Concavity of the Mutual Information Rate for Input-Restricted Memoryless Channels at High SNR *

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

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

September 29, 2011

Biography for Guangyue Han: Guangyue Han received the B.S. and M.S. degree in mathematics from Peking University, China, and the Ph.D. degree in mathematics from University of Notre Dame, U.S.A. in 1997, 2000 and 2004, respectively. After three years with the department of mathematics at University of British Columbia, Canada, he joined the department of mathematics at University of Hong Kong, China in 2007. His main research areas are analysis and combinatorics, with an emphasis on their applications to coding and information theory.

Biography for Brian Marcus: Brian H. Marcus received his B.A. from Pomona College in 1971 and Ph.D. in mathematics from UC-Berkeley in 1975. He held the IBM T.J. Watson Postdoctoral Fellowship in mathematical sciences in 1976-7. From 1975-1985 he was Assistant Professor and then Associate Professor of Mathematics (with tenure) at the University of North Carolina – Chapel Hill. From 1984-2002 he was a Research Staff Member at the IBM Almaden Research Center. He is currently Head and Professor of Mathematics at the University of British Columbia. He has been Consulting Associate Professor in Electrical Engineering at Stanford University (2000-2003) and Visiting Associate Professor in Mathematics at UC-Berkeley (1986). He was a co-recipient of the Leonard G. Abraham Prize Paper Award of the IEEE Communications Society in 1993 and gave an invited plenary lecture at the 1995 International Symposium on Information Theory. He is the author of more than fifty research papers in refereed mathematics and engineering journals, as well as a textbook, An Introduction to Symbolic Dynamics and Coding (Cambridge University Press, 1995, co-authored with Doug Lind). He also holds ten US patents. He is an IEEE Fellow, a member of Phi Beta Kappa, and served as an elected Member of the American Mathematical Society Council (2003-2006). His current research interests include constrained coding, error-correction coding, information theory and symbolic dynamics.

*This work is partially supported by a grant from the University Grants Committee of the Hong Kong Special Administrative Region, China (Project No. AoE/E-02/08) and a grant from Research Grants Council of the Hong Kong Special Administrative Region, China under grant No HKU 701708P.

Abstract

We consider a memoryless channel with an input Markov process supported on a mixing finite-type constraint. We continue the development of asymptotics for the entropy rate of the output hidden Markov chain and deduce that, at high signal-to-noise ratio (SNR), the mutual information rate of such a channel is concave with respect to "almost" all input Markov chains of a given order.

Index Terms–concavity, entropy rate, hidden Markov chain, mutual information rate

1 Channel Model

In this paper, we show that for certain input-restricted memoryless channels, the mutual information rate, at high signal-to-noise ratio, is concave with respect to almost all input Markov chains, in the following sense: let $\mathcal{M}_0$ denote the set of all allowed (by the input constraint) first-order Markov processes; at a given noise level, the mutual information rate is strictly concave on a subset of $\mathcal{M}_0$ which increases to the entire $\mathcal{M}_0$ as the noise level approaches zero. Here, we remark that $\mathcal{M}_0$ will be defined precisely immediately following Example 2.1 below, and a corresponding result holds for input Markov chains for any fixed given order.

This partially establishes a very special case of a conjecture of Vontobel et al. [17]. Namely, part of Conjecture 74 of that paper states that for a very general class of finite-state joint source/channel models, the mutual information rate is concave. A proof of the full conjecture (together with other mild assumptions) would imply global convergence of the generalized Blahut-Arimoto algorithm developed in that paper. Our results apply only to certain input-restricted discrete memoryless channels, only at high SNR, with a mild restriction on the class of Markov input processes.

Our approach depends heavily on results regarding asymptotics and smoothness of the entropy rate in special parameterized families of hidden Markov chains, such as those developed in [9], [13], [6], [7], [19], [14], [16] and continued here. The new results along these lines in our paper are of interest, independent of the application to concavity.

We first discuss the nature of the constraints on the input. Let $\mathcal{X}$ be a finite alphabet. Let $\mathcal{X}^n$ denote the set of words over $\mathcal{X}$ of length n and let $\mathcal{X}^* = \cup_n \mathcal{X}^n$. We use the notation $w_{n_1}^{n_2}$ to denote a sequence $w_{n_1} \dots w_{n_2}$.

A *finite-type* constraint $\mathcal{S}$ is a subset of $\mathcal{X}^*$ defined by a finite list $\mathcal{F}$ of forbidden words [11, 12]; equivalently, $\mathcal{S}$ 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 a non-negative integer 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 $uwv \in \mathcal{S}$. To avoid trivial cases, we do not allow $\mathcal{S}$ to consist entirely of constant sequences $a \dots a$ for some symbol a.

In magnetic recording, input sequences are required to satisfy certain constraints in order to eliminate the most damaging error events [12]. The constraints are often mixing finite-type constraints. The most well known example is the (d,k)-RLL constraint $\mathcal{S}(d,k)$, which forbids any sequence with fewer than d or more than k consecutive zeros in between two successive 1's. For $\mathcal{S}(d,k)$ with $k<\infty$, a forbidden set $\mathcal{F}$ is

$$\mathcal{F}=\{1\underbrace{0\cdots0}_{l}1:0\le l<d\}\cup\{\underbrace{0\cdots0}_{k+1}\}.$$

When $k=\infty$, one can choose $\mathcal{F}$ to be

$$\mathcal{F}=\{1\underbrace{0\cdots0}_{l}1:0\le l<d\};$$

in particular when $d=1,k=\infty$, $\mathcal{F}$ can be chosen to be $\{11\}$.

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 $\tilde{m}=\tilde{m}(\mathcal{S})$ where $\tilde{m}+1$ 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

$$\mathcal{A}(X)=\{w_{n_1}^{n_2}\in\mathcal{X}^*:n_1\le n_2,p(X_{n_1}^{n_2}=w_{n_1}^{n_2})>0\};$$

that is, the allowed words are those that occur with strictly positive probability.

Note that for any m-th order stationary Markov process X, the constraint $\mathcal{S}=\mathcal{A}(X)$ is necessarily of finite type with order $\tilde{m}\le m$, and we say that X is *supported* on $\mathcal{S}$. Also, X is mixing iff $\mathcal{S}$ 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). Note that a Markov chain with support contained in a finite-type constraint $\mathcal{S}$ may have order $m<\tilde{m}$.

Now, consider a memoryless channel with inputs $x\in\mathcal{X}$, outputs $z\in\mathcal{Z}$ and input sequences restricted to a mixing finite-type constraint $\mathcal{S}$. Any stationary input process X must satisfy $\mathcal{A}(X)\subseteq\mathcal{S}$. Let Z denote the stationary output process corresponding to X; then at any time slot, the channel is characterized by the conditional probability

$$p(z|x)=p(Z=z|X=x).$$

We are actually interested in families of channels, as above, parameterized by $\varepsilon\ge0$ such that for each x and z, $p(z|x)(\varepsilon)$ is an analytic function of $\varepsilon\ge0$. Recall that an analytic function is one that can be "locally" expressed as a convergent power series (p. 182 of [3]).

We assume that for all x and z, the probability $p(z|x)(\varepsilon)$ is not identically 0 as a function of ε. By a standard result in complex analysis (p. 240 of [3]), this means that for sufficiently small $\varepsilon>0$, $p(z|x)(\varepsilon)\neq0$; it follows that for any input x and sufficiently small $\varepsilon>0$, any output z can occur. We also assume that there is a one-to-one (not necessarily onto) mapping from $\mathcal{X}$ into $\mathcal{Z}$, $z=z(x)$, such that for any $x\in\mathcal{X}$, $p(z(x)|x)(0)=1$; so, ε can be regarded as a parameter that quantifies noise, and $z(x)$ is the noiseless output corresponding to input x. The regime of "small ε" corresponds to high SNR.

Note that the output process $Z=Z(X,\varepsilon)$ depends on the input process X and the parameter value ε; we will often suppress the notational dependence on ε or X, when it is clear from the context. Prominent examples of such families include input-restricted versions of the binary symmetric channel with crossover probability ε (denoted by BSC(ε)), and the binary erasure channel with erasure rate ε (denoted by BEC(ε)).

Recall that the *entropy rate* of $Z = Z(X, \varepsilon)$ is, as usual, defined as

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

where

$$H_n(Z) = H(Z_0|Z_{-n}^{-1}) = -\sum_{z_{-n}^0} p(z_{-n}^0) \log p(z_0|z_{-n}^{-1}).$$

The *mutual information rate* between Z and X can be defined as

$$I(Z;X) = \lim_{n\to\infty} I_n(Z;X),$$

where

$$I_n(Z;X) = H_n(Z) - \frac{1}{n+1} H(Z_{-n}^0|X_{-n}^0).$$

Given the memoryless assumption, one can check that the second term above is simply $H(Z_0|X_0)$ and in particular does not depend on n.

Under our assumptions, if X is a Markov chain, then for each $\varepsilon \geq 0$, the output process $Z = Z(X, \varepsilon)$ is a *hidden Markov chain* and in fact satisfies the "weak Black Hole" assumption of [7], where an asymptotic formula for $H(Z)$ is developed; the asymptotics are given as an expansion in ε around $\varepsilon = 0$. In Section 2, we further develop these ideas to establish smoothness properties of $H(Z)$ as a function of ε and the input Markov chain X of a fixed order. In particular, we show that for small $\varepsilon > 0$, $H(Z)$ can be expressed as $G(X, \varepsilon) + F(X, \varepsilon)\log(\varepsilon)$, where $G(X, \varepsilon)$ and $F(X, \varepsilon)$ are smooth (i.e., infinitely differentiable) functions of ε and of the parameters of the first-order Markov chain X supported on $\mathcal{S}$ (Theorem 2.18). The $\log(\varepsilon)$ term arises from the fact that the support of X will be contained in a non-trivial finite-type constraint and so X will necessarily have some zero transition probabilities; this prevents $H(Z)$ from being smooth in ε at 0. It is natural to ask if $F(X, \varepsilon)$ and $G(X, \varepsilon)$ are in fact analytic; we are only able to show that $F(X, \varepsilon)$ is analytic.

It is well known that for a discrete input random variable over a memoryless channel, mutual information is concave as a function of the input probability distribution (see Theorem 2.7.4 of [4]). In Section 3, we apply the above smoothness results to show that for a mixing finite-type constraint of order 1, and sufficiently small $\varepsilon_0 > 0$, for each $0 \leq \varepsilon \leq \varepsilon_0$, both $I_n(Z(\varepsilon, X); X)$ and the mutual information *rate* $I(Z(X, \varepsilon); X)$ are strictly concave on the set of all first-order Markov chains X whose non-zero transition probabilities are not "too small" (here, the input processes are parameterized by their *joint* probability distributions). This implies that there are unique first-order Markov chains $X_n = X_n(\varepsilon)$, $X_\infty = X_\infty(\varepsilon)$ such that X_n maximizes $I_n(Z(X, \varepsilon), X)$ and X_∞ maximizes $I(Z(X, \varepsilon), X)$. It also follows that $X_n(\varepsilon)$ converges exponentially to $X_\infty(\varepsilon)$ uniformly over $0 \leq \varepsilon \leq \varepsilon_0$. These results are contained in Theorem 3.1. The restriction to first-order constraints and first-order Markov chains is for simplicity only. By a simple recoding via enlarging the state spaces, the results apply to arbitrary mixing finite-type constraints and Markov chains of arbitrary fixed order m. As $m \to \infty$, the maxima converge to channel capacity [1].

2 Asymptotics of the Entropy Rate

2.1 Key Ideas and Lemmas

For simplicity, we consider only mixing finite-type constraints $\mathcal{S}$ of order 1, and correspondingly only first-order input Markov processes X with transition probability matrix Π such that $\mathcal{A}(X) \subseteq \mathcal{S}$ (the higher order case is easily reduced to this). For any $z \in \mathcal{Z}$, define the matrix Ω_z with entries

$$\Omega_z(x, y) = \Pi_{x,y} p(z|y). \tag{1}$$

Note that Ω_z implicitly depends on ε through $p(z|y)$. One checks that

$$\sum_{z \in \mathcal{Z}} \Omega_z = \Pi,$$

and

$$p(z_{-n}^0) = \pi \Omega_{z_{-n}} \Omega_{z_{-n+1}} \cdots \Omega_{z_0} \mathbf{1}, \tag{2}$$

where π is the stationary vector of Π and $\mathbf{1}$ is the all 1's column vector.

For a given analytic function $f(\varepsilon)$ around $\varepsilon = 0$, let $\mathrm{ord}\,(f(\varepsilon))$ denote its order with respect to ε, i.e., the degree of the first non-zero term of its Taylor series expansion around $\varepsilon = 0$. Thus, the orders $\mathrm{ord}\,(p(z|x))$ determine the orders $\mathrm{ord}\,(p(z_{-n}^0))$ and similarly the orders of conditional probabilities $\mathrm{ord}\,(p(z_0|z_{-n}^{-1}))$.

Example 2.1. Consider a binary symmetric channel with crossover probability ε and a binary input Markov chain X supported on the $(1, \infty)$-RLL constraint with transition probability matrix

$$\Pi = \begin{bmatrix} 1-p & p \\ 1 & 0 \end{bmatrix},$$

where $0 < p < 1$. The channel is characterized by the conditional probability

$$p(z|x) = p(z|x)(\varepsilon) = \begin{cases} 1-\varepsilon & \text{if } z = x \\ \varepsilon & \text{if } z \neq x \end{cases}$$

Let Z be the corresponding output binary hidden Markov chain. Now we have

$$\Omega_0 = \begin{bmatrix} (1-p)(1-\varepsilon) & p\varepsilon \\ 1-\varepsilon & 0 \end{bmatrix}, \quad \Omega_1 = \begin{bmatrix} (1-p)\varepsilon & p(1-\varepsilon) \\ \varepsilon & 0 \end{bmatrix}.$$

The stationary vector is $\pi = (1/(p+1), p/(p+1))$, and one computes, for instance,

$$p(z_{-2}z_{-1}z_0 = 110) = \pi \Omega_1 \Omega_1 \Omega_0 \mathbf{1} = \frac{2p - p^2}{1+p} \varepsilon + O(\varepsilon^2),$$

which has order 1 with respect to ε.

Let $\mathcal{M}$ denote the set of all first-order stationary Markov chains X satisfying $\mathcal{A}(X) \subseteq \mathcal{S}$. Let $\mathcal{M}_\delta$, $\delta \geq 0$, denote the set of all $X \in \mathcal{M}$ such that $p(w_{-1}^0) > \delta$ for all $w_{-1}^0 \in \mathcal{S}_2$. Note that whenever $X \in \mathcal{M}_0$, i.e., $\mathcal{A}(X) = \mathcal{S}$, X is mixing (thus its transition probability matrix

Π is primitive) since $\mathcal{S}$ is mixing, so X is completely determined by its transition probability matrix Π. For the purposes of this paper, however, we find it convenient to identify each $X \in \mathcal{M}_0$ with its vector of *joint* probabilities $\vec{p} = \vec{p}_X$ on words of length 2 instead:

$$\vec{p} = \vec{p}_X = (p(X_{-1}^0 = w_{-1}^0) : w_{-1}^0 \in \mathcal{S}_2);$$

sometimes we write $X = X(\vec{p})$. This is the same parameterization of Markov chains as in Definition 33 of [17].

In the following, for any parameterized sequence of functions $f_{n,\lambda}(\varepsilon)$ (ε is real or complex) with λ ranging within a parameter space Λ, we use

$$f_{n,\lambda}(\varepsilon) = \hat{O}(\varepsilon^n) \text{ on } \Lambda$$

to mean that there exist constants $C, \beta_1, \beta_2 > 0$, $\varepsilon_0 > 0$ such that for all n, all $\lambda \in \Lambda$ and all $0 \le |\varepsilon| \le \varepsilon_0$,

$$|f_{n,\lambda}(\varepsilon)| \le n^{\beta_1}(C|\varepsilon|^{\beta_2})^n.$$

Note that $f_{n,\lambda}(\varepsilon) = \hat{O}(\varepsilon^n)$ on Λ implies that there exists $\varepsilon_0 > 0$ and $0 < \rho < 1$ such that $|f_{n,\lambda}(\varepsilon)| < \rho^n$ for all $|\varepsilon| \le \varepsilon_0$, all $\lambda \in \Lambda$ and large enough n. One also checks that a $\hat{O}(\varepsilon^n)$-term is unaffected by multiplication by an exponential function in n (and thus a polynomial function in n, since, roughly speaking, a polynomial function does not grow as fast as an exponential function as n tends to infinity) and a polynomial function in $1/\varepsilon$; in particular, note that

Remark 2.2. For any given $f_{n,\lambda}(\varepsilon) = \hat{O}(\varepsilon^n)$, there exists $\varepsilon_0 > 0$ and $0 < \rho < 1$ such that $|g_1(n)g_2(1/\varepsilon)f_{n,\lambda}(\varepsilon)| \le \rho^n$, for all $|\varepsilon| \le \varepsilon_0$, all $\lambda \in \Lambda$, all polynomial functions $g_1(n), g_2(1/\varepsilon)$ and large enough n.

Of course, the output joint probabilities $p(z_{-n}^0)$ and conditional probabilities $p(z_0|z_{-n}^{-1})$ implicitly depend on $\vec{p} \in \mathcal{M}_0$ and ε. The following result asserts that for small ε, the total probability of output sequences with "large" order is exponentially small, uniformly over all input processes.

Lemma 2.3. *For any fixed* $0 < \alpha < 1$,

$$\sum_{z_{-n}^{-1}:\, \mathrm{ord}\,(p(z_{-n}^{-1})) \ge \alpha n} p(z_{-n}^{-1}) = \hat{O}(\varepsilon^n) \text{ on } \mathcal{M}_0.$$

Proof. Note that for any hidden Markov chain sequence z_{-n}^{-1}, we have

$$p(z_{-n}^{-1}) = \sum_{x_{-n}^{-1}} p(x_{-n}^{-1}) \prod_{i=-n}^{-1} p(z_i|x_i). \tag{3}$$

Now consider z_{-n}^{-1} with $k = \mathrm{ord}\,(p(z_{-n}^{-1})) \ge \alpha n$. One checks that for ε small enough there exists a positive constant C such that $p(z|x) \le C\varepsilon$ for all x, z with $\mathrm{ord}\,(p(z|x)) \ge 1$, and thus the term $\prod_{i=-n}^{-1} p(z_i|x_i)$ as in (3) is upper bounded by $C^k \varepsilon^k$, which is upper bounded by $C^{\alpha n}\varepsilon^{\alpha n}$ for $\varepsilon < 1/C$. Noticing that $\sum_{x_{-n}^{-1}} p(x_{-n}^{-1}) = 1$, we then have, for ε small enough,

$$\sum_{z_{-n}^{-1}:\ \mathrm{ord}\,(p(z_{-n}^{-1}))\geq \alpha n} p(z_{-n}^{-1}) \leq \sum_{z_{-n}^{-1}} \sum_{x_{-n}^{-1}} p(x_{-n}^{-1}) C^{\alpha n}\varepsilon^{\alpha n} \leq |\mathcal{Z}|^n C^{\alpha n}\varepsilon^{\alpha n},$$

which immediately implies the lemma. □

Remark 2.4. Note that for any z_{-n}^{-1} with $\mathrm{ord}\,(p(z_{-n}^{-1})) \geq \alpha n$, one immediately has

$$p(z_{-n}^{-1}) \leq K\varepsilon^{\alpha n}, \tag{4}$$

for a suitable K and small enough ε. However, this K may depend on z_{-n}^{-1} and n, so (4) does not imply Lemma 2.3.

By Lemma 2.3 the probability measure is concentrated mainly on the set of output sequences with relatively small order, and so we can focus on those sequences. For a fixed positive α, a sequence $z_{-n}^{-1} \in \mathcal{Z}^n$ is said to be *α-typical* if $\mathrm{ord}\,(p(z_{-n}^{-1})) \leq \alpha n$; let T_n^α denote the set of all α-typical $\mathcal{Z}$-sequences with length n. Note that this definition is independent of $\vec{p} \in \mathcal{M}_0$.

For a smooth mapping $f(\vec{x})$ from $\mathbb{R}^k$ to $\mathbb{R}$ and a nonnegative integer ℓ, $D_{\vec{x}}^\ell f$ denotes the ℓ-th total derivative with respect to $\vec{x}$; for instance,

$$D_{\vec{x}} f = \left(\frac{\partial f}{\partial x_i}\right)_i \text{ and } D_{\vec{x}}^2 f = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right)_{i,j}.$$

In particular, if $\vec{x} = \vec{p} \in \mathcal{M}_0$ or $\vec{x} = (\vec{p}, \varepsilon) \in \mathcal{M}_0 \times [0,1]$, this defines the derivatives $D_{\vec{p}}^\ell p(z_0|z_{-n}^{-1})$ or $D_{\vec{p},\varepsilon}^\ell p(z_0|z_{-n}^{-1})$. We shall use $|\cdot|$ to denote the Euclidean norm of a vector or a matrix (for a matrix $A = (a_{ij})$, $|A| = \sqrt{\sum_{i,j} a_{ij}^2}$), and we shall use $\|A\|$ to denote the matrix norm, that is,

$$\|A\| = \sup_{x \neq \vec{0}} \frac{|Ax|}{|x|}.$$

It is well known that $\|A\| \leq |A|$.

In this paper, we are interested in functions of $\vec{q} = (\vec{p}, \varepsilon)$. For any $\vec{n} = (n_1, n_2, \ldots, n_{|\mathcal{S}_2|+1}) \in \mathbb{Z}_+^{|\mathcal{S}_2|+1}$ and any smooth function f of $\vec{q}$, define

$$f^{(\vec{n})} = \frac{\partial^{|\vec{n}|} f}{\partial q_1^{n_1} \partial q_2^{n_2} \cdots \partial q_{|\mathcal{S}_2|+1}^{n_{|\mathcal{S}_2|+1}}},$$

here $|\vec{n}|$ denotes the order of the $\vec{n}$-th derivative of f with respect to $\vec{q}$, and is defined as

$$|\vec{n}| = n_1 + n_2 + \cdots + n_{|\mathcal{S}_2|+1}.$$

The next result shows, in a precise form, that for α-typical sequences z_{-n}^0, the derivatives, of all orders, of the difference between $p(z_0|z_{-n}^{-1})$ and $p(z_0|z_{-n-1}^{-1})$ converge exponentially in n, uniformly in $\vec{p}$ and ε. For $n \leq m, \hat{m} \leq 2n$, define

$$T_{n,m,\hat{m}}^\alpha = \{(z_{-m}^0, \hat{z}_{-\hat{m}}^0) \in \mathcal{Z}^{m+1} \times \mathcal{Z}^{\hat{m}+1} | z_{-n}^{-1} = \hat{z}_{-n}^{-1} \text{ is } \alpha\text{-typical}\}.$$

We then have the following proposition, whose proof is deferred to Section 2.2.

Proposition 2.5. *Assume $n \le m, \hat{m} \le 2n$. Given $\delta_0 > 0$, there exists $\alpha > 0$ such that for any ℓ*

$$|D^{\ell}_{\vec{p},\varepsilon} p(z_0|z_{-m}^{-1}) - D^{\ell}_{\vec{p},\varepsilon} p(\hat{z}_0|\hat{z}_{-\hat{m}}^{-1})| = \hat{O}(\varepsilon^n) \text{ on } \mathcal{M}_{\delta_0} \times T^{\alpha}_{n,m,\hat{m}}.$$

The proof of Proposition 2.5 depends on estimates of derivatives of certain induced maps on a simplex, which we now describe. Let $\mathcal{W}$ denote the unit simplex in $\mathbb{R}^{|\mathcal{X}|}$, i.e., the set of nonnegative vectors, which sum to 1, indexed by the joint input-state space $\mathcal{X}$. For any $z \in \mathcal{Z}$, Ω_z induces a mapping f_z defined on $\mathcal{W}$ by

$$f_z(w) = \frac{w\Omega_z}{w\Omega_z \mathbf{1}}. \tag{5}$$

Note that Ω_z implicitly depends on the input Markov chain $\vec{p} \in \mathcal{M}_0$ and ε, and thus so does f_z. While $w\Omega_z\mathbf{1}$ can vanish at $\varepsilon = 0$, it is easy to check that for all $w \in \mathcal{W}$, $\lim_{\varepsilon \to 0} f_z(w)$ exists, and so f_z can be defined at $\varepsilon = 0$. Let $O_{\max}$ denote the largest order of all entries of Ω_z (with respect to ε) for all $z \in \mathcal{Z}$, or equivalently, the largest order of $p(z|x)(\varepsilon)$ over all possible x, z.

For $\varepsilon_0, \delta_0 > 0$, let

$$U_{\delta_0,\varepsilon_0} = \{\vec{p} \in \mathcal{M}_{\delta_0}, \varepsilon \in [0, \varepsilon_0]\}.$$

Lemma 2.6. *Given $\delta_0 > 0$, there exists $\varepsilon_0 > 0$ and $C_{\mathrm{a}} > 0$ such that on $U_{\delta_0,\varepsilon_0}$ for all $z \in \mathcal{Z}$, $|D_w f_z| \le C_{\mathrm{a}}/\varepsilon^{2O_{\max}}$ on the entire simplex $\mathcal{W}$.*

Proof. Given $\delta_0 > 0$, there exist $\varepsilon_0 > 0$ and $C > 0$ such that for any $z \in \mathcal{Z}$, $w \in \mathcal{W}$, we have, for all $0 \le \varepsilon \le \varepsilon_0$,

$$|w\Omega_z \mathbf{1}| \ge C\varepsilon^{O_{\max}}.$$

We then apply the quotient rule for derivatives to establish the lemma. □

For any sequence $z_{-N}^{-1} \in \mathcal{Z}^N$, define

$$\Omega_{z_{-N}^{-1}} = \Omega_{z_{-N}} \Omega_{z_{-N+1}} \cdots \Omega_{z_{-1}}.$$

Similar to (5), $\Omega_{z_{-N}^{-1}}$ induces a mapping $f_{z_{-N}^{-1}}$ on $\mathcal{W}$ by:

$$f_{z_{-N}^{-1}}(w) = \frac{w\Omega_{z_{-N}^{-1}}}{w\Omega_{z_{-N}^{-1}} \mathbf{1}}.$$

By the chain rule, Lemma 2.6 gives upper bounds on derivatives of $f_{z_{-N}^{-1}}$. However, these bounds can be improved considerably in certain cases, as we now describe. A sequence $z_{-N}^{-1} \in \mathcal{Z}^N$ is Z-*allowed* if there exists $x_{-N}^{-1} \in \mathcal{A}(X)$ such that

$$z_{-N}^{-1} = z(x_{-N}^{-1}).$$

where $z(x_{-N}^{-1}) = (z(x_{-N}), z(x_{-N+1}), \ldots, z(x_{-1}))$. Note that z_{-N}^{-1} is Z-allowed iff $\mathrm{ord}\,(p(z_{-N}^{-1})) = 0$. So, the Z-allowed sequences are those output sequences resulting from noiseless transmission of input sequences that satisfy the constraint.

Since Π is a primitive matrix, by definition there exists a positive integer e such that $\Pi^e > 0$ (i.e., all entries of the matrix power are strictly positive). We then have the following lemma.

Lemma 2.7. *Assume that $X \in \mathcal{M}_0$. For any Z-allowed sequence $z_{-N}^{-1} = z(x_{-N}^{-1}) \in \mathcal{Z}^N$ (here $x_{-N}^{-1} \in \mathcal{S}$), if $N \geq 2eO_{\max}$, we have*

$$\mathrm{ord}\,(\Omega_{z_{-N}^{-1}}(\hat{x}_{-N-1}, x_{-1})) < \mathrm{ord}\,(\Omega_{z_{-N}^{-1}}(\hat{x}_{-N-1}, \tilde{x}_{-1})),$$

for any $\hat{x}_{-N-1} \in \mathcal{X}$ and any $\tilde{x}_{-1}$ with $\tilde{x}_{-1} \neq x_{-1}$.

Proof. The rough idea is that to minimize the order, a sequence must match x_{-N}^{-1} as closely as possible. Given the restrictions on initial and terminal states, the length N must be sufficiently long to overwhelm edge effects.

For any $\hat{x}_{-N-1}, \hat{x}_{-1} \in \mathcal{X}$, we have

$$\Omega_{z_{-N}^{-1}}(\hat{x}_{-N-1}, \hat{x}_{-1}) = p(X_{-1} = \hat{x}_{-1}, Z_{-N}^{-1} = z_{-N}^{-1} | X_{-N-1} = \hat{x}_{-N-1}) = p(\hat{x}_{-1}, z_{-N}^{-1} | \hat{x}_{-N-1}).$$

It then follows that

$$\mathrm{ord}\,(\Omega_{z_{-N}^{-1}}(\hat{x}_{-N-1}, \hat{x}_{-1})) = \mathrm{ord}\,(p(\hat{x}_{-1}, z_{-N}^{-1} | \hat{x}_{-N-1})) = \mathrm{ord}\,(p(\hat{x}_{-N-1}, z_{-N}^{-1}, \hat{x}_{-1})).$$

Since

$$p(\hat{x}_{-N-1}, z_{-N}^{-1}, \hat{x}_{-1}) = \sum_{\hat{x}_{-N}^{-2}} p(\hat{x}_{-N-1}^{-1}, z_{-N}^{-1}),$$

we have

$$\mathrm{ord}\,(\Omega_{z_{-N}^{-1}}(\hat{x}_{-N-1}, \hat{x}_{-1})) = \min \sum_{i=-N}^{-1} \mathrm{ord}\,(p(z_i | \hat{x}_i)),$$

where the minimization is over all sequences $\hat{x}_{-N}^{-2}$ such that $\hat{x}_{-N-1}^{-1} \in \mathcal{S}$.

Since $\Pi^e > 0$, there exists some $\hat{x}_{-N}^{-N-1+e}$ such that $\hat{x}_{-N-1+e} = x_{-N-1+e}$ and $p(\hat{x}_{-N-1}^{-N-1+e}) > 0$, and there exists some $\hat{x}_{-e}^{-2}$ such that $\hat{x}_{-e} = x_{-e}$ and $p(\hat{x}_{-e}^{-1}) > 0$. It then follows from $\mathrm{ord}\,(p(z|x)) \leq O_{\max}$ that, as long as $N \geq 2eO_{\max}$, for any fixed $\hat{x}_{-1}$ and any choice of order minimizing sequence $\hat{x}_{-N}^{-2}(\hat{x}_{-1})$, there exist $0 \leq i_0 = i_0(\hat{x}_{-1}), j_0 = j_0(\hat{x}_{-1}) \leq eO_{\max}$ such that $z(\hat{x}_i^j(\hat{x}_{-1})) = z_i^j$ if and only if $i \geq -N - 1 + i_0(\hat{x}_{-1})$ and $j \leq -1 - j_0(\hat{x}_{-1})$. One further checks that, for any choice of order minimizing sequences corresponding to $\hat{x}_{-1}$, $\hat{x}_{-N}^{-2}(\hat{x}_{-1})$,

$$\sum_{i=-N}^{-N-1+i_0(\hat{x}_{-1})} \mathrm{ord}\,(p(z_i | \hat{x}_i(\hat{x}_{-1}))),$$

does not depend on $\hat{x}_{-1}$, whereas $j_0(\hat{x}_{-1}) = 0$ if and only if $\hat{x}_{-1} = x_{-1}$. This immediately implies the lemma.

□

Example 2.8. (continuation of Example 2.1)

Recall that

$$\Omega_0 = \begin{bmatrix} (1-p)(1-\varepsilon) & p\varepsilon \\ 1-\varepsilon & 0 \end{bmatrix}, \qquad \Omega_1 = \begin{bmatrix} (1-p)\varepsilon & p(1-\varepsilon) \\ \varepsilon & 0 \end{bmatrix}.$$

First, observe that the only Z-allowed sequences are $00, 01, 10$; then straightforward computations show that

$$\begin{aligned}
\Omega_0\Omega_0 &= \begin{bmatrix} (1-p)^2(1-\varepsilon)^2 + p\varepsilon(1-\varepsilon) & p(1-p)\varepsilon(1-\varepsilon) \\ (1-p)(1-\varepsilon)^2 & p\varepsilon(1-\varepsilon) \end{bmatrix}, \\
\Omega_0\Omega_1 &= \begin{bmatrix} (1-p)^2\varepsilon(1-\varepsilon) + p\varepsilon^2 & p(1-p)(1-\varepsilon)^2 \\ (1-p)\varepsilon(1-\varepsilon) & p(1-\varepsilon)^2 \end{bmatrix}, \\
\Omega_1\Omega_0 &= \begin{bmatrix} (1-p)^2\varepsilon(1-\varepsilon) + p(1-\varepsilon)^2 & p(1-p)\varepsilon^2 \\ (1-p)\varepsilon(1-\varepsilon) & p\varepsilon^2 \end{bmatrix}.
\end{aligned}$$

One checks that for each of these three matrices, there is a unique column, each of whose entries minimizes the orders over all the entries in the same row. Note that, putting this example in the context of Lemma 2.7, we have $N = 2$, which is smaller than $2eO_{\max} = 2 \times 2 \times 1 = 4$.

Now fix $N \geq 2eO_{\max}$. Note that the mapping $f_{z_{-N}^{-1}}$ implicitly depends on ε, so for any $w \in \mathcal{W}$, $v = f_{z_{-N}^{-1}}(w)$ is in fact a function of ε. Let $q(z) \in \mathcal{W}$ be the point defined by $q(z)_x = 1$ for x with $z(x) = z$ and 0 otherwise. If z_{-N}^{-1} is Z-allowed, then by Lemma 2.7, we have

$$\lim_{\varepsilon \to 0} f_{z_{-N}^{-1}}(w) = q(z_{-1});$$

thus, in this limiting sense, at $\varepsilon = 0$, $f_{z_{-N}^{-1}}$ maps the entire simplex $\mathcal{W}$ to a single point $q(z_{-1})$. The following lemma says that if z_{-N-1}^{-1} is Z-allowed, then in a small neighbourhood of $q(z_{-N-1})$, the derivative of $f_{z_{-N}^{-1}}$ is much smaller than what would be given by repeated application of Lemma 2.6.

Lemma 2.9. *Given $\delta_0 > 0$, there exists $\varepsilon_0 > 0$ and $C_{\mathrm{b}} > 0$ such that on $U_{\delta_0,\varepsilon_0}$, if z_{-N-1}^{-1} is Z-allowed, then $|D_w f_{z_{-N}^{-1}}| \leq C_{\mathrm{b}}\varepsilon$ on some neighbourhood of $q(z_{-N-1})$.*

Proof. By the observations above, for all $w \in \mathcal{W}$, we have

$$f_{z_{-N}^{-1}}(w) = q(z_{-1}) + \varepsilon r(w),$$

where $r(w)$ is a rational vector-valued function with common denominator of order 0 (in ε) and leading coefficient uniformly bounded away from 0 near $w = q(z_{-N-1})$ over all $\vec{p} \in \mathcal{M}_{\delta_0}$. The lemma then immediately follows. □

2.2 Proof of Proposition 2.5

Before giving the detailed proof of Proposition 2.5, let us roughly explain the proof only for the special case $\ell = 0$, i.e., convergence of the difference between $p(z_0|z_{-n}^{-1})$ and $p(z_0|z_{-n-1}^{-1})$. Let N be as above and for simplicity consider only output sequences of length a multiple N: $n = n_0 N$. We can compute an estimate of $D_w f_{z_{-n}^0}$ by using the chain rule (with appropriate care at $\varepsilon = 0$) and multiplying the estimates on $|D_w f_{z_{-iN}^{(-i+1)N}}|$ given by Lemmas 2.6 and 2.9. This yields an estimate of the form, $|D_w f_{z_{-n}^0}| \leq (A\varepsilon^{1-B\alpha})^n$ for some constants A and B, on the entire simplex $\mathcal{W}$. If α is sufficiently small and z_{-n}^{-1} is α-typical, then the estimate from Lemma 2.9 applies enough of the time that $f_{z_{-n}^0}$ exponentially contracts the simplex. Then, interpreting elements of the simplex as conditional probabilities $p(X_i = \cdot | z_{-m}^i)$, we obtain exponential convergence of the difference $|p(z_0|z_{-n}^{-1}) - p(z_0|z_{-n-1}^{-1})|$ in n, as desired.

Proof of Proposition 2.5. For simplicity, we only consider the special case that $n = n_0 N, m = m_0 N, \hat{m} = \hat{m}_0 N$ for a fixed $N \geq 2eO_{\max}$; the general case can be easily reduced to this special case. For the sequences $z_{-m}^{-1}, \hat{z}_{-\hat{m}}^{-1}$, define their "blocked" versions $[z]_{-m_0}^{-1}, [\hat{z}]_{-\hat{m}_0}^{-1}$ by setting

$$[z]_i = z_{iN}^{(i+1)N-1}, i = -m_0, -m_0+1, \dots, -1, \qquad [\hat{z}]_j = \hat{z}_{jN}^{(j+1)N-1}, j = -\hat{m}_0, -\hat{m}_0+1, \dots, -1.$$

We first consider the case $\ell = 0$.

Let

$$w_{i,-m} = w_{i,-m}(z_{-m}^i) = p(X_i = \cdot \, | z_{-m}^i),$$

where $\cdot$ denotes the possible states of the Markov chain X. Then one checks that

$$p(z_0|z_{-m}^{-1}) = w_{-1,-m}\Omega_{z_0}\mathbf{1} \tag{6}$$

and $w_{i,-m}$ satisfies the following iteration

$$w_{i+1,-m} = f_{z_{i+1}}(w_{i,-m}) \qquad -n \leq i \leq -1,$$

and the following iteration (corresponding to the blocked chain $[z]_{-m_0}^{-1}$)

$$w_{(i+1)N-1,-m} = f_{[z]_i}(w_{iN-1,-m}) \qquad -n_0 \leq i \leq -1, \tag{7}$$

starting with

$$w_{-n-1,-m} = p(X_{-n-1} = \cdot \, | z_{-m}^{-n-1}).$$

Similarly let

$$\hat{w}_{i,-\hat{m}} = \hat{w}_{i,-\hat{m}}(\hat{z}_{-\hat{m}}^i) = p(X_i = \cdot \, | \hat{z}_{-\hat{m}}^i),$$

which also satisfies the same iterations as above, however starting with

$$\hat{w}_{-n-1,-\hat{m}} = p(X_{-n-1} = \cdot \, | \hat{z}_{-\hat{m}}^{-n-1}).$$

For any $-n_0 < i \leq -1$, we say $[z]_{-n_0}^{-1}$ *continues* between $[z]_{i-1}$ and $[z]_i$ if $[z]_{i-1}^i$ is Z-allowed; on the other hand, we say $[z]_{-n_0}^{-1}$ *breaks* between $[z]_{i-1}$ and $[z]_i$ if it does not continue between $[z]_{i-1}$ and $[z]_i$, namely, if any one of the following occurs:

1. $[z]_{i-1}$ is not Z-allowed;
2. $[z]_i$ is not Z-allowed;
3. both $[z]_{i-1}$ and $[z]_i$ are Z-allowed, however $[z]_{i-1}^i$ is not Z-allowed.

Iteratively applying Lemma 2.6, there is a positive constant $C_{\rm a}$ such that

$$|D_w f_{[z]_i}| \le C_{\rm a}^N/\varepsilon^{2NO_{\max}}, \tag{8}$$

on the entire simplex $\mathcal{W}$. In particular, this holds when $[z]_{-n_0}^{-1}$ "breaks" between $[z]_{i-1}$ and $[z]_i$. When $[z]_{-n_0}^{-1}$ "continues" between $[z]_{i-1}$ and $[z]_i$, by Lemma 2.9, we have that if ε is small enough, there is a constant $C_{\rm b} > 0$ such that

$$|D_w f_{[z]_i}| \le C_{\rm b}\varepsilon \tag{9}$$

on $f_{[z]_{i-1}}(\mathcal{W})$.

Now, applying the mean value theorem, we deduce that there exist ξ_i, $-n_0 \le i \le -1$, (here ξ_i is a convex combination of $w_{-iN-1,-m}$ and $\hat{w}_{-iN-1,-\hat{m}}$) such that

$$\begin{aligned}|w_{-1,-m} - \hat{w}_{-1,-\hat{m}}| &= |f_{[z]_{-n_0}^{-1}}(w_{-n_0N-1,-m}) - f_{[z]_{-n_0}^{-1}}(\hat{w}_{-n_0N-1,-\hat{m}})| \\ &\le \prod_{i=-n_0}^{-1} \|D_w f_{[z]_i}(\xi_i)\| \cdot |w_{-n_0N-1,-m} - \hat{w}_{-n_0N-1,-\hat{m}}|.\end{aligned}$$

If z_{-n}^{-1} satisfies the hypothesis of Proposition 2.5, then it is α-typical (recall the definition of $T^{\alpha}_{n,m,\hat{m}}$). It follows that $[z]_{-n_0}^{-1}$ breaks for at most $2\alpha n$ values of i (since, roughly speaking, each non-Z-allowed block $[z]_i$ contributes at most twice to the number of breakings); in other words, there are at least $(1/N - 2\alpha)n$ i's corresponding to (9) and at most $2\alpha n$ i's corresponding to (8). We then have

$$\prod_{i=-n_0}^{-1} \|D_w f_{[z]_i}(\xi_i)\| \le C_{\rm b}^{(1/N-2\alpha)n} C_{\rm a}^{2\alpha N n} \varepsilon^{(1/N-2\alpha-4NO_{\max}\alpha)n}. \tag{10}$$

Let $\alpha_0 = 1/(N(2 + 4NO_{\max}))$. Evidently, when $\alpha < \alpha_0$, $1/N - 2\alpha - 4NO_{\max}\alpha$ is strictly positive. We then have

$$|w_{-1,-m} - \hat{w}_{-1,-\hat{m}}| = \hat{O}(\varepsilon^n) \text{ on } \mathcal{M}_{\delta_0} \times T^{\alpha}_{n,m,\hat{m}}. \tag{11}$$

It then follows from (6) that

$$|p(z_0|z_{-m}^{-1}) - p(\hat{z}_0|\hat{z}_{-\hat{m}}^{-1})| = \hat{O}(\varepsilon^n) \text{ on } \mathcal{M}_{\delta_0} \times T^{\alpha}_{n,m,\hat{m}}.$$

This completes the proof for the special case $\ell = 0$.

The general case $\ell > 0$ follows along the same lines as in the special case, together with the following lemmas, whose proofs are deferred to the appendix.

Lemma 2.10. *For each $\vec{k}$, there is a positive constant $C_{|\vec{k}|}$ such that*

$$|w_{i,-m}^{(\vec{k})}|, |\hat{w}_{i,-\hat{m}}^{(\vec{k})}| \le n^{|\vec{k}|} C_{|\vec{k}|}/\varepsilon^{|\vec{k}|};$$

here, the superscript $^{(\vec{k})}$ denotes the $\vec{k}$-th order derivative with respect to $\vec{q} = (\vec{p}, \varepsilon)$. In fact, the partial derivatives with respect to $\vec{p}$ are upper bounded in norm by $n^{|\vec{k}|} C_{|\vec{k}|}$.

Lemma 2.11. *For each $\vec{k}$,*

$$|w_{-1,-m}^{(\vec{k})} - \hat{w}_{-1,-\hat{m}}^{(\vec{k})}| = \hat{O}(\varepsilon^n) \text{ on } \mathcal{M}_{\delta_0} \times T^{\alpha}_{n,m,\hat{m}}.$$

Note that Proposition 2.5 in full generality does indeed follow from (6) and Lemma 2.11.

□

2.3 Asymptotic Behavior of the Entropy Rate

The parameterization of Z as a function of ε fits in the framework of [7] in a more general setting. Consequently, we have the following three propositions.

Proposition 2.12. *Assume that $\vec{p} \in \mathcal{M}_0$. For any sequence $z_{-n}^0 \in \mathcal{Z}^{n+1}$, $p(X_{-1} = \cdot\ z_{-n}^{-1})$ and $p(z_0|z_{-n}^{-1})$ are analytic around $\varepsilon = 0$. Moreover,* $\operatorname{ord}(p(z_0|z_{-n}^{-1})) \le O_{\max}$.

Proof. Analyticity of $p(X_{-1} = \cdot\ |z_{-n}^{-1})$ follows from Proposition 2.4 in [7]. It then follows from $p(z_0|z_{-n}^{-1}) = p(X_{-1} = \cdot\ |z_{-n}^{-1})\Omega_{z_0}\mathbf{1}$ and the fact that any row sum of Ω_{z_0} is non-zero when $\varepsilon > 0$ that $p(z_0|z_{-n}^{-1})$ is analytic with $\operatorname{ord}(p(z_0|z_{-n}^{-1})) \le O_{\max}$. □

Proposition 2.13. *(see Proposition 2.7 in [7]) Assume that $\vec{p} \in \mathcal{M}_0$. For two fixed hidden Markov chain sequences $z_{-m}^0, \hat{z}_{-\hat{m}}^0$ such that*

$$z_{-n}^0 = \hat{z}_{-n}^0, \qquad \operatorname{ord}(p(z_{-n}^{-1}|z_{-m}^{-n-1})),\ \ \operatorname{ord}(p(\hat{z}_{-n}^{-1}|\hat{z}_{-\hat{m}}^{-n-1})) \le k$$

for some $n \le m, \hat{m}$ and some k, we have for j with $0 \le j \le n - 4k - 1$,

$$p^{(j)}(z_0|z_{-m}^{-1})(0) = p^{(j)}(\hat{z}_0|\hat{z}_{-\hat{m}}^{-1})(0),$$

where the derivatives are taken with respect to ε.

Remark 2.14. It follows from Proposition 2.13 that for any α-typical sequence z_{-n}^{-1} with α small enough and n large enough, $\operatorname{ord}(p(z_0|z_{-n}^{-1})) = \operatorname{ord}(p(z_0|z_{-n-1}^{-1}))$

Proposition 2.15. *(see Theorem 2.8 in [7]) Assume that $\vec{p} \in \mathcal{M}_0$. For any $k \ge 0$,*

$$H(Z) = H(Z)|_{\varepsilon=0} + \sum_{j=1}^{k} g_j \varepsilon^j + \sum_{j=1}^{k+1} f_j \varepsilon^j \log \varepsilon + O(\varepsilon^{k+1}), \tag{12}$$

where f_j's and g_j's depend on Π (but not on ε), the transition probability matrix of X.

For any $\delta > 0$, consider a first-order Markov chain $X \in \mathcal{M}_\delta$ with transition probability matrix Π (note that X is necessarily mixing). We will need the following complexified version of Π.

Let $\Pi^{\mathbb{C}}$ denote a complex "transition probability matrix" obtained by perturbing all entries of Π to complex numbers, while satisfying $\sum_y \Pi_{xy}^{\mathbb{C}} = 1$ for all x in $\mathcal{X}$. Then through solving the following system of equations

$$\pi^{\mathbb{C}}\Pi^{\mathbb{C}} = \pi^{\mathbb{C}}, \qquad \sum_y \pi_y^{\mathbb{C}} = 1,$$

one can obtain a complex "stationary probability" $\pi^{\mathbb{C}}$, which is uniquely defined if the perturbation of Π is small enough. It then follows that under a complex perturbation of Π, for any Markov chain sequence x_{-n}^0, one can obtain a complex version of $p(x_{-n}^0)$ through complexifying all terms in the following expression:

$$p(x_{-n}^0) = \pi_{x_{-n}} \Pi_{x_{-n},x_{-n+1}} \cdots \Pi_{x_{-1},x_0},$$

namely,

$$p^{\mathbb{C}}(x_{-n}^0) = \pi_{x_{-n}}^{\mathbb{C}} \Pi_{x_{-n},x_{-n+1}}^{\mathbb{C}} \cdots \Pi_{x_{-1},x_0}^{\mathbb{C}};$$

in particular, the joint probability vector $\vec{p}$ can be complexified to $\vec{p}^{\mathbb{C}}$ as well. We then use $\mathcal{M}_\delta^{\mathbb{C}}(\eta)$, $\eta > 0$, to denote the η-perturbed complex version of $\mathcal{M}_\delta$; more precisely,

$$\mathcal{M}_\delta^{\mathbb{C}}(\eta) = \{(\vec{p}^{\mathbb{C}}(w_{-1}^0) : w_{-1}^0 \in \mathcal{S}_2) : \ |\vec{p}^{\mathbb{C}} - \vec{p}| \leq \eta \text{ for some } \vec{p} \in \mathcal{M}_\delta\},$$

which is well-defined if η is small enough. Furthermore, together with a small complex perturbation of ε, one can obtain a well-defined complex version $p^{\mathbb{C}}(z_{-n}^0)$ of $p(z_{-n}^0)$ through complexifying (1) and (2).

Using the same argument as in Lemma 2.3 and applying the triangle inequality to the absolute value of (3), we have

Lemma 2.16. *For any $\delta > 0$, there exists $\eta > 0$ such that for any fixed $0 < \alpha < 1$,*

$$\sum_{z_{-n}^{-1}:\ \mathrm{ord}\,(p^{\mathbb{C}}(z_{-n}^{-1})) \geq \alpha n} |p^{\mathbb{C}}(z_{-n}^{-1})| = \hat{O}(|\varepsilon|^n) \ \text{ on } \mathcal{M}_\delta^{\mathbb{C}}(\eta).$$

We will also need the following result, which may be well-known. We give a proof for completeness.

Lemma 2.17. *Fix $\varepsilon_0 > 0$. As n tends to infinity, $H_n(Z)$ converges to $H(Z)$ uniformly over all $(\vec{p}, \varepsilon) \in \mathcal{M} \times [0, \varepsilon_0]$.*

Proof. Let $\tilde{H}_n(Z) = H(Z_0 | Z_{-n}^{-1}, X_{-n})$ and fix $(\vec{p}, \varepsilon) \in \mathcal{M} \times [0, \varepsilon_0]$. By Theorem 4.4.1 of [4], we have for any n

$$\tilde{H}_n(Z) \leq H(Z) \leq H_n(Z), \tag{13}$$

and

$$\lim_{n \to \infty} \tilde{H}_n(Z) = H(Z) = \lim_{n \to \infty} H_n(Z). \tag{14}$$

Moreover, $H_n(Z)$ is monotonically decreasing in n, and $\tilde{H}_n(Z)$ is monotonically increasing in n. It then follows from (13) and (14) that, for any $\delta > 0$, there exists n_0 such that

$$0 \leq H_{n_0}(Z) - \tilde{H}_{n_0}(Z) \leq \frac{\delta}{2}.$$

Since $H_n(Z), \tilde{H}_n(Z)$ are continuous functions of $(\vec{p}, \varepsilon)$, there exists a neighborhood $N_{\vec{p},\varepsilon}$ of $(\vec{p}, \varepsilon)$ such that on $N_{\vec{p},\varepsilon}$

$$0 \leq H_{n_0}(Z) - \tilde{H}_{n_0}(Z) \leq \delta.$$

which, together with (13) and the monotonicity of $H_n(Z)$ and $\tilde{H}_n(Z)$, implies that for all $n \geq n_0$

$$0 \leq H_n(Z) - H(Z) \leq H_n(Z) - \tilde{H}_n(Z) \leq \delta$$

on $N_{\vec{p},\varepsilon}$. The lemma then follows from the compactness of $\mathcal{M} \times [0, \varepsilon_0]$. □

The following theorem strengthens Proposition 2.15 in the sense that it describes how the coefficients f_j's and g_j's vary with respect to the input Markov chain. We first introduce some necessary notation. We shall break $H_n(Z)$ into a sum of $G_n(Z)$ and $F_n(Z)\log(\varepsilon)$ where $G_n(Z) = G_n(\vec{p}, \varepsilon)$ and $F_n(Z) = F_n(\vec{p}, \varepsilon)$ are smooth; precisely, we have

$$H_n(Z) = G_n(\vec{p}, \varepsilon) + F_n(\vec{p}, \varepsilon)\log\varepsilon$$

where

$$F_n(\vec{p}, \varepsilon) = -\sum_{z_{-n}^0} \mathrm{ord}\,(p(z_0|z_{-n}^{-1}))p(z_{-n}^0) \tag{15}$$

and

$$G_n(\vec{p}, \varepsilon) = -\sum_{z_{-n}^0} p(z_{-n}^0)\log p^\circ(z_0|z_{-n}^{-1}), \tag{16}$$

and

$$p^\circ(z_0|z_{-n}^{-1}) = p(z_0|z_{-n}^{-1})/\varepsilon^{\mathrm{ord}\,(p(z_0|z_{-n}^{-1}))}.$$

(note that $\mathrm{ord}\,(p(z_0|z_{-n}^{-1}))$ is well-defined since $p(z_0|z_{-n}^{-1})$ is analytic with respect to ε; see Proposition 2.12; note also that $\mathrm{ord}\,(p^\circ(z_0|z_{-n}^{-1})) = 0$).

Theorem 2.18. *Let $\delta_0 > 0$. For sufficiently small $\varepsilon_0 > 0$, we have:*

1. *On $U_{\delta_0,\varepsilon_0}$, there is an analytic function $F(\vec{p}, \varepsilon)$ and a smooth (i.e., infinitely differentiable) function $G(\vec{p}, \varepsilon)$ such that*

$$H(Z(\vec{p}, \varepsilon)) = G(\vec{p}, \varepsilon) + F(\vec{p}, \varepsilon)\log\varepsilon. \tag{17}$$

   Moreover,

$$G(\vec{p}, \varepsilon) = H(Z)|_{\varepsilon=0} + \sum_{j=1}^{k} g_j(\vec{p})\varepsilon^j + O(\varepsilon^{k+1}), \qquad F(\vec{p}, \varepsilon) = \sum_{j=1}^{k} f_j(\vec{p})\varepsilon^j + O(\varepsilon^{k+1}),$$

   here f_j's and g_j's are the corresponding functions as in Proposition 2.15.

2. *Define $\hat{F}(\vec{p}, \varepsilon) = F(\vec{p}, \varepsilon)/\varepsilon$. Then $\hat{F}(\vec{p}, \varepsilon)$ is analytic on $U_{\delta_0,\varepsilon_0}$.*

3. *For any ℓ, there exists $0 < \rho < 1$ (possibly depending on ℓ) such that on $U_{\delta_0,\varepsilon_0}$*

$$|D^\ell_{\vec{p},\varepsilon} F_n(\vec{p}, \varepsilon) - D^\ell_{\vec{p},\varepsilon} F(\vec{p}, \varepsilon)| < \rho^n,$$

$$|D^\ell_{\vec{p},\varepsilon} \hat{F}_n(\vec{p}, \varepsilon) - D^\ell_{\vec{p},\varepsilon} \hat{F}(\vec{p}, \varepsilon)| < \rho^n,$$

   and

$$|D^\ell_{\vec{p},\varepsilon} G_n(\vec{p}, \varepsilon) - D^\ell_{\vec{p},\varepsilon} G(\vec{p}, \varepsilon)| < \rho^n,$$

   for sufficiently large n.

Proof. **Part 1.** Recall that

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

We now define

$$H_n^\alpha(Z) = -\sum_{z_{-n}^{-1} \in T_n^\alpha,\, z_0} p(z_{-n}^0) \log p(z_0|z_{-n}^{-1});$$

here recall that T_n^α denotes the set of all α-typical $\mathcal{Z}$-sequences with length n. It follows from a compactness argument as in Lemma 2.17 that $H_n(Z)$ uniformly converges to $H(Z)$ on the parameter space $U_{\delta_0,\varepsilon_0}$ for any positive ε_0; applying Lemma 2.3, we deduce that $H_n^\alpha(Z)$ uniformly converges to $H(Z)$ on $U_{\delta_0,\varepsilon_0}$ as well.

By Proposition 2.12, $p(z_0|z_{-n}^{-1})$ is analytic with $\mathrm{ord}\,(p(z_0|z_{-n}^{-1})) \leq O_{\max}$. It then follows that for any α with $0 < \alpha < 1$ (we will choose α to be smaller later if necessary),

$$H_n^\alpha(Z) = G_n^\alpha(\vec{p}, \varepsilon) + F_n^\alpha(\vec{p}, \varepsilon) \log \varepsilon,$$

where

$$F_n^\alpha(\vec{p}, \varepsilon) = -\sum_{z_{-n}^{-1} \in T_n^\alpha,\, z_0} \mathrm{ord}\,(p(z_0|z_{-n}^{-1}))p(z_{-n}^0),$$

and

$$G_n^\alpha(\vec{p}, \varepsilon) = -\sum_{z_{-n}^{-1} \in T_n^\alpha,\, z_0} p(z_{-n}^0) \log p^\circ(z_0|z_{-n}^{-1}).$$

The idea of the proof is as follows. We first show that $F_n^\alpha(\vec{p}, \varepsilon)$ uniformly converges to a real analytic function $F(\vec{p}, \varepsilon)$. We then prove that $G_n^\alpha(\vec{p}, \varepsilon)$ and its derivatives with respect to $(\vec{p}, \varepsilon)$ also uniformly converge to a smooth function $G(\vec{p}, \varepsilon)$. Since $H_n^\alpha(Z)$ uniformly converges to $H(Z)$, $F(\vec{p}, \varepsilon)$, $G(\vec{p}, \varepsilon)$ satisfy (17). The "Moreover" part then immediately follows by equating (12) and (17) to compare the coefficients.

We now show that $F_n^\alpha(\vec{p}, \varepsilon)$ uniformly converges to a real analytic function $F(\vec{p}, \varepsilon)$. Note that

$$|F_n^\alpha(\vec{p}, \varepsilon) - F_{n+1}^\alpha(\vec{p}, \varepsilon)| = \left| \sum_{z_{-n}^{-1} \in T_n^\alpha,\, z_0} \mathrm{ord}\,(p(z_0|z_{-n}^{-1}))p(z_{-n}^0) - \sum_{z_{-n-1}^{-1} \in T_{n+1}^\alpha,\, z_0} \mathrm{ord}\,(p(z_0|z_{-n-1}^{-1}))p(z_{-n-1}^0) \right|$$

$$= \left| \left(\sum_{z_{-n}^{-1} \in T_n^\alpha, z_{-n-1}^{-1} \in T_{n+1}^\alpha,\, z_0} + \sum_{z_{-n}^{-1} \in T_n^\alpha, z_{-n-1}^{-1} \notin T_{n+1}^\alpha,\, z_0} \right) \mathrm{ord}\,(p(z_0|z_{-n}^{-1}))p(z_{-n-1}^0) \right.$$

$$\left. - \left(\sum_{z_{-n}^{-1} \in T_n^\alpha,\, z_{-n-1}^{-1} \in T_{n+1}^\alpha,\, z_0} + \sum_{z_{-n}^{-1} \notin T_n^\alpha,\, z_{-n-1}^{-1} \in T_{n+1}^\alpha,\, z_0} \right) \mathrm{ord}\,(p(z_0|z_{-n-1}^{-1}))p(z_{-n-1}^0) \right|.$$

By Remark 2.14, we have

$$|F_n^\alpha(\vec{p}, \varepsilon) - F_{n+1}^\alpha(\vec{p}, \varepsilon)| = \left| \sum_{z_{-n}^{-1} \in T_n^\alpha,\, z_{-n-1}^{-1} \notin T_{n+1}^\alpha,\, z_0} \mathrm{ord}\,(p(z_0|z_{-n}^{-1}))p(z_{-n-1}^0) \right.$$

$$- \sum_{z_{-n}^{-1} \notin T_n^\alpha,\, z_{-n-1}^{-1} \in T_{n+1}^\alpha,\, z_0} \text{ord}\,(p(z_0|z_{-n-1}^{-1}))p(z_{-n-1}^0) \Bigg| .$$

Applying Lemma 2.3, we have

$$|F_n^\alpha(\vec{p}, \varepsilon) - F_{n+1}^\alpha(\vec{p}, \varepsilon)| = \hat{O}(\varepsilon^n) \text{ on } \mathcal{M}_{\delta_0}, \tag{18}$$

which implies that there exists $\varepsilon_0 > 0$ such that $F_n^\alpha(\vec{p}, \varepsilon)$ is exponentially Cauchy (i.e., the difference between two successive terms in the sequence is exponentially small) and thus uniformly converges on $U_{\delta_0, \varepsilon_0}$ to a continuous function $F(\vec{p}, \varepsilon)$.

Let $F_n^{\alpha, \mathbb{C}}(\vec{p}, \varepsilon)$ denote the complexified $F_n^\alpha(\vec{p}, \varepsilon)$ on $(\vec{p}, \varepsilon)$ with $\vec{p} \in \mathcal{M}_{\delta_0}^{\mathbb{C}}(\eta_0)$ and $|\varepsilon| \leq \varepsilon_0$. Then, using Lemma 2.16 and a similar argument as above, we can prove that

$$|F_n^{\alpha, \mathbb{C}}(\vec{p}, \varepsilon) - F_{n+1}^{\alpha, \mathbb{C}}(\vec{p}, \varepsilon)| = \hat{O}(|\varepsilon|^n) \text{ on } \mathcal{M}_{\delta_0}^{\mathbb{C}}(\eta_0), \tag{19}$$

and hence for a complex analytic function $F^{\mathbb{C}}(\vec{p}, \varepsilon)$ (which is necessarily the complexified version of $F(\vec{p}, \varepsilon)$)

$$|F_n^{\alpha, \mathbb{C}}(\vec{p}, \varepsilon) - F^{\mathbb{C}}(\vec{p}, \varepsilon)| = \hat{O}(|\varepsilon|^n) \text{ on } \mathcal{M}_{\delta_0}^{\mathbb{C}}(\eta_0). \tag{20}$$

In other words, for some $\eta_0, \varepsilon_0 > 0$, $F_n^{\alpha, \mathbb{C}}(\vec{p}, \varepsilon)$ is exponentially Cauchy and thus uniformly converges to $F^{\mathbb{C}}(\vec{p}, \varepsilon)$ on all $(\vec{p}, \varepsilon)$ with $\vec{p} \in \mathcal{M}_{\delta_0}^{\mathbb{C}}(\eta_0)$ and $|\varepsilon| \leq \varepsilon_0$. Therefore, $F(\vec{p}, \varepsilon)$ is analytic with respect to $(\vec{p}, \varepsilon)$ on $U_{\delta_0, \varepsilon_0}$.

We now prove that $G_n^\alpha(\vec{p}, \varepsilon)$ and its derivatives with respect to $(\vec{p}, \varepsilon)$ uniformly converge to a smooth function $G^\alpha(\vec{p}, \varepsilon)$ and its derivatives.

Although the convergence of $G_n^\alpha(\vec{p}, \varepsilon)$ and its derivatives can be proven through the same argument at once, we first prove the convergence of $G_n^\alpha(\vec{p}, \varepsilon)$ only for illustrative purposes.

For any $\alpha, \beta > 0$, we have

$$|\log \alpha - \log \beta| \leq \max\{|(\alpha - \beta)/\beta|, |(\alpha - \beta)/\alpha|\}. \tag{21}$$

Note that the following is contained in Proposition 2.5 ($\ell = 0$)

$$|p^\circ(z_0|z_{-n}^{-1}) - p^\circ(z_0|z_{-n-1}^{-1})| = \hat{O}(\varepsilon^n) \text{ on } \mathcal{M}_{\delta_0} \times T_{n,n,n+1}^\alpha. \tag{22}$$

One further checks that by Proposition 2.12, there exists a positive constant C such that for ε small enough and for any sequence z_{-n}^{-1},

$$p(z_0|z_{-n}^{-1}) \geq C\varepsilon^{O_{\max}},$$

and thus,

$$p^\circ(z_0|z_{-n}^{-1}) \geq C\varepsilon^{O_{\max}}. \tag{23}$$

Using (21), (22), (23) and Lemma 2.3, we have

$$|G_n^\alpha(\vec{p}, \varepsilon) - G_{n+1}^\alpha(\vec{p}, \varepsilon)| = \left| \sum_{z_{-n}^{-1} \in T_n^\alpha,\, z_0} p(z_{-n}^0) \log p^\circ(z_0|z_{-n}^{-1}) - \sum_{z_{-n-1}^{-1} \in T_{n+1}^\alpha,\, z_0} p(z_{-n-1}^0) \log p^\circ(z_0|z_{-n-1}^{-1}) \right|$$

$$= \left| \left(\sum_{z_{-n}^{-1} \in T_n^{\alpha},\, z_{-n-1}^{-1} \in T_{n+1}^{\alpha},\, z_0} + \sum_{z_{-n}^{-1} \in T_n^{\alpha},\, z_{-n-1}^{-1} \notin T_{n+1}^{\alpha},\, z_0} \right) p(z_{-n-1}^0) \log p^{\circ}(z_0|z_{-n}^{-1}) \right.$$

$$\left. - \left(\sum_{z_{-n}^{-1} \in T_n^{\alpha},\, z_{-n-1}^{-1} \in T_{n+1}^{\alpha},\, z_0} + \sum_{z_{-n}^{-1} \notin T_n^{\alpha},\, z_{-n-1}^{-1} \in T_{n+1}^{\alpha},\, z_0} \right) p(z_{-n-1}^0) \log p^{\circ}(z_0|z_{-n-1}^{-1}) \right|$$

$$\leq \left| \sum_{z_{-n}^{-1} \in T_n^{\alpha},\, z_{-n-1}^{-1} \in T_{n+1}^{\alpha},\, z_0} p(z_{-n-1}^0)(\log p^{\circ}(z_0|z_{-n}^{-1}) - \log p^{\circ}(z_0|z_{-n-1}^{-1})) \right|$$

$$+ \left| \sum_{z_{-n}^{-1} \in T_n^{\alpha},\, z_{-n-1}^{-1} \notin T_{n+1}^{\alpha},\, z_0} p(z_{-n-1}^0) \log p^{\circ}(z_0|z_{-n}^{-1}) \right| + \left| \sum_{z_{-n}^{-1} \notin T_n^{\alpha},\, z_{-n-1}^{-1} \in T_{n+1}^{\alpha},\, z_0} p(z_{-n-1}^0) \log p^{\circ}(z_0|z_{-n-1}^{-1}) \right|$$

$$\leq \sum_{z_{-n}^{-1} \in T_n^{\alpha},\, z_{-n-1}^{-1} \in T_{n+1}^{\alpha},\, z_0} p(z_{-n-1}^0) \max \left\{ \left| \frac{p^{\circ}(z_0|z_{-n}^{-1}) - p^{\circ}(z_0|z_{-n-1}^{-1})}{p^{\circ}(z_0|z_{-n-1}^{-1})} \right|, \left| \frac{p^{\circ}(z_0|z_{-n}^{-1}) - p^{\circ}(z_0|z_{-n-1}^{-1})}{p^{\circ}(z_0|z_{-n}^{-1})} \right| \right\}$$

$$+ \left| \sum_{z_{-n}^{-1} \in T_n^{\alpha},\, z_{-n-1}^{-1} \notin T_{n+1}^{\alpha},\, z_0} p(z_{-n-1}^0) \log p^{\circ}(z_0|z_{-n}^{-1}) \right| + \left| \sum_{z_{-n}^{-1} \notin T_n^{\alpha},\, z_{-n-1}^{-1} \in T_{n+1}^{\alpha},\, z_0} p(z_{-n-1}^0) \log p^{\circ}(z_0|z_{-n-1}^{-1}) \right| = \hat{O}(\varepsilon^n) \text{ on } \mathcal{M}_{\delta_0}, \tag{24}$$

which implies that there exists $\varepsilon_0 > 0$ such that $G_n^{\alpha}(\vec{p}, \varepsilon)$ uniformly converges on $U_{\delta_0, \varepsilon_0}$. With this, the existence of $G(\vec{p}, \varepsilon)$ immediately follows.

Applying the multivariate Faa Di Bruno formula [2, 10] to the function $f(y) = \log y$, we have for $\vec{\ell}$ with $|\vec{\ell}| \neq 0$,

$$f(y)^{(\vec{\ell})} = \sum D(\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_k)(y^{(\vec{a}_1)}/y)(y^{(\vec{a}_2)}/y) \cdots (y^{(\vec{a}_k)}/y),$$

where the summation is over the set of unordered sequences of non-negative vectors $\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_k$ with $\vec{a}_1 + \vec{a}_2 + \cdots + \vec{a}_k = \vec{\ell}$ and $D(\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_k)$ is the corresponding coefficient. Then for any $\vec{m}$, applying the multivariate Leibniz rule, we have

$$(G_n^{\alpha})^{(\vec{m})}(\vec{p}, \varepsilon) = - \sum_{z_{-n}^{-1} \in T_n^{\alpha},\, z_0} \sum_{\vec{\ell} \preceq \vec{m}} C_{\vec{m}}^{\vec{\ell}} p^{(\vec{m}-\vec{\ell})}(z_{-n}^0)(\log p^{\circ}(z_0|z_{-n}^{-1}))^{(\vec{\ell})}$$

$$= - \sum_{z_{-n}^{-1} \in T_n^{\alpha},\, z_0} \sum_{|\vec{\ell}| \neq 0, \vec{\ell} \preceq \vec{m}} \sum_{\vec{a}_1 + \vec{a}_2 + \cdots + \vec{a}_k = \vec{\ell}} C_{\vec{m}}^{\vec{\ell}} D(\vec{a}_1, \ldots, \vec{a}_k) p^{(\vec{m}-\vec{\ell})}(z_{-n}^0) \frac{p^{\circ}(z_0|z_{-n}^{-1})^{(\vec{a}_1)}}{p^{\circ}(z_0|z_{-n}^{-1})} \cdots \frac{p^{\circ}(z_0|z_{-n}^{-1})^{(\vec{a}_k)}}{p^{\circ}(z_0|z_{-n}^{-1})}$$

$$- \sum_{z_{-n}^{-1} \in T_n^{\alpha},\, z_0} p^{(\vec{m})}(z_{-n}^0) \log p^{\circ}(z_0|z_{-n}^{-1}). \tag{25}$$

We tackle the last term of (25) first. Using (21) and (22) and with a parallel argument obtained through replacing $p(z_{-n}^0), p(z_{-n-1}^0)$ in (24) by $p^{(\vec{m})}(z_{-n}^0), p^{(\vec{m})}(z_{-n-1}^0)$, respectively, we can show that

$$\left|\sum_{z_{-n}^{-1}\in T_n^{\alpha},\, z_0} p^{(\vec m)}(z_{-n}^0)\log p^{\circ}(z_0|z_{-n}^{-1}) - \sum_{z_{-n-1}^{-1}\in T_{n+1}^{\alpha},\, z_0} p^{(\vec m)}(z_{-n-1}^0)\log p^{\circ}(z_0|z_{-n-1}^{-1})\right| = \hat O(\varepsilon^n) \text{ on } \mathcal{M}_{\delta_0}\times T^{\alpha}_{n,n,n+1},$$

where we used the fact that for any z_{-n}^0 and $\vec m$, $p^{(\vec m)}(z_{-n}^0)/p(z_{-n}^0)$ is $O(n^{|\vec m|}/\varepsilon^{|\vec m|})$ (see (40)). And using the identity

$$\alpha_1\alpha_2\cdots\alpha_n-\beta_1\beta_2\cdots\beta_n=(\alpha_1-\beta_1)\alpha_2\cdots\alpha_n+\beta_1(\alpha_2-\beta_2)\alpha_3\cdots\alpha_n+\cdots+\beta_1\cdots\beta_{n-1}(\alpha_n-\beta_n),$$

we have

$$\left|\frac{p^{\circ}(z_0|z_{-n}^{-1})^{(\vec a_1)}}{p^{\circ}(z_0|z_{-n}^{-1})}\cdots\frac{p^{\circ}(z_0|z_{-n}^{-1})^{(\vec a_k)}}{p^{\circ}(z_0|z_{-n}^{-1})}-\frac{p^{\circ}(z_0|z_{-n-1}^{-1})^{(\vec a_1)}}{p^{\circ}(z_0|z_{-n-1}^{-1})}\cdots\frac{p^{\circ}(z_0|z_{-n-1}^{-1})^{(\vec a_k)}}{p^{\circ}(z_0|z_{-n-1}^{-1})}\right|$$

$$\le\left|\left(\frac{p^{\circ}(z_0|z_{-n}^{-1})^{(\vec a_1)}}{p^{\circ}(z_0|z_{-n}^{-1})}-\frac{p^{\circ}(z_0|z_{-n-1}^{-1})^{(\vec a_1)}}{p^{\circ}(z_0|z_{-n-1}^{-1})}\right)\frac{p^{\circ}(z_0|z_{-n}^{-1})^{(\vec a_2)}}{p^{\circ}(z_0|z_{-n}^{-1})}\cdots\frac{p^{\circ}(z_0|z_{-n}^{-1})^{(\vec a_k)}}{p^{\circ}(z_0|z_{-n}^{-1})}\right|$$

$$+\left|\frac{p^{\circ}(z_0|z_{-n-1}^{-1})^{(\vec a_1)}}{p^{\circ}(z_0|z_{-n-1}^{-1})}\left(\frac{p^{\circ}(z_0|z_{-n}^{-1})^{(\vec a_2)}}{p^{\circ}(z_0|z_{-n}^{-1})}-\frac{p^{\circ}(z_0|z_{-n-1}^{-1})^{(\vec a_2)}}{p^{\circ}(z_0|z_{-n-1}^{-1})}\right)\frac{p^{\circ}(z_0|z_{-n}^{-1})^{(\vec a_3)}}{p^{\circ}(z_0|z_{-n}^{-1})}\cdots\frac{p^{\circ}(z_0|z_{-n}^{-1})^{(\vec a_k)}}{p^{\circ}(z_0|z_{-n}^{-1})}\right|+\cdots$$

$$+\left|\frac{p^{\circ}(z_0|z_{-n-1}^{-1})^{(\vec a_1)}}{p^{\circ}(z_0|z_{-n-1}^{-1})}\cdots\frac{p^{\circ}(z_0|z_{-n-1}^{-1})^{(\vec a_{k-1})}}{p^{\circ}(z_0|z_{-n-1}^{-1})}\left(\frac{p^{\circ}(z_0|z_{-n}^{-1})^{(\vec a_k)}}{p^{\circ}(z_0|z_{-n}^{-1})}-\frac{p^{\circ}(z_0|z_{-n-1}^{-1})^{(\vec a_k)}}{p^{\circ}(z_0|z_{-n-1}^{-1})}\right)\right|.$$

Now applying the inequality

$$\left|\frac{\beta_1}{\alpha_1}-\frac{\beta_2}{\alpha_2}\right|=\left|\frac{\beta_1}{\alpha_1}-\frac{\beta_1}{\alpha_2}+\frac{\beta_1}{\alpha_2}-\frac{\beta_2}{\alpha_2}\right|\le|\beta_1/(\alpha_1\alpha_2)||\alpha_1-\alpha_2|+|1/\alpha_2||\beta_1-\beta_2|,$$

we have for any $1\le i\le k$,

$$\left|\frac{p^{\circ}(z_0|z_{-n}^{-1})^{(\vec a_i)}}{p^{\circ}(z_0|z_{-n}^{-1})}-\frac{p^{\circ}(z_0|z_{-n-1}^{-1})^{(\vec a_i)}}{p^{\circ}(z_0|z_{-n-1}^{-1})}\right|$$

$$\le\left|\frac{p^{\circ}(z_0|z_{-n}^{-1})^{(\vec a_i)}}{p^{\circ}(z_0|z_{-n}^{-1})p^{\circ}(z_0|z_{-n-1}^{-1})}\right||p^{\circ}(z_0|z_{-n}^{-1})-p^{\circ}(z_0|z_{-n-1}^{-1})|+\left|\frac{1}{p^{\circ}(z_0|z_{-n-1}^{-1})}\right||p^{\circ}(z_0|z_{-n}^{-1})^{(\vec a_i)}-p^{\circ}(z_0|z_{-n-1}^{-1})^{(\vec a_i)}|.$$

It follows from multivariate Leibniz rule and Lemma 2.10 that there exists a positive constant $C_{\vec a}$ such that for sufficiently small ε and and for any $z_{-n}^{-1}\in\mathcal{Z}^n$,

$$|p(z_0|z_{-n}^{-1})^{(\vec a)}|=|(w_{-1,-n}\Omega_{z_0}\mathbf{1})^{(\vec a)}|\le n^{|\vec a|}C_{\vec a}/\varepsilon^{|\vec a|},\tag{26}$$

and furthermore there exists a positive constant $C_{\vec a}^{\circ}$ such that for sufficiently small ε and for any $z_{-n}^{-1}\in\mathcal{Z}^n$,

$$p^{\circ}(z_0|z_{-n}^{-1})^{(\vec a)}\le n^{|\vec a|}C_{\vec a}^{\circ}/\varepsilon^{|\vec a|+O_{\max}}.\tag{27}$$

Combining (23), (25), (26), (27) and Proposition 2.5 gives us

$$|(G_n^{\alpha})^{(\vec m)}(\vec p,\varepsilon)-(G_{n+1}^{\alpha})^{(\vec m)}(\vec p,\varepsilon)|=\hat O(\varepsilon^n)\text{ on }\mathcal{M}_{\delta_0}.\tag{28}$$

This implies that there exists $\varepsilon_0 > 0$ such that $G_n^\alpha(\vec{p}, \varepsilon)$ and its derivatives with respect to $(\vec{p}, \varepsilon)$ uniformly converge on $U_{\delta_0, \varepsilon_0}$ to a smooth function $G(\vec{p}, \varepsilon)$ and correspondingly its derivatives (here, by Remark 2.2, ε_0 does not depend on $\vec{m}$).

Part 2. This statement immediately follows from the analyticity of $F(\vec{p}, \varepsilon)$ and the fact that $\mathrm{ord}\,(F(\vec{p}, \varepsilon)) \geq 1$.

Part 3. Note that

$$F_n(\vec{p}, \varepsilon) - F_n^\alpha(\vec{p}, \varepsilon) = - \sum_{z_{-n}^{-1} \notin T_n^\alpha,\, z_0} \mathrm{ord}\,(p(z_0 | z_{-n}^{-1})) p(z_{-n}^0).$$

Applying the multivariate Leibniz rule, then by Proposition 2.12, (26), (40) and Lemma 2.3, we have for any ℓ,

$$\left| D_{\vec{p},\varepsilon}^\ell F_n(\vec{p}, \varepsilon) - D_{\vec{p},\varepsilon}^\ell F_n^\alpha(\vec{p}, \varepsilon) \right| = \left| \sum_{z_{-n}^{-1} \notin T_n^\alpha,\, z_0} \mathrm{ord}\,(p(z_0 | z_{-n}^{-1})) D_{\vec{p},\varepsilon}^\ell (p(z_0 | z_{-n}^{-1}) p(z_{-n}^{-1})) \right| = \hat{O}(\varepsilon^n) \text{ on } \mathcal{M}_{\delta_0}. \tag{29}$$

It follows from (19), (20) and the Cauchy integral formula (p. 157 of [3]) that

$$\left| D_{\vec{p},\varepsilon}^\ell F_{n+1}^\alpha(\vec{p}, \varepsilon) - D_{\vec{p},\varepsilon}^\ell F_n^\alpha(\vec{p}, \varepsilon) \right| = \hat{O}(\varepsilon^n) \text{ on } \mathcal{M}_{\delta_0},$$

and

$$\left| D_{\vec{p},\varepsilon}^\ell F_n^\alpha(\vec{p}, \varepsilon) - D_{\vec{p},\varepsilon}^\ell F(\vec{p}, \varepsilon) \right| = \hat{O}(\varepsilon^n) \text{ on } \mathcal{M}_{\delta_0},$$

which, together with (29), implies that

$$\left| D_{\vec{p},\varepsilon}^\ell F_n(\vec{p}, \varepsilon) - D_{\vec{p},\varepsilon}^\ell F(\vec{p}, \varepsilon) \right| = \hat{O}(\varepsilon^n) \text{ on } \mathcal{M}_{\delta_0}.$$

It then follows that there exists $\varepsilon_0 > 0$ such that, for any ℓ, there exists $0 < \rho < 1$ (here ρ depends on ℓ) such that on $U_{\delta_0, \varepsilon_0}$

$$|D_{\vec{p},\varepsilon}^\ell F_n(\vec{p}, \varepsilon) - D_{\vec{p},\varepsilon}^\ell F(\vec{p}, \varepsilon)| < \rho^n,$$

and further

$$|D_{\vec{p},\varepsilon}^\ell \hat{F}_n(\vec{p}, \varepsilon) - D_{\vec{p},\varepsilon}^\ell \hat{F}(\vec{p}, \varepsilon)| < \rho^n,$$

for sufficiently large n.

Similarly note that

$$G_n(\vec{p}, \varepsilon) - G_n^\alpha(\vec{p}, \varepsilon) = - \sum_{z_{-n}^{-1} \notin T_n^\alpha,\ z_0} p(z_{-n}^0) \log p^\circ(z_0 | z_{-n}^{-1}).$$

Then by (26), (27), (23) and Lemma 2.3, we have for any ℓ,

$$\left| D_{\vec{p},\varepsilon}^\ell G_n(\vec{p}, \varepsilon) - D_{\vec{p},\varepsilon}^\ell G_n^\alpha(\vec{p}, \varepsilon) \right|$$

$$= \left| \sum_{z_{-n}^{-1} \notin T_n^\alpha,\, z_0} D_{\vec{p},\varepsilon}^\ell (p(z_{-n}^{-1}) p(z_0 | z_{-n}^{-1}) \log p^\circ(z_0 | z_{-n}^{-1})) \right| = \hat{O}(\varepsilon^n) \text{ on } \mathcal{M}_{\delta_0},$$

which, together with (28), implies that there exists $\varepsilon_0 > 0$ such that for any ℓ, there exists $0 < \rho < 1$ such that on $U_{\delta_0,\varepsilon_0}$

$$|D^{\ell}_{\vec{p},\varepsilon} G_n(\vec{p},\varepsilon) - D^{\ell}_{\vec{p},\varepsilon} G(\vec{p},\varepsilon)| < \rho^n,$$

for sufficiently large n.

□

3 Concavity of the Mutual Information

Recall that we are considering a parameterized family of finite-state memoryless channels with inputs restricted to a mixing finite-type constraint $\mathcal{S}$. Again for simplicity, we assume that $\mathcal{S}$ has order 1.

For parameter value ε, the channel capacity is the supremum of the mutual information of $Z(X,\varepsilon)$ and X over all stationary input processes X such that $A(X) \subseteq \mathcal{S}$. Here, we use only first-order Markov input processes. While this will typically not achieve the true capacity, one can approach the true capacity by using Markov input processes of higher order. As in Section 2, we identify a first-order input Markov process X with its joint probability vector $\vec{p} = \vec{p}_X \in \mathcal{M}$, and we write $Z = Z(\vec{p},\varepsilon)$, thereby sometimes notationally suppressing dependence on X and ε.

Precisely, the *first-order capacity* is

$$C^1(\varepsilon) = \sup_{\vec{p}\in\mathcal{M}} I(Z;X) = \sup_{\vec{p}\in\mathcal{M}} (H(Z) - H(Z|X)) \tag{30}$$

and its n-th approximation is

$$C^1_n(\varepsilon) = \sup_{\vec{p}\in\mathcal{M}} I_n(Z;X) = \sup_{\vec{p}\in\mathcal{M}} \left(H_n(Z) - \frac{1}{n+1} H(Z^0_{-n}|X^0_{-n}) \right). \tag{31}$$

As mentioned earlier, since the channel is memoryless, the second terms in (30) and (31) both reduce to $H(Z_0|X_0)$, which can be written as:

$$-\sum_{x\in\mathcal{X},\, z\in\mathcal{Z}} p(x)p(z|x)\log p(z|x).$$

Note that this expression is a linear function of $\vec{p}$ and for all $\vec{p}$ it vanishes when $\varepsilon = 0$. Using this and the fact that for a mixing finite-type constraint there is a unique Markov chain of maximal entropy supported on the constraint (see [15] or Section 13.3 of [11]), one can show that for sufficiently small $\varepsilon_1 > 0, \delta_1 > 0$ and all $0 \le \varepsilon \le \varepsilon_1$,

$$C^1_n(\varepsilon) = \sup_{\vec{p}\in\mathcal{M}_{\delta_1}} (H_n(Z) - H(Z_0|X_0)) > \sup_{\vec{p}\in\mathcal{M}\backslash\mathcal{M}_{\delta_1}} (H_n(Z) - H(Z_0|X_0)), \tag{32}$$

$$C^1(\varepsilon) = \sup_{\vec{p}\in\mathcal{M}_{\delta_1}} (H(Z) - H(Z_0|X_0)) > \sup_{\vec{p}\in\mathcal{M}\backslash\mathcal{M}_{\delta_1}} (H(Z) - H(Z_0|X_0)). \tag{33}$$

For instance, to see (33), we argue as follows.

First, it follows from the fact that for any n, $H_n(Z)$ is a continuous function of $(\vec{p}, \varepsilon)$ and uniform convergence (Lemma 2.17) that $H(Z)$ is a continuous function of $(\vec{p}, \varepsilon)$ (the continuity was also noted in [8]). Let $X_{\max}$ denote the unique Markov chain of maximal entropy for the constraint. It is well known that $X_{\max} \in \mathcal{M}_0$ and $H(X_{\max}) > 0$ (see Section 13.3 of [11]). Thus, there exists $\delta_0 > 0$ and $0 < \eta < 1$ such that

$$\sup_{\vec{p} \in \mathcal{M} \backslash \mathcal{M}_{\delta_0}} H(Z)|_{\varepsilon=0} = \sup_{\vec{p} \in \mathcal{M} \backslash \mathcal{M}_{\delta_0}} H(X) < \eta H(X_{\max});$$

here, note that $H(Z)|_{\varepsilon=0} = H(X)$, since we assumed that there is a one-to-one mapping from $\mathcal{X}$ into $\mathcal{Z}$, $z = z(x)$, such that for any $x \in \mathcal{X}$, $p(z(x)|x)(0) = 1$.

Thus, there exists $\varepsilon_0 > 0$ such that for all $0 \leq \varepsilon \leq \varepsilon_0$,

$$\sup_{\vec{p} \in \mathcal{M} \backslash \mathcal{M}_{\delta_0}} H(Z) < (1/2 + \eta/2) H(X_{\max})$$

and

$$\sup_{\vec{p} \in \mathcal{M}_{\delta_0}} H(Z) > (1/2 + \eta/2) H(X_{\max}).$$

This gives the inequality (33) without the conditional entropy term. In order to incorporate the latter, notice that $H(Z_0|X_0)$ vanishes at $\varepsilon = 0$ and simply replace δ_0 and ε_0 with appropriate smaller numbers δ_1 and ε_1.

Theorem 3.1. *Let δ_1 be as in (32) and (33). For any $0 < \delta_0 < \delta_1$, there exist $\varepsilon_0 > 0$ such that for all $0 \leq \varepsilon \leq \varepsilon_0$,*

1. *the functions $I_n(Z(\vec{p}, \varepsilon); X(\vec{p}))$ and $I(Z(\vec{p}, \varepsilon); X(\vec{p}))$ are strictly concave on $\mathcal{M}_{\delta_0}$, with unique maximizing $\vec{p}_n(\varepsilon)$ and $\vec{p}_\infty(\varepsilon)$;*

2. *the functions $I_n(Z(\vec{p}, \varepsilon); X(\vec{p}))$ and $I(Z(\vec{p}, \varepsilon); X(\vec{p}))$ uniquely achieve their maxima on all of $\mathcal{M}$ at $\vec{p}_n(\varepsilon)$ and $\vec{p}_\infty(\varepsilon)$;*

3. *there exists $0 < \rho < 1$ such that*

$$|\vec{p}_n(\varepsilon) - \vec{p}_\infty(\varepsilon)| \leq \rho^n.$$

Proof. **Part 1:** Recall that

$$H(Z(\vec{p}, \varepsilon)) = G(\vec{p}, \varepsilon) + \hat{F}(\vec{p}, \varepsilon)(\varepsilon \log \varepsilon).$$

By Part 1 of Theorem 2.18, for any given $\delta_0 > 0$, there exists $\varepsilon_0 > 0$, such that $G(\vec{p}, \varepsilon)$ and $\hat{F}(\vec{p}, \varepsilon)$ are smooth on $U_{\delta_0, \varepsilon_0}$, and moreover

$$\lim_{\varepsilon \to 0} D^2_{\vec{p}} G(\vec{p}, \varepsilon) = D^2_{\vec{p}} G(\vec{p}, 0), \qquad \lim_{\varepsilon \to 0} D^2_{\vec{p}} \hat{F}(\vec{p}, \varepsilon) = D^2_{\vec{p}} \hat{F}(\vec{p}, 0),$$

uniformly on $\vec{p} \in \mathcal{M}_{\delta_0}$. Thus,

$$\lim_{\varepsilon \to 0} D^2_{\vec{p}} H(Z(\vec{p}, \varepsilon)) = D^2_{\vec{p}} G(\vec{p}, 0) = D^2_{\vec{p}} H(Z(\vec{p}, 0)), \qquad (34)$$

again uniformly on $\mathcal{M}_{\delta_0}$. Since $D^2_{\vec{p}} H(Z(\vec{p}, 0))$ is negative definite on $\mathcal{M}_{\delta_0}$ (see [6]), it follows from (34) that for sufficiently small ε, $D^2_{\vec{p}} H(Z(\vec{p}, \varepsilon))$ is also negative definite on $\mathcal{M}_{\delta_0}$, and thus $H(Z(\vec{p}, \varepsilon))$ is also strictly concave on $\mathcal{M}_{\delta_0}$.

Since for all $\varepsilon \geq 0$, $H(Z_0|X_0)$ is a linear function of $\vec{p}$, $I(Z(\vec{p}, \varepsilon); X(\vec{p}))$ is strictly concave on $\mathcal{M}_{\delta_0}$. This establishes Part 1 for $I(Z(\vec{p}, \varepsilon); X(\vec{p}))$. By Part 3 of Theorem 2.18, for sufficiently large n ($n \geq N_1$), we obtain the same result (with the same ε_0 and δ_0) for $I_n(Z(\vec{p}, \varepsilon); X(\vec{p}))$. For each $1 \leq n < N_1$, one can easily establish strict concavity on $U_{\delta^{(n)}, \varepsilon^{(n)}}$ for some $\delta^{(n)}, \varepsilon^{(n)} > 0$, and then replace δ_0 by $\min\{\delta_0, \delta^{(n)}\}$ and replace ε_0 by $\min\{\varepsilon_0, \varepsilon^{(n)}\}$.

Part 2: Choose $\delta_0 < \delta_1$ and further $\varepsilon_0 < \varepsilon_1$, where ε_1 is as in (32) and (33). Part 2 then follows from Part 1 and (32) and (33).

Part 3: For notational simplicity, for fixed $0 \leq \varepsilon \leq \varepsilon_0$, we rewrite $I(Z(\vec{p}, \varepsilon); X(\vec{p}))$, $I_n(Z(\vec{p}, \varepsilon); X(\vec{p}))$ as function $f(\vec{p}), f_n(\vec{p})$, respectively. By the Taylor formula with remainder, there exist $\eta_1, \eta_2 \in \mathcal{M}_{\delta_0}$ such that

$$\begin{aligned} f(\vec{p}_n(\varepsilon)) &= f(\vec{p}_\infty(\varepsilon)) + D_{\vec{p}} f(\vec{p}_\infty(\varepsilon))(\vec{p}_n(\varepsilon) - \vec{p}_\infty(\varepsilon)) \\ &\quad + (\vec{p}_n(\varepsilon) - \vec{p}_\infty(\varepsilon))^T D^2_{\vec{p}} f(\eta_1)(\vec{p}_n(\varepsilon) - \vec{p}_\infty(\varepsilon)), \qquad (35) \\ f_n(\vec{p}_\infty(\varepsilon)) &= f_n(\vec{p}_n(\varepsilon)) + D_{\vec{p}} f_n(\vec{p}_n(\varepsilon))(\vec{p}_\infty(\varepsilon) - \vec{p}_n(\varepsilon)) \\ &\quad + (\vec{p}_n(\varepsilon) - \vec{p}_\infty(\varepsilon))^T D^2_{\vec{p}} f_n(\eta_2)(\vec{p}_n(\varepsilon) - \vec{p}_\infty(\varepsilon)), \qquad (36) \end{aligned}$$

here the superscript T denotes the transpose.

By Part 2 of Theorem 3.1

$$D_{\vec{p}} f(\vec{p}_\infty(\varepsilon)) = 0, \quad D_{\vec{p}} f_n(\vec{p}_n(\varepsilon)) = 0. \qquad (37)$$

By Part 3 of Theorem 2.18, with $\ell = 0$, there exists $0 < \rho_0 < 1$ such that

$$|f(\vec{p}_\infty(\varepsilon)) - f_n(\vec{p}_\infty(\varepsilon))| \leq \rho_0^n, \quad |f(\vec{p}_n(\varepsilon)) - f_n(\vec{p}_n(\varepsilon))| \leq \rho_0^n. \qquad (38)$$

Combining (35), (36), (37), (38), we have

$$|(\vec{p}_n(\varepsilon) - \vec{p}_\infty(\varepsilon))^T (D^2_{\vec{p}} f(\eta_1) + D^2_{\vec{p}} f_n(\eta_2))(\vec{p}_n(\varepsilon) - \vec{p}_\infty(\varepsilon))| \leq 2\rho_0^n.$$

Since f and f_n are strictly concave on $\mathcal{M}_{\delta_0}$ (see Part 1), $D^2_{\vec{p}} f(\eta_1), D^2_{\vec{p}} f_n(\eta_2)$ are both negative definite. Thus there exists some positive constant K such that

$$K|\vec{p}_n(\varepsilon) - \vec{p}_\infty(\varepsilon)|^2 \leq 2\rho_0^n,$$

which implies the existence of ρ.

□

Example 3.2. Consider Example 2.1. For sufficiently small ε and p bounded away from 0 and 1, Part 1 of Theorem 2.18 gives an expression for $H(Z(\vec{p}, \varepsilon))$ and Part 1 of Theorem 3.1 shows that $I(Z(\vec{p}, \varepsilon))$ is strictly concave and thus has negative second derivative. In this case, the results boil down to the strict concavity of the binary entropy function; that is, when $\varepsilon = 0$, $H(Z) = H(X) = -p \log p - (1-p) \log(1-p)$, and one computes with the second derivative with respect to p

$$H''(Z)|_{\varepsilon=0} = -\frac{1}{p} - \frac{1}{1-p} \leq -4.$$

So, there is an ε_0 such that whenever $0 \leq \varepsilon \leq \varepsilon_0$, $H''(Z) < 0$.

Appendices

A Proof of Lemma 2.10

To illustrate the idea behind the proof, we first prove the lemma for $|\vec{k}| = 1$. Recall that

$$w_{i,-m} = p(X_i = \cdot\,|z_{-m}^i) = \frac{p(X_i = \cdot\,, z_{-m}^i)}{p(z_{-m}^i)}.$$

Let q be a component of $\vec{q} = (\vec{p}, \varepsilon)$. Then,

$$\left|\frac{\partial}{\partial q}\left(\frac{p(x_i, z_{-m}^i)}{p(z_{-m}^i)}\right)\right| = \left|\frac{p(x_i, z_{-m}^i)}{p(z_{-m}^i)}\left(\frac{\frac{\partial}{\partial q}p(x_i, z_{-m}^i)}{p(x_i, z_{-m}^i)} - \frac{\frac{\partial}{\partial q}p(z_{-m}^i)}{p(z_{-m}^i)}\right)\right|$$

$$\leq \left|\frac{p(X_i = \cdot\,, z_{-m}^i)}{p(z_{-m}^i)}\right|\left(\left|\frac{\frac{\partial}{\partial q}p(X_i = \cdot\,, z_{-m}^i)}{p(X_i = \cdot\,, z_{-m}^i)}\right| + \left|\frac{\frac{\partial}{\partial q}p(z_{-m}^i)}{p(z_{-m}^i)}\right|\right).$$

We first consider the partial derivative with respect to ε, i.e., $q = \varepsilon$. Since the first factor is bounded above by 1, it suffices to show that both terms of the second factor are $mO(1/\varepsilon)$ (applying the argument to both z_{-m}^i and $\hat{z}_{-\hat{m}}^i$ and recalling that $n \leq m, \hat{m} \leq 2n$). We will prove this only for $\left|\frac{\partial}{\partial \varepsilon}p(z_{-m}^i)/p(z_{-m}^i)\right|$, with the proof for the other term being similar. Now

$$p(z_{-m}^i) = \sum_{x_{-m}^{-1}} g(x_{-m}^{-1}), \tag{39}$$

where

$$g(x_{-m}^{-1}) = p(x_{-m}) \prod_{j=-m}^{i-1} p(x_{j+1}|x_j) \prod_{j=-m}^{i} p(z_j|x_j).$$

Clearly, $\frac{\partial}{\partial \varepsilon}p(z_j|x_j)/p(z_j|x_j)$ is $O(1/\varepsilon)$. Thus each $\frac{\partial}{\partial \varepsilon}g(x_{-m}^{-1})$ is $mO(1/\varepsilon)$. Each $g(x_{-m}^{-1})$ is lower bounded by a positive constant, uniformly over all $p \in \mathcal{M}_{\delta_0}$. Thus, each $\frac{\partial}{\partial \varepsilon}g(x_{-m}^{-1})/g(x_{-m}^{-1})$ is $mO(1/\varepsilon)$. It then follows from (39) that $\frac{\partial}{\partial q}p(z_{-m}^i)/p(z_{-m}^i) = mO(1/\varepsilon)$, as desired.

For the partial derivatives with respect to $\vec{p}$, we observe that $\frac{\partial}{\partial q}p(x_{-m})/p(x_{-m})$ and $\frac{\partial}{\partial q}p(x_{j+1}|x_j)/p(x_{j+1}|x_j)$ (here, q is a component of $\vec{p}$) are $O(1)$, with uniform constant over all $p \in \mathcal{M}_{\delta_0}$. We then immediately establish the lemma for $|\vec{k}| = 1$.

We now prove the lemma for a generic $\vec{k}$.

Applying the multivariate Faa Di Bruno formula (for the derivatives of a composite function) [2, 10] to the function $f(y) = 1/y$ (here, y is a function), we have for $\vec{\ell}$ with $|\vec{\ell}| \neq 0$,

$$f(y)^{(\vec{\ell})} = \sum D(\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_t)(1/y)(y^{(\vec{a}_1)}/y)(y^{(\vec{a}_2)}/y)\cdots(y^{(\vec{a}_t)}/y),$$

where the summation is over the set of unordered sequences of non-negative vectors $\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_t$ with $\vec{a}_1 + \vec{a}_2 + \cdots + \vec{a}_t = \vec{\ell}$ and $D(\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_t)$ is the corresponding coefficient. For any $\vec{\ell}$, define $\vec{\ell}! = \prod_{i=1}^{|\mathcal{S}_2|+1} l_i!$; and for any $\vec{\ell} \preceq \vec{k}$ (every component of $\vec{\ell}$ is less than or equal to the corresponding one of $\vec{k}$), define $C_{\vec{k}}^{\vec{\ell}} = \vec{k}!/(\vec{\ell}!(\vec{k}-\vec{\ell})!)$. Then for any $\vec{k}$, applying the multivariate Leibniz rule, we have

$$\left(\frac{p(x_i, z_{-m}^i)}{p(z_{-m}^i)}\right)^{(\vec{k})} = \sum_{\vec{\ell} \preceq \vec{k}} C_{\vec{k}}^{\vec{\ell}} (p(x_i, z_{-m}^i))^{(\vec{k}-\vec{\ell})} (1/p(z_{-m}^i))^{(\vec{\ell})}$$

$$= \sum_{\vec{\ell} \preceq \vec{k}} \sum_{\vec{a}_1+\vec{a}_2+\cdots+\vec{a}_t=\vec{\ell}} C_{\vec{k}}^{\vec{\ell}} D(\vec{a}_1, \ldots, \vec{a}_t) \frac{p(x_i, z_{-m}^i)}{p(z_{-m}^i)} \frac{p(x_i, z_{-m}^i)^{(\vec{k}-\vec{\ell})}}{p(x_i, z_{-m}^i)} \frac{p(z_{-m}^i)^{(\vec{a}_1)}}{p(z_{-m}^i)} \cdots \frac{p(z_{-m}^i)^{(\vec{a}_t)}}{p(z_{-m}^i)}.$$

Then, similarly as above, one can show that

$$p(z_{-m}^i)^{(\vec{a})}/p(z_{-m}^i) = m^{|\vec{a}|} O(1/\varepsilon^{|\vec{a}|}), \qquad p(x_i, z_{-m}^i)^{(\vec{a})}/p(x_i, z_{-m}^i) = m^{|\vec{a}|} O(1/\varepsilon^{|\vec{a}|}), \tag{40}$$

which implies that there is a positive constant $C_{|\vec{k}|}$ such that

$$|w_{i,-m}^{(\vec{k})}| \leq n^{|\vec{k}|} C_{|\vec{k}|}/\varepsilon^{|\vec{k}|}.$$

Obviously, the same argument can be applied to upper bound $|\hat{w}_{i,-\hat{m}}^{(\vec{k})}|$.

B Proof of Lemma 2.11

We first prove this for $|\vec{k}| = 1$. Again, let q be a component of $\vec{q} = (\vec{p}, \varepsilon)$. Then, for $i = -1, -2, \ldots, -n_0$, we have

$$\frac{\partial}{\partial q} w_{(i+1)N-1,-m} = \frac{\partial f_{[z]_i}}{\partial w}(\vec{q}, w_{iN-1,-m}) \frac{\partial}{\partial q} w_{iN-1,-m} + \frac{\partial f_{[z]_i}}{\partial q}(\vec{q}, w_{iN-1,-m}), \tag{41}$$

and

$$\frac{\partial}{\partial q} \hat{w}_{(i+1)N-1,-\hat{m}} = \frac{\partial f_{[z]_i}}{\partial w}(\vec{q}, \hat{w}_{iN-1,-\hat{m}}) \frac{\partial}{\partial q} \hat{w}_{iN-1,-\hat{m}} + \frac{\partial f_{[z]_i}}{\partial q}(\vec{q}, \hat{w}_{iN-1,-\hat{m}}). \tag{42}$$

Taking the difference, we then have

$$\frac{\partial}{\partial q} w_{(i+1)N-1,-m} - \frac{\partial}{\partial q} \hat{w}_{(i+1)N-1,-\hat{m}} = \frac{\partial f_{[z]_i}}{\partial q}(\vec{q}, w_{iN-1,-m}) - \frac{\partial f_{[z]_i}}{\partial q}(\vec{q}, \hat{w}_{iN-1,-\hat{m}})$$

$$+ \frac{\partial f_{[z]_i}}{\partial w}(\vec{q}, w_{iN-1,-m}) \frac{\partial}{\partial q} w_{iN-1,-m} - \frac{\partial f_{[z]_i}}{\partial w}(\vec{q}, \hat{w}_{iN-1,-\hat{m}}) \frac{\partial}{\partial q} \hat{w}_{iN-1,-\hat{m}}$$

$$= \left(\frac{\partial f_{[z]_i}}{\partial q}(\vec{q}, w_{iN-1,-m}) - \frac{\partial f_{[z]_i}}{\partial q}(\vec{q}, \hat{w}_{iN-1,-\hat{m}}) \right)$$

$$+ \left(\frac{\partial f_{[z]_i}}{\partial w}(\vec{q}, w_{iN-1,-m}) \frac{\partial}{\partial q} w_{iN-1,-m} - \frac{\partial f_{[z]_i}}{\partial w}(q, \hat{w}_{iN-1,-\hat{m}}) \frac{\partial}{\partial q} w_{iN-1,-m} \right)$$

$$+ \left(\frac{\partial f_{[z]_i}}{\partial w}(\vec{q}, \hat{w}_{iN-1,-\hat{m}}) \frac{\partial}{\partial q} w_{iN-1,-m} - \frac{\partial f_{[z]_i}}{\partial w}(\vec{q}, \hat{w}_{iN-1,-\hat{m}}) \frac{\partial}{\partial q} \hat{w}_{iN-1,-\hat{m}} \right).$$

This last expression is the sum of three terms, which we will refer to as $T_{i,1}$, $T_{i,2}$ and $T_{i,3}$.

From Lemma 2.6, one checks that for all $[z]_i \in \mathcal{Z}^N$, $w \in \mathcal{W}$ and $\vec{q} \in U_{\delta_0,\varepsilon_0}$,

$$\left|\frac{\partial^2 f_{[z]_i}}{\partial \vec{q}\partial w}(\vec{q}, w)\right|, \left|\frac{\partial^2 f_{[z]_i}}{\partial w\partial w}(\vec{q}, w)\right| \le C/\varepsilon^{4NO_{\max}}.$$

(Here, we remark that there are many different constants in this proof, which we will often refer to using the same notation C, making sure that the dependence of these constants on various parameters is clear.) It then follows from the mean value theorem that for each $i = -1, -2, \ldots, -n_0$

$$T_{i,1} \le (C/\varepsilon^{4NO_{\max}})|w_{iN-1,-m} - \hat{w}_{iN-1,-\hat{m}}|.$$

By the mean value theorem and Lemma 2.10,

$$T_{i,2} \le (C/\varepsilon^{4NO_{\max}})(nC_1/\varepsilon)|w_{iN-1,-m} - \hat{w}_{iN-1,-\hat{m}}|.$$

And finally

$$T_{i,3} \le \left\|\frac{\partial f_{[z]_i}}{\partial w}(\vec{q}, \hat{w}_{iN-1,-\hat{m}})\right\| \cdot |\frac{\partial}{\partial q} w_{iN-1,-m} - \frac{\partial}{\partial q}\hat{w}_{iN-1,-\hat{m}}|.$$

Thus,

$$|\frac{\partial}{\partial q} w_{(i+1)N-1,-m} - \frac{\partial}{\partial q}\hat{w}_{(i+1)N-1,-\hat{m}}| \le \left\|\frac{\partial f_{[z]_i}}{\partial w}(\vec{q}, \hat{w}_{iN-1,-\hat{m}})\right\| \cdot |\frac{\partial}{\partial q} w_{iN-1,-m} - \frac{\partial}{\partial q}\hat{w}_{iN-1,-\hat{m}}|$$
$$+(1 + nC_1/\varepsilon)C\varepsilon^{-4NO_{\max}}|w_{iN-1,-m} - \hat{w}_{iN-1,-\hat{m}}|.$$

Iteratively apply this inequality to obtain

$$\begin{aligned}|\frac{\partial}{\partial q} w_{-1,-m} - \frac{\partial}{\partial q}\hat{w}_{-1,-\hat{m}}| \le{}& \prod_{i=-n_0}^{-1} \left\|\frac{\partial f_{[z]_i}}{\partial w}(\vec{q}, \hat{w}_{iN-1,-\hat{m}})\right\| \cdot |\frac{\partial}{\partial q} w_{-n_0N-1,-m} - \frac{\partial}{\partial q}\hat{w}_{-n_0N-1,-\hat{m}}| \\
&+ \prod_{i=-n_0+1}^{-1} \left\|\frac{\partial f_{[z]_i}}{\partial w}(\vec{q}, \hat{w}_{iN-1,-\hat{m}})\right\| (1 + nC_1/\varepsilon)C\varepsilon^{-4NO_{\max}}|w_{-n_0N-1,-m} - \hat{w}_{-n_0N-1,-\hat{m}}| \\
&+ \cdots + \prod_{i=-j}^{-1} \left\|\frac{\partial f_{[z]_i}}{\partial w}(\vec{q}, \hat{w}_{iN-1,-\hat{m}})\right\| (1 + nC_1/\varepsilon)C\varepsilon^{-4NO_{\max}}|w_{(-j-1)N-1,-m} - \hat{w}_{(-j-1)N-1,-\hat{m}}| + \\
&+ \cdots + \left\|\frac{\partial f_{[z]_{-1}}}{\partial w}(\vec{q}, \hat{w}_{-N-1,-\hat{m}})\right\| (1 + nC_1/\varepsilon)C\varepsilon^{-4NO_{\max}}|w_{-2N-1,-m} - \hat{w}_{-2N-1,-\hat{m}}| \\
&+ (1 + nC_1/\varepsilon)C\varepsilon^{-4NO_{\max}}|w_{-N-1,-m} - \hat{w}_{-N-1,-\hat{m}}|. \end{aligned} \tag{43}$$

Now, applying the mean value theorem, we deduce that there exist ξ_i, $-n_0 \le i \le -j-2$ (here ξ_i is a convex combination of $w_{-iN-1,-m}$ and $\hat{w}_{-iN-1,-\hat{m}}$) such that

$$|w_{(-j-1)N-1,-m} - \hat{w}_{(-j-1)N-1,-\hat{m}}| = |f_{[z]_{-n_0}^{-j-2}}(w_{-n_0N-1,-m}) - f_{[z]_{-n_0}^{-j-2}}(\hat{w}_{-n_0N-1,-\hat{m}})|$$

$$\leq \prod_{i=-n_0}^{-j-2} \|D_w f_{[z]_i}(\xi_i)\| \cdot |w_{-n_0N-1,-m} - \hat{w}_{-n_0N-1,-\hat{m}}|.$$

Then, recall that an α-typical sequence z_{-n}^{-1} breaks at most $2\alpha n$ times. Thus there are at least $(1-2\alpha)n$ i's where we can use the estimate (9) and at most $2\alpha n$ i's where we can only use the weaker estimates (8). Similar to the derivation of (10), with Remark 2.2, we derive that for any $\alpha < \alpha_0$, every term on the right-hand side of (43) is $\hat{O}(\varepsilon^n)$ on $\mathcal{M}_{\delta_0} \times T^{\alpha}_{n,m,\hat{m}}$ (we use Lemma 2.10 to upper bound the first term). Again, with Remark 2.2, we conclude that

$$\left| \frac{\partial w_{-1,-m}}{\partial \vec{q}} - \frac{\partial \hat{w}_{-1,-\hat{m}}}{\partial \vec{q}} \right| = \hat{O}(\varepsilon^n) \text{ on } \mathcal{M}_{\delta_0} \times T^{\alpha}_{n,m,\hat{m}},$$

which, by (6), implies the proposition for $|\vec{k}| = 1$, as desired.

The proof of the lemma for a generic $\vec{k}$ is rather similar, however very tedious. We next briefly illustrate the idea of the proof. Note that (compare the following two equations with (41), (42) for $|\vec{k}| = 1$)

$$w^{(\vec{k})}_{(i+1)N-1,-m} = \frac{\partial f_{[z]_i}}{\partial w}(\vec{q}, w_{iN-1,-m}) w^{(\vec{k})}_{iN-1,-m} + \text{ others}$$

and

$$\hat{w}^{(\vec{k})}_{(i+1)N-1,-\hat{m}} = \frac{\partial f_{[z]_i}}{\partial w}(\vec{q}, \hat{w}_{iN-1,-\hat{m}}) \hat{w}^{(\vec{k})}_{iN-1,-\hat{m}} + \text{ others},$$

where the first "others" is a linear combination of terms taking the following forms (below, t can be 0, which corresponds to the partial derivatives of f with respect to the first argument $\vec{q}$):

$$f^{(\vec{k}')}_{[z]_i}(\vec{q}, w_{iN-1,-m}) w^{(\vec{a}_1)}_{iN-1,-m} \cdots w^{(\vec{a}_t)}_{iN-1,-m},$$

and the second "others" is a linear combination of terms taking the following forms:

$$f^{(\vec{k}')}_{[z]_i}(\vec{q}, \hat{w}_{iN-1,-\hat{m}}) \hat{w}^{(\vec{a}_1)}_{iN-1,-\hat{m}} \cdots \hat{w}^{(\vec{a}_t)}_{iN-1,-\hat{m}},$$

here $\vec{k}' \preceq \vec{k}$, $t \leq |\vec{k}|$ and $|\vec{a}_i| < |\vec{k}|$ for all i. Using Lemma 2.10 and the fact that there exists a constant C (by Lemma 2.6) such that

$$|f^{(\vec{k}')}_{[z]_i}(\vec{q}, w_{iN-1,-m})| \leq C/\varepsilon^{4NO_{\max}|\vec{k}'|},$$

we then can establish (compare the following inequality with (43) for $|\vec{k}| = 1$)

$$\left| w^{(k)}_{(i+1)N-1,-m} - \hat{w}^{(k)}_{(i+1)N-1,-\hat{m}} \right| \leq \left\| \frac{\partial f_{[z]_i}}{\partial w}(\vec{q}, \hat{w}_{iN-1,-\hat{m}}) \right\| \cdot \left| w^{\vec{k}}_{iN-1,-m} - \hat{w}^{(\vec{k})}_{iN-1,-\hat{m}} \right| + \text{ others},$$

where "others" is the sum of finitely many terms, each of which takes the following form (see the j-th term of (43) for $|\vec{k}| = 1$)

$$n^{D_{\vec{k}'}} O(1/\varepsilon^{D_{\vec{k}'}}) \prod_{i=-j}^{-1} \left\| \frac{\partial f_{[z]_i}}{\partial w}(\vec{q}, \hat{w}_{iN-1,-\hat{m}}) \right\| \cdot \left| w^{(\vec{a})}_{(-j-1)N-1,-m} - \hat{w}^{(\vec{a})}_{(-j-1)N-1,-\hat{m}} \right|, \qquad (44)$$

where $|\vec{a}| < |\vec{k}|$, $D_{\vec{k}'}$ is a constant dependent on $\vec{k}'$. Then inductively, one can use the similar approach to establish that (44) is $\hat{O}(\varepsilon^n)$ on $\mathcal{M}_{\delta_0} \times T^{\alpha}_{n,m,\hat{m}}$, which implies the lemma for a generic $\vec{k}$.

Acknowledgement

We are grateful to the anonymous referee and especially the associate editor Pascal Vontobel for numerous comments that helped greatly to improve this paper.

References

[1] J. Chen and P. H. Siegel. Markov processes asymptotically achieve the capacity of finite-state intersymbol interference channels. *IEEE Trans. Inf. Theory*, Vol. 54, No. 3, 2008, pp. 1295-1303.

[2] G. Constantine and T. Savits. A multivariate Faa Di Bruno formula with applications. *Transactions of the American Mathematical Society*, Vol. 348, No. 2, 1996, pp. 503-520.

[3] J. Brown and R. Churchill. *Complex Variables and Applications.* McGraw-Hill, 7th edition, 2004.

[4] T. Cover and J. Thomas. *Elements of Information Theory.* Wiley Series in Telecommunications, 1991.

[5] G. Han and B. Marcus. Analyticity of entropy rate of hidden Markov chains. *IEEE Trans. Inf. Theory*, Vol. 52, No. 12, 2006, pp. 5251-5266.

[6] G. Han and B. Marcus. Asymptotics of input-constrained binary symmetric channel capacity. *Annals of Applied Probability*, Vol. 19, No. 3, 2009, pp. 1063-1091.

[7] G. Han and B. Marcus Asymptotics of entropy rate in special families of hidden Markov chains. *IEEE Trans. Inf. Theory*, Vol. 56, No. 3, 2010, pp. 1287-1295.

[8] T. Holliday, A. Goldsmith, and P. Glynn. Capacity of finite state channels based on Lyapunov exponents of random matrices. *IEEE Trans. Inf. Theory*, Vol. 52, No. 8, 2006, pp. 3509 - 3532.

[9] P. Jacquet, G. Seroussi, and W. Szpankowski. On the entropy of a hidden Markov process. *Theoretical Computer Science*, Vol. 395, 2008, pp. 203-219.

[10] R. Leipnik and T. Reid. Multivariable Faa 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.

[11] D. Lind and B. Marcus. *An introduction to symbolic dynamics and coding.* Cambridge University Press, 1995 (reprinted 1999).

[12] 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.

[13] E. Ordentlich and T. Weissman. New bounds on the entropy rate of hidden Markov processes. *IEEE Inf. Theory Workshop*, San Antonio, Texas, 2004, pp. 117 - 122.

[14] E. Ordentlich and T. Weissman. Bounds on the entropy rate of binary hidden Markov processes. in *Entropy of Hidden Markov Processes and Connections to Dynamical Systems*, London Math. Soc. Lecture Notes, Vol. 385, Cambridge University Press, 2011, pp. 117 - 171.

[15] W. Parry. Intrinsic Markov chains. *Trans. Amer. Math. Soc.* Vol. 112, 1964, pp. 55-66.

[16] Y. Peres and A. Quas. Entropy rate for hidden Markov chains with rare transistions. *Entropy of Hidden Markov Processes and Connections to Dynamical Systems*, London Math. Soc. Lecture Notes, Vol. 385, Cambridge University Press, 2011, pp. 172 - 178.

[17] P. O. Vontobel, A. Kavcic, D. Arnold and H.-A. Loeliger. A generalization of the Blahut-Arimoto algorithm to finite-state channels. *IEEE Trans. Inf. Theory*, vol. 54, No. 5, 2008, pp. 1887-1918,

[18] E. Zehavi and J. Wolf. On runlength codes. *IEEE Trans. Inf. Theory*, Vol. 34, No. 1, 1988, pp. 45-54.

[19] O. Zuk, E. Domany, I. Kantor and M. Aizenman. From finite-system entropy to entropy rate for a hidden Markov process. *IEEE Signal Processing Letters*, Vol. 13, No. 9, 2006, pp. 517 - 520.