Concavity of Mutual Information Rate for Input-Restricted Finite-State 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

July 23, 2010

Abstract

We consider a finite-state memoryless channel with i.i.d. channel state and the input Markov process supported on a mixing finite-type constraint. We discuss the asymptotic behavior of entropy rate of the output hidden Markov chain and deduce that the mutual information rate of such a channel is concave with respect to the parameters of the input Markov processes at high signal-to-noise ratio. In principle, the concavity result enables good numerical approximation of the maximum mutual information rate and capacity of such a channel.

1 Channel Model

In this paper, we show that for certain input-restricted finite-state memoryless channels, the mutual information rate, at high SNR, is effectively a concave function of Markov input processes of a given order. While not directly addressed here, the goal is to help estimate the maximum of this function and ultimately the capacity of such channels (see, for example, the algorithm of Vontobel, et. al. [11]).

Our approach depends heavily on results regarding asymptotics and smoothness of entropy rate in special parameterized families of hidden Markov chains, such as those developed in [5], [9], [3], [4], and continued here.

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 $\mathcal{X}^* = \cup_n \mathcal{X}^n$. A *finite-type* constraint $\mathcal{S}$ 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 $uwv \in \mathcal{S}$.

In magnetic recording, input sequences are required to satisfy certain constraints in order to eliminate the most damaging error events [8]. 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 1's. For $\mathcal{S}(d,k)$ with $k < \infty$, a forbidden set $\mathcal{F}$ is:

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

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

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

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 $\hat{m} = \hat{m}(\mathcal{S})$ 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

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

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 $\hat{m} \le m$, and we say that X is *supported* on $\mathcal{S}$. Also, X is mixing iff 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 $m < \hat{m}$.

Now, consider a finite-state memoryless channel with finite sets of channel states $c \in \mathcal{C}$, inputs $x \in \mathcal{X}$, outputs $z \in \mathcal{Z}$ and input sequences restricted to a mixing finite-type constraint $\mathcal{S}$. The channel state process C is assumed to be i.i.d. with $P(C = c) = q_c$. 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

$$p(z|x,c) = P(Z = z|X = x, C = c).$$

We are actually interested in families of channels, as above, parameterized by $\varepsilon \ge 0$ such that for each x, c, and z, $p(z|x,c)(\varepsilon)$ is an analytic function of $\varepsilon \ge 0$. We assume that for all x, c, z, $p(z|x,c)(\varepsilon)$ is not identically 0 as a function of ε, so that for small $\varepsilon > 0$, for any input x and channel state c, by analyticity, 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 all c and x, $p(z(x)|x,c)(0) = 1$; so, ε can be regarded as noise, and $z(x)$ is the noiseless output corresponding to input x. 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 context.

Prominent examples of such families include input-restricted versions of the binary symmetric channel with crossover probability ε (denoted by BSC(ε)), the binary erasure channel with erasure rate ε (denoted by BEC(ε)), and some special Gilbert-Elliott Channels, where the channel state process is a 2-state i.i.d. process, with one state acting as BSC(ε) and the other state acting as BSC($k\varepsilon$) for some fixed k; see Section 3 of [4].

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 [4], 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 Markov chain input X of a fixed order. In particular, we show that $H(Z)$ 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 near 0 for any first order X supported on $\mathcal{S}$ (in fact, $F(X,\varepsilon)$ will be analytic); 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.

In Section 3, we apply the smoothness results to show that for a mixing finite-type constraint $\mathcal{S}$ of order 1, and sufficiently small $\varepsilon_0 > 0$, for each $0 \leq \varepsilon \leq \varepsilon_0$, $I_n(Z(\varepsilon, X); X)$ and $I(Z(X,\varepsilon); X)$ are strictly concave on the set of all first order X whose non-zero transition probabilities are not "too small". This will imply 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 will also follow that $X_n(\varepsilon)$ converges exponentially to $X_\infty(\varepsilon)$ uniformly over $0 \leq \varepsilon \leq \varepsilon_0$. In principle, the concavity result enables (via any convex optimization algorithm) good numerical approximation of $X_n(\varepsilon)$ and $X_\infty(\varepsilon)$ and therefore the maximum mutual information rate over first order X. This can be generalized to m-th order Markov chains, and as $m \to \infty$, this maximum converges to channel capacity; furthermore it can be generalized to higher order constraints.

2 Asymptotics of Entropy Rate

2.1 Key ideas and lemmas

For simplicity, we consider only mixing finite-type constraints S of order 1, and correspondingly only first order input Markov processes X such that $\mathcal{A}(X) \subseteq S$ (the higher order case is easily reduced to this). For such X with transition probability matrix Π, (X, C) is also a first order Markov chain, with transition probability matrix:

$$\Omega((x, c), (y, d)) = \Pi_{x,y} q_d.$$

For any $z \in \mathcal{Z}$, define

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

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

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

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 Ω and $\mathbf{1}$ is the all 1's column vector.

For a given analytic function $f(\varepsilon)$ around $\varepsilon = 0$, let $\operatorname{ord}(f(\varepsilon))$ 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 $\operatorname{ord}(p(z|x, c))$ determine the orders $\operatorname{ord}(p(z_{-n}^0))$ and similarly orders of conditional probabilities $\operatorname{ord}(p(z_0|z_{-n}^{-1}))$.

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 $0 < p < 1$. Here there is only one channel state, and so we can suppress dependence on the channel state. 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

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

The stationary vector $\pi = (1/(p+1), p/(p+1))$, 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.

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 S is mixing, so X is completely determined by its transition probability matrix Π. For the purpose 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 $X = X(\vec{p})$.

In the following, for any parameterized sequence of functions $f_{n,\lambda}(\varepsilon)$ (ε is real or complex), 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 $0 \leq |\varepsilon| \leq \varepsilon_0$,

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

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 $|\varepsilon| \leq \varepsilon_0$, all $\lambda \in \Lambda$ and large enough n. One also checks that a term $\hat{O}(\varepsilon^n)$ is unaffected by a multiplication of a exponential function (thus polynomial function) in n and a polynomial function in $1/\varepsilon$;

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 for any $|\varepsilon| \leq \varepsilon_0$, $|g_1(n)g_2(1/\varepsilon)f_{n,\lambda}(\varepsilon)| \leq \rho^n$, for all $|\varepsilon| \leq \varepsilon_0$, 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* $0 < \alpha < 1$*,*

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

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

$$p(z_{-n}^{-1}) = \sum p(x_{-n}^{-1}, c_{-n}^{-1}) \prod_{i=-n}^{-1} p(z_i|x_i, c_i), \tag{3}$$

where the summation is over all $(x_{-n}^{-1}, c_{-n}^{-1})$. Now consider z_{-n}^{-1} with $k = \mathrm{ord}\,(p(z_{-n}^{-1})) \geq \alpha n$. One checks that for ε small enough there exists a positive constant C such that $p(z|x,c) \leq C\varepsilon$ for (x, c, z) with $\mathrm{ord}\,(p(z|x,c)) \geq 1$, and thus the term $\prod_{i=-n}^{-1} p(z_i|x_i, c_i)$ 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 $\sum_{x_{-n}^{-1}, c_{-n}^{-1}} p(x_{-n}^{-1}, c_{-n}^{-1}) = 1$, we then have, for ε small enough,

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

which immediately implies the lemma. $\square$

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

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

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,

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

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

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

By Lemma 2.3 and Lemma 2.4 means that we can focus our attention on output sequences with relatively small order. 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, $D^{\ell}_{\vec{x}} f$ denotes the ℓ-th total derivative with respect to $\vec{x}$; for instance,

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

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 $D^l_{\vec{p}} p(z_0|z_{-n}^{-1})$ or $D^l_{\vec{p},\varepsilon} p(z_0|z_{-n}^{-1})$. We shall use $|\cdot|$ to denote the Euclidean norm (of a vector or a matrix), and we shall use $\|A\|$ to denote the norm of a matrix A as a linear map under the Euclidean norm, i.e.,

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

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

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

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

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

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

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 $\mathbb{R}^{|\mathcal{X}|\cdot|\mathcal{C}|}$, i.e., the set of nonnegative vectors, which sum to 1, indexed by the joint input-state space $\mathcal{X} \times \mathcal{C}$. For any $z \in \mathcal{Z}$, Ω_z induces a mapping f_z defined on $\mathcal{W}$ by

$$f_z(w) = \frac{w\Omega_z}{w\Omega_z \mathbf{1}}. \tag{4}$$

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_M 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 $p(z|x,c)(\varepsilon)$ over all possible x, c, 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 $C_e > 0$ such that on $U_{\delta_0,\varepsilon_0}$ for all $z \in \mathcal{Z}$, $|D_w f_z| \le C_e/\varepsilon^{2O_M}$ 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 $0 \le \varepsilon \le \varepsilon_0$,

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

We then apply the quotient rule to establish the lemma. $\square$

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

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

Similar to (4), $\Omega_{z_{-N}^{-1}}$ 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

$$z_{-N}^{-1} = z(x_{-N}^{-1}) \triangleq (z(x_{-N}), z(x_{-N+1}), \cdots, z(x_{-1})).$$

Note that z_{-N}^{-1} is Z-allowed iff $\operatorname{ord}(p(z_{-N}^{-1})) = 0$.

Since Π is a primitive matrix, there exists a positive integer e such that $\Pi^e > 0$. For any $z \in \mathcal{Z}$, let I_z denote the set of indices of the columns (x, c) of Ω_z such that $z = z(x)$; note that I_z can be empty for some $z \in \mathcal{Z}$.

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 $N \ge 2eO_M$, we have*

$$\operatorname{ord}((\Omega_{z_{-N}^{-1}})(s, t_1)) = \operatorname{ord}((\Omega_{z_{-N}^{-1}})(s, t_2)),$$

for all s, and any $t_1, t_2 \in I_{z_{-1}}$, and

$$\operatorname{ord}((\Omega_{z_{-N}^{-1}})(s, t_1)) < \operatorname{ord}((\Omega_{z_{-N}^{-1}})(s, t_2)),$$

for all s, and any $t_1 \in I_{z_{-1}}$, $t_2 \notin I_{z_{-1}}$.

Proof. Let $s = (\hat{x}_{-N-1}, \hat{c}_{-N-1})$, $t = (\hat{x}_{-1}, \hat{c}_{-1}) \in \mathcal{X} \times \mathcal{C}$. Then

$$\Omega_{z_{-N}^{-1}}(s, t) = P((X_{-1}, C_{-1}) = (\hat{x}_{-1}, \hat{c}_{-1}), Z_{-N}^{-1} = z_{-N}^{-1} | (X_{-N-1}, C_{-N-1}) = (\hat{x}_{-N-1}, \hat{c}_{-N-1}))$$

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

It then follows that

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

Since

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

we have

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

where the minimization is over all sequences $(\hat{x}_{-N}^{-2}, \hat{c}_{-N}^{-2})$ 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 $\mathrm{ord}\,(p(z|x,c)) \le O_M$ that, as long as $N \ge 2eO_M$, for any fixed t and any choice of order minimizing sequence $(\hat{x}_{-N}^{-2}(t), \hat{c}_{-N}^{-2}(t))$, there exist $0 \le i_0 = i_0(t), j_0 = j_0(t) \le eO_M$ such that $z(\hat{x}_i^j(t)) = z_i^j$ if and only if $i \ge -N-1+i_0(t)$ and $j \le -1-j_0(t)$. One further checks that, for any choice of order minimizing sequences corresponding to t, $(\hat{x}_{-N}^{-2}(t), \hat{c}_{-N}^{-2}(t))$,

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

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

$\square$

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

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

Note that in the spirit of Lemma 2.7, 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.

Now fix $N \ge 2eO_M$. For any $w \in \mathcal{W}$, let $v = f_{z_{-N}^{-1}}(w)$. Note that the mapping $f_{z_{-N}^{-1}}$ implicitly depends on ε, so v is in fact a function of ε. If $z_{-N}^{-1} = z(x_{-N}^{-1}) \in \mathcal{Z}^N$ is Z-allowed, by Lemma 2.7, when $\varepsilon = 0$,

- $v_i = 0$ if and only if $i \notin I_{z_{-1}}$,

- for each $i = (x_{-1}, c_{-1}) \in I_{z_{-1}}$, $v_i = q_{c_{-1}}$, which does not depend on w.

Let $q(z) \in \mathcal{W}$ be the point defined by $q(z)_{(x,c)} = q_c$ for all (x, c) with $z(x) = z$ and 0 otherwise. If z_{-N}^{-1} is Z-allowed, then

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

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 $C_c > 0$ such that on $U_{\delta_0,\varepsilon_0}$, if z_{-N-1}^{-1} is Z-allowed, then $|D_w f_{z_{-N}^{-1}}| \le C_c \varepsilon$ 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. $\square$

2.2 Proof of Proposition 2.5

We now explain the rough idea of the proof of Proposition 2.5, for only the special case $\ell = 0$, i.e., exponential 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, $|D_w f_{z_{-n}^0}| \le (A\varepsilon^{1-B\alpha})^n$ 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 $p((x_i, c_i) = \cdot|z_{-m}^i)$, we obtain exponential convergence of the difference $|p(z_0|z_{-n}^{-1}) - p(z_0|z_{-n-1}^{-1})|$, 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 $N \ge 2eO_M$; 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" version $[z]_{-m_0}^{-1}, [\hat{z}]_{-\hat{m}_0}^{-1}$ by setting

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

Let

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

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

$$p(z_0|z_{-m}^{-1}) = w_{-1,-m}\Omega_{z_0}\mathbf{1} \tag{5}$$

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

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

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

$$w_{(i+1)N-1,-m} = f_{[z]_i}(w_{iN-1,-m}) \qquad -n_0 \le i \le -1, \tag{6}$$

starting with

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

Similarly let

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

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

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

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 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_e such that

$$|D_w f_{[z]_i}| \le C_e^N / \varepsilon^{2NO_M}, \tag{7}$$

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 $C_c > 0$ such that

$$|D_w f_{[z]_i}| \le C_c \varepsilon \tag{8}$$

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

Now, apply 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

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

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

Since z_{-n}^{-1} is α-typical, $[z]_{-n_0}^{-1}$ breaks at most $3\alpha n$ times; in other words, there are at least $(1/N - 3\alpha)n$ i's corresponding to (8) and at most $3\alpha n$ i's corresponding to (7). We then have

$$\prod_{i=-n_0}^{-1} \|D_w f_{[z]_i}(\xi_i)\| \le C_c^{(1/N-3\alpha)n} C_e^{3\alpha N n} \varepsilon^{(1/N-3\alpha-6NO_M\alpha)n}. \tag{9}$$

Let $\alpha_0 = 1/(N(3+6NO_M))$. Evidently, when $\alpha < \alpha_0$, $1/N - 3\alpha - 6NO_M\alpha$ is strictly positive, we then have

$$|w_{-1,-m} - \hat{w}_{-1,-\hat{m}}| = \hat{O}(\varepsilon^n) \text{ on } \mathcal{M}_{\delta_0} \times T^{\alpha}_{n,m,\hat{m}}. \tag{10}$$

It then follows from (5) that

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

We next show that for each $\vec{k}$, there is a positive constant $C_{|\vec{k}|}$ such that

$$|w_{i,-m}^{(\vec{k})}|, |\hat{w}_{i,-\hat{m}}^{(\vec{k})}| \le n^{|\vec{k}|} C_{|\vec{k}|}/\varepsilon^{|\vec{k}|}; \tag{11}$$

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 $n^{|\vec{k}|} C_{|\vec{k}|}$.

To illustrate the idea, we first prove (11) for $|\vec{k}| = 1$. Recall that

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

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

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

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

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 $n \le m, \hat{m} \le 2n$). 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

$$p(z_{-m}^i) = \sum g(x_{-m}^{-1}, c_{-m}^{-1}), \tag{12}$$

where

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

and the summation is over all Markov chain sequences x_{-m}^i and channel state sequences c_{-m}^i. Clearly, $\frac{\partial}{\partial \varepsilon} p(z_j|x_j, c_j)/p(z_j|x_j, c_j)$ is $O(1/\varepsilon)$. Thus each $\frac{\partial}{\partial \varepsilon} g(x_{-m}^{-1}, c_{-m}^{-1})$ is $mO(1/\varepsilon)$. Each $g(x_{-m}^{-1}, c_{-m}^{-1})$ is lower bounded by a positive constant, uniformly over all $p \in \mathcal{M}_{\delta_0}$. Thus, each $\frac{\partial}{\partial \varepsilon} g(x_{-m}^{-1}, c_{-m}^{-1})/g(x_{-m}^{-1}, c_{-m}^{-1})$ is $mO(1/\varepsilon)$. It then follows from (12) 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 (11) for $|\vec{k}| = 1$.

We now prove (11) for a generic $\vec{k}$.

Apply the multivariate Faa Di Bruno formula (for the derivatives of a composite function) [1, 6] to the function $f(y) = 1/y$ (here, y is a function), we have for $\vec{l}$ with $|\vec{l}| \neq 0$,

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

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

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

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

Then, similarly as above, one can show that

$$p(z_{-m}^i)^{(\vec{a})}/p(z_{-m}^i), \qquad p((x_i, c_i), z_{-m}^i)^{(\vec{a})}/p((x_i, c_i), z_{-m}^i) = m^{|\vec{a}|} O(1/\varepsilon^{|\vec{a}|}), \tag{13}$$

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 $|\hat{w}_{i,-\hat{m}}^{(\vec{k})}|$.

We next prove that, for each $\vec{k}$,

$$|w_{-1,-m}^{(\vec{k})} - \hat{w}_{-1,-\hat{m}}^{(\vec{k})}| = \hat{O}(\varepsilon^n) \text{ on } \mathcal{M}_{\delta_0} \times T_{n,m,\hat{m}}^{\alpha}. \tag{14}$$

Proposition 2.5 will then follow from (5).

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

$$\frac{\partial}{\partial q} w_{(i+1)N-1,-m} = \frac{\partial f_{[z]_i}}{\partial w}(\vec{q}, w_{iN-1,-m}) \frac{\partial}{\partial q} w_{iN-1,-m} + \frac{\partial f_{[z]_i}}{\partial q}(\vec{q}, w_{iN-1,-m}), \tag{15}$$

and

$$\frac{\partial}{\partial q} \hat{w}_{(i+1)N-1,-\hat{m}} = \frac{\partial f_{[z]_i}}{\partial w}(\vec{q}, \hat{w}_{iN-1,-\hat{m}}) \frac{\partial}{\partial q} \hat{w}_{iN-1,-\hat{m}} + \frac{\partial f_{[z]_i}}{\partial q}(\vec{q}, \hat{w}_{iN-1,-\hat{m}}). \tag{16}$$

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_1, T_2 and T_3.

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

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

(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 $i = -1, -2, \cdots, -n_0$

$$T_1 \leq (C/\varepsilon^{4NO_M})|w_{iN-1,-m} - \hat{w}_{iN-1,-\hat{m}}|.$$

By the mean value theorem and (11),

$$T_2 \leq (C/\varepsilon^{4NO_M})(nC_1/\varepsilon)|w_{iN-1,-m} - \hat{w}_{iN-1,-\hat{m}}|.$$

And finally

$$T_3 \leq \left\|\frac{\partial f_{[z]_i}}{\partial w}(\vec{q}, \hat{w}_{iN-1,-\hat{m}})\right\| |\frac{\partial}{\partial q}w_{iN-1,-m} - \frac{\partial}{\partial q}\hat{w}_{iN-1,-\hat{m}}|.$$

Thus,

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

$$+(1 + nC_1/\varepsilon)C\varepsilon^{-4NO_M}|w_{iN-1,-m} - \hat{w}_{iN-1,-\hat{m}}|.$$

Iteratively apply this inequality to obtain

$$|\frac{\partial}{\partial q}w_{-1,-m} - \frac{\partial}{\partial q}\hat{w}_{-1,-\hat{m}}| \leq \prod_{i=-n_0}^{-1}\left\|\frac{\partial f_{[z]_i}}{\partial w}(\vec{q}, \hat{w}_{iN-1,-\hat{m}})\right\| |\frac{\partial}{\partial q}w_{-n_0N-1,-m} - \frac{\partial}{\partial q}\hat{w}_{-n_0N-1,-\hat{m}}|$$

$$+\prod_{i=-n_0+1}^{-1}\left\|\frac{\partial f_{[z]_i}}{\partial w}(\vec{q}, \hat{w}_{iN-1,-\hat{m}})\right\| (1 + nC_1/\varepsilon)C\varepsilon^{-4NO_M}|w_{-n_0N-1,-m} - \hat{w}_{-n_0N-1,-\hat{m}}|$$

$$+\cdots+\prod_{i=-j}^{-1}\left\|\frac{\partial f_{[z]_i}}{\partial w}(\vec{q}, \hat{w}_{iN-1,-\hat{m}})\right\| (1 + nC_1/\varepsilon)C\varepsilon^{-4NO_M}|w_{(-j-1)N-1,-m} - \hat{w}_{(-j-1)N-1,-\hat{m}}|+$$

$$+\cdots+\left\|\frac{\partial f_{[z]_{-1}}}{\partial w}(\vec{q},\hat{w}_{-N-1,-\hat{m}})\right\|(1+nC_1/\varepsilon)C\varepsilon^{-4NO_M}|w_{-2N-1,-m}-\hat{w}_{-2N-1,-\hat{m}}|$$
$$+(1+nC_1/\varepsilon)C\varepsilon^{-4NO_M}|w_{-N-1,-m}-\hat{w}_{-N-1,-\hat{m}}|. \tag{17}$$

Now, apply 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

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

Then, recall that an α-typical sequence z_{-n}^{-1} breaks at most $3\alpha n$ times. Thus there are at least $(1-3\alpha)n$ i's where we can use the estimate (8) and at most $3\alpha n$ i's where we can only use the weaker estimates (7). Similar to the derivation of (9), with Remark 2.2, we derive that for any $\alpha<\alpha_0$, every term in the right hand side of (17) is $\hat{O}(\varepsilon^n)$ on $\mathcal{M}_{\delta_0}\times T^{\alpha}_{n,m,\hat{m}}$ (we use (11) 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 (5), implies the proposition for $\ell=1$, as desired.

The proof of (14) for a generic $\vec{k}$ is rather similar, however very tedious. We next briefly illustrate the idea of the proof. Note that (compare with (15), (16) 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 $\vec{k}'\preceq\vec{k}$, $t\le|\vec{k}|$ and $|\vec{a}_i|<|\vec{k}|$ for all i. Using (11) and the fact that there exists a constant C (by Lemma 2.6) such that

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

we then can establish (compare with (17) for $|\vec{k}|=1$)

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

where "others" is the sum of finitely many terms, each of which takes the following form (see the j-th term of (17) for $|\vec{k}| = 1$)

$$n^{D_{\vec{k}'}} O(1/\varepsilon^{D_{\vec{k}'}}) \prod_{i=-j}^{-1} \left\| \frac{\partial f_{[z]_i}}{\partial w}(\vec{q}, \hat{w}_{iN-1,-\hat{m}}) \right\| |w^{(\vec{a})}_{(-j-1)N-1,-m} - \hat{w}^{(\vec{a})}_{(-j-1)N-1,-\hat{m}}|, \tag{18}$$

where $|\vec{a}| < |\vec{k}|$, $D_{\vec{k}'}$ is a constant dependent on $\vec{k}'$. Then inductively, one can use the similar dichotomy approach to establish that (18) is $\hat{O}(\varepsilon^n)$ on $\mathcal{M}_{\delta_0} \times T^{\alpha}_{n,m,\hat{m}}$, which implies (14) for a generic $\vec{k}$, and thus the proposition for a generic ℓ.

□

2.3 Asymptotic behavior of entropy rate

The parameterization of Z as a function of ε fits in the framework of [4] in a more general setting. Consequently, we have the following three propositions.

Proposition 2.10. *Assume that $\vec{p} \in \mathcal{M}_0$. For any sequence $z_{-n}^0 \in \mathcal{Z}^{n+1}$, $p((x_{-1}, c_{-1}) = \cdot|z_{-n}^{-1})$ and $p(z_0|z_{-n}^{-1})$ are analytic around $\varepsilon = 0$. Moreover,* $\mathrm{ord}\,(p(z_0|z_{-n}^{-1})) \le O_M$.

Proof. Analyticity of $p((x_{-1}, c_{-1}) = \cdot|z_{-n}^{-1})$ follows from Proposition 2.4 in [4]. It then follows from $p(z_0|z_{-n}^{-1}) = p((x_{-1}, c_{-1}) = \cdot|z_{-n}^{-1})\Omega_{z_0}\mathbf{1}$ and the fact that any row sum of Ω_{z_0} is non-zero that $p(z_0|z_{-n}^{-1})$ is analytic with $\mathrm{ord}\,(p(z_0|z_{-n}^{-1})) \le O_M$. □

Proposition 2.11. *(see Proposition 2.7 in [4]) 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*

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

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 2.12. It follows from Proposition 2.11 that for any α-typical sequence z_{-n}^{-1} with α small enough and n large enough, $\mathrm{ord}\,(p(z_0|z_{-n}^{-1})) = \mathrm{ord}\,(p(z_0|z_{-n-1}^{-1}))$

Proposition 2.13. *(see Theorem 2.8 in [4]) Assume that $\vec{p} \in \mathcal{M}_0$. For any $k \ge 0$,*

$$H(Z) = H(Z)|_{\varepsilon=0} + \sum_{j=1}^{k} g_j \varepsilon^j + \sum_{j=1}^{k+1} f_j \varepsilon^j \log \varepsilon + O(\varepsilon^{k+1}), \tag{19}$$

where f_j's and g_j's are functions of Π, the transition probability matrix of X.

The following theorem strengthens Proposition 2.13 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 (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 2.10)

$$F_n(\vec{p},\varepsilon)=\sum_{z_{-n}^0}-\mathrm{ord}\,(p(z_0|z_{-n}^{-1}))p(z_{-n}^0) \tag{20}$$

and

$$G_n(\vec{p},\varepsilon)=\sum_{z_{-n}^0}-p(z_{-n}^0)\log p^\circ(z_0|z_{-n}^{-1}), \tag{21}$$

where

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

Theorem 2.14. *Given $\delta_0>0$, for sufficiently small ε_0,*

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

$$H(Z(\vec{p},\varepsilon))=G(\vec{p},\varepsilon)+F(\vec{p},\varepsilon)\log\varepsilon. \tag{22}$$

   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 2.13;

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 $0<\rho<1$ such that on $U_{\delta_0,\varepsilon_0}$*

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

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

   and

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

   for sufficiently large n.

Proof. **1)** Recall that

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

It follows from a compactness argument that $H_n(Z)$ uniformly converges to $H(Z)$ on the parameter space $U_{\delta_0,\varepsilon_0}$ for any positive ε_0. 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. 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 2.10, $p(z_0|z_{-n}^{-1})$ is analytic with $\mathrm{ord}\,(p(z_0|z_{-n}^{-1})) \leq O_M$. 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 (22). Then the "Moreover" part then immediately follows by equating (19) and (22) to compare the coefficients.

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

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

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

By Remark 2.12, we have

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

Applying Lemma 2.3, we have

$$|F_n^\alpha(\vec{p},\varepsilon) - F_{n+1}^\alpha(\vec{p},\varepsilon)| = \hat{O}(\varepsilon^n) \text{ on } \mathcal{M}_{\delta_0}, \tag{23}$$

which implies that there exists $\varepsilon_0 > 0$ such that $F_n^\alpha(\vec{p},\varepsilon)$ are exponentially Cauchy 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 $|\varepsilon| \le \varepsilon_0$. Then, using Lemma 2.4 and a similar argument as above, we can prove that

$$|F_n^{\alpha,\mathbb{C}}(\vec{p},\varepsilon) - F_{n+1}^{\alpha,\mathbb{C}}(\vec{p},\varepsilon)| = \hat{O}(|\varepsilon|^n) \text{ on } \mathcal{M}_{\delta_0}^{\mathbb{C}}(\eta_0); \tag{24}$$

in other words, for some $\eta_0, \varepsilon_0 > 0$, $F_n^{\alpha,\mathbb{C}}(\vec{p},\varepsilon)$ are exponentially Cauchy and thus uniformly converges on all $(\vec{p},\varepsilon)$ with $\vec{p} \in \mathcal{M}_{\delta_0}^{\mathbb{C}}(\eta_0)$ and $|\varepsilon| \le \varepsilon_0$. 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 purpose.

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

$$|\log\alpha - \log\beta| \le \max\{|(\alpha-\beta)/\beta|, |(\alpha-\beta)/\alpha|\}. \tag{25}$$

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

$$|p^\circ(z_0|z_{-n}^{-1}) - p^\circ(z_0|z_{-n-1}^{-1})| = \hat{O}(\varepsilon^n) \text{ on } \mathcal{M}_{\delta_0} \times T_{n,n,n+1}^\alpha. \tag{26}$$

One further checks that by Proposition 2.10, there exists a positive constant C such that for ε small enough and for any sequence z_{-n}^{-1},

$$p(z_0|z_{-n}^{-1}) \ge C\varepsilon^{O_M},$$

and thus,

$$p^\circ(z_0|z_{-n}^{-1}) \ge C\varepsilon^{O_M}. \tag{27}$$

Using (25), (26), (27) and Lemma 2.3, we have

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

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

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

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

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

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

$$+ \left| \sum_{z_{-n}^{-1} \in T_n^\alpha, z_{-n-1}^{-1} \notin T_{n+1}^\alpha, z_0} -p(z_{-n-1}^0) \log p^\circ(z_0|z_{-n}^{-1}) \right| + \left| \sum_{z_{-n}^{-1} \notin T_n^\alpha, z_{-n-1}^{-1} \in T_{n+1}^\alpha, z_0} -p(z_{-n-1}^0) \log p^\circ(z_0|z_{-n-1}^{-1}) \right| = \hat{O}(\varepsilon^n) \text{ on } \mathcal{M}_{\delta_0}, \tag{28}$$

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

Apply the multivariate Faa Di Bruno formula [1, 6] to the function $f(y) = \log y$, we have for $\vec{l}$ with $|\vec{l}| \neq 0$,

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

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

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

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

$$+ \sum_{z_{-n}^{-1} \in T_n^\alpha, z_0} -p^{(\vec{m})}(z_{-n}^0) \log p^\circ(z_0|z_{-n}^{-1}). \tag{29}$$

We tackle the last term of (29) first. Using (25) and (26) and with a parallel argument obtained through replacing $p(z_{-n}^0), p(z_{-n-1}^0)$ in (28) by $p^{(\vec{m})}(z_{-n}^0), p^{(\vec{m})}(z_{-n-1}^0)$, respectively, we can show that

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

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 (13)). 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

$$\left|\frac{p^\circ(z_0|z_{-n}^{-1})^{(\vec{a}_1)}}{p^\circ(z_0|z_{-n}^{-1})}\cdots\frac{p^\circ(z_0|z_{-n}^{-1})^{(\vec{a}_k)}}{p^\circ(z_0|z_{-n}^{-1})}-\frac{p^\circ(z_0|z_{-n-1}^{-1})^{(\vec{a}_1)}}{p^\circ(z_0|z_{-n-1}^{-1})}\cdots\frac{p^\circ(z_0|z_{-n-1}^{-1})^{(\vec{a}_k)}}{p^\circ(z_0|z_{-n-1}^{-1})}\right|$$

$$\leq\left|\left(\frac{p^\circ(z_0|z_{-n}^{-1})^{(\vec{a}_1)}}{p^\circ(z_0|z_{-n}^{-1})}-\frac{p^\circ(z_0|z_{-n-1}^{-1})^{(\vec{a}_1)}}{p^\circ(z_0|z_{-n-1}^{-1})}\right)\frac{p^\circ(z_0|z_{-n}^{-1})^{(\vec{a}_2)}}{p^\circ(z_0|z_{-n}^{-1})}\cdots\frac{p^\circ(z_0|z_{-n}^{-1})^{(\vec{a}_k)}}{p^\circ(z_0|z_{-n}^{-1})}\right|$$

$$+\left|\frac{p^\circ(z_0|z_{-n-1}^{-1})^{(\vec{a}_1)}}{p^\circ(z_0|z_{-n-1}^{-1})}\left(\frac{p^\circ(z_0|z_{-n}^{-1})^{(\vec{a}_2)}}{p^\circ(z_0|z_{-n}^{-1})}-\frac{p^\circ(z_0|z_{-n-1}^{-1})^{(\vec{a}_2)}}{p^\circ(z_0|z_{-n-1}^{-1})}\right)\frac{p^\circ(z_0|z_{-n}^{-1})^{(\vec{a}_3)}}{p^\circ(z_0|z_{-n}^{-1})}\cdots\frac{p^\circ(z_0|z_{-n}^{-1})^{(\vec{a}_k)}}{p^\circ(z_0|z_{-n}^{-1})}\right|+\cdots$$

$$+\left|\frac{p^\circ(z_0|z_{-n-1}^{-1})^{(\vec{a}_1)}}{p^\circ(z_0|z_{-n-1}^{-1})}\cdots\frac{p^\circ(z_0|z_{-n-1}^{-1})^{(\vec{a}_{k-1})}}{p^\circ(z_0|z_{-n-1}^{-1})}\left(\frac{p^\circ(z_0|z_{-n}^{-1})^{(\vec{a}_k)}}{p^\circ(z_0|z_{-n}^{-1})}-\frac{p^\circ(z_0|z_{-n-1}^{-1})^{(\vec{a}_k)}}{p^\circ(z_0|z_{-n-1}^{-1})}\right)\right|.$$

Now apply the inequality

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

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

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

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

It follows from multivirate Leibnitz rule and (11) that there exists a positive constant $C_{\vec{a}}$ such that

$$|p(z_0|z_{-n}^{-1})^{(\vec{a})}|=|(w_{-1,-n}\Omega_{z_0}\mathbf{1})^{(\vec{a})}|\leq n^{|\vec{a}|}C_{\vec{a}}/\varepsilon^{|\vec{a}|},\tag{30}$$

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

$$p^\circ(z_0|z_{-n}^{-1})^{(\vec{a})}\leq n^{|\vec{a}|}C_{\vec{a}}^\circ/\varepsilon^{|\vec{a}|+O_M}.\tag{31}$$

Combining (27), (29), (30) and (31) gives us

$$|(G_n^\alpha)^{(\vec{m})}(\vec{p},\varepsilon)-(G_{n+1}^\alpha)^{(\vec{m})}(\vec{p},\varepsilon)|=\hat{O}(\varepsilon^n)\text{ on }\mathcal{M}_{\delta_0}.\tag{32}$$

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 $\vec{m}$).

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

3) Note that,

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

Apply the multivariate Leibnitz rule, then by Proposition 2.10, (30), (13) and Lemma 2.3, we have for any ℓ,

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

It follows from (24) and the Cauchy integral formula that

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

we then have

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

and thus

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

which imply that for any ℓ, there exist $\varepsilon_0 > 0$, $0 < \rho < 1$ such that on $U_{\delta_0,\varepsilon_0}$

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

and further

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

for sufficiently large n.

Similarly note that

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

Then by (30), (31), (27) and Lemma 2.3, we have for any ℓ,

$$\left|D^{\ell}_{\vec{p},\varepsilon} G_n(\vec{p},\varepsilon) - D^{\ell}_{\vec{p},\varepsilon} G^{\alpha}_n(\vec{p},\varepsilon)\right|$$

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

which, together with (32), implies that for any ℓ, there exists $\varepsilon_0 > 0$, $0 < \rho < 1$ 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.

□

Remark 2.15. We don't know if $G(\vec{p},\varepsilon)$ is analytic or not with respect to $(\vec{p},\varepsilon)$.

3 Concavity of 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 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

$$C^1(\varepsilon) = \sup_{\vec{p} \in \mathcal{M}} I(Z; X) = \sup_{\vec{p} \in \mathcal{M}} (H(Z) - H(Z|X)) \tag{33}$$

and its n-th approximation

$$C_n^1(\varepsilon) = \sup_{\vec{p} \in \mathcal{M}} I_n(Z; X) = \sup_{\vec{p} \in \mathcal{M}} \left(H_n(Z) - \frac{1}{n+1} H(Z_{-n}^0 | X_{-n}^0) \right). \tag{34}$$

As mentioned earlier, since the channel is memoryless, the second terms in (33) and (34) both reduce to $H(Z_0|X_0)$, which can be written as:

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

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 [10], one can show that for sufficiently small $\varepsilon_1 > 0, \delta_1 > 0$ and all $0 \le \varepsilon \le \varepsilon_1$,

$$C_n^1(\varepsilon) = \sup_{\vec{p} \in \mathcal{M}_{\delta_1}} (H_n(Z) - H(Z_0|X_0)) > \sup_{\vec{p} \in \mathcal{M} \setminus \mathcal{M}_{\delta_1}} (H_n(Z) - H(Z_0|X_0)), \tag{35}$$

$$C^1(\varepsilon) = \sup_{\vec{p} \in \mathcal{M}_{\delta_1}} (H(Z) - H(Z_0|X_0)) > \sup_{\vec{p} \in \mathcal{M} \setminus \mathcal{M}_{\delta_1}} (H(Z) - H(Z_0|X_0)). \tag{36}$$

Theorem 3.1. *There exist $\varepsilon_0 > 0, \delta_0 > 0$ such that for all $0 \le \varepsilon \le \varepsilon_0$,*

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*

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

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 2.14, for some $\varepsilon_0 > 0, \delta_0 > 0$, $G(\vec{p}, \varepsilon)$ and $\hat{F}(\vec{p}, \varepsilon)$ are smooth on $U_{\delta_0, \varepsilon_0}$, and so

$$\lim_{\varepsilon \to 0} D_{\vec{p}}^2 G(\vec{p}, \varepsilon) = D_{\vec{p}}^2 G(\vec{p}, 0)$$

and

$$\lim_{\varepsilon \to 0} D_{\vec{p}}^2 \hat{F}(\vec{p}, \varepsilon) = D_{\vec{p}}^2 \hat{F}(\vec{p}, 0),$$

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

$$\lim_{\varepsilon \to 0} D_{\vec{p}}^2 H(Z(\vec{p}, \varepsilon)) = D_{\vec{p}}^2 G(\vec{p}, 0) = D_{\vec{p}}^2 H(Z(\vec{p}, 0)),$$

again uniformly on $\mathcal{M}_{\delta_0}$. Since $D_{\vec{p}}^2 H(Z(\vec{p}, 0))$ is negative definite on $\mathcal{M}_{\delta_0}$ (see [3]), it follows that for sufficiently small ε, $D_{\vec{p}}^2 H(Z(\vec{p}, \varepsilon))$ 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 2 of Theorem 2.14, 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 $\delta_n, \varepsilon_n > 0$.

Part 2: This follows from part 1 and statements (35) and (36).

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

$$\begin{aligned} f(\vec{p}_n(\varepsilon)) &= f(\vec{p}_\infty(\varepsilon)) + D_{\vec{p}} f(\vec{p}_\infty(\varepsilon))(\vec{p}_n(\varepsilon) - \vec{p}_\infty(\varepsilon)) \\ &\quad + (\vec{p}_n(\varepsilon) - \vec{p}_\infty(\varepsilon))^T D_{\vec{p}}^2 f(\eta_1)(\vec{p}_n(\varepsilon) - \vec{p}_\infty(\varepsilon)), \end{aligned} \tag{37}$$

$$\begin{aligned} f_n(\vec{p}_\infty(\varepsilon)) &= f_n(\vec{p}_n(\varepsilon)) + D_{\vec{p}} f_n(\vec{p}_n(\varepsilon))(\vec{p}_\infty(\varepsilon) - \vec{p}_n(\varepsilon)) \\ &\quad + (\vec{p}_n(\varepsilon) - \vec{p}_\infty(\varepsilon))^T D_{\vec{p}}^2 f_n(\eta_2)(\vec{p}_n(\varepsilon) - \vec{p}_\infty(\varepsilon)), \end{aligned} \tag{38}$$

here the superscript T denotes the transpose.

By part 2 of Theorem 3.1

$$D_{\vec{p}} f(\vec{p}_\infty(\varepsilon)) = 0, \quad D_{\vec{p}} f_n(\vec{p}_n(\varepsilon)) = 0. \tag{39}$$

By part 2 of Theorem 2.14, with $\ell = 0$, there exists $0 < \rho_0 < 1$ such that

$$|f(\vec{p}_\infty(\varepsilon)) - f_n(\vec{p}_\infty(\varepsilon))| \leq \rho_0^n, |f(\vec{p}_n(\varepsilon)) - f_n(\vec{p}_n(\varepsilon))| \leq \rho_0^n. \tag{40}$$

Combining (37), (38), (39), (40), we have

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

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

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

This, together with part 1 of Lemma 2.4, 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 2.14 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

$$H''(Z)|_{\varepsilon=0} = -\frac{1}{p} - \frac{1}{1-p} \leq -2.$$

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

References

[1] G. Constantine and T. Savits. A Multivariate Faa Di Bruno Formula With Applications. *Transactions of the American Mathematical Society*, Vol. 348, No. 2., Feb, 1996, pp. 503-520.

[2] 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.

[3] G. Han and B. Marcus. Asymptotics of Input-Constrained Binary Symmetric Channel Capacity. *Annals of Applied Probability*, Vol. 19, No. 3, June, 2009, Page(s):1063-1091.

[4] G. Han and B. Marcus Asymptotics of Entropy Rate in Special Families of Hidden Markov Chains. *IEEE Transactions on Information Theory*, Volume 56, Issue 3, March 2010, Page(s):1287-1295.

[5] P. Jacquet, G. Seroussi, and W. Szpankowski. On the Entropy of a Hidden Markov Process (extended abstract). *Data Compression Conference*, 362–371, Snowbird, 2004.

[6] 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.

[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. Ordentlich and T. Weissman. New bounds on the entropy rate of hidden Markov process. *IEEE Information Theory Workshop*, San Antonio, Texas, 24-29 Oct. 2004, Page(s):117 - 122.

[10] W. Parry. Intrinsic Markov chains. *Trans. Amer. Math. Soc.* 112 (1964), 55-66.

[11] P. Vontobel, A. Kavcic, D. Arnold and Hans-Andrea Loeliger. A Generalization of the Blahut-Arimoto Algorithm to Finite-State Channels. *IEEE Trans. Inform. Theory*, vol. 54, no. 5, pp. 1887–1918, May, 2008.

[12] E. Zehavi and J. Wolf. On runlength codes. *IEEE Trans. Information Theory*, **34**, 45–54, 1988.