly, \( \hat{V}_1 \hat{W}_1 \hat{V}_1^{-1} \neq 0 \). Then by a further perturbation of \( r \) to \( \hat{r} \), we can simultaneously require that \( \hat{r}_1 \hat{V}_1 \hat{V}_1^{-1} \neq 0 \), \( \hat{r}_2 \hat{V}_2 \hat{V}_2^{-1} \neq 0 \). This is possible, and it follows that \( \hat{B} = \hat{B} \hat{V} \hat{W}^{-1} \hat{V}^{-1} \hat{W} \). For any \( \theta \) and \( \eta > 0 \), it can be checked that

\[
\sum_{k=0}^{\infty} \{ |z^k| < \eta \} = C \setminus \{0\}.
\]
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 at the bottom of the page, with \( [\tilde{\lambda}_2] \neq [\tilde{\lambda}_1] \) and \( [\tilde{\lambda}_2] \) strictly greater than \( [\tilde{\lambda}_1] \) for \( j \geq 3 \). Thus we prove the claim.

We now pick a positive matrix \( \Delta \in N_\Delta \) with corresponding \( \tilde{\varphi} \) and \( \tilde{B} \). Then pick \( \Delta \in N_\Delta \) with corresponding \( \tilde{\varphi} \) and \( \tilde{B} \) (with distinct eigenvalues in modulus) such that \( \tilde{\varphi} \tilde{B} \tilde{1} = 0 \) for some \( k_1 \), and we can further require that \( \tilde{B} \tilde{v}_1 \tilde{w}_1 \neq 0, \tilde{B} \tilde{v}_1 \tilde{c} \neq 0 \) (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 \( \tilde{B} \) to \( \tilde{B} \); we then connect \( \tilde{\varphi} \) and \( \tilde{\varphi} \) using an arc \( I_2 \) in \( C^{d-1} \). According to Lemma 7.7, we can choose the arc \( I = (I_1, I_2) \) to avoid the hypersurface

\[
V((rv_1w_11)rv_1w_1c) \subset C^{(d-1)^2} \times C^{d-1};
\]

in other words, we can assume that along the path \( I, rv_1w_11 \neq 0 \) 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 \( B \)).

Then we claim that there is a neighborhood \( N_I \) of \( I \) such that \( V_k \cap N_I \neq \phi \) and \( W_k \cap N_I \neq \phi \) hold only for many \( k \), where \( V_k = \{ (B, r) : rBk \neq 0 \} \) and \( W_k = \{ (B, r) : rBk \neq 0 \} \). Indeed, for any \( \Delta \in I \) with corresponding \( B \in S(d-1) \), by the Jordan form we have

\[
rBk = rv_1w_11k \Lambda + o(\lambda_1^k)
\]

where \( \Lambda \) is the isolated maximum eigenvalue and \( v_1, w_1 \) are appropriately scaled right and left eigenvectors of \( B \), respectively. Since \( rv_1w_11 \neq 0 \) on \( I \), there exists a complex connected neighborhood \( N_I \) of \( I \) such that \( rv_1w_11 \neq 0 \) on \( N_I \) and \( rv_1w_11k \Lambda \) dominates uniformly on \( N_I \) (see Lemma 7.5).

Consequently, \( [rBk] > 0 \) on \( N_I \) for large enough \( k \). In other words, \( V_k \cap N_I \neq \phi \) 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 \( W_k \cap N_I \neq \phi \) 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 \( C^{d-1} \times S(d-1) \) of \( I \) such that \( V_k \cap N_I \neq \phi \) holds only for \( k = k_1 \) and \( W_k \cap N_I \neq \phi \) for all \( k \). Consider \( \Delta \) with corresponding \( \tilde{B} \); let \( F_j = F_j(\tilde{B}) = \{ r : r\tilde{B}k = 0 \} \), which is a hyperplane orthogonal to the vector \( \tilde{B}k \) in \( C^{d-1} \). Similarly, we define \( G_i = G_i(\tilde{B}) = \{ r : r\tilde{B}c = 0 \} \). Recall that \( \tilde{B} = \tilde{U} \text{diag}(\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_{d-1}) \tilde{U}^{-1} \); we can require that \( \tilde{U}^{-1}1 \) has no zero coordinates by a small perturbation of \( \tilde{U} \) if necessary. We then show that \( F_j \)'s and \( G_i \)'s define different hyperplanes in \( C^{d-1} \). Indeed, suppose \( F_i = F_j \).

It follows that \( \tilde{U} \text{diag}(\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_{d-1}) \tilde{U}^{-1} \) is proportional to \( \tilde{U} \text{diag}(\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_{d-1}) \tilde{U}^{-1} \). It then follows that \( \tilde{U} \text{diag}(\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_{d-1}) \tilde{U}^{-1} \).

Since \( \tilde{\lambda}_2^{1/\tilde{\lambda}_1} \) is a perturbation of \( \tilde{\lambda}_1 \), it follows that for large enough \( k \), one can perturb \( \tilde{\lambda}_2 \) to satisfy the equation at the bottom of the page, with \( [\tilde{\lambda}_2] \neq [\tilde{\lambda}_1] \) and \( [\tilde{\lambda}_2] \) strictly greater than \( [\tilde{\lambda}_1] \) for \( j \geq 3 \). Thus we prove the claim.

In this section, we show that if \( \Delta \) is analytically parameterized by a real variable vector \( \varepsilon \), and at \( \varepsilon_0 \) \( \Delta \) satisfies Conditions 1 and 2 of Theorem 1.1, then the hidden Markov chain itself is a real analytic function of \( \varepsilon \) at \( \varepsilon_0 \) 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 $\{x_0^n : x_0 = z_0, \ldots, x_{-n} = z_{-n}\}$. The Borel sigma-algebra is the smallest sigma-algebra containing the cylinder sets. A Borel probability measure (BPM) $\nu$ 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 $\varepsilon$, consider the measure $\nu^\varepsilon$ on $\mathcal{X}$ defined by

$$
\nu^\varepsilon(\{x_0^n : x_0 = z_0, \ldots, x_{-n} = z_{-n}\}) = \nu^\varepsilon(\{z_0^\varepsilon_{-n}\}).
$$

(8.30)

Note that $H(\nu)$ can be rewritten as

$$
H^\varepsilon(\nu) = \int -\log \nu^\varepsilon(\{z_0^\varepsilon_{-n}\}) \, d\nu^\varepsilon.
$$

(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 $\xi$ and $\eta$ 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 $\nu$ acts as a bounded linear functional on $C(\mathcal{X})$, namely, $\nu(f) = \int f \, d\nu$. As such, the set of BPMs is a subset of the dual space $C(\mathcal{X})^*$, which is itself a Banach space; the norm of a BPM $\nu$ is defined as $\|\nu\| = \sup \{|f(\xi)| : f \in C(\mathcal{X}) \land \|f\|_{\infty} = 1\} \int f \, d\nu$. In fact, since $\mathcal{X}$ is compact, $C(\mathcal{X})^*$ is the linear span of the BPMs.

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

Let $\mathcal{X}$ be the set of binary left infinite sequences. Let $\nu_p$ denote the i.i.d. $(p, 1-p)$ measure, with $0 < p < 1$. We claim that, for fixed $p > q$ and $\varepsilon$ with $0 < \varepsilon < (p - q)/2$, by application of the law of large numbers assert that for $n$ sufficiently large

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

To see this, first say that a word $w = w_1 \ldots w_n$ is $(p, \varepsilon, n)$-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 $(p, \varepsilon, n)$-typical words. The law of large numbers asserts that for $n$ sufficiently large

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

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

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

(8.29)

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

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

For $f \in C(\mathcal{X})$, define

$$
\text{var}_n(f) = \sup \{|f(\xi) - f(\xi')| : \xi - \xi'\text{ for } i \leq n\}.
$$

We denote by $F^\theta$ the subset of $f \in C(\mathcal{X})$ such that

$$
\|f\|_\theta = \sup_{n \geq 0} (\theta^{-n} \text{var}_n(f)) < +\infty.
$$

$F^\theta$ is a Banach space with the norm $\|f\|_\theta = \max(\|f\|_{\infty}, \|f\|_\theta)$. Let $(F^\theta)^*$ denote the dual space (i.e., the set of bounded linear functionals) on $F^\theta$. For any $\nu \in (F^\theta)^*$, the norm of $\nu$ is naturally defined as $\|\nu\|_\theta = \sup_{f \in F^\theta : \|f\|_{\infty} = 1} \nu(f)$. Using complex functions instead of real functions, one defines $F^\theta_C$ and $(F^\theta)^*_C$ 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 $\Delta$ are analytically parameterized by a real variable $\varepsilon$. If at $\varepsilon = \varepsilon_0$, $\Delta$ satisfies Conditions 1 and 2 in Theorem 1.1, then the mapping $\varepsilon \mapsto \log \nu^\varepsilon(\{z_0^n\})$ is analytic at $\varepsilon_0$ from the real parameter space to $F^\theta$ (here $\rho$ is the contraction constant in the proof of Theorem 1.1). Moreover, the mapping $\varepsilon \mapsto \nu^\varepsilon$ is analytic at $\varepsilon_0$ from the real parameter space to $(F^\theta)^*_C$.

Proof: For complex $\varepsilon$, by (4.16), one shows that $\log \nu^\varepsilon(\{z_0^n\})$ can be defined on $\mathbb{C}$ as the uniform (in $\varepsilon$ and $z \in \mathcal{X}$) limit of $\log \nu^\varepsilon(\{z_0^n\})$ as $n \to \infty$, and $\log \nu^\varepsilon(\{z_0^n\})$ belongs to $F^\theta_C$. By (4.5), (4.6), (4.7), and (4.14) it follows that $\nu^\varepsilon(\{z_0^n\})$ is analytic on $\mathbb{C}$. As a result of (4.16), if $\Delta$ satisfies Conditions 1 and 2, for fixed $z \in \mathcal{X}$, $\log \nu^\varepsilon(\{z_0^n\})$ is the uniform limit of analytic functions and hence is analytic on $\mathbb{C}$ (see [32, Theorem 2.4.1]).

For a given sequence $z = z_0^n$, let $f(\varepsilon(z)) = \log \nu^\varepsilon(\{z_0^n\})$. Let $Df|_\varepsilon(\varepsilon(\cdot))$ denote the vector of partial derivatives of $f(\varepsilon(z))$ with respect to $\varepsilon$ at $\varepsilon = \varepsilon_0$. 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 $z \in \varepsilon_0$, for all $\varepsilon_0 \in \mathbb{C}$

$$
|Df|_\varepsilon(\varepsilon(z)) - Df|_\varepsilon(\varepsilon(z))| \leq C^*|\varepsilon|^n.(8.32)
$$

For a direction $\varepsilon_0$ in the parameter space, let $Df|_\varepsilon(\varepsilon_0) = \delta$ denote the directional derivative of $f(\varepsilon(z))$ at $\varepsilon = \varepsilon_0$ in direction $\varepsilon_0$. Let $Df|_\varepsilon(\varepsilon_0) \delta$ denote the function on $\mathcal{X}$, whose value on $z = z_0^n$ in $\mathcal{X}$ is given by $Df|_\varepsilon(\varepsilon_0)z$. By (8.32), $Df|_\varepsilon(\varepsilon_0)\delta$ belongs to $F^\theta_C$. 


Now, we must prove that the mapping $\mathcal{E} \mapsto \log f(\vec{z}; \mathcal{E})$ is analytic at $\vec{z}_0$, we show that $f(\vec{z}; \mathcal{E})$ is analytic at $\vec{z}_0$, as desired. In particular, we recover Theorem 5.2. Let $f(\vec{z}; \mathcal{E})$ denote the second directional derivative in direction $\mathcal{E}$ of $f(\vec{z}; \mathcal{E})$ at $\vec{z} = \vec{z}_0$. Again, applying the Cauchy integral formula in several variables, it follows that there exists a positive constant $C'$ such that for all $\vec{z}_0 \in \Omega_{\mathcal{E}}$, we have

$$\|\mathcal{E} \mapsto \log f(\vec{z}; \mathcal{E}) \| \leq C'\|\mathcal{E}\|^2.$$

(8.35)

and whenever $\vec{z} \in \mathcal{E}$,

$$\int_0^1 (1-t)[\mathcal{E} \mapsto \log f(\vec{z}; \mathcal{E})] \, dt \leq C'\|\mathcal{E}\|^2.$$

(8.36)

From the Taylor formula with integral remainder, we have

$$f(\vec{z}_0 + \mathcal{E}; \vec{z}) - f(\vec{z}_0; \mathcal{E}) = \int_0^1 (1-t)[\mathcal{E} \mapsto \log f(\vec{z}; \mathcal{E})] \, dt.$$

(8.37)

To prove (8.33), use (8.35) and (8.37). To prove (8.34), use (8.36) and (8.37). Therefore, $\mathcal{E} \mapsto \log f(\vec{z}; \mathcal{E})$ is analytic as a mapping from $\Omega_{\mathcal{E}}$ to $F^p$. Restricting the mapping $\mathcal{E} \mapsto \log f(\vec{z}; \mathcal{E})$ to the real parameter space, we conclude that it is real analytic (as a mapping into $F^p$). Using this and the theory of equilibrium states [26], the “Moreover” is proven in Appendix C.

Corollary 8.2: Suppose that at $\vec{z}_0$, $\Delta$ satisfies Conditions 1 and 2 in Theorem 1.1, and $\mathcal{E} \mapsto f(\vec{z}; \mathcal{E})$ is analytic at $\vec{z}_0$, then $\mathcal{E} \mapsto \nu(\vec{z})$ is analytic at $\vec{z}_0$. In particular, we recover Theorem 1.1: $\mathcal{E} \mapsto H^\mathcal{E}(Z)$ is analytic at $\vec{z}_0$.

Proof: The map

$$\Omega \rightarrow F^p \times (F^p)^n \rightarrow \mathbb{R}$$

$$\mathcal{E} \mapsto (f^\mathcal{E}, \nu^\mathcal{E}) \mapsto \nu(\vec{z})$$

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

Appendix 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 and sum norms are equivalent. More specifically, for another metric $d_1$ defined by

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

we have $d_B \sim d_1$. For metric $d_2$ defined by

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

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

$$u_i - v_i = \frac{u_i}{u_1 + u_2 + \cdots + u_k} - \frac{v_i}{v_1 + v_2 + \cdots + v_k}$$

$$= \frac{1}{v_1/v_i + v_2/v_i + \cdots + v_k/v_i}.$$ 

Applying the mean value theorem to function $f$, defined as

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

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

$$u_i - v_i = \nabla f(\xi) \cdot (u_1/v_i - v_1/v_i, \ldots, u_k/v_i - v_k/v_i).$$

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

$$d_E(u, v) < D_1d_1(u, v).$$

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

$$d_2(u, v) < D_2d_E(u, v).$$

Namely, $d_2 \sim d_E$. Thus, the claim in this proposition follows, namely, there exist two positive constant $C_1 < C_2$ such that for any two points $u, v \in S$,

$$C_1d_B(u, v) < d_E(u, v) < C_2d_B(u, v).$$

Appendix B

Proof of Lemma 7.9

Recall that for a nonnegative matrix $B$, the canonical form of $B$ is

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

where $B_{ii}$ is either an irreducible matrix (called irreducible components) or a $1 \times 1$ zero matrix.

Condition 2 in Theorem 7.8 is equivalent to the statement that $B = B(\vec{z}_0)$ 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 $\lambda_1$ denote the spectral radius of $B$, equivalently, the spectral radius of $C$.

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

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

- $x(\varepsilon_0), y(\varepsilon_0) \geq 0$, and they are positive on $S_C$;
- $y(\varepsilon) \cdot x(\varepsilon) = 1$;
- $x(\varepsilon)$ and $y(\varepsilon)$ are analytic in $\varepsilon \in \Omega$.

Let

$$V(\varepsilon) = \lambda_1(\varepsilon)x(\varepsilon) \cdot y(\varepsilon)$$

and

$$U(\varepsilon) = B(\varepsilon) - V(\varepsilon).$$

Then

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

And similarly

$$V(\varepsilon)U(\varepsilon) = 0.$$

Let $\mu(\varepsilon)$ denote the spectral radius of $U(\varepsilon)$. By Condition 2, $\mu(\varepsilon_0) < \lambda_1(\varepsilon_0)$. Thus, there is a constant $\nu > 0$ such that the neighborhood $\Omega$ is sufficiently small, then for all $\varepsilon \in \Omega$

$$\mu(\varepsilon) < \nu < |\lambda_1(\varepsilon)|.$$

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

$$|U^n_{ij}(\varepsilon)| \leq K_1 \mu^n.$$

(B.38)

Let $r = r(\varepsilon_0). c = c(\varepsilon_0), x = x(\varepsilon_0)$ and $y = y(\varepsilon_0)$.

Let $s_0 \in S_C$. Since $\Delta(\varepsilon_0)$ 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 $B_{s_1}^{j_1} c_{s_1} > 0$. Now

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

Since $C$ is primitive, by Perron–Frobenius theory, $C_{n-j_0-j_1}^{j_1}$ 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 $n$

$$rr \cdot yc_{\lambda_1^n} + rU^n c = rV^n c + rU^n c = rB^n c \geq K_2 \lambda_1^n.$$

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

$$|r(\varepsilon)x(\varepsilon) \cdot y(\varepsilon)c(\varepsilon)| > K_4$$

for $\varepsilon \in \Omega$.

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

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

$$|r(\varepsilon)B^n(\varepsilon)c(\varepsilon)| \leq |r(\varepsilon)U^n(\varepsilon)c(\varepsilon)| + |r(\varepsilon)V^n(\varepsilon)c(\varepsilon)| \leq |B|^2 K_3 K_1 \nu^n + |B|^2 K_3 \lambda(\varepsilon)^n$$

and

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

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

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

and

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

are uniformly bounded from above and away from zero. By Condition 1, for any finite collection of $n$, there is a (possibly smaller) neighborhood $\Omega$ of $\varepsilon_0$, such that for all $\varepsilon \in \Omega$, 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).

**APPENDIX C**

$\varepsilon \mapsto \Gamma(\varepsilon)$ Is Analytic

In this appendix, we follow the notation in Section VIII. 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]

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

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

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

$$(L_f)h(x) = \sum_{y \in \pi^{-1}(x)} e^{f(y)} h(y).$$

The connection between pressure and the Ruelle operator is as follows [26], [28]. When $f \in F^\theta$, $P(f) = \log \lambda$, where $\lambda$ is the spectral radius of $L_f$. The restriction of $L_f$ to $F^\theta$ still has spectral radius $\lambda$, and $\lambda$ 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 $F^0$. Moreover, he showed that each $f \in F^0$ has a unique equilibrium state $\mu_f$ and the first-order derivative of $f \mapsto P(f)$ on $F^0$ is $\mu_f$, viewed as a linear functional on $F^0$. So, the analyticity of $P(f)$ implies that the equilibrium state $\mu_f$ is also analytic in $f \in F^0$.

We first claim that for $f(\xi, \omega) = \log P(\omega)$, we have $\mu_f(\xi, \omega) = \frac{\partial}{\partial \xi}$ as in (8.30).

To see this, first observe that the spectral radius $\lambda$ of $\mathcal{L} = \mathcal{L}_f(\xi, \omega)$ is 1; this follows from the following observations:

- the function $\mathcal{I}$ which is identically 1 on $\mathcal{X}$ is a fixed point of $\mathcal{L}$
- (see [26, Proposition 5.16]) $\mathcal{L}^n(\mathcal{I})/\lambda^n$ converges to a strictly positive function.

Thus, $P(f(\xi, \omega)) = 0$. So, for $\mu_f = \mu_f(\xi, \omega)$, we have $h_{\mu_f}(\mathcal{I}) + \int f(\xi, \omega) d\mu_f = 0$.

But from (8.31), we have $h_{\mu_f}(\mathcal{I}) + \int f(\xi, \omega) d\mu_f = 0$.

By uniqueness of the equilibrium state, we thus obtain $\mu_f(\xi, \omega) = \frac{\partial}{\partial \xi}$ as claimed.

Since $\mathcal{I} \mapsto f(\xi, \omega)$ is analytic, it then follows that $\mathcal{I} \mapsto \mu_f$ is analytic, thereby completing the proof of Theorem 8.1.

ACKNOWLEDGMENT
The authors 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.

REFERENCES