Strong Data Processing Inequalities via Sums of Squares

Oisín Faust
DAMTP, CCIMI
University of Cambridge, UK
opbf2@maths.cam.ac.uk

Hamza Fawzi
DAMTP
University of Cambridge, UK
h.fawzi@damtp.cam.ac.uk

***Abstract*—A hierarchy of semidefinite programming relaxations is described which gives certified upper bounds on the strong data processing (SDPI) constant of a discrete channel. The relaxations rely on a combination of tools from approximation theory and sum-of-squares techniques. By leveraging the properties of rational Padé approximants, we prove that the hierarchy converges to the true SDPI constant. Numerical experiments are performed which verify that these relaxations are very accurate even at low levels of the hierarchy.**

I. Introduction

A fundamental inequality in information theory is the *data processing inequality*, which states that if $W : \mathcal{X} \to \mathcal{Y}$ is a channel, then for any two input distributions μ, π on $\mathcal{X}$ we have

$$D(\mu W \| \pi W) \leq D(\mu \| \pi)$$

where $D(\mu\|\pi) = \sum_{i\in\mathcal{X}} \mu_i \log(\mu_i/\pi_i)$ is the relative entropy. For a specific channel W and reference input distribution π, such an inequality can usually be strengthened. We say that the pair (π, W) satisfies a *strong data processing inequality (SDPI)* if there is a constant $\delta < 1$ such that

$$D(\mu W \| \pi W) \leq \delta D(\mu \| \pi) \tag{1}$$

for all probability distributions μ on $\mathcal{X}$ [1], [2]. The smallest such $\delta = \delta^*(\pi, W)$ is the SDPI constant of (π, W). For example, the binary symmetric channel with noise ϵ and a uniform source is known to have SDPI constant $\delta^*(\mathrm{Bern}(\frac{1}{2}), \mathrm{BSC}_\epsilon) = (1-2\epsilon)^2$ [1].

Strong data processing inequalities, and more generally the contraction properties of discrete channels, have received a lot of attention recently in the information theory community [2]–[8], and have been applied e.g., to obtain various converse results.

However, computing the SDPI constant of a channel can be very difficult. When $|\mathcal{X}| = 2$ and π is fixed, the recent work [9] attempts to determine whether the problem (1) is convex under a certain natural parametrisation of the free variable μ. This approach is motivated by geodesically convex reformulations of the Brascamp-Lieb constant [10]. It turns out that this happens only when π is equal to a specific distribution depending on W, and it remains unclear whether this can be extended to $|\mathcal{X}| > 2$. Also, in the recent paper [8] it was shown that for the computation of $\delta^*(W) := \max_\pi \delta^*(\pi, W)$, it suffices to maximise $\delta^*(\pi, W)$ over input channels π supported on only two elements of $\mathcal{X}$, reducing the computational burden.

A. Contributions

In this work we propose a new method to compute accurate and certified upper bounds on the SDPI constant of a pair (π, W). We make use of the powerful *sum of squares* framework for global optimisation. Whereas sum of squares techniques are most directly applied to *polynomial* inequalities, we show how this framework can be brought to bear on entropy-based functional inequalities by means of an intermediary step involving rational Padé approximants to the logarithm. For integers $m \geq 2$, we describe a semidefinite program whose solution δ_m is an upper bound on $\delta^*(\pi, W)$. For fixed m the size of the semidefinite program grows polynomially in $|\mathcal{X}|$. The main result of the paper is the following:

Theorem 1. *Let W be a discrete channel and π a positive distribution on the input space $\mathcal{X} = \{1, \dots, n\}$. Then there are numbers $\delta_m \geq \delta^*(\pi, W)$ such that the following is true:*
(i) Each δ_m can be computed using a semidefinite program of size $\binom{n+m}{m} + n\binom{n+m-1}{m-1} + (n+n')(3m-1)$, where n' is the size of the output space.
(ii) $\lim_{m\to\infty} \delta_m = \delta^*(\pi, W)$.

Numerical experiments show that these relaxations are often very close to the true constant even for low levels of the hierarchy. We note that our approach can be used to bound other constants related to the SDPI, most notably the logarithmic Sobolev inequality. We give a brief explanation of this extension below, and refer the reader to [11] for more details about this extension.

B. Organisation

In Section II we review some background material on semidefinite programming and sums of squares. We present our approach to bounding the SDPI constant in Section III via Padé approximants of the logarithm function. In Section IV we illustrate our method on some particular discrete channels.

II. Semidefinite Programming and Sums of Squares

A. Notation and Definitions

Given a variable t, we let $\mathbb{R}[t]$ (resp. $\mathbb{R}[t]_d$) be the space of univariate real polynomials (resp. of degree at most d).

Given variables $x_1, x_2, \ldots, x_n$, $\mathbb{R}[x] = \mathbb{R}[x_1, \ldots, x_n]$ denotes the space of polynomials with real coefficients in $x = (x_1, x_2, \ldots, x_n)$, and $\mathbb{R}[x]_d$ is the (finite-dimensional) subspace of $\mathbb{R}[x]$ containing only polynomials of degree at most d.

A polynomial $p \in \mathbb{R}[x]$ is a *sum of squares* if we can find polynomials $p_1, \ldots, p_m$ such that $p = \sum_i p_i^2$. The set of polynomials that can be written as a sum of squares is a convex cone inside $\mathbb{R}[x]$ which we will denote by $\Sigma[x]$. A fundamental fact about $\Sigma[x]$, is that one can decide membership in $\Sigma[x]$ using *semidefinite programming* [12]–[14]. Recall that a semidefinite program (SDP) is an optimization problem of the form

$$\min_{X \in \mathbf{S}^n} \text{trace}(CX) \quad \text{s.t.} \quad \mathcal{A}(X) = b, X \succeq 0 \tag{2}$$

where $\mathbf{S}^n$ is the space of $n \times n$ real symmetric matrices, $X \succeq 0$ means that X is positive semidefinite; and $C \in \mathbf{S}^n$, $\mathcal{A} : \mathbf{S}^n \to \mathbb{R}^m$, $b \in \mathbb{R}^m$ are given. Whilst it is generally NP-hard to decide whether a polynomial takes only nonnegative values on $\mathbb{R}^n$ [15], there are efficient algorithms for solving SDPs to any desired precision using floating point arithmetic. Therefore, for many computational tasks involving constraints specifying that certain polynomials be nonnegative, we can construct relaxations based on the more tractable condition that these polynomials are sums of squares.

B. Constrained Polynomial Optimization

Given a collection of polynomials $g = (g_j)_1^J \subseteq \mathbb{R}[x]$, $d \in \mathbb{N}$

$$\mathbf{Q}_d(g) = \left\{ \sum_j g_j \sigma_j : \sigma_j \in \Sigma[x],\ \deg(g_j \sigma_j) \leq 2d \right\}$$

is the d^{th} *truncated quadratic module* generated by g. Observe that, by construction, any polynomial $p \in \mathbf{Q}_d(g)$ satisfies $p(x) \geq 0$ whenever $g_1(x) \geq 0, \ldots, g_J(x) \geq 0$.

Similarly, given a collection of polynomials $h = (h_k)_1^K \subseteq \mathbb{R}[x]$, $d \in \mathbb{N}$, let

$$\mathbf{I}_d(h) = \left\{ \sum_k h_k \phi_k : \phi_k \in \mathbb{R}[x],\ \deg(h_k \phi_k) \leq d \right\}$$

be the d^{th} *truncated ideal* generated by h. Observe that, by construction, any polynomial $p \in \mathbf{I}_d(h)$ satisfies $p(x) = 0$ whenever $h_1(x) = 0, \ldots, h_K(x) = 0$.

Now, given a polynomial $f \in \mathbb{R}[x]_{2d_0}$, suppose we are interested in determining whether f is nonnegative on the basic semialgebraic set defined by

$$\begin{aligned} g_j(x) &\geq 0 \quad j = 1 \ldots, J \\ h_k(x) &= 0 \quad k = 1 \ldots, K. \end{aligned} \tag{3}$$

where $(g_j)_j$ and (h_k) are polynomials. It is easy to see that a sufficient condition is that

$$f \in \mathbf{I}_{2d}(h) + \mathbf{Q}_d(g) \tag{4}$$

for some $d \geq d_0$. The condition (4) can be expressed as a semidefinite feasibility problem having J semidefinite constraints of size $\dim \mathbb{R}[x]_{d - \frac{1}{2}\deg g_j} \times \dim \mathbb{R}[x]_{d - \frac{1}{2}\deg g_j}$ each [16].

C. The Special Case of Univariate Polynomials

For univariate polynomials, global nonnegativity is equivalent to being a sum of squares. See for example [17, Theorem 2.3]. There is also a simple sum-of-squares characterisation of nonnegativity on a closed interval:

Theorem 2 (e.g. Theorem 2.4 in [17])**.** *Let $p \in \mathbb{R}[t]$ have degree k, and suppose $p(t) \geq 0$ for all $t \in [-1, 1]$. Then $p \in \mathbf{Q}_{\lceil k/2 \rceil}(t - 1, t + 1)$.*

III. Our approach

A. Main Idea

Consider a discrete memoryless channel $W : [n] \to [n']$ (we use the notation $[n] = \{1, \ldots, n\}$), and a strictly positive probability distribution $\pi \in \mathbb{R}^n_{>0}$. We are interested in bounding from above the smallest constant δ^* for which

$$D(\mu W \| \pi W) \leq \delta D(\mu \| \pi) \tag{5}$$

holds for all probability distributions $\mu \in \mathbb{R}^n_+$. We let

$$\pi_* = \min\{\pi_i \mid i \in \mathcal{X}\} > 0.$$

Let us parameterise μ by the variables $x_i = \frac{\mu_i}{\pi_i}$. Then x should take values in the $(n-1)$-dimensional simplex

$$\Delta^\pi := \{x \in \mathbb{R}^n_+ \text{ s.t. } \mathbb{E}_\pi[x] = \sum_i \pi_i x_i = 1\}.$$

Given a channel W, the backward channel is defined by $W^\sharp(i|j) = \frac{\pi_i W(j|i)}{(\pi W)_j}$. We have $(\mu W)_j = \sum_i \pi_i x_i W(j|i) = (W^\sharp x)_j (\pi W)_j$, where $(W^\sharp x)_j := \sum_i W^\sharp(i|j) x_i$. In other words, the density of μW with respect to πW is the vector $W^\sharp x \in \Delta^{\pi W} \subset \mathbb{R}^{n'}_+$.

We can now rewrite (5) in terms of x as

$$\delta \operatorname{Ent}_\pi[x] - \operatorname{Ent}_{\pi W}[W^\sharp x] \geq 0 \quad \forall x \in \Delta^\pi, \tag{6}$$

where for any probability distribution ν on $\mathcal{X}$, the functional $\operatorname{Ent}_\nu : \mathbb{R}^{\mathcal{X}}_+ \backslash \{0\} \to \mathbb{R}_+$ is defined by

$$y \mapsto \sum_i \nu_i y_i \log\left(\frac{y_i}{\mathbb{E}_\nu[y]}\right).$$

For $y \in \Delta^\nu$, the normalising factor inside the logarithm is of course unnecessary.

Evidently, the left hand side of (6) is not a polynomial in x. In order to make use of the sum-of-squares machinery described in Section II, we will replace the left hand side of (6) by a lower bound which is a polynomial. Specifically, if $P(x) \in \mathbb{R}[x]_n$ and $Q(y) \in \mathbb{R}[y]_{n'}$ are polynomials satisfying

$$\begin{cases} P(x) \leq \operatorname{Ent}_\pi[x] & \forall\, x \in \Delta^\pi \\ Q(y) \geq \operatorname{Ent}_{\pi W}[y] & \forall\, y \in \Delta^{\pi W} \end{cases} \tag{7}$$

then the polynomial inequality

$$\delta P(x) - Q(W^\sharp x) \geq 0 \quad \forall x \in \Delta^\pi \tag{8}$$

implies (6). Fixing $d \geq \frac{1}{2}\max\{\deg P, \deg Q\}$, the following sum-of-squares feasibility program can be phrased as a semidefinite program of size $O(n^d)$:

$$\delta P(x) - Q(W^\sharp x) \in \mathsf{SOS}_d(\Delta^\pi) \tag{9}$$

where $\mathsf{SOS}_d(\Delta^\pi)$ is a sum-of-squares relaxation for the set of nonnegative polynomials on Δ^π, which we define to be

$$\mathsf{SOS}_d(\Delta^\pi) := \mathbf{I}_{2d}(1 - \mathbb{E}_\pi x) + \mathbf{Q}_d(x). \tag{10}$$

Recall that $\mathbf{I}_{2d}(1 - \mathbb{E}_\pi x) = (1 - \mathbb{E}_\pi x)\mathbb{R}[x]_{2d-1}$ is the $2d^{\text{th}}$ truncated ideal generated by $1 - \mathbb{E}_\pi x \equiv 1 - \sum_{i=1}^n \pi_i x_i$, and $\mathbf{Q}_d(x) = \Sigma[x]_{2d} + \sum_{i=1}^n x_i \Sigma[x]_{2d-2}$ is the d^{th} truncated quadratic module generated by $\{x_i\}_{i=1}^n$. A solution to the semidefinite feasibility program provides a certificate that $\delta \geq \delta^*$.

Conversely, one can easily verify that if P and Q are polynomials s.t.

$$\begin{cases} P(x) \geq (1-\epsilon)\operatorname{Ent}_\pi[x] & \forall\, x \in \Delta^\pi \\ Q(y) \leq (1+\epsilon)\operatorname{Ent}_{\pi W}[y] & \forall\, y \in \Delta^{\pi W} \end{cases} \tag{11}$$

then, with $\delta = \left(\frac{1+\epsilon}{1-\epsilon}\right)\delta^*$ (8) holds. [Indeed, with $y = W^\sharp x$ we have $\delta P(x) - Q(y) \geq \delta(1-\epsilon)\operatorname{Ent}_\pi[x] - (1+\epsilon)\operatorname{Ent}_{\pi W}[y] = (1+\epsilon)(\delta^* \operatorname{Ent}_\pi[x] - \operatorname{Ent}_{\pi W}[y]) \geq 0$ by definition of δ^*.] In fact, one can prove the stronger statement that (9) will hold for large enough d, using recent results from convex algebraic geometry [18]–[20].

Theorem 3. *Let $W : [n] \to [n']$ be a discrete memoryless channel and $\pi \in \mathbb{R}^n_{>0}$ a positive probability distribution. Let δ^* be the SDPI constant of (π, W). Let $\epsilon \in (0,1)$, and suppose that $P(x)$ and $Q(y)$ are polynomials satisfying* (11)*. Then for large enough d, $\delta P(x) - Q(W^\sharp x) \in \mathsf{SOS}_d(\Delta^\pi)$, with $\delta = \left(\frac{1+2\epsilon}{1-\epsilon}\right)\delta^*$.*

Sketch of proof. If we let $f(x) = \delta P(x) - Q(W^\sharp x)$, we know from the above discussion that $f(x) \geq 0$ for all $x \in \Delta^\pi$. It is a foundational result in sum-of-squares programming that any polynomial which is positive on a basic semialgebraic set has a sum-of-squares representation on that set provided the semialgebraic set satisfies the *archimedean condition*, an algebraic strengthening of compactness which holds, for example, for simplices. This is known as Putinar's Positivstellensatz [21]. We cannot directly use this result however, as $f(x) = \delta P(x) - Q(W^\sharp x)$ cannot be positive at $x = \mathbb{1}$ if P and Q satisfy (7).

Instead we use a result from a more recent line of work [18]–[20]. More precisely, [20, Theorem 1.1] states that if a polynomial f is nonnegative on Δ^π, and at each zero of f in Δ^π the constraint qualification, strict complementarity, and second order sufficiency conditions for local minima from nonlinear programming hold, then $f \in \mathsf{SOS}_d(\Delta^\pi)$ for sufficiently large d. These conditions can be verified for our polynomial $f(x)$, however we omit the details due to limited space. □

B. Polynomial Bounds on Entropy

So far we have not given any indication as to how the polynomials $P(x)$ and $Q(y)$ should be chosen. This is the topic of this subsection. In the sum-of-squares program (9), we could imagine allowing $P(x) = \sum_i \pi_i p_i(x_i)$ and $Q(y) = \sum_j \pi_j q_j(y_j)$ to vary subject to each p_i (q_j) being a lower (upper) bound on $t\log(t)$, i.e.,

$$\begin{cases} p_i(t) \leq t\log(t)\ \forall t \in [0, {\pi_i}^{-1}], i \in [n] \\ q_j(t) \geq t\log(t)\ \forall t \in [0, {(\pi W)_j}^{-1}], j \in [n']. \end{cases} \tag{12}$$

However, we currently lack an algorithmic way to enforce the constraints above. A naive approach would be to impose the inequalities on a large but finite set of points, however this does not guarantee that the inequalities are valid on the whole interval. The approach we adopt here, is that given a rational function (i.e. ratio of polynomials) $R(t)/Q(t)$ where $Q > 0$ which is an upper bound for $t\log(t)$, we can enforce the constraint $q_j(t) \geq R(t)/Q(t)$ on the interval $[0, {(\pi W)_j}^{-1}]$ by requiring a sum-of-squares certificate for the *univariate* polynomial $Q(t)q_j(t) - R(t)$. Such a certificate will naturally imply the stronger condition $q_j(t) \geq t\log(t)$. Similarly, given a rational lower bound for $t\log(t)$, we can enforce $p_i(t) \leq t\log(t)$ by requiring a certain polynomial to be a sum of squares on the interval $[0, \pi_i^{-1}]$. This approach should be fruitful if we can find rational functions of modest degree which are relatively tight upper/lower bounds for $t\log(t)$ (or simply for $\log(t)$, since the rational function $tR(t)/Q(t)$ is then a bound for $t\log(t)$). In what follows, we will use the $(m+1, m)$ *Padé approximant* of $\log(t)$.

The $(m+1, m)$ Padé approximant of $\log(t)$ around $t = 1$ is the rational function

$$\overline{r_m}(t) := \frac{(t-1)R_m(t)}{S_m(t)}$$

where R_m and S_m are polynomials of degree m such that $S_m(1) = 1$, and such that derivatives of $\overline{r_m}(t)$ at $t = 1$ agree with those of $\log(t)$ to order $2m+1$. That is, $\log(t) - \overline{r_m}(t) = O\big((t-1)^{2m+2}\big)$. Although defined by the same category of data as the truncated Taylor series' (namely the first few derivatives of $\log(t)$ at $t = 1$), the Padé approximants are better approximations of $\log(t)$. More precisely, one can show that $\overline{r_m}$ converges to $\log$ as $m \to \infty$ with geometric rate pointwise on the positive real line, while the sequence of truncated Taylor approximations diverges for $t > 2$, see e.g., [22] or [23, Prop. 6]. For the purpose of this work however, the only fact we will need about $\overline{r_m}$ is the following:

Lemma 4. *For each $m \in \mathbb{N}$, we have*

$$0 \leq \overline{r_m}(t) - \log(t) \leq B_m(t-1)^2/t \qquad \forall t > 0$$

where $B_m \downarrow 0$ as $m \to \infty$.

Proof. This can be proved using the properties of Padé approximants, and the fact that $\overline{r_m}$ coincides with the Gauss-Radau quadrature rule applied to the integral representation $\log(x) = \int_0^1 (x-1)/(t(x-1)+1)dt$ [24, Remark 2.10]. We omit the details because of limitations on space. □

We will also need rational lower bounds on $t\log(t)$. Observe that $\log(t) = -\log(1/t) \geq -\overline{r_m}(1/t)$. Therefore

$$t\log(t) \geq -t\,\overline{r_m}(1/t) = \frac{(t-1)\overline{R_m}(t)}{\overline{S_m}(t)},$$

where $\overline{S_m}(t) := t^m S_m(1/t)$ and $\overline{R_m}(t) := t^m R_m(1/t)$ are degree-m polynomials.

Our strategy is to replace the two conditions in (12) with

$$\begin{aligned} p_i(t) &\leq -t\,\overline{r_m}(1/t) \qquad \forall\, t \in [0, \pi_i^{-1}], \\ q_j(t) &\geq t\,\overline{r_m}(t) \qquad \forall\, t \in [0, (\pi W)_j^{-1}]. \end{aligned}$$

By Theorem 2, these are *equivalent* to the sum-of-squares constraints

$$\begin{cases} (t-1)\overline{R_m}(t) - p_i(t)\overline{S_m}(t) \in \mathbf{Q}_{d+\lfloor\frac{m+1}{2}\rfloor}\big(t, 1-\pi_i t\big) \\ q_j(t)S_m(t) - t(t-1)R_m(t) \in \mathbf{Q}_{d+\lfloor\frac{m+1}{2}\rfloor}\big(t, 1-(\pi W)_j t\big). \end{cases} \tag{13}$$

The following theorem attests that for any $\epsilon \in (0,1)$, there is a large enough m that this system of constraints admits polynomials $P(x) = \sum_{i=1}^n \pi_i p(x_i)$ and $Q(y) = \sum_{j=1}^{n'} (\pi W)_j q(y_j)$ which also satisfy the condition (11). In combination with Theorem 3, it will allow us to prove Theorem 1, which says that the resulting hierarchy of upper approximations to the SDPI constant converges to it in the limit of taking m and d to infinity.

Theorem 5. *Let $\epsilon \in (0,1)$ and let π, ν be positive probability distributions on $[n]$, $[n']$ respectively. Write $\pi_* = \min_i \pi_i$ and $\nu_* = \min_j \nu_j$. There exists $m \in \mathbb{N}$ and univariate polynomials $p, q \in \mathbb{R}[t]$ such that*

$$\begin{cases} p(t) \leq -t\,\overline{r_m}(1/t) & t \in [0, {\pi_*}^{-1}] \\ q(t) \geq t\,\overline{r_m}(t) & t \in [0, \nu_*^{-1}], \end{cases} \tag{14}$$

and

$$P(x) = \sum_{i=1}^n \pi_i p(x_i) \geq (1-\epsilon)\operatorname{Ent}_\pi[x] \quad \forall x \in \Delta^\pi \tag{15}$$

$$Q(y) = \sum_{j=1}^{n'} \nu_j q(y_j) \leq (1+\epsilon)\operatorname{Ent}_\nu[y] \quad \forall y \in \Delta^\nu. \tag{16}$$

Proof. Write $N_* = \max\{\pi_*^{-1}, \nu_*^{-1}\}$, and let $\epsilon_1 > 0$ be a constant depending on ϵ and on N_* whose value we will determine later. Choose $m \in \mathbb{Z}_{\geq 0}$ large enough that $B_m < \epsilon_1$ (see Lemma 4). We obtain that for every $t \geq 0$

$$t\,\overline{r_m}(t) - t\log(t) \leq \epsilon_1 (t-1)^2 \tag{17}$$

and (substituting $t \leftarrow 1/t$)

$$t\log(t) + t\,\overline{r_m}(1/t) \leq \epsilon_1 (t-1)^2. \tag{18}$$

Applying the Weierstrass Approximation Theorem to the continuous functions[1] $t \mapsto \frac{t\,\overline{r_m}(t) - (t-1)}{(t-1)^2} + \frac{\epsilon_1}{2}$ and $t \mapsto \frac{-t\,\overline{r_m}(1/t) - (t-1)}{(t-1)^2} + \frac{\epsilon_1}{2}$, we deduce the existence of polynomials p, q satisfying

$$\begin{aligned} t\,\overline{r_m}(t) \leq q(t) &\leq t\,\overline{r_m}(t) + \epsilon_1 (t-1)^2 \\ -t\,\overline{r_m}(1/t) - \epsilon_1 (t-1)^2 \leq p(t) &\leq -t\,\overline{r_m}(1/t) \end{aligned}$$

[1]Note that $\overline{r_m}(t) = (t-1) + O((t-1)^2)$ as $t \to 1$ since $\overline{r_m}(t) - \log(t) = O((t-1)^{2m+2})$.

for all $t \in [0, N_*]$. These polynomials certainly satisfy (14). Combining the above bounds with (17) and (18), we have, for $x \in \Delta^\pi$ and $y \in \Delta^\nu$:

$$\sum\nolimits_{i=1}^n \pi_i p(x_i) \geq \operatorname{Ent}_\pi[x] - 2\epsilon_1 \sum\nolimits_i \pi_i (x_i - 1)^2 \tag{19}$$

$$\sum\nolimits_{j=1}^{n'} \nu_j q(y_j) \leq \operatorname{Ent}_\nu[y] + 2\epsilon_1 \sum\nolimits_j \nu_j (y_j - 1)^2. \tag{20}$$

It remains to bound the last terms in (19) and (20) by $\epsilon \operatorname{Ent}_\pi[x]$ and $\epsilon \operatorname{Ent}_\nu[y]$, in order to obtain (15) and (16). Pinsker's inequality yields, for $x \in \Delta^\pi$,

$$\operatorname{Ent}_\pi[x] \geq \frac{1}{2}\left(\sum\nolimits_i \pi_i |x_i - 1|\right)^2 \geq \frac{1}{2N_*}\sum\nolimits_i \pi_i (x_i - 1)^2,$$

since $\pi_i \geq 1/N_*$ by definition of N_*. Similarly $\operatorname{Ent}_\nu[y] \geq \frac{1}{2N_*}\sum_j \nu_j (y_j - 1)^2$ for $y \in \Delta^\nu$. Choosing $\epsilon_1 = \frac{\epsilon}{4N_*}$ completes the proof. □

C. Main Result

We can now present a consistent sum-of-squares hierarchy for estimating the SDPI constant of discrete channel from above. Given integers $d \geq 2$, $m \geq 1$, the $(d,m)^{\text{th}}$ level $\delta_{d,m}$ of our hierarchy is defined by

$$\delta_{d,m} := \min_{\substack{\delta \in \mathbb{R} \\ p_i \in \mathbb{R}[t]_{2d} \\ q_j \in \mathbb{R}[t]_{2d}}} \delta \text{ s.t. } \begin{cases} p_i, q_j \text{ satisfy (13)}, \\ \delta P(x) - Q(W^\sharp x) \in \mathsf{SOS}_d(\Delta^\pi) \\ \text{where } P(x) := \sum_{i=1}^n \pi_i p_i(x_i) \\ \qquad Q(y) := \sum_{j=1}^{n'} (\pi W)_j q_i(y_j). \end{cases} \tag{21}$$

Theorem 6. *Let W be a discrete channel and π a positive distribution on the input space. For every $d \geq 2$ and $m \geq 1$ we have $\delta_{d,m} \geq \delta^*(\pi, W)$, and*

$$\lim_{m\to\infty} \lim_{d\to\infty} \delta_{d,m} = \delta^*(\pi, W).$$

Proof. We have already seen that if the p_i and q_j satisfy (13) then $P(x) \leq \operatorname{Ent}_\pi[x]$ for $x \in \Delta^\pi$ and $Q(y) \leq \operatorname{Ent}_{\pi W}[y]$ for $y \in \Delta^{\pi W}$. The third constraint of (21) then implies $0 \leq \delta P(x) - Q(W^\sharp x) \leq \delta \operatorname{Ent}_\pi[x] - \operatorname{Ent}_{\pi W}[W^\sharp x]$, i.e., $\delta \geq \delta^*(\pi, W)$.

To prove the second part, let $\epsilon \in (0,1)$. By Theorem 5 with $\nu = \pi W$, there exists large enough m, and polynomials $p, q \in \mathbb{R}[t]$ that satisfy (13) and such that (11) holds. For such P, Q, we know from Theorem 3 that, with $\delta = \left(\frac{1+2\epsilon}{1-\epsilon}\right)\delta^*$, $\delta P(x) - Q(W^\sharp x) \in \mathsf{SOS}_d(\Delta^\pi)$ for large enough d. This shows that $\delta_{d,m} \leq \left(\frac{1+2\epsilon}{1-\epsilon}\right)\delta^*$ as desired. □

Remark 1. The semidefinite program corresponding to $\delta_{d,m}$, is over the product of positive semidefinite cones

$$\underbrace{\left(\mathbf{S}_+^{2d+m-1}\right)^n \times \left(\mathbf{S}_+^{2d+m-1}\right)^{n'}}_{(13)} \times \underbrace{\mathbf{S}_+^{\binom{n+d}{d}} \times \left(\mathbf{S}_+^{\binom{n+d-1}{d-1}}\right)^n}_{\text{second line of (21)}}$$

which has dimension $O(n^{2d} + m^2(n+n'))$ for fixed d.

Remark 2. $\delta_{d,m}$ is separately nonincreasing in both d and in m, so taking $d = m$ yields a sequence indexed by one parameter that also converges to $\delta^*(\pi, W)$. This justifies our earlier statement of this result in Theorem 1.

D. Extension to Logarithmic Sobolev Inequalities

Consider an irreducible Markov kernel K on a finite state space $\mathcal{X}$, and let π be its stationary distribution. We associate with K a nonnegative quadratic form

$$\mathcal{E}(x) = (1/2)\sum\nolimits_{i,j\in\mathcal{X}} \pi_i K_{ij}(x_i - x_j)^2$$

known as the *Dirichlet form*. A *logarithmic Sobolev inequality* for (K,π) has the form

$$\mathcal{E}(x) \geq \alpha \operatorname{Ent}_\pi[x^2] \qquad \forall\, x \in \mathbb{R}^{\mathcal{X}}\,; \tag{22}$$

the largest constant α for which this inequality holds is called the logarithmic Sobolev (LSI) constant of (K,π). The logarithmic Sobolev constant characterises the hypercontractivity of the Markov semigroup $P_t = e^{-t(I-K)}$ associated to the kernel [25], [26]. It is also known to give good bounds on the mixing time of the Markov process [27], [28]. LSI is closely related to SDPI – in [3] it is shown that for a discrete channel W and arbitrary source π, the SDPI constant $\delta^*(\pi,W)$ is bounded by $\delta^*(\pi,W) \leq 1 - \alpha(W^\sharp W,\pi)$. One can use similar techniques as the ones developed in this paper to obtain rigorous lower bounds on α using semidefinite programming. We refer the reader to the preprint [11] for more details.

IV. Numerical Results

A. Binary Symmetric Channel

For the binary symmetric channel with noise ϵ and a uniform source, the SDPI constant is known to be $\delta^*(\mathrm{Bern}(\frac{1}{2}),\mathrm{BSC}_\epsilon) = (1-2\epsilon)^2$. We observed numerically that our hierarchy is exact at a level (d,m) which depends on the noise level ϵ. For $\epsilon = 0.4$, the first level is already exact: $\delta_{2,1} = 0.04$. For lower levels of noise, higher levels are needed. Table I lists the relative error of the approximation for the first few levels of the hierarchy for $\epsilon = 0.01$ (relative error $= \frac{\delta_{d,m} - (1-2\epsilon)^2}{(1-2\epsilon)^2}$). We see that in this case the relaxation is exact when $d \geq 6$ and $m \geq 4$.

TABLE I
Relative Approximation Error for $\delta^*\left(\mathrm{Bern}(\frac{1}{2}),\ \mathrm{BSC}_{0.01}\right)$

$m\backslash d$	2	3	4	5	6
1	2.36e−1	2.21e−1	2.20e−1	2.20e−1	2.20e−1
2	1.56e−1	6.74e−2	5.77e−2	5.55e−2	5.55e−2
3	1.54e−1	4.17e−2	1.99e−2	1.36e−2	1.06e−2
4	1.54e−1	3.89e−2	1.03e−2	1.18e−3	0
5	1.54e−1	3.49e−2	8.88e−3	2.01e−4	0

B. Binary Erasure Channel

We consider the channel BEC_ϵ on a binary input space which leaves its input unchanged with probability $1-\epsilon$ and otherwise outputs a third "blank" letter. For a uniform input distribution, the SDPI constant of this channel is $\delta^*(\mathrm{Bern}(\frac{1}{2}),\mathrm{BEC}_\epsilon) = 1-\epsilon$. In fact, (1) holds *with equality for every* μ. This is problematic for our methods, since our polynomial bounds can only ever be tight at a discrete number of points. It is not hard to verify directly that for fixed (d,m), $\delta_{d,m}(\mathrm{Bern}(\frac{1}{2}),\mathrm{BEC}_\epsilon)$ is proportional to $1-\epsilon$. Thus the relative error $\frac{\delta_{d,m}-(1-\epsilon)}{1-\epsilon}$ of our approximations does not depend on ϵ, so we need only measure it for $\epsilon = \frac{1}{2}$, say. Table II illustrates the convergence of our method.

TABLE II
Relative Approximation Error for $\delta^*\left(\mathrm{Bern}(\frac{1}{2}),\ \mathrm{BEC}_{1/2}\right)$

d,m	**rel. error**	d,m	**rel. error**	d,m	**rel. error**
2	1.63e−1	7	1.14e−2	12	4.29e−3
3	5.55e−2	8	8.99e−3	13	3.69e−3
4	2.99e−2	9	7.27e−3	14	3.22e−3
5	2.04e−2	10	6.00e−3	15	2.82e−3
6	1.49e−2	11	5.03e−3	16	2.50e−3

C. Toy Channel

Consider the following toy example a discrete memoryless channel $W : [4] \to [5]$:

$$W = \begin{pmatrix} 0.1 & 0.2 & 0.0 & 0.5 \\ 0.0 & 0.5 & 0.2 & 0.1 \\ 0.2 & 0.0 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.2 & 0.3 \\ 0.5 & 0.1 & 0.4 & 0.0 \end{pmatrix}$$

and a source distribution $\pi = (0.2, 0.2, 0.4, 0.2)$ on the input space.

In order to judge the accuracy of our relaxations, we need a way of obtaining lower bounds on the SDPI constant of $\delta^*(\pi,W)$. One way to do this is to locally maximize the nonconcave function $\operatorname{Ent}_{\pi W}[W^\sharp x]$ subject to the constraint $\operatorname{Ent}_\pi[x] \leq 1$, using e.g. a Newton-type method with random initialisation. This can repeated for several different random initialisations and the best lower bound recorded. By doing this, we obtained $\delta^* \geq \underline{\delta} = 0.374806998$. This value is used to determine the relative errors, listed in Table III, of the first few levels of the hierarchy.

TABLE III
Relative Approximation Error for $\delta^*(\pi,\ W)$

$m\backslash d$	2	3	4	5
2	2.21e−1	5.87e−2	4.22e−2	4.06e−2
3	2.15e−1	3.52e−2	9.73e−2	2.62e−3
4	2.15e−1	3.27e−2	3.55e−3	6.76e−6
5	2.14e−1	3.23e−2	2.63e−3	3.44e−7

V. Conclusion

In this paper we have presented a new hierarchy of convergent upper bounds $(\delta_{d,m})$ on the SDPI constant of any discrete memoryless channel, based on semidefinite programming, and sum-of-squares methods. Numerical experiments suggest that the quality of the upper bound is already very good at the first levels of the hierarchy. An interesting question would be to obtain quantitative rates on the convergence of $\delta_{d,m}$ to the true constant δ^*.

REFERENCES

[1] R. Ahlswede and P. Gacs, "Spreading of Sets in Product Spaces and Hypercontraction of the Markov Operator," *The Annals of Probability*, vol. 4, no. 6, pp. 925 – 939, 1976. [Online]. Available: https://doi.org/10.1214/aop/1176995937

[2] Y. Polyanskiy and Y. Wu, "Strong data-processing inequalities for channels and Bayesian networks," in *Convexity and Concentration*, E. Carlen, M. Madiman, and E. M. Werner, Eds. New York, NY: Springer New York, 2017, pp. 211–249.

[3] M. Raginsky, "Logarithmic Sobolev Inequalities and Strong Data Processing Theorems for Discrete Channels," in *2013 IEEE International Symposium on Information Theory*, 2013, pp. 419–423.

[4] T. A. Courtade, "Outer bounds for multiterminal source coding via a strong data processing inequality," in *2013 IEEE International Symposium on Information Theory*. IEEE, 2013, pp. 559–563.

[5] V. Anantharam, A. Gohari, S. Kamath, and C. Nair, "On hypercontractivity and a data processing inequality," in *2014 IEEE International Symposium on Information Theory*. IEEE, 2014, pp. 3022–3026.

[6] A. Xu and M. Raginsky, "Converses for distributed estimation via strong data processing inequalities," in *2015 IEEE International Symposium on Information Theory (ISIT)*, 2015, pp. 2376–2380.

[7] Y. Polyanskiy and Y. Wu, "Application of the information-percolation method to reconstruction problems on graphs," *Mathematical Statistics and Learning*, vol. 2, no. 1, pp. 1–24, 2020.

[8] O. Ordentlich and Y. Polyanskiy, "Strong data processing constant is achieved by binary inputs," *IEEE Transactions on Information Theory*, pp. 1–1, 2021.

[9] Q. Ding, C. W. Lau, C. Nair, and Y. N. Wang, "Concavity of output relative entropy for channels with binary inputs," in *2021 IEEE International Symposium on Information Theory (ISIT)*, 2021, pp. 2738–2743.

[10] S. Sra, N. K. Vishnoi, and O. Yildiz, "On Geodesically Convex Formulations for the Brascamp-Lieb Constant," in *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2018)*, ser. Leibniz International Proceedings in Informatics (LIPIcs), E. Blais, K. Jansen, J. D. P. Rolim, and D. Steurer, Eds., vol. 116. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2018, pp. 25:1–25:15. [Online]. Available: http://drops.dagstuhl.de/opus/volltexte/2018/9429

[11] O. Faust and H. Fawzi, "Sum-of-Squares proofs of logarithmic Sobolev inequalities on finite Markov chains," 2021. [Online]. Available: http://arxiv.org/abs/2101.04988

[12] N. Shor, "Class of global minimum bounds of polynomial functions," *Cybernetics*, vol. 23, pp. 731–734, 1987.

[13] P. A. Parrilo, "Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization," Ph.D. dissertation, Caltech, 2000.

[14] J. B. Lasserre, "Global optimization with polynomials and the problem of moments," 2001.

[15] K. G. Murty and S. N. Kabadi, "Some NP-complete problems in quadratic and nonlinear programming," *Mathematical Programming*, vol. 39, pp. 117–129, 1987.

[16] P. A. Parrilo, "Semidefinite programming relaxations for semialgebraic problems," *Mathematical Programming*, vol. 96, no. 2, pp. 293–320, 2003.

[17] J. B. Lasserre, *Positive Polynomials and Moment Problems*, ser. Cambridge Texts in Applied Mathematics. Cambridge University Press, 2015, p. 15–53.

[18] C. Scheiderer, "Sums of squares on real algebraic curves," *Mathematische Zeitschrift*, vol. 245, pp. 725–760, 2003.

[19] M. Marshall, "Representations of non-negative polynomials having finitely many zeros," *Annales de la faculté des sciences de Toulouse Mathématiques*, vol. 15, no. 3, pp. 599–609, 2006. [Online]. Available: http://eudml.org/doc/10014

[20] J. Nie, "Optimality conditions and finite convergence of Lasserre's hierarchy," *Mathematical Programming*, vol. 146, pp. 97–121, 2014.

[21] M. Putinar, "Positive polynomials on compact semi-algebraic sets," *Indiana University Mathematics Journal*, vol. 42, no. 3, pp. 969–984, 1993. [Online]. Available: http://www.jstor.org/stable/24897130

[22] F. Topsøe, "Some bounds for the logarithmic function," *Inequality theory and applications*, vol. 4, p. 137, 2006.

[23] H. Fawzi, J. Saunderson, and P. A. Parrilo, "Semidefinite approximations of the matrix logarithm," *Foundations of Computational Mathematics*, vol. 2, pp. 259–296, 2019.

[24] P. Brown, H. Fawzi, and O. Fawzi, "Device-independent lower bounds on the conditional von neumann entropy," *arXiv preprint arXiv:2106.13692*, 2021.

[25] L. Gross, "Logarithmic Sobolev inequalities," *American Journal of Mathematics*, vol. 97, no. 4, pp. 1061–1083, 1975. [Online]. Available: http://www.jstor.org/stable/2373688

[26] M. Ledoux, *The concentration of measure phenomenon*, ser. Mathematical Surveys and Monographs. Providence, RI.: American Mathematical Society, Providence, RI., 2001, vol. 89.

[27] P. Diaconis and L. Saloff-Coste, "Logarithmic Sobolev inequalities for finite Markov chains," *The Annals of Applied Probability*, vol. 6, no. 3, pp. 695–750, 1996. [Online]. Available: http://www.jstor.org/stable/2245210

[28] L. Saloff-Coste, *Lectures on finite Markov chains*. Berlin, Heidelberg: Springer, 1997, pp. 301–413. [Online]. Available: https://doi.org/10.1007/BFb0092621