A concavity result for output relative entropy

Devon Ding, Ken Lau, Chandra Nair, Dustin Wang

I. Introduction

The optimality of certain achievable rate regions for communication settings in multiuser information theory, such as Marton's region for the two-receiver broadcast channel, can be verified by establishing that product distributions are the global maximizers of a corresponding non-convex functional on product spaces, [1]. A functional satisfying the above property is said to satisfy global tensorization. As stated in [2] a curious connection has been repeatedly observed between functionals that satisfy global tensorization and those that satisfy a so-called local tensorization property. One way to reconcile this apparent relationship is to determine if all the local maximizers of non-convex functionals that satisfy global tensorization have the property that all local maximizers are also product distributions.

On a related note, information inequalities concerning non-convex functionals have also been established [3] by determining all the local maximizers. Additionally, certain non-convex functionals, such as the one arising in the capacity region computation of the vector Gaussian channel [4] is shown to have a unique local maximum. Inspired by these observations, we seek to understand the geometric structure of certain information functionals and determine its set of local extremizers. The family considered in this paper can be considered as an elementary but non-trivial sub-class of functionals. The results in this paper extend the celebrated convexity result, sometimes referred to as Mrs. Gerber's lemma, of Wyner and Ziv to broader family of channels.

This paper addresses a particular family of non-convex optimization problems that arises often-times in information theory. Given a conditional distribution $W_{Y|X}$, a reference distribution P_X, and a non-negative parameter λ we will be investigating non-convex optimization problems of form

$$\min_{\Phi_X} \{\lambda D(\Phi_X \| P_X) - D((W\Phi)_Y \| (WP)_Y)\}, \tag{1}$$

and see if these problems can be reparameterized into convex optimization problems. In the above expression $D(\Phi_X \| P_X)$ denotes the relative entropy, and the logarithms are assumed to be with respect to base e. If such a reparameterization exists, then any local minimizer would also be a global minimizer (similar to the observation in the MIMO Gaussian broadcast channel). The main idea is to choose a parameterization of Φ_X so that $D(\Phi_X \| P_X)$ is linear in the parameter and determine whether the output relative entropy, $D((W\Phi)_Y \| (WP)_Y)$, is concave.

A. Motivation

Consider the following optimization scenarios originating in multiuser information theory.

(i) In the Ahlswede-Korner source coding problem [5], to compute the minimal weight sum-rate, one is faced with the following optimization problem: Given a conditional distribution $W_{Y|X}$ and an input distribution Φ_X, one seeks to compute the value of the following optimization problem (parameterized by λ, $\lambda \geq 0$):

$$\min_{U: U \to X \to Y} H(Y|U) + \lambda I(U;X).$$

(ii) In the degraded broadcast channel, to compute the maximum weighted sum-rate $R_Z + \lambda R_Y$, one seeks to compute the value of the following optimization problem (parameterized by λ, $\lambda \leq 1$):

$$\max_{U,X: U \to X \to Y \to Z} I(U;Z) + \lambda I(X;Y|U).$$

Both of these problems result in the computation of the lower convex envelope with respect to Φ_X for the functionals $H(Y) - \lambda H(X)$ and $H(Z) - \lambda H(Y)$, respectively. Observe that in the latter case, the channel $W_{Z|Y}$ is fixed, and in the former case the conditional distribution $W_{Y|X}$ is fixed. Note that, when $\lambda = 0$ both functionals are concave in Φ_X and when $\lambda = 1$ both functionals are convex in Φ_X. For $\lambda \in (0,1)$ (the interesting regime), the function is not necessarily convex nor concave. Therefore computation of the lower convex envelope does not reduce to a convex optimization problem and apriori the functionals may have multiple local minimizers. Therefore it is natural to ask if there is a subset of the above family of problems for which under a suitable reparameterization, the problem reduces to a convex optimization problem.

Characterization of the lower convex envelope can be done via Fenchel duality by computing its supporting hyperplanes. To this end we seek to compute the minimum of

$$G(P_X) := \min_{\Phi_X} \{\lambda D(\Phi_X \| P_X) - D((W\Phi)_Y \| (WP)_Y)\}$$

$$= \min_{\Phi_X} \left\{ H_\Phi(Y) - \lambda H_\Phi(X) - \sum_x a_x \Phi_X(x) \right\},$$

where $(W\Phi)_Y$ denotes the distribution on Y induced by the input distribution Φ_X and the channel $W_{Y|X}$, $H_\Phi(X)$ denotes the Shannon entropy of X when $X \sim \Phi_X$, and $a_x = \sum_y W(y|x) \log \frac{(WP)_y}{P(x)^\lambda}$. Thus $G(P_X)$ denotes the Fenchel dual for the convex envelope of $H(Y) - H(X)$, with $a_x = \sum_y W(y|x) \log \frac{(WP)_y}{P(x)^\lambda}$ being the dual variables. This is one way in which optimization problems of the type described in (1) arise.

Another motivation for such optimization problems lie in determining the optimal constants for Strong-Data-Processing inequalities and in turn to determining limiting hypercontractivity parameters [6]. It has been shown in [7] that given $p_X, W_{Y|X}$, the inequality

$$I(U;Y) - \eta I(U;X) \leq 0,$$

holds for all $U : U \to X \to Y$ is Markov, if and only if, the inequality

$$\min_{\Phi_X} \{\lambda D(\Phi_X \| P_X) - D((W\Phi)_Y \| (WP)_Y)\} \geq 0$$

holds. Note that the range of η depends on P_X. One can also define a similar η that holds for all input distributions (and thus depends only on the channel) to be

$$\eta_W := \min\{\eta : I(U;Y) - \eta I(U;X) \leq 0, \quad \forall p_{UX} : U \to X \to Y \text{ is Markov}\}$$

or equivalently (see Exercise in [8])

$$\eta_W := \min\{\eta : \lambda D(\Phi_X \| P_X) - D((W\Phi)_Y \| (WP)_Y) \geq 0, \quad \forall q_X, p_X\}.$$

It has recently been shown [9] that for any $W_{Y|X}$ it suffices to consider P_X having support on two alphabets and $\Phi_X \ll P_X$ to compute η_W.

Remark 1. In light of this result, the case of $\mathcal{X}$ being binary takes particular significance while considering the family of optimization problems of the form

$$\min_{\Phi_X} \{\lambda D(\Phi_X \| P_X) - D((W\Phi)_Y \| (WP)_Y)\}, \tag{2}$$

$$\min_{P_X, \Phi_X} \{\lambda D(\Phi_X \| P_X) - D((W\Phi)_Y \| (WP)_Y)\}. \tag{3}$$

B. A convexity result due to Wyner and Ziv

While trying to compute the superposition coding region of a degraded binary-symmetric broadcast channel (see item (ii) in the Motivation), Wyner and Ziv showed that for any $\alpha \in [0, \frac{1}{2}]$, the function $H_2(\alpha * H_2^{-1}(u))$ is convex in u, where $H_2 : [0, \frac{1}{2}] \mapsto [0, \log 2]$ is binary entropy function given by $H_2(x) = -x \log_2(x) - (1-x)\log_2(1-x)$ and $H_2^{-1} : [0, \log 2] \mapsto [0, \frac{1}{2}]$ is its inverse. Here $a * b = a(1-b) + b(1-a)$ denotes a two-point convolution.

We can interpret this result alternatively as the following. Let $W_{Y|X}$ be the binary symmetric channel with crossover probability α. Let p_X be the uniform distribution and parameterize $\Phi_{X,t}(0) = H_2^{-1}(t)$. Now observe that under this parameterization, $D(\Phi_{X,t} \| P_X) = \log 2 - t$ is linear in t, and $D((W\Phi_{X,t})_Y \| (WP)_Y) = \log 2 - H_2(\alpha * H_2^{-1}(t))$ is concave in t. Therefore the function

$$\{\lambda D(\Phi_{X,t} \| P_X) - D((W\Phi_{X,t})_Y \| (WP)_Y)\}$$

is convex in t, reducing the computation of (1) to a convex optimization problem. One way to interpret this is, $\Phi_{X,t}$ determines a path along the binary simplex such that $D(\Phi_{X,t} \| P_X)$ is linear in t and $D((W\Phi_{X,t})_Y \| (WP)_Y)$ is concave in t.

Thus the question we seek to address is: given any channel $W_{Y|X}$, a reference distribution P_X, and an initial distribution Φ_X, is it possible to parameterize the path from Φ_X to P_X according to $\Phi_{X,t}$, where $\Phi_{X,0} = \Phi_X$ and $\Phi_{X,1} = P_X$, with the property that $D(\Phi_{X,t} \| P_X)$ is linear in t and $D((W\Phi_{X,t})_Y \| (WP)_Y)$ is concave in t. We will answer this question for channels with binary inputs, and partially for channels with higher input cardinalities. As stated in Remark 1, the case of binary inputs (and outputs of arbitrary cardinality) is particularly useful when computing η_W for a channel with arbitrary input alphabet.

II. Channels with binary inputs and binary outputs

Let us denote a binary-input binary-output channel as

$$W_{Y|X} = \begin{bmatrix} a & b \\ \bar{a} & \bar{b} \end{bmatrix}. \tag{4}$$

Here the matrix entry $W_{ij} = P(Y = i|X = j)$. Let us denote $\Phi_{X,t} = (\phi(t), 1-\phi(t))$ and $P_X = (p, 1-p)$ to characterize the parameterized path and the reference distribution. Further we denote, for $a, b \in [0,1]$,

$$D_2(a\|b) = a \log \frac{a}{b} + (1-a)\log\frac{1-a}{1-b}$$

to be the relative entropy between the two two-point distributions characterized by $(a, 1-a)$ and $(b, 1-b)$ respectively. We also use $\bar{\phi}$ to represent $1-\phi$ for brevity. We also assume that the reference measure satisfies $p > 0$; otherwise $D_2(\phi\|0) = \infty$ for all $\phi \neq 0$.

Note that $D_2(\phi(t)\|p)$ is monotonically increasing (resp. decreasing) when $\phi(t) \geq p$ (resp. $\phi(t) \leq p$). Hence if we enforce the linear dependence of input divergence on t, we obtain

$$\frac{d^2}{dt^2} D_2(\phi\|p) = \phi'' \log\frac{\phi\bar{p}}{\bar{\phi}p} + \frac{\phi'^2}{\phi\bar{\phi}} = 0. \tag{5}$$

Imposing the boundary conditions $\phi(0) = 1$ (resp. $\phi(0) = 1$) and $\phi(1) = p$, then $\phi(t)$ can be uniquely determined (due to the monotonicity of $D_2(\phi(t)\|p)$).

Remark 2. For an arbitrary reference P_X, this reparameterization $\phi(t)$ replaces the parameterization $H_2^{-1}(t)$ employed by Wyner and Ziv.

Let $(WP)_Y = (ap + b\bar{p}, \bar{a}p + \bar{b}\bar{p})$ and $(W\Phi)_Y = (a\phi + b\bar{\phi}, \bar{a}\phi + \bar{b}\bar{\phi})$. We define $q \triangleq ap + b\bar{p}$, and $\psi \triangleq a\phi + b\bar{\phi}$. Now we can calculate the second order derivative $\frac{d^2}{dt^2} D_2(\psi\|q)$ as the following:

$$\begin{aligned} \frac{d^2}{dt^2} D_2(\psi\|q) &= \psi'' \log\frac{\psi\bar{q}}{\bar{\psi}q} + \frac{\psi'^2}{\psi\bar{\psi}} \\ &= -(a-b)\frac{\phi'^2}{\phi\bar{\phi}} \log\frac{\psi\bar{q}}{\bar{\psi}q} / \log\frac{\phi\bar{p}}{\bar{\phi}p} + (a-b)^2 \frac{\phi'^2}{\psi\bar{\psi}} \end{aligned} \tag{6}$$

In the second step we used Equation (5). Suppose $\phi' \neq 0$, then concavity of $D((W\Phi_{X,t})_Y\|(WP)_Y)$ is equivalent to $\frac{d^2}{dt^2} D_2(\psi\|q) \leq 0$. This, in turn, is equivalent to

$$f(\phi; p) := (a-b)^2 \phi\bar{\phi} \log\frac{\phi\bar{p}}{\bar{\phi}p} - (a-b)\psi\bar{\psi}\log\frac{\psi\bar{q}}{\bar{\psi}q} \begin{cases} \geq 0, & \phi \leq p; \\ \leq 0, & \phi \geq p; \end{cases} \tag{7}$$

since $\log\frac{\phi}{p} \leq 0$ (resp. ≥ 0) when $\phi \leq p$ (resp. $\phi \geq p$).

Specifically, the condition (7) now can be expressed as $f(\phi; p) \geq 0$ for $\phi \in [0, p]$ and $f(\phi; p) \leq 0$ for $\phi \in [p, 1]$. We remark that this condition now does not depend on t. One may calculate the derivatives of $f(\phi; p)$ w.r.t. ϕ as the following.

$$\begin{aligned} \frac{d}{d\phi} f(\phi; p) &= (a-b)^2 (1-2\phi)\log\frac{\phi\bar{p}}{\bar{\phi}p} - (a-b)^2(1-2\psi)\log\frac{\psi\bar{q}}{\bar{\psi}q} \\ \frac{d^2}{d\phi^2} f(\phi; p) &= (a-b)^2 [-2\log\frac{\phi\bar{p}}{\bar{\phi}p} + \frac{(1-2\phi)}{\phi\bar{\phi}}] - (a-b)^3[-2\log\frac{\psi\bar{q}}{\bar{\psi}q} + \frac{(1-2\psi)}{\psi\bar{\psi}}] \\ \frac{d^3}{d\phi^3} f(\phi; p) &= -\frac{(a-b)^2}{\phi^2\bar{\phi}^2} + \frac{(a-b)^4}{\psi^2\bar{\psi}^2}. \end{aligned} \tag{8}$$

We will show that $\frac{d^2}{dt^2} f(\phi)$ is decreasing w.r.t. ϕ in the following lemma.

Lemma 1. *The second-order derivative $\frac{d^2}{d\phi^2} f(\phi; p)$ is monotonically decreasing in $\phi \in [0,1]$.*

Proof. Note that $\frac{d^3}{d\phi^3} f(\phi; p) \leq 0$ is equivalent to $\psi^2\bar{\psi}^2 \geq (a-b)^2\phi^2\bar{\phi}^2$ or

$$(\psi\bar{\psi} + (a-b)\phi\bar{\phi})(\psi\bar{\psi} - (a-b)\phi\bar{\phi}) \geq 0.$$

When $0 \leq b \leq a \leq 1$, we have $\psi\bar{\psi} + (a-b)\phi\bar{\phi} \geq 0$. Hence we only need to argue $\psi\bar{\psi} - (a-b)\phi\bar{\phi} \geq 0$. Note we have

$$\begin{aligned}
\psi\bar{\psi} - (a-b)\phi\bar{\phi} &= (b + (a-b)\phi)(\bar{b} + (b-a)\phi) - (a-b)\phi(1-\phi) \\
&= (a-b)(\bar{a}+b)\phi^2 - 2b(a-b)\phi + b\bar{b} \\
&= (a-b)(\bar{a}+b)\Big(\phi - \frac{b}{\bar{a}+b}\Big)^2 + \frac{\bar{a}b}{\bar{a}+b} \geq 0.
\end{aligned}$$

Similarly, when $0 \leq a \leq b \leq 1$, we have $(\psi\bar{\psi} - (a-b)\phi\bar{\phi}) \geq 0$. Hence we need to check

$$\psi\bar{\psi} + (a-b)\phi\bar{\phi} = (b-a)(a+\bar{b})\Big(\phi - \frac{\bar{b}}{a+\bar{b}}\Big)^2 + \frac{a\bar{b}}{a+\bar{b}} \geq 0.$$

This proves the required inequality. $\square$

Theorem 1. *Consider a binary channel represented as Equation* (4)*, with $a \neq b$ and $a, b \in (0,1)$. Assume that the input distribution is reparametrized according to Equation* (5)*, then $D((W\Phi_{X,t})_Y \| (WP)_Y) = D_2(\psi\|q)$ is concave w.r.t. t under such a reparametrization, if and only if p is equal to*

$$p^* := \frac{\sqrt{b\bar{b}}}{\sqrt{b\bar{b}} + \sqrt{a\bar{a}}}.$$

Proof. We will first show that $p = p^*$ is necessary. Calculate the Taylor expansion of $f(\phi;p)$ at $\phi = p$, and observe that $f(p;p) = \frac{d}{d\phi}f(p;p) = 0$, we have $f(p+\epsilon;p) = \frac{\epsilon^2}{2}\frac{d^2}{d\phi^2}f(p;p) + O(\epsilon^3)$. Hence to satisfy the condition in (7). i.e. for

$$f(\phi;p) \begin{cases} \geq 0, & \phi \leq p; \\ \leq 0, & \phi \geq p; \end{cases}$$

we must have $\frac{d^2}{d\phi^2}f(p;p) = 0$. By Equation (8), this is equivalent to

$$(a-b)\frac{1-2q}{q\bar{q}} - \frac{1-2p}{p\bar{p}} = 0.$$

The above equation is quadratic in p and the only feasible solution is p^*. Hence $p = p^*$ is necessary.

To show that it is sufficient, assume $p = p^*$. From Lemma 1, we have that $\frac{d^2}{d\phi^2}f(\phi;p^*)$ is decreasing. Since $\frac{d^2}{d\phi^2}f(p^*;p^*) = 0$, then $\frac{d^2}{d\phi^2}f(\phi;p^*) \leq 0$ for $\phi \geq p^*$. This implies that $\frac{d}{d\phi}f(\phi;p^*)$ is decreasing for $\phi \geq p^*$. As $\frac{d}{d\phi}f(p^*;p^*) = 0$, we have $\frac{d}{d\phi}f(\phi;p^*) \leq 0$ for $\phi \geq p^*$. Consequently $f(\phi;p^*)$ is decreasing for $\phi \geq p^*$. Finally, as $f(p^*;p^*) = 0$, we obtain $f(\phi;p^*) \leq 0$ when $\phi \geq p^*$. The analysis of $\phi \leq p^*$ is similar. This completes the proof. $\square$

Remark 3. This theorem implies that for the binary symmetric channel, the only P_X for which we have the concavity of $D((W\Phi_{X,t})_Y \| (WP)_Y)$ with respect to t is the uniform distribution.

When $p \neq p^*$, the next proposition establishes a one-sided concavity result for the output relative entropy.

Proposition 1. *In the same setting as Theorem 1, if $p > p^*$, $D_2(\psi\|q)$ is concave for all $\phi \geq p$. Similarly, if $p < p^*$, $D_2(\psi\|q)$ is concave for all $\phi \leq p$.*

Proof. We will prove the claim when $p > p^*$. The case where $p < p^*$ is analogous. We will show that $\frac{d^2}{d\phi^2}f(p;p)$ is decreasing w.r.t. p first. By Equation (8), we have

$$g(p) := \frac{d^2}{d\phi^2}f(p;p) = (a-b)^2\frac{1-2p}{p\bar{p}} - (a-b)^3\frac{1-2q}{q\bar{q}}.$$

Since q is a function of p, we deduce that

$$\begin{aligned}
g(p) &= (a-b)^2\Big(\frac{1}{p} - \frac{1}{\bar{p}}\Big) - (a-b)^3\Big(\frac{1}{q} - \frac{1}{\bar{q}}\Big) \\
\frac{d}{dp}g(p) &= (a-b)^2\Big(-\frac{1}{p^2} - \frac{1}{\bar{p}^2}\Big) - (a-b)^4\Big(-\frac{1}{q^2} - \frac{1}{\bar{q}^2}\Big) \\
&= -(a-b)^2\frac{(2(a-b)bp + b^2)}{p^2q^2} - (a-b)^2\frac{(2(a-b)\bar{a}\bar{p} + \bar{a}^2)}{\bar{p}^2\bar{q}^2} \qquad (9) \\
&= -(b-a)^2\frac{(2(b-a)\bar{b}p + \bar{b}^2)}{p^2\bar{q}^2} - (b-a)^2\frac{(2(b-a)a\bar{p} + a^2)}{\bar{p}^2q^2} \qquad (10)
\end{aligned}$$

Therefore, irrespective of the sign of $a-b$ (see (9) or (10)) , we have $\frac{d}{dp}g(p) \le 0$. Since $g(p^*) = \frac{d^2}{d\phi^2}f(p^*;p^*) = 0$ and $g(p)$ is decreasing w.r.t p, $g(p) \le 0$ when $p \ge p^*$. Moreover, by Lemma 1, we have $\frac{d^2}{d\phi^2}f(\phi;p)$ is decreasing w.r.t ϕ and hence $\frac{d^2}{d\phi^2}f(\phi;p) \le 0$ for all $\phi \ge p$. Since $f(p;p) = 0$ and $\frac{d}{d\phi}f(p;p) = 0$, we have $f(\phi;p) \le 0$ for all $\phi \ge p$. This implies $D(\boldsymbol{\psi}\|\boldsymbol{q})$ is concave with t when $\phi \ge p$. □

III. Concavity over a 2-to-n Channel

We now generalize our result from binary outputs to 2-to-n channels for arbitrary finite output dimension n. To do so, we follow the same approach to find the p such that when we make the input divergence linear in t, the output divergence becomes concave in t. The key different is that one is unable to explicitly identify the p^*, We denote the channel as

$$W(y|x) = \begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \\ \vdots & \vdots \\ a_n & b_n \end{bmatrix} \tag{11}$$

Here the matrix entry $W_{ij} = P(Y=i|X=j)$. The differential equation that makes the input divergence $D_2(\phi\|p)$ linear is the same as the binary case, as is shown in Equation (5). However, the expression for the output divergence $D((W\Phi_{X,t})_Y\|(WP)_Y)$ is different. Define $q_i = a_i p + b_i \bar{p}$ and $\psi_i = a_i\phi + b_i\bar{\phi}$. Denote $D((W\Phi_{X,t})_Y\|(WP)_Y) = D(\boldsymbol{\psi}\|\boldsymbol{q})$. We have

$$\begin{aligned} D(\boldsymbol{\psi}\|\boldsymbol{q}) &= \sum_{i=1}^n \psi_i \log\frac{\psi_i}{q_i} \\ \frac{d^2}{dt^2}D(\boldsymbol{\psi}\|\boldsymbol{q}) &= \sum_{i=1}^n \psi_i'' \log\frac{\psi_i}{q_i} + \frac{\psi_i'^2}{\psi_i} \\ &= \sum_{i=1}^n \left((a_i-b_i)\phi''\log\frac{\psi_i}{q_i} + \frac{(a_i-b_i)^2\phi'^2}{\psi_i}\right) \\ &= \sum_{i=1}^n \left(-(a_i-b_i)\frac{\phi'^2}{\phi\bar{\phi}}\log\frac{\psi_i}{q_i}\left(\log\frac{\phi\bar{p}}{\bar{\phi}p}\right)^{-1} + \frac{(a_i-b_i)^2\phi'^2}{\psi_i}\right). \end{aligned} \tag{12}$$

Here we used Equation (5) in the final step. Requiring, the output relative entropy to be concave, i.e. the second-order derivative to be negative, is then equivalent to

$$f(\phi;p) := \sum_{i=1}^n \left(-(a_i-b_i)\log\frac{\psi_i}{q_i} + (a_i-b_i)^2\frac{\phi\bar{\phi}}{\psi_i}\log\frac{\phi\bar{p}}{\bar{\phi}p}\right) \quad \begin{cases} \ge 0, & 0 \le \phi \le p; \\ \le 0, & p \le \phi \le 1. \end{cases}$$

Taking derivatives of $f(\phi;p)$ w.r.t. ϕ, we have

$$\begin{aligned} \frac{d}{d\phi}f(\phi;p) &= \log\frac{\phi\bar{p}}{\bar{\phi}p}\sum\frac{(a_i-b_i)^2(-a_i\phi^2+b_i\bar{\phi}^2)}{(a_i\phi+b_i\bar{\phi})^2} \\ &= \log\frac{\phi\bar{p}}{\bar{\phi}p}\sum\frac{(a_i-b_i)(-a_i^2\phi^2+a_ib_i(\phi^2+\bar{\phi}^2)-b_i^2\bar{\phi}^2)}{(a_i\phi+b_i\bar{\phi})^2} \\ &= \log\frac{\phi\bar{p}}{\bar{\phi}p}\sum\frac{(a_i-b_i)(-a_i^2\phi^2+a_ib_i(1-2\phi\bar{\phi})-b_i^2\bar{\phi}^2)}{(a_i\phi+b_i\bar{\phi})^2} \\ &= \log\frac{\phi\bar{p}}{\bar{\phi}p}\sum\frac{(a_i-b_i)(-(a_i\phi+b_i\bar{\phi})^2+a_ib_i)}{(a_i\phi+b_i\bar{\phi})^2} \\ &\overset{(a)}{=} \log\frac{\phi\bar{p}}{\bar{\phi}p}\sum\frac{a_ib_i(a_i-b_i)}{(a_i\phi+b_i\bar{\phi})^2}. \end{aligned}$$

The second derivative can be expressed in terms of the first derivative according to

$$\begin{aligned} \frac{d^2}{d\phi^2}f(\phi;p) &= \frac{1}{\phi\bar{\phi}}\sum\frac{a_ib_i(a_i-b_i)}{(a_i\phi+b_i\bar{\phi})^2} - 2\log\frac{\phi\bar{p}}{\bar{\phi}p}\sum\frac{(a_i-b_i)^2a_ib_i}{(a_i\phi+b_i\bar{\phi})^3} \\ &= \frac{1}{\phi\bar{\phi}}\left(\frac{d}{d\phi}f(\phi;p)\right)\left(\log\frac{\phi\bar{p}}{\bar{\phi}p}\right)^{-1} - 2\log\frac{\phi\bar{p}}{\bar{\phi}p}\sum\frac{(a_i-b_i)^2a_ib_i}{(a_i\phi+b_i\bar{\phi})^3}. \end{aligned} \tag{13}$$

Finally the third derivative can be expressed as

$$\frac{d^3}{d\phi^3} f(\phi;p) = \left(\frac{1}{\bar{\phi}^2} - \frac{1}{\phi^2}\right) \sum \frac{a_i b_i (a_i - b_i)}{(a_i\phi + b_i\bar{\phi})^2} + 6 \log \frac{\phi \bar{p}}{\bar{\phi} p} \sum \frac{(a_i - b_i)^3 a_i b_i}{(a_i \phi + b_i \bar{\phi})^4}$$
$$- \frac{4}{\phi\bar{\phi}} \sum \frac{a_i b_i (a_i - b_i)^2}{(a_i\phi + b_i\bar{\phi})^3}.$$

We can now generalize Theorem 1 to 2-to-n channels.

Theorem 2. *For a 2-to-n channel represented as Equation (11), if $\boldsymbol{a} \neq \boldsymbol{b}$, and we reparametrize the input distribution according to Equation (5), then of $D((W\Phi_{X,t})_Y \| (WP)_Y) = D(\boldsymbol{\psi}\|\boldsymbol{q})$ is concave w.r.t. t under such reparametrization, if and only if $p = p^*$ where p^* is the unique solution to*

$$\sum_{i=1}^{n} \frac{(a_i - b_i) a_i b_i}{(pa_i + \bar{p} b_i)^2} = 0. \tag{14}$$

Proof. Let $g(p) := \sum_{i=1}^n \frac{(a_i-b_i)a_ib_i}{(a_ip+b_i\bar{p})^2} = 0$. Observe that $g(p)$ is decreasing since

$$\frac{d}{dp} g(p) = -\sum_{i=1}^{n} \frac{2(a_i - b_i)^2 a_i b_i}{(a_i p + b_i \bar{p})^3} \leq 0.$$

Since $g(0) = \sum_{i=1}^n \frac{(a_i-b_i)^2}{b_i} \geq 0$, $g(1) = -\sum_{i=1}^n \frac{(a_i-b_i)^2}{a_i} \leq 0$, we conclude that $g(p)$ has a unique zero over $[0,1]$ so long as $\boldsymbol{a} \neq \boldsymbol{b}$.

Since $f(p;p) = \frac{d}{d\phi} f(p;p) = 0$ for any p, by considering the Taylor expansion at $\phi = p$ we see that the condition

$$f(\phi;p) \quad \begin{cases} \geq 0, & \phi \leq p; \\ \leq 0, & \phi \geq p; \end{cases}$$

forces $\frac{d^2}{d\phi^2} f(p;p) = 0$. Therefore from Equation (13), require $g(p) = 0$ or that $p = p^*$ is necessary.

We now argue that the above condition is also sufficient. By Equation(13), we have $f(p^*;p^*) = \frac{d}{d\phi} f(p^*;p^*) = \frac{d^2}{d\phi^2} f(p^*;p^*) = 0$, and

$$\frac{d^3}{d\phi^3} f(p^*;p^*) = -\frac{4}{p^* \bar{p}^*} \sum \frac{(a_i - b_i)^2 a_i b_i}{(a_i p^* + b_i \bar{p}^*)^3} \leq 0. \tag{15}$$

Using Lemma 2 completes the proof. □

Lemma 2. *Consider a real function $f(\phi) : (0,1) \to \mathbf{R}$ and assume $f \in \mathbf{C}^4$, i.e. four times differentiable,and satisfies the following properties:*

1) $f(p) = f'(p) = f''(p) = 0$, and $f'''(p) < 0$ for some $p \in (0,1)$;

2) $f''(\phi) = a(\phi) \cdot f'(\phi) + b(\phi)$, where $a(\phi) > 0$ and $b(\phi) \leq 0$ for $\phi \in (p,1)$; while $a(\phi) < 0$ and $b(\phi) \geq 0$ for $\phi \in (0,p)$.

Then we have $f(\phi) \leq 0$ for $\phi \in (p,1)$. Similarly, $f(\phi) \geq 0$ for $\phi \in (0,p)$.

Proof. From the Taylor expansion at p, we have $f'(\phi) = \frac{f'''(p)}{2}(\phi - p)^2 + O((\phi - p)^3)$. Since $f'''(p)$ is strictly less than zero, then there must exist some positive constant $q \in (p,1)$, such that for $p < \phi \leq q$, we have $f'(\phi) < 0$. Suppose there is some $s \in (q,1)$, such that $f'(s) > 0$. We then deduce that the minimum of $f'(\phi)$ over $\phi \in [p,s]$ must be attained by some interior minimizer $\phi_0 \in (p,s)$, and $f'(\phi_0) < 0$. Also we have $f''(\phi_0) = 0$ by local optimality conditions for interior minimizers. Since $a(\phi) > 0$ and $b(\phi) \leq 0$ for $\phi \in (p,1)$, we obtain

$$0 = f''(\phi_0) = a(\phi_0) \cdot f'(\phi_0) + b(\phi_0) \leq a(\phi_0) \cdot f'(\phi_0) < 0.$$

Contradiction! Hence such an s cannot exist. This guarantees $f'(\phi) \leq 0$ for $\phi \in (p,1)$ and therefore $f(\phi) \leq 0$ for $\phi \in (p,1)$. The other side can be proved by similar arguments. □

We then give an alternate proof of the sufficiency part in Theorem 2 as the following, without needing the above Lemma.

Alternate Proof of sufficiency of $p = p^*$. We note that $\frac{d}{d\phi} f(\phi;p^*) = g(\phi) \log \frac{\phi \bar{p}^*}{\bar{\phi} p^*}$. Here $g(\phi) := \sum_{i=1}^n \frac{(a_i-b_i)a_ib_i}{(a_ip+b_i\bar{p})^2}$ as is defined in the previous proof. Then we know that $g(\phi)$ is decreasing over $\phi \in [0,1]$, and hence $g(\phi) \leq 0$ for $\phi \geq p^*$. Also note that $\log \frac{\phi \bar{p}^*}{\bar{\phi} p^*} \geq 0$ for $\phi \geq p^*$. Then we have $\frac{d}{d\phi} f(\phi;p^*) \leq 0$ for $\phi \geq p^*$, which further guarantees $f(\phi;p^*) \leq 0$ for $\phi \geq p^*$. The other side ($\phi \leq p^*$) can be analyzed analogously. □

Proposition 2. *In the same setting as Theorem 2, if $p > p^*$, $D(\boldsymbol{\psi}\|\boldsymbol{q})$ is concave for all $\phi \geq p$. Similarly, if $p < p^*$, $D(\boldsymbol{\psi}\|\boldsymbol{q})$ is concave for all $\phi \leq p$.*

Proof. Note that $\frac{d}{d\phi}f(\phi;p) = \ln\frac{\phi\bar{p}}{\bar{\phi}p}\cdot g(p)$. Consider the case $p \leq p^*$. When $\phi \leq p$, we have $g(p) \geq 0$, $\ln\frac{\phi\bar{p}}{\bar{\phi}p} \leq 0$, hence $\frac{d}{d\phi}f(\phi;p) \leq 0$. But $f(p;p) = 0$, so $f(\phi;p) \geq 0$ in $[0,p]$. The other side can be proved similarly. □

IV. Concavity over an m-to-n Channel

When the cardinality of the input alphabet is increased from binary we lose the uniqueness of the path from a given q_X to the reference p_X. In the following lemma, we show that there are multiple paths that pass through q_X (but not necessarily through p_X) which turn the optimization problem in (1) into a convex optimization problem. This is done by identifying each trajectory as a corresponding path between two inputs for a related channel with binary inputs.

Given a m-to-n channel $W_{Y|X}$, where $W_{ij} = P(Y=i|X=j)$, and fix any two input co-ordinates, say the first two. Let $P_X = (p_1, ..., p_m)$ be an arbitrary reference distribution. Let $\Phi_{X,t} = (\phi_1(t), \phi_2(t), p_3, .., p_m)$ denote a path along the input distributions. Note that $\phi_1(t) + \phi_2(t) = p_1 + p_2$. Let $\hat{\Phi}(t) = \left(\frac{\phi_1(t)}{\phi_1(t)+\phi_2(t)}, \frac{\phi_2(t)}{\phi_1(t)+\phi_2(t)}\right)$ and $\hat{P} = \left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right)$. Define a binary input channel $\hat{W}_{Y|\hat{X}}$ (observe that this depends on on $W_{Y|X}$ and the reference measure p_X) according to

$$a_i = \hat{W}(Y=i|X=1) = W_{i1}(p_1+p_2) + \sum_{j=3}^{m} W_{ij}p_j$$
$$b_i = \hat{W}(Y=i|X=2) = W_{i2}(p_1+p_2) + \sum_{j=3}^{m} W_{ij}p_j.$$

Lemma 3. *The following hold:*

$$D(\Phi_{X,t}\|P_x) = (p_1+p_2)D(\hat{\Phi}(t)\|\hat{P})$$
$$D((W\Phi)_Y\|(WP)_Y) = D((\hat{W}\hat{\Phi})_Y\|(\hat{W}\hat{P})_Y)$$

Proof. The first equality is immediate by direct substitution. For the second observe that

$$\begin{aligned}
D((\hat{W}\hat{\Phi})_Y\|(\hat{W}\hat{P})_Y) &= \left(\sum_{i=1}^{n} a_i\frac{\phi_1(t)}{\phi_1(t)+\phi_2(t)} + b_i\frac{\phi_1(t)}{\phi_1(t)+\phi_2(t)}\right)\log\frac{a_i\phi_1(t)+b_i\phi_2(t)}{a_ip_1+b_ip_2} \\
&= \sum_{i=1}^{n}\left(W_{i1}\phi_1(t) + W_{i2}\phi_2(t) + \sum_{j=3}^{m}W_{ij}p_j\right)\log\frac{W_{i1}\phi_1(t)+W_{i2}\phi_2(t)+\sum_{j=3}^{m}W_{ij}p_j}{W_{i1}p_1+W_{i2}p_2+\sum_{j=3}^{m}W_{ij}p_j} \\
&= D((W\Phi)_Y\|(WP)_Y).
\end{aligned}$$

□

References

[1] V. Anantharam, A. Gohari, and C. Nair, "On the evaluation of marton's inner bound for two-receiver broadcast channels," *IEEE Transactions on Information Theory*, vol. 65, no. 3, pp. 1361–1371, March 2019.
[2] C. Nair, "On Marton's achievable region: Local tensorization for product channels with a binary component," in *2020 Information Theory and Applications Workshop (ITA)*, 2020.
[3] Y. Geng, V. Jog, C. Nair, and Z. V. Wang, "An information inequality and evaluation of Marton's inner bound for binary input broadcast channels," *IEEE Transactions on Information Theory*, vol. 59, no. 7, pp. 4095–4105, 2013.
[4] H. Weingarten, Y. Steinberg, and S. S. Shamai, "The capacity region of the Gaussian multiple-input multiple-output broadcast channel," *IEEE Transactions on Information Theory*, vol. 52, no. 9, pp. 3936–3964, Sept 2006.
[5] R. F. Ahlswede and J. Körner, "Source coding with side information and a converse for degraded broadcast channels," *IEEE Trans. Info. Theory*, vol. IT-21(6), pp. 629–637, November, 1975.
[6] R. Ahlswede and P. Gács, "Spreading of sets in product spaces and hypercontraction of the markov operator," *The Annals of Probability*, pp. 925–939, 1976.
[7] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, "On hypercontractivity and a data processing inequality," in *2014 IEEE International Symposium on Information Theory (ISIT'2014)*, Honolulu, USA, Jun. 2014, pp. 3022–3026.
[8] I. Csiszar and J. Korner, *Information theory: Coding theorems for discrete memoryless systems*. Cambridge University Press, 1 2011.
[9] O. Ordentlich and Y. Polyanskiy, "Strong data processing constant is achieved by binary inputs," 2020.