ISIT'2013 1569719153

Logarithmic Sobolev Inequalities and Strong Data Processing Theorems for Discrete Channels

Maxim Raginsky

***Abstract*—The noisiness of a channel can be measured by comparing suitable functionals of the input and output distributions. For instance, if we fix a reference input distribution, then the worst-case ratio of output relative entropy to input relative entropy for any other input distribution is bounded by one, by the data processing theorem. However, for a fixed reference input distribution, this quantity may be strictly smaller than one, giving so-called strong data processing inequalities (SDPIs). This paper shows that the problem of determining both the best constant in an SDPI and any input distributions that achieve it can be addressed using so-called logarithmic Sobolev inequalities, which relate input relative entropy to certain measures of input-output correlation. Another contribution is a proof of equivalence between SDPIs and a limiting case of certain strong data processing inequalities for the Rényi divergence.**

I. Introduction

Let X and Y be two finite alphabets. Consider a source P over X and a channel $W : \mathsf{X} \to \mathsf{Y}$. By the well-known data processing theorem for relative entropy, $D(QW\|PW) \leq D(Q\|P)$ for any other source Q over X, where QW and PW denote the corresponding output distributions over Y. However, in many cases it is possible to show that $D(QW\|PW)$ is *strictly* smaller than $D(Q\|P)$ unless $Q \equiv P$. For example, if $\mathsf{X} = \mathsf{Y} = \{0,1\}$, $P = \mathrm{Bern}(1/2)$, and $W = \mathrm{BSC}(\varepsilon)$, then $D(QW\|PW) \leq (1-2\varepsilon)^2 D(Q\|P)$ for all other Q [1]. This motivates the following definition:

Definition 1. *A* strong data processing inequality *with constant* $\delta \in [0,1)$*, or* $\mathrm{SDPI}(\delta)$*, holds for the channel* W *at* P *if* $D(QW\|PW) \leq \delta D(Q\|P)$ *for all sources* $Q \neq P$ *on* X*.*

We are interested in the best *distribution-dependent* constant

$$\delta^*(P,W) \triangleq \sup_{Q \neq P} \frac{D(QW\|PW)}{D(Q\|P)}. \tag{1}$$

In a remarkable paper [1], Ahlswede and Gács have uncovered deep relationships between $\delta^*(P,W)$ and several other quantities, such as the Hirschfeld–Gebelein–Rényi (HGR) maximal correlation [2] and so-called *hypercontraction constants* of a certain Markov operator associated to the pair (P,W).

In this paper, we revisit the problem of characterizing $\delta^*(P,W)$ from a different perspective, namely that of *logarithmic Sobolev inequalities*. These inequalities, which are well-known in the theory of probability and Markov chains (see, e.g., [3]–[7] and references therein), quantify the "noisiness" of a Markov operator (transition kernel) by relating certain "entropy-like" functionals of the input to the rate of increase of suitable "energy-like" quantities from the input to the output. There is, in fact, a whole hierarchy of log-Sobolev inequalities, generated by different ways of measuring entropy and energy. We make two contributions. First, we construct a hierarchy of log-Sobolev inequalities for an arbitrary source-channel pair (P,W) and then show that SDPIs fit naturally into this hierarchy. This interpretation leads to several new bounds on the constant $\delta^*(P,W)$ in terms of best constants in log-Sobolev inequalities; it allows us to recover several results of [1] in a more transparent manner; and it leads to a new variational characterization of extremal distributions that achieve the supremum in (1). Second, we show that hypercontractivity of a certain Markov operator associated to the pair (P,W) [specifically, the induced backward channel from Y to X; see Section II for details] is itself equivalent to a strong form of data processing inequalities for the Rényi divergence [8]–[10], and the SDPI for the relative entropy emerges as a limiting case. This result complements recent findings of Kamath and Anantharam [11].

Notation. We denote by $\mathcal{P}(\mathsf{X})$ the set of all probability distributions on X and by $\mathcal{P}(\mathsf{Y}|\mathsf{X})$ the set of all channels (stochastic matrices) $W : \mathsf{X} \to \mathsf{Y}$. For $P \in \mathcal{P}(\mathsf{X})$, $W \in \mathcal{P}(\mathsf{Y}|\mathsf{X})$, we denote by $P \otimes W$ the joint distribution of $(X,Y) \in \mathsf{X} \times \mathsf{Y}$ with $P_X = P$ and $P_{Y|X} = W$, and by PW the induced output distribution P_Y. We denote by $\mathcal{F}_{\mathsf{X}}$ the space of all functions $\varphi : \mathsf{X} \to \mathbb{R}$. The spaces of nonnegative and strictly positive functions $\mathsf{X} \to \mathbb{R}$ are denoted by $\mathcal{F}_{\mathsf{X}}^{\geq 0}$ and $\mathcal{F}_{\mathsf{X}}^{>0}$, respectively. The *entropy functional* of a random variable (RV) $U \geq 0$ is[1]

$$\mathrm{Ent}(U) \triangleq \mathbb{E}[U \log U] - (\mathbb{E}U)\log(\mathbb{E}U)$$

with the convention $0 \log 0 = 0$. By convexity of the function $u \mapsto u \log u$, $\mathrm{Ent}(U) \geq 0$.

II. Logarithmic Sobolev inequalities

We start by introducing log-Sobolev inequalities for a source-channel pair (P,W). Due to space limitations, we omit the proofs of the results presented in this section. Without loss of generality, we assume throughout that both P and PW are strictly positive (i.e., there are no useless input or output letters). As we will now see, log-Sobolev inequalities for the pair (P,W) relate the entropy $\mathrm{Ent}(f(X))$ of an arbitrary nonnegative function of the input $X \sim P$ to some measure of correlation between $f(X)$ and the output $Y \sim PW$.

The author is with the Department of Electrical and Computer Engineering and the Coordinated Science Laboratory, University of Illinois, Urbana, IL 61801, USA. E-mail: maxim@illinois.edu.

Research supported in part by the U.S. National Science Foundation under CAREER award no. CCF-1254041.

[1] Here and in the sequel, we use natural logarithms.

For any triple (U, V, Z) of jointly distributed RVs with $U, V \in \mathbb{R}$, we define

$$\mathcal{E}(U, V|Z) \triangleq \mathbb{E}\left[\big(U - \mathbb{E}[U|Z]\big)\big(V - \mathbb{E}[V|Z]\big)\right]. \quad (2)$$

This quantity has an estimation-theoretic interpretation: since $E(U|Z) \triangleq U - \mathbb{E}[U|Z]$ is the error of a minimum mean-square error (MMSE) estimator of U given Z, and $\mathbb{E}[E(U|Z)] = 0$, $\mathcal{E}(U, V|Z)$ is the covariance between $E(U|Z)$ and $E(V|Z)$: $\mathcal{E}(U, V|Z) = \mathrm{Cov}(E(U|Z), E(V|Z))$. In particular, $\mathcal{E}(U, U|Z) = \mathrm{MMSE}(U|Z)$, the MMSE achievable in estimating U from Z. We follow [3], [6], [7] and introduce the log-Sobolev inequalities:

Definition 2. (P, W) *satisfies* log-Sobolev inequality of order $p \in \mathbb{R}^+\backslash\{0, 1\}$ with constant c, *or* $\mathrm{LSI}_p(c)$, *if*

$$\mathrm{Ent}(f^p(X)) \le \frac{cp^2}{4(p-1)}\mathcal{E}(f^{p-1}(X), f(X)|Y), \quad \forall f \in \mathcal{F}_{\mathsf{X}}^{\ge 0};$$

$\mathrm{LSI}_1(c)$ *if*

$$\mathrm{Ent}(f(X)) \le \frac{c}{4}\mathcal{E}(f(X), \log f(X)|Y), \qquad \forall f \in \mathcal{F}_{\mathsf{X}}^{>0};$$

and $\mathrm{LSI}_0(c)$ *if*

$$\mathrm{Var}(\log f) \le -\frac{c}{2}\mathcal{E}(f(X), 1/f(X)|Y), \qquad \forall f \in \mathcal{F}_{\mathsf{X}}^{>0}.$$

Another important functional inequality relates the variance of $f(X)$ to the variance of the error $E(f(X)|Y)$:

Definition 3. (P, W) *satisfies a* Poincaré inequality *with constant* $c \ge 0$ *if for all* $f \in \mathcal{F}_{\mathsf{X}}^{\ge 0}$

$$\mathrm{Var}(f(X)) \le c\,\mathcal{E}(f(X), f(X)|Y) \equiv c\,\mathrm{MMSE}(f(X)|Y). \quad (3)$$

We are interested in the tightest constants in log-Sobolev inequalities for $p \in [0, 2]$. With that in mind, we define

$$\rho_p(P, W) \triangleq \frac{p^2}{4(p-1)} \inf_{f \in \mathcal{F}_{\mathsf{X}}^{>0}} \frac{\mathcal{E}(f^{p-1}(X), f(X)|Y)}{\mathrm{Ent}(f^p(X))}$$

for $p \notin \{0, 1\}$, with the convention $\frac{0}{0} = \infty$. The constants ρ_0, ρ_1 are defined analogously. The *Poincaré constant* is

$$\lambda(P, W) \triangleq \inf_{f \in \mathcal{F}_{\mathsf{X}}} \frac{\mathcal{E}(f(X), f(X)|Y)}{\mathrm{Var}(f(X))}.$$

Using the results of Witsenhausen [2], we can establish

Theorem 1. $\lambda(P, W) = 1 - S^2(P, W)$, *where*

$$S(P, W) \triangleq \sup_{f,g} \mathbb{E}[f(X)g(Y)]$$

is the HGR maximal correlation [2]; the supremum is over all $f \in \mathcal{F}_{\mathsf{X}}, g \in \mathcal{F}_{\mathsf{Y}}$ *with zero mean and unit variance.*

We will now show that the function $p \mapsto \rho_p(P, W)$ is monotonically decreasing, with $\rho_0(P, W) = \lambda(P, W)/2$. To that end, we first express the above inequalities in an equivalent "standard" form by marginalizing over the output Y. The general scheme for log-Sobolev inequalities on a discrete set X (see, e.g., [4], [6], [7]) involves a pair (π, K) with $\pi \in \mathcal{P}(\mathsf{X})$ and $K \in \mathcal{P}(\mathsf{X}|\mathsf{X})$ satisfying the *detailed balance condition*

$$\pi(x)K(x'|x) = \pi(x')K(x|x'), \quad \forall x, x' \in \mathsf{X}. \quad (4)$$

Define the *Dirichlet form* $\mathcal{D}_{\pi,K} : \mathcal{F}_{\mathsf{X}} \times \mathcal{F}_{\mathsf{X}} \to \mathbb{R}$ by

$$\mathcal{D}_{\pi,K}(f, g) \triangleq \frac{1}{2}\mathbb{E}\left[\big(f(X) - f(X')\big)\big(g(X) - g(X')\big)\right],$$

where the expectation is w.r.t. $P_{XX'} = \pi \otimes K$. Then from (4) it follows that K leaves π invariant, $\pi = \pi K$, and that $\mathcal{D}_{\pi,K}$ is symmetric, $\mathcal{D}_{\pi,K}(f, g) = \mathcal{D}_{\pi,K}(g, f)$ for all f, g. The pair (π, K) satisfies $\mathrm{LSI}_p(c)$ for $p \in \mathbb{R}^+\backslash\{0, 1\}$ if

$$\mathrm{Ent}_\pi(f^p) \le \frac{cp^2}{4(p-1)}\mathcal{D}_{\pi,K}(f^{p-1}, f), \ \forall f \in \mathcal{F}_{\mathsf{X}}^{\ge 0}$$

where $\mathrm{Ent}_\pi(f)$ is shorthand for $\mathrm{Ent}(f(X))$ with $X \sim \pi$. The definitions for $p = 0$ or 1 are analogous [7]. Going back to our setting, we have the following lemma:

Lemma 1. *Define the* backward channel $W^\sharp \in \mathcal{P}(\mathsf{X}|\mathsf{Y})$ *by*

$$W^\sharp(x|y) \triangleq \frac{W(y|x)P(x)}{PW(y)}, \qquad \forall (x, y) \in \mathsf{X} \times \mathsf{Y}.$$

Then the pair $(P, W^\sharp W)$ *satisfies the detailed balance condition* (4).[2] *Moreover, for any two* $f, g \in \mathcal{F}_{\mathsf{X}}$,

$$\mathcal{E}(f(X), g(X)|Y) = \mathcal{D}_{P,W^\sharp W}(f, g).$$

Here, $W^\sharp W$ *is the cascade channel* $X \xrightarrow{W} Y \xrightarrow{W^\sharp} X'$.

We can now deduce the following from the results of [7]:

Theorem 2. *If* (P, W) *satisfies* $\mathrm{LSI}_p(c)$ *for some* $p \in [0, 2]$, *then it satisfies* $\mathrm{LSI}_q(c)$ *for all* $q \in [0, p]$. *Moreover, the function* $p \mapsto \rho_p(P, W)$ *is decreasing, and*

$$2\rho_p(P, W) \le 2\rho_0(P, W) = 1 - S^2(P, W), \quad \forall p \in [0, 2]$$

Example 1. Let $\mathsf{X} = \mathsf{Y} = \{0, 1\}$, $P = \mathrm{Bern}(p)$ with $p \ge 1/2$, $W = \mathrm{BSC}(\varepsilon)$. By Lemma 1 and [4, Corollary A.3],

$$\rho_2(p, \varepsilon) \triangleq \rho_2(P, W) = \frac{\varepsilon\bar{\varepsilon}(2p-1)}{2(\varepsilon p + \bar{\varepsilon}\bar{p})p(\log p - \log \bar{p})}$$
$$\rho_0(p, \varepsilon) \triangleq \rho_0(P, W) = \frac{\varepsilon\bar{\varepsilon}}{2}\left(\frac{1}{\varepsilon p + \bar{\varepsilon}\bar{p}} + \frac{1}{\bar{\varepsilon}p + \varepsilon\bar{p}}\right)$$

where for $a \in [0, 1]$ we have let $\bar{a} \triangleq 1 - a$. Since the limits of $\rho_0(p, \varepsilon)$ and $\rho_2(p, \varepsilon)$ as $p \to 1/2$ are both equal to $2\varepsilon\bar{\varepsilon}$, Theorem 2 implies that all log-Sobolev constants $\rho_p, p \in [0, 2]$, for the pair $(\mathrm{Bern}(1/2), \mathrm{BSC}(\varepsilon))$ are equal to $2\varepsilon\bar{\varepsilon} = 2\varepsilon(1-\varepsilon)$.

III. Strong data processing theorems: an approach through functional inequalities

We now turn to the problem of characterizing the SDPI constant $\delta^*(P, W)$. A nontrivial upper bound can be derived if the pair (P, W) satisfies LSI_2. In particular, the following theorem generalizes a similar result by Miclo [5], which deals with the special case $\mathsf{X} = \mathsf{Y}$ and $P = PW$:

[2] This property is key in the analysis of the Gibbs sampler [12].

Theorem 3. $\delta^*(P,W) \le 1-\rho_2(P,W)$.

Proof: We use the following delicate convexity bound for the function $\eta(t) \triangleq t\log t, t\ge 0$ [5]: for all $s,t\ge 0$

$$\eta(s) \ge \eta(t) + (1+\log t)(s-t) + \left(\sqrt{s}-\sqrt{t}\right)^2. \quad (5)$$

Let us define the functions $f = \mathrm{d}Q/\mathrm{d}P \in \mathcal{F}_{\mathsf{X}}^{\ge 0}$ and $g = \mathrm{d}(QW)/\mathrm{d}(PW) \in \mathcal{F}_{\mathsf{Y}}^{\ge 0}$ and use the bound (5) to get

$$\eta(f(x)) \ge \eta(g(y)) + (1+\log g(y))\,(f(x)-g(y)) + \left(\sqrt{f(x)}-\sqrt{g(y)}\right)^2 \quad (6)$$

Noting that $g(y) = \mathbb{E}[f(X)|Y=y]$ and taking conditional expectation $\mathbb{E}[\cdot|Y]$ of both sides of (6), we obtain

$$\begin{aligned}\mathbb{E}[\eta(f(X))|Y] &\ge \eta(\mathbb{E}[f(X)|Y]) \\ &\quad + \mathbb{E}\Big[\left(\sqrt{f(X)}-\sqrt{\mathbb{E}[f(X)|Y]}\right)^2\Big|Y\Big] \\ &\ge \eta(\mathbb{E}[f(X)|Y]) + \mathbb{E}\Big[\left(\sqrt{f(X)}-\mathbb{E}\big[\sqrt{f(X)}\big|Y\big]\right)^2\Big|Y\Big],\end{aligned}$$

where we have used the fact that $\mathbb{E}[(U-h(Y))^2|Y] \ge \mathbb{E}[(U-\mathbb{E}[U|Y])^2|Y]$. Next we take the expectation w.r.t. Y to get

$$\mathrm{Ent}(f(X)) \ge \mathrm{Ent}(\mathbb{E}[f(X)|Y]) + \mathcal{E}\Big(\sqrt{f(X)}, \sqrt{f(X)}\Big|Y\Big),$$

where we have used the fact that $\mathrm{Ent}(U) = \mathbb{E}[\eta(U)]$ for all nonnegative RVs U, as well as the definition (2) of $\mathcal{E}$. Using this and the definition of $\rho_2(P,W)$, we get $\mathrm{Ent}(\mathbb{E}[f(X)|Y]) \le (1-\rho_2)\,\mathrm{Ent}(f(X))$. The proof is concluded by noting that $D(Q\|P) = \mathrm{Ent}(f(X))$ and $D(QW\|PW) = \mathrm{Ent}(g(Y)) = \mathrm{Ent}(\mathbb{E}[f(X)|Y])$. ∎

Unfortunately, Theorem 3 is not tight. For instance, it gives

$$\delta^*(\mathrm{Bern}(1/2), \mathrm{BSC}(\varepsilon)) \le 1-2\varepsilon\bar{\varepsilon} = \frac{1+(1-2\varepsilon)^2}{2}$$

(cf. Example 1), whereas we have $\delta^*(\mathrm{Bern}(1/2), \mathrm{BSC}(\varepsilon)) = (1-2\varepsilon)^2$ [1, Theorem 9]. We now show that the SDPI constant $\delta^*(P,W)$ arises from a stronger form of LSI_1. For a pair (U,V) of jointly distributed RVs with $U\ge 0$, let us define

$$\Gamma(U|V) \triangleq \mathrm{Ent}(U) - \mathrm{Ent}(\mathbb{E}[U|V]).$$

Definition 4. *We say that a* strong log-Sobolev inequality of order 1 *with constant c, or* $\text{s-LSI}_1(c)$*, holds for (P,W) if*

$$\mathrm{Ent}(f(X)) \le \frac{c}{4}\Gamma(f(X)|Y), \qquad \forall f\in\mathcal{F}_{\mathsf{X}}^{\ge 0} \quad (7)$$

The reason for calling (7) a strong LSI_1 is stated in

Theorem 4. *The pair (P,W) satisfies* $\text{s-LSI}_1(c)$ *if and only if* $\mathrm{SDPI}(1-4/c)$ *holds for W at P. Moreover, in that case (P,W) also satisfies* $\mathrm{LSI}_p(c)$ *for all $p\in[0,1]$.*

Proof: We begin by noting that the inequality (7) holds for all $f\in\mathcal{F}_{\mathsf{X}}^{\ge 0}$ if and only if it holds for all $f\in\mathcal{F}_{\mathsf{X}}^{\ge 0}$ with $\mathbb{E}[f(X)]=1$. The set of all such f is in a one-to-one correspondence with the set $\{\mathrm{d}Q/\mathrm{d}P : Q\in\mathcal{P}(\mathsf{X})\}$. Moreover, if we fix an arbitrary $Q\in\mathcal{P}(\mathsf{X})$ and consider the function $f = \mathrm{d}Q/\mathrm{d}P \in \mathcal{F}_{\mathsf{X}}^{\ge 0}$, then $\mathrm{Ent}(f(X)) = D(Q\|P)$ and

$$\begin{aligned}\Gamma(f(X)|Y) &= \mathrm{Ent}(f(X)) - \mathrm{Ent}(\mathbb{E}[f(X)|Y]) \\ &= \mathrm{Ent}\left(\frac{\mathrm{d}Q}{\mathrm{d}P}(X)\right) - \mathrm{Ent}\left(\frac{\mathrm{d}(QW)}{\mathrm{d}(PW)}(Y)\right) \\ &= D(Q\|P) - D(QW\|PW).\end{aligned}$$

Applying (7) to this f and rearranging, we get $D(QW\|PW) \le (1-4/c)D(Q\|P)$.

To prove the second part of the theorem, we use concavity of the function $t\mapsto\log t$ to deduce that

$$\begin{aligned}&\Gamma(f(X)|Y) \\ &= \mathbb{E}[f(X)\log f(X)] - \mathbb{E}[\mathbb{E}[f(X)|Y]\log\mathbb{E}[f(X)|Y]] \\ &\le \mathbb{E}[f(X)\log f(X)] - \mathbb{E}[\mathbb{E}[f(X)|Y]\mathbb{E}[\log f(X)|Y]] \\ &= \mathcal{E}(f(X), \log f(X)|Y).\end{aligned}$$

Consequently, $\text{s-LSI}_1(c)$ implies $\mathrm{LSI}_1(c)$, which in turn implies $\mathrm{LSI}_p(c)$ for all $p\in[0,1]$ by Theorem 2. ∎

Having established the equivalence between s-LSI_1 and SDPI, we turn to the problem of characterizing the best constant $\delta^*(P,W)$ and the extremal distributions $Q\in\mathcal{P}(\mathsf{X})$ that achieve the supremum in (1). To that end, we define

$$\gamma(P,W) \triangleq \inf_{f\in\mathcal{F}_{\mathsf{X}}^{\ge 0}} \frac{\Gamma(f(X)|Y)}{4\,\mathrm{Ent}(f(X))}. \quad (8)$$

From Theorem 4, $\delta^*(P,W) = 1-4\gamma(P,W)$ and

$$\gamma(P,W) \le \rho_1(P,W) \le \ldots \le \rho_0(P,W) = \frac{1-S^2(P,W)}{2}.$$

We can get a tighter upper bound on $\gamma(P,W)$ in terms of the maximal correlation $S(P,W)$: applying the s-LSI_1 inequality (7) to functions of the form $1+\varepsilon f$ with $f\in\mathcal{F}_{\mathsf{X}}$ and small enough $\varepsilon>0$ (in which case $1+\varepsilon f\in\mathcal{F}_{\mathsf{X}}^{\ge 0}$) and using the fact that $\mathrm{Ent}(1+\varepsilon U) = (\varepsilon^2/2)\,\mathrm{Var}(U) + O(\varepsilon^3)$, we get

$$\begin{aligned}\mathrm{Var}(f(X)) &\le \frac{c}{4}\left\{\mathrm{Var}(f(X)) - \mathrm{Var}(\mathbb{E}[f(X)|Y])\right\} \\ &= \frac{c}{4}\,\mathrm{MMSE}(f(X)|Y),\end{aligned}$$

which is the Poincaré inequality with constant $c/4$. Consequently, $4\gamma(P,W) \le \lambda(P,W) = 1-S^2(P,W)$. (The same bound was derived in [1] using a more complicated argument.)

Next, we characterize the extremal functions $f\in\mathcal{F}_{\mathsf{X}}^{\ge 0}$ that attain the infimum in (8). As we shall see shortly, these functions are solutions of the variational equation

$$\mathbb{E}\big[\log\mathbb{E}[f(X)|Y]\big|X=x\big] = (1-4\gamma)\log f(x) \quad (9)$$

with $\gamma = \gamma(P,W)$ under the constraint $f\in\mathcal{F}_{\mathsf{X}}^{\ge 0}$ and $\mathbb{E}[f(X)]=1$. To account for the possibility that the only solution of (9) is the trivial one $f\equiv 1$, we follow [6] and say that $f\in\mathcal{F}_{\mathsf{X}}^{\ge 0}$ is a *generalized* solution of (9) if it satisfies (9) with $\gamma=\gamma(P,W)$ or if the equation

$$\mathbb{E}\big[\mathbb{E}[g(X)|Y]\big|X=x\big] = (1-4\gamma)g(x). \quad (10)$$

with $\gamma = \gamma(P, W)$ has a solution $g \in \mathcal{F}_{\mathsf{X}}$ with $\mathbb{E}[g(X)] = 0$ and $\mathbb{E}[g^2(X)] = 1$. This definition arises from the observation that if we apply (9) to functions of the form $1 + \varepsilon g$ with $\mathbb{E}[g(X)] = 0, \mathbb{E}[g^2(X)] = 1$ for sufficiently small $\varepsilon > 0$, we get (10) in the limit $\varepsilon \to 0$. In fact, if such a solution g to (10) exists, then $4\gamma(P, W) = S^2(P, W)$. This follows from the fact that $\lambda(P, W) = 1 - S^2(P, W)$ is the smallest nonnegative constant λ, for which the equation $\mathbb{E}\big[\mathbb{E}[g(X)|Y]\big|X = x\big] = \lambda g(x)$ has a solution $g \in \mathcal{F}_{\mathsf{X}}$ with $\mathbb{E}[g(X)] = 0$ and $\mathbb{E}[g^2(X)] = 1$ [2].

With these preliminaries out of the way, we have the following theorem, which parallels the results of Bobkov and Tetali [6] for log-Sobolev inequalities with $p \in \{1, 2\}$:

Theorem 5. *Suppose that $S^2(P, W) < 1$.*

1) *There exists a number $\gamma > 0$ such that the generalized Eq.* (9) *has a non-constant solution $f \in \mathcal{F}_{\mathsf{X}}^{\geq 0}$.*
2) *Among such numbers there is a minimal value.*
3) *This minimal value represents the optimal constant $\gamma(P, W)$ in the log-Sobolev inequality* (7)*.*

Proof: We assume that $4\gamma(P, W) < \lambda(P, W)$, for otherwise (10) holds with $\gamma = \gamma(P, W)$ and there is nothing to prove. We seek to minimize the functional

$$W(f) \triangleq \frac{\Gamma(f(X)|Y)}{4\,\mathrm{Ent}(f(X))} = \frac{\mathrm{Ent}(f(X)) - \mathrm{Ent}(\mathbb{E}[f(X)|Y])}{4\,\mathrm{Ent}(f(X))}$$

over all $f \in \mathcal{F}_{\mathsf{X}}^{\geq 0}$. Since $W(af) = W(f)$ for all positive constants a, we may assume that $\max_{x \in \mathsf{X}} f(x) = 1$, and let $\mathcal{P}$ denote the class of all such functions. Let $\{f_n\} \subset \mathcal{P}$ be a minimizing sequence, i.e., $W(f_n) \to \gamma(P, W)$ as $n \to \infty$. The same argument as in the proof of Theorem 6.2 in [6] can be used to show that this sequence converges pointwise to a nonconstant function $f : \mathsf{X} \to [0, 1]$, and that $W(f_n) \to W(f)$. So, the infimum in (8) is actually achieved at f.

So now let $f \in \mathcal{F}_{\mathsf{X}}^{\geq 0}$ attain the minimum in (8). Fix an arbitrary $g \in \mathcal{F}_{\mathsf{X}}$. For small enough $\varepsilon > 0$, the function $f + \varepsilon g$ will be nonnegative. Then by definition

$$4\gamma(P, W)\,\mathrm{Ent}(f(X) + \varepsilon g(X)) \leq \Gamma(f(X) + \varepsilon g(X)|Y). \quad (11)$$

Applying the Taylor expansion $\mathrm{Ent}(U + \varepsilon V) = \mathrm{Ent}(U) + \varepsilon\mathbb{E}[V \log U] + O(\varepsilon^2)$ first to $U = f(X), V = g(X)$ and then to $U = \mathbb{E}[f(X)|Y], V = \mathbb{E}[g(X)|Y]$ and using the fact that $\Gamma(f(X)|Y) = 4\gamma(P, W)\,\mathrm{Ent}(f(X))$, we have

$$\begin{aligned} &4\gamma(P, W)\,\mathrm{Ent}(f(X) + \varepsilon g(X)) - \Gamma(f(X) + \varepsilon g(X)|Y) \\ &= \varepsilon\Big((4\gamma(P, W) - 1)\mathbb{E}[g(X)\log f(X)] \\ &\qquad + \mathbb{E}[g(X)\mathbb{E}[\log\mathbb{E}[f(X)|Y]|X]]\Big) + o(\varepsilon) \end{aligned} \quad (12)$$

The left-hand side of (12) is nonpositive, cf. (11), while the right-hand side is nonpositive for all small $\varepsilon > 0$ iff

$$\begin{aligned} \mathbb{E}\Big[g(X)\Big((1 - 4\gamma(P, W))\log f(X) \\ - \mathbb{E}[\log\mathbb{E}[f(X)|Y]|X]\Big)\Big] = 0. \end{aligned}$$

Since g is arbitrary and $P > 0$, the minimizing function f must satisfy (9) with $\gamma = \gamma(P, W)$. Hence, the log-Sobolev constant $\gamma(P, W)$ is among the numbers $\gamma > 0$ for which (9) has a nonconstant solution $f \in \mathcal{F}_{\mathsf{X}}^{\geq 0}$ with $\mathbb{E}[f(X)] = 1$.

It remains to show minimality. To that end, let $\gamma' > 0$ be another constant such that there exists some function $f' \in \mathcal{F}_{\mathsf{X}}^{\geq 0}$ with $\mathbb{E}[f'(X)] = 1$ satisfying

$$\mathbb{E}[\log\mathbb{E}[f'(X)|Y]|X = x] = (1 - 4\gamma')\log f'(x) \quad (13)$$

Multiplying both sides of (13) by $f'(x)$ and taking expectations, we get $\mathrm{Ent}(\mathbb{E}[f'(X)|Y]) = (1 - 4\gamma')\,\mathrm{Ent}(f'(X))$. By definition of $\gamma(P, W)$, we must have $\gamma' \geq \gamma(P, W)$. ∎

The above proof shows that if $4\gamma(P, W) < 1 - S^2(P, W)$, then Eq. (9) admits a nontrivial (i.e., nonconstant) solution. The contrapositive of this statement gives:

Corollary 1. *If the infimum in* (8) *is not achieved, i.e., if the LSI* (7) *is strict unless $f \equiv 1$, then $4\gamma(P, W) = 1 - S^2(P, W)$.*

Remark 1. Equivalently, $4\gamma(P, W) = 1 - S^2(P, W)$ if for an arbitrary $\gamma > 0$ the only solution to Eq. (9) among $\mathcal{F}_{\mathsf{X}}^{\geq 0}$ with $\mathbb{E}[f(X)] = 1$ is the trivial solution $f \equiv 1$.

We can now characterize extremal distributions Q that achieve the supremum in (1). There are two cases to consider:

1) There exists a nonconstant function $f \in \mathcal{F}_{\mathsf{X}}^{\geq 0}$ with $\mathbb{E}[f(X)] = 1$ that satisfies (9) with $\gamma = \gamma(P, W)$. Define $Q \in \mathcal{P}(\mathsf{X})$ by $\mathrm{d}Q = f\mathrm{d}P$. Multiplying both sides of (9) by $f(x)$ and taking expectations, we get $D(QW\|PW) = (1 - 4\gamma(P, W))D(Q\|P)$.
2) Eq. (9) has no solution other than $f \equiv 1$. In that case, by Corollary 1, $\delta^*(P, W) = S^2(P, W)$ and $D(QW\|PW) < S^2(P, W)D(Q\|P)$ for all $Q \neq P$. This is the case for $P = \mathrm{Bern}(1/2)$ and $W = \mathrm{BSC}(\varepsilon)$.

IV. Hypercontraction of Rényi divergences

Another important result from [1] is that strong data processing inequalities for a pair (P, W) emerge as limiting cases of so-called *hypercontractivity* of the backward channel $W^\sharp$, viewed as a mapping of $\mathcal{F}_{\mathsf{X}}$ into $\mathcal{F}_{\mathsf{Y}}$ with

$$W^\sharp f(y) \triangleq \mathbb{E}[f(X)|Y = y], \quad \forall y \in \mathsf{Y}, f \in \mathcal{F}_{\mathsf{X}}.$$

A recent paper by Kamath and Anantharam [11] (see also [13]) contains several interesting new results along these lines. Here, we show that hypercontractivity can itself be interpreted as a strong data processing property of *Rényi divergences* [8], [9].

The Rényi divergence of order $\alpha \in \mathbb{R}^+\backslash\{0, 1\}$ between two probability measures $Q, P \in \mathcal{P}(\mathsf{X})$ is [8]

$$D_\alpha(Q\|P) \triangleq \frac{1}{\alpha - 1}\log\mathbb{E}_P\left[\left(\frac{\mathrm{d}Q}{\mathrm{d}P}\right)^\alpha\right] \quad (14)$$

(assuming $Q \ll P$, which is the case when $P > 0$). Some key properties of the Rényi divergence are [9], [10]:

- $D_1(Q\|P) \triangleq \lim_{\alpha\uparrow 1} D_\alpha(Q\|P) = D(Q\|P)$
- the function $\alpha \mapsto D_\alpha(Q\|P)$ is nondecreasing
- if $D(Q\|P) = \infty$ or if there exists some $\beta > 1$ such that $D_\beta(Q\|P) < \infty$, then $D_1(Q\|P) = D(Q\|P) = \lim_{\alpha\downarrow 1} D_\alpha(Q\|P) = \lim_{\alpha\to 1} D_\alpha(Q\|P)$.

- data processing inequality: $D_\alpha(QW\|PW) \le D_\alpha(Q\|P)$ for any channel W with input alphabet X

We now introduce hypercontractivity. To that end, let us define for any $\alpha \in \mathbb{R}^+\backslash\{0,1\}$ its *Hölder conjugate* α' via the relation $1/\alpha + 1/\alpha' = 1$, and let $g(\alpha,\beta) \triangleq \alpha'/\beta'$ if $\alpha \neq \beta$ and 1 otherwise. Also, we define the region $\mathbb{T} \triangleq \{(\alpha,\beta) : 1 \le \beta \le \alpha\} \cup \{(\alpha,\beta) : 1 \ge \beta \ge \alpha > 0\} \subset \mathbb{R}^2$.

Definition 5. *The* Rényi hypercontractivity constants *of* (P,W)*, for* $(\alpha,\beta) \in \mathbb{T}$*, are*

$$\delta^*_{\beta\to\alpha}(P,W) \triangleq \begin{cases} \sup\limits_{Q\neq P} \dfrac{D_\alpha(QW\|PW)}{g(\alpha,\beta)D_\beta(Q\|P)}, & \alpha,\beta \ge 1 \\ \sup\limits_{Q\neq P,\, Q>0} \dfrac{D_\alpha(QW\|PW)}{g(\alpha,\beta)D_\beta(Q\|P)}, & \alpha,\beta \le 1 \end{cases},$$

so in particular $\delta^*_{1\to 1}(P,W) \equiv \delta^*(P,W)$. *The* Rényi hypercontractivity region *of* (P,W) *is the set*

$$\mathcal{H}(P,W) \triangleq \left\{(\alpha,\beta) \in \mathbb{T} : \delta^*_{\beta\to\alpha} \le 1\right\}.$$

If $\alpha,\beta \in \mathbb{T} \cap [1,\infty)$, then $D_\alpha(QW\|PW) \le D_\alpha(Q\|P)$ for any Q by the data processing property of Rényi divergences. Suppose, however, that $\delta^*_{\beta\to\alpha} \le 1$. Then we have

$$D_\alpha(QW\|PW) \le g(\alpha,\beta)D_\beta(Q\|P) \le D_\alpha(Q\|P),$$

where the second inequality uses the fact that $g(\alpha,\beta) \le 1$ for $\alpha \ge \beta \ge 1$, as well as monotonicity of the function $\alpha \mapsto D_\alpha(Q\|P)$. Notice that $g(\alpha,\beta)D_\beta(Q\|P)$ is sandwiched between $D_\alpha(QW\|PW)$ and $D_\alpha(Q\|P)$, so the bound $D_\alpha(QW\|PW) \le g(\alpha,\beta)D_\beta(Q\|P)$ is tighter than, and implies, the DPI $D_\alpha(QW\|PW) \le D_\alpha(Q\|P)$. Moreover, we can prove the following:

Theorem 6. *The SDPI constant* $\delta^*(P,W)$ *is the infimum of all* $\sigma \in (0,1]$*, such that* $(1+\varepsilon/\sigma, 1+\varepsilon)$ *and* $(1-\varepsilon/\sigma, 1-\varepsilon)$ *are in* $\mathcal{H}(P,W)$ *for all sufficiently small* $\varepsilon > 0$*.*

Proof: The proof hinges on the equivalence between Rényi hypercontractivity of (P,W) and "ordinary" hypercontractivity of the backward channel $W^\sharp$ [1], [11], which we now introduce. For any $p > 0$ and a real-valued RV U with $\mathbb{E}|U|^p < \infty$, let $\|U\|_p \triangleq (\mathbb{E}|U|^p)^{1/p}$ (for $p \ge 1$, this is the usual L^p norm; for $p < 1$, this is a quasinorm). Let us define

$$\|W^\sharp\|_{p\to q} \triangleq \begin{cases} \sup\limits_{f\in\mathcal{F}_{\mathsf{X}}^{\ge 0}; f\neq 0} \dfrac{\|W^\sharp f(Y)\|_q}{\|f(X)\|_p}, & q \ge p \ge 1 \\ \inf\limits_{f\in\mathcal{F}_{\mathsf{X}}^{>0}} \dfrac{\|W^\sharp f(Y)\|_q}{\|f(X)\|_p}, & 0 < q \le p \le 1 \end{cases}$$

We have $\|W^\sharp\|_{p\to p} \le 1$ for $p \ge 1$ (i.e., $W^\sharp$ is a contraction for $p \ge 1$) and $\|W^\sharp\|_{p\to p} \ge 1$ for $p \le 1$ (reverse contraction for $p \le 1$). On the other hand, we say that $W^\sharp$ is *hypercontractive* at (p,q) with $q \ge p \ge 1$ if $\|W^\sharp\|_{p\to q} \le 1$, and *reverse hypercontractive* at (p,q) with $1 \ge p \ge q > 0$ if $\|W^\sharp\|_{p\to q} \ge 1$. (Hypercontractivity and reverse hypercontractivity are properties of general Markov operators [7], [11].) We claim that

$$\forall(\alpha,\beta) \in \mathbb{T}:\ \delta^*_{\beta\to\alpha} \le 1 \iff \begin{cases} \|W^\sharp\|_{\beta\to\alpha} \le 1, & \alpha,\beta \ge 1 \\ \|W^\sharp\|_{\beta\to\alpha} \ge 1, & \alpha,\beta \le 1 \end{cases}$$

This is easy to prove. Consider the case $\alpha \ge \beta \ge 1$. By homogeneity of norms, we may restrict the supremum in the definition of $\|W^\sharp\|_{\beta\to\alpha}$ to $f \in \mathcal{F}_{\mathsf{X}}^{\ge 0}$ with $\mathbb{E}[f(X)] = 1$. Any such f has the form $\mathrm{d}Q/\mathrm{d}P$ for some $Q \in \mathcal{P}(\mathsf{X})$. If $\delta^*_{\beta\to\alpha} \le 1$, then from the fact that $\mathrm{d}(QW)/\mathrm{d}(PW) = W^\sharp f$ and from (14)

$$\begin{aligned}\|W^\sharp f(Y)\|_\alpha &= \exp\left(\frac{\alpha-1}{\alpha}D_\alpha(QW\|PW)\right) \\ &\le \exp\left(\frac{\beta-1}{\beta}D_\beta(Q\|P)\right) = \|f(X)\|_\beta, \quad (15)\end{aligned}$$

where the inequality uses the fact that $\delta^*_{\beta\to\alpha} \le 1$, so

$$\begin{aligned}\frac{\alpha-1}{\alpha}D_\alpha(QW\|PW) &\le \frac{(\alpha-1)g(\alpha,\beta)}{\alpha}D_\beta(Q\|P) \\ &= \frac{\beta-1}{\beta}D_\beta(Q\|P).\end{aligned}$$

Therefore, $\|W^\sharp\|_{\beta\to\alpha} \le 1$. The converse $\|W^\sharp\|_{\beta\to\alpha} \le 1 \Rightarrow \delta^*_{\beta\to\alpha} \le 1$ is proved similarly. The case $1 > \beta \ge \alpha > 0$ is the same, except the inequality in (15) is reversed because $\alpha < 1$. The desired statement now follows immediately from [11, Theorem 3.3]. ∎

Acknowledgment. The author thanks V. Anantharam, S. Kamath, Y. Polyanskiy, P. Tetali, and Y. Wu for stimulating discussions.

References

[1] R. Ahlswede and P. Gács, "Spreading of sets in product spaces and hypercontraction of the Markov operator," *Ann. Probab.*, vol. 4, no. 6, pp. 925–939, 1976.

[2] H. S. Witsenhausen, "On sequences of pairs of dependent random variables," *SIAM J. Appl. Math.*, vol. 28, no. 1, pp. 100–113, January 1975.

[3] D. Bakry, "L'hypercontractivité et son utilisation en théorie des semigroupes," in *Lectures on Probability Theory*. Springer, 1994, vol. 1581, pp. 1–114.

[4] P. Diaconis and L. Saloff-Coste, "Logarithmic Sobolev inequalities for finite Markov chains," *Ann. Appl. Probab.*, vol. 6, no. 3, pp. 695–750, 1996.

[5] L. Miclo, "Remarques sur l'hypercontractivité et l'évolution de l'entropie pour des chaînes de Markov finies," *Séminaire de probabilités (Strasbourg)*, vol. 31, pp. 136–167, 1997.

[6] S. G. Bobkov and P. Tetali, "Modified logarithmic Sobolev inequalities in discrete settings," *J. Theor. Prob.*, vol. 19, no. 2, pp. 289–336, 2006.

[7] E. Mossel, K. Oleszkiewicz, and A. Sen, "On reverse hypercontractivity," 2013, to appear in *Geom. Funct. Anal.*

[8] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math., Statist. and Probab.*, vol. 1, 1961, pp. 547–561.

[9] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Trans. Inform. Theory*, vol. 52, no. 10, pp. 4394–4412, October 2006.

[10] T. van Erven and P. Harremoës, "Rényi divergence and Kullback–Leibler divergence," 2012, submitted.

[11] S. Kamath and V. Anantharam, "Non-interactive simulation of joint distributions: The Hirschfeld–Gebelein–Rényi maximal correlation and the hypercontractivity ribbon," in *Proc. 50th Annu. Allerton Conf. on Commun., Control, and Comput.*, Monticello, IL, October 2012.

[12] P. Diaconis, K. Khare, and L. Saloff-Coste, "Stochastic alternating projections," *Illinois J. Math.*, vol. 54, no. 3, pp. 963–979, 2010.

[13] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, "On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover," 2013, arXiv preprint. [Online]. Available: http://arxiv.org/abs/1304.6133