

Entropy Rate of Continuous-State Hidden Markov Chains

Guangyue Han
University of Hong Kong
Email: ghan@maths.hku.hk

Brian Marcus
University of British Columbia
Email: marcus@math.ubc.ca

Abstract—We prove that under mild assumptions, the entropy rate of a continuous-state hidden Markov chain, observed when passing a finite-state Markov chain through a discrete-time continuous-output channel, is analytic with respect to the transition probabilities of the underlying Markov chain. We further prove that the entropy rate of a continuous-state hidden Markov chain, observed when passing a mixing finite-type constrained Markov chain through a discrete-time Gaussian channel, is smooth with respect to the transition probabilities of the underlying Markov chain.

I. MAIN RESULTS

Consider a discrete-time memoryless channel with a finite input alphabet \mathcal{Y} and a continuous output alphabet $\mathcal{Z} = \mathbb{R}$. Assume that the input process is a \mathcal{Y} -valued first order stationary Markov chain Y with transition probability matrix $\Pi = (\pi_{ij})_{|\mathcal{Y}| \times |\mathcal{Y}|}$ and stationary vector $\pi = (\pi_i)_{|\mathcal{Y}|}$. Assume that the channel is memoryless in the sense that at each time, the distribution of the output $z \in \mathcal{Z}$, given the input $y \in \mathcal{Y}$, is independent of the previous inputs and outputs, and is distributed according to probability density function $q(z|y)$.

Let Z denote the corresponding output process of this channel. Then the entropy rate $H(Z)$ can be written as

$$H(Z) = \lim_{n \rightarrow \infty} H_n(Z),$$

where

$$H_n(Z) = - \int_{\mathcal{Z}^{n+1}} p(z_{-n}^0) \log p(z_0 | z_{-n}^{-1}) dz_{-n}^0, \quad (1)$$

here $z_{-n}^0 := (z_{-n}, z_{-n+1}, \dots, z_0)$ denotes an instance of $Z_{-n}^0 := (Z_{-n}, Z_{-n+1}, \dots, Z_0)$, and $p(z_{-n}^0)$ denotes the probability density of z_{-n}^0 , and $p(z_0 | z_{-n}^{-1})$ denotes the conditional density of z_0 given z_{-n}^{-1} . For any stochastic process Z , it is well known that (e.g., see page 60 of [3]) if for any n , the integral

$$H(Z_{-n}^0) = - \int_{\mathcal{Z}^{n+1}} p(z_{-n}^0) \log p(z_{-n}^0) dz_{-n}^0$$

is well-defined, then $H(Z)$ is well-defined. Since Z in this paper is stationary (recall that Y is stationary), $H(Z)$ is well-defined, provided the integral

$$\begin{aligned} H(Z_0) &= - \int_{\mathcal{Z}} p(z_0) \log p(z_0) dz_0 \\ &= - \int_{\mathcal{Z}} \sum_{i \in \mathcal{Y}} \pi_i q(z_0|i) \log \sum_{i \in \mathcal{Y}} \pi_i q(z_0|i) dz_0 \end{aligned}$$

is well-defined.

The following theorem states that under positivity assumptions, $H(Z)$ is analytic with respect to Π . More specifically, we state our first result of this paper as follows.

Theorem 1.1: Consider a discrete-time memoryless continuous-output channel as above. Assume that Π is analytically parameterized by $\vec{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_m) \in \Omega$, where Ω denotes an open and bounded subset of \mathbb{R}^m , and assume that the integral

$$\int_{\mathcal{Z}} \sum_{i \in \mathcal{Y}} \pi_i q(z_0|i) \log \sum_{i \in \mathcal{Y}} \pi_i q(z_0|i) dz_0$$

is well-defined. If Π is strictly positive at $\vec{\varepsilon}_0$, then $H(Z)$ is analytic around $\vec{\varepsilon}_0$.

Our next result deals with a discrete-time memoryless Gaussian channel, a special type of discrete-time memoryless continuous-output channel. We shall relax the positivity assumptions in Theorem 1.1, and we assume that the input Markov chain is supported on a mixing finite-type constraint. The consideration of such channels mainly comes from practice: apparently Gaussian channels are of great importance in a variety of scenarios in real applications, and often (particularly in magnetic recording) input sequences are required to satisfy certain constraints in order to eliminate the most damaging error events [8] and the constraints are often mixing finite-type constraints.

Let \mathcal{X} be a finite alphabet, and let \mathcal{X}^n denote the set of words over \mathcal{X} of length n . Let $\mathcal{X}^* = \cup_n \mathcal{X}^n$. A *finite-type constraint* \mathcal{S} over \mathcal{X} is a subset of \mathcal{X}^* defined by a finite list \mathcal{F} of forbidden words [7], [8]; in other words, \mathcal{S} is the set of words over \mathcal{X} that do not contain any element in \mathcal{F} as a contiguous subsequence. We define $\mathcal{S}_n = \mathcal{S} \cap \mathcal{X}^n$. The constraint \mathcal{S} is said to be *mixing* if there exists N such that, for any $u, v \in \mathcal{S}$ and any $n \geq N$, there is a $w \in \mathcal{S}_n$ such that $uwv \in \mathcal{S}$.

The *maximal length* of a forbidden list \mathcal{F} is the length of the longest word in \mathcal{F} . In general, there can be many forbidden lists \mathcal{F} which define the same finite type constraint \mathcal{S} . However, we may always choose a list with smallest maximal length. The *(topological) order* of \mathcal{S} is defined to be $\hat{m} = \hat{m}(\mathcal{S})$ where $\hat{m} + 1$ is the smallest maximal length of any forbidden list that defines \mathcal{S} (the order of the trivial constraint \mathcal{X}^* is taken to be 0). It is easy to see that the order of $\mathcal{S}(d, k)$ is k when $k < \infty$, and is d when $k = \infty$; $\mathcal{S}(d, k)$ is mixing when $d < k$.

For a stationary stochastic process X over \mathcal{X} , the set of *allowed* words with respect to X is defined as

$$\mathcal{A}(X) = \{w_{-n}^0 : n \geq 0, P(X_{-n}^0 = w_{-n}^0) > 0\}.$$

For any m -th order Markov process X , we say X is *supported* on a constraint \mathcal{S} if $\mathcal{S} = \mathcal{A}(X)$; note that in this case, the constraint \mathcal{S} is necessarily of finite-type with order $\hat{m} \leq m$. Also, X is mixing if and only if \mathcal{S} is mixing (recall that a Markov chain is mixing if its transition probability matrix (obtained by appropriately enlarging the state space) is irreducible and aperiodic)).

Now, consider a discrete-time memoryless Gaussian channel, which is a special case of the generic channel model described in the beginning of this paper. More specifically, for any $y \in \mathcal{Y}$, the channel is characterized by the transition probability density function

$$q(z|y) = \frac{1}{\sqrt{2\pi}\sigma_y} e^{-(z-\mu_y)^2/(2\sigma_y^2)},$$

where $z \in \mathcal{Z}$ denotes a possible output of the channel.

The following theorem states that under certain assumptions, $H(Z)$ is smooth with respect to the transition probabilities of Y . More specifically, we state our second result of this paper as follows.

Theorem 1.2: Consider a discrete-time memoryless Gaussian channel as above. Assume that Π is analytically parameterized by $\vec{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_m) \in \Omega$, where Ω denotes an open and bounded subset of \mathbb{R}^m , and assume that for any $\varepsilon \in \Omega$, the input Markov chain Y is supported on a mixing finite-type constraint \mathcal{S} , i.e., $\mathcal{A}(X) = \mathcal{S}$. Then for any $\vec{\varepsilon}_0 \in \Omega$, $H(Z)$ is smooth around $\vec{\varepsilon}_0$.

The rest of the paper is organized as follows. In Section II, we briefly review the Hilbert metric and introduce a complex Hilbert. In Section III, using the complex Hilbert metric, we prove that for any continuous channel, under mild positivity assumptions, $H(Z)$ is analytic with respect to the input Markov parameters (Theorem 1.1). In Section IV, we prove that for a Gaussian channel, where the input Markov chain is supported on a mixing finite-type constraint, $H(Z)$ is smooth with respect to the input Markov parameters (Theorem 1.2).

II. A COMPLEX HILBERT METRIC

Let W be the standard simplex in $|\mathcal{Y}|$ -dimensional real Euclidean space,

$$W = \{w = (w_1, w_2, \dots, w_{|\mathcal{Y}|}) \in \mathbb{R}^{|\mathcal{Y}|} : w_i \geq 0, \sum_i w_i = 1\},$$

and let W° denote its interior, consisting of the vectors with positive coordinates. For any two vectors $v, w \in W^\circ$, the Hilbert metric [9] is defined as

$$d_H(w, v) = \max_{i,j} \log \left(\frac{w_i/w_j}{v_i/v_j} \right). \quad (2)$$

For a $|\mathcal{Y}| \times |\mathcal{Y}|$ strictly positive matrix $T = (t_{ij})$, the mapping f_T induced by T on W is defined by

$$f_T(w) = \frac{wT}{(wT\mathbf{1})}, \quad (3)$$

where $\mathbf{1}$ is the all 1 column vector. It is well known that f_T is a contraction mapping under the Hilbert metric [9]. The contraction coefficient of T , which is also called the Birkhoff coefficient, is given by

$$\tau(T) = \sup_{v \neq w} \frac{d_H(vT, wT)}{d_H(v, w)} = \frac{1 - \sqrt{\phi(T)}}{1 + \sqrt{\phi(T)}}, \quad (4)$$

where $\phi(T) = \min_{i,j,k,l} \frac{t_{ik}t_{jl}}{t_{jk}t_{il}}$.

Let \hat{W} denote the complex version of W ,

$$\hat{W} = \{w = (w_1, w_2, \dots, w_{|\mathcal{Y}|}) \in \mathbb{C}^{|\mathcal{Y}|} : \sum_i w_i = 1\}.$$

Let $\hat{W}^+ = \{v \in \hat{W} : \Re(v_i/v_j) > 0 \text{ for all } i, j\}$. For $v, w \in \hat{W}^+$, let

$$\hat{d}_H(v, w) = \max_{i,j} \left| \log \left(\frac{w_i/w_j}{v_i/v_j} \right) \right|, \quad (5)$$

where \log is taken as the principal branch of the complex $\log(\cdot)$ function (i.e., the branch whose branch cut is the negative real axis). Since the principal branch of \log is additive on the right-half plane, \hat{d}_H is a metric on \hat{W}^+ , which we call a *complex Hilbert metric*.

Let M denote the set of all stochastic matrices with dimension $|\mathcal{Y}| \times |\mathcal{Y}|$, i.e.,

$$M = \{\Pi = (\pi_{ij}) \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{Y}|} : \sum_{j=1}^{|\mathcal{Y}|} \pi_{ij} = 1\}.$$

Let \hat{M} denote the complex version of M , defined as

$$\hat{M} = \{\Pi = (\pi_{ij}) \in \mathbb{C}^{|\mathcal{Y}| \times |\mathcal{Y}|} : \sum_{j=1}^{|\mathcal{Y}|} \pi_{ij} = 1\}.$$

For a given positive Π and a small δ , let $\hat{M}_\Pi(\delta)$ denote the δ -neighborhood around Π within \hat{M} . For an element $\hat{\Pi} \in \hat{M}_\Pi(\delta)$, similar to (3), $\hat{\Pi}$ will induce a mapping $f_{\hat{\Pi}}$ on \hat{W} . The main theorem in [5] says:

Theorem 2.1: There exist $\delta_1, \delta_2 > 0$ such that for any $\hat{\Pi} \in \hat{M}_\Pi(\delta_1)$, $f_{\hat{\Pi}}$ is a contraction mapping on $\hat{W}_H^\circ(\delta_2)$ under the complex Hilbert metric in (5), where $\hat{W}_H^\circ(\delta_2)$ denotes the δ_2 -neighborhood of W° within \hat{W} under the complex Hilbert metric, i.e.,

$$\hat{W}_H^\circ(\delta_2) = \{v = (v_1, v_2, \dots, v_{|\mathcal{Y}|}) \in \hat{W} : \exists u \in W^\circ, \hat{d}_H(v, u) \leq \delta_2\}.$$

III. PROOF OF THEOREM 1.1

In this section, we consider a discrete-time memoryless continuous-output channel as in Theorem 1.1, which has been described in the beginning of Section I.

For each $z \in \mathcal{Z}$, define $\Pi(z)$ as a $|\mathcal{Y}| \times |\mathcal{Y}|$ matrix with the entries

$$\Pi(z)_{ij} = \pi_{ij}(\vec{\varepsilon})q(z|j), \text{ for all } i, j. \quad (6)$$

Apparently $\Pi(z)$, as in (3), will induce a mapping $f_z^{\vec{\varepsilon}} := f_{\Pi(z)}$ from W to W . For any fixed n and z_{-n}^0 , define

$$x_i^{\vec{\varepsilon}} = x_i^{\vec{\varepsilon}}(z_{-n}^i) = p(y_i = \cdot | z_i, z_{i-1}, \dots, z_{-n}), \quad (7)$$

(here \cdot represent the states of the Markov chain Y .) then similar to Blackwell [1], $\{x_i^{\vec{\varepsilon}}\}$ satisfies the random dynamical iteration

$$x_{i+1}^{\vec{\varepsilon}} = f_{z_{i+1}}^{\vec{\varepsilon}}(x_i^{\vec{\varepsilon}}), \quad (8)$$

starting with

$$x_{-n-1}^{\vec{\varepsilon}} = \pi(\vec{\varepsilon}). \quad (9)$$

And obviously we have

$$p^{\vec{\varepsilon}}(z_0|z_{-n}) = x_{-1}^{\vec{\varepsilon}}\Pi(z_0)\mathbf{1}, \quad (10)$$

and

$$p^{\vec{\varepsilon}}(z_{-n}^0) = \pi(\vec{\varepsilon})\Pi(z_{-n})\Pi(z_{-n+1}) \cdots \Pi(z_0)\mathbf{1}. \quad (11)$$

Apparently $x_i^{\vec{\varepsilon}}$, $p^{\vec{\varepsilon}}(z_0|z_{-n})$ and $p^{\vec{\varepsilon}}(z_{-n}^0)$ all depend on the real vector $\vec{\varepsilon} \in \Omega$. In what follows, we shall show that they can be ‘‘complexified’’. For $r > 0$, let $\mathbb{C}_{\vec{\varepsilon}_0}(r)$ denote a r -ball around $\vec{\varepsilon}_0$ in \mathbb{C}^m . For any $\vec{\varepsilon} \in \mathbb{C}_{\vec{\varepsilon}_0}(r)$, one checks that for r small enough, the following system of equations with respect to $\pi(\vec{\varepsilon})$

$$\pi(\vec{\varepsilon})\Pi = \pi(\vec{\varepsilon}), \quad \sum_y \pi(\vec{\varepsilon})_y = 1$$

has a unique solution $\pi(\vec{\varepsilon})$, which is analytic on $\mathbb{C}_{\vec{\varepsilon}_0}(r)$. Then through (15) and (14), $x_i^{\vec{\varepsilon}}$ can analytically continued to $\mathbb{C}_{\vec{\varepsilon}_0}(r)$; furthermore, through (16) and (17), $p^{\vec{\varepsilon}}(z_0|z_{-n})$ and $p^{\vec{\varepsilon}}(z_{-n}^0)$ can be analytically continued to $\mathbb{C}_{\vec{\varepsilon}_0}(r)$. Eventually, $H_n^{\vec{\varepsilon}}(Z)$ can be analytically continued to $\mathbb{C}_{\vec{\varepsilon}_0}(r)$ as well.

For any $z \in \mathcal{Z}$, by the definition of $\Pi(z)$, one checks that for any $u, v \in \hat{W}$, we have

$$\hat{d}_H(u\Pi(z), v\Pi(z)) \leq \hat{d}_H(u\Pi, v\Pi).$$

Then immediately by Lemma 2.1, we have the following lemma, which says that if we perturb $\vec{\varepsilon}_0$ ‘‘a bit’’ to $\vec{\varepsilon}$, $f_z^{\vec{\varepsilon}}$ is contraction mapping on certain complex neighborhood of W° , regardless of the value of z .

Lemma 3.1: There exist $r, \delta > 0$ such that for any $\vec{\varepsilon} \in \mathbb{C}_{\vec{\varepsilon}_0}(r)$ and any $z \in \mathcal{Z}$, $f_z^{\vec{\varepsilon}}$ is a contraction mapping on $\hat{W}_H^\circ(\delta)$ under the complex Hilbert metric in (5), where $\hat{W}_H^\circ(\delta)$ denotes the δ -neighborhood of W° within \hat{W} under the complex Hilbert metric, i.e.,

$$\hat{W}_H^\circ(\delta) = \{v = (v_1, v_2, \dots, v_{|Y|}) \in \hat{W} : \exists u \in W^\circ, \hat{d}_H(v, u) \leq \delta\}.$$

The following lemma, roughly speaking, says that if we perturb $\vec{\varepsilon}_0$ ‘‘a bit’’ to $\vec{\varepsilon}$, the whole image of W under $f_z^{\vec{\varepsilon}}$, for any $z \in \mathcal{Z}$, does not change much from that of $f_z^{\vec{\varepsilon}_0}$.

Lemma 3.2: Consider any $\vec{\varepsilon}_0 \in \Omega$ with $\pi_{ij}(\vec{\varepsilon}_0) > 0$ for all i, j . For any $\delta > 0$, there exists $r > 0$ such that for any $\vec{\varepsilon} \in \mathbb{C}_{\vec{\varepsilon}_0}(r)$, any $z \in \mathcal{Z}$ and any $x \in W$, we have

$$\hat{d}_H(f_z^{\vec{\varepsilon}}(x), f_z^{\vec{\varepsilon}_0}(x)) \leq \delta.$$

The following lemma is implied by Theorem 1.3 in [5].

Lemma 3.3: 1) For any $\delta > 0$, there exists $\xi > 0$ such that for any $\hat{x} \in \hat{W}$, $x \in W$ with $\hat{d}_H(\hat{x}, x) \leq \xi$, we have $|\hat{x} - x| \leq \delta$.

2) For any $\zeta > 0$ and any $\delta > 0$, there exists $\xi > 0$ such that for any $\hat{x}, \hat{y} \in \hat{W}$ with $|\hat{x} - \hat{y}| \leq \zeta$ for certain $x, y \in W$, and $\hat{d}_H(\hat{x}, \hat{y}) \leq \xi$, we have $|\hat{x} - \hat{y}| \leq \delta$.

Lemma 3.4: For any $\delta > 0$, there exists $r > 0$ such that for any n and for all $\vec{\varepsilon} \in \mathbb{C}_{\vec{\varepsilon}_0}(r)$, we have

$$|p^{\vec{\varepsilon}}(z_{-n}^0)| \leq (1 + \delta)^n p^{\vec{\varepsilon}_0}(z_{-n}^0).$$

Lemma 3.5: For sufficiently small $r > 0$, there exist $0 < \rho_1 < 1$ and a positive constant L_1 such that for any two \mathcal{Z} -valued sequences $\{a_{-n_1}^0\}$ and $\{b_{-n_2}^0\}$ with $a_{-n}^0 = b_{-n}^0$ and for all $\vec{\varepsilon} \in \mathbb{C}_{\vec{\varepsilon}_0}(r)$, we have

$$|p^{\vec{\varepsilon}}(a_0|a_{-n_1}^{-1}) - p^{\vec{\varepsilon}}(b_0|b_{-n_2}^{-1})| \leq L_1 \rho_1^n p^{\vec{\varepsilon}_0}(a_0).$$

We are now ready for the proof of Theorem 1.1.

Proof of Theorem 1.1:

We only need to prove that there is a $r > 0$ such that the $H_n^{\vec{\varepsilon}}(Z)$, as $n \rightarrow \infty$, uniformly converge on $\mathbb{C}_{\vec{\varepsilon}_0}(r)$. Note that

$$\begin{aligned} |H_{n+1}^{\vec{\varepsilon}}(Z) - H_n^{\vec{\varepsilon}}(Z)| &= \left| \int_{\mathcal{Z}_{-n-1}^0} p^{\vec{\varepsilon}}(z_{-n-1}^0) \log p^{\vec{\varepsilon}}(z_0|z_{-n-1}^{-1}) dz_{-n-1}^0 \right. \\ &\quad \left. - \int_{\mathcal{Z}_{-n}^0} p^{\vec{\varepsilon}}(z_{-n}^0) \log p^{\vec{\varepsilon}}(z_0|z_{-n}^{-1}) dz_{-n}^0 \right| \\ &= \left| \int_{\mathcal{Z}_{-n-1}^0} p^{\vec{\varepsilon}}(z_{-n-1}^0) (\log f^{\vec{\varepsilon}}(z_0|z_{-n-1}^{-1}) - \log p^{\vec{\varepsilon}}(z_0|z_{-n-1}^{-1})) dz_{-n-1}^0 \right|. \end{aligned}$$

Note that for complex α, β , we have the following inequality

$$|\log \alpha - \log \beta| \leq \max \left(\frac{|\alpha - \beta|}{|\alpha|}, \frac{|\alpha - \beta|}{|\beta|} \right).$$

Now fix $\vec{\varepsilon} \in \mathbb{C}_{\vec{\varepsilon}_0}(r)$, then by Lemma 3.5, either we have, for some $\delta_1, L_1 > 0$ and some $0 < \rho_1 < 1$,

$$\begin{aligned} &|p^{\vec{\varepsilon}}(z_{-n-1}^0) (\log p^{\vec{\varepsilon}}(z_0|z_{-n-1}^{-1}) - \log p^{\vec{\varepsilon}}(z_0|z_{-n}^{-1}))| \\ &\leq \left| p^{\vec{\varepsilon}}(z_{-n-1}^0) \frac{p^{\vec{\varepsilon}}(z_0|z_{-n-1}^{-1}) - p^{\vec{\varepsilon}}(z_0|z_{-n}^{-1})}{p^{\vec{\varepsilon}}(z_0|z_{-n-1}^{-1})} \right| \\ &\leq |p^{\vec{\varepsilon}}(z_{-n-1}^{-1})| L_1 \rho_1^n p^{\vec{\varepsilon}_0}(z_0) \leq L_1 \rho_1^n (1 + \delta_1)^n p^{\vec{\varepsilon}_0}(z_{-n-1}^{-1}) p^{\vec{\varepsilon}_0}(z_0), \end{aligned}$$

or we have, for some $\delta_1, L_1 > 0$ and some $0 < \rho_1 < 1$,

$$\begin{aligned} &|p^{\vec{\varepsilon}}(z_{-n-1}^0) (\log p^{\vec{\varepsilon}}(z_0|z_{-n-1}^{-1}) - \log p^{\vec{\varepsilon}}(z_0|z_{-n}^{-1}))| \\ &\leq \left| p^{\vec{\varepsilon}}(z_{-n-1}^0) \frac{p^{\vec{\varepsilon}}(z_0|z_{-n-1}^{-1}) - p^{\vec{\varepsilon}}(z_0|z_{-n}^{-1})}{p^{\vec{\varepsilon}}(z_0|z_{-n}^{-1})} \right| \\ &\leq |p^{\vec{\varepsilon}}(z_{-n}^{-1}) p^{\vec{\varepsilon}}(z_{-n-1}|z_{-n}^0)| L_1 \rho_1^n p^{\vec{\varepsilon}_0}(z_0) \\ &\leq L_1 \rho_1^n (1 + \delta_1)^n p^{\vec{\varepsilon}_0}(z_0) p^{\vec{\varepsilon}_0}(z_{-n-1}) p^{\vec{\varepsilon}_0}(z_{-n}^{-1}). \end{aligned}$$

Combining all the inequalities above gives us some $L > 0$ and some $0 < \rho < 1$ such that for all $\vec{\varepsilon} \in \mathbb{C}_{\vec{\varepsilon}_0}(r)$,

$$\begin{aligned} &|H_{n+1}^{\vec{\varepsilon}}(Z) - H_n^{\vec{\varepsilon}}(Z)| \\ &\leq \int_{\mathcal{Z}_{-n-1}^0} |p^{\vec{\varepsilon}}(z_{-n-1}^0) (\log p^{\vec{\varepsilon}}(z_0|z_{-n-1}^{-1}) - \log p^{\vec{\varepsilon}}(z_0|z_{-n}^{-1}))| dz_{-n-1}^0 \\ &\leq L \rho^n, \end{aligned}$$

which implies the analyticity of $H^{\vec{\varepsilon}}(Z)$ around $\vec{\varepsilon}_0$. ■

Remark 3.6: Consider a discrete-time memoryless discrete output (with a possibly infinite output alphabet) channel with channel transition probability $q(z|y)$. With essentially the same proof, we can show that if the integral

$$\int_{\mathcal{Z}} \sum_{i \in \mathcal{Y}} \pi_i q(z_0|i) \log \sum_{i \in \mathcal{Y}} \pi_i q(z_0|i) dz_0$$

is well-defined, and the transition probability matrix Π of the input Markov chain Y , analytically parameterized by $\vec{\varepsilon}$, is strictly positive at $\vec{\varepsilon}_0$, then for the corresponding output discrete hidden Markov chain Z , $H(Z)$ is analytic around $\vec{\varepsilon}_0$. More precisely, all the lemmas above still hold; one only has to replace the integral sign \int in the main proof with a summation sign \sum .

For such a channel with only finite output alphabet, analyticity of $H(Z)$ is implied by Theorem 1.1 of [4]. Although the proof of Theorem 1.1 of [4] does not translate to the infinite output (either discrete or continuous) case, the flow of the proof of Theorem 1.1, in fact, mainly follows that of the proof of Theorem 1.1 of [4].

IV. SKETCH OF PROOF OF THEOREM 1.2

In this section, we are concerned with a Gaussian channel, as in Theorem 1.2. Recall that the channel transition probability is given by

$$q(z|y) = \frac{1}{\sqrt{2\pi}\sigma_y} e^{-(z-\mu_y)^2/(2\sigma_y^2)},$$

where $\mu_y, \sigma_y > 0$ and z denotes the output corresponding to the input y . The transition probability matrix Π of the input Markov chain Y is parameterized by $\vec{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_m) \in \Omega$. We assume that at any $\vec{\varepsilon} \in \Omega$, the input Markov chain Y is supported on a mixing finite-type constraint \mathcal{S} , i.e., $\mathcal{A}(Y^{\vec{\varepsilon}}) = \mathcal{S}$. For simplicity, we assume the order of the constraint \mathcal{S} is 1, so is the order of the input Markov chain Y ; the higher order case can be reduced to order 1 case by the usual ‘‘blocking’’ technique.

Assume that e is the smallest integer such that at $\vec{\varepsilon}_0$, Π^e is strictly positive. For the Markov chain Y , define $\tilde{Y} = \{\tilde{Y}_i : i \in \mathbb{Y}\}$ to be a ‘‘blocked’’ process taking values in $\tilde{\mathcal{Y}} = \mathcal{Y}^e$ by

$$\tilde{Y}_i = (Y_{ei}, Y_{ei-1}, \dots, Y_{ei-e+1});$$

correspondingly, for the hidden Markov chain Z , define $\tilde{Z} = \{\tilde{Z}_i : i \in \mathbb{Z}\}$ to be a ‘‘blocked’’ process taking values in $\tilde{\mathcal{Z}} = \mathcal{Z}^e$ by

$$\tilde{Z}_i = (Z_{ei}, Z_{ei-1}, \dots, Z_{ei-e+1}).$$

Apparently $H_n(\tilde{Z})/e$ will converge to $H(Z)$ as n goes to ∞ , thus to prove the smoothness of $H(Z)$, it suffices to prove that $H_n(\tilde{Z})$ and all its derivatives uniformly converge within certain neighborhood of $\vec{\varepsilon}_0$.

For each $\tilde{z} \in \tilde{\mathcal{Z}}$, define $\Pi(\tilde{z})$ by

$$\Pi(\tilde{z}) = \Pi(\tilde{z}_1)\Pi(\tilde{z}_2) \cdots \Pi(\tilde{z}_e). \quad (12)$$

Apparently $\Pi(\tilde{z})$, as in (3), will induce a mapping $f_{\tilde{z}} := f_{\Pi(\tilde{z})}$ from W to W . For any fixed n and \tilde{z}_{-n}^0 , define

$$\tilde{x}_i = \tilde{x}_i(\tilde{z}_{-n}^i) = p(\tilde{y}_i = \cdot | \tilde{z}_i, \tilde{z}_{i-1}, \dots, \tilde{z}_{-n}), \quad (13)$$

(here \cdot represent the states of the Markov chain \tilde{Y} .) then similar to Blackwell [1], $\{\tilde{x}_i^{\vec{\varepsilon}}\}$ satisfies the random dynamical iteration

$$\tilde{x}_{i+1} = f_{\tilde{z}_{i+1}}(\tilde{x}_i), \quad (14)$$

starting with

$$\tilde{x}_{-n-1} = \pi(\vec{\varepsilon}). \quad (15)$$

And obviously we have

$$p(\tilde{z}_0 | \tilde{z}_{-n}) = \tilde{x}_{-1} \Pi(\tilde{z}_0) \mathbf{1}, \quad (16)$$

and

$$p(\tilde{z}_{-n}^0) = \pi(\vec{\varepsilon}) \Pi(\tilde{z}_{-n}) \Pi(\tilde{z}_{-n+1}) \cdots \Pi(\tilde{z}_0) \mathbf{1}. \quad (17)$$

For any fixed $M > 0$, $0 < \alpha < 1$, an instance (with finite length) \tilde{z}_{-n}^{-1} of the above-mentioned \tilde{Z} -process, is said to be (M, α) -typical if the number of i ($-n \leq i \leq -1$) with $|\tilde{z}_i| \leq M$ (here $|\cdot|_{\infty}$ denotes ℓ_{∞} norm of a sequence) is bigger than αn . Let $T_n^{M, \alpha}$ denote the set of all the (M, α) -typical \tilde{Z} -sequences with length n .

Lemma 4.1: Fix $0 < \alpha < 1$. For sufficiently large M , there exists $0 < \rho < 1$ such that

$$\int_{\tilde{z}_{-n}^{-1} \notin T_n^{M, \alpha}} p(\tilde{z}_{-n}^{-1}) d\tilde{z}_{-n}^{-1} = O(\rho^n).$$

The above lemma says that non- (M, α) -typical sequences only occur with exponentially small probability, thus we only have to focus on (M, α) -typical sequences. More precisely, define

$$H_n^{M, \alpha}(\tilde{Z}) = \int_{\tilde{z}_{-n}^{-1} \in T_n^{M, \alpha}, \tilde{z}_0} -p(\tilde{z}_{-n}^0) \log p(\tilde{z}_0 | \tilde{z}_{-n}^{-1}) d\tilde{z}_{-n}^0.$$

Note that for any z_i^j , we have

$$\begin{aligned} \min_l \sum_k \pi_{l,k} e^{-(z_j - \mu_k)^2 / (2\sigma_k^2)} &\leq p(z_j | z_i^{j-1}) \\ &= x_{j-1} \Pi_{z_j} \mathbf{1} \leq \max_l \sum_k \pi_{l,k} e^{-(z_j - \mu_k)^2 / (2\sigma_k^2)}. \end{aligned}$$

It then follows from

$$\begin{aligned} &p(\tilde{z}_0 | \tilde{z}_{-n}^{-1}) \log p(\tilde{z}_0 | \tilde{z}_{-n}^{-1}) \\ &= \prod_{i=-e+1}^0 p(z_i | z_{-en-e+1}^{i-1}) \sum_{i=-e+1}^0 \log p(z_i | z_{-en-e+1}^{i-1}) \end{aligned}$$

that $|p(z_0 | z_{-n-1}^{-1}) \log p(z_0 | z_{-n-1}^{-1})|$ is upper bounded by an integrable function $g(\tilde{z}_0)$, which is independent of \tilde{z}_{-n}^{-1} . It then follows from Lemma 4.1 that there exists $0 < \rho < 1$ such that

$$\begin{aligned} |H_n^{M, \alpha}(\tilde{Z}) - H_n(\tilde{Z})| &= \left| \int_{\tilde{z}_{-n}^{-1} \notin T_n^{M, \alpha}, \tilde{z}_0} -p(\tilde{z}_{-n}^0) \log p(\tilde{z}_0 | \tilde{z}_{-n}^{-1}) d\tilde{z}_{-n}^0 \right| \\ &= \left| \int_{\tilde{z}_{-n}^{-1} \notin T_n^{M, \alpha}, \tilde{z}_0} -p(\tilde{z}_{-n}^{-1}) p(\tilde{z}_0 | \tilde{z}_{-n}^{-1}) \log p(\tilde{z}_0 | \tilde{z}_{-n}^{-1}) d\tilde{z}_{-n}^0 \right| \\ &\leq \left| \int_{\tilde{z}_{-n}^{-1} \notin T_n^{M, \alpha}, \tilde{z}_0} -p(\tilde{z}_{-n}^{-1}) d\tilde{z}_{-n}^{-1} \int_{\tilde{z}_0} g(\tilde{z}_0) d\tilde{z}_0 \right| = O(\rho^n), \end{aligned}$$

which implies that, like $H_n(\tilde{Z})$, $H_n^{M,\alpha}(\tilde{Z})$ converge to $H(\tilde{Z})$, as $n \rightarrow \infty$.

To prove smoothness of $H(Z)$ at $\vec{\varepsilon}_0$, it suffices to prove that $H_n^{M,\alpha}(\tilde{Z})$ and all its derivatives uniformly converge on certain neighborhood of $\vec{\varepsilon}_0$. In the following, we only prove $H_n^{M,\alpha}(\tilde{Z})$ uniformly converge on certain neighborhood of $\vec{\varepsilon}_0$. The proof of uniform convergence of derivatives of $H_n^{M,\alpha}(\tilde{Z})$ is very similar, however much more tedious and technical, thus omitted due to space limit.

Sketch of Proof of Theorem 1.2:

In the following, we prove the convergence of $H_n^{M,\alpha}(\tilde{Z})$ only for illustrative purpose.

Now,

$$\begin{aligned} |H_n^{M,\alpha}(\tilde{Z}) - H_{n+1}^{M,\alpha}(\tilde{Z})| &= \left| \int_{\tilde{z}_{-n}^{-1} \in T_n^{M,\alpha}, \tilde{z}_0} -p(\tilde{z}_{-n}^0) \log p(\tilde{z}_0 | \tilde{z}_{-n}^{-1}) d\tilde{z}_{-n}^0 \right. \\ &\quad \left. - \int_{\tilde{z}_{-n-1}^{-1} \in T_{n+1}^{M,\alpha}, \tilde{z}_0} -p(\tilde{z}_{-n-1}^0) \log p(\tilde{z}_0 | \tilde{z}_{-n-1}^{-1}) d\tilde{z}_{-n-1}^0 \right| \\ &\leq \left| \int_{\tilde{z}_{-n}^{-1} \in T_n^{M,\alpha}, \tilde{z}_{-n-1}^{-1} \in T_{n+1}^{M,\alpha}, \tilde{z}_0} -p(\tilde{z}_{-n-1}^0) \left(\log \frac{p(\tilde{z}_0 | \tilde{z}_{-n}^{-1})}{p(\tilde{z}_0 | \tilde{z}_{-n-1}^{-1})} \right) d\tilde{z}_{-n-1}^0 \right| \\ &\quad + \left| \int_{\tilde{z}_{-n}^{-1} \in T_n^{M,\alpha}, \tilde{z}_{-n-1}^{-1} \notin T_{n+1}^{M,\alpha}, \tilde{z}_0} -p(\tilde{z}_{-n-1}^0) \log p(\tilde{z}_0 | \tilde{z}_{-n-1}^{-1}) d\tilde{z}_{-n-1}^0 \right| \\ &\quad + \left| \int_{\tilde{z}_{-n}^{-1} \notin T_n^{M,\alpha}, \tilde{z}_{-n-1}^{-1} \in T_{n+1}^{M,\alpha}, \tilde{z}_0} -p(\tilde{z}_{-n-1}^0) \log p(\tilde{z}_0 | \tilde{z}_{-n-1}^{-1}) d\tilde{z}_{-n-1}^0 \right| \end{aligned}$$

We first show that the second and the third terms above are $O(\rho^n)$ for some $0 < \rho < 1$. One checks that

$$|p(\tilde{z}_0 | \tilde{z}_{-n-1}^{-1}) \log p(\tilde{z}_0 | \tilde{z}_{-n}^{-1})|$$

is upper bounded by an integrable function $g_0(\tilde{z}_0)$, which is independent of \tilde{z}_{-n}^{-1} . So we have, by Lemma 4.1, for some $0 < \rho < 1$,

$$\begin{aligned} &\left| \int_{\tilde{z}_{-n}^{-1} \in T_n^{M,\alpha}, \tilde{z}_{-n-1}^{-1} \notin T_{n+1}^{M,\alpha}, \tilde{z}_0} -p(\tilde{z}_{-n-1}^0) \log p(\tilde{z}_0 | \tilde{z}_{-n}^{-1}) d\tilde{z}_{-n-1}^0 \right| \\ &= \left| \int_{\tilde{z}_{-n}^{-1} \in T_n^{M,\alpha}, \tilde{z}_{-n-1}^{-1} \notin T_{n+1}^{M,\alpha}, \tilde{z}_0} -p(\tilde{z}_{-n-1}^{-1}) p(\tilde{z}_0 | \tilde{z}_{-n-1}^{-1}) \log p(\tilde{z}_0 | \tilde{z}_{-n}^{-1}) d\tilde{z}_{-n-1}^0 \right| \\ &\leq \left| \int_{\tilde{z}_{-n-1}^{-1} \notin T_{n+1}^{M,\alpha}} -p(\tilde{z}_{-n-1}^{-1}) d\tilde{z}_{-n-1}^{-1} \int_{\tilde{z}_0} g_0(\tilde{z}_0) d\tilde{z}_0 \right| = O(\rho^n). \end{aligned}$$

Similarly $|p(\tilde{z}_{-n-1}^{-1} | \tilde{z}_{-n}^{-1})|$, $|p(\tilde{z}_0 | \tilde{z}_{-n-1}^{-1}) \log p(\tilde{z}_0 | \tilde{z}_{-n-1}^{-1})|$ are upper bounded integrable functions $g_1(\tilde{z}_{-n-1})$, $g_2(\tilde{z}_0)$, respectively. So we have, again by Lemma 4.1, for some $0 < \rho < 1$,

$$\begin{aligned} &\left| \int_{\tilde{z}_{-n}^{-1} \notin T_n^{M,\alpha}, \tilde{z}_{-n-1}^{-1} \in T_{n+1}^{M,\alpha}, \tilde{z}_0} -p(\tilde{z}_{-n-1}^0) \log p(\tilde{z}_0 | \tilde{z}_{-n-1}^{-1}) d\tilde{z}_{-n-1}^0 \right| \\ &= \left| \int_{\tilde{z}_{-n}^{-1} \notin T_n^{M,\alpha}, \tilde{z}_{-n-1}^{-1} \in T_{n+1}^{M,\alpha}, \tilde{z}_0} -p(\tilde{z}_{-n-1}^{-1}) p(\tilde{z}_{-n-1}^{-1} | \tilde{z}_{-n}^{-1}) \right. \end{aligned}$$

$$\begin{aligned} &\left. p(\tilde{z}_0 | \tilde{z}_{-n-1}^{-1}) \log p(\tilde{z}_0 | \tilde{z}_{-n-1}^{-1}) d\tilde{z}_{-n-1}^0 \right| \leq \int_{\tilde{z}_{-n}^{-1} \notin T_n^{M,\alpha}} p(\tilde{z}_{-n}^{-1}) d\tilde{z}_{-n}^{-1} \\ &\int_{\tilde{z}_{-n-1}} g_1(\tilde{z}_{-n-1}) d\tilde{z}_{-n-1} \int_{\tilde{z}_0} g_2(\tilde{z}_0) d\tilde{z}_0 = O(\rho^n). \end{aligned}$$

To show the first term is also $O(\rho^n)$, we need to estimate $|\tilde{x}_i^a - \tilde{x}_i^b|$ where we rewrite $\tilde{x}_i(\tilde{z}_{-n}^i)$, $\tilde{x}_i(\tilde{z}_{-n-1}^i)$ as \tilde{x}_i^a , \tilde{x}_i^b , respectively. Note that for $|\tilde{z}_i|_\infty \leq M$, there is a $0 < \rho_1 < 1$ such that

$$d_H(\tilde{x}_i^a, \tilde{x}_i^b) \leq \rho_1 d_H(\tilde{x}_{i-1}^a, \tilde{x}_{i-1}^b),$$

while otherwise trivially we have

$$d_H(\tilde{x}_i^a, \tilde{x}_i^b) \leq d_H(\tilde{x}_{i-1}^a, \tilde{x}_{i-1}^b).$$

Then for any sequence $\tilde{z}_{-n}^{-1} \in T_n^{M,\alpha}$, let i_0 denote the smallest index such that $|\tilde{z}_{i_0}|_\infty \leq M$, then we have

$$d_H(\tilde{x}_{-1}^a, \tilde{x}_{-1}^b) \leq \rho_1^{\alpha n - 1} d_H(\tilde{x}_{i_0}^a, \tilde{x}_{i_0}^b),$$

which implies that there exists $0 < \rho < 1$ such that $|\tilde{x}_{-1}^a - \tilde{x}_{-1}^b| \leq O(\rho^n)$. It then follows that there exists $0 < \rho < 1$ such that $|p(\tilde{z}_0 | \tilde{z}_{-n}^{-1}) - p(\tilde{z}_0 | \tilde{z}_{-n-1}^{-1})| \leq \rho^n g_3(\tilde{z}_0)$, where $g_3(\tilde{z}_0)$ is an integrable function of \tilde{z}_0 . This, together with the fact that $p(\tilde{z}_{-n-1} | \tilde{z}_{-n}^0)$ is upper bounded by $g_4(\tilde{z}_{-n-1})$, which is an integrable function of \tilde{z}_{-n-1} , will establish the case when $|\vec{l}| = 0$, thus implying that $H_n^{M,\alpha}(Z)$ uniformly converge to $H(\tilde{Z})$. ■

Remark 4.2: One checks that Theorem 1.2 still holds if we only assume that at $\vec{\varepsilon}_0$, the input Markov chain Y is supported on a mixing finite-type constraint \mathcal{S} , i.e., $\mathcal{A}(Y^{\vec{\varepsilon}_0}) = \mathcal{S}$.

REFERENCES

- [1] 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.
- [2] G. Constantine and T. Savits. A Multivariate Faà Di Bruno Formula With Applications. *Transactions of the American Mathematical Society*, Vol. 348, No. 2., Feb, 1996, pp. 503–520.
- [3] S. Ihara. *Information theory for continuous systems*. World Scientific, 1993.
- [4] G. Han and B. Marcus. Analyticity of entropy rate of hidden Markov chains. *IEEE Transactions on Information Theory*, Volume 52, Issue 12, December, 2006, pages: 5251–5266.
- [5] G. Han and B. Marcus. A note on a complex Hilbert metric with application to domain of analyticity for entropy rate of hidden Markov processes. To appear in *Entropy of Hidden Markov Processes and Connections to Dynamical Systems*.
- [6] R. Leipnik and T. Reid. Multivariable Faà di Bruno Formulas. *Electronic Proceedings of the Ninth Annual International Conference on Technology in Collegiate Mathematics*, <http://archives.math.utk.edu/ICTCM/EP-9.html#C23>.
- [7] D. Lind and B. Marcus. *An Introduction to Symbolic Dynamics and Coding*. Cambridge University Press, 1995.
- [8] B. Marcus, R. Roth and P. Siegel. Constrained Systems and Coding for Recording Channels. Chap. 20 in *Handbook of Coding Theory* (eds. V. S. Pless and W. C. Huffman), Elsevier Science, 1998.
- [9] E. Seneta. *Springer Series in Statistics. Non-negative Matrices and Markov Chains*. Springer-Verlag, New York Heidelberg Berlin, 1980.