Strong data-processing inequalities for channels and Bayesian networks

Yury Polyanskiy∗ and Yihong Wu†
MIT, UIUC

Abstract

The data-processing inequality, that is, $I(U;Y) \leq I(U;X)$ for a Markov chain $U \rightarrow X \rightarrow Y$, 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 does one 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 Ahlswede-Gács type contraction coefficients (mutual information), Dobrushin’s contraction coefficients (total variation), and finally the $F_I$-curve (the best possible SDPI for a given channel). 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 SDPI and comparison to erasure channel (in the sense of “less noisy” order). This leads to a simple proof of a curious inequality of Samorodnitsky (2015).

∗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.
1 Introduction

The observation that multiplying a componentwise non-negative vector by a stochastic matrix results in a vector that is “more uniform” dates back to Birkhoff [Bir57]. We will not discuss here this type of distance-contraction (sometimes known as geometric ergodicity). Instead, we will be interested in contraction of statistical distances and information measures involving probability distributions.

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.

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

\[
d_{TV}(P,Q) \triangleq \sup_{E} |P[E] - Q[E]| = \frac{1}{2} \int |dP - dQ|, \tag{1}
\]

\[
D(P\|Q) \triangleq \int \log \frac{dP}{dQ} dP, \tag{2}
\]

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

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:

\[
d_{TV}(P_{Y|X} \circ P, P_{Y|X} \circ Q) < d_{TV}(P,Q). \tag{1}
\]

2. Divergence contraction:

\[
D(P_{Y|X} \circ P \| P_{Y|X} \circ Q) < D(P\|Q). \tag{2}
\]

3. Information loss: For any Markov chain\(^1\) \( U \to X \to Y \) we have

\[
I(U;Y) < I(U;X). \tag{3}
\]

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.

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 B 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].

\(^1\)The notation \( A \to B \to C \) simply means that \( A \perp \perp C|B \).
In Section 3 we present the essence of the Evans-Schulman method and derive upper bounds
on the mutual information contraction coefficient \( \eta_{\text{KL}} \) for Bayesian networks (directed graphical
models). Then in Section 4 we derive analogous estimates for Dobrushin’s coefficient \( \eta_{\text{TV}} \) 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 \( P^n_{Y|X} = P^n_{Y|X^n} \) 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

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

and denoted by \( \text{BSC}(\delta)^n \).

## 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].
For example, the total variation (1) and the KL divergence (2) correspond to \( f(x) = |x - 1| \) and \( f(x) = x \log x \) respectively;
taking \( f(x) = (x - 1)^2 \) we obtain the \( \chi^2 \)-divergence: \( \chi^2(P||Q) \triangleq \int \frac{dP}{dQ}^2 dQ - 1 \).

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

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

\[
\eta_f(P_{Y|X}) \triangleq \sup_Q \eta_f(Q).
\]

It is easy to show that the supremum is over a non-empty set whenever \( Q \) is not a point mass
(see Appendix B). For notational simplicity when the channel is clear from context we abbreviate
\( \eta_f(P_{Y|X}) \) as \( \eta_f \). For contraction coefficients of total variation, \( \chi^2 \) and KL divergence, we write
\( \eta_{\text{TV}}, \eta_{\chi^2} \) and \( \eta_{\text{KL}} \), 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, \( \eta_{\text{TV}}(P_{Y|X}) \) is known as the Dobrushin’s coefficient of the kernel \( P_{Y|X} \). Dobrushin [Dob56]
showed that the supremum in the definition of \( \eta_{\text{TV}} \) can be restricted to point masses, namely,

\[
\eta_{\text{TV}}(P_{Y|X}) = \sup_{x,x'} \text{d}_{\text{TV}}(P_{Y|X=x}, P_{Y|X=x'}),
\]

\[\text{More generally, } D_f(P||Q) \triangleq \mathbb{E}_Q[f(\frac{dP}{dQ})], \text{ where } \mu \text{ is a dominating probability measure of } P \text{ and } Q, \text{ e.g., } \]

\[
\mu = (P + Q)/2, \text{ with the understanding that } f(0) = f(0+), \text{ } 0f(\frac{a}{2}) = 0 \text{ and } 0f(\frac{a}{2}) = \lim_{x \to 0} xf(\frac{a}{x}) \text{ for } a > 0.
\]
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

$$\eta_f(P_{Y|X}) \leq \eta_{TV}(P_{Y|X}). \quad (8)$$

For the opposite direction, lower bounds on $\eta_f$ typically involves $\eta_{\chi^2}$, the contraction coefficient of the $\chi^2$-divergence. It is well-known, e.g. Sarmanov [Sar58], that $\eta_{\chi^2}(P_{Y|X}, P_X)$ is the squared maximal correlation coefficient of the joint distribution $P_{XY}$:

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

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 $\eta_{KL}$ and $\eta_{\chi^2}$ on finite alphabets has been systematically studied by Ahlswede and Gács [AG76]. In particular, [AG76] proved

$$\eta_{\chi^2}(P_{Y|X}, P_X) \leq \eta_{KL}(P_{Y|X}, P_X), \quad (10)$$

and noticed that the inequality is frequently strict.\(^3\) Furthermore, for finite alphabets, the following equivalence is demonstrated in [AG76]:

$$\eta_{\chi^2}(P_X, P_{Y|X}) < 1 \iff \eta_{KL}(P_X, P_{Y|X}) < 1 \iff \text{graph } \{(x, y) : P_X(x) > 0, P_{Y|X}(y|x) > 0\} \text{ is connected.} \quad (11)\quad (12)$$

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

**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,

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

and

$$\eta_{\chi^2}(P_Y|X) \leq \eta_f(P_Y|X). \quad (14)$$

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

Moreover, (13) 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.**

$$\eta_{\chi^2}(P_Y|X) = \eta_{KL}(P_Y|X), \quad (15)$$

\(^3\)See [AG76, Theorem 9] and [AGKN13] for examples.
See Appendix B.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]:
\[ \eta_f(P_{Y|X}) = 1 \iff \eta_{TV}(P_{Y|X}) = 1 \]
for any non-linear \( f \).

As an illustrating example, for BSC(\( \delta \)) defined in (4), we have cf. [AG76]
\[ \eta_{K^2} = \eta_{KL} = (1 - 2\delta)^2 < \eta_{TV} = |1 - 2\delta|. \] (16)

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 \( \eta_{KL}(P_{Y|X}, P_X) \) as the maximal ratio of hyper-contractivity of the conditional expectation operator \( E[\cdot|X] \).

The fixed-input contraction coefficient \( \eta_{KL}(Q) \) is closely related to the (modified) log-Sobolev inequalities. Indeed, if \( \eta_{KL}(Q) < 1 \) 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
\[ D(P_{Y|X} \circ P||Q) \leq \eta_{KL}^2(P_{Y|X}, Q) D(P||Q), \]
where the exponent \( \eta_{KL}(P_{Y|X}, Q) \) can in turn be estimated from log-Sobolev inequalities, e.g. [Led99]. When \( Q \) is not invariant, it was shown [DMLM03] that
\[ 1 - \alpha(Q) \leq \eta_{KL}(P_{Y|X}, Q) \leq 1 - C\alpha(Q) \]
holds for some universal constant \( C \), where \( \alpha(Q) \) is a modified log-Sobolev (also known as 1-log-Sobolev) constant:
\[ \alpha(Q) = \inf_{f \perp 1, \|f\|=1} \frac{E\left[f^2(X) \log \frac{f^2(X)}{f^2(\cdot)}\right]}{E[f^2(X) \log f^2(X)]}, \quad P_{XX'} = Q \times (P_{X|Y} \circ P_{Y|X}). \]

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

There exist several other characterizations of \( \eta_{KL} \), such as the following one in terms of the contraction of mutual information (cf. [CK81, Exercise III.3.12, p. 350] for finite alphabet):
\[ \eta_{KL}(P_{Y|X}) = \sup \frac{I(U; Y)}{I(U; X)}, \]
where the supremum is over all Markov chains \( U \to X \to Y \) with fixed \( P_{Y|X} \) (or equivalently, over all joint distributions \( P_{XX'} \)) such that \( I(U; X) < \infty \). This result is an immediate consequence of the following input-dependent version (see Appendix B.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,
\[ \eta_{KL}(P_{Y|X}, P_X) = \sup \frac{I(U; Y)}{I(U; X)}, \]
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 \( \eta_{KL} \), in view of (15) and (9), is
\[ \eta_{KL}(P_{Y|X}) = \sup \rho^2(f(X), g(Y)), \]
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 $\eta_{\text{KL}} = 1$ 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

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

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

$$F^c_I(t, P_{Y|X}) \triangleq \sup_{P_{VUX}} \left\{ I(U; Y|V) : I(U; X|V) \leq t, P_{VUXY} = P_{VUX}P_{Y|X} \right\}.$$ (20)

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 $F_I'(0) = \eta_{\text{KL}}(P_{Y|X})$ 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 $F^c_I$ as it allows for some natural extension of the results about contraction coefficients. Proposition 17 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 $\eta_{\text{KL}}(P_{Y|X}) \leq \eta < 1$ and

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

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

Classically and more recently SDPIs have been increasingly popular for establishing lower (impossibility) bounds in various setups. 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 \( P_{X|Y_{−∞}} \) 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 \( \sum \eta^n < ∞ \). 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 \( \{Y_v : v \in V\} \) denoting random variables taking values in a fixed finite alphabet.\(^4\) We assume that each vertex \( Y_v \) is associated with a conditional distribution \( P_{Y_v|Y_{pa(v)}} \) where \( pa(v) \) 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 \( V \subset V \) is an arbitrary set of nodes we can progressively chain together all the random transformations and compute unequivocally \( P_{V,X} \) (here and below we use \( V \) and \( Y_V = \{Y_v : v \in V\} \) interchangeably). We assume that vertices in \( V \) 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

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

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 \( W \in V \) and \( V \subset V \) such that \( W > V \). Then

\[ \eta_{KL}(P_{V,W|X}) \leq \eta_W \cdot \eta_{KL}(P_{V,pa(W)|X}) + (1 - \eta_W) \cdot \eta_{KL}(P_{V|X}). \]  

(22)

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

**Proof.** Consider an arbitrary random variable \( U \) such that

\[ U \rightarrow X \rightarrow (V,W). \]

\(^4\)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.
Let \( A = \text{pa}(W) \setminus V \). 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:

\[
\begin{align*}
U & \rightarrow X \rightarrow V \\
& \downarrow \quad \downarrow \\
& A \rightarrow W
\end{align*}
\]

From the characterization (17) it is sufficient to show

\[
I(U;V,W) \leq (1 - \eta_W)I(U;V) + \eta_W I(U;V,A). \tag{23}
\]

Denote \( B = V \setminus \text{pa}(W) \) and \( C = V \cap \text{pa}(W) \). Then \( \text{pa}(W) = (A,C) \) and \( V = (B,C) \). To verify (23) notice that by assumption we have

\[
U \rightarrow X \rightarrow (V,A) \rightarrow W.
\]

Therefore conditioned on \( V \) we have the Markov chain

\[
U \rightarrow X \rightarrow A \rightarrow W \quad |V
\]

and the channel \( A \rightarrow W \) is a restriction of the original \( P_{W|\text{pa}(W)} \) to a subset of the inputs. Indeed, \( P_{W|A,V} = P_{W|\text{pa}(W),B} = P_{W|\text{pa}(W)} \) by 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

\[
I(U;W|V = v) \leq \eta(P_{W|A,C=c})I(U;A|V = v) \leq \eta(P_{W|A,C})I(U;A|V = v), \tag{24}
\]

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

\[
\sup_c \eta(P_{W|A,C=c}) \leq \eta(P_{W|A,C}). \tag{25}
\]

Averaging both sides of (24) over \( v \sim P_V \) and using the definition \( \eta_W = \eta(P_{W|\text{pa}(W)}) = \eta(P_{W|A,C}) \), we have

\[
I(U;W|V) \leq \eta_W I(U;A|V). \tag{26}
\]

Adding \( I(U;V) \) to both sides yields (23).

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 \( f_j \)

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

which corresponds to the graphical model:

\[
W \rightarrow Y_1 \rightarrow Y_2 \rightarrow Y_3 \cdots
\]

Then

\[
\eta_{\text{KL}}(P_{Y^n|W}) \leq 1 - (1 - \eta_{\text{KL}}(P_{Y|X}))^n < n \cdot \eta_{\text{KL}}(P_{Y|X})
\]

\[
\eta_{\text{KL}}(P_{Y^n|W}) \leq 1 - (1 - \eta_{\text{KL}}(P_{Y|X}))^n < n \cdot \eta_{\text{KL}}(P_{Y|X})
\]
Proof. Apply Theorem 5 $n$ times.

Let us call a path $\pi = \{X, v, \cdots, v\}$ with $v \in V$ to be shortcut-free from $X$ to $V$, denoted $X \xrightarrow{s.f.} V$, if there does not exist another path $\pi'$ from $X$ to any node in $V$ such that $\pi'$ is a subset of $\pi$. (In particular $v$ necessarily is the first node in $V$ that $\pi$ 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

$$\eta_{KL}(P_{V|X}) \leq \sum_{\pi: X \xrightarrow{s.f.} V} \eta^\pi.$$  

(27)

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

$$\eta_{KL}(P_{V|X}) \leq \sum_{\pi: X \rightarrow V} \eta^\pi.$$  

(28)

**Proof.** First notice the following two self-evident observations:

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

$$\sum_{\pi: X \xrightarrow{s.f.} A \cup B} \eta^\pi = \sum_{\pi: X \xrightarrow{s.f.} A, \text{avoid } B} \eta^\pi + \sum_{\pi: X \xrightarrow{s.f.} B, \text{avoid } A} \eta^\pi.$$  

(29)

2. Let $\pi : X \rightarrow V$ and $\pi_1$ be $\pi$ without the last node, then

$$\pi : X \xrightarrow{s.f.} V \iff \pi_1 : X \xrightarrow{s.f.} \{\text{pa}(V) \setminus V\}.$$  

(30)

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

$$\eta_{KL}(P_{V'|X}) \leq \sum_{\pi: X \xrightarrow{s.f.} V'} \eta^\pi.$$  

(31)

$$\eta_{KL}(P_{V', P|X}) \leq \sum_{\pi: X \xrightarrow{s.f.} (V', P)} \eta^\pi.$$  

(32)

By (29) and (30) we have

$$\sum_{\pi: X \xrightarrow{s.f.} V} \eta^\pi = \sum_{\pi: X \xrightarrow{s.f.} V'} \eta^\pi + \sum_{\pi: X \xrightarrow{s.f.} W, \text{avoid } V'} \eta^\pi$$

$$= \sum_{\pi: X \xrightarrow{s.f.} V'} \eta^\pi + \eta_W \sum_{\pi: X \xrightarrow{s.f.} P, \text{avoid } V'} \eta^\pi$$  

(33)
Then by Theorem 5 and induction hypotheses (31)-(32) we get
\[
\eta_{KL}(P_{V|X}) \leq \eta_W \left( \sum_{\pi : X \xrightarrow{\beta} V, P} \eta^\pi + (1 - \eta_W) \sum_{\pi : X \xrightarrow{\beta} V'} \eta^\pi \right)
\]
\[
= \eta_W \left( \sum_{\pi : X \xrightarrow{\beta} P, avoid V'} \eta^\pi - \sum_{\pi : X \xrightarrow{\beta} V', pass P} \eta^\pi \right) + \sum_{\pi : X \xrightarrow{\beta} V'} \eta^\pi
\]
\[
\leq \eta_W \sum_{\pi : X \xrightarrow{\beta} P, avoid V'} \eta^\pi + \sum_{\pi : X \xrightarrow{\beta} V'} \eta^\pi
\]
where in (36) we applied (29) and split the summation over \(\pi : X \xrightarrow{\beta} V'\) into paths that avoid and pass nodes in \(P\). Comparing (34) and (37) the conclusion follows.

Both estimates (27) and (28) are compared to that of Theorem 5 in Table 1 in various graphical models.

### 3.1 Evaluation for the BSC

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

\[
\eta_{KL} \leq 1 - (4\delta(1 - \delta))^n.
\]

A simple lower bound for \(\eta_{KL}\) can be obtained by considering (17) and taking \(U \sim \text{Bern}(1/2)\) and \(U \rightarrow X\) being an \(n\)-letter repetition code, namely, \(X = (U, \ldots, U)\). Let\(^5\) \(\epsilon = \mathbb{P}[|Z| \geq n/2]\) be

\[^5\]For elements of \(F_2^n\) \(\cdot\) is the Hamming weight.
the probability of error for the maximal likelihood decoding of $U$ based on $Y$, which satisfies the Chernoff bound $\epsilon \leq (4\delta(1 - \delta))^{n/2}$. We have from Jensen’s inequality

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

and thus

$$\eta_{KL} \geq 1 - (4\delta(1 - \delta))^{\frac{n}{2} + O(\log n)}. \quad (39)$$

Comparing (38) and (39) we see that $\eta_{KL} \to 1$ exponentially fast. To get the exact exponent we need to replace (38) by the following improvement:

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

where the first inequality is from (8) and the second is from (46) below. Thus, all in all we have for BSC$(\delta)^n$ as $n \to \infty$

$$\eta_{KL}, \eta_{TV} = 1 - (4\delta(1 - \delta))^{\frac{n}{2} + O(\log n)}. \quad (40)$$

## 4 Dobrushin’s coefficients in networks

The proof of Theorem 5 relies on the characterization (17) of $\eta_{KL}$ 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,

$$\eta_{TV}(P_{V,W|X}) \leq (1 - \eta_W)\eta_{TV}(P_{V|X}) + \eta_W\eta_{TV}(P_{pa(W),V|X}), \quad (41)$$

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

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

$$d_{TV}(P_{VV'}, Q_{VV'}) \leq (1 - \eta_W)d_{TV}(P_V, Q_V) + \eta_Wd_{TV}(P_{ZV}, Q_{ZV}). \quad (42)$$

which, by the arbitrariness of $x, x'$ and in view of the characterization of $\eta$ in (7), yields the desired (41). By Lemma 20 in Appendix A, there exists a coupling of $P_{ZV}$ and $Q_{ZV}$, denoted by $\pi_{ZVZ'}$, such that

$$\pi([Z, V] \neq [Z', V']) = d_{TV}(P_{ZV}, Q_{ZV}),$$

$$\pi[V \neq V'] = d_{TV}(P_V, Q_V)$$

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 $Q_{W|Z} = P_{W|Z} = P_{W|pa(W)}$ by definition). This defines a joint distribution $\pi_{ZVZ'W'VV'}$, under which we have the Markov chain $VV' \to ZZ' \to WW'$. Then

$$\pi[W \neq W'|ZZ'] = \pi[W \neq W'|ZZ'] = d_{TV}(P_{W|pa(W) = Z}, P_{W|pa(W) = Z'}) \leq \eta_W 1_{Z \neq Z'}.$$
Therefore we have
\[
\pi[W \neq W'|V = V'] = \mathbb{E}[\pi[W \neq W'|ZZ']|V = V'] \\
\leq \eta_W \pi[Z \neq Z'|V = V'].
\]

Multiplying both sides by \(\pi[V = V']\) and then adding \(\pi[V \neq V']\), we obtain
\[
\pi[(W,V) \neq (W',V')] \leq (1 - \eta_W)\pi[V \neq V'] + \eta_W \pi[(Z,V) \neq (Z',V')] \\
= (1 - \eta_W)d_{TV}(P_V, Q_V) + \eta_W d_{TV}(P_{ZV}, Q_{ZV}),
\]
where the LHS is lower bounded by \(d_{TV}(P_{WV}, Q_{WV})\) and the equality is due to the choice of \(\pi\). This yields the desired (42), completing the proof.

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

**Corollary 9.** In the setting of Corollary 6 we have
\[
\eta_{TV}(P_{Y|W}) \leq 1 - (1 - \eta_{TV}(P_{Y|X}))^n < n \cdot \eta_{KL}(P_{Y|X}). \tag{43}
\]

**Corollary 10.** In the setting of Corollary 7 we have
\[
\eta_{TV}(P_{V|X}) \leq \sum_{\pi: X \rightarrow V} \eta_{TV}^\pi \leq \sum_{\pi: X \rightarrow V} \eta_{TV}^\pi,
\]
where for any path \(\pi = (X, v_1, \ldots, v_m)\) we denoted \(\eta_{TV}^\pi \triangleq \prod_{j=1}^{m} \eta_{TV}(P_{v_j|pa(v_j)})\).

### 4.1 Evaluation for the BSC

Consider the \(n\)-letter BSC defined in (4), where \(Y = X + Z\) with \(Z \sim \text{Bern}(\delta)\) and \(|Z| \sim \text{Binom}(n, \delta)\). By Dobrushin's characterization (7), we have
\[
\eta_{TV} = \max_{x,x'} d_{TV}(P_{Y|X=x}, P_{Y|X=x'}) \\
= d_{TV}(\text{Bern}(\delta)^n, \text{Bern}(1 - \delta)^n) \\
= d_{TV}(\text{Binom}(n, \delta), \text{Binom}(n, 1 - \delta)) \tag{44} \\
= 1 - 2\mathbb{P}[|Z| > n/2] - \mathbb{P}[|Z| = n/2] \tag{45} \\
= 1 - (4\delta(1 - \delta))^{\frac{n}{2}} + O(\log n), \tag{46}
\]
where (44) follows from the sufficiency of \(|Z|\) for testing the two distributions, (45) follows from \(d_{TV}(P, Q) = 1 - \int P \wedge Q\) and (46) 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:
\[
\eta_{TV} \leq 1 - (2\delta)^n. \tag{47}
\]
Although (47) correctly predicts the exponential convergence of \(\eta_{TV} \to 1\) 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 $P_{v|\text{pa}(v)}$ by $F_{v|\text{pa}(v)}$, abbreviated by $F_v$, and the concavified version by $F_v^c$.

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

$$F_{V|W|X} \leq F_{V|X} + F_W^c \circ (F_{\text{pa}(W),V|X} - F_{V|X}),$$  \hspace{1cm} (48)

$$F_{V|W|X}^c \leq F_{V|X}^c + F_W^c \circ (F_{\text{pa}(W),V|X}^c - F_{V|X}^c).$$  \hspace{1cm} (49)

Furthermore, the right-hand side of (49) is non-negative, concave, nondecreasing and upper bounded by the identity mapping $\text{id}$.

**Remark 1.** The $F_I$-curve estimate in Theorem 11 implies that of contraction coefficients of Theorem 5. To see this, note that since $F_{\text{pa}(W),V|X} \leq \text{id}$, the following is a relaxation of (48):

$$\text{id} - F_{V|W|X} \geq (\text{id} - F_W) \circ (\text{id} - F_{V|X}).$$  \hspace{1cm} (50)

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., $F_v < \text{id}$ for all $v \in V$, then the channel $X \rightarrow V$ also has a strict contractive property, i.e., $F_{V|X} < \text{id}$.

Furthermore, since $F_W^c(t) \leq \eta_W t$, noting the fact that $F_{V|X}'(0) = \eta_{\text{KL}}(P_{V|X})$ and taking the derivative on both sides of (48) 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

$$t_2 - F_Y|X(t_2) \geq t_2 - \frac{t_2}{t_1} F_Y|X(t_1)$$

$$= \frac{t_2}{t_1} (t_1 - F_Y|X(t_1))$$

$$\geq t_1 - F_Y|X(t_1),$$

where the last step follows from the fact that $F_Y|X(t) \leq t$. Similarly, for any concave function $\Phi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ s.t. $\Phi(0) = 0$ we have $\frac{\Phi(t_2)}{t_2} \leq \frac{\Phi(t_1)}{t_1}$. 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 $P_{U|X}$ be such that $I(U;V) \leq t$ and $I(U;W,V) = F_{V|W|X}(t)$. By the same argument that leads to (26) we obtain

$$I(U;W|V = v_0) \leq F_W(I(U;A|V = v_0))$$

$$\leq F_W^c(I(U;A|V = v_0)).$$

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

$$I(U;W,V) \leq F_W^c(I(U;\text{pa}(W),V) - I(U;V)) + I(U;V).$$
Therefore,
\[
F_{V,W|X}(t) \leq F_W^c(I(U; \text{pa}(W), V) - I(U; V)) + I(U; V)
\]
\[
\leq F_W^c(F_{\text{pa}(W),V|X}(t) - I(U; V)) + I(U; V)
\]
\[
= F_{\text{pa}(W),V|X}(t) - (\text{id} - F_W^c)(F_{\text{pa}(W),V|X}(t) - I(U; V))
\]
\[
\leq F_{\text{pa}(W),V|X}(t) - (\text{id} - F_W^c)(F_{V|X}(t) - F_{V|X}(t))
\]
\[
\leq F_{V|X}(t) + F_W^c(F_{\text{pa}(W),V|X}(t) - F_{V|X}(t))
\]
\[
\leq F_{V|X}(t) + F_W^c(F_{\text{pa}(W),V|X}(t) - F_{V|X}(t))
\]
(53)
where (51) and (52) follow from the facts that \( t \mapsto F_W(t) \) and \( t \mapsto t - F_W(t) \) are both nondecreasing, and (53) 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 (53) is nondecreasing and concave (this automatically implies that (53) is an upper-bound to the concavification \( F_{V|X}^c \)). To that end, denote \( t_\lambda = \lambda t_1 + (1 - \lambda)t_2 \), \( f_\lambda = F_{V|X}^c(t_\lambda) \), \( g_\lambda = F_{\text{pa}(W),V|X}(t_\lambda) \) and notice the chain
\[
f_\lambda + F_W^c(g_\lambda - f_\lambda) \geq \lambda f_1 + (1 - \lambda)f_0 + F_W^c(\lambda(g_1 - f_1) + (1 - \lambda)(g_0 - f_0))
\]
\[
\geq \lambda(f_1 + F_W^c(g_1 - f_1)) + (1 - \lambda)(f_0 + F_W^c(g_0 - f_0))
\]
where (54) is from concavity of \( F_{V|X}^c \), \( F_{\text{pa}(W),V|X}^c \) and monotonicity of \( (a, b) \mapsto a + F_W^c(b - a) \), and (55) is from concavity of \( F_W^c \).

**Corollary 12.** In the setting of Corollary 6 we have
\[
F_{Y^n|W}(t) \leq t - \psi^{(n)}(t),
\]
where \( \psi^{(1)} = \psi \), \( \psi^{(k+1)} = \psi^{(k)} \circ \psi \) and \( \psi : \mathbb{R}_+ \to \mathbb{R}_+ \) is a convex function such that
\[
F_{Y|X}(t) \leq t - \psi(t).
\]

**Proof.** The case of \( n = 1 \) follows from the assumption on \( \psi \). 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 \( t \leq \log q \) is computed in Theorem 16 below.

### 5.1 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 \( p \ast q \triangleq p(1 - q) + q(1 - p) \) for \( p, q \in [0, 1] \) denote binary convolution and define
\[
\psi(t) \triangleq t - 1 + h(\delta \ast h^{-1}(\max(1 - t, 0)))
\]
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 \( \text{BSC}(\delta) \) is given by
\[
F_I(t, \text{BSC}(\delta)) = F_I^c(t, \text{BSC}(\delta)) = t - \psi(t).
\]

Applying Corollary 12 we obtain the following bound on the \( F_I \)-curve of BSC of blocklength \( n \) (even with feedback):
Proposition 13. Let $Z_1, \ldots, Z_n \overset{i.i.d.}{\sim} \text{Bern}(\delta)$ 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

$$I(U; Y^n) \leq I(U; X^n) - \psi(n)(I(U; X^n)),$$  

where $\psi(1) = \psi, \psi(k+1) = \psi(k) \circ \psi$ and $\psi$ is defined in (56).

Remark 2. The estimate (57) 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 (57) generally even with feedback.

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

$$F^*_t(t, \text{BSC}(\delta)^n) \leq t - t(4\delta(1 - \delta))^{n+o(n)}$$

as $n \to \infty$ for any fixed $t$. A simple lower bound, for comparison purposes, can be inferred from (39) after noticing that there we have $I(U; X) = 1$, and so

$$F^*_t(1, \text{BSC}(\delta)^n) \geq 1 - (4\delta(1 - \delta))\frac{n}{2} + O(\log n),$$

This shows that the bound of Proposition 13 is order-optimal: $F(t) \to t$ exponentially fast. Exact exponent is given by (40).

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

$$\lim_{n \to \infty} \frac{1}{n} F^*_t(n\theta, \text{BSC}(\delta)^n) = \min(\theta, C),$$

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

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

and therefore we conclude that in this sense the bound (57) 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 $\delta$ 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 $\text{EC}_q(\delta)$. In the binary, we denote the binary erasure channel by $\text{BEC}(\delta) \triangleq \text{EC}_2(\delta)$. A simple calculation shows that for every $P_{UX}$ we have

$$I(U; E) = (1 - \delta)I(U; X)$$

and therefore for $\text{EC}_q(\delta)$ we have $\eta_{\text{KL}}(P_{E|X}) = 1 - \delta$ and $F^*_t(t) = \min((1 - \delta)t, \log q)$.  

16
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 $P_{Y|X} \leq_{l.n.} P'_{Y'|X}$ if for all joint distributions $P_{UX}$ we have

\[ I(U;Y) \leq I(U;Y') \, . \]

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

**Proposition 14.** For an arbitrary channel $P_{Y|X}$ we have

\[ \eta_{\text{KL}}(P_{Y|X}) \leq \eta \iff P_{Y|X} \leq_{l.n.} P_{E|X} \, , \tag{59} \]

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

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

\[ I(U;Y) \leq (1 - \delta)I(U;X) \, , \tag{60} \]

where the right-hand side is precisely $I(U;E)$ by (58).

It turns out the notion of less-noisiness tensorizes:

**Proposition 15.** If $P_{Y_1|X_1} \leq_{l.n.} P'_{Y'_1|X_1}$ and $P_{Y_2|X_2} \leq_{l.n.} P'_{Y'_2|X_2}$ then

\[ P_{Y_1|X_1} \times P_{Y_2|X_2} \leq_{l.n.} P'_{Y'_1|X_1} \times P'_{Y'_2|X_2} \, . \]

In particular,

\[ \eta_{\text{KL}}(P_{Y|X}) \leq \eta \implies P^n_{Y|X} \leq_{l.n.} P^n_{E|X} \, . \tag{61} \]

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

**Proof.** Construct a relevant joint distribution $U \rightarrow X \rightarrow (Y_1, Y'_2)$ and consider

\[ I(U;Y_1, Y_2) = I(U;Y_1) + I(U;Y_2|Y_1) \, . \tag{62} \]

Now since $U \perp Y_2|Y_2$ we have by $P_{Y_2|X_2} \leq_{l.n.} P'_{Y'_2|X_2}$

\[ I(U;Y_2|Y_1) \leq I(U;Y'_2|Y_1) \]

and putting this back into (62) we get

\[ I(U;Y_1, Y_2) \leq I(U;Y_1) + I(U;Y'_2|Y_1) = I(U;Y_1, Y'_2) \, . \]

Repeating the same argument, but conditioning on $Y'_2$ we get

\[ I(U;Y_1, Y_2) \leq I(U;Y'_1, Y'_2) \, , \]

as required. The last claim of the proposition follows from Proposition 14.

Consequently, everything that has been said in this paper about $\eta_{\text{KL}}(P_{Y|X})$ can be restated in terms of seeking to compare a given channel in the sense of the $\leq_{l.n.}$ order to an erasure channel. It seems natural, then, to consider erasure channel in somewhat greater details.
6.1 $F_1$-curve of erasure channels

**Theorem 16.** Consider the $q$-ary erasure channel of blocklength $n$ and erasure probability $\delta$. Its $F_1$-curve is bounded by

\[
F_1^c(t, EC_q(\delta)^n) \leq \mathbb{E}[\min(B \log q, t)], \quad B \sim \text{Binom}(n, 1-\delta).
\]  

The bound is tight in the following cases:

1. at $t = k \log q$ with integral $k \leq n$ if and only if an $(n, k, n-k+1)_q$ MDS code exists\(^6\)
2. for $t \leq \log q$ and $t \geq (n-1) \log q$;
3. for all $t$ when $n = 1, 2, 3$.

**Remark 3.** Introducing $B' \sim \text{Binom}(n-1, 1-\delta)$ and using the identity $\mathbb{E}[B_1 1_{B \leq a}] = n(1-\delta) \mathbb{P}[B' \leq a - 1]$, we can express the right-hand side of (63) in terms of binomial CDFs:

\[
\mathbb{E}[\min(B, x)] = x + \mathbb{P}[B' \leq \lceil x \rceil - 1](1-\delta)(n-x) - x\delta \mathbb{P}[B' \leq \lfloor x \rfloor].
\]

This implies that the upper bound (63) is piecewise-linear, increasing and concave.

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

\[
I(U; E^n) = I(U; X_S, S) = \sum_{\sigma \subset [n]} \mathbb{P}[S = \sigma] I(U; X_\sigma) \leq \sum_{\sigma \subset [n]} \mathbb{P}[S = \sigma] \min(|\sigma, \log q, t) = \mathbb{E}[\min(B \log q, t)].
\]  

From here (63) 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 $F_1(t)$. 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 (63) 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 (63) is a piecewise-linear and concave function. Thus the bound being tight for $F_1(t)$ (that is a concave-envelope of $F_1(t)$) should also be tight as a bound for $F_1(t)$. Consequently, there must exist $U \rightarrow X^n \rightarrow E^n$ such that the bound (65) is tight with $t = I(U; X^n)$. This implies that we should have

\[
I(U; X_\sigma) = \min(\sigma \log q, t)
\]  

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 (66) 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 $(n, k, n-k+1)_q$.

\(^6\)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 injective.
We also formulate some interesting observations for binary erasure channels:

**Proposition 17.** For BEC$(n, \delta)$ we have:

1. For $n \geq 3$ we have that $F_1(t)$ is not concave. More exactly, $F_1(t) < F_1^c(t)$ for $t \in (1, 2)$.

2. For arbitrary $n$ and $t \leq \log 2$ or $t \geq (n - 1) \log 2$ we have $F_1(t) = F_1^c(t) = \mathbb{E}[\min(B \log 2, t)]$ with $B$ defined in in (63).

3. For $t = 2, n = 4$ the bound (63) is not tight and $F_1^c(t) < \mathbb{E}[\min(B \log 2, t)]$.

**Proof.** First note that in Definition 1 of $F_1(t)$ the supremum is a maximum and and $U$ can be restricted to alphabet of size $|X| + 2$. So in particular, $F_1(t) = f$ 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 $H(X^3|U) = 0$. Consequently, without loss of generality we may take $U = X^n$. So for the bound to be tight we need to have distribution s.t.

$$H(X^3) = H(X_1, X_2) = H(X_2, X_3) = H(X_1, X_3) = t, \quad H(X_1) = H(X_2) = H(X_3) = 1. \quad (67)$$

Note that since it is straightforward to verify that this set of entropies satisfies Shannon inequalities (easy to check), the main result of [ZY97] shows that there exist a sequence of triples $X^3$ (over large alphabets) which attains this point. We will show 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 $\mathbb{R}_+^3$ (as a continuous image of a compact set). Thus, it is sufficient to show that (67) itself is not achievable.

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

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

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 16 we know that the bound (63) is tight for $F_1(t)$ we conclude that

$$F_1(t) < F_1^c(t), \quad \forall t \in (1, 2).$$

To show the second claim consider $U = X^n$ and $X_1 = \cdots = X_n \sim \text{Bern}(p)$ for $t \leq \log 2$. For $t \geq (n - 1) \log 2$ take $X^{n-1}$ to be iid 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 (63) must be tight.

Finally, third claim follows from Theorem 16 and the fact that there is no $[4, 2, 3]$ binary code, e.g. [MS77, Corollary 7, Chapter 11].

Putting together (61) and (63) we get the following upper bound on the concavified $F_1$-curve of $n$-letter product channels in terms of the contraction coefficient of the single-letter channel.
Corollary 18. If $\eta_{KL}(P_{Y|X}) = \eta$, then

$$ F_{\eta}(t, P_{Y|X}^n) \leq \mathbb{E}[\min(B \log q, t)], \quad B \sim \text{Binom}(n, 1 - \delta). $$

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

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

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

The following bound can be distilled from [Sam15]:

**Theorem 19 (Samorodnitsky).** Consider the Bayesian network

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

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

$$ I(U; Y^n) \leq I(U; X_S | S) = I(U; X_S, S), \quad (68) $$

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

$$ I(U; Y^n) \leq \sum_{\sigma \subset [n]} \eta^{|\sigma|} (1 - \eta)^{|n| - |\sigma|} I(U; X_\sigma) \quad (69) $$

**Proof.** Just put together characterization (59), tensorization property Proposition 15 to get $I(U; Y^n) \leq I(U; E^n)$, where $E^n$ is the output of the product of erasure channels with erasure probabilities $1 - \eta_i$. Then the calculation (64) 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 (69) says (replacing $\text{Binom}(n, \eta)$ by its mean value $n\eta$):

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

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 $\eta n$ (or smaller).

**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 (40).
A Simultaneously maximal couplings

Lemma 20. 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 $\pi$ 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 $\pi$, such that

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

(70)

Remark 5. 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 based on Kantorovich’s dual representation (though non-constructive).

Remark 6. A triply-optimal coupling achieving in addition to (70) also $\pi[Y \neq Y'] = d_{TV}(P_Y, Q_Y)$ 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, $X, Y \sim \text{Bern}(1/2)$ under both $P$ and $Q$; however, $X = Y$ under $P$ and $X = 1 - Y$ under $Q$. Furthermore, since $d_{TV}(P_X, Q_X) = d_{TV}(P_Y, Q_Y) = 0$, under any coupling $\pi_{XY \times Y'}$ 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 20) exists: just take $X = X' = Y = \text{Bern}(1/2)$ 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 $c(x,y,x',y') \triangleq 1_{\{(x,y) \neq (x',y')\}} + 1_{\{x \neq x'\}} = 21_{\{x = x', y = y'\}} + 1_{\{x = x', y \neq y'\}}$. 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

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

(71)

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

$$f(x,y) - g(x',y') \leq c(x,y,x',y').$$  

(72)

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 $0 \leq f, g \leq 2$. Note that constraint (72) is equivalent to

$$f(x,y) - g(x',y') \leq 2, \forall x \neq x', \forall y \neq y'$$  

$$f(x,y) - g(x,y') \leq 1, \forall x, \forall y \neq y'$$  

$$f(x,y) - g(x,y) \leq 0, \forall x, \forall y$$

where the first condition is redundant given the range of $f, g$. In summary, the maximum on the right-hand side of (71) can be taken over all $f, g$ satisfying the following constraints:

$$0 \leq f, g \leq 2$$  

$$f(x,y) - g(x,y') \leq 1, \forall x, y \neq y'$$  

$$f(x,y) - g(x,y) \leq 0, \forall x, y$$
Then
\[
\max_{f,g} \mathbb{E}_P[f(X,Y)] - \mathbb{E}_Q[g(X,Y)] = \int_X \max_{\phi,\psi} \left\{ \int_Y p(x,y)\phi(y) - q(x,y)\psi(y) \right\}
\] (73)

where the maximum on the right-hand side is over \( \phi, \psi : \mathcal{Y} \to \mathbb{R} \) satisfying
\[
0 \leq \phi, \psi \leq 2
\]
\[
\phi(y) - \psi(y') \leq 1, \forall y \neq y'
\]
\[
\phi(y) - \psi(y) \leq 0, \forall y
\] (74)

The optimization problem in the bracket on the RHS of (73) can be solved using the following lemma:

**Lemma 21.** Let \( p, q \geq 0 \). Let \( (x)_+ \triangleq \max\{x, 0\} \). Then
\[
\max_{\phi,\psi} \left\{ \int_Y p\phi - q\psi : 0 \leq \phi \leq \psi \leq 2, \sup \phi \leq 1 + \inf \psi \right\} = \int (p - q)_+ + \left( \int (p - q) \right)_+. \] (75)

**Proof.** First we show that it suffices to consider \( \phi = \psi \). Given any feasible pair \((\phi, \psi)\), set \( \phi' = \max\{\phi, \inf \psi\} \). To check that \((\phi', \phi')\) is a feasible pair, note that clearly \( \phi' \) takes values in \([0, 2]\). Furthermore, \( \sup \phi' \leq \sup \phi \leq 1 + \inf \psi \leq 1 + \inf \phi' \). Therefore the maximum on the left-hand side of (75) is equal to
\[
\max_{\phi} \left\{ \int_Y (p - q)\phi : 0 \leq \phi \leq 2, \sup \phi \leq 1 + \inf \phi \right\}.
\]

Let \( a = \inf \phi \). Then
\[
\max_{\phi} \left\{ \int_Y (p - q)\phi : 0 \leq \phi \leq 2, \sup \phi \leq 1 + \inf \phi \right\} = \sup_{0 \leq a \leq 2} \max_{\phi} \left\{ \int_Y (p - q)\phi : a \leq \phi \leq 2 \wedge (1 + a) \right\}
\]
\[
= \sup_{0 \leq a \leq 1} \max_{\phi} \left\{ \int_Y (p - q)\phi : a \leq \phi \leq 1 + a \right\}
\]
\[
= \sup_{0 \leq a \leq 1} \left\{ (1 + a) \int (p - q)_+ + a \int (p - q)_- \right\}
\]
\[
= \sup_{0 \leq a \leq 1} \left\{ \int (p - q)_+ + a \int (p - q) \right\}
\]
\[
= \int (p - q)_+ + \left( \int (p - q) \right)_+.
\]

\( \square \)

Applying Lemma 21 to (73) for fixed \( x \), we have
\[
\max_{f,g} \mathbb{E}_P[f(X,Y)] - \mathbb{E}_Q[g(X,Y)]
\]
\[
= \int_X \left( \int_Y (p(x,y) - q(x,y))_+ + (p(x) - q(x))_+ \right)
\]
\[
= \int_X \int_Y (p(x,y) - q(x,y))_+ + \int_X (p(x) - q(x))_+ = d_{TV}(P_{XY}, Q_{XY}) + d_{TV}(P_X, Q_X)
\]

22
We show that
\[
\min_{\pi_{XY|X'}} \pi\{(X, Y) \neq (X', Y')\} + \pi\{X \neq X'\} = d_{TV}(P_{XY}, Q_{XY}) + d_{TV}(P_X, Q_X).
\]
Since \(\pi\{(X, Y) \neq (X', Y')\} \geq d_{TV}(P_{XY}, Q_{XY})\) and \(\pi\{X \neq X'\} \geq d_{TV}(P_X, Q_X)\) for any \(\pi\), the minimizer of the sum on the left-hand side achieves equality simultaneously for both terms, proving the theorem.

\[\Box\]

### B Contraction coefficients on general spaces

#### B.1 Proof of Theorem 2

We show that
\[
\eta_{f}(P_{Y|X}, P_X) = \sup_{Q_X} \frac{D_f(Q_Y\|P_Y)}{D_f(Q_X\|P_X)} \geq \eta_{x^2}(Q_{Y|X}, P_X) = \sup_{Q_X} \chi^2(Q_Y\|P_Y), \tag{76}
\]
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 \(Q_X = \frac{1}{2}(P_X + P_{X|X \in E})\), then \(\frac{1}{2} \leq \frac{dQ_X}{dP_X} \leq \frac{1}{2}(1 + \frac{1}{P_X(E)})\) and hence \(D_f(Q_X\|P_X) < \infty\) since \(f\) is continuous. Furthermore, \(Q_X \neq P_X\) implies that \(D_f(Q_X\|P_X) \neq 0\) [Csi67]. Let
\[
\eta_{f}(P_X) = \sup I(U; Y) \tag{77}
\]
where the supremum is over all Markov chains \(U \to X \to Y\) with fixed \(P_{XY}\) such that \(0 < I(U; X) < \infty\). First note that

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 \(Q_X(A) \geq 1/2\). Let \(Q'_X = Q_{X|X \in A}\) and \(Q''_Y = P_{Y|X \circ Q'_X}\). Then \(\frac{dQ''_Y}{dQ_X(A)} \leq \frac{a}{Q_X(A)} \leq 2a\). Let \(Q''_X = \frac{1}{a}P_X + (1 - \frac{1}{a})Q''_Y\) and \(Q''_Y = P_{Y|X \circ Q''_X} = \frac{1}{a}P_Y + (1 - \frac{1}{a})Q''_Y\). Then we have
\[
\frac{1}{a} \leq \frac{dQ''_X}{dP_X} \leq 2a + \frac{1}{a}, \quad \frac{1}{a} \leq \frac{dQ''_Y}{dP_Y} \leq 2a + \frac{1}{a}. \tag{77}
\]
Note that \(\chi^2(Q''_X\|P_X) = \frac{1}{Q_X(A)} \mathbb{E}_{Q_X(A)}[\log \left(\frac{dQ''_X}{dP_X}\right)^2 1_{\{X \in A\}}] - 1\). 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 \(\chi^2\)-divergence yields \(\liminf_{a \to \infty} \chi^2(Q''_Y\|P_Y) \geq \chi^2(Q_Y\|P_Y)\). 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 (76), it suffices to show for each fixed \(a\), for any \(\delta > 0\), there exists \(\tilde{P}_X\) such that
\[
\frac{D_f(\tilde{P}_X\|P_X)}{D_f(Q''_X\|P_X)} \geq \frac{\chi^2(Q''_X\|P_X)}{\chi^2(Q''_Y\|P_Y)} - \delta.
\]
For \(0 < \epsilon < 1\), let \(\tilde{P}_X = \epsilon P_X + (1 - \epsilon)Q''_X\), which induces \(\tilde{P}_Y = P_{Y|X \circ \tilde{P}_X} = \epsilon P_Y + (1 - \epsilon)Q''_Y\). Then
\[
D_f(\tilde{P}_X\|P_X) = \mathbb{E}_{P_X}[f(1 + \epsilon \frac{dQ''_X}{dP_X}(x) - 1))].
\]
Recall from (77) 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 \(\eta_{f}(P_X) \geq \eta_{x^2}(P_X)\).
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 $D(\tilde{P}_X \| P_X) = e^2 \chi^2(Q_X \| P_X)/2 + o(\epsilon^2)$ and $D(\tilde{P}_Y \| P_Y) = e^2 \chi^2(Q_Y \| P_Y)/2 + o(\epsilon^2)$, as $\epsilon \to 0$. Therefore $rac{\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_{KL}(P_X)$. Therefore $\eta_{KL}(P_X) \geq \eta_x^2(P_X)$.

B.2 Proof of Theorem 3

We prove

$$\eta_{KL} = \eta_x^2. \quad (78)$$

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

$$D(Q \| P) = \int_0^\infty \chi^2(Q \| P^t)dt, \quad (79)$$

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

$$D(Q_Y \| P_Y) = \int_0^\infty \chi^2(Q_Y \| P^t_Y)dt$$

$$\leq \int_0^\infty \eta_x^2 \cdot \chi^2(Q_X \| P^t_X)dt = \eta_x^2 D(Q_X \| P_X),$$

where we used $P^t_Y = P_Y | X \circ P^t_X$. It remains to check (79). Note that

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

Therefore

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

Note that $t\mathbb{E}_Q \left[ \frac{dQ - dP}{dP + tdQ} \right] = -\mathbb{E}_P \left[ \frac{dQ - dP}{dP + tdQ} \right]$. Therefore $\mathbb{E}_Q \left[ \frac{dQ - dP}{dP + tdQ} \right] = \frac{1}{1 + t} \int \frac{(dQ - dP)^2}{dP + tdQ} = (1 + t)\chi^2(Q \| P^t)$, completing the proof of (79).

B.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

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

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., $U = 1_{\{X \in E\}}$ and then $I(U; X) = h(P_X(E)) \in (0, \log 2)$. The goal of this appendix is to prove (18), namely

$$\eta_{KL}(P_X) = \eta_I(P_X).$$
The inequality $\eta(P_X) \leq \eta_{KL}(P_X)$ follows trivially:

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

For the other direction, fix $Q_X$ such that $0 < D(Q_X \| P_X) < \infty$. First, consider the case where $\frac{dQ_X}{dP_X}$ is bounded, namely, $\frac{dQ_X}{dP_X} \leq a$ for some $a > 0$ $Q_X$-a.s. For any $\epsilon \leq \frac{1}{2a}$, let $U \sim \text{Bern}(\epsilon)$ 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 $U \rightarrow X \rightarrow Y$ such that $X, Y$ is distributed as the desired $P_{XY}$. Note that

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

where $\tilde{P}_Y = P_{Y|X} \circ \tilde{P}_X$. We claim that

$$D(\tilde{P}_X \| P_X) = o(\epsilon), \quad (80)$$

which, in view of the data processing inequality $D(\tilde{P}_X \| P_X) \leq D(\tilde{P}_Y \| P_Y)$, implies $I(U; Y) \rightarrow_{\epsilon \downarrow 0} \frac{D(Q_X \| P_X)}{D(Q_X \| P_X)}$ as desired. To establish (80), 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,

$$\frac{1}{\epsilon} D(\tilde{P}_X \| P_X) = \mathbb{E}_{P_X} \left[ f \left( \frac{dQ_X}{dP_X}, \epsilon \right) \right] \rightarrow \mathbb{E}_{P_X} \left[ \frac{dQ_X}{dP_X} \right] \log e = 0.$$

To drop the boundedness assumption on $\frac{dQ_X}{dP_X}$ we simply consider the conditional distribution $Q'_X \triangleq Q_{X|X \in A}$ where $A = \{x : \frac{dQ_X}{dP_X}(x) < a\}$ and $a > 0$ is sufficiently large so that $Q_X(A) > 0$. Clearly, as $a \rightarrow \infty$, we have $Q'_X \rightarrow Q_X$ and $Q'_Y \rightarrow Q_Y$ pointwise, where $Q'_Y \triangleq P_{Y|X} \circ Q'_X$. Hence the lower-semicontinuity of divergence yields

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

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

$$D(Q'_X \| P_X) = \log \frac{1}{Q_X(A)} + \frac{1}{Q_X(A)} \mathbb{E}_Q \left[ \log \frac{dQ_X}{dP_X} 1 \left\{ \frac{dQ_X}{dP_X} \leq a \right\} \right]. \quad (81)$$

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

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

completing the proof.
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/](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. Míclo. 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.

[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.*, pages 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. *Séminaire de probabilités XXXIII*, pages 120–216, 1999.

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

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

[PW15] Y. Polyanskiy and Y. Wu. Lecture notes on information theory. 2015. http://www.ifp.illinois.edu/~yihongwu/teaching/itlectures.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 $\phi$-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.
[Vil03] C. Villani. *Topics in optimal transportation*. American Mathematical Society, Providence, RI, 2003.

[XR15] Aolin Xu and Maxim Raginsky. Conveses 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.