Strong Data-Processing Inequalities for Channels and Bayesian Networks

Citation: Polyanskiy, Yury and Yihong Wu. "Strong Data-Processing Inequalities for Channels and Bayesian Networks." Convexity and Concentration, IMA Volumes in Mathematics and its Applications, 161, Springer, 2017, 211-249. © 2017 Springer Science+Business Media LLC

As Published: http://dx.doi.org/10.1007/978-1-4939-7005-6_7

Publisher: Springer

Persistent URL: https://hdl.handle.net/1721.1/129982

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Strong data-processing inequalities for channels and Bayesian networks

Yury Polyanskiy* and Yihong Wu†

MIT, UIUC

Abstract

The data-processing inequality, that is, $I(U;Y) \le I(U;X)$ for a Markov chain $U \to X \to Y$, has been the method of choice for proving impossibility (converse) results in information theory and many other disciplines. Various channel-dependent improvements (called strong data-processing inequalities, or SDPIs) of this inequality have been proposed both classically and more recently. In this note we first survey known results relating various notions of contraction for a single channel. Then we consider the basic extension: given SDPI for each constituent channel in a Bayesian network, how to produce an end-to-end SDPI?

Our approach is based on the (extract of the) Evans-Schulman method, which is demonstrated for three different kinds of SDPIs, namely, the usual Ahslwede-Gács type contraction coefficients (mutual information), Dobrushin's contraction coefficients (total variation), and finally the F_I-curve (the best possible non-linear SDPI for a given channel). Resulting bounds on the contraction coefficients are interpreted as probability of site percolation. As an example, we demonstrate how to obtain SDPI for an n-letter memoryless channel with feedback given an SDPI for $n = 1$.

Finally, we discuss a simple observation on the equivalence of a linear SDPI and comparison to an erasure channel (in the sense of "less noisy" order). This leads to a simple proof of a curious inequality of Samorodnitsky (2015), and sheds light on how information spreads in the subsets of inputs of a memoryless channel.

*Y.P. is with the Department of EECS, MIT, Cambridge, MA, yp@mit.edu. His research has been supported by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-09-39370 and by the NSF CAREER award under grant agreement CCF-12-53205.

†Y.W. is with the Department of ECE, University of Illinois at Urbana-Champaign, Urbana, IL, yihongwu@illinois.edu. His research has been supported in part by NSF grants IIS-1447879, CCF-1423088 and the Strategic Research Initiative of the College of Engineering at the University of Illinois.

Contents

1 Introduction

Multiplication of a componentwise non-negative vector by a stochastic matrix results in a vector that is "more uniform". This observation appears in several classical works [Mar06, Doe37, Bir57] differing in their particular way of making quantitative estimates. For example, Birkhoff's work [Bir57] initiated a study (sometimes known as geometric ergodicity) of contraction of the projective distance $d_P(x,y) \triangleq \log \max_i \frac{x_i}{y_i} - \log \min_i \frac{x_i}{y_i}$ between vectors in $\mathbb{R}_+^n$. Here, instead, we will be interested in contraction of statistical distances and information measures involving probability distributions, which we define next.

Fix a transition probability kernel (channel) $P_{Y|X} : \mathcal{X} \to \mathcal{Y}$ acting between two measurable spaces. We denote by $P_{Y|X} \circ P$ the distribution on $\mathcal{Y}$ induced by the push-forward of the distribution P, which is the distribution of the output Y when the input X is distributed according to P, and by $P \times P_{Y|X}$ the joint distribution P_{XY} if $P_X = P$. We also denote by $P_{Z|Y} \circ P_{Y|X}$ the serial composition of channels.[1]

We define three quantities that will play key role in our discussion: the total variation, the Kullback-Leibler (KL) divergence and the mutual information

$$d_{\mathrm{TV}}(P,Q) \triangleq \sup_E |P[E] - Q[E]| = \frac{1}{2}\int |\mathrm{d}P - \mathrm{d}Q|, \tag{1}$$

$$D(P\|Q) \triangleq \int \log \frac{\mathrm{d}P}{\mathrm{d}Q}\,\mathrm{d}P, \tag{2}$$

$$I(A;B) \triangleq D(P_{AB}\|P_A P_B). \tag{3}$$

The purpose of this paper is to give exposition to the phenomenon that upon passing through a non-degenerate noisy channel distributions become strictly closer and this leads to a loss of information. Namely we have three effects:

1. Total-variation (or Dobrushin) contraction:

$$d_{\mathrm{TV}}(P_{Y|X} \circ P, P_{Y|X} \circ Q) < d_{\mathrm{TV}}(P,Q)\,.$$

2. Divergence contraction:

$$D(P_{Y|X} \circ P \| P_{Y|X} \circ Q) < D(P\|Q)$$

3. Information loss: For any Markov chain[2] $U \to X \to Y$ we

$$I(U;Y) < I(U;X)\,.$$

These strict inequalities are collectively referred to as *strong data-processing inequalities* (SDPIs). The goal of this paper is to show intricate interdependencies between these effects, as well as introducing tools for quantifying how strict these SDPIs are.

[1] More formally, we should have written $P_{Y|X} : \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{Y})$ as a map between spaces of probability measures $\mathcal{P}(\cdot)$ on respective bases. The rationale for our notation $P_{Y|X} : \mathcal{X} \to \mathcal{Y}$ is that we view Markov kernels as randomized functions. Then, a single distribution P on $\mathcal{X}$ is a randomized function acting from a space of a single point, i.e. $P : [1] \to \mathcal{X}$, and that in turn explains our notation $P_{Y|X} \circ P$ for denoting the induced marginal distribution.

[2] The notation $A \to B \to C$ simply means that $A \perp\!\!\!\perp C | B$.

Organization In Section 2 we overview the case of a single channel. Notably, most of the results in the literature are proved for finite alphabets, i.e., $|\mathcal{X}||\mathcal{Y}| < \infty$, with a few exceptions such as [CKZ98, PW16]. We provide in Appendix A a self-contained proof of some of these results for general alphabets.

From then on we focus on the question: *Given a multi-terminal network with a single source and multiple sinks, and given SDPIs for each of the channels comprising the network, how do we obtain an SDPI for the composite channel from source to sinks?* It turns out that this question has been addressed implicitly in the work of Evans and Schulman [ES99] on redundancy required in circuits of noisy gates. Rudiments also appeared in Dawson [Daw75] as well as Boyen and Koller [BK98].

In Section 3 we present the essence of the Evans-Schulman method and derive upper bounds on the mutual information contraction coefficient η_{KL} for Bayesian networks (directed graphical models). We also interpret the resulting bounds as probabilities of disupting end-to-end connectivity under independent removals of graph vertices (site percolation). Then in Section 4 we derive analogous estimates for Dobrushin's coefficient η_{TV} that governs the contraction of the total variation on networks. While the results exactly parallel those for mutual information, the proof relies on new arguments using coupling. Finally, Section 5 extends the technique to bounding the F_I-curves (the non-linear SDPIs). Section 6 concludes with an alternative point of view on mutual information contraction, namely that of comparison to an erasure channel. As an example we give a short proof of a result of Samorodnitsky [Sam15] about distribution of information in subsets of channel outputs.

Notation Elements of the Cartesian product $\mathcal{X}^n$ are denoted $x^n \triangleq (x_1, \ldots, x_n)$ to emphasize their dimension. Given a transition probability kernel from $P_{Y|X} : \mathcal{X} \to \mathcal{Y}$ we denote $P_{Y|X}^n = P_{Y^n|X^n}$ the kernel acting from $\mathcal{X}^n \to \mathcal{Y}^n$ componentwise independently:

$$P_{Y^n|X^n}(y^n|x^n) \triangleq \prod_{j=1}^{n} P_{Y|X}(y_j|x_j).$$

To demonstrate the general bounds we consider the running example of $P_{Y|X}$ being an n-letter binary symmetric channel (BSC), given by

$$Y = X + Z, \quad X, Y \in \mathbb{F}_2^n,\ Z \sim \text{Bern}(\delta)^n \tag{4}$$

and denoted by $\mathsf{BSC}(\delta)^n$. Throughout this paper $\bar{\delta} \triangleq 1 - \delta$.

2 SDPI for a single channel

2.1 Contraction coefficients for f-divergence and mutual information

Let $f : (0, \infty) \to \mathbb{R}$ be a convex function that is strictly convex at 1 and $f(1) = 0$. Let $D_f(P||Q) \triangleq \mathbb{E}_Q[f(\frac{dP}{dQ})]$ denote the f-divergence of P and Q with $P \ll Q$, cf. [Csi67].[3] For example, the total variation (1) and the KL divergence (2) correspond to $f(x) = \frac{1}{2}|x - 1|$ and $f(x) = x \log x$ respectively; taking $f(x) = (x - 1)^2$ we obtain the χ^2-divergence: $\chi^2(P\|Q) \triangleq \int (\frac{dP}{dQ})^2 dQ - 1$.

[3] More generally, $D_f(P||Q) \triangleq \mathbb{E}_\mu\left[f\left(\frac{dP/d\mu}{dQ/d\mu}\right)\right]$, where μ is a dominating probability measure of P and Q, e.g., $\mu = (P + Q)/2$, with the understanding that $f(0) = f(0+)$, $0f(\frac{0}{0}) = 0$ and $0f(\frac{a}{0}) = \lim_{x\downarrow 0} xf(\frac{a}{x})$ for $a > 0$.

For any Q that is not a point mass, define:

$$\eta_f(P_{Y|X}, Q) \triangleq \sup_{P: 0 < D_f(P\|Q) < \infty} \frac{D_f(P_{Y|X} \circ P \| P_{Y|X} \circ Q)}{D_f(P\|Q)}, \tag{5}$$

$$\eta_f(P_{Y|X}) \triangleq \sup_Q \eta_f(Q) \,. \tag{6}$$

It is easy to show that the supremum is over a non-empty set whenever Q is not a point mass (see Appendix A). For notational simplicity when the channel is clear from context we abbreviate $\eta_f(P_{Y|X})$ as η_f. For contraction coefficients of total variation, χ^2 and KL divergence, we write $\eta_{\mathrm{TV}}, \eta_{\chi^2}$ and η_{KL}, respectively, which play prominent roles in this exposition.

One of the main tools for studying ergodicity property of Markov chains as well as Gibbs measures, $\eta_{\mathrm{TV}}(P_{Y|X})$ is known as the *Dobrushin's coefficient* of the kernel $P_{Y|X}$. Dobrushin [Dob56] showed that the supremum in the definition of η_{TV} can be restricted to point masses, namely,

$$\eta_{\mathrm{TV}}(P_{Y|X}) = \sup_{x,x'} d_{\mathrm{TV}}(P_{Y|X=x}, P_{Y|X=x'}), \tag{7}$$

thus providing a simple criterion for strong ergodicity of Markov processes. Later [CKZ98, Proposition II.4.10(i)] (see also [CIR+93, Theorem 4.1] for finite alphabets) demonstrated that all other contraction coefficients are upper bounded by the Dobrushin's coefficient, with inequality being typically strict (cf. the BSC example below):

Theorem 1 ([CKZ98, Proposition II.4.10]). *For every f-divergence, we have*

$$\eta_f(P_{Y|X}) \le \eta_{\mathrm{TV}}(P_{Y|X}). \tag{8}$$

For the opposite direction, lower bounds on η_f typically involves η_{χ^2}, the contraction coefficient of the χ^2-divergence. It is well-known, e.g. Sarmanov [Sar58], that $\eta_{\chi^2}(P_{Y|X}, P_X)$ is the squared second largest eigenvalue of the conditional expectation operator, which in turn equals the *maximal correlation* coefficient of the joint distribution P_{XY}:

$$S(X;Y) \triangleq \sup_{f,g} \rho(f(X), g(Y)) = \sqrt{\eta_{\chi^2}(P_{Y|X}, P_X)}\,, \tag{9}$$

where $\rho(\cdot,\cdot)$ denotes the correlation coefficient and the supremum is over real-valued functions f, g such that $f(X)$ and $g(Y)$ are square integrable.

The relationship between η_{KL} and η_{χ^2} on finite alphabets has been systematically studied by Ahlswede and Gács [AG76]. In particular, [AG76] proved

$$\eta_{\chi^2}(P_{Y|X}, P_X) \le \eta_{\mathrm{KL}}(P_{Y|X}, P_X), \tag{10}$$

and noticed that the inequality is frequently strict.[4] Furthermore, for finite alphabets, the following equivalence is demonstrated in [AG76]:

$$\eta_{\chi^2}(P_X, P_{Y|X}) < 1 \iff \eta_{\mathrm{KL}}(P_X, P_{Y|X}) < 1 \tag{11}$$

$$\iff \text{graph } \{(x,y) : P_X(x) > 0, P_{Y|X}(y|x) > 0\} \text{ is connected.} \tag{12}$$

As a criterion for $\eta_f(P_{Y|X}, P_X) < 1$, this is an improvement of (8) only for channels with $\eta_{\mathrm{TV}}(P_{Y|X}) = 1$. The lower bound (10) can in fact be considerably generalized:

[4] See [AG76, Theorem 9] and [AGKN13] for examples.

Theorem 2. *Let f be twice continuously differentiable on $(0,\infty)$ with $f''(1) > 0$. Then for any P_X that is not a point mass,*

$$\eta_{\chi^2}(P_{Y|X}, P_X) \le \eta_f(P_{Y|X}, P_X)\,, \tag{13}$$

and

$$\eta_{\chi^2}(P_{Y|X}) \le \eta_f(P_{Y|X})\,. \tag{14}$$

See Appendix A.1 for a proof of (13) for the general case, which yields (14) by taking suprema over P_X on both sides. Note that (14) (resp. (13)) have been proved in [CKZ98, Proposition II.6.15] for the general alphabet (resp. in [Rag14, Theorem 3.3] for finite alphabets).

Moreover, (14) in fact holds with equality for all nonlinear and operator convex f, e.g., for KL divergence and for squared Hellinger distance; see [CRS94, Theorem 1] and [CKZ98, Proposition II.6.13 and Corollary II.6.16]. Therefore, we have:

Theorem 3.

$$\eta_{\chi^2}(P_{Y|X}) = \eta_{\mathrm{KL}}(P_{Y|X})\,. \tag{15}$$

See Appendix A.1 for a self-contained proof. This result was first obtained in [AG76] using different methods for discrete space. Rather naturally, we also have [CKZ98, Proposition II.4.12]:

$$\eta_f(P_{Y|X}) = 1 \qquad \Longleftrightarrow \qquad \eta_{\mathrm{TV}}(P_{Y|X}) = 1$$

for any non-linear f.

As an illustrating example, for $\mathsf{BSC}(\delta)$ defined in (4), we have cf. [AG76]

$$\eta_{\chi^2} = \eta_{\mathrm{KL}} = (1-2\delta)^2 < \eta_{\mathrm{TV}} = |1-2\delta|. \tag{16}$$

Appendix B present general results on the contraction coefficients for binary-input arbitrary-output channels, which can be bounded using Hellinger distance within a factor of two.

We next discuss the the fixed-input contraction coefficient $\eta_{\mathrm{KL}}(P_{Y|X}, Q)$. Unfortunately, there is no simple reduction to the χ^2-case as in (15). Besides the lower bound (10), there is a variety of upper bounds relating η_{KL} and η_{χ^2}. We quote [MZ15, Theorem 11], who show for finite input-alphabet case:

$$\eta_{\mathrm{KL}}(P_{Y|X}, Q) \le \frac{1}{\min_x Q(x)} \eta_{\chi^2}(P_{Y|X}, Q)\,.$$

Another bound (which also holds for all η_f with operator-convex f) is in [Rag14, Theorem 3.6]:

$$\eta_{\mathrm{KL}}(P_{Y|X}, Q) \le \max\left(\eta_{\chi^2}(P_{Y|X}, Q), \sup_{0<\beta<1} \eta_{\mathrm{LC}_\beta}(P_{Y|X}, Q)\right)\,,$$

where η_{LC_β} denotes contraction coefficient of an f-divergence $\mathrm{LC}_\beta(P\|Q) = \beta\bar\beta \int \frac{(P-Q)^2}{\beta P + \bar\beta Q}$ with $\beta \in (0,1)$ and $\bar\beta = 1-\beta$ (see also Appendix B).

We also note in passing that SDPIs are intimately related to hypercontractivity and maximal correlation, as discovered by Ahlswede and Gács [AG76] and recently improved by Anantharam et al. [AGKN13] and Nair [Nai14]. Indeed, the main result of [AG76] characterizes $\eta_{\mathrm{KL}}(P_{Y|X}, P_X)$ as the maximal ratio of hyper-contractivity of the conditional expectation operator $\mathbb{E}[\cdot|X]$.

The fixed-input contraction coefficient $\eta_{\mathrm{KL}}(Q)$ is closely related to the (modified) log-Sobolev inequalities. Indeed, if $\eta_{\mathrm{KL}}(Q) < 1$ where Q is the invariant measure for the Markov kernel $P_{Y|X}$, i.e., $P_{Y|X} \circ Q = Q$, then any initial distribution P such that $D(P\|Q) < \infty$ converges to Q exponentially fast since

$$D(P^n_{Y|X} \circ P \| Q) \le \eta^n_{\mathrm{KL}}(P_{Y|X}, Q) D(P\|Q),$$

where the exponent $\eta_{\mathrm{KL}}(P_{Y|X}, Q)$ can in turn be estimated from log-Sobolev inequalities, e.g. [Led99]. When Q is not invariant, it was shown [DMLM03] that

$$1 - \alpha(Q) \le \eta_{\mathrm{KL}}(P_{Y|X}, Q) \le 1 - C\alpha(Q)$$

holds for some universal constant C, where $\alpha(Q)$ is a modified log-Sobolev (also known as 1-log-Sobolev) constant:

$$\alpha(Q) = \inf_{f \neq 1, \|f\|_2 = 1} \frac{\mathbb{E}\left[f^2(X) \log \frac{f^2(X)}{f^2(X')}\right]}{\mathbb{E}[f^2(X) \log f^2(X)]}, \qquad P_{XX'} = Q \times (P_{X|Y} \circ P_{Y|X}).$$

For further connections between η_{KL} and log-Sobolev inequalities on finite alphabets see [Rag13, Rag14].

There exist several other characterizations of η_{KL}, such as the following one in terms of the contraction of mutual information (cf. [CK81, Exercise III.3.12, p. 350] for finite alphabet):

$$\eta_{\mathrm{KL}}(P_{Y|X}) = \sup \frac{I(U;Y)}{I(U;X)}, \tag{17}$$

where the supremum is over all Markov chains $U \to X \to Y$ with fixed $P_{Y|X}$ (or equivalently, over all joint distributions P_{XU}) such that $I(U;X) < \infty$. This result is an immediate consequence of the following input-dependent version (see Appendix A.3 for a proof in the general case; the finite alphabet case has been shown in [AGKN13])

Theorem 4. *For any P_X that is not a point mass,*

$$\eta_{\mathrm{KL}}(P_{Y|X}, P_X) = \sup \frac{I(U;Y)}{I(U;X)}, \tag{18}$$

where the supremum is taken over all Markov chains $U \to X \to Y$ with fixed $P_{XY} = P_X \circ P_{Y|X}$ such that $0 < I(U;X) < \infty$.

Another characterization of η_{KL}, in view of (15) and (9), is

$$\eta_{\mathrm{KL}}(P_{Y|X}) = \sup \rho^2(f(X), g(Y)),$$

where the supremum is over all P_X and real-valued square-integrable $f(X)$ and $g(Y)$.

2.2 Non-linear SDPI

How to quantify the information loss if $\eta_{\mathrm{KL}} = 1$ for the channel of interest? In fact this situation can arise in very basic settings, such as the additive-noise Gaussian channel under the moment constraint on the input distributions (cf. [PW16, Theorem 9, Section 4.5]), where the mutual information does not contract linearly as in (17), but can still contract *non-linearly*. In such cases, establishing a strong-data processing inequality can be done by following the joint-range idea of Harremoës and Vajda [HV11]. Namely, we aim to find (or bound) the *best possible data-processing function* F_I defined as follows.

Definition 1 (F_I-curve)**.** Fix $P_{Y|X}$ and define

$$F_I(t, P_{Y|X}) \triangleq \sup_{P_{UX}} \left\{ I(U;Y) \colon I(U;X) \le t, P_{UXY} = P_{UX} P_{Y|X} \right\}. \tag{19}$$

Equivalently, the supremum is taken over all joint distributions P_{UXY} with a given conditional $P_{Y|X}$ and satisfying $U \to X \to Y$. The upper concave envelope of F_I is denoted by F_I^c:

$$F_I^c(t, P_{Y|X}) \triangleq \inf\{f(t) : \forall t' \geq 0 \; F_I(t', P_{Y|X}) \leq f(t'), f\text{–concave}\}\,.$$

Equivalently, we have

$$F_I^c(t, P_{Y|X}) = \sup_{P_{VUX}} \left\{ I(U;Y|V) \colon I(U;X|V) \leq t, P_{VUXY} = P_{VUX}P_{Y|X} \right\}\,, \tag{20}$$

where $I(A;B|C) \triangleq I(A,C;B) - I(C;B)$ is the conditional mutual information, and averaging over V serves the role of concavification (so that V can be taken binary). Whenever it does not lead to confusion we will write $F_{Y|X}(t)$ instead of $F_I(t, P_{Y|X})$.

The operational significance of the F_I-curve is that it gives the optimal input-independent strong data processing inequality:

$$I(U;Y) \leq F_I(I(U;X)),$$

which generalizes (17) since $F_I'(0) = \eta_{\mathrm{KL}}(P_{Y|X})$ and $t \mapsto \frac{1}{t}F_I(t)$ is decreasing (see, e.g., [CPW15, Section I]). See [CPW15] for bounds and expressions for BSC and Gaussian channels.

Frequently it is more convenient to work with the concavified version F_I^c as it allows for some natural extension of the results about contraction coefficients. Proposition 18 shows that F_I may not be concave.

2.3 Some applications: classical and new

The main example of a strong data-processing inequality (SDPI) was discovered by Ahlswede and Gács [AG76]. They have shown, using the characterization (11), that whenever $P_{Y|X}$ is a discrete memoryless channel that does not admit zero-error communication, we have $\eta_{\mathrm{KL}}(P_{Y|X}) \leq \eta < 1$ and

$$I(W;Y) \leq \eta I(W;X) \tag{21}$$

for all Markov chains $W \to X \to Y$.

SDPIs have been popular for establishing lower (impossibility) bounds in various setups, in both classical and more recent works. We mention only a few of these applications:

- By Dobrushin for showing non-existence of multiple phases in Ising models at high temperatures [Dob70];
- By Erkip and Cover in portfolio theory [EC98];
- By Evans and Schulman in analysis of noise-resistant circuits [ES99];
- By Evans, Kenyon, Peres and Schulman in the analysis of inference on trees and percolation [EKPS00];
- By Courtade in distributed data-compression [Cou12];
- By Duchi, Wainwright and Jordan in statistical limitations of differential privacy [DJW13];
- By the authors to quantify optimal communication and optimal control in line networks [PW16];
- By Liu, Cuff and Verdú in key generation [LCV15];

- By Xu and Raginsky in distributed estimation [XR15].

All of the applications above use SDPI (21) to prove negative (impossibility) statements. A notable exception is the work of Boyen and Koller [BK98], who considered the basic problem of computing the posterior-belief vector of a hidden Markov model: that is, given a Markov chain $\{X_j\}$ observed over a memoryless channel $P_{Y|X}$, one aims to recompute $P_{X_j|Y_{-\infty}^j}$ as each new observation Y_j arrives. The problem arises when X is of large dimension and then for practicality one is constrained to approximate (quantize) the posterior. However, due to the recursive nature of belief computations, the cumulative effect of these approximations may become overwhelming. Boyen and Koller [BK98] proposed to use the SDPI similar to (21) with $\eta < 1$ for the Markov chain $\{X_j\}$ and show that this cumulative effect stays bounded since $\sum \eta^n < \infty$. Similar considerations also enable one to provide provable guarantees for simulation of inter-dependent stochastic processes.

3 Contraction of mutual information in networks

We start by defining a *Bayesian network* (also known as a *directed graphical model*). Let G be a finite directed acyclic graph with set of vertices $\{Y_v : v \in \mathcal{V}\}$ denoting random variables taking values in a fixed finite alphabet.[5] We assume that each vertex Y_v is associated with a conditional distribution $P_{Y_v|Y_{\mathrm{pa}(v)}}$ where $\mathrm{pa}(v)$ denotes parents of v, with the exception of one special "source" node X that has no inbound edges (there may be other nodes without inbound edges, but those have to have their marginals specified). Notice that if $V \subset \mathcal{V}$ is an arbitrary set of nodes we can progressively chain together all the random transformations and unequivocally compute $P_{V|X}$ (here and below we use V and $Y_V = \{Y_v : v \in V\}$ interchangeably). We assume that vertices in $\mathcal{V}$ are topologically sorted so that $v_1 > v_2$ implies there is no path from v_1 to v_2. Associated to each node we also define

$$\eta_v \triangleq \eta_{\mathrm{KL}}(P_{Y_v|Y_{\mathrm{pa}(v)}}) .$$

See the excellent book of Lauritzen [Lau96] for a thorough introduction to a graphical model language of specifying conditional independencies.

The following result can be distilled from [ES99]:

Theorem 5. *Let $W \in \mathcal{V}$ and $V \subset \mathcal{V}$ such that $W > V$. Then*

$$\eta_{\mathrm{KL}}(P_{V,W|X}) \le \eta_W \cdot \eta_{\mathrm{KL}}(P_{V,\mathrm{pa}(W)|X}) + (1-\eta_W) \cdot \eta_{\mathrm{KL}}(P_{V|X}) . \tag{22}$$

Furthermore, let $\mathrm{perc}(V)$ denote the probability that there is a path from X to V[6] *in the graph if each node v is removed independently with probability $1-\eta_v$ (site percolation). Then, we have for every $V \subset \mathcal{V}$*

$$\eta_{\mathrm{KL}}(P_{V|X}) \le \mathrm{perc}(V) . \tag{23}$$

In particular, if $\eta_v < 1$ for all $v \in \mathcal{V}$ then $\eta_{\mathrm{KL}}(P_{V|X}) < 1$.

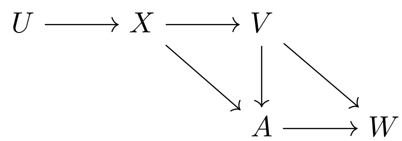
Proof. Consider an arbitrary random variable U such that

$$U \to X \to (V, W) .$$

[5] At the expense of technical details, the alphabet can be replaced with any countably-generated (e.g. Polish) measurable space. For clarity of presentation we focus here on finite alphabets.

[6] More formally, $\mathrm{perc}(V)$ equals probability that there exists a sequence of nodes $v_1, \ldots, v_n$ with $v_1 = X$, $v_n \in V$ satisfying two conditions: 1) for each $i \in [n-1]$ the pair (v_i, v_{i+1}) is a directed edge in G; and 2) each v_i is not removed.

Let $A = \mathrm{pa}(W) \setminus V$. Without loss of generality we may assume A does not contain X: indeed, if A includes X then we can introduce an artificial node X' such that $X' = X$ and include X' into A instead of X. Relevant conditional independencies are encoded in the following graph:



From the characterization (17) it is sufficient to show

$$I(U; V, W) \le (1 - \eta_W) I(U; V) + \eta_W I(U; V, A)\,. \tag{24}$$

Denote $B = V \backslash \mathrm{pa}(W)$ and $C = V \cap \mathrm{pa}(W)$. Then $\mathrm{pa}(W) = (A, C)$ and $V = (B, C)$. To verify (24) notice that by assumption we have

$$U \to X \to (V, A) \to W\,.$$

Therefore conditioned on V we have the Markov chain

$$U \to X \to A \to W \qquad | V$$

and the channel $A \to W$ is a restriction of the original $P_{W|\mathrm{pa}(W)}$ to a subset of the inputs. Indeed, $P_{W|A,V} = P_{W|\mathrm{pa}(W),B} = P_{W|\mathrm{pa}(W)}$ by the assumption of the graphical model. Thus, for every realization $v = (b, c)$ of V, we have $P_{W|A=a,V=v} = P_{W|A=a,C=c}$ and therefore

$$I(U; W|V = v) \le \eta(P_{W|A,C=c}) I(U; A|V = v) \le \eta(P_{W|A,C}) I(U; A|V = v), \tag{25}$$

where the last inequality uses the following property of the contraction coefficient which easily follows from either (6) or (17):

$$\sup_c \eta(P_{W|A,C=c}) \le \eta(P_{W|A,C}). \tag{26}$$

Averaging both sides of (25) over $v \sim P_V$ and using the definition $\eta_W = \eta(P_{W|\mathrm{pa}(W)}) = \eta(P_{W|A,C})$, we have

$$I(U; W|V) \le \eta_W I(U; A|V)\,. \tag{27}$$

Adding $I(U; V)$ to both sides yields (24).

We now move to proving the percolation bound (23). First, notice that if a vertex W satisfies $W > V$, then letting $\{\exists\, \pi : X \to V\}$ be the event that there exists a directed path from X to (any element of) the set V under the site percolation model, we notice that $\{W \text{ removed}\}$ is independent from $\{\exists\, \pi : X \to V\}$ and $\{\exists\, \pi : X \to V \cup \mathrm{pa}(W)\}$. Thus we have

$$\begin{aligned}
\mathrm{perc}(V \cup \{W\}) &\triangleq \mathbb{P}[\exists\, \pi : X \to V \cup \{W\}] \\
&= \mathbb{P}[\exists\, \pi : X \to V \cup \{W\}, W \text{ removed}] + \mathbb{P}[\exists\, \pi : X \to V \cup \{W\}, W \text{ kept}] \\
&= \mathbb{P}[\exists\, \pi : X \to V, W \text{ removed}] + \mathbb{P}[\exists\, \pi : X \to V \cup \mathrm{pa}(W), W \text{ kept}] \\
&= \mathbb{P}[\exists\, \pi : X \to V](1 - \eta_W) + \eta_W \mathbb{P}[\exists\, \pi : X \to V \cup \mathrm{pa}(W)] \\
&= (1 - \eta_W)\mathrm{perc}(V) + \eta_W \mathrm{perc}(V \cup \mathrm{pa}(W))\,.
\end{aligned}$$

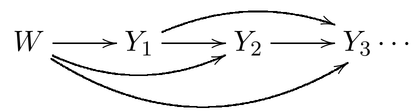
That is, the set-function $\mathrm{perc}(\cdot)$ satisfies the recursion given by the right-hand side of (22). Now notice that (23) holds trivially for $V = \{X\}$, since both sides are equal to 1. Then, by induction on the maximal element of V and applying (22) we get that (23) holds for all V. $\square$

Theorem 5 allows us to estimate contraction coefficients in arbitrary (finite) networks by peeling off last nodes one by one. Next we derive a few corollaries:

Corollary 6. *Consider a fixed (single-letter) channel $P_{Y|X}$ and assume that it is used repeatedly and with perfect feedback to send information from W to $(Y_1, \ldots, Y_n)$. That is, we have for some encoder functions f_j*

$$P_{Y^n|W}(y^n|w) = \prod_{j=1}^{n} P_{Y|X}(y_j|f_j(w, y^{j-1})),$$

which corresponds to the graphical model:



Then

$$\eta_{\mathrm{KL}}(P_{Y^n|W}) \le 1 - (1 - \eta_{\mathrm{KL}}(P_{Y|X}))^n < n \cdot \eta_{\mathrm{KL}}(P_{Y|X})$$

Proof. Apply Theorem 5 n times. □

Let us call a path $\pi = (X, \cdots, v)$ with $v \in V$ to be *shortcut-free from X to V*, denoted $X \stackrel{sf}{\to} V$, if there does not exist another path π' from X to any node in V such that π' is a subset of π. (In particular v necessarily is the first node in V that π visits.) Also for every path $\pi = (X, v_1, \ldots, v_m)$ we define

$$\eta^\pi \triangleq \prod_{j=1}^{m} \eta_{v_j} .$$

Corollary 7. *For any subset V we have*

$$\eta_{\mathrm{KL}}(P_{V|X}) \le \sum_{\pi : X \stackrel{sf}{\to} V} \eta^\pi . \tag{28}$$

In particular, we have the estimate of Evans-Schulman [ES99]:

$$\eta_{\mathrm{KL}}(P_{V|X}) \le \sum_{\pi : X \to V} \eta^\pi . \tag{29}$$

Proof. Both results are simple consequence of union-bounding the right-hand side of (23). But for completeness, we give an explicit proof. First, notice the following two self-evident observations:

1. If A and B are disjoint sets of nodes, then

$$\sum_{\pi : X \stackrel{sf}{\to} A \cup B} \eta^\pi = \sum_{\pi : X \stackrel{sf}{\to} A,\ \text{avoid } B} \eta^\pi + \sum_{\pi : X \stackrel{sf}{\to} B,\ \text{avoid } A} \eta^\pi . \tag{30}$$

2. Let $\pi : X \to V$ and π_1 be π without the last node, then

$$\pi : X \stackrel{sf}{\to} V \quad \Longleftrightarrow \quad \pi_1 : X \stackrel{sf}{\to} \{\mathrm{pa}(V) \setminus V\}. \tag{31}$$

Now represent $V = (V', W)$ with $W > V'$, denote $P = \mathrm{pa}(W) \setminus V$ and assume (by induction) that

$$\eta_{\mathrm{KL}}(P_{V'|X}) \le \sum_{\pi: X \overset{sf}{\to} V} \eta^\pi \tag{32}$$

$$\eta_{\mathrm{KL}}(P_{V',P|X}) \le \sum_{\pi: X \overset{sf}{\to} \{V', P\}} \eta^\pi \,. \tag{33}$$

By (30) and (31) we have

$$\sum_{\pi: X \overset{sf}{\to} V} \eta^\pi = \sum_{\pi: X \overset{sf}{\to} V'} \eta^\pi + \sum_{\pi: X \overset{sf}{\to} W,\ \text{avoid}\ V'} \eta^\pi \tag{34}$$

$$= \sum_{\pi: X \overset{sf}{\to} V'} \eta^\pi + \eta_W \sum_{\pi: X \overset{sf}{\to} P,\ \text{avoid}\ V'} \eta^\pi \tag{35}$$

Then by Theorem 5 and induction hypotheses (32)-(33) we get

$$\eta_{\mathrm{KL}}(P_{V|X}) \le \eta_W \sum_{\pi: X \overset{sf}{\to} \{V', P\}} \eta^\pi + (1 - \eta_W) \sum_{\pi: X \overset{sf}{\to} V'} \eta^\pi \tag{36}$$

$$= \eta_W \left(\sum_{\pi: X \overset{sf}{\to} P,\ \text{avoid}\ V'} \eta^\pi - \sum_{\pi: X \overset{sf}{\to} V',\ \text{pass}\ P} \eta^\pi \right) + \sum_{\pi: X \overset{sf}{\to} V'} \eta^\pi \tag{37}$$

$$\le \eta_W \sum_{\pi: X \overset{sf}{\to} P,\ \text{avoid}\ V'} \eta^\pi + \sum_{\pi: X \overset{sf}{\to} V'} \eta^\pi \tag{38}$$

where in (37) we applied (30) and split the summation over $\pi : X \overset{sf}{\to} V'$ into paths that avoid and pass nodes in P. Comparing (35) and (38) the conclusion follows. ☐

Both estimates (28) and (29) are compared to that of Theorem 5 in Table 1 in various graphical models.

Evaluation for the BSC We consider the contraction coefficient for the n-letter binary symmetric channel $\mathsf{BSC}(\delta)^n$ defined in (4). By (16), for $n = 1$ we have $\eta_{\mathrm{KL}} = (1 - 2\delta)^2$. Then by Corollary 6 we have for arbitrary n:

$$\eta_{\mathrm{KL}} \le 1 - (4\delta(1-\delta))^n \,. \tag{39}$$

A simple lower bound for η_{KL} can be obtained by considering (17) and taking $U \sim \mathrm{Bern}(1/2)$ and $U \to X$ being an n-letter repetition code, namely, $X = (U, \dots, U)$. Let[7] $\epsilon = \mathbb{P}[|Z| \ge n/2]$ be the probability of error for the maximal likelihood decoding of U based on Y, which satisfies the Chernoff bound $\epsilon \le (4\delta(1-\delta))^{n/2}$. We have from Jensen's inequality

$$I(U;Y) = H(U) - H(U|Y) \ge 1 - h(\epsilon) = 1 - (4\delta(1-\delta))^{\frac{n}{2} + O(\log n)} \,,$$

[7] For elements of $\mathbb{F}_2^n$, $|\cdot|$ is the Hamming weight.

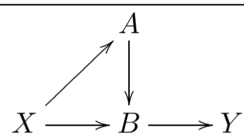
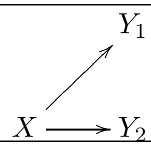
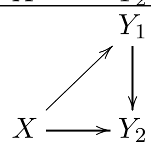
Name	Graph	Theorem 5	Estimate (28) via shortcut-free paths	Original Evans-Schulman estimate (29)
Markov chain 1	$X \to Y_1 \to B \to Y_2$	η	η	$\eta + \eta^3$
Markov chain 2		η^2	η^2	$\eta^2 + \eta^3$
Parallel channels		$2\eta - \eta^2$	2η	2η
Parallel channels with feedback		$2\eta - \eta^2$	2η	3η

Table 1: Comparing bounds on the contraction coefficient $\eta_{\mathrm{KL}}(P_{Y|X})$. For simplicity, we assume that the η_{KL} coefficients of all constituent kernels are bounded from above by η.

where we used the fact that the binary entropy $h(x) = -x\log x - (1-x)\log(1-x) = -x\log x + O(x^2)$ as $x \to 0$. Consequently, we get

$$\eta_{\mathrm{KL}} \geq 1 - (4\delta(1-\delta))^{\frac{n}{2}+O(\log n)} . \tag{40}$$

Comparing (39) and (40) we see that $\eta_{\mathrm{KL}} \to 1$ exponentially fast. To get the exact exponent we need to replace (39) by the following improvement:

$$\eta_{\mathrm{KL}} \leq \eta_{\mathrm{TV}} \leq 1 - (4\delta(1-\delta))^{\frac{n}{2}+O(\log n)} ,$$

where the first inequality is from (8) and the second is from (48) below. Thus, all in all we have for $\mathsf{BSC}(\delta)^n$ as $n \to \infty$

$$\eta_{\mathrm{KL}}, \eta_{\mathrm{TV}} = 1 - (4\delta(1-\delta))^{\frac{n}{2}+O(\log n)} . \tag{41}$$

4 Dobrushin's coefficients in networks

The proof of Theorem 5 relies on the characterization (17) of η_{KL} via mutual information, which satisfies the chain rule. Neither of these two properties is enjoyed by the total variation. Nevertheless, the following is an exact counterpart of Theorem 5 for total variation.

Theorem 8. *Under the same assumption of Theorem 5,*

$$\eta_{\mathrm{TV}}(P_{V,W|X}) \leq (1-\eta_W)\eta_{\mathrm{TV}}(P_{V|X}) + \eta_W \eta_{\mathrm{TV}}(P_{\mathrm{pa}(W),V|X}) , \tag{42}$$

where $\eta_W = \eta_{\mathrm{TV}}(P_{W|\mathrm{pa}(W)})$. Furthermore, let $\mathrm{perc}(V)$ denote the probability that there is a path from X to V in the graph if each node v is removed independently with probability $1-\eta_v$ (site percolation). Then, we have for every $V \subset \mathcal{V}$

$$\eta_{\mathrm{TV}}(P_{V|X}) \leq \mathrm{perc}(V) . \tag{43}$$

In particular, if $\eta_v < 1$ for all $v \in V$, then $\eta_{\mathrm{TV}}(P_{V|X}) < 1$.

Proof. Fix $x, \tilde{x}$ and denote by P (resp. Q) the distribution conditioned on $X = x$ (resp. x'). Denote $Z = \mathrm{pa}(W)$. The goal is to show

$$d_{\mathrm{TV}}(P_{VW}, Q_{VW}) \le (1 - \eta_W) d_{\mathrm{TV}}(P_V, Q_V) + \eta_W d_{\mathrm{TV}}(P_{ZV}, Q_{ZV}) . \tag{44}$$

which, by the arbitrariness of x, x' and in view of the characterization of η in (7), yields the desired (42). By Lemma 22 in Appendix C, there exists a coupling of P_{ZV} and Q_{ZV}, denoted by $\pi_{ZVZ'V'}$, such that

$$\pi[(Z, V) \ne (Z', V')] = d_{\mathrm{TV}}(P_{ZV}, Q_{ZV}) ,$$
$$\pi[V \ne V'] = d_{\mathrm{TV}}(P_V, Q_V)$$

simultaneously (that is, this coupling is jointly optimal for the total variation of the joint distributions and one pair of marginals).

Conditioned on $Z = z$ and $Z' = z'$ and independently of VV', let WW' be distributed according to a maximal coupling of the conditional laws $P_{W|Z=z}$ and $P_{W|Z=z'}$ (recall that $Q_{W|Z} = P_{W|Z} = P_{W|\mathrm{pa}(W)}$ by definition). This defines a joint distribution $\pi_{ZVWZ'V'W'}$, under which we have the Markov chain $VV' \to ZZ' \to WW'$. Then

$$\pi[W \ne W' | ZVZ'V'] = \pi[W \ne W' | ZZ'] = d_{\mathrm{TV}}(P_{W|\mathrm{pa}(W)=Z}, P_{W|\mathrm{pa}(W)=Z'}) \le \eta_W \mathbf{1}_{\{Z \ne Z'\}}.$$

Therefore we have

$$\pi[W \ne W' | V = V'] = \mathbb{E}[\pi[W \ne W' | ZZ'] | V = V']$$
$$\le \eta_W \pi[Z \ne Z' | V = V'].$$

Multiplying both sides by $\pi[V = V']$ and then adding $\pi[V \ne V']$, we obtain

$$\pi[(W, V) \ne (W', V')] \le (1 - \eta_W)\pi[V \ne V'] + \eta_W \pi[(Z, V) \ne (Z', V')]$$
$$= (1 - \eta_W) d_{\mathrm{TV}}(P_V, Q_V) + \eta_W d_{\mathrm{TV}}(P_{ZV}, Q_{ZV}),$$

where the LHS is lower bounded by $d_{\mathrm{TV}}(P_{WV}, Q_{WV})$ and the equality is due to the choice of π. This yields the desired (44), completing the proof of (42). The rest of the proof is done as in Theorem 5. □

As a consequence of Theorem 8, both Corollary 6 and 7 extend to total variation verbatim with η_{KL} replaced by η_{TV}:

Corollary 9. *In the setting of Corollary 6 we have*

$$\eta_{\mathrm{TV}}(P_{Y^n|W}) \le 1 - (1 - \eta_{\mathrm{TV}}(P_{Y|X}))^n < n \cdot \eta_{\mathrm{KL}}(P_{Y|X}) . \tag{45}$$

Corollary 10. *In the setting of Corollary 7 we have*

$$\eta_{\mathrm{TV}}(P_{V|X}) \le \sum_{\pi: X \stackrel{sf}{\to} V} \eta^{\pi}_{\mathrm{TV}} \le \sum_{\pi: X \to V} \eta^{\pi}_{\mathrm{TV}} ,$$

where for any path $\pi = (X, v_1, \ldots, v_m)$ *we denoted* $\eta^{\pi}_{\mathrm{TV}} \triangleq \prod_{j=1}^{m} \eta_{\mathrm{TV}}(P_{v_j|\mathrm{pa}(v_j)})$.

Evaluation for the BSC Consider the n-letter BSC defined in (4), where $Y = X + Z$ with $Z \sim \mathrm{Bern}(\delta)^n$ and $|Z| \sim \mathrm{Binom}(n, \delta)$. By Dobrushin's characterization (7), we have

$$\begin{aligned}
\eta_{\mathrm{TV}} &= \max_{x,x' \in \mathbb{F}_2^n} d_{\mathrm{TV}}(P_{Y|X=x}, P_{Y|X=x'}) \\
&= d_{\mathrm{TV}}(\mathrm{Bern}(\delta)^n, \mathrm{Bern}(1-\delta)^n) \\
&= d_{\mathrm{TV}}(\mathrm{Binom}(n,\delta), \mathrm{Binom}(n, 1-\delta)) \qquad (46)\\
&= 1 - 2\mathbb{P}[|Z| > n/2] - \mathbb{P}[|Z| = n/2] \qquad (47)\\
&= 1 - (4\delta(1-\delta))^{\frac{n}{2} + O(\log n)}, \qquad (48)
\end{aligned}$$

where (46) follows from the sufficiency of $|Z|$ for testing the two distributions, (47) follows from $d_{\mathrm{TV}}(P, Q) = 1 - \int P \wedge Q$ and (48) follows from standard binomial tail estimates (see, e.g., [Ash65, Lemma 4.7.2]). The above sharp estimate should be compared to the bound obtained by applying Corollary 9:

$$\eta_{\mathrm{TV}} \le 1 - (2\delta)^n . \qquad (49)$$

Although (49) correctly predicts the exponential convergence of $\eta_{\mathrm{TV}} \to 1$ whenever $\delta < \frac{1}{2}$, the exponent estimated is not optimal.

5 Bounding F_I-curves in networks

In this section our goal is to produce upper bound bounds on the F_I-curve of a Bayesian network $F_{V|X}$ in terms of those of the constituent channels. For any vertex v of the network, denote the F_I-curve of the channel $P_{v|\mathrm{pa}(v)}$ by $F_{v|\mathrm{pa}(v)}$, abbreviated by F_v, and the concavified version by F_v^c.

Theorem 11. *In the setting of Theorem 5,*

$$F_{V,W|X} \le F_{V|X} + F_W^c \circ (F_{\mathrm{pa}(W),V|X} - F_{V|X}), \qquad (50)$$

$$F_{V,W|X}^c \le F_{V|X}^c + F_W^c \circ (F_{\mathrm{pa}(W),V|X}^c - F_{V|X}^c). \qquad (51)$$

Furthermore, the right-hand side of (51) *is non-negative, concave, nondecreasing and upper bounded by the identity mapping* id.

Remark 1. The F_I-curve estimate in Theorem 11 implies that of contraction coefficients of Theorem 5. To see this, note that since $F_{\mathrm{pa}(W),V|X} \le \mathrm{id}$, the following is a relaxation of (50):

$$\mathrm{id} - F_{V,W|X} \ge (\mathrm{id} - F_W) \circ (\mathrm{id} - F_{V|X}). \qquad (52)$$

Consequently, if each channel in the network satisfies an SDPI, then the end-to-end SDPI is also satisfied. That is, if each vertex has a non-trivial F_I-curve, i.e., $F_v < \mathrm{id}$ for all $v \in \mathcal{V}$, then the channel $X \to V$ also has a strict contractive property, i.e., $F_{V|X} < \mathrm{id}$.

Furthermore, since $F_W^c(t) \le \eta_W t$, noting the fact that $F_{V|X}'(0) = \eta_{\mathrm{KL}}(P_{V|X})$ and taking the derivative on both sides of (50) we see that the latter implies (22).

Proof. We first show that for any channel $P_{Y|X}$, its $F_{Y|X}$-curve satisfies that $t \mapsto t - F_{Y|X}(t)$ is nondecreasing. Indeed, it is known, cf. [CPW15, Section I], that $t \mapsto \frac{F_{Y|X}(t)}{t}$ is nonincreasing. Thus, for $t_1 < t_2$ we have

$$\begin{aligned}
t_2 - F_{Y|X}(t_2) &\ge t_2 - \frac{t_2}{t_1} F_{Y|X}(t_1) \\
&= \frac{t_2}{t_1}\left(t_1 - F_{Y|X}(t_1)\right) \\
&\ge t_1 - F_{Y|X}(t_1),
\end{aligned}$$

where the last step follows from the fact that $F_{Y|X}(t) \le t$. Similarly, for any concave function $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ s.t. $\Phi(0) = 0$ we have $\frac{\Phi(t_2)}{t_2} \le \frac{\Phi(t_1)}{t_1}$. Therefore, the argument above implies $t \mapsto t - \Phi(t)$ is nondecreasing and, in particular, so is $t \mapsto t - F_W^c(t)$.

Let P_{UX} be such that $I(U; X) \le t$ and $I(U; W, V) = F_{V,W|X}(t)$. By the same argument that leads to (27) we obtain

$$\begin{aligned} I(U; W|V = v_0) &\le F_W(I(U; A|V = v_0)) \\ &\le F_W^c(I(U; A|V = v_0))\,. \end{aligned}$$

Averaging over $v_0 \sim P_V$ and applying Jensen's inequality we get

$$I(U; W, V) \le F_W^c(I(U; \mathrm{pa}(W), V) - I(U; V)) + I(U; V).$$

Therefore,

$$\begin{aligned} F_{V,W|X}(t) &\le F_W^c(I(U; \mathrm{pa}(W), V) - I(U; V)) + I(U; V) \\ &\le F_W^c(F_{\mathrm{pa}(W),V|X}(t) - I(U; V)) + I(U; V) \qquad (53) \\ &= F_{\mathrm{pa}(W),V|X}(t) - (\mathrm{id} - F_W^c)(F_{\mathrm{pa}(W),V|X}(t) - I(U; V)) \\ &\le F_{\mathrm{pa}(W),V|X}(t) - (\mathrm{id} - F_W^c)(F_{\mathrm{pa}(W),V|X}(t) - F_{V|X}(t)) \qquad (54) \\ &= F_{V|X}(t) + F_W^c(F_{\mathrm{pa}(W),V|X}(t) - F_{V|X}(t)) \\ &\le F_{V|X}^c(t) + F_W^c(F_{\mathrm{pa}(W),V|X}^c(t) - F_{V|X}^c(t)) \qquad (55) \end{aligned}$$

where (53) and (54) follow from the facts that $t \mapsto F_W(t)$ and $t \mapsto t - F_W(t)$ are both nondecreasing, and (55) follows from that $a + F_W^c(b - a)$ is nondecreasing in both a and b.

Finally, we need to show that the right-hand side of (55) is nondecreasing and concave (this automatically implies that (55) is an upper-bound to the concavification $F_{V|X}^c$). To that end, denote $t_\lambda = \lambda t_1 + (1 - \lambda)t_0$, $f_\lambda = F_{V|X}^c(t_\lambda)$, $g_\lambda = F_{\mathrm{pa}(W),V|X}^c(t_\lambda)$ and notice the chain

$$\begin{aligned} f_\lambda + F_W^c(g_\lambda - f_\lambda) &\ge \lambda f_1 + (1 - \lambda) f_0 + F_W^c(\lambda(g_1 - f_1) + (1 - \lambda)(g_0 - f_0)) \qquad (56) \\ &\ge \lambda(f_1 + F_W^c(g_1 - f_1)) + (1 - \lambda)(f_0 + F_W^c(g_0 - f_0)) \qquad (57) \end{aligned}$$

where (56) is from concavity of $F_{V|X}^c$, $F_{\mathrm{pa}(W),V|X}^c$ and monotonicity of $(a, b) \mapsto a + F_W^c(b - a)$, and (57) is from concavity of F_W^c. □

Corollary 12. *In the setting of Corollary 6 we have*

$$F_{Y^n|W}(t) \le t - \psi^{(n)}(t)\,,$$

where $\psi^{(1)} = \psi$, $\psi^{(k+1)} = \psi^{(k)} \circ \psi$ and $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ is a convex function such that

$$F_{Y|X}(t) \le t - \psi(t)\,.$$

Proof. The case of $n = 1$ follows from the assumption on ψ. The case of $n > 1$ is proved by induction, with the induction step being an application of Theorem 11 with $V = Y^{n-1}$ and $W = Y_n$. □

Generally, the bound of Corollary 12 cannot be improved in the vicinity of zero. As an example where this is tight, consider a parallel erasure channel, whose F_I-curve for $t \le \log q$ is computed in Theorem 17 below.

Evaluation for the BSC To ease the notation, all logarithms are with respect to base two in this section. Let $h(y) = y \log \frac{1}{y} + (1-y) \log \frac{1}{1-y}$ denote the binary entropy function and $h^{-1} : [0,1] \to [0, \frac{1}{2}]$ its functional inverse. Let $p * q \triangleq p(1-q) + q(1-p)$ for $p, q \in [0,1]$ denote binary convolution and define

$$\psi(t) \triangleq t - 1 + h(\delta * h^{-1}(\max(1-t, 0))) \tag{58}$$

which is convex and increasing in t on $\mathbb{R}_+$. For $n = 1$ it was shown in [CPW15, Section 2] that the F_I-curve of $\mathsf{BSC}(\delta)$ is given by

$$F_I(t, \mathsf{BSC}(\delta)) = F_I^c(t, \mathsf{BSC}(\delta)) = t - \psi(t)\,.$$

Applying Corollary 12 we obtain the following bound on the F_I-curve of BSC of blocklength n (even with feedback):

Proposition 13. *Let $Z_1, \ldots, Z_n \overset{i.i.d.}{\sim} Bern(\delta)$ be independent of U. For any (encoder) functions $f_j, j = 1, \ldots, n$, define*

$$X_j = f_j(U, Y^{j-1}), \quad Y_j = X_j + Z_j\,.$$

Then

$$I(U; Y^n) \le I(U; X^n) - \psi^{(n)}(I(U; X^n))\,, \tag{59}$$

where $\psi^{(1)} = \psi$, $\psi^{(k+1)} = \psi^{(k)} \circ \psi$ and ψ is defined in (58).

Remark 2. The estimate (59) was first shown by A. Samorodnitsky (private communication) under extra technical constraints on the joint distribution of (X^n, W) and in the absence of feedback. We have then observed that Evans-Schulman type of technique yields (59) generally.

Since $\psi(t) = 4\delta(1-\delta)t + o(t)$ as $t \to 0$ we get

$$F_I^c(t, \mathsf{BSC}(\delta)^n) \le t - t(4\delta(1-\delta))^{n+o(n)}$$

as $n \to \infty$ for any fixed t. A simple lower bound, for comparison purposes, can be inferred from (40) after noticing that there we have $I(U; X) = 1$, and so

$$F_I^c(1, \mathsf{BSC}(\delta)^n) \ge 1 - (4\delta(1-\delta))^{\frac{n}{2} + O(\log n)}\,,$$

This shows that the bound of Proposition 13 is order-optimal: $F(t) \to t$ exponentially fast. Exact exponent is given by (41).

As another point of comparison, we note the following. Existence of capacity-achieving error-correcting codes then easily implies

$$\lim_{n\to\infty} \frac{1}{n} F_I^c(n\theta, \mathsf{BSC}(\delta)^n) = \min(\theta, C)\,,$$

where $C = 1 - h(\delta)$ is the Shannon capacity of $\mathsf{BSC}(\delta)$. Since for $t > 1$ we have $\psi(t) = t - C$ one can show that

$$\lim_{n\to\infty} \frac{1}{n} \psi^{(n)}(n\theta) = |\theta - C|^+\,,$$

and therefore we conclude that in this sense the bound (59) is asymptotically tight.

6 SDPI via comparison to erasure channels

So far our leading example has been the binary symmetric channel (4). We now consider another important example:

Example 1. For any set $\mathcal{X}$, the *erasure channel* on $\mathcal{X}$ with erasure probability δ is a random transformation from $\mathcal{X}$ to $\mathcal{X} \cup \{?\}$, where $? \notin \mathcal{X}$ defined as

$$P_{E|X}(e|x) = \begin{cases} \delta, & e = ? \\ 1-\delta, & e = x \end{cases}.$$

For $\mathcal{X} = [q]$, we call it the *q-ary erasure channel* denoted by $\mathsf{EC}_q(\delta)$. In the binary case, we denote the binary erasure channel by $\mathsf{BEC}(\delta) \triangleq \mathsf{EC}_2(\delta)$. A simple calculation shows that for every P_{UX} we have

$$I(U;E) = (1-\delta) I(U;X) \tag{60}$$

and therefore for $\mathsf{EC}_q(\delta)$ we have $\eta_{\mathrm{KL}}(P_{E|X}) = 1-\delta$ and $F_I(t) = \min((1-\delta)t, \log q)$.

Next we recall a standard information-theoretic ordering on channels, cf. [EGK11, Section 5.6]:

Definition 2. Given two channels with common input alphabet, $P_{Y|X}$ and $P_{Y'|X}$, we say that $P_{Y'|X}$ is less noisy than $P_{Y|X}$, denoted by $P_{Y|X} \le_{l.n.} P_{Y'|X}$ if for all joint distributions P_{UX} we have

$$I(U;Y) \le I(U;Y'). \tag{61}$$

We also have an equivalent formulation in terms of divergence:

Proposition 14. *$P_{Y|X} \le_{l.n.} P_{Y'|X}$ if and only if for all P_X, Q_X we have*

$$D(Q_Y \| P_Y) \le D(Q_{Y'} \| P_{Y'}) \tag{62}$$

where $P_Y, P_{Y'}, Q_Y, Q_{Y'}$ are the output distributions induced by P_X, Q_X over $P_{Y|X}$ and $P_{Y'|X}$, respectively.

See Appendix A.4 for the proof.[8]

The following result shows that the contraction coefficient of KL divergence can be equivalently formulated as being less noisy than the corresponding erasure channel:[9]

Proposition 15. *For an arbitrary channel $P_{Y|X}$ we have*

$$\eta_{\mathrm{KL}}(P_{Y|X}) \le \eta \iff P_{Y|X} \le_{l.n.} P_{E|X}, \tag{63}$$

where $P_{E|X}$ is the erasure channel on the same input alphabet and erasure probability $1-\eta$.

Proof. The definition of $\eta_{\mathrm{KL}}(P_{Y|X})$ guarantees for every P_{UX}

$$I(U;Y) \le (1-\delta) I(U;X), \tag{64}$$

where the right-hand side is precisely $I(U;E)$ by (60). ☐

[8] It is tempting to put forward a fixed-P_X version of the previous criterion (similar to Theorem 4). That would, however, require some extra assumptions on P_X. Indeed, knowing that $I(W;Y) \le I(W;Y')$ for all $P_{W,X}$ with a given fixed P_X tells us nothing about how distributions $P_{Y|X=x}$ and $P_{Y'|X=x}$ compare outside the support of P_X. (For discrete channels and strictly positive P_X, however, it is easy to argue that indeed (62) holds for all Q_X if and only if (61) holds for all $P_{U,X}$ with a given marginal P_X.)

[9] Note that another popular partial order for random transformations – that of stochastic degradation – may also be related to contraction coefficients, see [Rag14, Remark 3.2].

It turns out that the notion of less-noisiness tensorizes:

Proposition 16. *If $P_{Y_1|X_1} \le_{l.n.} P_{Y_1'|X_1}$ and $P_{Y_2|X_2} \le_{l.n.} P_{Y_2'|X_2}$ then*

$$P_{Y_1|X_1} \times P_{Y_2|X_2} \le_{l.n.} P_{Y_1'|X_1} \times P_{Y_2'|X_2}$$

In particular,

$$\eta_{\mathrm{KL}}(P_{Y|X}) \le \eta \quad \implies \quad P_{Y|X}^n \le_{l.n.} P_{E|X}^n \,. \tag{65}$$

where $P_{E|X}$ is the erasure channel on the same input alphabet and erasure probability $1-\eta$.

Proof. Construct a relevant joint distribution $U \to X^2 \to (Y^2, Y'^2)$ and consider

$$I(U; Y_1, Y_2) = I(U; Y_1) + I(U; Y_2|Y_1) \,. \tag{66}$$

Now since $U \perp\!\!\!\perp Y_2|Y_1$ we have by $P_{Y_2|X_2} \le_{l.n.} P_{Y_2'|X_2}$

$$I(U; Y_2|Y_1) \le I(U; Y_2'|Y_1)$$

and putting this back into (66) we get

$$I(U; Y_1, Y_2) \le I(U; Y_1) + I(U; Y_2'|Y_1) = I(U; Y_1, Y_2') \,.$$

Repeating the same argument, but conditioning on Y_2' we get

$$I(U; Y_1, Y_2) \le I(U; Y_1', Y_2') \,,$$

as required. The last claim of the proposition follows from Proposition 15. □

Consequently, everything that has been said in this paper about $\eta_{\mathrm{KL}}(P_{Y|X})$ can be restated in terms of seeking to compare a given channel in the sense of the $\le_{l.n.}$ order to an erasure channel. It seems natural, then, to consider erasure channel in somewhat greater details.

6.1 F_I-curve of erasure channels

Theorem 17. *Consider the q-ary erasure channel of blocklength n and erasure probability δ. Its F_I-curve is bounded by*

$$F_I^c(t, \mathsf{EC}_q(\delta)^n) \le \mathbb{E}[\min(B \log q, t)], \qquad B \sim \mathrm{Binom}(n, 1-\delta) \,. \tag{67}$$

The bound is tight in the following cases:

1. *at $t = k \log q$ with integral $k \le n$ if and only if an $(n, k, n-k+1)_q$ MDS code exists*[10]
2. *for $t \le \log q$ and $t \ge (n-1)\log q$;*
3. *for all t when $n = 1, 2, 3$.*

Remark 3. Introducing $B' \sim \mathrm{Binom}(n-1, 1-\delta)$ and using the identity $\mathbb{E}[B \mathbf{1}_{\{B \le a\}}] = n(1-\delta)\mathbb{P}[B' \le a-1]$, we can express the right-hand side of (67) in terms of binomial CDFs:

$$\mathbb{E}[\min(B, x)] = x + \mathbb{P}[B' \le \lfloor x \rfloor - 1](1-\delta)(n-x) - x\delta\mathbb{P}[B' \le \lfloor x \rfloor]$$

This implies that the upper bound (67) is piecewise-linear, increasing and concave.

[10] We remind that a subset $\mathcal{C}$ of $[q]^n$ is called an $(n, k, d)_q$ code if $|\mathcal{C}| = q^k$ and Hamming distance between any two points from $\mathcal{C}$ is at least d. A code is called maximum-distance separable (MDS) if $d = n - k + 1$. This is equivalent to the property that projection of $\mathcal{C}$ onto any subset of k coordinates is bijective.

Proof. Consider arbitrary $U \to X^n \to E^n$ with $P_{E^n|X^n} = \mathsf{EC}_q(\delta)^n$. Let S be random subset of $[n]$ which includes each $i \in [n]$ independently with probability $1-\delta$. A direct computation, shows that

$$I(U; E^n) = I(U; X_S, S) = \sum_{\sigma \subset [n]} \mathbb{P}[S = \sigma] I(U; X_\sigma) \tag{68}$$

$$\leq \sum_{\sigma \subset [n]} \mathbb{P}[S = \sigma] \min(|\sigma| \log q, t) = \mathbb{E}[\min(B \log q, t)] \,. \tag{69}$$

From here (67) follows by taking supremum over P_{U,X^n}.

Claims about tightness follow by constructing $U = X^n$ and taking X^n to be the output of the MDS code (so that $H(X_\sigma) = \min(|\sigma| \log q, t))$ and invoking the concavity of $F_I(t)$. One also notes that $[n, 1, n]_q$ (repetition code) and $[n, n-1, 2]$ (single parity check code) show tightness at $t = \log q$ and $t = (n-1)\log q$.

Finally, we prove that when $t = k \log q$ and the bound (67) is tight then a (possibly non-linear) $(n, k, n-k+1)_q$ MDS code must exist. First, notice that the right-hand side of (67) is a piecewise-linear and concave function. Thus the bound being tight for $F_I(t)$ (that is a concave-envelope of $F_I(t)$) should also be tight as a bound for $F_I(t)$. Consequently, there must exist $U \to X^n \to E^n$ such that the bound (69) is tight with $t = I(U; X^n)$. This implies that we should have

$$I(U; X_\sigma) = \min(\sigma \log q, t) \tag{70}$$

for all $\sigma \subset [n]$. In particular, we have $I(U; X_i) = \log q$ and thus $H(X_i|U) = 0$ and without loss of generality we may assume that $U = X^n$. Again from (70) we have that $H(X^n) = H(X^k) = k \log q$. This implies that X^n is a uniform distribution on a set of size q^k and projection on any k coordinates is injective. This is exactly the characterization of an MDS code (possibly non-linear) with parameters $(n, k, n-k+1)_q$. ☐

We also formulate some interesting observations for binary erasure channels:

Proposition 18. *For* $\mathsf{BEC}(n, \delta)$ *we have:*

1. *For* $n \geq 3$ *we have that* $F_I(t)$ *is not concave. More exactly,* $F_I(t) < F_I^c(t)$ *for* $t \in (1, 2)$*.*
2. *For arbitrary* n *and* $t \leq \log 2$ *or* $t \geq (n-1)\log 2$ *we have* $F_I(t) = F_I^c(t) = \mathbb{E}[\min(B \log 2, t)]$ *with* B *defined in in* (67)*.*
3. *For* $t = 2, n = 4$ *the bound* (67) *is not tight and* $F_I^c(t) < \mathbb{E}[\min(B \log 2, t)]$*.*

Proof. First note that in Definition 1 of $F_I(t)$ the supremum is a maximum and and U can be restricted to alphabet of size $|\mathcal{X}| + 2$. So in particular, $F_I(t) = f$ if and only if there exists $I(U; Y^n) = f$, $I(U; X^n) \leq t$.

Now consider $t \in (1, 2)$ and $n = 3$ and suppose (U, X^n) achieves the bound. For the bound to be tight we must have $I(U; X^3) = t$. For the bound to be tight we must have $I(U; X_i) = 1$ for all i, that is $H(X_i) = 1$, $H(X_i|U) = 0$ and $H(X^n|U) = 0$. Consequently, without loss of generality we may take $U = X^n$. So for the bound to be tight we need to find a distribution s.t.

$$H(X^3) = H(X_1, X_2) = H(X_2, X_3) = H(X_1, X_3) = t, H(X_1) = H(X_2) = H(X_3) = 1. \tag{71}$$

It is straightforward to verify that this set of entropies satisfies Shannon inequalities (i.e. submodularity of entropy checks), so the main result of [ZY97] shows that there does exist a sequence of triples X^3 (over large alphabets) which attains this point. We will show, however, that this is impossible for binary-valued random variables. First, notice that the set of achievable entropy vectors by binary triplets is a closed subset of $\mathbb{R}_+^7$ (as a continuous image of a compact set). Thus, it is sufficient to show that (71) itself is not achievable.

Second, note that for any pair A, B of binary random variables with uniform marginals we must have

$$A = B + Z, \qquad B \perp\!\!\!\perp Z \sim \text{Bern}(p)\,.$$

Without loss of generality, assume that $X_2 = X_1 + Z$ where $H(Z) = t - 1 > 0$. Moreover, $H(X_3|X_1, X_2) = 0$ implies that $X_3 = f(X_1, X_2)$ for some function f.

Given X_1 we have $H(X_3|X_1 = x) = H(X_3|X_2 = x) = t - 1 > 0$. So the function $X_1 \mapsto f(X_1, x)$ should not be constant for either choice of $x \in \{0, 1\}$ and the same holds for $X_2 \mapsto f(x, X_2)$. Eliminating cases leaves us with $f = X_1 + X_2$ or $f = X_1 + X_2 + 1$. But then $X_3 = X_1 + X_2 = Z$ and $H(X_3) < 1$, which is a contradiction.

Since by Theorem 17 we know that the bound (67) is tight for $F_I(t)$ we conclude that

$$F_I(t) < F_I^c(t), \quad \forall t \in (1, 2)\,.$$

To show the second claim consider $U = X^n$ and $X_1 = \cdots = X_n \sim \text{Bern}(p)$ for $t \le \log 2$. For $t \ge (n-1)\log 2$ take X^{n-1} to be iid $\text{Bern}(\frac{1}{2})$ and

$$X_n = X_1 + \cdots + X_{n-1} + Z\,,$$

where $Z \sim \text{Bern}(p)$. This yields $I(U; X_\sigma) = H(X_\sigma) = |\sigma| \log 2$ for every subset $\sigma \subset [n]$ of size up to $n - 1$. Consequently, the bound (67) must be tight.

Finally, third claim follows from Theorem 17 and the fact that there is no $[4, 2, 3]$ binary code, e.g. [MS77, Corollary 7, Chapter 11]. □

Putting together (65) and (67) we get the following upper bound on the concavified F_I-curve of n-letter product channels in terms of the contraction coefficient of the single-letter channel.

Corollary 19. *If $\eta_{\text{KL}}(P_{Y|X}) = \eta$, then*

$$F_I^c(t, P_{Y|X}^n) \le \mathbb{E}[\min(B \log q, t)], \qquad B \sim \text{Binom}(n, 1-\delta)\,.$$

This gives an alternative proof of Corollary 6 for the case of no feedback.

6.2 Samorodnitsky's SDPI

So far, we have been concerned with bounding the "output" mutual information in terms of a certain "input" one. However, frequently, one is interested in bounding some "output" information given knowledge of several input ones. For example, for the parallel channel we have shown that

$$I(W; Y^n) \le (1 - (1 - \eta_{\text{KL}}(P_{Y|X}))^n) I(W; X^n)\,.$$

But it turns out that a stronger bound can be given if we have finer knowledge about the joint distribution of W and X^n.

The following bound can be distilled from [Sam15]:

Theorem 20 (Samorodnitsky)**.** *Consider the Bayesian network*

$$U \to X^n \to Y^n\,,$$

where $P_{Y^n|X^n} = \prod_{i=1}^n P_{Y_i|X_i}$ *is a memoryless channel with* $\eta_i \triangleq \eta_{KL}(P_{Y_i|X_i})$*. Then we have*

$$I(U;Y^n) \le I(U;X_S|S) = I(U;X_S,S)\,, \tag{72}$$

where $S \perp\!\!\!\perp (U, X^n, Y^n)$ *is a random subset of* $[n]$ *generated by independently sampling each element* i *with probability* η_i*. In particular, if* $\eta_i = \eta$ *for all* i*, then*

$$I(U;Y^n) \le \sum_{\sigma\subset[n]} \eta^{|\sigma|}(1-\eta)^{n-|\sigma|} I(U;X_\sigma) \tag{73}$$

Proof. Just put together characterization (63), tensorization property Proposition 16 to get $I(U;Y^n) \le I(U;E^n)$, where E^n is the output of the product of erasure channels with erasure probabilities $1-\eta_i$. Then the calculation (68) completes the proof. □

Remark 4. Let us say that "total" information $I(U;X^n)$ is distributed among subsets of $[n]$ as given by the following numbers:

$$I_k \triangleq \binom{n}{k}^{-1} \sum_{T\in\binom{[n]}{k}} I(U;X_T)\,.$$

Then bound (73) says (replacing $\mathrm{Binom}(n,\eta)$ by its mean value ηn):

$$I(U;Y^n) \lesssim I_{\eta n}\,.$$

Informally: the only kind of information about U that has a chance to be inferred on the basis of Y^n is one that is contained in subsets of X of size at most ηn.

Remark 5. Another implication of the Theorem is a strengthening of the Mrs. Gerber's Lemma. Fix a single-letter channel $P_{Y|X}$ and suppose that for some increasing *convex* function $m(\cdot)$ and all random variables X we have

$$H(Y) \ge m(H(X))\,.$$

Then, in the setting of the Theorem we have

$$H(Y^n) \ge m\left(\frac{1}{\eta n} H(X_S|S)\right)\,. \tag{74}$$

Note that by Han's inequality (74) is strictly better than the simple consequence of the chain rule: $H(Y^n) \ge nm(H(X^n)/n)$. For the case of $P_{Y|X} = \mathsf{BSC}(\delta)$ the bound (74) is a sharpening of the Mrs. Gerber's Lemma, and has been the focus of [Sam15], see also [Ord16]. To prove (74) let $X^n \to E^n$ be $\mathsf{EC}(1-\eta)$. Then, by Theorem 20 applied to $U = X_i$, $n = i-1$ we have

$$H(X_i|Y^{i-1}) \ge H(X_i|E^{i-1})\,.$$

Thus, from the chain rule and convexity of $m(\cdot)$ we obtain

$$H(Y^n) = \sum_i H(Y_i|Y^{i-1}) \ge nm\left(\frac{1}{n}\sum_i H(X_i|E^{i-1})\right)\,,$$

and the proof is completed by computing $H(E^n)$ in two ways:

$$\begin{aligned} nh(\eta) + H(X_S|S) &= H(E^n) \\ &= \sum_i H(E_i|E^{i-1}) = \sum_i h(\eta) + \eta H(X_i|E^{i-1})\,. \end{aligned}$$

Remark 6. Using Proposition 14 we may also state a divergence version of the Theorem: In the setting of Theorem 20 for any pair of distributions P_{X^n} and Q_{X^n} we have

$$D(P_{Y^n}\|Q_{Y^n}) \le D(P_{X_S|S}\|Q_{X_S|S}|P_S)\,.$$

Similarly, we may extend the argument in the previous remark: If for a fixed Q_X, Q_Y (not necessarily related by $P_{Y|X}$) there exists an increasing concave function f such that for all P_X and $P_Y = P_{Y|X}\circ P_X$ we have

$$D(P_X\|Q_X) \le f(D(P_Y\|Q_Y)) \qquad \forall P_X$$

then

$$D(P_{Y^n}\|(Q_Y)^n) \le nf\left(\frac{1}{\eta n}D(P_{X_S|S}\|\prod_{i\in S}Q_X|P_S)\right).$$

Acknowledgment

We thank Prof. M. Raginsky for references [BK98, Daw75, Gol79] and Prof. A. Samorodnitsky for discussions on Proposition 13 with us. We also thank Aolin Xu for pointing out (41). We are grateful to an anonymous referee for helpful comments.

A Contraction coefficients on general spaces

A.1 Proof of Theorem 2

We show that

$$\eta_f(P_{Y|X}, P_X) = \sup_{Q_X}\frac{D_f(Q_Y\|P_Y)}{D_f(Q_X\|P_X)} \ge \eta_{\chi^2}(P_{Y|X}, P_X) = \sup_{Q_X}\frac{\chi^2(Q_Y\|P_Y)}{\chi^2(Q_X\|P_X)}, \tag{75}$$

where both suprema are over all Q_X such that the respective denominator is in $(0,\infty)$. With the assumption that P_X is not a point mass, namely, there exists a measurable set E such that $P_X(E)\in(0,1)$, it is clear that such Q_X always exists. For example, let $Q_X = \frac{1}{2}(P_X + P_{X|X\in E})$, where $P_{X|X\in E}(\cdot) \triangleq \frac{P_X(\cdot\cap E)}{P_X(E)}$. Then $\frac{1}{2} \le \frac{dQ_X}{dP_X} \le \frac{1}{2}(1+\frac{1}{P_X(E)})$ and hence $D_f(Q_X\|P_X) < \infty$ since f is continuous. Furthermore, $Q_X \ne P_X$ implies that $D_f(Q_X\|P_X)\ne 0$ [Csi67].

The proof follows that of [CIR+93, Theorem 5.4] using the local quadratic behavior of f-divergence; however, in order to deal with general alphabets, additional approximation steps are needed to ensure the likelihood ratio is bounded away from zero and infinity.

Fix Q_X such that $\chi^2(Q_X\|P_X)<\infty$. Let $A = \{x : \frac{dQ_X}{dP_X}(x) < a\}$ where $a>0$ is sufficiently large such that $Q_X(A)\ge 1/2$. Let $Q'_X = Q_{X|X\in A}$ and $Q'_Y = P_{Y|X}\circ Q'_X$. Then $\frac{dQ'_Y}{dP_Y} \le \frac{a}{Q_X(A)} \le 2a$. Let $Q''_X = \frac{1}{a}P_X + (1-\frac{1}{a})Q'_X$ and $Q''_Y = P_{Y|X}\circ Q'_X = \frac{1}{a}P_Y + (1-\frac{1}{a})Q'_Y$. Then we have

$$\frac{1}{a} \le \frac{dQ''_X}{dP_X} \le 2a + \frac{1}{a}, \quad \frac{1}{a} \le \frac{dQ''_Y}{dP_Y} \le 2a + \frac{1}{a}. \tag{76}$$

Note that $\chi^2(Q'_X\|P_X) = \frac{1}{Q(X\in A)}\mathbb{E}_P[(\frac{dQ_X}{dP_X})^2\mathbf{1}_{\{X\in A\}}] - 1$. By dominated convergence theorem, $\chi^2(Q'_X\|P_X)\to\chi^2(Q_X\|P_X)$ as $a\to\infty$. On the other hand, since $Q'_Y\to Q_Y$ pointwise, the weak lower-semicontinuity of χ^2-divergence yields $\liminf_{a\to\infty}\chi^2(Q'_Y\|P_Y)\ge\chi^2(Q_Y\|P_Y)$. Furthermore, using the simple fact that $\chi^2(\epsilon P + (1-\epsilon)Q\|P) = (1-\epsilon)^2\chi^2(Q\|P)$, we have $\frac{\chi^2(Q''_X\|P_X)}{\chi^2(Q''_Y\|P_Y)} = \frac{\chi^2(Q'_X\|P_X)}{\chi^2(Q'_Y\|P_Y)}$.

Therefore, to prove (75), it suffices to show for each fixed a, for any $\delta > 0$, there exists $\tilde{P}_X$ such that $\frac{D_f(\tilde{P}_X\|\tilde{P}_Y)}{D_f(Q_X\|P_X)} \geq \frac{\chi^2(Q''_X\|P_X)}{\chi^2(Q''_Y\|P_Y)} - \delta$.

For $0 < \epsilon < 1$, let $\tilde{P}_X = \bar{\epsilon}P_X + \epsilon Q''_X$, which induces $\tilde{P}_Y = P_{Y|X} \circ \tilde{P}_X = \bar{\epsilon}P_Y + \epsilon Q''_Y$. Then $D_f(\tilde{P}_X\|P_X) = \mathbb{E}_{P_X}[f(1 + \epsilon(\frac{dQ''_X}{dP_X} - 1))]$. Recall from (76) that $\frac{dQ''_X}{dP_X} \in [\frac{1}{a}, \frac{1}{a} + 2a]$. Since f'' is continuous and $f''(1) = 1$, by Taylor's theorem and dominated convergence theorem, we have $D_f(\tilde{P}_X\|P_X) = \frac{\epsilon^2}{2}\chi^2(Q''_X\|P_X)(1+o(1))$. Analogously, $D_f(\tilde{P}_Y\|P_Y) = \frac{\epsilon^2}{2}\chi^2(Q''_Y\|P_Y)(1+o(1))$. This completes the proof of $\eta_f(P_X) \geq \eta_{\chi^2}(P_X)$.

Remark 7. In the special case of KL divergence, we can circumvent the step of approximating by bounded likelihood ratio: By [PW15, Lemma 4.2], since $\chi^2(Q_Y\|P_Y) \leq \chi^2(Q_X\|P_X) < \infty$, we have $D(\tilde{P}_X\|P_X) = \epsilon^2\chi^2(Q_X\|P_X)/2 + o(\epsilon^2)$ and $D(\tilde{P}_Y\|P_Y) = \epsilon^2\chi^2(Q_Y\|P_Y)/2 + o(\epsilon^2)$, as $\epsilon \to 0$. Therefore $\frac{\chi^2(Q_Y\|P_Y)}{\chi^2(Q_X\|P_X)} \leq \lim_{\epsilon\to 0} \frac{D(\tilde{P}_Y\|P_Y)}{D(\tilde{P}_X\|P_X)} \leq \eta_{\mathrm{KL}}(P_X)$. Therefore $\eta_{\mathrm{KL}}(P_X) \geq \eta_{\chi^2}(P_X)$

A.2 Proof of Theorem 3

We prove

$$\eta_{\mathrm{KL}} = \eta_{\chi^2}. \tag{77}$$

First of all, $\eta_{\mathrm{KL}} \geq \eta_{\chi^2}$ follows from Theorem 2. For the other direction we closely follow the argument of [CRS94, Theorem 1]. Below we prove the following integral representation:

$$D(Q\|P) = \int_0^\infty \chi^2(Q\|P^t)\mathrm{d}t, \tag{78}$$

where $P^t \triangleq \frac{tQ+P}{1+t}$. Then

$$\begin{aligned} D(Q_Y\|P_Y) &= \int_0^\infty \chi^2(Q_Y\|P^t_Y)\mathrm{d}t \\ &\leq \int_0^\infty \eta_{\chi^2} \cdot \chi^2(Q_X\|P^t_X)\mathrm{d}t = \eta_{\chi^2} D(Q_X\|P_X). \end{aligned}$$

where we used $P^t_Y = P_{Y|X} \circ P^t_X$. It remains to check (78). Note that

$$-\log x = \int_0^\infty \frac{1-x}{(x+t)(1+t)}\mathrm{d}t$$

Therefore

$$D(Q\|P) = \int_0^\infty \frac{1}{1+t}\mathbb{E}_Q\left[\frac{\mathrm{d}Q - \mathrm{d}P}{\mathrm{d}P + t\mathrm{d}Q}\right]\mathrm{d}t$$

Note that $t\mathbb{E}_Q\left[\frac{\mathrm{d}Q-\mathrm{d}P}{\mathrm{d}P+t\mathrm{d}Q}\right] = -\mathbb{E}_P\left[\frac{\mathrm{d}Q-\mathrm{d}P}{\mathrm{d}P+t\mathrm{d}Q}\right]$. Therefore $\mathbb{E}_Q\left[\frac{\mathrm{d}Q-\mathrm{d}P}{\mathrm{d}P+t\mathrm{d}Q}\right] = \frac{1}{1+t}\int \frac{(\mathrm{d}Q-\mathrm{d}P)^2}{\mathrm{d}P+t\mathrm{d}Q} = (1+t)\chi^2(Q\|P^t)$, completing the proof of (78).

It is instructive to remark how this result was established for finite alphabets originally in [AG76]. Consider the map

$$P_X \mapsto V_r(P_X, Q_X) \triangleq D(P_{Y|X} \circ P_X\|P_{Y|X} \circ Q_X) - rD(P_X\|Q_X)\,.$$

A simple differentiation shows that Hessian of this map at P_X is negative-definite if and only if $r > \eta_{\chi^2}(P_{Y|X}, P_X)$ and negative semidefinite if and only if $r \geq \eta_{\chi^2}(P_{Y|X}, P_X)$ (note that this does not depend on Q_X). Thus, taking $r = \eta_{\chi^2}(P_{Y|X})$ the map $P_X \mapsto V_r(P_X, Q_X)$ is concave in P_X for all Q_X. Thus, its local extremum at $P_X = Q_X$ is a global maximum and hence $V_r(P_X, Q_X) \leq 0$.

A.3 Proof of Theorem 4

We shall assume that P_X is not a point mass, namely, there exists a measurable set E such that $P_X(E) \in (0,1)$. Define

$$\eta_{\mathrm{KL}}(P_X) = \sup_{Q_X} \frac{D(Q_Y \| P_Y)}{D(Q_X \| P_X)}$$

where the supremum is over all Q_X such that $0 < D(Q_X \| P_X) < \infty$. It is clear that such Q_X always exists (e.g., $Q_X = P_{X|X\in E}$ and $D(Q_X\|P_X) = \log \frac{1}{P_X(E)} \in (0,\infty)$). Let

$$\eta_I(P_X) = \sup \frac{I(U;Y)}{I(U;X)}$$

where the supremum is over all Markov chains $U \to X \to Y$ with fixed P_{XY} such that $0 < I(U;X) < \infty$. Such Markov chains always exist, e.g., $U = \mathbf{1}_{\{X\in E\}}$ and then $I(U;X) = h(P_X(E)) \in (0, \log 2)$. The goal of this appendix is to prove (18), namely

$$\eta_{\mathrm{KL}}(P_X) = \eta_I(P_X)\,.$$

The inequality $\eta_I(P_X) \le \eta_{\mathrm{KL}}(P_X)$ follows trivially:

$$I(U;Y) = D(P_{Y|U}\|P_Y|P_U) \le \eta_{\mathrm{KL}}(P_X) D(P_{X|U}\|P_X|P_U) = \eta_{\mathrm{KL}}(P_X) I(X;U)\,.$$

For the other direction, fix Q_X such that $0 < D(Q_X\|P_X) < \infty$. First, consider the case where $\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}$ is bounded, namely, $\frac{\mathrm{d}Q_X}{\mathrm{d}P_X} \le a$ for some $a > 0$ Q_X-a.s. For any $\epsilon \le \frac{1}{2a}$, let $U \sim \mathrm{Bern}(\epsilon)$ and define the probability measure $\tilde{P}_X = \frac{P_X - \epsilon Q_X}{1-\epsilon}$. Let $P_{X|U=0} = \tilde{P}_X$ and $P_{X|U=1} = Q_X$, which defines a Markov chain $U \to X \to Y$ such that X, Y is distributed as the desired P_{XY}. Note that

$$\frac{I(U;Y)}{I(U;X)} = \frac{\bar{\epsilon} D(\tilde{P}_Y\|P_Y) + \epsilon D(Q_Y\|P_Y)}{\bar{\epsilon} D(\tilde{P}_X\|P_X) + \epsilon D(Q_X\|P_X)}$$

where $\bar{\epsilon} = 1-\epsilon$ and $\tilde{P}_Y = P_{Y|X} \circ \tilde{P}_X$. We claim that

$$D(\tilde{P}_X\|P_X) = o(\epsilon), \tag{79}$$

which, in view of the data processing inequality $D(\tilde{P}_X\|P_X) \le D(\tilde{P}_Y\|P_Y)$, implies $\frac{I(U;Y)}{I(U;X)} \xrightarrow{\epsilon\downarrow 0} \frac{D(Q_Y\|P_Y)}{D(Q_X\|P_X)}$ as desired. To establish (79), define the function

$$f(x,\epsilon) \triangleq \begin{cases} \frac{1-\epsilon x}{\epsilon(1-\epsilon)} \log \frac{1-\epsilon x}{1-\epsilon}\,, & \epsilon > 0 \\ (x-1)\log e, & \epsilon = 0\,. \end{cases}$$

One easily notices that f is continuous on $[0,a]\times[0,\frac{1}{2a}]$ and thus bounded. So we get, by bounded convergence theorem,

$$\frac{1}{\epsilon} D(\tilde{P}_X\|P_X) = \mathbb{E}_{P_X}\left[f\left(\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}, \epsilon\right)\right] \to \mathbb{E}_{P_X}\left[\frac{\mathrm{d}Q_X}{\mathrm{d}P_X} - 1\right]\log e = 0\,.$$

To drop the boundedness assumption on $\frac{\mathrm{d}Q_X}{\mathrm{d}P_X}$ we simply consider the conditional distribution $Q'_X \triangleq Q_{X|X\in A}$ where $A = \{x : \frac{\mathrm{d}Q_X}{\mathrm{d}P_X}(x) < a\}$ and $a > 0$ is sufficiently large so that $Q_X(A) > 0$.

Clearly, as $a \to \infty$, we have $Q'_X \to Q_X$ and $Q'_Y \to Q_Y$ pointwise (i.e. $Q'_Y(E) \to Q_Y(E)$ for every measurable set E), where $Q'_Y \triangleq P_{Y|X} \circ Q'_X$. Hence the lower-semicontinuity of divergence yields

$$\liminf_{a\to\infty} D(Q'_Y \| P_Y) \geq D(Q_Y \| P_Y)\,.$$

Furthermore, since $\frac{dQ'_X}{dP_X} = \frac{1}{Q_X(A)} \frac{dQ_X}{dP_X} \mathbf{1}_A$, we have

$$D(Q'_X \| P_X) = \log \frac{1}{Q_X(A)} + \frac{1}{Q_X(A)} \mathbb{E}_Q \left[\log \frac{dQ_X}{dP_X} \mathbf{1}\left\{ \frac{dQ_X}{dP_X} \leq a \right\} \right]. \tag{80}$$

Since $Q_X(A) \to 1$, by dominated convergence (note: $\mathbb{E}_Q[|\log \frac{dQ_X}{dP_X}|] < \infty$) we have $D(Q'_X \| P_X) \to D(Q_X \| P_X)$. Therefore,

$$\liminf_{a\to\infty} \frac{D(Q'_Y \| P_Y)}{D(Q'_X \| P_X)} \geq \frac{D(Q_Y \| P_Y)}{D(Q_X \| P_X)}\,,$$

completing the proof.

A.4 Proof of Proposition 14

First, notice the following simple result:

$$D(Q \| \lambda P + \bar{\lambda} Q) = o(\lambda), \lambda \to 0 \qquad \Longleftrightarrow \qquad P \ll Q \tag{81}$$

Indeed, if $P \not\ll Q$ then there is a set E with $p = P[E] > 0 = Q[E]$. Denote the binary divergence by $d(p\|q) \triangleq D(\mathrm{Bern}(p)\|\mathrm{Bern}(q))$. Applying data-processing for divergence to $X \mapsto 1_E(X)$, we get

$$D(Q \| \lambda P + \bar{\lambda} Q) \geq d(0 \| \lambda p) = \log \frac{1}{1 - \lambda p}$$

and the derivative at $\lambda \to 0$ is non-zero. If $P \ll Q$, then let $f = \frac{dP}{dQ}$ and notice

$$\log \bar{\lambda} \leq \log(\bar{\lambda} + \lambda f) \leq \lambda (f - 1) \log e\,.$$

Dividing by λ and assuming $\lambda < \frac{1}{2}$ we get

$$\left| \frac{1}{\lambda} \log(\bar{\lambda} + \lambda f) \right| \leq C_1 f + C_2\,,$$

for some absolute constants C_1, C_2. Thus, by the dominated convergence theorem we get

$$\frac{1}{\lambda} D(Q \| \lambda P + \bar{\lambda} Q) = -\int dQ \left(\frac{1}{\lambda} \log(\bar{\lambda} + \lambda f) \right) \to \int dQ (1 - f) = 0\,.$$

Another observation is that

$$\lim_{\lambda\to 0} D(P \| \lambda P + \bar{\lambda} Q) = D(P \| Q)\,, \tag{82}$$

regardless of the finiteness of the right-hand side (this is a property of all convex lower-semicontinuous functions).

Now, we prove Proposition 14. One direction is easy: if $D(Q_Y \| P_Y) \leq D(Q_{Y'} \| P_{Y'})$ then

$$I(W;Y) = D(P_{Y|W} \| P_Y | P_W) \leq D(P_{Y'|W} \| P_{Y'} | P_W) = I(W;Y')\,.$$

For the other direction, consider an arbitrary pair (P_X, Q_X). Let $W = \text{Bern}(\epsilon)$ and $P_{X|W=0} = P_X$, $P_{X|W=1} = Q_X$. Then, we get

$$I(W;Y) = \bar{\epsilon} D(P_Y \| \bar{\epsilon} P_Y + \epsilon Q_Y) + \epsilon D(Q_Y \| \bar{\epsilon} P_Y + \epsilon Q_Y)\,,$$

and similarly for $I(W;Y')$. Assume that $D(Q_{Y'} \| P_{Y'}) < \infty$, for otherwise (62) holds trivially. Then $Q_{Y'} \ll P_{Y'}$ and we get from (81) and (82) that

$$I(W;Y') = \epsilon D(Q_{Y'} \| P_{Y'}) + o(\epsilon)\,. \tag{83}$$

On the other hand, again from (82)

$$I(W;Y) \geq \epsilon D(Q_Y \| \bar{\epsilon} P_Y + \epsilon Q_Y) = \epsilon D(Q_Y \| P_Y) + o(\epsilon)\,. \tag{84}$$

Since by assumption $I(W;Y) \leq I(W;Y')$ we conclude from comparing (83) to (84) that $D(Q_Y \| P_Y) \leq D(Q_{Y'} \| P_{Y'}) < \infty$, completing the proof.

B Contraction coefficients for binary-input channels

In this appendix we provide a tight characterization of the KL contraction coefficient for binary-input channel $P_{Y|X}$, where $X \in \{0,1\}$ and Y is arbitrary. Clearly, $\eta_{\text{KL}}(P_{Y|X})$ is a function of $P \triangleq P_{Y|X=0}$ and $Q \triangleq P_{Y|X=1}$, which we abbreviate as $\eta(\{P,Q\})$. The behavior of this quantity closely resembles that of divergence between distributions. Indeed, we expect $\eta(\{P,Q\})$ to be bigger if P and Q are more dissimilar and, furthermore, $\eta(\{P,Q\}) = 0$ (resp. 1) if and only if $P = Q$ (resp. $P \perp Q$). Next we show that $\eta(\{P,Q\})$ is essentially equivalent to Hellinger distance:

Theorem 21. *Consider a binary input channel $P_{Y|X} : \{0,1\} \to \mathcal{Y}$ with $P_{Y|X=0} = P$ and $P_{Y|X=1} = Q$. Then, its contraction coefficient $\eta_{\text{KL}}(P_{Y|X}) = \eta_{\chi^2}(P_{Y|X}) \triangleq \eta(\{P,Q\})$ satisfies*

$$\frac{H^2(P,Q)}{2} \leq \eta(\{P,Q\}) \leq H^2(P,Q) - \frac{H^4(P,Q)}{2}\,, \tag{85}$$

where Hellinger distance is defined as $H^2(P,Q) \triangleq 2 - 2\int \sqrt{dPdQ}$.

Remark 8. An obvious upper bound is $\eta(\{P,Q\}) \leq d_{\text{TV}}(P,Q)$ by Theorem 1, which is worse than Theorem 21 since d_{TV} is small than the square-root of the right-hand side of (85). In fact it is straightforward to verify that the upper bound holds with equality when the output Y is also binary-valued. In particular, Theorem 21 implies that $\eta(\{P,Q\})$ is always within *a factor of two* of $H^2(P,Q)$.

Proof. First notice the identities:

$$\chi^2(\text{Bern}(\alpha) \| \text{Bern}(\beta)) = \frac{(\alpha-\beta)^2}{\beta\bar{\beta}},$$
$$\chi^2(\alpha P + \bar{\alpha} Q \| \beta P + \bar{\beta} Q) = (\alpha - \beta)^2 \int \frac{(P-Q)^2}{\beta P + \bar{\beta} Q}\,,$$

where we denote $\bar{\alpha} = 1 - \alpha$. Therefore the (input-dependent) χ^2-contraction coefficient is given by

$$\eta_{\chi^2}(\text{Bern}(\beta), P_{Y|X}) = \sup_{\alpha \neq \beta} \frac{\chi^2(\alpha P + \bar{\alpha} Q \| \beta P + \bar{\beta} Q)}{\chi^2(\text{Bern}(\alpha) \| \text{Bern}(\beta))} = \beta\bar{\beta} \int \frac{(P-Q)^2}{\beta P + \bar{\beta} Q} \triangleq \text{LC}_\beta(P\|Q),$$

where $\mathrm{LC}_\beta(P\|Q)$, clearly an f-divergence, is known as the Le Cam divergence (see, e.g., [Vaj09, p. 889]). In view of Theorem 3, the input-independent KL-contraction coefficient coincides with that of χ^2 and hence

$$\eta(\{P,Q\}) = \sup_{\beta\in(0,1)} \mathrm{LC}_\beta(P\|Q).$$

Thus the desired bound (85) follows from the characterization of the joint range between pairs of f-divergence [HV11], namely, H^2 versus LC_β, by taking the convex hull of their joint range restricted to Bernoulli distributions. Instead of invoking this general result, next we prove (85) using elementary arguments. Since $\mathrm{LC}_{1/2}(P\|Q) = 1-2\int\frac{dPdQ}{dP+dQ} \geq 1-\int\sqrt{dPdQ} = \frac{1}{2}H^2(P,Q)$, the left inequality of (85) follows immediately. To prove the right inequality, by Cauchy-Schwartz, note that we have $(1-\frac{1}{2}H^2(P,Q))^2 = (\int\sqrt{dPdQ})^2 = (\int\sqrt{\beta dP+\bar\beta dQ}\sqrt{\frac{dPdQ}{\beta dP+\bar\beta dQ}})^2 \leq \int\frac{dPdQ}{\beta dP+\bar\beta dQ} = 1-\mathrm{LC}_\beta(P\|Q)$, for any $\beta\in(0,1)$. ☐

C Simultaneously maximal couplings

Lemma 22. *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish spaces. Given any pair of Borel probability measures P_{XY}, Q_{XY} on $\mathcal{X}\times\mathcal{Y}$, there exists a coupling π of P_{XY} and Q_{XY}, namely, a joint distribution of (X,Y,X',Y') such that $\mathcal{L}(X,Y)=P_{XY}$ and $\mathcal{L}(X',Y')=Q_{XY}$ under π, such that*

$$\pi\{(X,Y)\neq(X',Y')\} = d_{\mathrm{TV}}(P_{XY},Q_{XY}) \quad \text{and} \quad \pi\{X\neq X'\} = d_{\mathrm{TV}}(P_X,Q_X). \tag{86}$$

Remark 9. After submitting this manuscript, we were informed that this result is the main content of [Gol79]. For interested reader we keep our original proof which is different from [Gol79] by relying on Kantorovich's dual representation and, thus, is non-constructive.

Remark 10. A triply-optimal coupling achieving in addition to (86) also $\pi[Y\neq Y'] = d_{\mathrm{TV}}(P_Y,Q_Y)$ need not exist. Indeed, consider the example where X,Y are $\{0,1\}$-valued and

$$P_{XY} = \begin{pmatrix}\frac{1}{2} & 0\\ 0 & \frac{1}{2}\end{pmatrix}, \quad Q_{XY} = \begin{pmatrix}0 & \frac{1}{2}\\ \frac{1}{2} & 0\end{pmatrix}.$$

In other words, $X,Y\sim\mathrm{Bern}(1/2)$ under both P and Q; however, $X=Y$ under P and $X=1-Y$ under Q. Furthermore, since $d_{\mathrm{TV}}(P_X,Q_X)=d_{\mathrm{TV}}(P_Y,Q_Y)=0$, under any coupling $\pi_{XYX'Y'}$ of P_{XY} and Q_{XY} that simultaneously couples P_X to Q_X and P_Y to Q_Y maximally, we have $X=X'$ and $Y=Y'$, which contradicts $X=Y$ and $X'=1-Y'$. On the other hand, it is clear that a doubly-optimal coupling (as claimed by Lemma 22) exists: just take $X=X'=Y\sim\mathrm{Bern}(1/2)$ and $Y'=1-X'$. It is not hard to show that such a coupling also attains the minimum

$$\min_\pi \pi[(X,Y)\neq(X',Y')] + \pi[X\neq X'] + \pi[Y\neq Y'] = 2.$$

Proof. Define the cost function $c(x,y,x',y') \triangleq \mathbf{1}_{\{(x,y)\neq(x',y')\}} + \mathbf{1}_{\{x\neq x'\}} = 2\mathbf{1}_{\{x\neq x'\}} + \mathbf{1}_{\{x=x',y\neq y'\}}$. Since the indicator of any open set is lower semicontinuous, so is $(x,y,x',y')\mapsto c(x,y,x',y')$. Applying Kantorovich's duality theorem (see, e.g., [Vil03, Theorem 1.3]), we have

$$\min_{\pi\in\Pi(P_{XY},Q_{XY})} \mathbb{E}_\pi c(X,Y,X',Y') = \max_{f,g} \mathbb{E}_P[f(X,Y)] - \mathbb{E}_Q[g(X,Y)]. \tag{87}$$

where $f\in L_1(P), g\in L_1(Q)$ and

$$f(x,y) - g(x',y') \leq c(x,y,x',y'). \tag{88}$$

Since the cost function is bounded, namely, c takes values in $[0, 2]$, applying [Vil03, Remark 1.3], we conclude that it suffices to consider $0 \le f, g \le 2$. Note that constraint (88) is equivalent to

$$\begin{aligned} f(x,y) - g(x',y') &\le 2, \forall x \neq x', \forall y \neq y' \\ f(x,y) - g(x,y') &\le 1, \forall x, \forall y \neq y' \\ f(x,y) - g(x,y) &\le 0, \forall x, \forall y \end{aligned}$$

where the first condition is redundant given the range of f, g. In summary, the maximum on the right-hand side of (87) can be taken over all f, g satisfying the following constraints:

$$\begin{aligned} 0 \le f, g &\le 2 \\ f(x,y) - g(x,y') &\le 1, \forall x, y \neq y' \\ f(x,y) - g(x,y) &\le 0, \forall x, y \end{aligned}$$

Then

$$\max_{f,g} \mathbb{E}_P[f(X,Y)] - \mathbb{E}_Q[g(X,Y)] = \int_{\mathcal{X}} \max_{\phi,\psi} \left\{ \int_{\mathcal{Y}} p(x,y)\phi(y) - q(x,y)\psi(y) \right\} \tag{89}$$

where the maximum on the right-hand side is over $\phi, \psi : \mathcal{Y} \to \mathbb{R}$ satisfying

$$\begin{aligned} 0 \le \phi, \psi &\le 2 \\ \phi(y) - \psi(y') &\le 1, \forall y \neq y' \\ \phi(y) - \psi(y) &\le 0, \forall y \end{aligned} \tag{90}$$

The optimization problem in the bracket on the RHS of (89) can be solved using the following lemma:

Lemma 23. *Let $p, q \ge 0$. Let $(x)_+ \triangleq \max\{x, 0\}$. Then*

$$\max_{\phi,\psi} \left\{ \int_{\mathcal{Y}} p\phi - q\psi : 0 \le \phi \le \psi \le 2, \sup \phi \le 1 + \inf \psi \right\} = \int (p-q)_+ + \left(\int (p-q) \right)_+. \tag{91}$$

Proof. First we show that it suffices to consider $\phi = \psi$. Given any feasible pair (ϕ, ψ), set $\phi' = \max\{\phi, \inf \psi\}$. To check that (ϕ', ϕ') is a feasible pair, note that clearly ϕ' takes values in $[0, 2]$. Furthermore, $\sup \phi' \le \sup \phi \le 1 + \inf \psi \le 1 + \inf \phi'$. Therefore the maximum on the left-hand side of (91) is equal to

$$\max_{\phi} \left\{ \int_{\mathcal{Y}} (p-q)\phi : 0 \le \phi \le 2, \sup \phi \le 1 + \inf \phi \right\}.$$

Let $a = \inf \phi$. Then

$$\begin{aligned} \max_{\phi} \left\{ \int (p-q)\phi : 0 \le \phi \le 2, \sup \phi \le 1 + \inf \phi \right\} &= \sup_{0 \le a \le 2} \max_{\phi} \left\{ \int (p-q)\phi : a \le \phi \le 2 \wedge (1+a) \right\} \\ &= \sup_{0 \le a \le 1} \max_{\phi} \left\{ \int (p-q)\phi : a \le \phi \le 1+a \right\} \\ &= \sup_{0 \le a \le 1} \left\{ (1+a) \int (p-q)_+ + a \int (p-q)_- \right\} \\ &= \sup_{0 \le a \le 1} \left\{ \int (p-q)_+ + a \int (p-q) \right\} \\ &= \int (p-q)_+ + \left(\int (p-q) \right)_+. \end{aligned}$$

□

Applying Lemma 23 to (89) for fixed x, we have

$$\begin{aligned}
&\max_{f,g} \mathbb{E}_P[f(X,Y)] - \mathbb{E}_Q[g(X,Y)] \\
&= \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} (p(x,y) - q(x,y))_+ + (p(x) - q(x))_+ \right) \\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}} (p(x,y) - q(x,y))_+ + \int_{\mathcal{X}} (p(x) - q(x))_+ = d_{\mathrm{TV}}(P_{XY}, Q_{XY}) + d_{\mathrm{TV}}(P_X, Q_X)
\end{aligned}$$

Combining the above with (87), we have

$$\min_{\pi_{XYX'Y'}} \pi\{(X,Y) \neq (X',Y')\} + \pi\{X \neq X'\} = d_{\mathrm{TV}}(P_{XY}, Q_{XY}) + d_{\mathrm{TV}}(P_X, Q_X).$$

Since $\pi\{(X,Y) \neq (X',Y')\} \geq d_{\mathrm{TV}}(P_{XY}, Q_{XY})$ and $\pi\{X \neq X'\} \geq d_{\mathrm{TV}}(P_X, Q_X)$ for any π, the minimizer of the sum on the left-hand side achieves equality simultaneously for both terms, proving the theorem. □

References

[AG76] R. Ahlswede and P. Gács. Spreading of sets in product spaces and hypercontraction of the Markov operator. *Ann. Probab.*, pages 925–939, 1976.

[AGKN13] Venkat Anantharam, Amin Gohari, Sudeep Kamath, and Chandra Nair. On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover. *arXiv preprint arXiv:1304.6133*, 2013.

[Ash65] Robert B. Ash. *Information Theory*. Dover Publications Inc., New York, NY, 1965.

[Bir57] G. Birkhoff. Extensions of Jentzsch's theorem. *Trans. of AMS*, 85:219–227, 1957.

[BK98] Xavier Boyen and Daphne Koller. Tractable inference for complex stochastic processes. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence—UAI 1998*, pages 33–42. San Francisco: Morgan Kaufmann, 1998. Available at http://www.cs.stanford.edu/~xb/uai98/.

[CIR+93] J.E. Cohen, Yoh Iwasa, Gh. Rautu, M.B. Ruskai, E. Seneta, and Gh. Zbaganu. Relative entropy under mappings by stochastic matrices. *Linear algebra and its applications*, 179:211–235, 1993.

[CK81] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic, New York, 1981.

[CKZ98] J. E. Cohen, J. H. B. Kempermann, and Gh. Zbăganu. *Comparisons of Stochastic Matrices with Applications in Information Theory, Statistics, Economics and Population*. Springer, 1998.

[Cou12] T. Courtade. *Two Problems in Multiterminal Information Theory*. PhD thesis, U. of California, Los Angeles, CA, 2012.

[CPW15] F. Calmon, Y. Polyanskiy, and Y. Wu. Strong data processing inequalities for input-constrained additive noise channels. *arXiv*, December 2015. arXiv:1512.06429.

[CRS94] M. Choi, M.B. Ruskai, and E. Seneta. Equivalence of certain entropy contraction coefficients. *Linear algebra and its applications*, 208:29–36, 1994.

[Csi67] I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Sci. Math. Hungar.*, 2:229–318, 1967.

[Daw75] DA Dawson. Information flow in graphs. *Stoch. Proc. Appl.*, 3(2):137–151, 1975.

[DJW13] John C Duchi, Michael Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 429–438. IEEE, 2013.

[DMLM03] P. Del Moral, M. Ledoux, and L. Miclo. On contraction properties of Markov kernels. *Probab. Theory Relat. Fields*, 126:395–420, 2003.

[Dob56] R. L. Dobrushin. Central limit theorem for nonstationary Markov chains. I. *Theory Probab. Appl.*, 1(1):65–80, 1956.

[Dob70] R. L. Dobrushin. Definition of random variables by conditional distributions. *Theor. Probability Appl.*, 15(3):469–497, 1970.

[Doe37] Wolfgang Doeblin. *Le cas discontinu des probabilités en chaîne.* na, 1937.

[EC98] Elza Erkip and Thomas M. Cover. The efficiency of investment information. *IEEE Trans. Inf. Theory*, 44(3):1026–1040, 1998.

[EGK11] Abbas El Gamal and Young-Han Kim. *Network information theory.* Cambridge university press, 2011.

[EKPS00] William Evans, Claire Kenyon, Yuval Peres, and Leonard J Schulman. Broadcasting on trees and the Ising model. *Ann. Appl. Probab.*, 10(2):410–433, 2000.

[ES99] William S Evans and Leonard J Schulman. Signal propagation and noisy circuits. *IEEE Trans. Inf. Theory*, 45(7):2367–2373, 1999.

[Gol79] Sheldon Goldstein. Maximal coupling. *Probability Theory and Related Fields*, 46(2):193–204, 1979.

[HV11] P. Harremoës and I. Vajda. On pairs of f-divergences and their joint range. *IEEE Trans. Inf. Theory*, 57(6):3230–3235, Jun. 2011.

[Lau96] Steffen L Lauritzen. *Graphical Models.* Oxford University Press, 1996.

[LCV15] Jingbo Liu, Paul Cuff, and Sergio Verdu. Secret key generation with one communicator and a zero-rate one-shot via hypercontractivity. *arXiv preprint arXiv:1504.05526*, 2015.

[Led99] M. Ledoux. Concentration of measure and logarithmic Sobolev inequalities. *Seminaire de probabilites XXXIII*, pages 120–216, 1999.

[Mar06] Andrey Andreyevich Markov. Extension of the law of large numbers to dependent quantities. *Izv. Fiz.-Matem. Obsch. Kazan Univ.(2nd Ser)*, 15:135–156, 1906.

[MS77] Florence Jessie MacWilliams and Neil James Alexander Sloane. *The theory of error correcting codes.* Elsevier, 1977.

[MZ15] Anuran Makur and Lizhong Zheng. Bounds between contraction coefficients. *arXiv preprint arXiv:1510.01844*, 2015.

[Nai14] C. Nair. Equivalent formulations of hypercontractivity using information measures. In *Proc. 2014 Zurich Seminar on Comm.*, 2014.

[Ord16] Or Ordentlich. Novel lower bounds on the entropy rate of binary hidden Markov processes. In *Proc. 2016 IEEE Int. Symp. Inf. Theory (ISIT)*, Barcelona, Spain, July 2016.

[PW15] Y. Polyanskiy and Y. Wu. Lecture notes on information theory. 2016. http://people.lids.mit.edu/yp/homepage/data/itlectures_v4.pdf

[PW16] Yury Polyanskiy and Yihong Wu. Dissipation of information in channels with input constraints. *IEEE Trans. Inf. Theory*, 62(1):35–55, January 2016. also arXiv:1405.3629.

[Rag13] Maxim Raginsky. Logarithmic Sobolev inequalities and strong data processing theorems for discrete channels. In *2013 IEEE International Symposium on Information Theory Proceedings (ISIT)*, pages 419–423, 2013.

[Rag14] Maxim Raginsky. Strong data processing inequalities and ϕ-sobolev inequalities for discrete channels. *arXiv preprint arXiv:1411.3575*, November 2014.

[Sam15] Alex Samorodnitsky. On the entropy of a noisy function. *arXiv preprint arXiv:1508.01464*, August 2015.

[Sar58] O. V. Sarmanov. Maximal correlation coefficient (non-symmetric case). *Dokl. Akad. Nauk SSSR*, 121(1):52–55, 1958.

[Vaj09] I. Vajda. On metric divergences of probability measures. *Kybernetika*, 45(6):885–900, 2009.

[Vil03] C. Villani. *Topics in optimal transportation.* American Mathematical Society, Providence, RI, 2003.

[XR15] Aolin Xu and Maxim Raginsky. Converses for distributed estimation via strong data processing inequalities. In *Proc. 2015 IEEE Int. Symp. Inf. Theory (ISIT)*, Hong Kong, CN, July 2015.

[ZY97] Zhen Zhang and Raymond W Yeung. A non-Shannon-type conditional inequality of information quantities. *IEEE Trans. Inf. Theory*, 43(6):1982–1986, 1997.