Concavity of Mutual Information Rate of Finite-State Channels

Yonglong Li and Guangyue Han

The University of Hong Kong

Email:{liyonglong, ghan}@hku.hk

Abstract—The computation of the capacity of a finite-state channel (FSC) is a fundamental and long-standing open problem in information theory. The capacity of a memoryless channel can be effectively computed via the classical Blahut-Arimoto algorithm (BAA), which, however, does not apply to a general FSC. Recently Vontobel *et al.* [1] generalized the BAA to compute the capacity of a finite-state machine channel with a Markovian input. Their proof of the convergence of this algorithm, however, depends on the concavity conjecture posed in their paper. In this paper, we confirm the concavity conjecture for some special FSCs. On the other hand, we give examples to show that the conjecture is not true in general.

I. Introduction

Discrete-time FSCs are a broad class of channels whose transition probabilities at any time slot are dictated by a finite number of channel states. They can be used to model channels with memory since the input and output signals in the past can be regarded as a single current state of the channel. The most prevalent examples include partial response channels [2], Gilbert-Elliot channels [3] and noisy input-restricted channels [4], which are widely used in various real-life applications such as magnetic and optical recording [5] and communications over band-limited channels with inter-symbol interference [6].

The computation of the capacity of an FSC is notoriously difficult and has been open for decades. For a discrete memoryless channel, Shannon gave a closed-form formula of the capacity in his classical paper [7], and Blahut [8] and Arimoto [9], independently proposed an algorithm which can efficiently compute the capacity and the capacity-achieving distribution simultaneously. There are also some well-known methods to compute the capacity for some other special FSCs, such as water-filling methods for Gaussian channels with ISI and average power constraint [10] and methods in [3], [11] for Gilbert-Elliot channels. However, little is known about the efficient computation of the capacity of a general FSC.

Recently Vontobel *et al.* [1] proposed a generalized Blahut-Arimoto algorithm (GBAA) to compute the capacity of a class of finite-state machine channels. They showed that

- the stationary points of the GBAA correspond to the critical points of the mutual information rate; and
- if the convergence of the GBAA is observed, then it will converge to a local maximum of the mutual information rate of the channel.

It has been observed that the GBAA fairly precisely approximates the channel capacity for a number of practical channels. On the theoretical side, however, the proof of the convergence of the GBAA in [1] depends on the concavity conjecture posed in their paper, which claims the concavity of both the mutual information rate and some conditional entropy with respect to certain prescribed parameterization. In a recent work [12], Han and Marcus partially confirmed this conjecture for a class of input-restricted memoryless channels in the high signal-to-noise ratio (SNR) regime. In this paper, we continue to develop the ideas in [12], [13], [19], [15] to show that for a special family of FSCs, the mutual information rate is indeed a concave function with respect to the above-mentioned parameterization. And we give examples to show the concavity conjecture in [1] may hold true or fail in different regimes.

II. Channel Model

In this section, we assume the reader is familiar with the standard notation and terminologies in information theory.

Consider a discrete-time FSC and let X, Z, C denote its stationary input, output and state processes over the finite alphabets $\mathcal{X}, \mathcal{Z}, \mathcal{C}$, respectively. Assume that X is a stationary first-order input Markov chain with transition probability matrix $\Pi = (\pi_{xy})_{x,y \in \mathcal{X}}$, which is parameterized as in [1], that is, X is parameterized by $\vec{p} = \vec{p}(X) \triangleq (p_{ij} : i, j \in \mathcal{X})$, where $p_{ij} = p_{ij}(X) \triangleq P(X_1 = i, X_2 = j)$. Unless specified otherwise, we assume that

(i) for all i, j, $p_{ij} > 0$;

(ii) $p(z|x, c)$, analytically parameterized by ε, is not identically 0 around $\varepsilon = 0$ for any x, c, z;

(iii) $p(z|x, c)(0) = \begin{cases} 1 & z = \Phi(x) \\ 0 & \text{otherwise} \end{cases}$, where $\Phi(x)$ is a one-to-one mapping from $\mathcal{X}$ to $\mathcal{Z}$;

(iv) the joint distribution $p(z_{-n}^0, c_{-n}^0, x_{-n}^0 | x_{-n-1}, c_{-n-1})$ can be written in the form

$$\prod_{i=-n}^{0} p(x_i|x_{i-1}) p(c_i|x_{i-1}^i, c_{i-1}) p(z_i|x_i, c_i),$$

where $p(z_i|x_i, c_i)$ and $p(c_i|x_i, x_{i-1}^i)$ is independent of i and $p(c_i|x_{i-1}^i, c_{i-1}) > 0$ for any x_{i-1}^i, c_{i-1}^i.

Remark II.1. In this remark, we give some further explanations of the above assumptions.

- ε can be regarded as the channel parameter and $\Phi(x)$ is the noiseless output corresponding to the input x.

- Assumption (iv) implies that X, (X, C), (X, C, Z) are Markov processes, and by definition, Z is a *hidden Markov process* [16].
- Assumption (iv) implies that, at any time slot, the channel is characterized by the conditional probability:

$$p(z_n|x_n, c_n) = P(Z_n = z_n|X_n = x_n, C_n = c_n).$$

- Prominent examples of our model include binary symmetric channels with crossover probability ε and the binary erasure channel with erasure rate ε.

We next claim that in the high SNR regime, that is, when ε is sufficiently small, the capacity of our channel can only be achieved by some X with $p_{ij} > 0$, which justifies Assumption (i).

Indeed, for a first-order Markov chain Y over alphabet $\mathcal{X}$, define $S(Y)$, the so-called *structure matrix* of Y, as

$$S(Y)(i, j) = \begin{cases} 1, & \text{if } p_{ij}(Y) > 0 \\ 0, & \text{if } p_{ij}(Y) = 0 \end{cases}.$$

It has been shown [17] that for any first-order irreducible Markov chain Y, there is a unique first-order Markov chain Y^* such that
(a) $S(Y^*) = S(Y)$;
(b) for any first-order irreducible Markov chain Y_1 with $S(Y_1) = S(Y)$, we have $H(Y_1) \leq H(Y^*)$;
(c) $H(Y^*) = \log \lambda(S(Y))$, where $\lambda(S(Y))$ is the largest eigenvalue of $S(Y)$ in modulus.

By Assumption (i), $S(X)$ is an all one matrix with the largest eigenvalue $|\mathcal{X}|$. This, together with the known fact that $\lambda(S(Y)) < |\mathcal{X}|$ (see, e.g., [18]), implies that if $\vec{p}(Y)$ has some zero entries, we will have

$$H(Y) \leq H(Y^*) = \log \lambda(S(Y)) < \log |\mathcal{X}| = H(X^*),$$

where X^* is a first-order Markov chain with

$$\vec{p}(X^*) = (p_{ij} = 1/|\mathcal{X}|^2 : i, j \in \mathcal{X}).$$

It then follows from Assumption (iii) that

$$I(X^*; Z)|_{\varepsilon=0} = H(X^*) > H(Y) = I(Y; Z)|_{\varepsilon=0}. \quad (1)$$

Using the known fact that for the hidden Markov process Z, $H(Z)$ is continuous with respect to $(\vec{p}, \varepsilon)$ and thus so are $I(X^*; Z)$ and $I(Y; Z)$, we establish the claim.

Here we remark that, in this paper, X, Z and many other quantities depends on the parameters $(\vec{p}, \varepsilon)$; however, as in (1), we will often suppress the notational dependence for convenience.

III. Concavity of Mutual Information Rate

Throughout the paper, we will use the superscript $''$ to denote the Hessian with respect to $\vec{p}$. We will show that for the channel models described in Section II, $I(X; Z)$ is a concave function with respect to $\vec{p}$ in the high SNR regime. In more detail, Proposition III.3 shows that $H''(Z)$, $H''(X, Z)$ approach to $H''(X)$ as $\varepsilon \to 0$, which, in conjunction with the

concavity of $H(X)$, implies the concavity of $I(X; Z)$ with respect to $\vec{p}$ for sufficiently small ε.

We first introduce some notation and terminologies. By Assumption (iv), the transition probability matrix of (X, C, Z) can be computed as

$$\Omega((x, c, \hat{w}), (y, d, w)) = \pi_{xy} p(d|x, y, c) p(w|y, d),$$

where $x, y \in \mathcal{X}, c, d \in \mathcal{C}$ and $\hat{w}, w \in \mathcal{Z}$. For any $z \in \mathcal{Z}$, define Ω_z to be the matrix of the same size as Ω such that

$$\Omega_z((x, c, \hat{w}), (y, d, w)) = \begin{cases} \Omega((x, c, \hat{w}), (y, d, w)) & \text{if } w = z \\ 0 & \text{otherwise} \end{cases}.$$

One then checks that

$$p(z_{-n}^0) = \pi \Omega_{z_{-n}^0} \mathbf{1}. \quad (2)$$

where π denotes the stationary vector of Ω, $\mathbf{1}$ denotes the all one column vector and

$$\Omega_{z_{-n}^0} \triangleq \prod_{-n}^{0} \Omega_{z_i},$$

Let $A_1 = \{z \in \mathcal{Z} : \Phi(x) = z \text{ for some } x\}$ and $A_2 = \mathcal{Z} \setminus A_1$. For an analytic function $g(\varepsilon)$, let $\text{ord}(g(\varepsilon))$ denote the order of $g(\varepsilon)$ around $\varepsilon = 0$, i.e., the degree of the first non-zero term of its Taylor series expansion around $\varepsilon = 0$. We then have the following lemma.

Lemma III.1. *For all z_{-n}^0, we have*

$$ord(p(z_0|z_{-n}^{-1})) = ord(p(z_0)).$$

Proof: Applying (2), we have

$$p(z_0|z_{-n}^{-1}) = \frac{p(z_{-n}^0)}{p(z_{-n}^{-1})} = \frac{\pi \Omega_{z_{-n}^0} \mathbf{1}}{\pi \Omega_{z_{-n}^{-1}} \mathbf{1}} = \frac{(\pi \Omega_{z_{-n}^{-1}})(\Omega_{z_0} \mathbf{1})}{\pi \Omega_{z_{-n}^{-1}} \mathbf{1}}.$$

It follows from the definition of Ω_{z_0} and positivity of $\vec{p}$ and $p(c|x, y, d)$ that, all the entries in the column vector $\Omega_{z_0} \mathbf{1}$ have the same order, which, by (2), must be $\text{ord}(p(z_0))$. The lemma then follows from

$$p(z_0|z_{-n}^{-1}) = \varepsilon^{\text{ord}(p(z_0))} \frac{(\pi \Omega_{z_{-n}^{-1}})(\Omega_{z_0} \mathbf{1}/\varepsilon^{\text{ord}(p(z_0))})}{\pi \Omega_{z_{-n}^{-1}} \mathbf{1}}.$$

$\square$

Remark III.2. For $z_0 \in A_1$, by Assumptions (i) and (iii), there exists $x \in \mathcal{X}$ such that $z = \Phi(x)$ and

$$p(z_0)|_{\varepsilon=0} = p(x) > 0.$$

So, for $z_0 \in A_1$, $\text{ord}(p(z_0|z_{-n}^{-1})) = 0$. On the other hand, it is obvious that $\text{ord}(p(z_0|z_{-n}^{-1})) = \text{ord}(p(z_0)) > 0$ for $z_0 \in A_2$.

In what follows, for any $\delta_0 > 0$, $\varepsilon_0 > 0$, define

$$\mathcal{M}_{\delta_0} \triangleq \{\vec{p} : p_{ij} \geq \delta_0\},$$

and

$$\mathcal{U}_{\delta_0, \varepsilon_0} = \{(\vec{p}, \varepsilon) : p_{ij} > \delta_0, \varepsilon < \varepsilon_0\}.$$

Proposition III.3. *For any $\delta_0 > 0$, there exist $\varepsilon_0 > 0$ and analytic functions $F(\vec{p}, \varepsilon)$, $\tilde{F}(\vec{p}, \varepsilon)$, $G(\vec{p}, \varepsilon)$ and $\tilde{G}(\vec{p}, \varepsilon)$ on $\mathcal{U}_{\delta_0, \varepsilon_0}$ such that*

$$H(Z) = F(\vec{p}, \varepsilon) + G(\vec{p}, \varepsilon)(\varepsilon \log \varepsilon) \tag{3}$$

and

$$H(X, Z) = \tilde{F}(\vec{p}, \varepsilon) + \tilde{G}(\vec{p}, \varepsilon)(\varepsilon \log \varepsilon). \tag{4}$$

Proof: We will only prove (3), the proof of (4) being similar. By Lemma III.1, $\mathrm{ord}(p(z_0|z_{-n}^{-1})) = \mathrm{ord}(p(z_0))$, which will be rewritten as $c(z_0)$ for notational simplicity.

Recall that the n-th order entropy rate of Z is defined as

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

which can be decomposed as follows:

$$
\begin{aligned}
H_n(Z) &= -\Big(\sum_{z_{-n}^{-1}, z_0 \in A_1} + \sum_{z_{-n}^{-1}, z_0 \in A_2}\Big) p(z_{-n}^0) \log p(z_0|z_{-n}^{-1}) \\
&\overset{(a)}{=} -\sum_{z_{-n}^{-1}, z_0 \in A_1} p(z_{-n}^0) \log p(z_0|z_{-n}^{-1}) \\
&\quad - \sum_{z_{-n}^{-1}, z_0 \in A_2} p(z_{-n}^0)[\log \hat{p}(z_0|z_{-n}^{-1}) + c(z_0)\log \varepsilon] \\
&= F_n(\vec{p}, \varepsilon) + G_1(\vec{p}, \varepsilon)\log \varepsilon
\end{aligned}
$$

where (a) follows from Lemma III.1 and Remark III.2, and

$$\hat{p}(z_0|z_{-n}^{-1}) \triangleq \frac{p(z_0|z_{-n}^{-1})}{\varepsilon^{c(z_0)}}, \quad G_1(\vec{p}, \epsilon) \triangleq -\sum_{z_0 \in A_2} p(z_0)c(z_0)$$

and

$$
\begin{aligned}
F_n(\vec{p}, \varepsilon) \triangleq &-\sum_{z_{-n}^{-1}, z_0 \in A_1} p(z_{-n}^0) \log p(z_0|z_{-n}^{-1}) \\
&- \sum_{z_{-n}^{-1}, z_0 \in A_2} p(z_{-n}^0) \log \hat{p}(z_0|z_{-n}^{-1}).
\end{aligned}
$$

Now, fix $\vec{v}_0 = (\vec{p}_0, 0)$, where $\vec{p}_0 \in \mathcal{M}_{\delta_0}$. For some r, let $N_{\vec{v}_0}(r)$ denote a r-ball in $\mathbb{C}^{|\mathcal{X}|^2 + 1}$ centered at $\vec{v}_0$. As in [13], for r sufficiently small, $p(z_0|z_{-n}^{-1}), \hat{p}(z_0|z_{-n}^{-1}), p(z_{-n}^0), F_n$ can be naturally extended to complex analytic functions of $\vec{v} \in N_{\vec{v}_0}(r)$, which will be denoted by $p^{\vec{v}}(z_0|z_{-n}^{-1}), \hat{p}^{\vec{v}}(z_0|z_{-n}^{-1}), p^{\vec{v}}(z_{-n}^0), F_n^{\vec{v}}$, respectively. We will show that

(I) there exists $r > 0$ such that on $N_{\vec{v}_0}(r)$, $F_n^{\vec{v}}$ uniformly converges to a limit function $F^{\vec{v}}$, which is necessarily analytic.

Then, we define F to be $F^{\vec{v}}$ restricted to the real parameter $(\vec{p}, \varepsilon)$ and set $G = G_1/\varepsilon$. With the obvious fact that G is analytic, the theorem then follows from a compactness argument (to obtain a "common" r for all $\vec{p} \in \mathcal{M}_{\delta_0}$).

Under the positivity conditions in Assumptions (i) and (iv), similar arguments as the proof of Theorem 1.1 of [13] can be applied to establish (I). So, in the following, we only outline the major steps needed.

Step 1. Applying the mean-value theorem, we can show that, there exists some constant C such that for any two output sequences $z_{-n_1}^0$ and $\hat{z}_{-n_2}^0$ with a common tail $z_{-n}^0 = \hat{z}_{-n}^0$, $n \leq n_1, n_2$, and for all $\vec{v} \in N_{\vec{v}_0}(r)$

$$|\log p^{\vec{v}}(z_0|z_{-n_1}^{-1}) - \log p^{\vec{v}}(\hat{z}_0|\hat{z}_{-n_2}^{-1})| < C\rho^n,$$

and

$$|\log \hat{p}^{\vec{v}}(z_0|z_{-n_1}^{-1}) - \log \hat{p}^{\vec{v}}(\hat{z}_0|\hat{z}_{-n_2}^{-1})| < C\rho^n.$$

Step 2. Next, we can show that there exists $1 < \sigma < 1/\rho$ such that for all $\vec{v} \in N_{\vec{v}_0}(r)$

$$\sum_{z_{-n-1}^{-1}, z_0 \in A_1} |p^{\vec{v}}(z_{-n-1}^0)| \leq \sigma^{n+2}, \quad \sum_{z_{-n-1}^{-1}, z_0 \in A_2} |p^{\vec{v}}(z_{-n-1}^0)| \leq \sigma^{n+2}.$$

Step 3. Letting $\rho_1 = \rho\sigma < 1$ and $L = 2C\sigma^2$, we then have for any $\vec{v} \in N_{\vec{v}_0}(r)$

$$
\begin{aligned}
&|F_{n+1}^{\vec{v}} - F_n^{\vec{v}}| \\
&\leq \sum_{z_{-n-1}^0, z_0 \in A_1} |p^{\vec{v}}(z_{-n-1}^0)(\log p^{\vec{v}}(z_0|z_{-n-1}^{-1}) - \log p^{\vec{v}}(z_0|z_{-n}^{-1}))| \\
&\quad + \sum_{z_{-n-1}^0, z_0 \in A_2} |p^{\vec{v}}(z_{-n-1}^0)(\log \hat{p}^{\vec{v}}(z_0|z_{-n-1}^{-1}) \log \hat{p}^{\vec{v}}(z_0|z_{-n}^{-1}))| \\
&\leq 2C\sigma^{n+2}\rho^n = L\rho_1^n.
\end{aligned}
$$

Thus, for all $m > n$,

$$|F_m^{\vec{v}} - F_n^{\vec{v}}| \leq L(\rho_1^n + \cdots + \rho_1^{m-1}) \leq \frac{L\rho_1^n}{1 - \rho_1}.$$

This establishes the uniform convergence of $F_n^{\vec{v}}(\vec{p}, \varepsilon)$ to a limit $F^{\vec{v}}(\vec{p}, \varepsilon)$, which, by Theorem 2.4.1 of [20], is necessarily analytic on $N_{\vec{v}_0}(r)$. The proof is then complete. $\square$

Theorem III.4. *Given $\delta_0 > 0$, there exists $\varepsilon_0 > 0$ such that for any $\varepsilon < \varepsilon_0$, $I(X; Z)$ is concave with respect to $\vec{p} \in \mathcal{M}_{\delta_0}$.*

Proof: By Proposition III.3, there exists $\varepsilon_0 > 0$ such that

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

and

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

where $F, \tilde{F}, G$ and $\tilde{G}$ are analytic functions on $\mathcal{U}_{\delta_0, \varepsilon_0}$. Therefore, we can switch the limit and derivative,

$$\lim_{\varepsilon \to 0} H''(Z) = (\lim_{\varepsilon \to 0} H(Z))'' = H''(X),$$

and

$$\lim_{\varepsilon \to 0} H''(X, Z) = (\lim_{\varepsilon \to 0} H(X, Z))'' = H''(X).$$

It then follows that

$$
\begin{aligned}
\lim_{\varepsilon \to 0} I''(X; Z) &= \lim_{\varepsilon \to 0}(H''(Z) - H''(X, Z) + H''(X)) \\
&= H''(X) - H''(X) + H''(X) \\
&= H''(X). \tag{5}
\end{aligned}
$$

Moreover, by the compactness of $\mathcal{M}_{\delta_0}$, the convergence in (5) is uniform over $\mathcal{M}_{\delta_0}$. This, in conjunction with the fact that $H''(X)$ is negative definite (see [14]), implies that, for sufficiently small $\varepsilon_0 > 0$, $I''(X, Z)$ is negative definite over $\mathcal{M}_{\delta_0}$ for all $\varepsilon < \varepsilon_0$. We then have established the theorem. $\square$

IV. EXAMPLES

Consider a binary symmetric channel with crossover probability ε, denoted by BSC(ε). As before, let X, Z denote the input, output processes of the BSC(ε), respectively. Then, at time n, the channel is characterized by

$$Z_n = X_n \oplus E_n,$$

where $\oplus$ denotes modulo 2 addition and $\{E_n\}$ is the i.i.d. binary noise with $p_E(0) = 1 - \varepsilon$. In this section, using a BSC(ε) at different parameters, we will give examples to show that concavity of $H(X|Z)$ and $I(X;Z)$ may hold true or fail in different regimes. Here, let us note that BSC(ε) can be viewed as degenerated Gilbert-Elliot channels with only one channel state, and by a continuity argument, the concavity assertions in Examples IV.2 and IV.4 hold true for non-degenerated Gilbert-Elliot channels with slightly perturbed parameters.

Suppose X is a first-order Markov chain with the transition probability matrix

$$\left[\begin{array}{cc} \pi_{00} & \pi_{01} \\ \pi_{10} & \pi_{11} \end{array} \right].$$

Then $Y = \{Y_n\} = \{(X_n, E_n)\}$ is a Markov chain with transition probability matrix

$$\Delta = \left[\begin{array}{cc} \pi_{00} & \pi_{01} \\ \pi_{10} & \pi_{11} \end{array} \right] \otimes \left[\begin{array}{cc} 1-\varepsilon & \varepsilon \\ 1-\varepsilon & \varepsilon \end{array} \right],$$

where $\otimes$ denotes the Kronecker product of matrices.

As in Section III, the input Markov chain X is parameterized by $\vec{p} = (p_{00}, p_{11})$, where

$$p_{ii} = P(X_0 = i, X_1 = i), i = 0, 1.$$

Then the transition probability matrix of $\{(X_n, E_n)\}$ can be written as

$$\Delta(\vec{p}, \varepsilon) = \Pi(p_{00}, p_{11}) \otimes \left[\begin{array}{cc} 1-\varepsilon & \varepsilon \\ 1-\varepsilon & \varepsilon \end{array} \right],$$

where

$$\Pi(p_{00}, p_{11}) \triangleq \left[\begin{array}{cc} \frac{2p_{00}}{1+p_{00}-p_{11}} & \frac{1-p_{00}-p_{11}}{1+p_{00}-p_{11}} \\ \frac{1-p_{00}-p_{11}}{1-p_{00}+p_{11}} & \frac{2p_{11}}{1-p_{00}+p_{11}} \end{array} \right].$$

For $i = 0, 1$, define Δ_i as a 4×4 matrix with entries

$$\Delta_i((s,t),(u,v)) = \begin{cases} \Delta((s,t),(u,v)) & \text{if } u \oplus v = i, \\ 0 & \text{otherwise} \end{cases},$$

where $s, t, u, v \in \{0, 1\}$.

The following theorem gives an explicit expression of the derivatives of $H(Z)$ at some special parameters, which is a key result for proving/disproving the concavity conjecture in [1].

Theorem IV.1 (see Theorem 2.1 in [15])**.** *If at $(\vec{p}_0, \varepsilon_0)$, for $i = 0, 1$, Δ_i is a rank one matrix and each of its column is either a positive or a zero column, then we have*

$$\left. \frac{\partial^{\alpha_1 + \alpha_2 + \alpha_3} H(Z)}{\partial \varepsilon^{\alpha_1} \partial p_{00}^{\alpha_2} \partial p_{11}^{\alpha_3}} \right|_{\vec{p}_0, \varepsilon_0} = \left. \frac{\partial^{\alpha_1 + \alpha_2 + \alpha_3} H_{\alpha_1 + \alpha_2 + \alpha_3}(Z)}{\partial \varepsilon^{\alpha_1} \partial p_{00}^{\alpha_2} \partial p_{11}^{\alpha_3}} \right|_{\vec{p}_0, \varepsilon_0},$$

where α_1, α_2 and α_3 are non-negative integers.

The following example shows that while the concavity of $I(X;Z)$ can hold true, the concavity of $H(X|Z)$ may fail in the high SNR regime.

Example IV.2. In the high SNR regime, it follows from Theorem III.4 that $I(X;Z)$ is concave with respect to all $\vec{p}$ whose entries are all bounded away from zero. More precisely, for any $0 < \delta_1 < \delta_2$, when ε is sufficiently small, $I(X;Z)$ is concave with respect to all $\vec{p} = (p_{00}, p_{11})$ with $\delta_1 < p_{00} < \delta_2$.

Next, we show that the concavity of $H(X|Z)$ at some strictly positive $\vec{p}_0$ fails. By Theorem 1.1 in [13], $H(Z)$ is analytic at $(\vec{p}_0, 0)$; moreover, by Theorem IV.1, the Taylor series expansion of $H(Z)$ at $\varepsilon = 0$ can be computed as

$$H(Z) = H(Z)|_{\varepsilon=0} + \sum_{k=1}^{\infty} f_k(\vec{p}) \varepsilon^k,$$

where

$$f_k(\vec{p}) = \left. \frac{\partial^k H(Z)}{\partial \varepsilon^k} \right|_{\varepsilon=0} = \left. \frac{\partial^k H_k(Z)}{\partial \varepsilon^k} \right|_{\varepsilon=0}.$$

Then, around $(\vec{p}_0, 0)$,

$$H(X|Z) = H(X) - H(Z) + H(Z|X) = - \sum_{k=1}^{\infty} f_k(\vec{p}) \varepsilon^k + H(\varepsilon), \tag{6}$$

where $H(\varepsilon) = -\varepsilon \log \varepsilon - (1-\varepsilon) \log(1-\varepsilon)$ and we have used the fact that for a BSC(ε),

$$H(Z)|_{\varepsilon=0} = H(X) \text{ and } H(Z|X) = H(\varepsilon).$$

From (6), we deduce that

$$H''(X|Z) = - \sum_{k=1}^{\infty} f_k''(\vec{p}) \varepsilon^k.$$

So, as shown in the Table I, if $-f_1''(p_{00}, p_{11})$ at some $\vec{p}$ fails to be negative-definite, so does $H''(X|Z)$ for sufficiently small ε, which implies the concavity of $H(Z|X)$ fails in the high SNR regime.

TABLE I

(p_{00}, p_{11})	$(0.2, 0.7)$		$(0.4, 0.5)$	
$-f_1''$	$\begin{array}{cc} -255.4 & -253.5 \\ -253.5 & -238.8 \end{array}$		$\begin{array}{cc} -246.5 & -248.6 \\ -248.6 & -244.3 \end{array}$	
Eigenvalues of $-f_1''$	$(-500.7, 6.6)$		$(-494.0, 3.1)$	

Remark IV.3. For the same settings as in Example IV.2, both $I(X;Z)$ and $H(X|Z)$ can be concave with respect to a different parameterization. More precisely, consider the following parameterization of X:

$$\vec{q} = \{q_{ijk} : i, j, k \in \{0, 1\}\},$$

where $q_{ijk} = q_{ijk}(X) \triangleq P(X_1 = i, X_2 = j, X_3 = k)$.

Again, it follows from Theorem III.4 that in the high SNR regime $I(X;Y)$ is concave with respect to all $\vec{q}$ whose entries are all bounded away from zero. For the concavity of $H(X|Z)$, as in Example IV.2, we have

$$H(X|Z) = - \sum_{k=1}^{\infty} f_k(\vec{q}) \varepsilon^k + H(\varepsilon),$$

where Jacquet *et al.* [22] showed that

$$f_1(\vec{q}) = \sum_{z_1, z_2, z_3} q_{z_1 z_2 z_3} \log \frac{q_{z_1 z_2 z_3}}{q_{z_1 (1 - z_2) z_3}}.$$

It then follows that f_1 is convex with respect to $\vec{q}$, and by (6), $H(X|Z)$ is concave with respect to $\vec{q}$ in the high SNR regime.

The following example shows that while the concavity of $H(X|Z)$ can hold true, the concavity of $I(X;Z)$ may fail in the low SNR regime.

Example IV.4. Notice that for any positive $\vec{p}$ with

$$p_{11} = (1 - \sqrt{p_{00}})^2,$$

$\Pi(p_{00}, p_{11})$ is of rank 1 and so are Δ_0 and Δ_1. It then follows from Theorem 1.1 of [13] that $H(Z)$ is analytic in $(\vec{p}, \varepsilon)$, and by Theorem IV.1, we have

$$H''(Z) = H_2''(Z),$$

which implies that

$$I''(X;Z) = H_2''(Z),$$

where we have used the fact that for a BSC(ε)

$$I(X;Z) = H(Z) - H(Z|X) = H(Z) - H(\varepsilon).$$

As shown in Table II, when $\varepsilon = 0.03$, $I(X;Z)$ is neither convex nor concave at $\vec{p} = (0.01, 0.81)$. Furthermore, when $\varepsilon = 0.2$, $I(X;Z)$ is convex around $\vec{p} = (0.01, 0.81)$, which together with

$$H(X|Z) = H(X) - I(X;Z), \tag{7}$$

implies that $H(X|Z)$ is concave around $\vec{p} = (0.01, 0.81)$.

TABLE II

ε	0.2		0.03	
$H_2''(Z)$	9.751	1.659	111.377	1.285
	1.659	0.342	1.285	0.248
Eigenvalues of $H_2''(Z)$	(10.035, 0.058)		(2.216, -0.591)	

With a BSC(ε) whose input is supported on a $(1, \infty)$ run-length-limited constraint [21] (and therefore Assumption (i) is not satisfied), the following example shows that the conjecture may be true in low SNR regime.

Example IV.5. Consider a BSC(ε) with a first-order input Markov chain supported on a $(1, \infty)$ run-length-limited constraint [21], which simply means that the string "11" is forbidden, a special case of the channel model considered in [1].

In the low SNR regime, that is, when ε is close to $1/2$, it has been shown in [23], [15] that

$$I(X;Z) = \frac{8(1 - \pi_{00})}{(2 - \pi_{00})^2} (1/2 - \varepsilon)^2 + o((1/2 - \varepsilon)^2). \tag{8}$$

For any $0 < \delta_1 < \delta_2 < 1$, consider p_{00} with $\delta_1 \le p_{00} \le \delta_2$. By Theorem 6.1 of [13], $H(Z)$ is analytic with respect to (p_{00}, ε) around $(p_{00}, 1/2)$, and so is $I(X;Z)$. So, the second-order derivative of the $o((1/2 - \varepsilon)^2)$-term in (3) with respect to p_{00} is still an $o((1/2 - \varepsilon)^2)$-term. It then follows from

$$\frac{d^2}{dp_{00}^2} \left(\frac{8(1 - \pi_{00})}{(2 - \pi_{00})^2} \right) = -4 < 0$$

that $I(X;Z)$ is concave with respect to p_{00} over $\delta_1 \le p_{00} \le \delta_2$ for ε sufficiently close to $1/2$. On the other hand, together with the fact that $H(X)$ is analytic in (p_{00}, ε) and strictly concave with respect to p_{00}, it follows from (7) and (8) that $H(X|Z)$ is also concave with respect to p_{00} over $\delta_1 \le p_{00} \le \delta_2$ for ε sufficiently close to $1/2$.

REFERENCES

[1] P. O. Vontobel, A. Kavčić, D. M. Arnold, and H.-A. Loeliger, "A generalization of the Blahut-Arimoto algorithm to finite-state channels," *IEEE Trans. Info. Theory*, vol. 54, no. 5, pp. 1887–1918, 2008.

[2] H. Kobayashi and D. T. Tang, "Application of partial-response channel coding to magnetic recording systems," *IBM Journal of Research and Development*, vol. 14, no. 4, pp. 368–375, 1970.

[3] M. Mushkin and I. Bar-David, "Capacity and coding for the gilbert-elliott channels," *IEEE Trans. Info. Theory*, vol. 35, no. 6, pp. 1277–1290, 1989.

[4] E. Zehavi and J. Wolf, "On runlength codes," *IEEE Trans. Info. Theory*, vol. 34, no. 1, pp. 45–54, 1988.

[5] B. Marcus, R. Roth, and P. Siegel, "Constrained systems and coding for recording channels," *Handbook of coding theory*, Vol. I, II. Amsterdam: North-Holland, pp. 1635–1764, 1998.

[6] G. Forney, Jr., "Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference," *IEEE Trans. Info. Theory*, vol. 18, no. 3, pp. 363–378, 1972.

[7] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech.J.*, vol. 27, pp. 379–423, 623–656, 1948.

[8] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Info. Theory*, vol. 18, no. 4, pp. 460–473, 1972.

[9] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Info. Theory*, vol. 18, no. 1, pp. 14–20, 1972.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[11] A. Goldsmith and P. Varaiya, "Capacity, mutual information, and coding for finite-state markov channels," *IEEE Trans. Info. Theory*, vol. 42, no. 3, pp. 868–886, 1996.

[12] G. Han and B. Marcus, "Concavity of the mutual information rate for input-restricted memoryless channels at high SNR," *IEEE Trans. Info. Theory*, vol. 58, no. 3, pp. 1534–1548, 2012.

[13] G. Han and B. Marcus, "Analyticity of entropy rate of hidden Markov chains," *IEEE Trans. Info. Theory*, vol. 52, no. 12, pp. 5251–5266, 2006.

[14] G. Han and B. Marcus, "Asymptotics of input-constrained binary symmetric channel capacity," *Ann. Appl. Probab.*, vol. 19, no. 3, pp. 1063–1091, 2009.

[15] G. Han and B. Marcus, "Derivatives of entropy rate in special families of hidden Markov chains," *IEEE Trans. Info. Theory*, vol. 53, no. 7, pp. 2642–2652, 2007.

[16] Y. Ephraim and N. Merhav, "Hidden Markov processes," *IEEE Trans. Info. Theory*, vol. 48, no. 6, pp. 1518–1569, 2002.

[17] W. Parry, "Intrinsic markov chains," *Transactions of the American Mathematical Society*, vol. 112, no. 1, pp. 55–66, 1964.

[18] E. Seneta, *Non-negative matrices and Markov chains*, Springer Series in Statistics. New York: Springer, 1981.

[19] G. Han and B. Marcus, "Asymptotics of entropy rate in special families of hidden Markov chains," *IEEE Trans. Info. Theory*, vol. 56, no. 3, pp. 1287–1295, 2010.

[20] J. Taylor, *Several Complex Variables With Connections to Algebraic Geometry and Lie Groups*, Graduate Studies in Mathematics. American Mathematical Society, 2002.

[21] D. Lind and B. Marcus, *An Introduction to Symbolic Dynamics and Coding*. Cambridge University Press, 1995.

[22] P. Jacquet, G. Seroussi, and W. Szpankowski, "On the entropy of a hidden Markov process," *Theoret. Comput. Sci.*, vol. 395, no. 2-3, pp. 203–219, 2008.

[23] H. D. Pfister, "The capacity of finite-state channels in the high-noise regime," *Entropy of hidden Markov processes and connections to dynamical systems*, London Math. Soc. Lecture Note Series. Cambridge: Cambridge Univ. Press, 2011, vol. 385, pp. 179–222.