On Data-Processing and Majorization Inequalities for f-Divergences

Igal Sason
Andrew and Erna Viterbi Faculty of Electrical Engineering
Technion-Israel Institute of Technology
Haifa 32000, Israel
E-mail: sason@ee.technion.ac.il

***Abstract*—This work introduces new strong data-processing and majorization inequalities for f-divergences, and it studies some of their applications in information theory and statistics. The full paper version [16] will be published soon in the *Entropy* journal, including all proofs and further results, discussions, and information-theoretic applications. One application refers to the performance analysis of list decoding with either fixed or variable list sizes. Another application is related to a study of the quality of approximating a probability mass function, induced by the leaves of a Tunstall tree, by an equiprobable distribution. The compression rates of finite-length Tunstall codes are further analyzed for asserting their closeness to the Shannon entropy of a memoryless and stationary discrete source.**

Index Terms – Contraction coefficient, data-processing inequalities, f-divergences, hypothesis testing, list decoding, majorization, Rényi information measures, Tsallis entropy, Tunstall trees.

I. Introduction

Divergences are non-negative measures of the dissimilarity between pairs of probability measures which are defined on the same measurable space. They play a key role in the development of information theory, probability theory, statistics, learning, signal processing, and other related fields. One important class of divergence measures is defined by means of convex functions f, and it is called the class of f-divergences. It unifies fundamental and independently-introduced concepts in several branches of mathematics such as the chi-squared test for the goodness of fit in statistics, the total variation distance in functional analysis, the relative entropy in information theory and statistics, and it is also closely related to the Rényi divergence which generalizes the relative entropy. The class of f-divergences was independently introduced in the sixties by Ali and Silvey [2], and Csiszár [5]. This class satisfies pleasing features such as the data-processing inequality, convexity, continuity and duality properties, and it finds nice applications in information theory and statistics (see, e.g., [6], [7], [8], [17], [19], [20], [21]).

The full paper version of this work [16] is a research paper which is focused on the derivation of data-processing and majorization inequalities for f-divergences, and a study of some of their potential applications in information theory and statistics. Preliminaries are next provided.

II. Preliminaries

A. Preliminaries and Related Works

We provide here definitions which serve as a background to the presentation in this paper. We first provide a definition for the family of f-divergences.

Definition 1: [9, p. 4398] Let P and Q be probability measures, let μ be a dominating measure of P and Q (i.e., $P, Q \ll \mu$), and let $p := \frac{\mathrm{d}P}{\mathrm{d}\mu}$ and $q := \frac{\mathrm{d}Q}{\mathrm{d}\mu}$. The f-divergence from P to Q is given, independently of μ, by

$$D_f(P\|Q) := \int q\, f\Bigl(\frac{p}{q}\Bigr)\, \mathrm{d}\mu, \tag{1}$$

where

$$f(0) := \lim_{t\to 0^+} f(t), \tag{2}$$

$$0 f\Bigl(\frac{0}{0}\Bigr) := 0, \tag{3}$$

$$0 f\Bigl(\frac{a}{0}\Bigr) := \lim_{t\to 0^+} t f\Bigl(\frac{a}{t}\Bigr) = a \lim_{u\to\infty} \frac{f(u)}{u}, \quad a > 0. \tag{4}$$

Definition 2: Let Q_X be a probability distribution which is defined on a set $\mathcal{X}$, and that is not a point mass, and let $W_{Y|X}\colon \mathcal{X} \to \mathcal{Y}$ be a stochastic transformation. The contraction coefficient for f-divergences is defined as

$$\mu_f(Q_X, W_{Y|X}) := \sup_{P_X \colon D_f(P_X\|Q_X) \in (0,\infty)} \frac{D_f(P_Y\|Q_Y)}{D_f(P_X\|Q_X)}, \tag{5}$$

where, for all $y \in \mathcal{Y}$,

$$P_Y(y) = (P_X W_{Y|X})\,(y) := \int_{\mathcal{X}} \mathrm{d}P_X(x)\, W_{Y|X}(y|x), \tag{6}$$

$$Q_Y(y) = (Q_X W_{Y|X})\,(y) := \int_{\mathcal{X}} \mathrm{d}Q_X(x)\, W_{Y|X}(y|x). \tag{7}$$

Contraction coefficients for f-divergences play a key role in strong data-processing inequalities (see [1], [12], [13]).

Definition 3: Pearson's χ^2-divergence from P to Q is defined to be the f-divergence from P to Q (see Definition 1) with $f(t) = (t-1)^2$ or $f(t) = t^2 - 1$ for all $t > 0$,

$$\begin{aligned} \chi^2(P\|Q) &:= D_f(P\|Q) && (8)\\ &= \int \frac{(p-q)^2}{q}\, \mathrm{d}\mu && (9)\\ &= \int \frac{p^2}{q}\, \mathrm{d}\mu - 1 && (10) \end{aligned}$$

independently of the dominating measure μ (i.e., $P, Q \ll \mu$, e.g., $\mu = P + Q$).

Neyman's χ^2-divergence from P to Q is the Pearson's χ^2-divergence from Q to P, i.e., it is equal to

$$\chi^2(Q\|P) = D_g(P\|Q) \tag{11}$$

with $g(t) = \frac{(t-1)^2}{t}$ or $g(t) = \frac{1}{t} - t$ for all $t > 0$.

For the presentation of our majorization inequalities for f-divergences and related entropy bounds, essential definitions and basic results are next provided (see, e.g., [11]). Let P be a probability mass function defined on a finite set $\mathcal{X}$, let $p_{\max}$ be the maximal mass of P, and let $G_P(k)$ be the sum of the k largest masses of P for $k \in \{1, \ldots, |\mathcal{X}|\}$ (hence, it follows that $G_P(1) = p_{\max}$ and $G_P(|\mathcal{X}|) = 1$).

Definition 4: Consider discrete probability mass functions P and Q defined on a finite set $\mathcal{X}$. It is said that P is majorized by Q (or Q majorizes P), and it is denoted by $P \prec Q$, if $G_P(k) \leq G_Q(k)$ for all $k \in \{1, \ldots, |\mathcal{X}|\}$ (recall that $G_P(|\mathcal{X}|) = G_Q(|\mathcal{X}|) = 1$).

A unit mass majorizes any other distribution; on the other hand, the equiprobable distribution on a finite set is majorized by any other distribution defined on the same set.

Definition 5: Let $\mathcal{P}_n$ denote the set of all the probability mass functions that are defined on $\mathcal{A}_n := \{1, \ldots, n\}$. A function $f \colon \mathcal{P}_n \to \mathbb{R}$ is said to be *Schur-convex* if for every $P, Q \in \mathcal{P}_n$ such that $P \prec Q$, we have $f(P) \leq f(Q)$. Likewise, f is said to be *Schur-concave* if $-f$ is Schur-convex, i.e., $P, Q \in \mathcal{P}_n$ and $P \prec Q$ imply that $f(P) \geq f(Q)$.

Finally, what is the connection between data processing and majorization, and why these types of inequalities are both considered in the same manuscript ? This connection is provided in the following fundamental well-known result (see, e.g., [11, Theorem B.2]):

Proposition 1: Let P and Q be probability mass functions defined on a finite set $\mathcal{A}$. Then, $P \prec Q$ if and only if there exists a doubly-stochastic transformation $W_{Y|X} \colon \mathcal{A} \to \mathcal{A}$ (i.e., $\sum\limits_{x \in \mathcal{A}} W_{Y|X}(y|x) = 1$ for all $y \in \mathcal{A}$, and $\sum\limits_{y \in \mathcal{A}} W_{Y|X}(y|x) = 1$ for all $x \in \mathcal{A}$ with $W_{Y|X}(\cdot|\cdot) \geq 0$) such that

$$Q \to W_{Y|X} \to P.$$

In other words, $P \prec Q$ if and only if in their representation as column vectors, there exists a doubly-stochastic matrix $\mathbf{W}$ (i.e., a square matrix with non-negative entries such that the sum of each column or each row in $\mathbf{W}$ is equal to 1) such that $P = \mathbf{W}Q$.

B. Contributions

This work (see the full paper version in [16]) is focused on the derivation of data-processing and majorization inequalities for f-divergences, and it applies these inequalities to information theory and statistics.

The starting point for obtaining strong data-processing inequalities in this paper relies on the derivation of lower and upper bounds on the difference $D_f(P_X\|Q_X) - D_f(P_Y\|Q_Y)$ where (P_X, Q_X) and (P_Y, Q_Y) denote, respectively, pairs of input and output probability distributions with a given stochastic transformation $W_{Y|X}$ (i.e., $P_X \to W_{Y|X} \to P_Y$, $Q_X \to W_{Y|X} \to Q_Y$). These bounds are expressed in terms of the respective difference in the Pearson's or Neyman's χ^2-divergence, and they hold for all f-divergences (see Theorem 1).

This paper also derives majorization inequalities for f-divergences where part of these inequalities rely on the earlier data-processing inequalities (see Theorem 3). A different approach, which relies on the concept of majorization, serves to derive tight bounds on the maximal value of an f-divergence from a probability mass function P to an equiprobable distribution; the maximization is carried over all P with a fixed finite support where the ratio of their maximal to minimal probability masses does not exceed a given value (see Theorem 4). These bounds lead to accurate asymptotic results which apply to general f-divergences, and they strengthen and generalize recent results of this type with respect to the relative entropy [4], and the Rényi divergence [15].

As an application of the data-processing inequalities for f-divergences, the setup of list decoding is further studied in [16], reproducing in a unified way some known bounds on the list decoding error probability, and deriving new bounds for fixed and variable list sizes.

As an application of the majorization inequalities in this paper, we study in [16] properties of a measure which is used to quantify the quality of approximating probability mass functions, induced by the leaves of a Tunstall tree, by an equiprobable distribution. An application of majorization inequalities for the relative entropy is used to derive a sufficient condition, expressed in terms of the principal and secondary real branches of the Lambert W function, for asserting the proximity of compression rates of finite-length (lossless and variable-to-fixed) Tunstall codes to the Shannon entropy of a memoryless and stationary discrete source.

III. Main Results on f-divergences

A. Data-processing inequalities for f-divergences

Strong data-processing inequalities are provided in the following, bounding the difference $D_f(P_X\|Q_X) - D_f(P_Y\|Q_Y)$ and ratio $\frac{D_f(P_Y\|Q_Y)}{D_f(P_X\|Q_X)}$ where (P_X, Q_X) and (P_Y, Q_Y) denote, respectively, pairs of input and output probability distributions with a given stochastic transformation.

Theorem 1: Let $\mathcal{X}$ and $\mathcal{Y}$ be finite or countably infinite sets, let P_X and Q_X be probability mass functions that are supported on $\mathcal{X}$, and let

$$\xi_1 := \inf_{x \in \mathcal{X}} \frac{P_X(x)}{Q_X(x)} \in [0, 1], \tag{12}$$

$$\xi_2 := \sup_{x \in \mathcal{X}} \frac{P_X(x)}{Q_X(x)} \in [1, \infty]. \tag{13}$$

Let $W_{Y|X} \colon \mathcal{X} \to \mathcal{Y}$ be a stochastic transformation such that for every $y \in \mathcal{Y}$, there exists $x \in \mathcal{X}$ with $W_{Y|X}(y|x) > 0$, and let (see (6) and (7))

$$P_Y := P_X W_{Y|X}, \tag{14}$$

$$Q_Y := Q_X W_{Y|X}. \tag{15}$$

Furthermore, let $f\colon (0,\infty) \to \mathbb{R}$ be a convex function with $f(1)=0$, and let the non-negative constant $c_f := c_f(\xi_1,\xi_2)$ satisfy

$$f'_+(v) - f'_+(u) \geq 2c_f\,(v-u), \quad \forall\, u,v \in \mathcal{I},\ u<v \tag{16}$$

where f'_+ denotes the right-side derivative of f, and

$$\mathcal{I} := \mathcal{I}(\xi_1,\xi_2) = [\xi_1,\xi_2] \cap (0,\infty). \tag{17}$$

Then,

a)

$$\begin{aligned} & D_f(P_X\|Q_X) - D_f(P_Y\|Q_Y) \\ & \geq c_f(\xi_1,\xi_2)\left[\chi^2(P_X\|Q_X) - \chi^2(P_Y\|Q_Y)\right] \qquad (18) \\ & \geq 0, \qquad (19) \end{aligned}$$

where equality holds in (18) if $D_f(\cdot\|\cdot)$ is Pearson's χ^2-divergence with $c_f \equiv 1$.

b) If f is twice differentiable on $\mathcal{I}$, then the largest possible coefficient in the right side of (16) is given by

$$c_f(\xi_1,\xi_2) = \tfrac{1}{2} \inf_{t\in\mathcal{I}(\xi_1,\xi_2)} f''(t). \tag{20}$$

c) Under the assumption in Item b), the following dual inequality also holds:

$$\begin{aligned} & D_f(P_X\|Q_X) - D_f(P_Y\|Q_Y) \\ & \geq c_{f^*}\Bigl(\tfrac{1}{\xi_2},\tfrac{1}{\xi_1}\Bigr)\left[\chi^2(Q_X\|P_X) - \chi^2(Q_Y\|P_Y)\right] \qquad (21) \\ & \geq 0, \qquad (22) \end{aligned}$$

where $f^*\colon (0,\infty)\to\mathbb{R}$ is the dual convex function which is given by

$$f^*(t) := t\,f\biggl(\frac{1}{t}\biggr), \quad \forall\, t>0, \tag{23}$$

and the coefficient in the right side of (21) satisfies

$$c_{f^*}\Bigl(\tfrac{1}{\xi_2},\tfrac{1}{\xi_1}\Bigr) = \tfrac{1}{2} \inf_{t\in\mathcal{I}(\xi_1,\xi_2)} \{t^3 f''(t)\} \tag{24}$$

with the convention that $\frac{1}{\xi_1} = \infty$ if $\xi_1 = 0$. Equality holds in (21) if $D_f(\cdot\|\cdot)$ is Neyman's χ^2-divergence (i.e., $D_f(P\|Q) := \chi^2(Q\|P)$ for all P and Q) with $c_{f^*} \equiv 1$.

d) Under the assumption in Item b), if

$$e_f(\xi_1,\xi_2) := \tfrac{1}{2} \sup_{t\in\mathcal{I}(\xi_1,\xi_2)} f''(t) < \infty, \tag{25}$$

then,

$$\begin{aligned} & D_f(P_X\|Q_X) - D_f(P_Y\|Q_Y) \\ & \leq e_f(\xi_1,\xi_2)\left[\chi^2(P_X\|Q_X) - \chi^2(P_Y\|Q_Y)\right]. \qquad (26) \end{aligned}$$

Furthermore,

$$\begin{aligned} & D_f(P_X\|Q_X) - D_f(P_Y\|Q_Y) \\ & \leq e_{f^*}\Bigl(\tfrac{1}{\xi_2},\tfrac{1}{\xi_1}\Bigr)\left[\chi^2(Q_X\|P_X) - \chi^2(Q_Y\|P_Y)\right] \qquad (27) \end{aligned}$$

where the coefficient in the right side of (27) satisfies

$$e_{f^*}\Bigl(\tfrac{1}{\xi_2},\tfrac{1}{\xi_1}\Bigr) = \tfrac{1}{2} \sup_{t\in\mathcal{I}(\xi_1,\xi_2)} \{t^3 f''(t)\}, \tag{28}$$

which is assumed to be finite. Equalities hold in (26) and (27) if $D_f(\cdot\|\cdot)$ is Pearson's or Neyman's χ^2-divergence with $e_f \equiv 1$ or $e_{f^*} \equiv 1$, respectively.

e) The lower and upper bounds in (18), (21), (26) and (27) are locally tight. More precisely, let $\{P_X^{(n)}\}$ be a sequence of probability mass functions defined on $\mathcal{X}$ and pointwise converging to Q_X which is supported on $\mathcal{X}$, and let $P_Y^{(n)}$ and Q_Y be the probability mass functions defined on $\mathcal{Y}$ via (14) and (15) with inputs $P_X^{(n)}$ and Q_X, respectively. Suppose that

$$\lim_{n\to\infty} \inf_{x\in\mathcal{X}} \frac{P_X^{(n)}(x)}{Q_X(x)} = 1, \tag{29}$$

$$\lim_{n\to\infty} \sup_{x\in\mathcal{X}} \frac{P_X^{(n)}(x)}{Q_X(x)} = 1. \tag{30}$$

If f has a continuous second derivative at unity, then

$$\lim_{n\to\infty} \frac{D_f(P_X^{(n)}\|Q_X) - D_f(P_Y^{(n)}\|Q_Y)}{\chi^2(P_X^{(n)}\|Q_X) - \chi^2(P_Y^{(n)}\|Q_Y)} = \tfrac{1}{2} f''(1), \tag{31}$$

$$\lim_{n\to\infty} \frac{D_f(P_X^{(n)}\|Q_X) - D_f(P_Y^{(n)}\|Q_Y)}{\chi^2(Q_X\|P_X^{(n)}) - \chi^2(Q_Y\|P_Y^{(n)})} = \tfrac{1}{2} f''(1), \tag{32}$$

and these limits indicate the local tightness of the lower and upper bounds in Items a)–d).

Proof: See [16]. ■

In continuation to [10, Theorem 8], we next provide an upper bound on the contraction coefficient for another subclass of f-divergences. Although the first part of the next result is stated for finite or countably infinite alphabets, it is clear from its proof that it also holds in the general alphabet setting. Connections to the literature are provided in [16].

Theorem 2: Let $f\colon (0,\infty)\to\mathbb{R}$ satisfy the conditions:

- f is a convex function, differentiable at 1, $f(1)=0$, and $f(0) := \lim_{t\to 0^+} f(t) < \infty$;
- The function $g\colon (0,\infty)\to\mathbb{R}$, defined by $g(t) := \frac{f(t)-f(0)}{t}$ for all $t>0$, is convex.

Let

$$\kappa(\xi_1,\xi_2) := \sup_{t\in(\xi_1,1)\cup(1,\xi_2)} \frac{f(t) + f'(1)\,(1-t)}{(t-1)^2} \tag{33}$$

where, for P_X and Q_X which are non-identical probability mass functions, $\xi_1 \in [0,1)$ and $\xi_2 \in (1,\infty]$ are given in (12) and (13). Then, in the setting of (14) and (15),

$$\frac{D_f(P_Y\|Q_Y)}{D_f(P_X\|Q_X)} \leq \frac{\kappa(\xi_1,\xi_2)}{f(0)+f'(1)} \cdot \frac{\chi^2(P_Y\|Q_Y)}{\chi^2(P_X\|Q_X)}. \tag{34}$$

Consequently, if Q_X is finitely supported on $\mathcal{X}$,

$$\mu_f(Q_X, W_{Y|X}) \le \frac{1}{f(0)+f'(1)} \cdot \kappa\left(0, \frac{1}{\min\limits_{x\in\mathcal{X}} Q_X(x)}\right) \cdot \mu_{\chi^2}(Q_X, W_{Y|X}). \tag{35}$$

Proof: See [16]. ■

We refer the reader to a parametric subclass of f-divergences with interesting properties which is introduced in [16], and which satisfies the conditions of Theorem 2.

B. *f-divergence Inequalities via Majorization*

Let U_n denote an equiprobable probability mass function on $\{1, \dots, n\}$ with $n \in \mathbb{N}$, i.e., $U_n(i) := \frac{1}{n}$ for all $i \in \{1, \dots, n\}$. By majorization theory and Theorem 1, the next result strengthens the Schur-convexity property of the f-divergence $D_f(\cdot\|U_n)$ (see [3, Lemma 1]).

Theorem 3: Let P and Q be probability mass functions which are supported on $\{1, \dots, n\}$, and suppose that $P \prec Q$. Let $f\colon (0,\infty) \to \mathbb{R}$ be twice differentiable and convex with $f(1) = 0$, and let $q_{\max}$ and $q_{\min}$ be, respectively, the maximal and minimal positive masses of Q. Then,

a)

$$\begin{aligned} & n e_f(nq_{\min}, nq_{\max}) \left(\|Q\|_2^2 - \|P\|_2^2\right) \\ & \ge D_f(Q\|U_n) - D_f(P\|U_n) \qquad (36) \\ & \ge n c_f(nq_{\min}, nq_{\max}) \left(\|Q\|_2^2 - \|P\|_2^2\right) \ge 0, \qquad (37) \end{aligned}$$

where $c_f(\cdot,\cdot)$ and $e_f(\cdot,\cdot)$ are given in (20) and (25), respectively, and $\|\cdot\|_2$ denotes the Euclidean norm. Furthermore, (36) and (37) hold with equality if $D_f(\cdot\|\cdot) = \chi^2(\cdot\|\cdot)$.

b) If $P \prec Q$ and $\frac{q_{\max}}{q_{\min}} \le \rho$ for an arbitrary $\rho \ge 1$, then

$$0 \le \|Q\|_2^2 - \|P\|_2^2 \le \frac{(\rho-1)^2}{4\rho n}. \tag{38}$$

Proof: See [16]. ■

The next result provides upper and lower bounds on f-divergences from any probability mass function to an equiprobable distribution. It relies on majorization theory, and it follows in part from Theorem 3.

Theorem 4: Let $\mathcal{P}_n$ denote the set of all the probability mass functions that are defined on $\mathcal{A}_n := \{1, \dots, n\}$. For $\rho \ge 1$, let $\mathcal{P}_n(\rho)$ be the set of all $Q \in \mathcal{P}_n$ which are supported on $\mathcal{A}_n$ with $\frac{q_{\max}}{q_{\min}} \le \rho$, and let $f\colon (0,\infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$. Then,

a) The set $\mathcal{P}_n(\rho)$, for any $\rho \ge 1$, is a non-empty, convex and compact set.

b) For a given $Q \in \mathcal{P}_n$, which is supported on $\mathcal{A}_n$, the f-divergences $D_f(\cdot\|Q)$ and $D_f(Q\|\cdot)$ attain their maximal values over the set $\mathcal{P}_n(\rho)$.

c) For $\rho \ge 1$ and an integer $n \ge 2$, let

$$u_f(n,\rho) := \max_{Q\in\mathcal{P}_n(\rho)} D_f(Q\|U_n), \tag{39}$$

$$v_f(n,\rho) := \max_{Q\in\mathcal{P}_n(\rho)} D_f(U_n\|Q), \tag{40}$$

let

$$\Gamma_n(\rho) := \left[\frac{1}{1+(n-1)\rho}, \frac{1}{n}\right], \tag{41}$$

and let the probability mass function $Q_\beta \in \mathcal{P}_n(\rho)$ be defined on the set $\mathcal{A}_n$ as follows:

$$Q_\beta(j) := \begin{cases} \rho\beta, & j \in \{1, \dots, i_\beta\}, \\ 1 - \big(n + i_\beta(\rho-1) - 1\big)\beta, & j = i_\beta + 1, \\ \beta, & i_\beta + 2 \le j \le n \end{cases} \tag{42}$$

where

$$i_\beta := \left\lfloor \frac{1-n\beta}{(\rho-1)\beta} \right\rfloor. \tag{43}$$

Then,

$$u_f(n,\rho) = \max_{\beta\in\Gamma_n(\rho)} D_f(Q_\beta\|U_n), \tag{44}$$

$$v_f(n,\rho) = \max_{\beta\in\Gamma_n(\rho)} D_f(U_n\|Q_\beta). \tag{45}$$

d) For $\rho \ge 1$ and an integer $n \ge 2$, let the non-negative function $g_f^{(\rho)}\colon [0,1] \to \mathbb{R}_+$ be given by

$$g_f^{(\rho)}(x) := x f\left(\frac{\rho}{1+(\rho-1)x}\right) + (1-x) f\left(\frac{1}{1+(\rho-1)x}\right), \tag{46}$$

for all $x \in [0,1]$. Then,

$$\max_{m\in\{0,\dots,n\}} g_f^{(\rho)}\left(\tfrac{m}{n}\right) \le u_f(n,\rho) \le \max_{x\in[0,1]} g_f^{(\rho)}(x), \tag{47}$$

$$\max_{m\in\{0,\dots,n\}} g_{f^*}^{(\rho)}\left(\tfrac{m}{n}\right) \le v_f(n,\rho) \le \max_{x\in[0,1]} g_{f^*}^{(\rho)}(x) \tag{48}$$

with the convex function $f^*\colon (0,\infty) \to \mathbb{R}$ in (23).

e) The right-side inequalities in (47) and (48) are asymptotically tight ($n \to \infty$). Namely,

$$\lim_{n\to\infty} u_f(n,\rho) = \max_{x\in[0,1]} g_f^{(\rho)}(x), \tag{49}$$

$$\lim_{n\to\infty} v_f(n,\rho) = \max_{x\in[0,1]} g_{f^*}^{(\rho)}(x). \tag{50}$$

f) If $g_f^{(\rho)}(\cdot)$ in (46) is differentiable on $(0,1)$ and its derivative is upper bounded by $K_f(\rho) \ge 0$, then for every integer $n \ge 2$

$$0 \le \lim_{n'\to\infty} \left\{u_f(n',\rho)\right\} - u_f(n,\rho) \le \frac{K_f(\rho)}{n}. \tag{51}$$

g) Let $f(0) := \lim\limits_{t\to 0} f(t) \in (-\infty, +\infty]$, and let $n \ge 2$ be an integer. Then,

$$\lim_{\rho\to\infty} u_f(n,\rho) = \left(1 - \frac{1}{n}\right) f(0) + \frac{f(n)}{n}. \tag{52}$$

Furthermore, if $f(0) < \infty$, f is differentiable on $(0, n)$, and $K_n := \sup_{t \in (0,n)} \bigl|f'(t)\bigr| < \infty$, then, for every $\rho \geq 1$,

$$0 \leq \lim_{\rho' \to \infty} \bigl\{u_f(n, \rho')\bigr\} - u_f(n, \rho) \leq \frac{2K_n\,(n-1)}{n+\rho-1}. \quad (53)$$

h) For $\rho \geq 1$, let the function f be also twice differentiable, and let M and m be constants such that the following condition holds:

$$0 \leq m \leq f''(t) \leq M, \quad \forall\, t \in \bigl[\tfrac{1}{\rho}, \rho\bigr]. \quad (54)$$

Then, for all $Q \in \mathcal{P}_n(\rho)$,

$$\begin{aligned} 0 &\leq \tfrac{1}{2} m \bigl(n\|Q\|_2^2 - 1\bigr) && (55) \\ &\leq D_f(Q\|U_n) && (56) \\ &\leq \tfrac{1}{2} M \bigl(n\|Q\|_2^2 - 1\bigr) && (57) \\ &\leq \frac{M(\rho-1)^2}{8\rho} && (58) \end{aligned}$$

with equalities in (56) and (57) for the χ^2 divergence (with $M = m = 2$).

i) Let $d > 0$. If $f''(t) \leq M_f \in (0, \infty)$ for all $t > 0$, then $D_f(Q\|U_n) \leq d$ for all $Q \in \mathcal{P}_n(\rho)$, if

$$\rho \leq 1 + \frac{4d}{M_f} + \sqrt{\frac{8d}{M_f} + \frac{16d^2}{M_f^2}}. \quad (59)$$

Proof: See [16]. ∎

Tsallis entropy was introduced in [18] as a generalization of the Shannon entropy (similarly to the Rényi entropy [14]), and it was applied to statistical physics in [18].

Definition 6: [18] Let P_X be a probability mass function defined on a discrete set $\mathcal{X}$. The *Tsallis entropy of order* $\alpha \in (0,1) \cup (1, \infty)$ of X, denoted by $S_\alpha(X)$ or $S_\alpha(P_X)$, is

$$S_\alpha(X) = \frac{\|P_X\|_\alpha^\alpha - 1}{1-\alpha}, \quad (60)$$

where $\|P_X\|_\alpha := \left(\sum_{x \in \mathcal{X}} P_X^\alpha(x)\right)^{\frac{1}{\alpha}}$. The Tsallis entropy is continuously extended at orders 0, 1, and ∞; at order 1, it coincides with the Shannon entropy in nats.

Theorem 3 enables to strengthen the Schur-concavity property of the Tsallis entropy (see [11, Theorem 13.F.3.a.]).

Theorem 5: Let P and Q be probability mass functions which are supported on a finite set, and let $P \prec Q$. Then, for all $\alpha > 0$,

a)

$$0 \leq L(\alpha, P, Q) \leq S_\alpha(P) - S_\alpha(Q) \leq U(\alpha, P, Q), \quad (61)$$

where

$$L(\alpha, P, Q) := \begin{cases} \frac{1}{2}\,\alpha q_{\min}^{\alpha-2}\,\bigl(\|Q\|_2^2 - \|P\|_2^2\bigr), & \text{if } \alpha \in (0, 2], \\ \frac{1}{2}\,\alpha q_{\min}^{\alpha-2}\,\bigl(\|Q\|_2^2 - \|P\|_2^2\bigr), & \alpha \in (2, \infty), \end{cases} \quad (62)$$

$$U(\alpha, P, Q) := \begin{cases} \frac{1}{2}\,\alpha q_{\min}^{\alpha-2}\,\bigl(\|Q\|_2^2 - \|P\|_2^2\bigr), & \text{if } \alpha \in (0, 2], \\ \frac{1}{2}\,\alpha q_{\max}^{\alpha-2}\,\bigl(\|Q\|_2^2 - \|P\|_2^2\bigr), & \alpha \in (2, \infty), \end{cases} \quad (63)$$

and the bounds in (62) and (63) are attained at $\alpha = 2$.

b)

$$\inf_{P \prec Q} \frac{S_\alpha(P) - S_\alpha(Q)}{L(\alpha, P, Q)} = \sup_{P \prec Q} \frac{S_\alpha(P) - S_\alpha(Q)}{U(\alpha, P, Q)} = 1,$$

where the inf. and sup. in (b) can be restricted to PMFs P and Q $(P \neq Q)$ supported on a binary alphabet.

References

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

[2] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistics Society*, Series B, vol. 28, no. 1, pp. 131–142, 1966.

[3] F. Cicalese, L. Gargano and U. Vaccaro, "A note on approximation of uniform distributions from variable-to-fixed length codes," *IEEE Transactions on Information Theory*, vol. 52, no. 8, pp. 3772–3777, August 2006.

[4] F. Cicalese, L. Gargano, and U. Vaccaro, "Bounds on the entropy of a function of a random variable and their applications," *IEEE Transactions on Information Theory*, vol. 64, no. 4, pp. 2220–2230, April 2018.

[5] I. Csiszár, "Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Bewis der Ergodizität von Markhoffschen Ketten," *Publ. Math. Inst. Hungar. Acad. Sci.*, vol. 8, pp. 85–108, January 1963.

[6] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 299–318, January 1967.

[7] I. Csiszár, "A class of measures of informativity of observation channels," *Periodica Mathematicarum Hungarica*, vol. 2, no. 1, pp. 191–213, March 1972.

[8] F. Liese and I. Vajda, *Convex Statistical Distances* (Teubner-Texte Zur Mathematik), vol. 95. Leipzig, Germany, 1987.

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

[10] A. Makur and L. Zheng, "Linear bounds between contraction coefficients for f-divergences," *preprint*, July 2018. [Online]. Available at https://arxiv.org/pdf/1510.01844.pdf.

[11] A. W. Marshall, I. Olkin and B. C. Arnold, *Inequalities: Theory of Majorization and Its Applications*, second edition, Springer, 2011.

[12] Y. Polyanskiy and Y. Wu, "Strong data processing inequalities for channels and Bayesian networks," *Convexity and Concentration*, the IMA Volumes in Mathematics and its Applications (Editors: E. Carlen, M. Madiman and E. M. Werner), vol. 161, pp. 211–249, Springer, 2017.

[13] M. Raginsky, "Strong data processing inequalities and Φ-Sobolev inequalities for discrete channels," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3355–3389, June 2016.

[14] A. Rényi, "On measures of entropy and information," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pp. 547–561, University of California Press, Berkeley, California, USA, 1961.

[15] I. Sason, "Tight bounds on the Rényi entropy via majorization with applications to guessing and compression," *Entropy*, vol. 20, no. 12, paper 896, pp. 1–25, November 2018.

[16] I. Sason, "On data-processing and majorization inequalities for f-divergences with applications," *Entropy*, vol. 21, no. 10, paper 1022, pp. 1–80, October 2019.

[17] W. Stummer and I. Vajda, "On divergences of finite measures and their applicability in statistics and information theory," *Statistics*, vol. 44, no. 2, pp. 169–187, April 2010.

[18] C. Tsallis, "Possible generalization of the Boltzmann-Gibbs statistics," *Journal of Statistical Physics*, vol. 52, no. 1–2, pp. 479–487, July 1988.

[19] I. Vajda, *Theory of Statistical Inference and Information*, Kluwer Academic Publishers, 1989.

[20] M. Zakai and J. Ziv, "A generalization of the rate-distortion theory and applications," *Information Theory - New Trends and Open Problems* (Editor: G. Longo), pp. 87–123, Springer, 1975.

[21] J. Ziv and M. Zakai, "On functionals satisfying a data-processing theorem," *IEEE Transactions on Information Theory*, vol. 19, no. 3, pp. 275–283, May 1973.