EDIC Semester Project: Spring 2023
On Characterizing Nonlinear Strong Data Processing Inequalities

Adway Girish
Supervisor: Prof. Emre Telatar
Information Theory Laboratory (LTHI)

Last Updated: June 14, 2023

Abstract

The data processing inequality is a fundamental result in information theory. It states that no further processing of data can generate new information that is not already present, or equivalently, the information contained in any dataset can only decrease on further processing. This statement can be strengthened by quantifying this decrease in information as a scaling by a constant factor strictly smaller than 1, leading to *strong* data processing inequalities. However, it has since been observed that such linear relations do not fully capture the true characteristics of the decrease in information. Thus we study *nonlinear* data processing inequalities, which we call data processing functions, and conjecture that these functions are concave. We also study a specific example of these functions, namely for the KL divergence over the binary symmetric channel, and see that even for this simple case, the conjecture is difficult to prove.

1 Introduction

The data processing inequality (DPI) [1] states that for any random variables $U - X - Y$ forming a Markov chain, $I(U;X) \geq I(U;Y)$, i.e., further processing can only decrease the statistical dependence as measured by the mutual information. It is natural to ask if we can say something stronger about how much the information decreases, than simply that it does — the answer is yes, through functions that exactly track how much the information decreases. The chain of developments leading to these functions, through strong data processing inequalities, is described in Section 3, after we explain the notation used and provide definitions of standard quantities in Section 2. We then conjecture that this curve is concave, and attempt to prove it in a special case, in Section 4.

The DPI in its basic form is a seemingly simple observation has been used in information theory to great effect, particularly in proving several classical converse results. More recently, there has also been increasing interest from the machine learning community, particularly since the introduction of the Information Bottleneck (IB) paradigm [2], [3] to provide theoretical justifications for machine learning algorithms. The IB problem asks how much information can be extracted about a random variable by observing a function of it, without overfitting to the observations. This is made formal and the connection is explained at the end, in Section 5.

2 Preliminaries

For completeness, we first provide a summary of the notation used, and then define some information measures.

2.1 Notation

Let $\mathcal{X}$ and $\mathcal{Y}$ be two finite sets. For any pair of random variables (X, Y) on $\mathcal{X} \times \mathcal{Y}$, with the joint probability mass function (pmf) P_{XY}, which we denote as $(X, Y) \sim P_{XY}$, we let P_X and P_Y be the marginal pmfs of X and Y respectively. Further, let $\mathsf{W}_{Y|X}$ or simply W denote the conditional pmf of Y given X, i.e., $\mathsf{W}(y \mid x) = \frac{P_{XY}(x,y)}{P_X(x)}$, called the *channel* from X to Y. Given any *input* distribution P_X on $\mathcal{X}$ and a channel W from $\mathcal{X}$ to $\mathcal{Y}$, we define the corresponding *output* distribution $P_Y = \mathsf{W} \circ P_X$ as the marginal distribution induced by the channel, i.e., $P_Y(y) = \sum_{x \in \mathcal{X}} P_X(x)\, \mathsf{W}(y \mid x)$. The expectation of any function f of a random variable X with distribution P is given by $\mathbb{E}_P[f(X)] \triangleq \sum_{x \in \mathcal{X}} P(x)\, f(x)$; we simply say $\mathbb{E}[f(X)]$ when the distribution is clear. We denote the support of a distribution P by $\mathrm{supp}(P) = \{x \in \mathcal{X} : P(x) > 0\}$. For $a, b \in [0, 1]$, we write $\bar{a}$ to denote $1 - a$ and $a * b$ for $a\bar{b} + \bar{a}b$. All logarithms are taken to the base 2.

2.2 Information measures

Given two distributions P and Q on $\mathcal{X}$, we say that P is *absolutely continuous* w.r.t. Q, if $Q(x) = 0$ for some $x \in \mathcal{X}$ implies that $P(x) = 0$; this is denoted as $P \ll Q$. Then, for any convex function $f : (0, \infty) \to \mathbb{R}$ that is strictly convex at 1 and $f(1) = 0$, we define the *f-divergence* of P and Q with $P \ll Q$ as

$$D_f(P \,||\, Q) \triangleq \mathbb{E}_Q\left[f\left(\frac{P}{Q}\right)\right] = \sum_{x \in \mathrm{supp}(Q)} Q(x) f\left(\frac{P(x)}{Q(x)}\right).$$

In particular, when $f(x) = x \log x$, we have the KL divergence,

$$D_{\mathsf{KL}}(P \,||\, Q) \triangleq D_{t \mapsto t \log t}(P \,||\, Q) = \mathbb{E}_Q\left[\frac{P}{Q} \log \frac{P}{Q}\right] = \sum_{x \in \mathrm{supp}(Q)} P(x) \log \frac{P(x)}{Q(x)},$$

and when $f(x) = \frac{1}{2}|x - 1|$, we have the total variation (TV) distance,

$$D_{\mathsf{TV}}(P \,||\, Q) \triangleq D_{t \mapsto \frac{1}{2}|t-1|}(P \,||\, Q) = \mathbb{E}_Q\left[\frac{1}{2}\left|\frac{P}{Q} - 1\right|\right] = \frac{1}{2} \sum_{x \in \mathcal{X}} |P(x) - Q(x)|.$$

The mutual information between random variables X and Y is related to the KL divergence as $I(X; Y) = D_{\mathsf{KL}}(P_{XY} \,||\, P_X P_Y)$, and the entropy of a random variable is the self-information $I(X; X)$. All these quantities can also be extended to arbitrary random variables by suitably replacing pmfs with the Radon-Nikodym derivative, as

$$D_f(P \,||\, Q) \triangleq \mathbb{E}_Q\left[f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\right] = \int_{\mathcal{X}} f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}Q(x).$$

The binary entropy function is the entropy of a Bernoulli(p) random variable (i.e., one that takes the value 1 with probability p and 0 otherwise), given by

$$h_2(p) \triangleq -p \log p - (1 - p) \log(1 - p).$$

3 Data Processing Inequalities: From Classical to Strong to Non-linear

We define the input-dependent and independent *contraction coefficients* [4], [5] as

$$\eta_I(\mathsf{W}, P_X) \triangleq \sup_{P_{U|X}: \substack{I(U;X)>0, \\ U - X - Y}} \frac{I(U; Y)}{I(U; X)},$$

$$\eta_I(\mathsf{W}) \triangleq \sup_{P_X} \eta_I(\mathsf{W}, P_X).$$

By the DPI, $\eta_I(\mathsf{W}, P_X)$ is at most 1, but surprisingly, in many cases, it is strictly smaller than 1. As an example, when W is defined from $\{0,1\}$ to $\{0,1\}$ as a binary symmetric channel (BSC) with crossover probability $\alpha \in (0,1)$, i.e., $\mathsf{W}(1-x \mid x) = \alpha$ and $\mathsf{W}(x \mid x) = 1-\alpha$, $\eta_I(\mathrm{BSC}(\alpha)) = (1-2\alpha)^2$, which is strictly smaller than 1, implying that $I(U;Y) \leq (1-2\alpha)^2 I(U;X)$ for any distribution P_X on $\mathcal{X}$ and any U such that $U - X - Y$ form a Markov Chain. Such inequalities are called *strong data processing inequalities* (SDPIs).

In order to better understand this contraction phenomenon, it is useful to define and study the *data processing function* [6], [7], given by

$$F_I^{\mathsf{W}}(t) \triangleq \sup_{P_{UX}} \{I(U;Y) : I(U;X) \leq t,\ U - X - Y \text{ form a Markov Chain}\}. \tag{1}$$

We could also consider an input-dependent data processing function, where the distribution P_X is also fixed, and we have $F_I^{\mathsf{W}}(t, P_X)$ as the supremum over all conditional distributions on U given X, $P_{U|X}$, but we use the input-independent version in the following. Plotting $y = F_I^{\mathsf{W}}(x)$ on the (x,y)-plane, we have that this curve lies under the $y = \eta_I(\mathsf{W})\, x$ line in the first quadrant, and this is all that the contraction coefficient can tell us about this function. Indeed, the SDPI provides a linear upper bound, but in many cases, this is not useful enough to describe the contraction properties of channels, as we shall see by means of examples.

Example 1. Consider two channels: W_1 is a BSC(α) and W_2 is defined from {0,1} to {0,1,?} as a binary erasure channel (BEC) with erasure probability ϵ, i.e., $\mathsf{W}_2(? \mid x) = \epsilon$ and $\mathsf{W}_2(x \mid x) = 1-\epsilon$. For these channels, the data processing functions are known [6] to be the following:

$$F_I^{\mathrm{BSC}(\alpha)}(t) = \begin{cases} 1 - h_2(\alpha * h_2^{-1}(1-t)) & \text{if } t \leq 1, \\ 1 - h_2(\alpha) & \text{else,} \end{cases}$$

$$F_I^{\mathrm{BEC}(\delta)}(t) = \begin{cases} (1-\delta)t & \text{if } t \leq 1, \\ 1-\delta & \text{else.} \end{cases}$$

Additionally, the contraction coefficients are known to be $\eta_I(\mathrm{BSC}(\alpha)) = (1-2\alpha)^2$ and $\eta_I(\mathrm{BEC}(\delta)) = 1-\delta$. For any α, it is possible to choose δ such that the two contraction coefficients are equal, but $F_I^{\mathrm{BSC}(\alpha)}$ is always strictly concave (by Mrs. Gerber's lemma [8]), while $F_I^{\mathrm{BEC}(\delta)}$ remains linear.

More generally, the DPI says that given any fixed channel W from $\mathcal{X}$ to $\mathcal{Y}$, for any input distributions P_X and Q_X on $\mathcal{X}$, the output distributions will be "closer" to each other, i.e., $D_f(P_X \,||\, Q_X) \geq D_f(P_Y \,||\, Q_Y)$ for any f-divergence. Thus, we define more general contraction coefficients and the associated data processing function as

$$\eta_f(\mathsf{W}, Q_X) \triangleq \sup_{P_X : 0 < D_f(P_X \,||\, Q_X) < \infty} \frac{D_f(\mathsf{W} \circ P_X \,||\, \mathsf{W} \circ Q_X)}{D_f(P_X \,||\, Q_X)},$$

$$\eta_f(\mathsf{W}) \triangleq \sup_{Q_X} \eta_f(\mathsf{W}, Q_X).$$

$$F_f^{\mathsf{W}}(t) \triangleq \sup_{P_X, Q_X} \{D_f(\mathsf{W} \circ P_X \,||\, \mathsf{W} \circ Q_X) : D_f(P_X \,||\, Q_X) \leq t\}.$$

Example 2. For the Gaussian channel with an input power-constraint a, i.e., $Y = X + Z$, $Z \sim \mathcal{N}(0, \sigma^2)$ for all X such that $\mathbb{E}[X^2] \leq a$, the contraction coefficients for the KL divergence and

the total variation distance, which we denote by $\eta_{\mathsf{KL}}(\mathcal{N}(0,\sigma^2),a)$ and $\eta_{\mathsf{TV}}(\mathcal{N}(0,\sigma^2),a)$ resp., are both equal to 1, which seems to suggest that the Gaussian channel does not contract its input distributions w.r.t. at least the KL and TV divergences. This is supported by the observation that $F_{\mathsf{KL}}^{\mathcal{N}(0,\sigma^2),a}(t) = t$ (at least for all $t < \frac{a}{8}$) [6]. It is also known [9] that for any channel, the contraction coefficient of any f-divergence is upper bounded by that of the total variation distance, i.e., $\eta_f(\mathsf{W}) \leq \eta_{\mathsf{TV}}(\mathsf{W})$. Thus, one may be tempted to think that the total variation distance is more "resistant" to contraction. However, it is also known [6] that $F_{\mathsf{TV}}^{\mathcal{N}(0,\sigma^2),a}(t) < t$ for $t > 0$.

It is worth clarifying that the results are not contradictory; the contraction coefficient only looks at the linear part of the data processing function. The takeaway is that this restriction may hide some of the finer details of the function, and these details may sometimes be significant. This motivates the study of such data processing functions, or, as they are more commonly referred to, nonlinear SDPIs.

4 Concavity of the Data Processing Function

We now pose our conjecture on the shape of the curve produced by these data processing functions.

4.1 Problem statement

Conjecture 1. *For any channel* W *and* f*-divergence, the data processing function* F_f^{W} *is concave.*

Consider the set of points $(D_f(P_Y \,||\, Q_Y), D_f(P_X \,||\, Q_X))$ over all distributions P_X, Q_X on $\mathcal{X}$, given a fixed W. This is the *joint range of input and output divergences* over the channel W, formally given by

$$\mathcal{D}_f^{\mathsf{W}} \triangleq \bigcup_{P_X,Q_X} (D_f(P_Y \,||\, Q_Y), D_f(P_X \,||\, Q_X)). \tag{2}$$

The curve F_f^{W} is exactly the upper boundary of this set. Hence, if this set is convex, then Conjecture 1 is true. We may pose this as a stronger conjecture as follows.

Conjecture 2. *For any channel* W *and* f*-divergence,* $\mathcal{D}_f^{\mathsf{W}}$ *is convex.*

This is motivated by a similar result that is known for the joint range of f-divergences.

Theorem 1. *(Joint range of* f*-divergences [10], [11]) For two functions* f *and* g *such that* $D_f(\cdot\,||\,\cdot)$ *and* $D_g(\cdot\,||\,\cdot)$ *are well-defined, consider the map* $(P,Q) \mapsto (D_f(P\,||\,Q), D_g(P\,||\,Q)) \subset \mathbb{R}^2$*, where* P *and* Q *range over all probability distributions such that* $P \ll Q$*. Then, the range of the map is a convex set in* $\mathbb{R}^2$*.*

The proof of this result relies heavily on the fact that the range is formed by *all possible distributions*. In particular, let P_0, Q_0, P_1, Q_1 be any distributions (such that the associated f-divergences are well-defined) on some set $\mathcal{X}$. Then any point that is a convex combination of the points $(D_f(P_0\,||\,Q_0), D_g(P_0\,||\,Q_0))$ and $(D_f(P_1\,||\,Q_1), D_g(P_1\,||\,Q_1))$, can be achieved by simply considering the distributions $P_i \times \delta_i$, where δ_i is the Dirac measure at i, for $i = 0, 1$. These distributions are now defined on $\mathcal{X} \times \{0,1\}$, a set of dimension twice that of the original, but these are still valid probability distributions. Such a trick does not work for the joint range of input and output divergences over a given channel, since the input and output sets are fixed.

Being unable to make any progress on this general conjecture, we decide to focus on the simpler case when the channel is a BSC(α), and $f(x) = x \log x$, i.e., the f-divergence is simply the KL

divergence. For the input distribution P_X given by $(p, 1-p)$ on $\mathcal{X} = \{0, 1\}$ for some $p \in [0, 1]$, the output distribution is given by $p * \alpha = p\bar{\alpha} + \bar{p}\alpha$[1]. Taking Q_X given by $(q, 1-q)$ to be the other input distribution, the input and output KL divergences can be written as

$$D_X(p, q) \triangleq D_{\mathsf{KL}}(P_X \,||\, Q_X) = p \log \frac{p}{q} + \bar{p} \log \frac{\bar{p}}{\bar{q}},$$
$$D_Y(p, q) \triangleq D_{\mathsf{KL}}(P_Y \,||\, Q_Y) = (p * \alpha) \log \frac{p * \alpha}{q * \alpha} + (\bar{p} * \alpha) \log \frac{\bar{p} * \alpha}{\bar{q} * \alpha},$$

where the limiting values are defined as follows: $p \log \frac{p}{0} = \infty$ for $p \neq 0$, $0 \log \frac{0}{q} = 0$ for any q. For this particular channel–divergence pair, Conjecture 3 becomes the following.

Conjecture 3. *For the* BSC(α)*, the data processing function* $F_{\mathsf{KL}}^{\mathrm{BSC}(\alpha)}$ *is concave, where*

$$F_{\mathsf{KL}}^{\mathrm{BSC}(\alpha)}(t) = \sup_{0 \leq p, q \leq 1} \{D_Y(p, q) : D_X(p, q) \leq t\}. \tag{3}$$

Clearly, Conjecture 2 is the strongest of the three; it implies Conjecture 1, which, in turn, implies Conjecture 3. We now restrict our attention to the BSC–KL-divergence scenario.

4.2 The data processing function for the BSC

To study the curve (3), it is useful to consider the joint range $\mathcal{D}_{\mathsf{KL}}^{\mathrm{BSC}(\alpha)}$, i.e., image of the set $[0, 1] \times [0, 1]$ under the mapping $(p, q) \mapsto (D_X(p, q), D_Y(p, q))$. The function $F_{\mathsf{KL}}^{\mathrm{BSC}(\alpha)}$ is then the upper boundary of $\mathcal{D}_{\mathsf{KL}}^{\mathrm{BSC}(\alpha)}$. Since $(D_X(p, q), D_Y(p, q)) = (D_X(\bar{p}, \bar{q}), D_Y(\bar{p}, \bar{q}))$, it suffices to consider the image of

$$\Delta = \{(p, q) : 0 \leq q \leq p \leq 1\}.$$

By the open mapping theorem, interior points of Δ will get mapped to interior points in the image unless the derivative matrix is singular at those points. Hence the upper boundary of the image must be due to either the boundary of Δ or the points at which the derivative matrix is singular. Let us first consider the former. The boundary of Δ consists of three line segments:

1. $0 \leq p = q \leq 1$: This gives $D_X(p, q) = D_Y(p, q) = 0$, i.e., this entire line segment is mapped to the point $(0, 0)$.

2. $0 < p \leq 1, q = 0$: Then we have $D_X(p, q) = \infty$, and

$$\begin{aligned} D_Y(p, q) &= (p * \alpha) \log \frac{p * \alpha}{\alpha} + (\bar{p} * \alpha) \log \frac{\bar{p} * \alpha}{\bar{\alpha}} \\ &= h_2(\alpha) - h_2(p * \alpha) + p\,(\alpha - \bar{\alpha}) \log \frac{\alpha}{\bar{\alpha}}, \end{aligned}$$

   which is increasing in p. When $p \to 0$, we have $D_Y(p, q) \to 0$, and when $p \to 1$, we have $D_Y(p, q) \to (\alpha - \bar{\alpha}) \log \frac{\alpha}{\bar{\alpha}}$.

3. $p = 1, 0 < q < 1$: This gives $D_X(p, q) = \log \frac{1}{q}$, and

$$D_Y(p, q) = \bar{\alpha} \log \frac{\bar{\alpha}}{q * \alpha} + \alpha \log \frac{\alpha}{\bar{q} * \alpha},$$

   then taking $q = 2^{-t}$, $0 < t < \infty$, we have that this line segment is mapped to the curve $\{(t, \bar{\alpha} \log \frac{\bar{\alpha}}{2^{-t} * \alpha} + \alpha \log \frac{\alpha}{2^{-t} * \bar{\alpha}}) : 0 < t < \infty\}$.

[1]Recall that we use the notation $\bar{x} = 1 - x$; we then have $\overline{p * \alpha} = \bar{p} * \alpha = p * \bar{\alpha}$.

These line segments (in green, red and blue respectively) and their images are shown in Figure 1. If it were the case that the derivative matrix is never singular, i.e., the boundary of the range is completely determined by the boundary of the domain, then the curve of interest to us is simply the blue line, which is the function mapping $t \mapsto \bar{\alpha} \log \frac{\bar{\alpha}}{2^{-t} * \alpha} + \alpha \log \frac{\alpha}{2^{-t} * \bar{\alpha}}$. It can be seen analytically (by differentiating) and also directly from the figure that this curve is not concave. However, it turns out that the derivative matrix is indeed singular at some points, and hence the blue curve is not the boundary of the image of Δ. This happens when

$$\begin{vmatrix} \frac{\partial D_X}{\partial p} & \frac{\partial D_Y}{\partial p} \\ \frac{\partial D_X}{\partial q} & \frac{\partial D_Y}{\partial q} \end{vmatrix} = 0 \implies \frac{\log \frac{p}{q} - \log \frac{\bar{p}}{\bar{q}}}{\frac{p}{q} - \frac{\bar{p}}{\bar{q}}} = \frac{\log \frac{p*\alpha}{q*\alpha} - \log \frac{\bar{p}*\alpha}{\bar{q}*\alpha}}{\frac{p*\alpha}{q*\alpha} - \frac{\bar{p}*\alpha}{\bar{q}*\alpha}},$$

for (p, q) in the interior of Δ, i.e., $0 < q < p < 1$. This expression has a nice interpretation: let $\ell_1 = \frac{p}{q}$ and $\ell_2 = \frac{\bar{p}}{\bar{q}}$ be the likelihood ratios. Then we have $\ell_1 > 1$ and $\ell_2 < 1$. Further, any pair $(\ell_1, \ell_2) \in (1, \infty) \times (0, 1)$ uniquely determines a pair (p, q), given by

$$q = \frac{1 - \ell_2}{\ell_1 - \ell_2}, \qquad p = \ell_1 \frac{1 - \ell_2}{\ell_1 - \ell_2},$$

so we have an isomorphic relation between the set of all (ℓ_1, ℓ_2) and the interior of Δ. Further, let $\tilde{\ell}_1 = \frac{p*\alpha}{q*\alpha}$ and $\tilde{\ell}_2 = \frac{\bar{p}*\alpha}{\bar{q}*\alpha}$ be the likelihood ratios at the output. The condition for the derivative matrix to be singular is then given by

$$\frac{\log \ell_1 - \log \ell_2}{\ell_1 - \ell_2} = \frac{\log \tilde{\ell}_1 - \log \tilde{\ell}_2}{\tilde{\ell}_1 - \tilde{\ell}_2},$$

i.e., the line joining the points $(\ell_1, \log \ell_1)$ and $(\ell_2, \log \ell_2)$ is parallel to the line joining $(\tilde{\ell}_1, \log \tilde{\ell}_1)$ and $(\tilde{\ell}_2, \log \tilde{\ell}_2)$. It is possible to use this interpretation to obtain a numerical solution, which is shown as the black curve in Figure 1, by taking $\alpha = 0.3$ — this choice of α is not special, and similar results can be obtained for all nontrivial values of α (i.e., $\alpha \neq 0, \frac{1}{2}, 1$). This curve is concave, which verifies Conjecture 3 numerically. However, the transformation $(\ell_1, \ell_2) \mapsto (\tilde{\ell}_1, \tilde{\ell}_2)$ is difficult to study analytically, and Conjecture 3 and, more generally, Conjectures 1 and 2 remain open.

Figure 1: The boundary of the region Δ and its image under the mapping $(p, q) \mapsto (D_X(p, q), D_Y(p, q))$ are shown in green, red and blue — had these been the only critical points, the image of the boundary would have also been the boundary of the image, but this is not the case, as seen by the numerically obtained boundary shown in black. We take $\alpha = 0.3$.

5 Revisiting the Information Bottleneck Problem

The IB problem was introduced as an information-theoretic approach to learning [12]. In particular, we consider a pair of correlated random variables (X, Y) with some joint distribution, where Y is some target variable that we wish to study, and X is an observation that depends on Y. The goal is to find a representation $U(X)$ (that may be randomized) which characterizes the trade-off between U depending too much on X itself (complexity) and not having enough dependence on Y (relevance). Formally, the IB problem for (X, Y) is

$$\sup_{P_{U|X}:\substack{U-X-Y \\ I(U;X)\leq t}} I(U;Y),$$

which is exactly the input-dependent form of $F_I(t)$ as in (1). This can be equivalently written in terms of a Lagrange multiplier β, as

$$\mathcal{L}_\beta(P_{U|X}) = I(U;Y) - \beta I(U;X).$$

For each β, maximizing $\mathcal{L}_\beta$ gives an optimal conditional distribution $P^{*,\beta}_{U|X}$; plotting the pair $(I(U;X), I(U;Y))$ for each such value of $P^{*,\beta}_{U|X}$ then gives the curve $F_I(t, P_X)$, parametrized by β.

The IB problem can be easily seen to be equivalent to various classical source coding setups [2] in information theory, but its recent popularity comes from its success in explaining various observations in machine learning such as generalization and layering in deep architectures, particularly with deep neural networks [3]. Insights gained from theoretical studies through IB have also led to practical algorithms with improved performance. Making progress on obtaining better characterizations of the data processing function would give us a better understanding of the trade-off between relevance and complexity in learning algorithms, and more generally, aid in the design and analysis of improved learning systems.

6 Conclusion

We have discussed the need for data processing functions instead of inequalities, and proposed Conjecture 1, which seems natural enough. However, we have been able to show even the simplest case, i.e., Conjecture 3, only numerically. Here are some related results that might be of use in the future.

1. It is known that F_I^{W}, the data processing function for the mutual information, is not concave in general, by a counter-example involving binary-erasure channels used thrice [4]. This seems to suggest that Conjecture 1 may also be false, but in the absence of a counter-example, we still believe that it is true, mainly due to Theorem 1.
2. It can be shown [13] that any point in the upper concave envelope of the joint range (2) is achievable by only considering P_X and Q_X of binary support (for any finite input set $\mathcal{X}$ and channel W). This adds support to the belief that the joint range is indeed convex, since it seems that we can obtain points on the boundary even without going to higher input dimensions (as in the proof of Theorem 1). It may be possible to fashion a similar proof in this case as well, but we have been unable to.
3. To show that the upper boundary of the joint range is concave, one approach could be to try to find suitable parametrizations that trace out concave paths between any pair of points within the range — some attempts have been made for the case with binary $\mathcal{X}$ [14].

References

[1] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[2] A. Zaidi, I. Estella-Aguerri, and S. Shamai (Shitz), "On the information bottleneck problems: Models, connections, applications and information theoretic views," *Entropy*, 2020.

[3] Z. Goldfeld and Y. Polyanskiy, "The information bottleneck problem and its applications in machine learning," *IEEE Journal on Selected Areas in Information Theory*, 2020.

[4] Y. Polyanskiy and Y. Wu, "Strong data-processing inequalities for channels and bayesian networks," in *Convexity and Concentration*, Springer New York, 2017.

[5] A. Makur, "Information contraction and decomposition," Ph.D. dissertation, Massachusetts Institute of Technology, 2019.

[6] Y. Polyanskiy and Y. Wu, "Dissipation of information in channels with input constraints," *IEEE Transactions on Information Theory*, 2016.

[7] F. d. P. Calmon, Y. Polyanskiy, and Y. Wu, "Strong data processing inequalities for input constrained additive noise channels," *IEEE Transactions on Information Theory*, 2018.

[8] A. Wyner and J. Ziv, "A theorem on the entropy of certain binary sequences and applications–i," *IEEE Transactions on Information Theory*, 1973.

[9] J. Cohen, J. H. Kempermann, and G. Zbaganu, *Comparisons of stochastic matrices with applications in information theory, statistics, economics and population*. Springer Science & Business Media, 1998.

[10] P. Harremoes and I. Vajda, "On pairs of f -divergences and their joint range," *IEEE Transactions on Information Theory*, 2011.

[11] Y. Wu, *Lecture notes: Information-theoretic methods for high-dimensional statistics*. [Online]. Available: http://www.stat.yale.edu/~yw562/teaching/it-stats.pdf.

[12] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. Allerton Conference on Communication, Control and Computing*, 1999.

[13] O. Ordentlich and Y. Polyanskiy, "Strong data processing constant is achieved by binary inputs," *IEEE Transactions on Information Theory*, 2022.

[14] Q. Ding, C. W. Lau, C. Nair, and Y. N. Wang, "Concavity of output relative entropy for channels with binary inputs," in *2021 IEEE International Symposium on Information Theory (ISIT)*, 2021.