entropy

MDPI

Article

On Data-Processing and Majorization Inequalities for f-Divergences with Applications

Igal Sason

Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 3200003, Israel; sason@ee.technion.ac.il; Tel.: +972-4-8294699

Received: 17 July 2019; Accepted: 17 October 2019; Published: 21 October 2019

Abstract: This paper is focused on the derivation of data-processing and majorization inequalities for f-divergences, and their applications in information theory and statistics. For the accessibility of the material, the main results are first introduced without proofs, followed by exemplifications of the theorems with further related analytical results, interpretations, and information-theoretic applications. One application refers to the performance analysis of list decoding with either fixed or variable list sizes; some earlier bounds on the list decoding error probability are reproduced in a unified way, and new bounds are obtained and exemplified numerically. 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. Almost all the analysis is relegated to the appendices, which form the major part of this manuscript.

Keywords: contraction coefficient; data-processing inequalities; f-divergences; hypothesis testing; list decoding; majorization theory; Rényi information measures; Tsallis entropy; Tunstall trees

1. Introduction

Divergences are non-negative measures of 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 closely related to the Rényi divergence which generalizes the relative entropy. The class of f-divergences was introduced in the sixties by Ali and Silvey [1], Csiszár [2–6], and Morimoto [7]. This class satisfies pleasing features such as the data-processing inequality, convexity, continuity and duality properties, finding interesting applications in information theory and statistics (see, e.g., [4,6,8–15]).

This manuscript 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.

1.1. Preliminaries and Related Works

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

Definition 1 ([16], 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 $p := \frac{dP}{d\mu}$ and $q := \frac{dQ}{d\mu}$. The f-divergence from P to Q is given, independently of μ, by*

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

where

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

$$0 f\left(\frac{0}{0}\right) := 0, \tag{3}$$

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

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 $y \in \mathcal{Y}$,

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

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

The notation in (6) *and* (7)*, and also in* (20), (21), (42), (43), (44) *in the continuation of this paper, is consistent with the standard notation used in information theory (see, e.g., the first displayed equation after (3.2) in [17]).*

Contraction coefficients for f-divergences play a key role in strong data-processing inequalities (see [18–20], ([21], Chapter II), [22–26]). The following are essential definitions and results which are related to maximal correlation and strong data-processing inequalities.

Definition 3. *The maximal correlation between two random variables X and Y is defined as*

$$\rho_{\mathrm{m}}(X;Y) := \sup_{f,g} \mathbb{E}[f(X)g(Y)], \tag{8}$$

where the supremum is taken over all real-valued functions f and g such that

$$\mathbb{E}[f(X)] = \mathbb{E}[g(Y)] = 0, \quad \mathbb{E}[f^2(X)] \le 1,\ \mathbb{E}[g^2(Y)] \le 1. \tag{9}$$

Definition 4. *Pearson's χ^2-divergence [27] 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 $t > 0$,*

$$\chi^2(P\|Q) := D_f(P\|Q) \tag{10}$$

$$= \int \frac{(p-q)^2}{q}\, d\mu \tag{11}$$

$$= \int \frac{p^2}{q}\, d\mu - 1 \tag{12}$$

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

Neyman's χ^2-divergence [28] from P to Q is the Pearson's χ^2-divergence from Q to P, i.e., it is equal to

$$\chi^2(Q\|P) = D_g(P\|Q) \tag{13}$$

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

Proposition 1 (([24], Theorem 3.2), [29])**.** *The contraction coefficient for the χ^2-divergence satisfies*

$$\mu_{\chi^2}(Q_X, W_{Y|X}) = \rho_{\mathrm{m}}^2(X;Y) \tag{14}$$

with $X \sim Q_X$ and $Y \sim Q_Y$ (see (7)).

Proposition 2 ([25], Theorem 2)**.** *Let $f\colon (0,\infty) \to \mathbb{R}$ be convex and twice continuously differentiable with $f(1) = 0$ and $f''(1) > 0$. Then, for any Q_X that is not a point mass,*

$$\mu_{\chi^2}(Q_X, W_{Y|X}) \le \mu_f(Q_X, W_{Y|X}), \tag{15}$$

i.e., the contraction coefficient for the χ^2-divergence is the minimal contraction coefficient among all f-divergences with f satisfying the above conditions.

Remark 1. *A weaker version of (15) was presented in ([21], Proposition II.6.15) in the general alphabet setting, and the result in (15) was obtained in ([24], Theorem 3.3) for finite alphabets.*

The following result provides an upper bound on the contraction coefficient for a subclass of f-divergences in the finite alphabet setting.

Proposition 3 ([26], Theorem 8)**.** *Let $f\colon [0,\infty) \to \mathbb{R}$ be a continuous convex function which is three times differentiable at unity with $f(1) = 0$ and $f''(1) > 0$, and let it further satisfy the following conditions:*

(a)

$$\Bigl(f(t) - f'(1)\,(t-1)\Bigr)\left(1 - \frac{f^{(3)}(1)(t-1)}{3f''(1)}\right) \ge \tfrac{1}{2}f''(1)(t-1)^2, \quad \forall\, t > 0. \tag{16}$$

(b) *The function $g\colon (0,\infty) \to \mathbb{R}$, given by $g(t) := \frac{f(t)-f(0)}{t}$ for all $t > 0$, is concave.*

Then, for a probability mass function Q_X supported over a finite set $\mathcal{X}$,

$$\mu_f(Q_X, W_{Y|X}) \le \left(\frac{f'(1) + f(0)}{f''(1)\,\min\limits_{x\in\mathcal{X}} Q_X(x)}\right)\mu_{\chi^2}(Q_X, W_{Y|X}). \tag{17}$$

For the presentation of our majorization inequalities for f-divergences and related entropy bounds (see Section 2.3), essential definitions and basic results are next provided (see, e.g., [30], ([31], Chapter 13) and ([32], Chapter 2)). 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 5. *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 $G_P(k) \le G_Q(k)$ 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 6. *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 $f(P) \geq f(Q)$.*

Characterization of Schur-convex functions is provided, e.g., in ([30], Chapter 3). For example, there exist some connections between convexity and Schur-convexity (see, e.g., ([30], Section 3.C) and ([32], Chapter 2.3)). However, a Schur-convex function is not necessarily convex ([32], Example 2.3.15).

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., ([32], Theorem 2.1.10), ([30], Theorem B.2) and ([31], Chapter 13)):

Proposition 4. *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 $P = \mathbf{W}Q$.*

1.2. Contributions

This paper 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., where $P_X \to W_{Y|X} \to P_Y$ and $Q_X \to W_{Y|X} \to Q_Y$). 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 Theorems 1 and 2). By a different approach, we derive an upper bound on the contraction coefficient for f-divergences of a certain type, which gives an alternative strong data-processing inequality for the considered type of f-divergences (see Theorems 3 and 4). In this framework, a parametric subclass of f-divergences is introduced, its interesting properties are studied (see Theorem 5), all the data-processing inequalities which are derived in this paper are applied to this subclass, and these inequalities are exemplified numerically to examine their tightness (see Section 3.1).

This paper also derives majorization inequalities for f-divergences where part of these inequalities rely on the earlier data-processing inequalities (see Theorem 6). 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 7). 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 [33], and the Rényi divergence [34]. Furthermore, we explore in Theorem 7 the convergence rates to the asymptotic results. Data-processing and majorization inequalities also serve to strengthen the Schur-concavity property of the Tsallis entropy (see Theorem 8), showing by a comparison to earlier bounds in [35,36] that none of these bounds is superseded by the other. Further analytical results which are related to the specialization of our central result on majorization inequalities in Theorem 7, applied to several important sub-classes of f-divergences, are provided in

Section 3.2 (including Theorem 9). A quantity which is involved in our majorization inequalities in Theorem 7 is interpreted by relying on a variational representation of f-divergences (see Theorem 10).

As an application of the data-processing inequalities for f-divergences, the setup of list decoding is further studied, reproducing in a unified way some known bounds on the list decoding error probability, and deriving new bounds for fixed and variable list sizes (see Theorems 11–13).

As an application of the majorization inequalities in this paper, we study 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 (see Theorem 14). 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 [37], 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 (see Theorem 15).

1.3. Paper Organization

The paper is structured as follows: Section 2 provides our main new results on data-processing and majorization inequalities for f-divergences and related entropy measures. Illustration of the theorems in Section 2, and further mathematical results which follow from these theorems are introduced in Section 3. Applications in information theory and statistics are considered in Section 4. Proofs of all theorems are relegated to the appendices, which form a major part of this paper.

2. Main Results on f-Divergences

This section provides strong data-processing inequalities for f-divergences (see Section 2.1), followed by a study of a new subclass of f-divergences (see Section 2.2) which later serves to exemplify our data-processing inequalities. The third part of this section (see Section 2.3) provides majorization inequalities for f-divergences, and for the Tsallis entropy, whose derivation relies in part on the new data-processing inequalities.

2.1. 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*

$$\xi_1 := \inf_{x\in\mathcal{X}} \frac{P_X(x)}{Q_X(x)} \in [0,1], \tag{18}$$

$$\xi_2 := \sup_{x\in\mathcal{X}} \frac{P_X(x)}{Q_X(x)} \in [1,\infty]. \tag{19}$$

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))

$$P_Y := P_X W_{Y|X}, \tag{20}$$

$$Q_Y := Q_X W_{Y|X}. \tag{21}$$

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

$$f'_+(v) - f'_+(u) \geq 2c_f\,(v-u), \quad \forall\, u,v \in \mathcal{I},\ u < v \tag{22}$$

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

$$\mathcal{I} := \mathcal{I}(\xi_1, \xi_2) = [\xi_1, \xi_2] \cap (0, \infty). \tag{23}$$

Then,

(a)

$$D_f(P_X \| Q_X) - D_f(P_Y \| Q_Y) \geq c_f(\xi_1, \xi_2) \left[\chi^2(P_X \| Q_X) - \chi^2(P_Y \| Q_Y) \right] \tag{24}$$

$$\geq 0, \tag{25}$$

where equality holds in (24) 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 (22) is given by

$$c_f(\xi_1, \xi_2) = \tfrac{1}{2} \inf_{t \in \mathcal{I}(\xi_1, \xi_2)} f''(t). \tag{26}$$

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

$$D_f(P_X \| Q_X) - D_f(P_Y \| Q_Y) \geq c_{f^*}\Big(\tfrac{1}{\xi_2}, \tfrac{1}{\xi_1}\Big) \left[\chi^2(Q_X \| P_X) - \chi^2(Q_Y \| P_Y) \right] \tag{27}$$

$$\geq 0, \tag{28}$$

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

$$f^*(t) := t\, f\left(\frac{1}{t}\right), \quad \forall\, t > 0, \tag{29}$$

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

$$c_{f^*}\Big(\tfrac{1}{\xi_2}, \tfrac{1}{\xi_1}\Big) = \tfrac{1}{2} \inf_{t \in \mathcal{I}(\xi_1, \xi_2)} \{ t^3 f''(t) \} \tag{30}$$

with the convention that $\frac{1}{\xi_1} = \infty$ if $\xi_1 = 0$. Equality holds in (27) 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

$$e_f(\xi_1, \xi_2) := \tfrac{1}{2} \sup_{t \in \mathcal{I}(\xi_1, \xi_2)} f''(t) < \infty, \tag{31}$$

then,

$$D_f(P_X \| Q_X) - D_f(P_Y \| Q_Y) \leq e_f(\xi_1, \xi_2) \left[\chi^2(P_X \| Q_X) - \chi^2(P_Y \| Q_Y) \right]. \tag{32}$$

Furthermore,

$$D_f(P_X \| Q_X) - D_f(P_Y \| Q_Y) \leq e_{f^*}\Big(\tfrac{1}{\xi_2}, \tfrac{1}{\xi_1}\Big) \left[\chi^2(Q_X \| P_X) - \chi^2(Q_Y \| P_Y) \right] \tag{33}$$

where the coefficient in the right side of (33) satisfies

$$e_{f^*}\Big(\tfrac{1}{\xi_2}, \tfrac{1}{\xi_1}\Big) = \tfrac{1}{2} \sup_{t \in \mathcal{I}(\xi_1, \xi_2)} \{ t^3 f''(t) \}, \tag{34}$$

which is assumed to be finite. Equalities hold in (32) and (33) 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* (24), (27), (32) *and* (33) *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* (20) *and* (21) *with inputs* $P_X^{(n)}$ *and* Q_X, *respectively. Suppose that*

$$\lim_{n\to\infty} \inf_{x\in\mathcal{X}} \frac{P_X^{(n)}(x)}{Q_X(x)} = 1, \tag{35}$$

$$\lim_{n\to\infty} \sup_{x\in\mathcal{X}} \frac{P_X^{(n)}(x)}{Q_X(x)} = 1. \tag{36}$$

If f *has a continuous second derivative at unity, then*

$$\lim_{n\to\infty} \frac{D_f(P_X^{(n)}\|Q_X) - D_f(P_Y^{(n)}\|Q_Y)}{\chi^2(P_X^{(n)}\|Q_X) - \chi^2(P_Y^{(n)}\|Q_Y)} = \tfrac{1}{2}f''(1), \tag{37}$$

$$\lim_{n\to\infty} \frac{D_f(P_X^{(n)}\|Q_X) - D_f(P_Y^{(n)}\|Q_Y)}{\chi^2(Q_X\|P_X^{(n)}) - \chi^2(Q_Y\|P_Y^{(n)})} = \tfrac{1}{2}f''(1), \tag{38}$$

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

Proof. See Appendix A. □

An application of Theorem 1 gives the following result.

Theorem 2. *Let* $\mathcal{X}$ *and* $\mathcal{Y}$ *be finite or countably infinite sets, let* $n \in \mathbb{N}$, *and let* $X^n := (X_1, \ldots, X_n)$ *and* $Y^n := (Y_1, \ldots, Y_n)$ *be random vectors taking values on* $\mathcal{X}^n$ *and* $\mathcal{Y}^n$, *respectively. Let* P_{X^n} *and* Q_{X^n} *be the probability mass functions of discrete memoryless sources where, for all* $\underline{x} \in \mathcal{X}^n$,

$$P_{X^n}(\underline{x}) = \prod_{i=1}^n P_{X_i}(x_i), \quad Q_{X^n}(\underline{x}) = \prod_{i=1}^n Q_{X_i}(x_i), \tag{39}$$

with P_{X_i} *and* Q_{X_i} *supported on* $\mathcal{X}$ *for all* $i \in \{1, \ldots, n\}$. *Let each symbol* X_i *be independently selected from one of the source outputs at time instant* i *with probabilities* λ *and* $1-\lambda$, *respectively, and let it be transmitted over a discrete memoryless channel with transition probabilities*

$$W_{Y^n|X^n}(\underline{y}\,|\,\underline{x}) = \prod_{i=1}^n W_{Y_i|X_i}(y_i|x_i), \quad \forall\, \underline{x} \in \mathcal{X}^n,\ \underline{y} \in \mathcal{Y}^n. \tag{40}$$

Let $R_{X^n}^{(\lambda)}$ *be the probability mass function of the symbols at the channel input, i.e.,*

$$R_{X^n}^{(\lambda)}(\underline{x}) = \prod_{i=1}^n \bigl(\lambda P_{X_i}(x_i) + (1-\lambda) Q_{X_i}(x_i)\bigr), \quad \forall\, \underline{x} \in \mathcal{X}^n,\ \lambda \in [0,1], \tag{41}$$

let

$$R_{Y^n}^{(\lambda)} := R_{X^n}^{(\lambda)} W_{Y^n|X^n}, \tag{42}$$

$$P_{Y^n} := P_{X^n} W_{Y^n|X^n}, \tag{43}$$

$$Q_{Y^n} := Q_{X^n} W_{Y^n|X^n}, \tag{44}$$

and let $f\colon (0,\infty) \to \mathbb{R}$ *be a convex and twice differentiable function with* $f(1) = 0$. *Then,*

(a) *For all* $\lambda \in [0,1]$,

$$D_f(R^{(\lambda)}_{X^n} \,\|\, Q_{X^n}) - D_f(R^{(\lambda)}_{Y^n} \,\|\, Q_{Y^n})$$
$$\ge c_f\big(\xi_1(n,\lambda),\, \xi_2(n,\lambda)\big) \left[\prod_{i=1}^n \big(1+\lambda^2\,\chi^2(P_{X_i}\|Q_{X_i})\big) - \prod_{i=1}^n \big(1+\lambda^2\,\chi^2(P_{Y_i}\|Q_{Y_i})\big)\right] \tag{45}$$
$$\ge c_f\big(\xi_1(n,\lambda),\, \xi_2(n,\lambda)\big)\,\lambda^2 \sum_{i=1}^n \big[\chi^2(P_{X_i}\|Q_{X_i}) - \chi^2(P_{Y_i}\|Q_{Y_i})\big] \ge 0, \tag{46}$$

where $c_f(\cdot,\cdot)$ *in the right sides of* (45) *and* (46) *is given in* (26), *and*

$$\xi_1(n,\lambda) := \prod_{i=1}^n \left(1-\lambda+\lambda \inf_{x\in\mathcal{X}} \frac{P_{X_i}(x)}{Q_{X_i}(x)}\right) \in [0,1], \tag{47}$$
$$\xi_2(n,\lambda) := \prod_{i=1}^n \left(1-\lambda+\lambda \sup_{x\in\mathcal{X}} \frac{P_{X_i}(x)}{Q_{X_i}(x)}\right) \in [1,\infty]. \tag{48}$$

(b) *For all* $\lambda \in [0,1]$,

$$D_f(R^{(\lambda)}_{X^n} \,\|\, Q_{X^n}) - D_f(R^{(\lambda)}_{Y^n} \,\|\, Q_{Y^n})$$
$$\le e_f\big(\xi_1(n,\lambda),\, \xi_2(n,\lambda)\big) \left[\prod_{i=1}^n \big(1+\lambda^2\,\chi^2(P_{X_i}\|Q_{X_i})\big) - \prod_{i=1}^n \big(1+\lambda^2\,\chi^2(P_{Y_i}\|Q_{Y_i})\big)\right] \tag{49}$$

where $e_f(\cdot,\cdot)$, $\xi_1(\cdot,\cdot)$ *and* $\xi_2(\cdot,\cdot)$ *in the right side of* (49) *are given in* (31), (47) *and* (48), *respectively.*

(c) *If* f *has a continuous second derivative at unity, and* $\sup\limits_{x\in\mathcal{X}} \frac{P_{X_i}(x)}{Q_{X_i}(x)} < \infty$ *for all* $i \in \{1,\dots,n\}$, *then*

$$\lim_{\lambda\to 0^+} \frac{D_f(R^{(\lambda)}_{X^n} \,\|\, Q_{X^n}) - D_f(R^{(\lambda)}_{Y^n} \,\|\, Q_{Y^n})}{\lambda^2}$$
$$= \tfrac{1}{2} f''(1) \sum_{i=1}^n \big[\chi^2(P_{X_i}\|Q_{X_i}) - \chi^2(P_{Y_i}\|Q_{Y_i})\big]. \tag{50}$$

The lower bounds in the right sides of (45) *and* (46), *and the upper bound in the right side of* (49) *are tight as we let* $\lambda \to 0^+$, *yielding the limit in the right side of* (50).

Proof. See Appendix B. □

Remark 2. *Similar upper and lower bounds on* $D_f(P_{X^n} \,\|\, R^{(\lambda)}_{X^n}) - D_f(P_{Y^n} \,\|\, R^{(\lambda)}_{Y^n})$ *can be obtained for all* $\lambda \in [0,1]$. *To that end, in* (45)–(49), *one needs to replace* f *with* f^*, *switch between* P_{X_i} *and* Q_{X_i} *for all* i, *and replace* λ *with* $1-\lambda$.

In continuation to ([26], Theorem 8) (see Proposition 3 in Section 1.1), we next provide an upper bound on the contraction coefficient for a subclass of f-divergences (this subclass is different from the one which is addressed in ([26], Theorem 8)). 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 Remarks A1–A3.

Theorem 3. *Let* $f\colon (0,\infty) \to \mathbb{R}$ *be a function which satisfies the following conditions:*

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

Let P_X and Q_X be non-identical probability mass functions which are defined on a finite or a countably infinite set $\mathcal{X}$, and let

$$\kappa(\xi_1, \xi_2) := \sup_{t \in (\xi_1, 1) \cup (1, \xi_2)} \frac{f(t) + f'(1)\,(1-t)}{(t-1)^2} \tag{51}$$

where $\xi_1 \in [0,1)$ and $\xi_2 \in (1, \infty]$ are given in (18) and (19). Then, in the setting of (20) and (21),

$$\frac{D_f(P_Y \| Q_Y)}{D_f(P_X \| Q_X)} \leq \frac{\kappa(\xi_1, \xi_2)}{f(0) + f'(1)} \cdot \frac{\chi^2(P_Y \| Q_Y)}{\chi^2(P_X \| Q_X)}. \tag{52}$$

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

$$\mu_f(Q_X, W_{Y|X}) \leq \frac{1}{f(0) + f'(1)} \cdot \kappa\Biggl(0, \frac{1}{\min\limits_{x \in \mathcal{X}} Q_X(x)}\Biggr) \cdot \mu_{\chi^2}(Q_X, W_{Y|X}). \tag{53}$$

Proof. See Appendix C.1. □

Similarly to the extension of Theorem 1 to Theorem 2, a similar extension of Theorem 3 leads to the following result.

Theorem 4. *In the setting of (39)–(44) in Theorem 2, and under the assumptions on f in Theorem 3, the following holds for all $\lambda \in (0,1]$ and $n \in \mathbb{N}$:*

$$\frac{D_f\bigl(R^{(\lambda)}_{Y^n} \,\|\, Q_{Y^n}\bigr)}{D_f\bigl(R^{(\lambda)}_{X^n} \,\|\, Q_{X^n}\bigr)} \leq \frac{\kappa\bigl(\xi_1(n,\lambda),\, \xi_2(n,\lambda)\bigr)}{f(0) + f'(1)} \; \frac{\prod\limits_{i=1}^{n} \Bigl(1 + \lambda^2\, \chi^2(P_{Y_i} \,\|\, Q_{Y_i})\Bigr) - 1}{\prod\limits_{i=1}^{n} \Bigl(1 + \lambda^2\, \chi^2(P_{X_i} \,\|\, Q_{X_i})\Bigr) - 1}, \tag{54}$$

with $\xi_1(n,\lambda)$ and $\xi_2(n,\lambda)$ and $\kappa(\cdot,\cdot)$ defined in (47), (48) and (51), respectively.

Proof. See Appendix C.2. □

2.2. A Subclass of f-Divergences

A subclass of f-divergences with interesting properties is introduced in Theorem 5. The data-processing inequalities in Theorems 2 and 4 are applied to these f-divergences in Section 3.

Theorem 5. *Let $f_\alpha \colon [0,\infty) \to \mathbb{R}$ be given by*

$$f_\alpha(t) := (\alpha + t)^2 \log(\alpha + t) - (\alpha + 1)^2 \log(\alpha + 1), \quad t \geq 0 \tag{55}$$

for all $\alpha \geq \mathrm{e}^{-\frac{3}{2}}$. Then,

(a) $D_{f_\alpha}(\cdot \| \cdot)$ is an f-divergence which is monotonically increasing and concave in α, and its first three derivatives are related to the relative entropy and χ^2-divergence as follows:

$$\frac{\partial}{\partial \alpha} \bigl\{ D_{f_\alpha}(P \| Q) \bigr\} = 2(\alpha + 1)\, D\Bigl(\tfrac{\alpha Q + P}{\alpha + 1} \,\|\, Q\Bigr), \tag{56}$$

$$\frac{\partial^2}{\partial \alpha^2} \bigl\{ D_{f_\alpha}(P \| Q) \bigr\} = -2\, D\Bigl(Q \,\|\, \tfrac{\alpha Q + P}{\alpha + 1}\Bigr), \tag{57}$$

$$\frac{\partial^3}{\partial \alpha^3} \bigl\{ D_{f_\alpha}(P \| Q) \bigr\} = \frac{2 \log \mathrm{e}}{\alpha + 1} \cdot \chi^2\Bigl(Q \,\|\, \tfrac{\alpha Q + P}{\alpha + 1}\Bigr). \tag{58}$$

(b) *For every $n \in \mathbb{N}$,*

$$(-1)^{n-1} \frac{\partial^n}{\partial \alpha^n} \{ D_{f_\alpha}(P\|Q) \} \geq 0, \tag{59}$$

and, in addition to (56)–(58), for all $n > 3$

$$\frac{\partial^n}{\partial \alpha^n} \{ D_{f_\alpha}(P\|Q) \} = \frac{2(-1)^{n-1}(n-3)! \log \mathrm{e}}{(\alpha+1)^{n-2}} \left[\exp\Big((n-2)\, D_{n-1}\big(Q \,\|\, \tfrac{\alpha Q+P}{\alpha+1}\big) \Big) - 1 \right], \tag{60}$$

where $D_{n-1}(\cdot\|\cdot)$ in the right side of (60) denotes the Rényi divergence of order $n-1$.

(c)

$$\begin{aligned} D_{f_\alpha}(P\|Q) &\geq k(\alpha)\, \chi^2(P\|Q) && (61) \\ &\geq k(\alpha)\, \big[\exp\big(D(P\|Q)\big) - 1 \big] && (62) \end{aligned}$$

where the function $k \colon [\mathrm{e}^{-\frac{3}{2}}, \infty) \to \mathbb{R}$ is defined as

$$k(\alpha) := \log(\alpha+1) + \tfrac{3}{2} \log \mathrm{e} - \frac{\log \mathrm{e}}{3\alpha}, \tag{63}$$

which is monotonically increasing in α, satisfying $k(\alpha) \geq 0.2075 \log \mathrm{e}$ for all $\alpha \geq \mathrm{e}^{-\frac{3}{2}}$, and it tends to infinity as we let $\alpha \to \infty$. Consequently, unless $P \equiv Q$,

$$\lim_{\alpha \to \infty} D_{f_\alpha}(P\|Q) = +\infty. \tag{64}$$

(d)

$$D_{f_\alpha}(P\|Q) \leq \left[\log(\alpha+1) + \tfrac{3}{2} \log \mathrm{e} - \frac{\log \mathrm{e}}{\alpha+1} \right] \chi^2(P\|Q) + \frac{\log \mathrm{e}}{3(\alpha+1)} \left[\exp\big(2 D_3(P\|Q)\big) - 1 \right]. \tag{65}$$

(e) *For every $\varepsilon > 0$ and a pair of probability mass functions (P, Q) where $D_3(P\|Q) < \infty$, there exists $\alpha^* := \alpha(P, Q, \varepsilon)$ such that for all $\alpha > \alpha^*$*

$$\left| D_{f_\alpha}(P\|Q) - \big[\log(\alpha+1) + \tfrac{3}{2} \log \mathrm{e} \big]\, \chi^2(P\|Q) \right| < \varepsilon. \tag{66}$$

(f) *If a sequence of probability measures $\{P_n\}$ converges to a probability measure Q such that*

$$\lim_{n \to \infty} \operatorname{ess\,sup} \frac{dP_n}{dQ}\,(Y) = 1, \quad Y \sim Q, \tag{67}$$

where $P_n \ll Q$ for all sufficiently large n, then

$$\lim_{n \to \infty} \frac{D_{f_\alpha}(P_n\|Q)}{\chi^2(P_n\|Q)} = \log(\alpha+1) + \tfrac{3}{2} \log \mathrm{e}. \tag{68}$$

(g) *If $\alpha > \beta \geq \mathrm{e}^{-\frac{3}{2}}$, then*

$$\begin{aligned} 0 &\leq (\alpha-\beta)(\alpha+\beta+2)\, D\big(\tfrac{\alpha Q+P}{\alpha+1} \,\|\, Q \big) && (69) \\ &\leq D_{f_\alpha}(P\|Q) - D_{f_\beta}(P\|Q) && (70) \\ &\leq (\alpha-\beta)\, \min\Big\{ (\alpha+\beta+2)\, D\big(\tfrac{\beta Q+P}{\beta+1} \,\|\, Q \big),\; 2D(P\|Q) \Big\}. && (71) \end{aligned}$$

(h) *The function $f_\alpha \colon [0,\infty) \to \mathbb{R}$, as given in (55), satisfies the conditions in Theorems 3 and 4 for all $\alpha \geq \mathrm{e}^{-\frac{3}{2}}$. Furthermore, the corresponding function in (51) is equal to*

$$\kappa_\alpha(\xi_1,\xi_2) := \sup_{t\in(\xi_1,1)\cup(1,\xi_2)} \frac{f_\alpha(t)+f'_\alpha(1)\,(1-t)}{(t-1)^2} \tag{72}$$

$$= \frac{f_\alpha(\xi_2)+f'_\alpha(1)\,(1-\xi_2)}{(\xi_2-1)^2} \tag{73}$$

for all $\xi_1 \in [0,1)$ and $\xi_2 \in (1,\infty)$.

Proof. See Appendix D. □

2.3. f-Divergence Inequalities via Majorization

Let U_n denote an equiprobable probability mass function on $\{1,\dots,n\}$ for an arbitrary $n \in \mathbb{N}$, 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 ([38], Lemma 1)).

Theorem 6. *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)

$$n e_f(nq_{\min}, nq_{\max}) \left(\|Q\|_2^2 - \|P\|_2^2\right)$$

$$\geq D_f(Q\|U_n) - D_f(P\|U_n) \tag{74}$$

$$\geq n c_f(nq_{\min}, nq_{\max}) \left(\|Q\|_2^2 - \|P\|_2^2\right) \geq 0, \tag{75}$$

where $c_f(\cdot,\cdot)$ and $e_f(\cdot,\cdot)$ are given in (26) and (31), respectively, and $\|\cdot\|_2$ denotes the Euclidean norm. Furthermore, (74) and (75) hold with equality if $D_f(\cdot\|\cdot) = \chi^2(\cdot\|\cdot)$.

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

$$0 \leq \|Q\|_2^2 - \|P\|_2^2 \leq \frac{(\rho-1)^2}{4\rho n}. \tag{76}$$

Proof. See Appendix E. □

Remark 3. *If P is not supported on $\{1,\dots,n\}$, then (74) and (75) hold if f is also right continuous at zero.*

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 6.

Theorem 7. *Let $\mathcal{P}_n$ denote the set of all the probability mass functions that are defined on $\mathcal{A}_n := \{1,\dots,n\}$. For $\rho \geq 1$, let $\mathcal{P}_n(\rho)$ be the set of all $Q \in \mathcal{P}_n$ which are supported on $\mathcal{A}_n$ with $\frac{q_{\max}}{q_{\min}} \leq \rho$, 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 $\rho \geq 1$, 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 $\rho \geq 1$ and an integer $n \geq 2$, let*

$$u_f(n,\rho) := \max_{Q \in \mathcal{P}_n(\rho)} D_f(Q \| U_n), \tag{77}$$

$$v_f(n,\rho) := \max_{Q \in \mathcal{P}_n(\rho)} D_f(U_n \| Q), \tag{78}$$

let

$$\Gamma_n(\rho) := \left[\frac{1}{1+(n-1)\rho}, \frac{1}{n} \right], \tag{79}$$

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

$$Q_\beta(j) := \begin{cases} \rho\beta, & \text{if } j \in \{1, \ldots, i_\beta\}, \\ 1 - \bigl(n + i_\beta(\rho-1) - 1\bigr)\beta, & \text{if } j = i_\beta + 1, \\ \beta, & \text{if } j \in \{i_\beta + 2, \ldots, n\} \end{cases} \tag{80}$$

where

$$i_\beta := \left\lfloor \frac{1-n\beta}{(\rho-1)\beta} \right\rfloor. \tag{81}$$

Then,

$$u_f(n,\rho) = \max_{\beta \in \Gamma_n(\rho)} D_f(Q_\beta \| U_n), \tag{82}$$

$$v_f(n,\rho) = \max_{\beta \in \Gamma_n(\rho)} D_f(U_n \| Q_\beta). \tag{83}$$

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

$$g_f^{(\rho)}(x) := x\, f\left(\frac{\rho}{1+(\rho-1)x} \right) + (1-x)\, f\left(\frac{1}{1+(\rho-1)x} \right), \quad x \in [0,1]. \tag{84}$$

Then,

$$\max_{m \in \{0,\ldots,n\}} g_f^{(\rho)}\left(\tfrac{m}{n}\right) \leq u_f(n,\rho) \leq \max_{x \in [0,1]} g_f^{(\rho)}(x), \tag{85}$$

$$\max_{m \in \{0,\ldots,n\}} g_{f^*}^{(\rho)}\left(\tfrac{m}{n}\right) \leq v_f(n,\rho) \leq \max_{x \in [0,1]} g_{f^*}^{(\rho)}(x) \tag{86}$$

with the convex function $f^* \colon (0,\infty) \to \mathbb{R}$ in (29).

(e) *The right-side inequalities in (85) and (86) are asymptotically tight ($n \to \infty$). More explicitly,*

$$\lim_{n\to\infty} u_f(n,\rho) = \max_{x \in [0,1]} \left\{ x\, f\left(\frac{\rho}{1+(\rho-1)x} \right) + (1-x)\, f\left(\frac{1}{1+(\rho-1)x} \right) \right\}, \tag{87}$$

$$\lim_{n\to\infty} v_f(n,\rho) = \max_{x \in [0,1]} \left\{ \frac{\rho x}{1+(\rho-1)x}\ f\left(\frac{1+(\rho-1)x}{\rho} \right) + \frac{(1-x)\, f\bigl(1+(\rho-1)x\bigr)}{1+(\rho-1)x} \right\}. \tag{88}$$

(f) *If $g_f^{(\rho)}(\cdot)$ in (84) is differentiable on $(0,1)$ and its derivative is upper bounded by $K_f(\rho) \geq 0$, then for every integer $n \geq 2$*

$$0 \leq \lim_{n' \to \infty} \{u_f(n', \rho)\} - u_f(n, \rho) \leq \frac{K_f(\rho)}{n}. \tag{89}$$

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

$$\lim_{\rho \to \infty} u_f(n, \rho) = \left(1 - \frac{1}{n}\right) f(0) + \frac{f(n)}{n}. \tag{90}$$

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

$$0 \leq \lim_{\rho' \to \infty} \{u_f(n, \rho')\} - u_f(n, \rho) \leq \frac{2K_n\,(n-1)}{n + \rho - 1}. \tag{91}$$

(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:*

$$0 \leq m \leq f''(t) \leq M, \quad \forall\, t \in \left[\tfrac{1}{\rho}, \rho\right]. \tag{92}$$

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

$$\begin{aligned} 0 &\leq \tfrac{1}{2} m \left(n \|Q\|_2^2 - 1\right) && (93)\\ &\leq D_f(Q \| U_n) && (94)\\ &\leq \tfrac{1}{2} M \left(n \|Q\|_2^2 - 1\right) && (95)\\ &\leq \frac{M(\rho - 1)^2}{8\rho} && (96)\end{aligned}$$

with equalities in (94) and (95) 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*

$$\rho \leq 1 + \frac{4d}{M_f} + \sqrt{\frac{8d}{M_f} + \frac{16d^2}{M_f^2}}. \tag{97}$$

Proof. See Appendix F. □

Tsallis entropy was introduced in [39] as a generalization of the Shannon entropy (similarly to the Rényi entropy [40]), and it was applied to statistical physics in [39].

Definition 7 ([39]). *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 defined as*

$$\begin{aligned} S_\alpha(X) &= \frac{1}{1-\alpha} \left(\sum_{x \in \mathcal{X}} P_X^\alpha(x) - 1 \right) && (98)\\ &= \frac{\|P_X\|_\alpha^\alpha - 1}{1 - \alpha}, && (99)\end{aligned}$$

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 on base* e *(expressed in nats).*

Theorem 6 enables to strengthen the Schur-concavity property of the Tsallis entropy (see ([30], Theorem 13.F.3.a.)) as follows.

Theorem 8. *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)

$$0 \leq L(\alpha, P, Q) \leq S_\alpha(P) - S_\alpha(Q) \leq U(\alpha, P, Q), \tag{100}$$

where

$$L(\alpha, P, Q) := \begin{cases} \frac{1}{2}\, \alpha q_{\max}^{\alpha-2} \left(\|Q\|_2^2 - \|P\|_2^2 \right), & \text{if } \alpha \in (0, 2], \\ \frac{1}{2}\, \alpha q_{\min}^{\alpha-2} \left(\|Q\|_2^2 - \|P\|_2^2 \right), & \text{if } \alpha \in (2, \infty), \end{cases} \tag{101}$$

$$U(\alpha, P, Q) := \begin{cases} \frac{1}{2}\, \alpha q_{\min}^{\alpha-2} \left(\|Q\|_2^2 - \|P\|_2^2 \right), & \text{if } \alpha \in (0, 2], \\ \frac{1}{2}\, \alpha q_{\max}^{\alpha-2} \left(\|Q\|_2^2 - \|P\|_2^2 \right), & \text{if } \alpha \in (2, \infty), \end{cases} \tag{102}$$

and the bounds in (101) *and* (102) *are attained at $\alpha = 2$.*

(b)

$$\inf_{P \prec Q,\, P \neq Q} \frac{S_\alpha(P) - S_\alpha(Q)}{L(\alpha, P, Q)} = \sup_{P \prec Q,\, P \neq Q} \frac{S_\alpha(P) - S_\alpha(Q)}{U(\alpha, P, Q)} = 1, \tag{103}$$

where the infimum and supremum in (103) *can be restricted to probability mass functions P and Q which are supported on a binary alphabet.*

Proof. See Appendix G. □

Remark 4. *The lower bound in ([36], Theorem 1) also strengthens the Schur-concavity property of the Tsallis entropy. It can be verified that none of the lower bounds in ([36], Theorem 1) and Theorem 8 supersedes the other. For example, let $\alpha > 0$, and let P_ε and Q_ε be probability mass functions supported on $\mathcal{A} := \{0, 1\}$ with $P_\varepsilon(0) = \frac{1}{2} + \varepsilon$ and $Q_\varepsilon(0) = \frac{1}{2} + \beta\varepsilon$ where $\beta > 1$ and $0 < \varepsilon < \frac{1}{2\beta}$. This yields $P_\varepsilon \prec Q_\varepsilon$. From* (A233) *(see Appendix G),*

$$\lim_{\varepsilon \to 0^+} \frac{S_\alpha(P_\varepsilon) - S_\alpha(Q_\varepsilon)}{L(\alpha, P_\varepsilon, Q_\varepsilon)} = 1. \tag{104}$$

If $\alpha = 1$, then $S_1(P_\varepsilon) - S_1(Q_\varepsilon) = \frac{1}{\log e} \big(H(P_\varepsilon) - H(Q_\varepsilon) \big)$, and the continuous extension of the lower bound in ([36], Theorem 1) at $\alpha = 1$ is specialized to the earlier result by the same authors in ([35], Theorem 3); it states that if $P \prec Q$, then $H(P) - H(Q) \geq D(Q \| P)$. In contrast to (104)*, it can be verified that*

$$\lim_{\varepsilon \to 0^+} \frac{S_1(P_\varepsilon) - S_1(Q_\varepsilon)}{\frac{1}{\log e} D(Q_\varepsilon \| P_\varepsilon)} = \frac{\beta + 1}{\beta - 1} > 1, \quad \forall\, \beta > 1, \tag{105}$$

which can be made arbitrarily large by selecting β to be sufficiently close to 1 (from above). This provides a case where the lower bound in Theorem 8 outperforms the one in ([35], Theorem 3).

Remark 5. *Due to the one-to-one correspondence between Tsallis and Rényi entropies of the same positive order, similar to the transition from ([36], Theorem 1) to ([36], Theorem 2), also Theorem 8 enables to strengthen the Schur-concavity property of the Rényi entropy. For information-theoretic implications of the Schur-concavity of the Rényi entropy, the reader is referred to, e.g., [34], ([41], Theorem 3) and ([42], Theorem 11).*

3. Illustration of the Main Results and Implications

3.1. Illustration of Theorems 2 and 4

We apply here the data-processing inequalities in Theorems 2 and 4 to the new class of f-divergences introduced in Theorem 5.

In the setup of Theorems 2 and 4, consider communication over a time-varying binary-symmetric channel (BSC). Consequently, let $\mathcal{X} = \mathcal{Y} = \{0, 1\}$, and let

$$P_{X_i}(1) = p_i, \quad Q_{X_i}(1) = q_i, \tag{106}$$

with $p_i \in (0, 1)$ and $q_i \in (0, 1)$ for every $i \in \{1, \dots, n\}$. Let the transition probabilities $P_{Y_i|X_i}(\cdot|\cdot)$ correspond to $\text{BSC}(\delta_i)$ (i.e., a BSC with a crossover probability δ_i), i.e.,

$$P_{Y_i|X_i}(y|x) = \begin{cases} 1 - \delta_i & \text{if } x = y, \\ \delta_i & \text{if } x \neq y. \end{cases} \tag{107}$$

For all $\lambda \in [0, 1]$ and $\underline{x} \in \mathcal{X}^n$, the probability mass function at the channel input is given by

$$R^{(\lambda)}_{X^n}(\underline{x}) = \prod_{i=1}^{n} R^{(\lambda)}_{X_i}(x_i), \tag{108}$$

with

$$R^{(\lambda)}_{X_i}(x) = \lambda P_{X_i}(x) + (1 - \lambda) Q_{X_i}(x), \quad x \in \{0, 1\}, \tag{109}$$

where the probability mass function in (109) refers to a Bernoulli distribution with parameter $\lambda p_i + (1 - \lambda) q_i$. At the output of the time-varying BSC (see (42)–(44) and (107)), for all $\underline{y} \in \mathcal{Y}^n$,

$$R^{(\lambda)}_{Y^n}(\underline{y}) = \prod_{i=1}^{n} R^{(\lambda)}_{Y_i}(y_i), \quad P_{Y^n}(\underline{y}) = \prod_{i=1}^{n} P_{Y_i}(y_i), \quad Q_{Y^n}(\underline{y}) = \prod_{i=1}^{n} Q_{Y_i}(y_i), \tag{110}$$

where

$$R^{(\lambda)}_{Y_i}(1) = \bigl(\lambda p_i + (1 - \lambda) q_i\bigr) * \delta_i, \tag{111}$$

$$P_{Y_i}(1) = p_i * \delta_i, \tag{112}$$

$$Q_{Y_i}(1) = q_i * \delta_i, \tag{113}$$

with

$$a * b := a(1 - b) + (1 - a) b, \quad 0 \leq a, b \leq 1. \tag{114}$$

The χ^2-divergence from $\text{Bernoulli}(p)$ to $\text{Bernoulli}(q)$ is given by

$$\chi^2(\text{Bernoulli}(p) \,\|\, \text{Bernoulli}(q)) = \frac{(p - q)^2}{q(1 - q)}, \tag{115}$$

and since the probability mass functions P_{X_i}, Q_{X_i}, P_{Y_i} and Q_{Y_i} correspond to Bernoulli distributions with parameters p_i, q_i, $p_i * \delta_i$ and $q_i * \delta_i$, respectively, Theorem 2 gives that

$$c_{f_\alpha}\big(\xi_1(n,\lambda),\, \xi_2(n,\lambda)\big) \left[\prod_{i=1}^{n} \left(1 + \frac{\lambda^2 (p_i - q_i)^2}{q_i(1-q_i)}\right) - \prod_{i=1}^{n} \left(1 + \frac{\lambda^2 (p_i * \delta_i - q_i * \delta_i)^2}{(q_i * \delta_i)(1 - q_i * \delta_i)}\right) \right]$$

$$\leq D_{f_\alpha}(R^{(\lambda)}_{X^n} \,\|\, Q_{X^n}) - D_{f_\alpha}(R^{(\lambda)}_{Y^n} \,\|\, Q_{Y^n}) \tag{116}$$

$$\leq e_{f_\alpha}\big(\xi_1(n,\lambda),\, \xi_2(n,\lambda)\big) \left[\prod_{i=1}^{n} \left(1 + \frac{\lambda^2 (p_i - q_i)^2}{q_i(1-q_i)}\right) - \prod_{i=1}^{n} \left(1 + \frac{\lambda^2 (p_i * \delta_i - q_i * \delta_i)^2}{(q_i * \delta_i)(1 - q_i * \delta_i)}\right) \right] \tag{117}$$

for all $\lambda \in [0,1]$ and $n \in \mathbb{N}$. From (26), (31) and (55), we get that for all $\xi_1 < 1 < \xi_2$,

$$c_{f_\alpha}(\xi_1, \xi_2) = \tfrac{1}{2} \inf_{t \in [\xi_1, \xi_2]} f''_\alpha(t) \tag{118}$$

$$= \log(\alpha + \xi_1) + \tfrac{3}{2} \log \mathrm{e}, \tag{119}$$

$$e_{f_\alpha}(\xi_1, \xi_2) = \tfrac{1}{2} \sup_{t \in [\xi_1, \xi_2]} f''_\alpha(t) \tag{120}$$

$$= \log(\alpha + \xi_2) + \tfrac{3}{2} \log \mathrm{e}, \tag{121}$$

and, from (47), (48) and (106), for all $\lambda \in (0,1]$,

$$\xi_1(n,\lambda) := \prod_{i=1}^{n} \left(1 - \lambda + \lambda \min\left\{\frac{p_i}{q_i}, \frac{1-p_i}{1-q_i}\right\}\right) \in [0,1), \tag{122}$$

$$\xi_2(n,\lambda) := \prod_{i=1}^{n} \left(1 - \lambda + \lambda \max\left\{\frac{p_i}{q_i}, \frac{1-p_i}{1-q_i}\right\}\right) \in (1,\infty), \tag{123}$$

provided that $p_i \neq q_i$ for some $i \in \{1, \ldots, n\}$ (otherwise, both f-divergences in the right side of (116) are equal to zero since $P_{X_i} \equiv Q_{X_i}$ and therefore $R^{(\lambda)}_{X_i} \equiv Q_{X_i}$ for all i and $\lambda \in [0,1]$). Furthermore, from Item (c) of Theorem 2, for every $n \in \mathbb{N}$ and $\alpha \geq \mathrm{e}^{-\frac{3}{2}}$,

$$\lim_{\lambda \to 0^+} \frac{D_{f_\alpha}(R^{(\lambda)}_{X^n} \,\|\, Q_{X^n}) - D_{f_\alpha}(R^{(\lambda)}_{Y^n} \,\|\, Q_{Y^n})}{\lambda^2}$$

$$= \left(\log(\alpha+1) + \tfrac{3}{2} \log \mathrm{e}\right) \sum_{i=1}^{n} \left\{ \frac{(p_i - q_i)^2}{q_i(1-q_i)} - \frac{(p_i * \delta_i - q_i * \delta_i)^2}{(q_i * \delta_i)(1 - q_i * \delta_i)} \right\}, \tag{124}$$

and the lower and upper bounds in the left side of (116) and the right side of (117), respectively, are tight as we let $\lambda \to 0$, and they both coincide with the limit in the right side of (124).

Figure 1 illustrates the upper and lower bounds in (116) and (117) with $\alpha = 1$, $p_i \equiv \frac{1}{4}$, $q_i \equiv \frac{1}{2}$ and $\delta_i \equiv 0.110$ for all i, and $n \in \{1, 10, 50\}$. In the special case where $\{\delta_i\}$ are fixed for all i, the communication channel is a time-invariant BSC whose capacity is equal to $\frac{1}{2}$ bit per channel use.

By referring to the upper and middle plots of Figure 1, if $n = 1$ or $n = 10$, then the exact values of the differences of the f_α-divergences in the right side of (116) are calculated numerically, being compared to the lower and upper bounds in the left side of (116) and the right side of (117) respectively. Since the f_α-divergence does not tensorize, the computation of the exact value of each of the two f_α-divergences in the right side of (116) involves a pre-computation of 2^n probabilities for each of the probability mass functions P_{X^n}, Q_{X^n}, P_{Y^n} and Q_{Y^n}; this computation is prohibitively complex unless n is small enough.

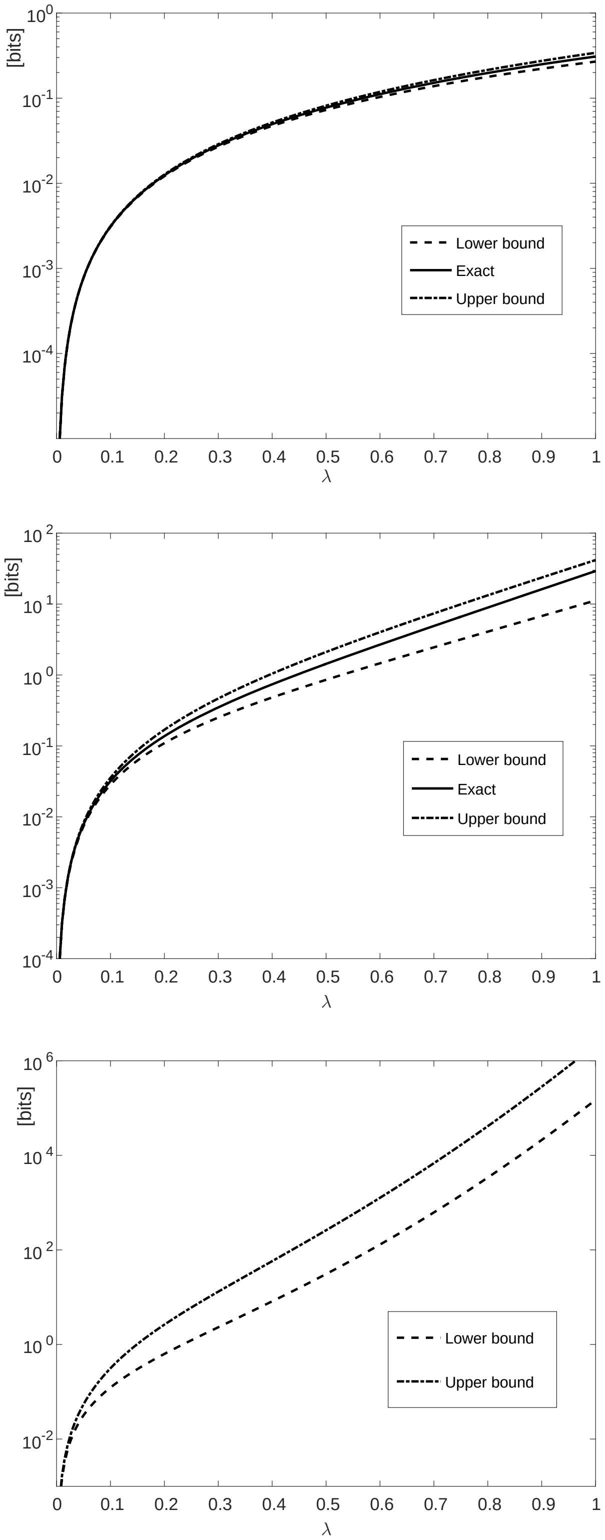

Figure 1. The bounds in Theorem 2 applied to $D_{f_\alpha}(R^{(\lambda)}_{X^n} \,\|\, Q_{X^n}) - D_{f_\alpha}(R^{(\lambda)}_{Y^n} \,\|\, Q_{Y^n})$ (vertical axis) versus $\lambda \in [0,1]$ (horizontal axis). The f_α-divergence refers to Theorem 5. The probability mass functions P_{X^n} and Q_{X^n} correspond, respectively, to discrete memoryless sources emitting n i.i.d. Bernoulli(p) and Bernoulli(q) symbols; the symbols are transmitted over BSC(δ) with $(\alpha, p, q, \delta) = \bigl(1, \frac{1}{4}, \frac{1}{2}, 0.110\bigr)$. The bounds in the upper and middle plots are compared to the exact values, being computationally feasible for $n = 1$ and $n = 10$, respectively. The upper, middle and lower plots correspond, respectively, to $n = 1$, $n = 10$, and $n = 50$.

We now apply the bound in Theorem 4. In view of (51), (54), (55) and (73), for all $\lambda \in (0,1]$ and $\alpha \geq \mathrm{e}^{-\frac{3}{2}}$,

$$\begin{aligned}
&\frac{D_{f_\alpha}\big(R^{(\lambda)}_{Y^n} \,\|\, Q_{Y^n}\big)}{D_{f_\alpha}\big(R^{(\lambda)}_{X^n} \,\|\, Q_{X^n}\big)} \\
&\leq \frac{\kappa_\alpha\big(\xi_1(n,\lambda),\, \xi_2(n,\lambda)\big)}{f_\alpha(0)+f'_\alpha(1)} \; \frac{\prod\limits_{i=1}^{n}\Big(1+\lambda^2\,\chi^2(P_{Y_i} \,\|\, Q_{Y_i})\Big)-1}{\prod\limits_{i=1}^{n}\Big(1+\lambda^2\,\chi^2(P_{X_i} \,\|\, Q_{X_i})\Big)-1} && (125)\\
&= \frac{f_\alpha\big(\xi_2(n,\lambda)\big)+f'_\alpha(1)\,\big(1-\xi_2(n,\lambda)\big)}{\big(\xi_2(n,\lambda)-1\big)^2\,\big(f_\alpha(0)+f'_\alpha(1)\big)} \cdot \frac{\prod\limits_{i=1}^{n}\left(1+\dfrac{\lambda^2(p_i*\delta_i-q_i*\delta_i)^2}{(q_i*\delta_i)(1-q_i*\delta_i)}\right)-1}{\prod\limits_{i=1}^{n}\left(1+\dfrac{\lambda^2(p_i-q_i)^2}{q_i(1-q_i)}\right)-1}, && (126)
\end{aligned}$$

where $\xi_1(n,\lambda) \in [0,1)$ and $\xi_2(n,\lambda) \in (1,\infty)$ are given in (122) and (123), respectively, and for $t \geq 0$,

$$\begin{aligned}
&f_\alpha(t)+f'_\alpha(1)(1-t) && (127)\\
&= (\alpha+t)^2\log(\alpha+t)-(\alpha+1)^2\log(\alpha+1)+\big[2(\alpha+1)\log(\alpha+1)+(\alpha+1)\log\mathrm{e}\big](1-t).
\end{aligned}$$

Figure 2 illustrates the upper bound on $\frac{D_{f_\alpha}(R^{(\lambda)}_{Y^n} \,\|\, Q_{Y^n})}{D_{f_\alpha}(R^{(\lambda)}_{X^n} \,\|\, Q_{X^n})}$ (see (125)–(127)) as a function of $\lambda \in (0,1]$. It refers to the case where $p_i \equiv \frac{1}{4}$, $q_i \equiv \frac{1}{2}$, and $\delta_i \equiv 0.110$ for all i (similarly to Figure 1). The upper and middle plots correspond to $n = 10$ with $\alpha = 10$ and $\alpha = 100$, respectively; the middle and lower plots correspond to $\alpha = 100$ with $n = 10$ and $n = 100$, respectively. The bounds in the upper and middle plots are compared to their exact values since their numerical computations are feasible for $n = 10$. It is observed from the numerical comparisons for $n = 10$ (see the upper and middle plots in Figure 2) that the upper bounds are informative, especially for large values of α where the f_α-divergence becomes closer to a scaled version of the χ^2-divergence (see Item (e) in Theorem 5).

3.2. Illustration of Theorems 3 and 5

Following the application of the data-processing inequalities in Theorems 2 and 4 to a class of f-divergences (see Section 3.1), some interesting properties of this class are introduced in Theorem 5.

For $\alpha \geq \mathrm{e}^{-\frac{3}{2}}$, let $d_{f_\alpha} \colon (0,1)^2 \to [0,\infty)$ be the binary f_α-divergence (see (55)), defined as

$$\begin{aligned}
d_{f_\alpha}(p\|q) &:= D_{f_\alpha}\big(\,\mathrm{Bernoulli}(p) \,\|\, \mathrm{Bernoulli}(q)\,\big) && (128)\\
&= q\left(\alpha+\frac{p}{q}\right)^2\log\left(\alpha+\frac{p}{q}\right)+(1-q)\left(\alpha+\frac{1-p}{1-q}\right)^2\log\left(\alpha+\frac{1-p}{1-q}\right)\\
&\quad -(\alpha+1)^2\log(\alpha+1), \quad \forall\,(p,q)\in(0,1)^2. && (129)
\end{aligned}$$

Theorem 5 is illustrated in Figure 3, showing that $d_{f_\alpha}(p\|q)$ is monotonically increasing as a function of $\alpha \geq \mathrm{e}^{-\frac{3}{2}}$ (note that the concavity in α is not reflected from these plots because the horizontal axis of α is in logarithmic scaling). The binary divergence $d_{f_\alpha}(p\|q)$ is also compared in Figure 3 with its lower and upper bounds in (61) and (65), respectively, illustrating that these bounds are both asymptotically tight for large values of α. The asymptotic approximation of $d_{f_\alpha}(p\|q)$ for large α, expressed as a function of α and $\chi^2(p\|q)$ (see (66)), is also depicted in Figure 3. The upper and lower plots in Figure 3 refer, respectively, to $(p,q) = (0.1, 0.9)$ and $(0.2, 0.8)$; a comparison of these plots show a better match between the exact value of the binary divergence, its upper and lower bounds, and its asymptotic approximation when the values of p and q are getting closer.

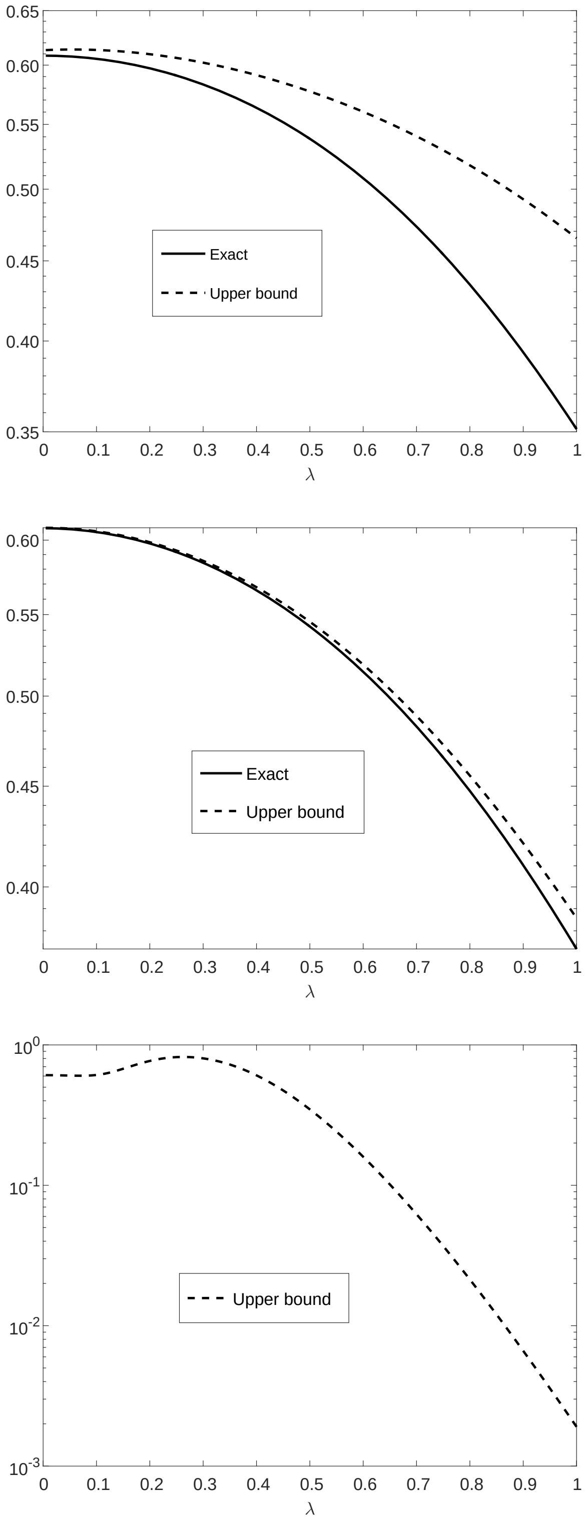

Figure 2. The upper bound in Theorem 4 applied to $\frac{D_{f_\alpha}(R^{(\lambda)}_{Y^n} \,\|\, Q_{Y^n})}{D_{f_\alpha}(R^{(\lambda)}_{X^n} \,\|\, Q_{X^n})}$ (see (125)–(127)) in the vertical axis versus $\lambda \in [0,1]$ in the horizontal axis. The f_α-divergence refers to Theorem 5. The probability mass functions P_{X_i} and Q_{X_i} are Bernoulli(p) and Bernoulli(q), respectively, for all $i \in \{1, \ldots, n\}$ with n uses of BSC(δ), and parameters $(p, q, \delta) = \left(\frac{1}{4}, \frac{1}{2}, 0.110\right)$. The upper and middle plots correspond to $n = 10$ with $\alpha = 10$ and $\alpha = 100$, respectively; the middle and lower plots correspond to $\alpha = 100$ with $n = 10$ and $n = 100$, respectively. The bounds in the upper and middle plots are compared to the exact values, being computationally feasible for $n = 10$.

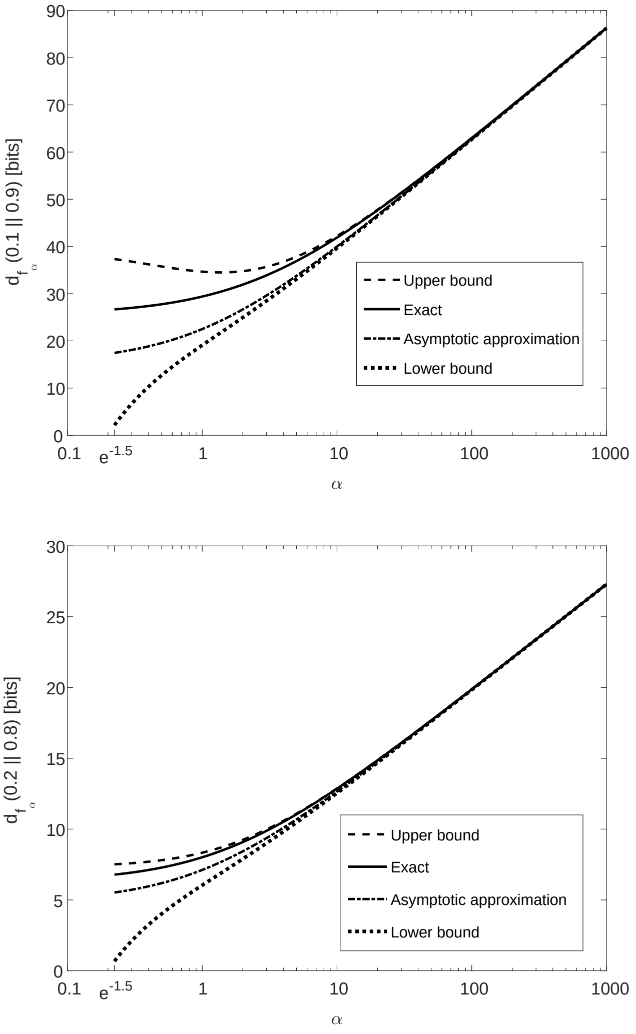

Figure 3. Plots of $d_{f_\alpha}(p\|q)$, its upper and lower bounds in (61) and (65), respectively, and its asymptotic approximation in (66) for large values of α. The plots are shown as a function of $\alpha \in \left[e^{-\frac{3}{2}}, 1000\right]$. The upper and lower plots refer, respectively, to $(p,q) = (0.1, 0.9)$ and $(p,q) = (0.2, 0.8)$.

In view of the results in (66) and (68), it is interesting to note that the asymptotic value of $D_{f_\alpha}(P\|Q)$ for large values of α is also the exact scaling of this f-divergence for *any finite* value of $\alpha \geq e^{-\frac{3}{2}}$ when the probability mass functions P and Q are close enough to each other.

We next consider the ratio of the contraction coefficients $\frac{\mu_{f_\alpha}(Q_X, W_{Y|X})}{\mu_{\chi^2}(Q_X, W_{Y|X})}$ where Q_X is finitely supported on $\mathcal{X}$ and it is not a point mass (i.e., $|\mathcal{X}| \geq 2$), and $W_{Y|X}$ is arbitrary. For all $\alpha \geq e^{-\frac{3}{2}}$,

$$1 \leq \frac{\mu_{f_\alpha}(Q_X, W_{Y|X})}{\mu_{\chi^2}(Q_X, W_{Y|X})} \leq \frac{f_\alpha(\xi) + f'_\alpha(1)(1-\xi)}{(\xi-1)^2\big(f_\alpha(0) + f'_\alpha(1)\big)}, \tag{130}$$

where $f_\alpha \colon (0,\infty) \to \mathbb{R}$ is given in (55), and

$$\xi := \frac{1}{\min\limits_{x\in\mathcal{X}} Q_X(x)} \in [|\mathcal{X}|, \infty). \tag{131}$$

The left-side inequality in (130) is due to ([25], Theorem 2) (see Proposition 2), and the right-side inequality in (130) holds due to (53) and (73).

Figure 4 shows the upper bound on the ratio of contraction coefficients $\frac{\mu_{f_\alpha}(Q_X, W_{Y|X})}{\mu_{\chi^2}(Q_X, W_{Y|X})}$, as it is given in the right-side inequality of (130), as a function of the parameter $\alpha \geq \mathrm{e}^{-\frac{3}{2}}$. The curves in Figure 4 correspond to different values of $\xi \in [|\mathcal{X}|, \infty)$, as it is given in (131); these upper bounds are monotonically decreasing in α, and they asymptotically tend to 1 as we let $\alpha \to \infty$. Hence, in view of the left-side inequality in (130), the upper bound on the ratio of the contraction coefficients (in the right-side inequality) is asymptotically tight in α. The fact that the ratio of the contraction coefficients in the middle of (130) tends asymptotically to 1, as α gets large, is not directly implied by Item (e) of Theorem 5. The latter implies that, for fixed probability mass functions P and Q and for sufficiently large α,

$$D_{f_\alpha}(P\|Q) \approx \left[\log(\alpha+1) + \tfrac{3}{2}\log \mathrm{e}\right] \chi^2(P\|Q); \tag{132}$$

however, there is no guarantee that for fixed Q and sufficiently large α, the approximation in (132) holds for all P. By the upper bound in the right side of (130), it follows however that $\mu_{f_\alpha}(Q_X, W_{Y|X})$ tends asymptotically (as we let $\alpha \to \infty$) to the contraction coefficient of the χ^2 divergence.

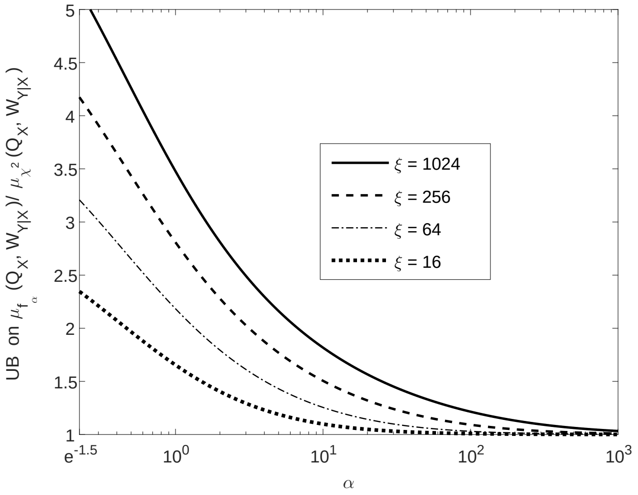

Figure 4. Curves of the upper bound on the ratio of contraction coefficients $\frac{\mu_{f_\alpha}(Q_X, W_{Y|X})}{\mu_{\chi^2}(Q_X, W_{Y|X})}$ (see the right-side inequality of (130)) as a function of the parameter $\alpha \geq \mathrm{e}^{-\frac{3}{2}}$. The curves correspond to different values of ξ in (131).

3.3. Illustration of Theorem 7 and Further Results

Theorem 7 provides upper and lower bounds on an f-divergence, $D_f(Q\|U_n)$, from any probability mass function Q supported on a finite set of cardinality n to an equiprobable distribution over this set. We apply in the following, the exact formula for

$$d_f(\rho) := \lim_{n\to\infty} \max_{Q \in \mathcal{P}_n(\rho)} D_f(Q\|U_n), \quad \rho \geq 1 \tag{133}$$

to several important f-divergences. From (87),

$$d_f(\rho) = \max_{x\in[0,1]} \left\{ x f\left(\frac{\rho}{1+(\rho-1)x}\right) + (1-x) f\left(\frac{1}{1+(\rho-1)x}\right) \right\}, \quad \rho \geq 1. \tag{134}$$

Since f is a convex function on $(0,\infty)$ with $f(1) = 0$, Jensen's inequality implies that the function which is subject to maximization in the right-side of (134) is non-negative over the interval $[0,1]$. It is equal to zero at the endpoints of the interval $[0,1]$, so the maximum over this interval is attained at an interior point. Note also that, in view of Items (d) and (e) of Theorem 7, the exact asymptotic expression in (134) satisfies

$$\max_{Q \in \mathcal{P}_n(\rho)} D_f(Q \| U_n) \leq d_f(\rho), \quad \forall \, n \in \{2, 3, \ldots\}, \; \rho \geq 1. \tag{135}$$

3.3.1. Total Variation Distance

This distance is an f-divergence with $f(t) := |t-1|$ for $t > 0$. Substituting f into (134) gives

$$d_f(\rho) = \max_{x \in [0,1]} \left\{ \frac{2(\rho-1)x(1-x)}{1+(\rho-1)x} \right\}. \tag{136}$$

By setting to zero the derivative of the function which is subject to maximization in the right side of (136), it can be verified that the maximizer over this interval is equal to $x = \frac{1}{1+\sqrt{\rho}}$, which implies that

$$d_f(\rho) = \frac{2(\sqrt{\rho}-1)}{\sqrt{\rho}+1}, \quad \forall \, \rho \geq 1. \tag{137}$$

3.3.2. Alpha Divergences

The class of Alpha divergences forms a parametric subclass of the f-divergences, which includes in particular the relative entropy, χ^2-divergence, and the squared-Hellinger distance. For $\alpha \in \mathbb{R}$, let

$$D_{\mathrm{A}}^{(\alpha)}(P \| Q) := D_{u_\alpha}(P \| Q), \tag{138}$$

where $u_\alpha \colon (0, \infty) \to \mathbb{R}$ is a non-negative and convex function with $u_\alpha(1) = 0$, which is defined for $t > 0$ as follows (see ([8], Chapter 2), followed by studies in, e.g., [10,16,43–45]):

$$u_\alpha(t) := \begin{cases} \dfrac{t^\alpha - \alpha(t-1) - 1}{\alpha(\alpha-1)}, & \alpha \in (-\infty, 0) \cup (0,1) \cup (1, \infty), \\ t \log_{\mathrm{e}} t + 1 - t, & \alpha = 1, \\ -\log_{\mathrm{e}} t, & \alpha = 0. \end{cases} \tag{139}$$

The functions u_0 and u_1 are defined in the right side of (139) by a continuous extension of u_α at $\alpha = 0$ and $\alpha = 1$, respectively. The following relations hold (see, e.g., ([44], (10)–(13))):

$$D_{\mathrm{A}}^{(1)}(P \| Q) = \tfrac{1}{\log \mathrm{e}} \, D(P \| Q), \tag{140}$$

$$D_{\mathrm{A}}^{(0)}(P \| Q) = \tfrac{1}{\log \mathrm{e}} \, D(Q \| P), \tag{141}$$

$$D_{\mathrm{A}}^{(2)}(P \| Q) = \tfrac{1}{2} \, \chi^2(P \| Q), \tag{142}$$

$$D_{\mathrm{A}}^{(-1)}(P \| Q) = \tfrac{1}{2} \, \chi^2(Q \| P), \tag{143}$$

$$D_{\mathrm{A}}^{(\frac{1}{2})}(P \| Q) = 4 \mathscr{H}^2(P \| Q). \tag{144}$$

Substituting $f := u_\alpha$ (see (139)) into the right side of (134) gives that

$$\begin{aligned} \Delta(\alpha, \rho) &:= d_{u_\alpha}(\rho) && (145) \\ &= \lim_{n \to \infty} \max_{Q \in \mathcal{P}_n(\rho)} D_{\mathrm{A}}^{(\alpha)}(Q \| U_n) && (146) \\ &= \max_{x \in [0,1]} \left\{ \frac{1+(\rho^\alpha - 1)x}{\left(1+(\rho-1)x\right)^\alpha} - 1 \right\}. && (147) \end{aligned}$$

Setting to zero the derivative of the function which is subject to maximization in the right side of (147) gives

$$x = x^* := \frac{1 + \alpha(\rho - 1) - \rho^\alpha}{(1-\alpha)(\rho-1)(\rho^\alpha - 1)}, \tag{148}$$

where it can be verified that $x^* \in (0,1)$ for all $\alpha \in (-\infty, 0) \cup (0,1) \cup (1,\infty)$ and $\rho > 1$. Substituting (148) into the right side of (147) gives that, for all such α and ρ,

$$\Delta(\alpha,\rho) = \frac{1}{\alpha(\alpha-1)} \left[\frac{(1-\alpha)^{\alpha-1}(\rho^\alpha - 1)^\alpha (\rho - \rho^\alpha)^{1-\alpha}}{(\rho-1)\alpha^\alpha} - 1 \right]. \tag{149}$$

By a continuous extension of $\Delta(\alpha,\rho)$ in (149) at $\alpha = 1$ and $\alpha = 0$, it follows that for all $\rho > 1$

$$\Delta(1,\rho) = \Delta(0,\rho) = \frac{\rho \log \rho}{\rho - 1} - \log \left(\frac{\mathrm{e}\rho \log_{\mathrm{e}} \rho}{\rho - 1} \right). \tag{150}$$

Consequently, for all $\rho > 1$,

$$\lim_{n\to\infty} \max_{Q \in \mathcal{P}_n(\rho)} D(Q\|U_n) = \log \mathrm{e} \lim_{n\to\infty} \max_{Q \in \mathcal{P}_n(\rho)} D_{\mathrm{A}}^{(1)}(Q\|U_n) \tag{151}$$

$$= \Delta(1,\rho) \ \log \mathrm{e} \tag{152}$$

$$= \frac{\rho \log \rho}{\rho - 1} - \log \left(\frac{\mathrm{e}\rho \log_{\mathrm{e}} \rho}{\rho - 1} \right), \tag{153}$$

where (151) holds due to (140); (152) is due to (146), and (153) holds due to (150). This sharpens the result in ([33], Theorem 2) for the relative entropy from the equiprobable distribution, $D(Q\|U_n) = \log n - H(Q)$, by showing that the bound in ([33], (7)) is asymptotically tight as we let $n \to \infty$. The result in ([33], Theorem 2) can be further tightened for finite n by applying the result in Theorem 7 (d) with $f(t) := u_1(t) \ \log \mathrm{e} = t \log t + (1-t) \log \mathrm{e}$ for all $t > 0$ (although, unlike the asymptotic result in (149), the refined bound for a finite n does not lend itself to a closed-form expression as a function of n; see also ([34], Remark 3), which provides such a refinement of the bound on $D(Q\|U_n)$ for finite n in a different approach).

From (141), (146) and (150), it follows similarly to (153) that for all $\rho > 1$

$$\lim_{n\to\infty} \max_{Q \in \mathcal{P}_n(\rho)} D(U_n\|Q) = \Delta(0,\rho) \ \log \mathrm{e} \tag{154}$$

$$= \frac{\rho \log \rho}{\rho - 1} - \log \left(\frac{\mathrm{e}\rho \log_{\mathrm{e}} \rho}{\rho - 1} \right). \tag{155}$$

It should be noted that in view of the one-to-one correspondence between the Rényi divergence and the Alpha divergence of the same order α where, for $\alpha \neq 1$,

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log\Big(1 + \alpha(\alpha-1) D_{\mathrm{A}}^{(\alpha)}(P\|Q)\Big), \tag{156}$$

the asymptotic result in (149) can be obtained from ([34], Lemma 4) and vice versa; however, in [34], the focus is on the Rényi divergence from the equiprobable distribution, whereas the result in (149) is obtained by specializing the asymptotic expression in (134) for a general f-divergence. Note also that the result in ([34], Lemma 4) is restricted to $\alpha > 0$, whereas the result in (149) and (150) covers all values of $\alpha \in \mathbb{R}$.

In view of (146), (149), (153), (155), and the special cases of the Alpha divergences in (140)–(144), it follows that for all $\rho > 1$ and for every integer $n \geq 2$

$$\max_{Q \in \mathcal{P}_n(\rho)} D(Q \| U_n) \leq \Delta(1, \rho) \, \log \mathrm{e} = \frac{\rho \log \rho}{\rho - 1} - \log \left(\frac{\mathrm{e} \rho \log_{\mathrm{e}} \rho}{\rho - 1} \right), \tag{157}$$

$$\max_{Q \in \mathcal{P}_n(\rho)} D(U_n \| Q) \leq \Delta(0, \rho) \, \log \mathrm{e} = \frac{\rho \log \rho}{\rho - 1} - \log \left(\frac{\mathrm{e} \rho \log_{\mathrm{e}} \rho}{\rho - 1} \right), \tag{158}$$

$$\max_{Q \in \mathcal{P}_n(\rho)} \chi^2(Q \| U_n) \leq 2\Delta(2, \rho) = \frac{(\rho - 1)^2}{4\rho}, \tag{159}$$

$$\max_{Q \in \mathcal{P}_n(\rho)} \chi^2(U_n \| Q) \leq 2\Delta(-1, \rho) = \frac{(\rho - 1)^2}{4\rho}, \tag{160}$$

$$\max_{Q \in \mathcal{P}_n(\rho)} \mathscr{H}^2(Q \| U_n) \leq \tfrac{1}{4} \, \Delta(\tfrac{1}{2}, \rho) = \frac{(\sqrt[4]{\rho} - 1)^2}{\sqrt{\rho} + 1}, \tag{161}$$

and the upper bounds on the right sides of (157)–(161) are asymptotically tight in the limit where n tends to infinity.

The next result characterizes the function $\Delta \colon (0, \infty) \times (1, \infty) \to \mathbb{R}$ as it is given in (149) and (150).

Theorem 9. *The function Δ satisfies the following properties:*

(a) *For every $\rho > 1$, $\Delta(\alpha, \rho)$ is a convex function of α over the real line, and it is symmetric around $\alpha = \frac{1}{2}$ with a global minimum at $\alpha = \frac{1}{2}$.*

(b) *The following inequalities hold:*

$$\alpha \, \Delta(\alpha, \rho) \leq \beta \, \Delta(\beta, \rho), \qquad 0 < \alpha \leq \beta < \infty, \tag{162}$$

$$(1 - \beta) \, \Delta(\beta, \rho) \leq (1 - \alpha) \, \Delta(\alpha, \rho), \qquad -\infty < \alpha \leq \beta < 1. \tag{163}$$

(c) *For every $\alpha \in \mathbb{R}$, $\Delta(\alpha, \rho)$ is monotonically increasing and continuous in $\rho \in (1, \infty)$, and $\lim_{\rho \to 1^+} \Delta(\alpha, \rho) = 0$.*

Proof. See Appendix H.1. □

Remark 6. *The symmetry of $\Delta(\alpha, \rho)$ around $\alpha = \frac{1}{2}$ (see Theorem 9 (a)) is not implied by the following symmetry property of the Alpha divergence around $\alpha = \frac{1}{2}$ (see, e.g., ([8], p. 36)):*

$$D_{\mathrm{A}}^{(\frac{1}{2} + \alpha)}(P \| Q) = D_{\mathrm{A}}^{(\frac{1}{2} - \alpha)}(Q \| P). \tag{164}$$

Relying on Theorem 9, the following corollary gives a similar result to (146) where the order of Q and U_n in $D_{\mathrm{A}}^{(\alpha)}(\cdot \| \cdot)$ is switched.

Corollary 1. *For all $\alpha \in \mathbb{R}$ and $\rho > 1$,*

$$\lim_{n \to \infty} \max_{Q \in \mathcal{P}_n(\rho)} D_{\mathrm{A}}^{(\alpha)}(U_n \| Q) = \Delta(\alpha, \rho). \tag{165}$$

Proof. See Appendix H.2. □

We next further exemplify Theorem 7 for the relative entropy. Let $f(t) := t \log t + (1-t) \log \mathrm{e}$ for $t > 0$. Then, $f''(t) = \frac{\log \mathrm{e}}{t}$, so the bounds on the second derivative of f over the interval $\left[\frac{1}{\rho}, \rho\right]$ are given by $M = \rho \log \mathrm{e}$ and $m = \frac{\log \mathrm{e}}{\rho}$. Theorem 7 (h) gives the following bounds:

$$\frac{\left(n\|Q\|_2^2 - 1\right) \log \mathrm{e}}{2\rho} \leq D(Q\|U_n) \leq \frac{\rho \left(n\|Q\|_2^2 - 1\right) \log \mathrm{e}}{2}. \tag{166}$$

From ([33], Theorem 2) (and (157)),

$$D(Q\|U_n) \leq \frac{\rho \log \rho}{\rho - 1} - \log \left(\frac{\mathrm{e}\rho \log_{\mathrm{e}} \rho}{\rho - 1} \right). \tag{167}$$

Furthermore, (96) gives that

$$D(Q\|U_n) \leq \tfrac{1}{8}(\rho - 1)^2 \log \mathrm{e}, \tag{168}$$

which, for $\rho > 1$, is a looser bound in comparison to (167). It can be verified, however, that the dominant term in the Taylor series expansion (around $\rho = 1$) of the right side of (167) coincides with the right side of (168), so the bounds scale similarly for small values of $\rho \geq 1$.

Suppose that we wish to assert that, for every integer $n \geq 2$ and for all probability mass functions $Q \in \mathcal{P}_n(\rho)$, the condition

$$D(Q\|U_n) \leq d \log \mathrm{e} \tag{169}$$

holds with a fixed $d > 0$. Due to the left side inequality in (89), this condition is equivalent to the requirement that

$$\lim_{n \to \infty} \max_{Q \in \mathcal{P}_n(\rho)} D(Q\|U_n) \leq d \log \mathrm{e}. \tag{170}$$

Due to the asymptotic tightness of the upper bound in the right side of (157) (as we let $n \to \infty$), requiring that this upper bound is not larger than $d \log \mathrm{e}$ is necessary and sufficient for the satisfiability of (169) for all n and $Q \in \mathcal{P}_n(\rho)$. This leads to the analytical solution $\rho \leq \rho_{\max}^{(1)}(d)$ with (see Appendix I)

$$\rho_{\max}^{(1)}(d) := \frac{W_{-1}\left(-\mathrm{e}^{-d-1}\right)}{W_0\left(-\mathrm{e}^{-d-1}\right)}, \tag{171}$$

where W_0 and W_{-1} denote, respectively, the principal and secondary real branches of the Lambert W function [37]. Requiring the stronger condition where the right side of (168) is not larger than $d \log \mathrm{e}$ leads to the sufficient solution $\rho \leq \rho_{\max}^{(2)}$ with the simple expression

$$\rho_{\max}^{(2)}(d) := 1 + \sqrt{8d}. \tag{172}$$

In comparison to $\rho_{\max}^{(1)}$ in (171), $\rho_{\max}^{(2)}$ in (172) is more insightful; these values nearly coincide for small values of $d > 0$, providing in that case the same range of possible values of ρ for asserting the satisfiability of condition (169). As it is shown in Figure 5, for $d \leq 0.01$, the difference between the maximal values of ρ in (171) and (172) is marginal, though in general $\rho_{\max}^{(1)}(d) > \rho_{\max}^{(2)}(d)$ for all $d > 0$.

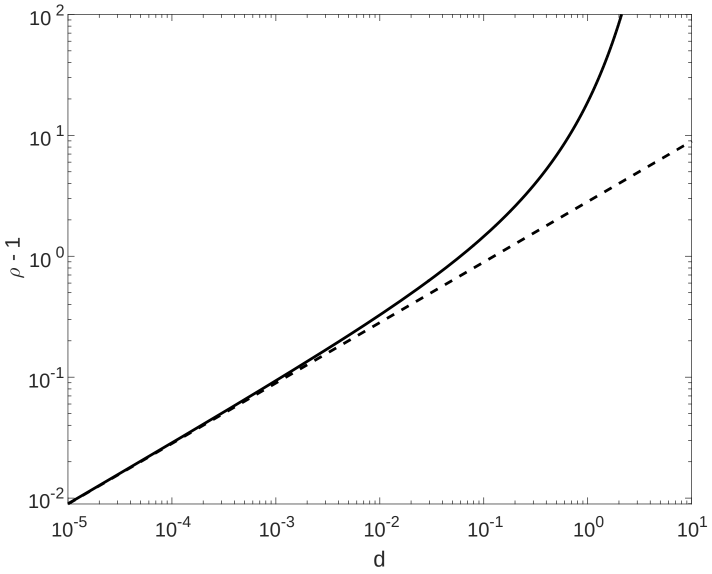

Figure 5. A comparison of the maximal values of ρ (minus 1) according to (171) and (172), asserting the satisfiability of the condition $D(Q\|U_n) \leq d \log \mathrm{e}$, with an arbitrary $d > 0$, for all integers $n \geq 2$ and probability mass functions Q supported on $\{1, \dots, n\}$ with $\frac{q_{\max}}{q_{\min}} \leq \rho$. The solid line refers to the necessary and sufficient condition which gives (171), and the dashed line refers to a stronger condition which gives (172).

3.3.3. The Subclass of f-Divergences in Theorem 5

This example refers to the subclass of f-divergences in Theorem 5. For these f_α-divergences, with $\alpha \geq \mathrm{e}^{-\frac{3}{2}}$, substituting $f := f_\alpha$ from (55) into the right side of (134) gives that for all $\rho \geq 1$

$$\begin{aligned}\Phi(\alpha, \rho) &:= d_{f_\alpha}(\rho) \qquad (173)\\ &= \lim_{n\to\infty} \max_{Q \in \mathcal{P}_n(\rho)} D_{f_\alpha}(Q\|U_n) \qquad (174)\\ &= \max_{x\in[0,1]} \Bigg\{ x\left(\alpha + \frac{\rho}{1+(\rho-1)x}\right)^2 \log\left(\alpha + \frac{\rho}{1+(\rho-1)x}\right) - (\alpha+1)^2 \log(\alpha+1)\\ &\qquad + (1-x)\left(\alpha + \frac{1}{1+(\rho-1)x}\right)^2 \log\left(\alpha + \frac{1}{1+(\rho-1)x}\right)\Bigg\}. \qquad (175)\end{aligned}$$

The exact asymptotic expression in the right side of (175) is subject to numerical maximization.

We next provide two alternative closed-form upper bounds, based on Theorems 5 and 7, and study their tightness. The two upper bounds, for all $\alpha \geq \mathrm{e}^{-\frac{3}{2}}$ and $\rho \geq 1$, are given by (see Appendix J)

$$\begin{aligned}\Phi(\alpha, \rho) \leq &\left[\log(\alpha+1) + \tfrac{3}{2}\log \mathrm{e} - \frac{\log \mathrm{e}}{\alpha+1}\right] \frac{(\rho-1)^2}{4\rho}\\ &+ \frac{\log \mathrm{e}}{81(\alpha+1)} \left(\frac{(\rho-1)(2\rho+1)(\rho+2)}{\rho(\rho+1)}\right)^2, \qquad (176)\end{aligned}$$

and

$$\Phi(\alpha, \rho) \leq \left[\log(\alpha+\rho) + \tfrac{3}{2}\log \mathrm{e}\right] \frac{(\rho-1)^2}{4\rho}. \qquad (177)$$

Suppose that we wish to assert that, for every integer $n \geq 2$ and for all probability mass functions $Q \in \mathcal{P}_n(\rho)$, the condition

$$D_{f_\alpha}(Q \| U_n) \leq d \log \mathrm{e} \tag{178}$$

holds with a fixed $d > 0$ and $\alpha \geq \mathrm{e}^{-\frac{3}{2}}$. Due to (173)–(174) and the left side inequality in (89), the satisfiability of the latter condition is equivalent to the requirement that

$$\Phi(\alpha, \rho) \leq d \log \mathrm{e}. \tag{179}$$

In order to obtain a sufficient condition for ρ to satisfy (179), expressed as an explicit function of α and d, the upper bound in the right side of (176) is slightly loosened to

$$\Phi(\alpha, \rho) \leq a(\rho - 1)^2 + b \min\{\rho - 1, (\rho - 1)^2\}, \tag{180}$$

where

$$a := \frac{4 \log \mathrm{e}}{81(\alpha + 1)}, \tag{181}$$

$$b := \tfrac{1}{4} \log(\alpha + 1) + \tfrac{3}{8} \log \mathrm{e}, \tag{182}$$

for all $\rho \geq 1$ and $\alpha \geq \mathrm{e}^{-\frac{3}{2}}$. The upper bounds in the right sides of (176), (177) and (180) are derived in Appendix J.

In comparison to (179), the stronger requirement that the right side of (180) is less than or equal to $d \log \mathrm{e}$ gives the sufficient condition

$$\rho \leq \rho_{\max}(\alpha, d) := \max\{\rho_1(\alpha, d), \rho_2(\alpha, d)\}, \tag{183}$$

with

$$\rho_1(\alpha, d) := 1 + \frac{\sqrt{b^2 + 4ad \log \mathrm{e}} - b}{2a}, \tag{184}$$

$$\rho_2(\alpha, d) := 1 + \sqrt{\frac{d \log \mathrm{e}}{a + b}}. \tag{185}$$

Figure 6 compares the exact expression in (175) with its upper bounds in (176), (177) and (180). These bounds show good match with the exact value, and none of the bounds in (176) and (177) is superseded by the other; the bound in (180) is looser than (176), and it is derived for obtaining the closed-form solution in (183)–(185). The bound in (176) is tighter than the bound in (177) for small values of $\rho \geq 1$, whereas the latter bound outperforms the first one for sufficiently large values of ρ. It has been observed numerically that the tightness of the bounds is improved by increasing the value of α, and the range of parameters of ρ over which the bound in (176) outperforms the second bound in (177) is enlarged when α is increased. It is also shown in Figure 6 that the bound in (176) and its loosened version in (180) almost coincide for sufficiently small values of ρ (i.e., for ρ is close to 1), and also for sufficiently large values of ρ.

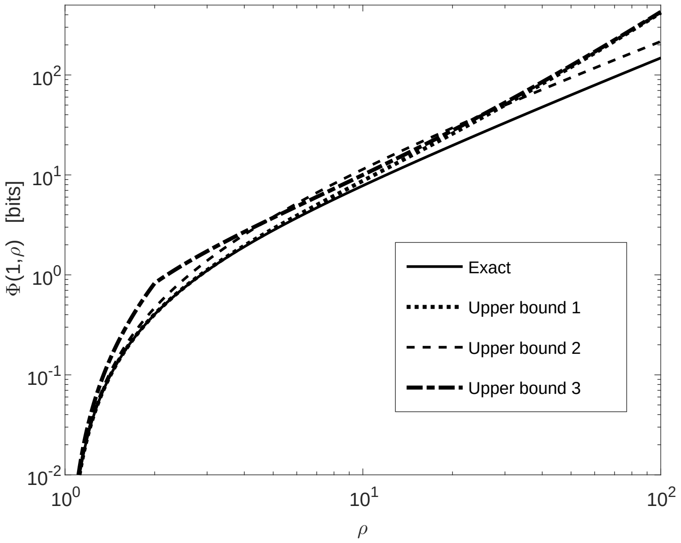

Figure 6. A comparison of the exact expression of $\Phi(\alpha, \rho)$ in (175), with $\alpha = 1$, and its three upper bounds in the right sides of (176), (177) and (180) (called 'Upper bound 1' (dotted line), 'Upper bound 2' (thin dashed line), and 'Upper bound 3' (thick dashed line), respectively).

3.4. An Interpretation of $u_f(\cdot, \cdot)$ in Theorem 7

We provide here an interpretation of $u_f(n, \rho)$ in (77), for $\rho > 1$ and an integer $n \geq 2$; note that $u_f(n, 1) \equiv 0$ since $\mathcal{P}_n(1) = \{U_n\}$. Before doing so, recall that (82) introduces an identity which significantly simplifies the numerical calculation of $u_f(n, \rho)$, and (85) gives (asymptotically tight) upper and lower bounds.

The following result relies on the variational representation of f-divergences.

Theorem 10. *Let $f \colon (0, \infty) \to \mathbb{R}$ be convex with $f(1) = 0$, and let $\overline{f} \colon \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ be the convex conjugate function of f (a.k.a. the Fenchel-Legendre transform of f), i.e.,*

$$\overline{f}(x) := \sup_{t>0} \{tx - f(t)\}, \quad x \in \mathbb{R}. \tag{186}$$

Let $\rho > 1$, and define $\mathcal{A}_n := \{1, \ldots, n\}$ for an integer $n \geq 2$. Then, the following holds:

(a) *For every $P \in \mathcal{P}_n(\rho)$, a random variable $X \sim P$, and a function $g \colon \mathcal{A}_n \to \mathbb{R}$,*

$$\mathbb{E}[g(X)] \leq u_f(n, \rho) + \frac{1}{n} \sum_{i=1}^{n} \overline{f}\big(g(i)\big). \tag{187}$$

(b) *There exists $P \in \mathcal{P}_n(\rho)$ such that, for every $\varepsilon > 0$, there is a function $g_\varepsilon \colon \mathcal{A}_n \to \mathbb{R}$ which satisfies*

$$\mathbb{E}[g_\varepsilon(X)] \geq u_f(n, \rho) + \frac{1}{n} \sum_{i=1}^{n} \overline{f}\big(g_\varepsilon(i)\big) - \varepsilon, \tag{188}$$

with $X \sim P$.

Proof. See Appendix K. □

Remark 7. *The proof suggests a constructive way to obtain, for an arbitrary $\varepsilon > 0$, a function g_ε which satisfies* (188).

4. Applications in Information Theory and Statistics

4.1. Bounds on the List Decoding Error Probability with f-Divergences

The minimum probability of error of a random variable X given Y, denoted by $\varepsilon_{X|Y}$, can be achieved by a deterministic function (*maximum-a-posteriori* decision rule) $\mathcal{L}^* \colon \mathcal{Y} \to \mathcal{X}$ (see [42]):

$$\begin{aligned} \varepsilon_{X|Y} &= \min_{\mathcal{L}\colon \mathcal{Y}\to\mathcal{X}} \mathbb{P}[X \neq \mathcal{L}(Y)] && (189)\\ &= \mathbb{P}[X \neq \mathcal{L}^*(Y)] && (190)\\ &= 1 - \mathbb{E}\left[\max_{x\in\mathcal{X}} P_{X|Y}(x|Y)\right]. && (191) \end{aligned}$$

Fano's inequality [46] gives an upper bound on the conditional entropy $H(X|Y)$ as a function of $\varepsilon_{X|Y}$ (or, otherwise, providing a lower bound on $\varepsilon_{X|Y}$ as a function of $H(X|Y)$) when X takes a finite number of possible values.

The list decoding setting, in which the hypothesis tester is allowed to output a subset of given cardinality, and an error occurs if the true hypothesis is not in the list, has great interest in information theory. A generalization of Fano's inequality to list decoding, in conjunction with the blowing-up lemma ([17], Lemma 1.5.4), leads to strong converse results in multi-user information theory. This approach was initiated in ([47], Section 5) (see also ([48], Section 3.6)). The main idea of the successful combination of these two tools is that, given a code, it is possible to blow-up the decoding sets in a way that the probability of decoding error can be as small as desired for sufficiently large blocklengths; since the blown-up decoding sets are no longer disjoint, the resulting setup is a list decoder with sub-exponential list size (as a function of the block length).

In statistics, Fano's-type lower bounds on Bayes and minimax risks, expressed in terms of f-divergences, are derived in [49,50].

In this section, we further study the setup of list decoding, and derive bounds on the average list decoding error probability. We first consider the special case where the list size is fixed (see Section 4.1.1), and then move to the more general case of a list size which depends on the channel observation (see Section 4.1.2).

4.1.1. Fixed-Size List Decoding

A generalization of Fano's inequality for fixed-size list decoding is given in ([42], (139)), expressed as a function of the conditional Shannon entropy (strengthening ([51], Lemma 1)). A further generalization in this setup, which is expressed as a function of the Arimoto-Rényi conditional entropy with an arbitrary positive order (see Definition 9), is provided in ([42], Theorem 8).

The next result provides a generalized Fano's inequality for fixed-size list decoding, expressed in terms of an arbitrary f-divergence. Some earlier results in the literature are reproduced from the next result, followed by its strengthening as an application of Theorem 1.

Theorem 11. *Let P_{XY} be a probability measure defined on $\mathcal{X} \times \mathcal{Y}$ with $|\mathcal{X}| = M$. Consider a decision rule $\mathcal{L}\colon \mathcal{Y} \to \binom{\mathcal{X}}{L}$, where $\binom{\mathcal{X}}{L}$ stands for the set of subsets of $\mathcal{X}$ with cardinality L, and $L < M$ is fixed. Denote the list decoding error probability by $P_{\mathcal{L}} := \mathbb{P}\big[X \notin \mathcal{L}(Y)\big]$. Let U_M denote an equiprobable probability mass function on $\mathcal{X}$. Then, for every convex function $f\colon (0,\infty) \to \mathbb{R}$ with $f(1) = 0$,*

$$\mathbb{E}\Big[D_f\big(P_{X|Y}(\cdot|Y) \,\|\, U_M\big)\Big] \geq \frac{L}{M}\, f\left(\frac{M\,(1-P_{\mathcal{L}})}{L}\right) + \left(1 - \frac{L}{M}\right) f\left(\frac{MP_{\mathcal{L}}}{M-L}\right). \tag{192}$$

Proof. See Appendix L. □

Remark 8. *The case where $L = 1$ (i.e., a decoder with a single output) gives ([50], (5)).*

As consequences of Theorem 11, we first reproduce some earlier results as special cases.

Corollary 2 ([42] (139))**.** *Under the assumptions in Theorem 11,*

$$H(X|Y) \le \log M - d\biggl(P_{\mathcal{L}} \,\|\, 1 - \frac{L}{M}\biggr) \tag{193}$$

where $d(\cdot\|\cdot)\colon [0,1] \times [0,1] \to [0,+\infty]$ *denotes the binary relative entropy, defined as the continuous extension of* $D([p,1-p]\|[q,1-q]) := p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$ *for* $p,q \in (0,1)$*.*

Proof. The choice $f(t) := t\log t + (1-t)\log \mathrm{e}$, for $t > 0$, (note that $f(t) = u_1(t)\log\mathrm{e}$ with $u_1(\cdot)$ defined in (139)) gives

$$\begin{aligned}\mathbb{E}\Bigl[D_f\bigl(P_{X|Y}(\cdot|Y) \,\|\, U_M\bigr)\Bigr] &= \int_{\mathcal{Y}} \mathrm{d}P_Y(y)\, D\bigl(P_{X|Y}(\cdot|y) \,\|U_M\bigr) &(194)\\ &= \int_{\mathcal{Y}} \mathrm{d}P_Y(y)\, \bigl[\log M - H(X|Y=y)\bigr] &(195)\\ &= \log M - H(X|Y), &(196)\end{aligned}$$

and

$$\frac{L}{M}\, f\biggl(\frac{M\,(1-P_{\mathcal{L}})}{L}\biggr) + \biggl(1 - \frac{L}{M}\biggr)\, f\biggl(\frac{MP_{\mathcal{L}}}{M-L}\biggr) = d\biggl(P_{\mathcal{L}} \,\|\, 1 - \frac{L}{M}\biggr). \tag{197}$$

Substituting (194)–(197) into (192) gives (193). □

Theorem 11 enables to reproduce a result in [42] which generalizes Corollary 2. It relies on Rényi information measures, and we first provide definitions for a self-contained presentation.

Definition 8 ([40])**.** *Let* P_X *be a probability mass function defined on a discrete set* $\mathcal{X}$*. The Rényi entropy of order* $\alpha \in (0,1) \cup (1,\infty)$ *of* X*, denoted by* $H_\alpha(X)$ *or* $H_\alpha(P_X)$*, is defined as*

$$\begin{aligned}H_\alpha(X) &:= \frac{1}{1-\alpha}\, \log \sum_{x\in\mathcal{X}} P_X^\alpha(x) &(198)\\ &= \frac{\alpha}{1-\alpha}\, \log \|P_X\|_\alpha. &(199)\end{aligned}$$

The Rényi entropy is continuously extended at orders 0, 1, and ∞; at order 1, it coincides with the Shannon entropy $H(X)$*.*

Definition 9 ([52])**.** *Let* P_{XY} *be defined on* $\mathcal{X} \times \mathcal{Y}$*, where X is a discrete random variable. The Arimoto-Rényi conditional entropy of order* $\alpha \in [0,\infty]$ *of X given Y is defined as follows:*

- *If* $\alpha \in (0,1) \cup (1,\infty)$*, then*

$$\begin{aligned}H_\alpha(X|Y) &= \frac{\alpha}{1-\alpha}\, \log\, \mathbb{E}\left[\left(\sum_{x\in\mathcal{X}} P_{X|Y}^\alpha(x|Y)\right)^{\frac{1}{\alpha}}\right] &(200)\\ &= \frac{\alpha}{1-\alpha}\, \log \mathbb{E}\Bigl[\|P_{X|Y}(\cdot|Y)\|_\alpha\Bigr] &(201)\\ &= \frac{\alpha}{1-\alpha}\, \log \int_{\mathcal{Y}} \mathrm{d}P_Y(y)\, \exp\biggl(\frac{1-\alpha}{\alpha}\, H_\alpha(X|Y=y)\biggr). &(202)\end{aligned}$$

- *The Arimoto-Rényi conditional entropy is continuously extended at orders 0, 1, and ∞; at order 1, it coincides with the conditional Shannon entropy* $H(X|Y)$*.*

Definition 10 ([42])**.** *For all $\alpha \in (0,1) \cup (1,\infty)$, the binary Rényi divergence of order α, denoted by $d_\alpha(p\|q)$ for $(p,q) \in [0,1]^2$, is defined as $D_\alpha([p,1-p] \,\|\, [q,1-q])$. It is the continuous extension to $[0,1]^2$ of*

$$d_\alpha(p\|q) = \frac{1}{\alpha-1} \log\Big(p^\alpha q^{1-\alpha} + (1-p)^\alpha (1-q)^{1-\alpha}\Big). \tag{203}$$

For $\alpha = 1$,

$$d_1(p\|q) := \lim_{\alpha \to 1} d_\alpha(p\|q) = d(p\|q). \tag{204}$$

The following result, generalizing Corollary 2, is shown to be a consequence of Theorem 11. It has been originally derived in ([42], Theorem 8) in a different way. The alternative derivation of this inequality relies on Theorem 11, applied to the family of Alpha-divergences (see (138)) as a subclass of the f-divergences.

Corollary 3 ([42] Theorem 8)**.** *Under the assumptions in Theorem 11, for $\alpha \in (0,1) \cup (1,\infty)$,*

$$\begin{aligned} H_\alpha(X|Y) &\le \log M - d_\alpha\left(P_{\mathcal{L}} \,\|\, 1 - \frac{L}{M}\right) && (205) \\ &= \frac{1}{1-\alpha} \log\Big(L^{1-\alpha}\,(1-P_{\mathcal{L}})^\alpha + (M-L)^{1-\alpha}\,P_{\mathcal{L}}^\alpha\Big), && (206) \end{aligned}$$

with equality in (205) if and only if

$$P_{X|Y}(x|y) = \begin{cases} \dfrac{P_{\mathcal{L}}}{M-L}, & x \notin \mathcal{L}(y), \\[2mm] \dfrac{1-P_{\mathcal{L}}}{L}, & x \in \mathcal{L}(y). \end{cases} \tag{207}$$

Proof. See Appendix M. □

Another application of Theorem 11 with the selection $f(t) := |t-1|^s$, for $t \in [0,\infty)$ and a parameter $s \ge 1$, gives the following result.

Corollary 4. *Under the assumptions in Theorem 11, for all $s \ge 1$,*

$$P_{\mathcal{L}} \ge 1 - \frac{L}{M} - \Big(L^{1-s} + (M-L)^{1-s}\Big)^{-\frac{1}{s}} \left(\mathbb{E}\left[\sum_{x \in \mathcal{X}} \left|P_{X|Y}(x|Y) - \frac{1}{M}\right|^s\right]\right)^{\frac{1}{s}}, \tag{208}$$

where (208) holds with equality if X and Y are independent with X being equiprobable. For $s = 1$ and $s = 2$, (208) respectively gives that

$$\begin{aligned} P_{\mathcal{L}} &\ge 1 - \frac{L}{M} - \frac{1}{2}\,\mathbb{E}\left[\sum_{x \in \mathcal{X}} \left|P_{X|Y}(x|Y) - \frac{1}{M}\right|\right], && (209) \\ P_{\mathcal{L}} &\ge 1 - \frac{L}{M} - \sqrt{\frac{L}{M}\left(1 - \frac{L}{M}\right)\big(M\,\mathbb{E}[P_{X|Y}(X|Y)] - 1\big)}. && (210) \end{aligned}$$

The following refinement of the generalized Fano's inequality in Theorem 11 relies on the version of the strong data-processing inequality in Theorem 1.

Theorem 12. *Under the assumptions in Theorem 11, let the convex function $f \colon (0,\infty) \to \mathbb{R}$ be twice differentiable, and assume that there exists a constant $m_f > 0$ such that*

$$f''(t) \geq m_f, \quad \forall\, t \in \mathcal{I}(\xi_1^*, \xi_2^*), \tag{211}$$

where

$$\xi_1^* := M \inf_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_{X|Y}(x|y), \tag{212}$$

$$\xi_2^* := M \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_{X|Y}(x|y), \tag{213}$$

and the interval $\mathcal{I}(\cdot,\cdot)$ is defined in (23). Let $u^+ := \max\{u, 0\}$ for $u \in \mathbb{R}$. Then,

(a)

$$\begin{aligned}\mathbb{E}\Big[D_f\big(P_{X|Y}(\cdot|Y)\,\|\,U_M\big)\Big] \geq \frac{L}{M}\, f\left(\frac{M\,(1-P_{\mathcal{L}})}{L}\right) + \left(1-\frac{L}{M}\right)\, f\left(\frac{MP_{\mathcal{L}}}{M-L}\right) \\ + \tfrac{1}{2} m_f\, M \left(\mathbb{E}\big[P_{X|Y}(X|Y)\big] - \frac{1-P_{\mathcal{L}}}{L} - \frac{P_{\mathcal{L}}}{M-L}\right)^{+}.\end{aligned} \tag{214}$$

(b) If the list decoder selects the L most probable elements from $\mathcal{X}$, given the value of $Y \in \mathcal{Y}$, then (214) is strengthened to

$$\begin{aligned}\mathbb{E}\Big[D_f\big(P_{X|Y}(\cdot|Y)\,\|\,U_M\big)\Big] \geq \frac{L}{M}\, f\left(\frac{M\,(1-P_{\mathcal{L}})}{L}\right) + \left(1-\frac{L}{M}\right)\, f\left(\frac{MP_{\mathcal{L}}}{M-L}\right) \\ + \tfrac{1}{2} m_f\, M \left(\mathbb{E}\big[P_{X|Y}(X|Y)\big] - \frac{1-P_{\mathcal{L}}}{L}\right),\end{aligned} \tag{215}$$

where the last term in the right side of (215) is necessarily non-negative.

Proof. See Appendix N. □

An application of Theorem 12 gives the following tightened version of Corollary 2.

Corollary 5. *Under the assumptions in Theorem 11, the following holds:*

(a) Inequality (193) is strengthened to

$$H(X|Y) \leq \log M - d\left(P_{\mathcal{L}} \,\|\, 1-\frac{L}{M}\right) - \frac{\log \mathrm{e}}{2} \, \frac{\left(\mathbb{E}\big[P_{X|Y}(X|Y)\big] - \frac{1-P_{\mathcal{L}}}{L} - \frac{P_{\mathcal{L}}}{M-L}\right)^{+}}{\sup\limits_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_{X|Y}(x|y)}. \tag{216}$$

(b) If the list decoder selects the L most probable elements from $\mathcal{X}$, given the value of $Y \in \mathcal{Y}$, then (216) is strengthened to

$$H(X|Y) \leq \log M - d\left(P_{\mathcal{L}} \,\|\, 1-\frac{L}{M}\right) - \frac{\log \mathrm{e}}{2} \cdot \frac{\left(\mathbb{E}\big[P_{X|Y}(X|Y)\big] - \frac{1-P_{\mathcal{L}}}{L}\right)^{+}}{\sup\limits_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_{X|Y}(x|y)}. \tag{217}$$

Proof. The choice $f(t) := t \log t + (1-t) \log \mathrm{e}$, for $t > 0$, gives (see (23) and (211)–(213))

$$\begin{aligned} m_f \, M &= M \inf_{t \in \mathcal{I}(\xi_1^*, \xi_2^*)} f''t) \\ &= \frac{M \log \mathrm{e}}{\xi_2^*} \\ &= \frac{\log \mathrm{e}}{\sup\limits_{(x,y) \in \mathcal{X} \times \mathcal{Y}} P_{X|Y}(x|y)}. \end{aligned} \tag{218}$$

Substituting (194)–(197) and (218) into (214) and (215) give, respectively, (216) and (217). □

Remark 9. *Similarly to the bounds on $P_{\mathcal{L}}$ in (193) and (205), which tensorize when $P_{X|Y}$ is replaced by a product probability measure $P_{X^n|Y^n}(\underline{x}|\underline{y}) = \prod\limits_{i=1}^{n} P_{X_i|Y_i}(x_i|y_i)$, this is also the case with the new bounds in (216) and (217).*

Remark 10. *The ceil operation in the right side of (217) is redundant with $P_{\mathcal{L}}$ denoting the list decoding error probability (see (A335)–(A341)). However, for obtaining a lower bound on $P_{\mathcal{L}}$ with (217), the ceil operation assures that the bound is at least as good as the lower bound which relies on the generalized Fano's inequality in (193).*

Example 1. *Let X and Y be discrete random variables taking values in $\mathcal{X} = \{0, 1, \ldots, 8\}$ and $\mathcal{Y} = \{0, 1\}$, respectively, and let P_{XY} be the joint probability mass function, given by*

$$[P_{XY}(x,y)]_{(x,y) \in \mathcal{X} \times \mathcal{Y}} = \frac{1}{512} \begin{pmatrix} 128 & 64 & 32 & 16 & 8 & 4 & 2 & 1 & 1 \\ 2 & 2 & 2 & 2 & 8 & 16 & 32 & 64 & 128 \end{pmatrix}^{\mathrm{T}}. \tag{219}$$

Let the list decoder select the L most probable elements from $\mathcal{X}$, given the value of $Y \in \mathcal{Y}$. Table 1 compares the list decoding error probability $P_{\mathcal{L}}$ with the lower bound which relies on the generalized Fano's inequality in (193), its tightened version in (217), and the closed-form lower bound in (210) for fixed list sizes of $L = 1, \ldots, 4$. For $L = 3$ and $L = 4$, (217) improves the lower bound in (193) (see Table 1). If $L = 4$, then the generalized Fano's lower bound in (193) and also (210) are useless, whereas (217) gives a non-trivial lower bound. It is shown here that none of the new lower bounds in (210) and (217) is superseded by the other.

Table 1. The lower bounds on $P_{\mathcal{L}}$ in (193), (210) and (217), and its exact value for fixed list size L (see Example 1).

L	Exact $P_{\mathcal{L}}$	(193)	(217)	(210)
1	0.500	0.353	0.353	0.444
2	0.250	0.178	0.178	0.190
3	0.125	0.065	0.072	5.34×10^{-5}
4	0.063	0	0.016	0

4.1.2. Variable-Size List Decoding

In the more general setting of list decoding where the size of the list may depend on the channel observation, Fano's inequality has been generalized as follows.

Proposition 5 (([48], Appendix 3.E) and [53])**.** *Let P_{XY} be a probability measure defined on $\mathcal{X} \times \mathcal{Y}$ with $|\mathcal{X}| = M$. Consider a decision rule $\mathcal{L} \colon \mathcal{Y} \to 2^{\mathcal{X}}$, and let the (average) list decoding error probability be given by $P_{\mathcal{L}} := \mathbb{P}\big[X \notin \mathcal{L}(Y)\big]$ with $|\mathcal{L}(y)| \geq 1$ for all $y \in \mathcal{Y}$. Then,*

$$H(X|Y) \leq h(P_{\mathcal{L}}) + \mathbb{E}[\log |\mathcal{L}(Y)|] + P_{\mathcal{L}} \log M, \tag{220}$$

where $h \colon [0,1] \to [0, \log 2]$ denotes the binary entropy function. If $|\mathcal{L}(Y)| \leq N$ almost surely, then also

$$H(X|Y) \leq h(P_{\mathcal{L}}) + (1 - P_{\mathcal{L}}) \log N + P_{\mathcal{L}} \log M. \tag{221}$$

By relying on the data-processing inequality for f-divergences, we derive in the following an alternative explicit lower bound on the average list decoding error probability $P_{\mathcal{L}}$. The derivation relies on the E_γ divergence (see, e.g., [54]), which forms a subclass of the f-divergences.

Theorem 13. *Under the assumptions in (220), for every $\gamma \geq 1$,*

$$P_{\mathcal{L}} \geq \frac{1+\gamma}{2} - \frac{\gamma \mathbb{E}[|\mathcal{L}(Y)|]}{M} - \frac{1}{2}\,\mathbb{E}\left[\sum_{x \in \mathcal{X}} \left| P_{X|Y}(x|Y) - \frac{\gamma}{M} \right|\right]. \tag{222}$$

Let $\gamma \geq 1$, and let $|\mathcal{L}(y)| \leq \frac{M}{\gamma}$ for all $y \in \mathcal{Y}$. Then, (222) holds with equality if, for every $y \in \mathcal{Y}$, the list decoder selects the $|\mathcal{L}(y)|$ most probable elements in $\mathcal{X}$ given $Y = y$; if $x_\ell(y)$ denotes the ℓ-th most probable element in $\mathcal{X}$ given $Y = y$, where ties in probabilities are resolved arbitrarily, then (222) holds with equality if

$$P_{X|Y}(x_\ell(y)\,|y) = \begin{cases} \alpha(y), & \forall\, \ell \in \big\{1, \ldots, |\mathcal{L}(y)|\big\}, \\[2ex] \dfrac{1 - \alpha(y)\,|\mathcal{L}(y)|}{M - |\mathcal{L}(y)|}, & \forall\, \ell \in \big\{|\mathcal{L}(y)| + 1, \ldots, M\big\}, \end{cases} \tag{223}$$

with $\alpha \colon \mathcal{Y} \to [0,1]$ being an arbitrary function which satisfies

$$\frac{\gamma}{M} \leq \alpha(y) \leq \frac{1}{|\mathcal{L}(y)|}, \quad \forall\, y \in \mathcal{Y}. \tag{224}$$

Proof. See Appendix O. □

Remark 11. *By setting $\gamma = 1$ and $|\mathcal{L}(Y)| = L$ (i.e., a decoding list of fixed size L), (222) is specialized to (209).*

Example 2. *Let X and Y be discrete random variables taking their values in $\mathcal{X} = \{0,1,2,3,4\}$ and $\mathcal{Y} = \{0,1\}$, respectively, and let P_{XY} be their joint probability mass function, which is given by*

$$\begin{cases} P_{XY}(0,0) = P_{XY}(1,0) = P_{XY}(2,0) = \frac{1}{8}, & P_{XY}(3,0) = P_{XY}(4,0) = \frac{1}{16}, \\[1ex] P_{XY}(0,1) = P_{XY}(1,1) = P_{XY}(2,1) = \frac{1}{24}, & P_{XY}(3,1) = P_{XY}(4,1) = \frac{3}{16}. \end{cases} \tag{225}$$

Let $\mathcal{L}(0) := \{0,1,2\}$ and $\mathcal{L}(1) := \{3,4\}$ be the lists in $\mathcal{X}$, given the value of $Y \in \mathcal{Y}$. We get $P_Y(0) = P_Y(1) = \frac{1}{2}$, so the conditional probability mass function of X given Y satisfies $P_{X|Y}(x|y) = 2P_{XY}(x,y)$ for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$. It can be verified that, if $\gamma = \frac{5}{4}$, then $\max\{|\mathcal{L}(0)|, |\mathcal{L}(1)|\} = 3 \leq \frac{M}{\gamma}$, and also (223) and (224) are satisfied (here, $M := |\mathcal{X}| = 5$, $\alpha(0) = \frac{1}{4} = \frac{\gamma}{M}$ and $\alpha(1) = \frac{3}{8} \in \big[\frac{1}{4}, \frac{1}{2}\big]$). By Theorem 13, it follows that (222) holds in this case with equality, and the list decoding error probability is equal to $P_{\mathcal{L}} = 1 - \mathbb{E}\big[\alpha(Y)\,|\mathcal{L}(Y)|\big] = \frac{1}{4}$ (i.e., it coincides with the lower bound in the right side of (222) with $\gamma = \frac{5}{4}$). On the other hand, the generalized Fano's inequality in (220) gives that $P_{\mathcal{L}} \geq 0.1206$ (the left side of (220) is $H(X|Y) = \frac{5}{2} \log 2 - \frac{1}{4} \log 3 = 2.1038$ bits); moreover, by letting $N := \max_{y \in \mathcal{Y}} |\mathcal{L}(y)| = 3$, (221) gives the *looser lower bound $P_{\mathcal{L}} \geq 0.0939$. This exemplifies a case where the lower bound in Theorem 13 is tight, whereas the generalized Fano's inequalities in* (220) *and* (221) *are looser.*

4.2. A Measure for the Approximation of Equiprobable Distributions by Tunstall Trees

The best possible approximation of equiprobable distributions, which one can get by using tree codes has been considered in [38]. The optimal solution is obtained by using Tunstall codes, which are variable-to-fixed lossless compression codes (see ([55], Section 11.2.3), [56]). The main idea behind Tunstall codes is parsing the source sequence into variable-length segments of roughly the same probability, and then coding all these segments with codewords of fixed length. This task is done by assigning the leaves of a Tunstall tree, which correspond to segments of source symbols with a variable length (according to the depth of the leaves in the tree), to codewords of fixed length. The following result links Tunstall trees with majorization theory.

Proposition 6 ([38] Theorem 1)**.** *Let P_ℓ be the probability measure generated on the leaves by a Tunstall tree $\mathcal{T}$, and let Q_ℓ be the probability measure generated by an arbitrary tree $\mathcal{S}$ with the same number of leaves as of $\mathcal{T}$. Then, $P_\ell \prec Q_\ell$.*

From Proposition 6, and the Schur-convexity of an f-divergence $D_f(\cdot \| U_n)$ (see ([38], Lemma 1)), it follows that (see ([38], Corollary 1))

$$D_f(P_\ell \| U_n) \leq D_f(Q_\ell \| U_n), \tag{226}$$

where n designates the joint number of leaves of the trees $\mathcal{T}$ and $\mathcal{S}$.

Before we proceed, it is worth noting that the strong data-processing inequality in Theorem 6 implies that if f is also twice differentiable, then (226) can be strengthened to

$$D_f(P_\ell \| U_n) + n c_f(n q_{\min}, n q_{\max}) \bigl(\|Q_\ell\|_2^2 - \|P_\ell\|_2^2 \bigr) \leq D_f(Q_\ell \| U_n), \tag{227}$$

where $q_{\max}$ and $q_{\min}$ denote, respectively, the maximal and minimal positive masses of Q_ℓ on the n leaves of a tree $\mathcal{S}$, and $c_f(\cdot, \cdot)$ is given in (26).

We next consider a measure which quantifies the quality of the approximation of the probability mass function P_ℓ, induced by the leaves of a Tunstall tree, by an equiprobable distribution U_n over a set whose cardinality (n) is equal to the number of leaves in the tree. To this end, consider the setup of Bayesian binary hypothesis testing where a random variable X has one of the two probability distributions

$$\begin{cases} \mathrm{H}_0: & X \sim P_\ell, \\ \mathrm{H}_1: & X \sim U_n, \end{cases} \tag{228}$$

with a-priori probabilities $\mathbb{P}[\mathrm{H}_0] = \omega$, and $\mathbb{P}[\mathrm{H}_1] = 1 - \omega$ for an arbitrary $\omega \in (0, 1)$. The measure being considered here is equal to the difference between the minimum a-priori and minimum a-posteriori error probabilities of the Bayesian binary hypothesis testing model in (228), which is close to zero if the two distributions are sufficiently close.

The difference between the minimum a-priori and minimum a-posteriori error probabilities of a general Bayesian binary hypothesis testing model with the two arbitrary alternative hypotheses $\mathrm{H}_0: X \sim P$ and $\mathrm{H}_1: X \sim Q$ with a-priori probabilities ω and $1 - \omega$, respectively, is defined to be the order-ω DeGroot statistical information $\mathcal{I}_\omega(P, Q)$ [57] (see also ([16], Definition 3)). It can be expressed as an f-divergence:

$$\mathcal{I}_\omega(P, Q) = D_{\phi_\omega}(P \| Q), \tag{229}$$

where $\phi_\omega \colon [0, \infty) \to \mathbb{R}$ is the convex function with $\phi_\omega(1) = 0$, given by (see ([16], (73)))

$$\phi_\omega(t) := \min\{\omega, 1-\omega\} - \min\{\omega, 1-\omega t\}, \quad t \geq 0. \tag{230}$$

The measure considered here for quantifying the closeness of P_ℓ to the equiprobable distribution U_n is therefore given by

$$d_{\omega,n}(P_\ell) := D_{\phi_\omega}(P_\ell \| U_n), \quad \forall\, \omega \in (0,1), \tag{231}$$

which is bounded in the interval $\big[0, \min\{\omega, 1-\omega\}\big]$.

The next result partially relies on Theorem 7.

Theorem 14. *The measure in* (231) *satisfies the following properties:*

(a) *It is the minimum of $D_{\phi_\omega}(P \| U_n)$ with respect to all probability measures $P \in \mathcal{P}_n$ that are induced by an arbitrary tree with n leaves.*

(b)

$$d_{\omega,n}(P_\ell) \leq \max_{\beta \in \Gamma_n(\rho)} D_{\phi_\omega}(Q_\beta \| U_n), \tag{232}$$

with the function $\phi_\omega(\cdot)$ in (230)*, the interval $\Gamma_n(\rho)$ in* (79)*, the probability mass function Q_β in* (80)*, and $\rho := \frac{1}{p_{\min}}$ is the reciprocal of the minimal probability of the source symbols.*

(c) *The following bound holds for every $n \in \mathbb{N}$, which is the asymptotic limit of the right side of* (232) *as we let $n \to \infty$:*

$$d_{\omega,n}(P_\ell) \leq \max_{x \in [0,1]} \left\{ x\, \phi_\omega\!\left(\frac{\rho}{1+(\rho-1)x}\right) + (1-x)\, \phi_\omega\!\left(\frac{1}{1+(\rho-1)x}\right) \right\}. \tag{233}$$

(d) *If $f \colon (0, \infty) \to \mathbb{R}$ is convex and twice differentiable, continuous at zero and $f(1) = 0$, then*

$$D_f(P_\ell \| U_n) = \int_0^1 \frac{d_{\omega,n}(P_\ell)}{\omega^3}\, f''\!\left(\frac{1-\omega}{\omega}\right) \mathrm{d}\omega. \tag{234}$$

Proof. See Appendix P.1. □

Remark 12. *The integral representation in* (234) *provides another justification for quantifying the closeness of P_ℓ to an equiprobable distribution by the measure in* (231)*.*

Figure 7 refers to the upper bound on the closeness-to-equiprobable measure $d_{\omega,n}(P_\ell)$ in (233) for Tunstall trees with n leaves. The bound holds for all $n \in \mathbb{N}$, and it is shown as a function of $\omega \in [0,1]$ for several values of $\rho \in [1, \infty]$. In the limit where $\rho \to \infty$, the upper bound is equal to $\min\{\omega, 1-\omega\}$ since the minimum a-posteriori error probability of the Bayesian binary hypothesis testing model in (228) tends to zero. On the other hand, if $\rho = 1$, then the right side of (233) is identically equal to zero (since $\phi_\omega(1) = 0$).

Theorem 14 gives an upper bound on the measure in (231), for the closeness of the probability mass function generated on the leaves by a Tunstall tree to the equiprobable distribution, where this bound is expressed as a function of the minimal probability mass of the source. The following result, which relies on ([33], Theorem 4) and our earlier analysis related to Theorem 7, provides a sufficient condition on the minimal probability mass for asserting the closeness of the compression rate to the Shannon entropy of a stationary and memoryless discrete source.

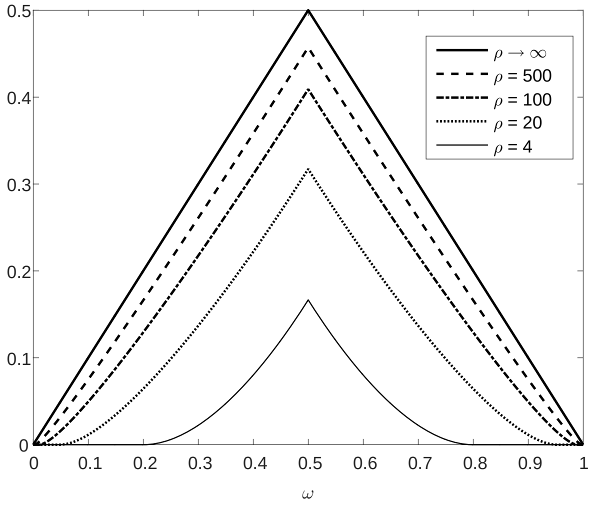

Figure 7. Curves of the upper bound on the measure $d_{\omega,n}(P_\ell)$ in (233), valid for all $n \in \mathbb{N}$, as a function of $\omega \in [0,1]$ for different values of $\rho := \frac{1}{p_{\min}}$.

Theorem 15. *Let P be a probability mass function of a stationary and memoryless discrete source, and let the emitted source symbols be from an alphabet of size $D \geq 2$. Let $\mathcal{C}$ be a Tunstall code which is used for source compression; let m and $\mathcal{X}$ denote, respectively, the fixed length and the alphabet of the codewords of $\mathcal{C}$ (where $|\mathcal{X}| \geq 2$), referring to a Tunstall tree of n leaves with $n \leq |\mathcal{X}|^m < n + (D-1)$. Let $p_{\min}$ be the minimal probability mass of the source symbols, and let*

$$d = d(m, \varepsilon) := \begin{cases} \dfrac{m\varepsilon \log_e |\mathcal{X}|}{1+\varepsilon} + \log_e\left(1 - \dfrac{D-1}{|\mathcal{X}|^m}\right), & \text{if } D > 2, \\ \dfrac{m\varepsilon \log_e |\mathcal{X}|}{1+\varepsilon}, & \text{if } D = 2, \end{cases} \tag{235}$$

with an arbitrary $\varepsilon > 0$ such that $d > 0$. If

$$p_{\min} \geq \frac{W_0\left(-e^{-d-1}\right)}{W_{-1}\left(-e^{-d-1}\right)}, \tag{236}$$

where W_0 and W_{-1} denote, respectively, the principal and secondary real branches of the Lambert W function [37], then the compression rate of the Tunstall code is larger than the Shannon entropy of the source by a factor which is at most $1+\varepsilon$.

Proof. See Appendix P.2. □

Remark 13. *The condition in* (236) *can be replaced by the stronger requirement that*

$$p_{\min} \geq \frac{1}{1+\sqrt{8d}}. \tag{237}$$

Unless d is a small fraction of unity, there is a significant difference between the condition in (236) *and the more restrictive condition in* (237) *(see Figure 8).*

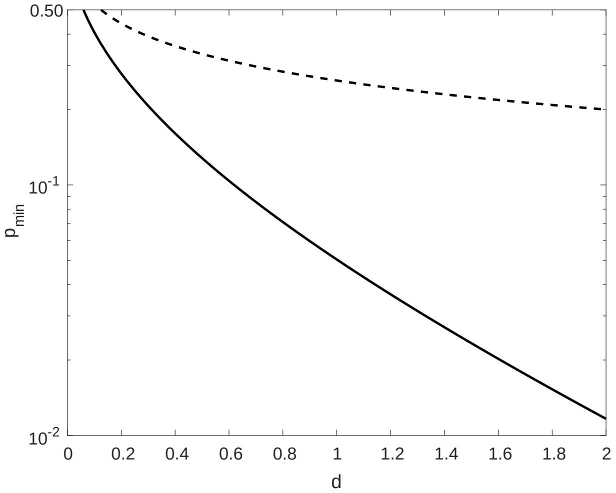

Figure 8. Curves for the smallest values of $p_{\min}$, in the setup of Theorem 15, according to the condition in (236) (solid line) and the more restrictive condition in (237) (dashed line) for binary Tunstall codes which are used to compress memoryless and stationary binary sources.

Example 3. *Consider a memoryless and stationary binary source, and a binary Tunstall code with codewords of length $m = 10$ referring to a Tunstall tree with $n = 2^m = 1024$ leaves. Letting $\varepsilon = 0.1$ in Theorem 15, it follows that if the minimal probability mass of the source satisfies $p_{\min} \geq 0.0978$ (see (235), and Figure 8 with $d = \frac{m\varepsilon \log_e 2}{1+\varepsilon} = 0.6301$), then the compression rate of the Tunstall code is at most 10% larger than the Shannon entropy of the source.*

Acknowledgments: The author wishes to thank the Guest Editor, Amos Lapidoth, and the two anonymous reviewers for an efficient process in reviewing and handling this paper.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Proof of Theorem 1

We start by proving Item (a). By our assumptions on Q_X and $W_{Y|X}$,

$$P_X(x), Q_X(x) > 0, \qquad \forall\, x \in \mathcal{X}, \tag{A1}$$

$$\sum_{x\in\mathcal{X}} W_{Y|X}(y|x) > 0, \quad \forall\, y \in \mathcal{Y}, \tag{A2}$$

$$\sum_{y\in\mathcal{Y}} W_{Y|X}(y|x) = 1, \quad \forall\, x \in \mathcal{X}, \tag{A3}$$

$$W_{Y|X}(y|x) \geq 0, \qquad \forall\, (x,y) \in \mathcal{X} \times \mathcal{Y}. \tag{A4}$$

From (20), (21), (A1), (A2) and (A4), it follows that

$$P_Y(y) = \sum_{x\in\mathcal{X}} P_X(x) W_{Y|X}(y|x) > 0, \quad \forall\, y \in \mathcal{Y}, \tag{A5}$$

$$Q_Y(y) = \sum_{x\in\mathcal{X}} Q_X(x) W_{Y|X}(y|x) > 0, \quad \forall\, y \in \mathcal{Y}, \tag{A6}$$

which imply that, for all $y \in \mathcal{Y}$,

$$\inf_{x\in\mathcal{X}} \frac{P_X(x)}{Q_X(x)} \leq \frac{P_Y(y)}{Q_Y(y)} \leq \sup_{x\in\mathcal{X}} \frac{P_X(x)}{Q_X(x)}. \tag{A7}$$

Since by assumption P_X and Q_X are supported on $\mathcal{X}$, and P_Y and Q_Y are supported on $\mathcal{Y}$ (see (A5) and (A6)), it follows that the left side inequality in (A7) is strict if the infimum in the left side is equal to 0, and the right side inequality in (A7) is strict if the supremum in the right side is equal to ∞. Hence, due to (18), (19) and (23),

$$\frac{P_X(x)}{Q_X(x)}, \frac{P_Y(y)}{Q_Y(y)} \in \mathcal{I}(\xi_1, \xi_2), \quad \forall\, (x,y) \in \mathcal{X} \times \mathcal{Y}. \tag{A8}$$

Since by assumption $f\colon (0,\infty) \to \mathbb{R}$ is convex, it follows that its right derivative $f'_+(\cdot)$ exists, and it is monotonically non-decreasing and finite on $(0,\infty)$ (see, e.g., ([58], Theorem 1.2) or ([59], Theorem 24.1)). A straightforward generalization of ([60], Theorem 1.1) (see ([60], Remark 1)) gives

$$D_f(P_X \| Q_X) - D_f(P_Y \| Q_Y) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \left\{ Q_X(x)\, W_{Y|X}(y|x)\, \Delta\left(\frac{P_X(x)}{Q_X(x)}, \frac{P_Y(y)}{Q_Y(y)} \right) \right\} \tag{A9}$$

where

$$\Delta(u,v) := f(u) - f(v) - f'_+(v)(u-v), \quad u,v > 0. \tag{A10}$$

In comparison to ([60], Theorem 1.1), the requirement that f is differentiable on $(0,\infty)$ is relaxed here, and the derivative of f is replaced by its right-side derivative. Note that if f is differentiable, then $\Delta\left(\frac{P_X(x)}{Q_X(x)}, \frac{P_Y(y)}{Q_Y(y)}\right)$ with $\Delta(\cdot,\cdot)$ as defined in (A10) is Bregman's divergence [61]. The following equality, expressed in terms of Lebesgue-Stieltjes integrals, holds by ([16], Theorem 1):

$$\begin{aligned} &\Delta\left(\frac{P_X(x)}{Q_X(x)}, \frac{P_Y(y)}{Q_Y(y)} \right) \\ &= \begin{cases} \displaystyle\int 1\left\{ s \in \left(\frac{P_Y(y)}{Q_Y(y)}, \frac{P_X(x)}{Q_X(x)} \right] \right\} \left(\frac{P_X(x)}{Q_X(x)} - s \right) \mathrm{d}f'_+(s), & \text{if } \dfrac{P_X(x)}{Q_X(x)} \geq \dfrac{P_Y(y)}{Q_Y(y)}, \\ \displaystyle\int 1\left\{ s \in \left(\frac{P_X(x)}{Q_X(x)}, \frac{P_Y(y)}{Q_Y(y)} \right] \right\} \left(s - \frac{P_X(x)}{Q_X(x)} \right) \mathrm{d}f'_+(s), & \text{if } \dfrac{P_X(x)}{Q_X(x)} < \dfrac{P_Y(y)}{Q_Y(y)}. \end{cases} \end{aligned} \tag{A11}$$

From (18), (19), (22), (A8) and (A11), if $\frac{P_X(x)}{Q_X(x)} \geq \frac{P_Y(y)}{Q_Y(y)}$, then

$$\begin{aligned} \Delta\left(\frac{P_X(x)}{Q_X(x)}, \frac{P_Y(y)}{Q_Y(y)} \right) &\geq 2c_f(\xi_1,\xi_2) \int_{\frac{P_Y(y)}{Q_Y(y)}}^{\frac{P_X(x)}{Q_X(x)}} \left(\frac{P_X(x)}{Q_X(x)} - s \right) \mathrm{d}s \\ &= c_f(\xi_1,\xi_2) \left(\frac{P_X(x)}{Q_X(x)} - \frac{P_Y(y)}{Q_Y(y)} \right)^2, \end{aligned} \tag{A12}$$

and similarly, if $\frac{P_X(x)}{Q_X(x)} < \frac{P_Y(y)}{Q_Y(y)}$, then

$$\begin{aligned} \Delta\left(\frac{P_X(x)}{Q_X(x)}, \frac{P_Y(y)}{Q_Y(y)} \right) &\geq 2c_f(\xi_1,\xi_2) \int_{\frac{P_X(x)}{Q_X(x)}}^{\frac{P_Y(y)}{Q_Y(y)}} \left(s - \frac{P_X(x)}{Q_X(x)} \right) \mathrm{d}s \\ &= c_f(\xi_1,\xi_2) \left(\frac{P_X(x)}{Q_X(x)} - \frac{P_Y(y)}{Q_Y(y)} \right)^2. \end{aligned} \tag{A13}$$

By combining (A9), (A12) and (A13), it follows that

$$D_f(P_X \| Q_X) - D_f(P_Y \| Q_Y) \geq c_f(\xi_1,\xi_2) \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \left\{ Q_X(x)\, W_{Y|X}(y|x) \left(\frac{P_X(x)}{Q_X(x)} - \frac{P_Y(y)}{Q_Y(y)} \right)^2 \right\}, \tag{A14}$$

and an evaluation of the sum in the right side of (A14) gives (see (20), (21) and (A3))

$$\begin{aligned}
&\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left\{ Q_X(x)\, W_{Y|X}(y|x) \left(\frac{P_X(x)}{Q_X(x)} - \frac{P_Y(y)}{Q_Y(y)} \right)^2 \right\} \\
&= \sum_{x\in\mathcal{X}} \left\{ \frac{P_X^2(x)}{Q_X(x)} \underbrace{\sum_{y\in\mathcal{Y}} W_{Y|X}(y|x)}_{=1} \right\} - 2 \sum_{y\in\mathcal{Y}} \left\{ \frac{P_Y(y)}{Q_Y(y)} \underbrace{\sum_{x\in\mathcal{X}} P_X(x) W_{Y|X}(y|x)}_{=P_Y(y)} \right\} \\
&\quad + \sum_{y\in\mathcal{Y}} \left\{ \frac{P_Y^2(y)}{Q_Y^2(y)} \underbrace{\sum_{x\in\mathcal{X}} Q_X(x) W_{Y|X}(y|x)}_{=Q_Y(y)} \right\} && \text{(A15)} \\
&= \sum_{x\in\mathcal{X}} \frac{P_X^2(x)}{Q_X(x)} - \sum_{y\in\mathcal{Y}} \frac{P_Y^2(y)}{Q_Y(y)} && \text{(A16)} \\
&= \sum_{x\in\mathcal{X}} \frac{\big(P_X(x)-Q_X(x)\big)^2}{Q_X(x)} - \sum_{y\in\mathcal{Y}} \frac{\big(P_Y(y)-Q_Y(y)\big)^2}{Q_Y(y)} && \text{(A17)} \\
&= \chi^2(P_X\|Q_X) - \chi^2(P_Y\|Q_Y). && \text{(A18)}
\end{aligned}$$

Combining (A14)–(A18) gives (24); (25) is due to the data-processing inequality for f-divergences (applied to the χ^2-divergence), and the non-negativity of $c_f(\xi_1, \xi_2)$ in (22).

The χ^2-divergence is an f-divergence with $f(t) = (t-1)^2$ for $t \geq 0$. The condition in (22) allows to set here $c_f(\xi_1, \xi_2) \equiv 1$, implying that (24) holds in this case with equality.

We next prove Item (b). Let f be twice differentiable on $\mathcal{I} := \mathcal{I}(\xi_1, \xi_2)$ (see (23)), and let $(u, v) \in \mathcal{I} \times \mathcal{I}$ with $v > u$. Dividing both sides of (22) by $v - u$, and letting $v \to u^+$, yields $c_f(\xi_1, \xi_2) \leq \frac{1}{2} f''(u)$. Since this holds for all $u \in \mathcal{I}$, it follows that $c_f(\xi_1, \xi_2) \leq \frac{1}{2} \inf_{t\in\mathcal{I}} f''(t)$. We next show that $c_f(\xi_1, \xi_2)$ in (26) fulfills the condition in (22), and therefore it is the largest possible value of c_f to satisfy (22). By the mean value theorem of Lagrange, for all $(u, v) \in \mathcal{I} \times \mathcal{I}$ with $v > u$, there exists an intermediate value $\xi \in (u, v)$ such that $f'(v) - f'(u) = f''(\xi)\,(v - u)$; hence, $f'(v) - f'(u) \geq 2c_f(\xi_1, \xi_2)\,(v - u)$, so the condition in (22) is indeed fulfilled with $c_f := c_f(\xi_1, \xi_2)$ as given in (26).

We next prove Item (c). Let $f^* \colon (0, \infty) \to \mathbb{R}$ be the dual convex function which is given by $f^*(t) := t f\big(\frac{1}{t}\big)$ for all $t > 0$ with $f^*(1) = f(1) = 0$. Since P_X, P_Y, Q_X and Q_Y are supported on $\mathcal{X}$ (see (A5) and (A6)), we have

$$\begin{aligned}
&D_f(P_X\|Q_X) = D_{f^*}(Q_X\|P_X), && \text{(A19)} \\
&D_f(P_Y\|Q_Y) = D_{f^*}(Q_Y\|P_Y), && \text{(A20)} \\
&\xi_1^* := \inf_{x\in\mathcal{X}} \frac{Q_X(x)}{P_X(x)} = \left(\sup_{x\in\mathcal{X}} \frac{P_X(x)}{Q_X(x)} \right)^{-1} = \frac{1}{\xi_2}, && \text{(A21)} \\
&\xi_2^* := \sup_{x\in\mathcal{X}} \frac{Q_X(x)}{P_X(x)} = \left(\inf_{x\in\mathcal{X}} \frac{P_X(x)}{Q_X(x)} \right)^{-1} = \frac{1}{\xi_1}. && \text{(A22)}
\end{aligned}$$

Consequently, it follows that

$$\begin{aligned}
D_f(P_X\|Q_X) - D_f(P_Y\|Q_Y) &= D_{f^*}(Q_X\|P_X) - D_{f^*}(Q_Y\|P_Y) && \text{(A23)} \\
&\geq c_{f^*}(\xi_1^*, \xi_2^*)\,\big[\chi^2(Q_X\|P_X) - \chi^2(Q_Y\|P_Y)\big] && \text{(A24)} \\
&= c_{f^*}\big(\tfrac{1}{\xi_2}, \tfrac{1}{\xi_1}\big)\,\big[\chi^2(Q_X\|P_X) - \chi^2(Q_Y\|P_Y)\big] && \text{(A25)}
\end{aligned}$$

where (A23) holds due to (A19) and (A20); (A24) follows from (24) with f, P_X and Q_X replaced by f^*, Q_X and P_X, respectively, which then implies that ξ_1 and ξ_2 in (18) and (19) are, respectively, replaced by ξ_1^* and ξ_2^* in (A21) and (A22); finally, (A25) holds due to (A21) and (A22). Since by assumption f is twice differentiable on $(0, \infty)$, so is f^*, and

$$(f^*)''(t) = \frac{1}{t^3} f\left(\frac{1}{t}\right), \quad t > 0. \tag{A26}$$

Hence,

$$c_{f^*}\left(\tfrac{1}{\xi_2}, \tfrac{1}{\xi_1}\right) = \tfrac{1}{2} \inf_{u \in \mathcal{I}\left(\frac{1}{\xi_2}, \frac{1}{\xi_1}\right)} (f^*)''(u) \tag{A27}$$

$$= \tfrac{1}{2} \inf_{u \in \mathcal{I}\left(\frac{1}{\xi_2}, \frac{1}{\xi_1}\right)} \left\{ \left(\frac{1}{u}\right)^3 f\left(\frac{1}{u}\right) \right\} \tag{A28}$$

$$= \tfrac{1}{2} \inf_{t \in \mathcal{I}(\xi_1, \xi_2)} \left\{ t^3 f(t) \right\} \tag{A29}$$

where (A27) follows from (24) with f, ξ_1 and ξ_2 replaced by f^*, $\frac{1}{\xi_2}$ and $\frac{1}{\xi_1}$, respectively; (A28) holds due to (A26), and (A29) holds by substituting $t =: \frac{1}{u}$. This proves (27) and (30), where (28) is due to the data-processing inequality for f-divergences, and the non-negativity of $c_{f^*}(\cdot, \cdot)$.

Similarly to the condition for equality in (24), equality in (27) is satisfied if $f^*(t) = (t-1)^2$ for all $t > 0$, or equivalently $f(t) = t f^*\left(\frac{1}{t}\right) = \frac{(t-1)^2}{t}$ for all $t > 0$. This f-divergence is Neyman's χ^2-divergence where $D_f(P \| Q) := \chi^2(Q \| P)$ for all P and Q with $c_{f^*} \equiv 1$ (due to (30), and since $t^3 f''(t) = 2$ for all $t > 0$).

The proof of Item (d) follows that same lines as the proof of Items (a)–(c) by replacing the condition in (22) with a complementary condition of the form

$$f'_+(v) - f'_+(u) \leq 2 e_f(\xi_1, \xi_2)\,(v - u), \quad \forall\, u, v \in \mathcal{I}(\xi_1, \xi_2),\ u < v. \tag{A30}$$

We finally prove Item (e) by showing that the lower and upper bounds in (24), (27), (32) and (33) 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 $\mathcal{X}$, let $P_Y^{(n)}$ and Q_Y be the probability mass functions defined on $\mathcal{Y}$ via (20) and (21) with inputs $P_X^{(n)}$ and Q_X, respectively, and let $\{\xi_{1,n}\}$ and $\{\xi_{2,n}\}$ be defined, respectively, by (18) and (19) with P_X being replaced by $P_X^{(n)}$. By the assumptions in (35) and (36),

$$\lim_{n \to \infty} \xi_{1,n} = \lim_{n \to \infty} \inf_{x \in \mathcal{X}} \frac{P_X^{(n)}(x)}{Q_X(x)} = 1, \tag{A31}$$

$$\lim_{n \to \infty} \xi_{2,n} = \lim_{n \to \infty} \sup_{x \in \mathcal{X}} \frac{P_X^{(n)}(x)}{Q_X(x)} = 1. \tag{A32}$$

Consequently, if f has a continuous second derivative at unity, then (24), (26), (31), (32), (A31) and (A32) imply that

$$\lim_{n \to \infty} \frac{D_f(P_X^{(n)} \| Q_X) - D_f(P_Y^{(n)} \| Q_Y)}{\chi^2(P_X^{(n)} \| Q_X) - \chi^2(P_Y^{(n)} \| Q_Y)}$$

$$= \lim_{n \to \infty} c_f(\xi_{1,n}, \xi_{2,n}) = \lim_{n \to \infty} e_f(\xi_{1,n}, \xi_{2,n}) = \tfrac{1}{2} f''(1), \tag{A33}$$

and similarly, from (27), (30), (33), (34), (A31) and (A32),

$$\begin{aligned}&\lim_{n\to\infty} \frac{D_f(P_X^{(n)}\|Q_X) - D_f(P_Y^{(n)}\|Q_Y)}{\chi^2(Q_X\|P_X^{(n)}) - \chi^2(Q_Y\|P_Y^{(n)})} \\ &= \lim_{n\to\infty} c_{f^*}\left(\frac{1}{\xi_{2,n}}, \frac{1}{\xi_{1,n}}\right) = \lim_{n\to\infty} e_{f^*}\left(\frac{1}{\xi_{2,n}}, \frac{1}{\xi_{1,n}}\right) = \tfrac{1}{2} f''(1), \end{aligned} \tag{A34}$$

which, respectively, prove (37) and (38).

Appendix B. Proof of Theorem 2

We start by proving Item (a). By the assumption that P_{X_i} and Q_{X_i} are supported on $\mathcal{X}$ for all $i \in \{1, \ldots, n\}$, it follows from (39) that the probability mass functions P_{X^n} and Q_{X^n} are supported on $\mathcal{X}^n$. Consequently, from (41), also $R_{X^n}^{(\lambda)}$ is supported on $\mathcal{X}^n$ for all $\lambda \in [0,1]$. Due to the product forms of Q_{X^n} and $R_{X^n}^{(\lambda)}$ in (39) and (41), respectively, we get from (47) that

$$\begin{aligned}\xi_1(n,\lambda) &= \prod_{i=1}^n \left(1 - \lambda + \lambda \inf_{x\in\mathcal{X}} \frac{P_{X_i}(x)}{Q_{X_i}(x)}\right) \\ &= \prod_{i=1}^n \left(\inf_{x\in\mathcal{X}} \frac{\lambda P_{X_i}(x) + (1-\lambda) Q_{X_i}(x)}{Q_{X_i}(x)}\right) \\ &= \inf_{\underline{x}\in\mathcal{X}^n} \left\{ \frac{\prod\limits_{i=1}^n \big(\lambda P_{X_i}(x_i) + (1-\lambda) Q_{X_i}(x_i)\big)}{\prod\limits_{i=1}^n Q_{X_i}(x_i)} \right\} \\ &= \inf_{\underline{x}\in\mathcal{X}^n} \frac{R_{X^n}^{(\lambda)}(\underline{x})}{Q_{X^n}(\underline{x})} \in (0,1], \end{aligned} \tag{A35}$$

and likewise, from (48),

$$\xi_2(n,\lambda) = \sup_{\underline{x}\in\mathcal{X}^n} \frac{R_{X^n}^{(\lambda)}(\underline{x})}{Q_{X^n}(\underline{x})} \in [1,\infty) \tag{A36}$$

for all $\lambda \in [0,1]$. In view of (24), (26), (A35) and (A36), replacing $(P_X, P_Y, Q_X, Q_Y, \xi_1, \xi_2)$ in (24) and (26) with $(R_{X^n}^{(\lambda)}, R_{Y^n}^{(\lambda)}, Q_{X^n}, Q_{Y^n}, \xi_1(n,\lambda), \xi_2(n,\lambda))$, we obtain that, for all $\lambda \in [0,1]$,

$$\begin{aligned}&D_f(R_{X^n}^{(\lambda)} \,\|\, Q_{X^n}) - D_f(R_{Y^n}^{(\lambda)} \,\|\, Q_{Y^n}) \\ &\geq c_f\big(\xi_1(n,\lambda), \xi_2(n,\lambda)\big) \left[\chi^2(R_{X^n}^{(\lambda)} \,\|\, Q_{X^n}) - \chi^2(R_{Y^n}^{(\lambda)} \,\|\, Q_{Y^n})\right]. \end{aligned} \tag{A37}$$

Due to the setting in (39)–(44), for all $\underline{y} \in \mathcal{Y}^n$ and $\lambda \in [0,1]$,

$$\begin{aligned} R^{(\lambda)}_{Y^n}(\underline{y}) &= \sum_{\underline{x} \in \mathcal{X}^n} R^{(\lambda)}_{X^n}(\underline{x})\, W_{Y^n|X^n}(\underline{y}|\underline{x}) \\ &= \sum_{\underline{x} \in \mathcal{X}^n} \left\{ \prod_{i=1}^n \bigl(\lambda P_{X_i}(x_i) + (1-\lambda) Q_{X_i}(x_i)\bigr) \prod_{i=1}^n W_{Y_i|X_i}(y_i|x_i) \right\} \\ &= \prod_{i=1}^n \left\{ \sum_{x_i \in \mathcal{X}} \Bigl\{ \bigl(\lambda P_{X_i}(x_i) + (1-\lambda) Q_{X_i}(x_i)\bigr)\, W_{Y_i|X_i}(y_i|x_i) \Bigr\} \right\} \\ &= \prod_{i=1}^n \left\{ \lambda \sum_{x \in \mathcal{X}} P_{X_i}(x) W_{Y_i|X_i}(y_i|x) + (1-\lambda) \sum_{x \in \mathcal{X}} Q_{X_i}(x) W_{Y_i|X_i}(y_i|x) \right\} \\ &= \prod_{i=1}^n \bigl(\lambda P_{Y_i}(y_i) + (1-\lambda) Q_{Y_i}(y_i)\bigr) \\ &= \prod_{i=1}^n R^{(\lambda)}_{Y_i}(y_i) \end{aligned} \tag{A38}$$

with

$$R^{(\lambda)}_{Y_i}(y) := \lambda P_{Y_i}(y) + (1-\lambda) Q_{Y_i}(y), \quad \forall\, i \in \{1, \ldots, n\},\ y \in \mathcal{Y},\ \lambda \in [0,1], \tag{A39}$$

and $R^{(\lambda)}_{Y_i}$ is the probability mass function at the channel output at time instant i. In particular, setting $\lambda = 0$ in (A38) gives

$$Q_{Y^n}(\underline{y}) = \prod_{i=1}^n Q_{Y_i}(y_i), \quad \forall\, \underline{y} \in \mathcal{Y}^n. \tag{A40}$$

Due to the tensorization property of the χ^2 divergence, and since $R^{(\lambda)}_{X^n}$, $R^{(\lambda)}_{Y^n}$, Q_{X^n} and Q_{Y^n} are product probability measures (see (39), (41), (A38) and (A40)), it follows that

$$\chi^2(R^{(\lambda)}_{X^n} \,\|\, Q_{X^n}) = \prod_{i=1}^n \Bigl(1 + \chi^2(R^{(\lambda)}_{X_i} \,\|\, Q_{X_i})\Bigr) - 1, \tag{A41}$$

and

$$\chi^2(R^{(\lambda)}_{Y^n} \,\|\, Q_{Y^n}) = \prod_{i=1}^n \Bigl(1 + \chi^2(R^{(\lambda)}_{Y_i} \,\|\, Q_{Y_i})\Bigr) - 1. \tag{A42}$$

Substituting (A41) and (A42) into the right side of (A37) gives that, for all $\lambda \in [0,1]$,

$$\begin{aligned} &D_f(R^{(\lambda)}_{X^n} \,\|\, Q_{X^n}) - D_f(R^{(\lambda)}_{Y^n} \,\|\, Q_{Y^n}) \\ &\geq c_f\bigl(\xi_1(n,\lambda), \xi_2(n,\lambda)\bigr) \left[\prod_{i=1}^n \Bigl(1 + \chi^2(R^{(\lambda)}_{X_i} \,\|\, Q_{X_i})\Bigr) - \prod_{i=1}^n \Bigl(1 + \chi^2(R^{(\lambda)}_{Y_i} \,\|\, Q_{Y_i})\Bigr) \right]. \end{aligned} \tag{A43}$$

Due to (41) and (A39), since

$$R^{(\lambda)}_{X_i} = \lambda P_{X_i} + (1-\lambda) Q_{X_i}, \tag{A44}$$

$$R^{(\lambda)}_{Y_i} = \lambda P_{Y_i} + (1-\lambda) Q_{Y_i}, \tag{A45}$$

and (see ([45], Lemma 5))

$$\chi^2(\lambda P + (1-\lambda)Q \,\|\, Q) = \lambda^2 \chi^2(P\|Q), \quad \forall\, \lambda \in [0,1] \tag{A46}$$

for every pair of probability measures (P,Q), it follows that

$$\chi^2(R_{X_i}^{(\lambda)} \,\|\, Q_{X_i}) = \lambda^2 \chi^2(P_{X_i} \,\|\, Q_{X_i}), \tag{A47}$$

$$\chi^2(R_{Y_i}^{(\lambda)} \,\|\, Q_{Y_i}) = \lambda^2 \chi^2(P_{Y_i} \,\|\, Q_{Y_i}). \tag{A48}$$

Substituting (A47) and (A48) into the right side of (A43) gives (45). For proving the looser bound (46) from (45), and also for later proving the result in Item (c), we rely on the following lemma.

Lemma A1. *Let $\{a_i\}_{i=1}^n$ and $\{b_i\}_{i=1}^n$ be non-negative with $a_i \geq b_i$ for all $i \in \{1,\ldots,n\}$. Then,*

(a) *For all $u \geq 0$,*

$$\prod_{i=1}^n (1+a_iu) - \prod_{i=1}^n (1+b_iu) \geq \sum_{i=1}^n (a_i - b_i)u. \tag{A49}$$

(b) *If $a_i > b_i$ for at least one index i, then*

$$\prod_{i=1}^n (1+a_iu) - \prod_{i=1}^n (1+b_iu) = \sum_{i=1}^n (a_i - b_i)u + O(u^2). \tag{A50}$$

Proof. Let $g\colon [0,\infty) \to \mathbb{R}$ be defined as

$$g(u) := \prod_{i=1}^n (1+a_iu) - \prod_{i=1}^n (1+b_iu), \quad \forall\, u \geq 0. \tag{A51}$$

We have $g(0) = 0$, and the first two derivatives of g are given by

$$g'(u) = \sum_{i=1}^n \Big\{ a_i \prod_{j\neq i} (1+a_ju) - b_i \prod_{j \neq i} (1+b_ju) \Big\}, \tag{A52}$$

and

$$g''(u) = \sum_{i=1}^n \sum_{j\neq i} \Big\{ a_ia_j \prod_{k\neq i,j} (1+a_ku) - b_ib_j \prod_{k \neq i,j} (1+b_ku) \Big\}. \tag{A53}$$

Since by assumption $a_i \geq b_i \geq 0$ for all i, it follows from (A53) that $g''(u) \geq 0$ for all $u \geq 0$, which asserts the convexity of g on $[0,\infty)$. Hence, for all $u \geq 0$,

$$g(u) \geq g(0) + g'(0)u = \sum_{i=1}^n (b_i - a_i)u \tag{A54}$$

where the right-side equality in (A54) is due to (A51) and (A52). This gives (A49).

We next prove Item (b) of Lemma A1. By the Taylor series expansion of the polynomial function g, we get

$$\begin{aligned} g(u) &= g(0) + g'(0)u + \tfrac{1}{2}g''(0)u^2 + \ldots \\ &= \sum_{i=1}^n (b_i - a_i)u + \tfrac{1}{2} \sum_{i=1}^n \sum_{j\neq i} (a_ia_j - b_ib_j)u^2 + \ldots \end{aligned} \tag{A55}$$

for all $u \geq 0$. Since by assumption $a_i \geq b_i \geq 0$ for all i, and there exists an index $i \in \{1, \ldots, n\}$ such that $a_i > b_i$, it follows that the coefficient of u^2 in the right side of (A55) is positive. This yields (A50). □

We obtain here (46) from (45) and Item (a) of Lemma A1. To that end, for $i \in \{1, \ldots, n\}$, let

$$a_i := \chi^2(P_{X_i} \| Q_{X_i}), \quad b_i := \chi^2(P_{Y_i} \| Q_{Y_i}), \quad u := \lambda^2 \tag{A56}$$

with $u \in [0,1]$ for every $\lambda \in [0,1]$. Since by (39), (40), (43) and (44),

$$P_{X_i} \to W_{Y_i|X_i} \to P_{Y_i}, \tag{A57}$$

$$Q_{X_i} \to W_{Y_i|X_i} \to Q_{Y_i}, \tag{A58}$$

it follows from the data-processing inequality for f-divergences, and their non-negativity, that

$$a_i \geq b_i \geq 0, \quad \forall\, i \in \{1, \ldots, n\}, \tag{A59}$$

which yields (46) from (45), (A49), (A56) and (A59).

We next prove Item (b) of Theorem 2. Similarly to the proof of (A37), we get from (32) (rather than (24)) that

$$\begin{aligned} & D_f(R^{(\lambda)}_{X^n} \,\|\, Q_{X^n}) - D_f(R^{(\lambda)}_{Y^n} \,\|\, Q_{Y^n}) \\ & \leq e_f\big(\xi_1(n,\lambda), \xi_2(n,\lambda)\big) \left[\chi^2(R^{(\lambda)}_{X^n} \,\|\, Q_{X^n}) - \chi^2(R^{(\lambda)}_{Y^n} \,\|\, Q_{Y^n})\right]. \end{aligned} \tag{A60}$$

Combining (A41), (A42), (A47), (A48) and (A60) gives (49).

We finally prove Item (c) of Theorem 2. In view of (47) and (48), and by the assumption that $\sup\limits_{x \in \mathcal{X}} \frac{P_{X_i}(x)}{Q_{X_i}(x)} < \infty$ for all $i \in \{1, \ldots, n\}$, we get

$$\lim_{\lambda \to 0^+} \xi_1(n,\lambda) = 1, \tag{A61}$$

$$\lim_{\lambda \to 0^+} \xi_2(n,\lambda) = 1. \tag{A62}$$

Since, by assumption f has a continuous second derivative at unity, (26), (31), (A61) and (A62) imply that

$$\lim_{\lambda \to 0^+} c_f\big(\xi_1(n,\lambda), \xi_2(n,\lambda)\big) = \tfrac{1}{2} f''(1), \tag{A63}$$

$$\lim_{\lambda \to 0^+} e_f\big(\xi_1(n,\lambda), \xi_2(n,\lambda)\big) = \tfrac{1}{2} f''(1). \tag{A64}$$

From (A56), (A59), and Item (b) of Lemma A1, it follows that

$$\begin{aligned} & \lim_{\lambda \to 0^+} \frac{1}{\lambda^2} \left[\prod_{i=1}^n \big(1 + \lambda^2 \chi^2(P_{X_i} \| Q_{X_i})\big) - \prod_{i=1}^n \big(1 + \lambda^2 \chi^2(P_{Y_i} \| Q_{Y_i})\big) \right] \\ & = \sum_{i=1}^n \big[\chi^2(P_{X_i} \| Q_{X_i}) - \chi^2(P_{Y_i} \| Q_{Y_i})\big]. \end{aligned} \tag{A65}$$

The result in (50) finally follows from (45), (49) and (A63)–(A65). This indeed shows that the lower bounds in the right sides of (45) and (46), and the upper bound in the right side of (49) yield a tight result as we let $\lambda \to 0^+$, leading to the limit in the right side of (50).

Appendix C. Proof of Theorems 3 and 4

Appendix C.1. Proof of Theorem 3

We first obtain a lower bound on $D_f(P_X \| Q_X)$, and then obtain an upper bound on $D_f(P_Y \| Q_Y)$.

$$\begin{aligned}
D_f(P_X \| Q_X) &= \sum_{x \in \mathcal{X}} Q_X(x)\, f\!\left(\frac{P_X(x)}{Q_X(x)}\right) && \text{(A66)}\\
&= \sum_{x \in \mathcal{X}} Q_X(x)\, \left[\frac{P_X(x)}{Q_X(x)}\, g\!\left(\frac{P_X(x)}{Q_X(x)}\right) + f(0)\right] && \text{(A67)}\\
&= f(0) + \sum_{x \in \mathcal{X}} P_X(x)\, g\!\left(\frac{P_X(x)}{Q_X(x)}\right) && \text{(A68)}\\
&\geq f(0) + g\left(\sum_{x \in \mathcal{X}} \frac{P_X^2(x)}{Q_X(x)}\right) && \text{(A69)}\\
&= f(0) + g\big(1 + \chi^2(P_X \| Q_X)\big) && \text{(A70)}\\
&\geq f(0) + g(1) + g'(1)\, \chi^2(P_X \| Q_X) && \text{(A71)}\\
&= g'(1)\, \chi^2(P_X \| Q_X) && \text{(A72)}\\
&= \big(f'(1) + f(0)\big)\, \chi^2(P_X \| Q_X), && \text{(A73)}
\end{aligned}$$

where (A67) holds by the definition of g in Theorem 3 with the assumption that $f(0) < \infty$; (A69) is due to Jensen's inequality and the convexity of g; (A70) holds by the definition of the χ^2-divergence; (A71) holds due to the convexity of g, and its differentiability at 1 (due to the differentiability of f at 1); (A72) holds since $f(0) + g(1) = f(1) = 0$; finally, (A73) holds since $f(1) = 0$ implies that $g'(1) = f'(1) + f(0)$.

By ([62], Theorem 5), it follows that

$$D_f(P_Y \| Q_Y) \leq \kappa(\xi_1, \xi_2)\, \chi^2(P_Y \| Q_Y), \tag{A74}$$

where $\kappa(\xi_1, \xi_2)$ is given in (51). Combining (A66)–(A74) yields (52). Taking suprema on both sides of (52), with respect to all probability mass functions P_X with $P_X \ll Q_X$ and $P_X \neq Q_X$, gives (53) since by the definition of $\kappa(\xi_1, \xi_2)$ in (51), it is monotonically decreasing in $\xi_1 \in [0, 1)$ and monotonically increasing in $\xi_2 \in (1, \infty]$, while (18) and (19) yield

$$\xi_1 \geq 0, \quad \xi_2 \leq \frac{1}{\min\limits_{x \in \mathcal{X}} Q_X(x)}. \tag{A75}$$

Remark A1. *The derivation in (A66)–(A73) is conceptually similar to the proof of ([24], Lemma A.2). However, the function g here is convex, and our derivation involves the χ^2-divergence.*

Remark A2. *The proof of ([26], Theorem 8) (see Proposition 3 in Section 1.1 here) relies on ([24], Lemma A.2), where the function g is required to be concave in [24,26]. This leads, in the proof of ([26], Theorem 8), to an upper bound on $D_f(P_Y \| Q_Y)$. One difference in the derivation of Theorem 3 is that our requirement on the convexity of g leads to a lower bound on $D_f(P_X \| Q_X)$, instead of an upper bound on $D_f(P_Y \| Q_Y)$. Another difference between the proofs of Theorem 3 and ([26], Theorem 8) is that we apply here the result in ([62], Theorem 5) to obtain an upper bound on $D_f(P_Y \| Q_Y)$, whereas the proof of ([26], Theorem 8) relies on a Pinsker-type inequality (see ([63], Theorem 3)) to obtain a lower bound on $D_f(P_X \| Q_X)$; the latter lower bound relies on the condition on f in (16), which is not necessary for the derivation of the bound in Theorem 3.*

Remark A3. *From ([62], Theorem 1 (b)), it follows that*

$$\sup_{P \neq Q} \frac{D_f(P\|Q)}{\chi^2(P\|Q)} = \kappa(\xi_1, \xi_2), \tag{A76}$$

with $\kappa(\xi_1, \xi_2)$ in the right side of (A76) as given in (51), and the supremum in the left side of (A76) is taken over all probability measures P and Q such that $P \neq Q$. In view of ([62], Theorem 1 (b)), the equality in (A76) holds since the functions $\widetilde{f}, \widetilde{g} \colon (0, \infty) \to \mathbb{R}$, defined as $\widetilde{f}(t) := f(t) + f'(1)(1-t)$ and $\widetilde{g}(t) := (t-1)^2$ for all $t > 0$, satisfy $D_{\widetilde{f}}(P\|Q) = D_f(P\|Q)$ and $D_{\widetilde{g}}(P\|Q) = \chi^2(P\|Q)$ for all probability measures P and Q, and since $\widetilde{f}'(1) = \widetilde{g}'(1) = 0$ while the function $\widetilde{g}$ is also strictly positive on $(0,1) \cup (1, \infty)$. Furthermore, from the proof of ([62], Theorem 1 (b)), restricting P and Q to be probability mass functions which are defined over a binary alphabet, the ratio $\frac{D_f(P\|Q)}{\chi^2(P\|Q)}$ can be made arbitrarily close to the supremum in the left side of (A76); such probability measures can be obtained as the output distributions P_Y and Q_Y of an arbitrary non-degenerate stochastic transformation $W_{Y|X} \colon \mathcal{X} \to \mathcal{Y}$, with $|\mathcal{Y}| = 2$, by a suitable selection of probability input distributions P_X and Q_X, respectively (see (A5) and (A6)). In the latter case where $|\mathcal{Y}| = 2$, this shows the optimality of the non-negative constant $\kappa(\xi_1, \xi_2)$ in the right side of (A74).

Appendix C.2. Proof of Theorem 4

Combining (A66)–(A73) gives that, for all $\lambda \in [0,1]$,

$$D_f\big(R_{X^n}^{(\lambda)} \,\|\, Q_{X^n}^{(\lambda)}\big) \geq \big(f'(1) + f(0)\big)\, \chi^2\big(R_{X^n}^{(\lambda)} \,\|\, Q_{X^n}\big), \tag{A77}$$

and from (A74)

$$D_f\big(R_{Y^n}^{(\lambda)} \,\|\, Q_{Y^n}\big) \leq \kappa\big(\xi_1(n,\lambda), \xi_2(n,\lambda)\big)\, \chi^2\big(R_{Y^n}^{(\lambda)} \,\|\, Q_{Y^n}\big). \tag{A78}$$

From (A41) and (A47),

$$\chi^2\big(R_{X^n}^{(\lambda)} \,\|\, Q_{X^n}\big) = \prod_{i=1}^{n} \Big(1 + \lambda^2 \chi^2(P_{X_i} \,\|\, Q_{X_i})\Big) - 1, \tag{A79}$$

and similarly, from (A42) and (A48),

$$\chi^2\big(R_{Y^n}^{(\lambda)} \,\|\, Q_{Y^n}\big) = \prod_{i=1}^{n} \Big(1 + \lambda^2 \chi^2(P_{Y_i} \,\|\, Q_{Y_i})\Big) - 1. \tag{A80}$$

Combining (A77)–(A80) yields (54).

Appendix D. Proof of Theorem 5

The function $f_\alpha \colon [0, \infty) \to \mathbb{R}$ in (55) satisfies $f_\alpha(1) = 0$, and for all $\alpha \geq \mathrm{e}^{-\frac{3}{2}}$

$$f_\alpha''(t) = 2\log(\alpha + t) + 3\log \mathrm{e} > 0, \quad \forall\, t > 0, \tag{A81}$$

which yields the convexity of $f_\alpha(\cdot)$ on $[0, \infty)$. This justifies the definition of the f-divergence

$$D_{f_\alpha}(P\|Q) := \sum_{x \in \mathcal{X}} Q(x)\, f_\alpha\left(\frac{P(x)}{Q(x)}\right) \tag{A82}$$

for probability mass functions P and Q, which are defined on a finite or countably infinite set $\mathcal{X}$, with Q supported on $\mathcal{X}$. In the general alphabet setting, sums and probability mass functions are, respectively, replaced by Lebesgue integrals and Radon-Nikodym derivatives. Differentiation of both sides of (A82) with respect to α gives

$$\frac{\partial}{\partial\alpha}\{D_{f_\alpha}(P\|Q)\} = \sum_{x\in\mathcal{X}} Q(x)\, r_\alpha\!\left(\frac{P(x)}{Q(x)}\right) \tag{A83}$$

where

$$\begin{align} r_\alpha(t) &:= \frac{\partial f_\alpha(t)}{\partial\alpha} \tag{A84}\\ &= 2(\alpha+t)\log(\alpha+t) - 2(\alpha+1)\log(\alpha+1) + (t-1)\log \mathrm{e}, \quad t>0. \tag{A85}\end{align}$$

The function $r_\alpha\colon (0,\infty)\to\mathbb{R}$ is convex since

$$r''_\alpha(t) = \frac{2\log \mathrm{e}}{\alpha+t} > 0, \quad \forall\, t>0, \tag{A86}$$

and $r_\alpha(1)=0$. Hence, $D_{r_\alpha}(\cdot\|\cdot)$ is an f-divergence, and it follows from (A83)–(A85) that

$$\begin{align} \frac{\partial}{\partial\alpha}\{D_{f_\alpha}(P\|Q)\} &= D_{r_\alpha}(P\|Q) \tag{A87}\\ &= 2\sum_{x\in\mathcal{X}}\left\{\big(\alpha Q(x)+P(x)\big)\,\log\left(\alpha+\frac{P(x)}{Q(x)}\right)\right\} - 2(\alpha+1)\log(\alpha+1) \tag{A88}\\ &= 2(\alpha+1)\sum_{x\in\mathcal{X}} \frac{\alpha Q(x)+P(x)}{\alpha+1}\,\log\left(\frac{\alpha Q(x)+P(x)}{(\alpha+1)\,Q(x)}\right) \tag{A89}\\ &= 2(\alpha+1)\,D\!\left(\frac{\alpha Q+P}{\alpha+1}\,\|\,Q\right) \geq 0, \tag{A90}\end{align}$$

which gives (56), so $D_{f_\alpha}(\cdot\|\cdot)$ is monotonically increasing in α. Double differentiation of both sides of (A82) with respect to α gives

$$\frac{\partial^2}{\partial\alpha^2}\{D_{f_\alpha}(P\|Q)\} = \sum_{x\in\mathcal{X}} Q(x)\, v_\alpha\!\left(\frac{P(x)}{Q(x)}\right) \tag{A91}$$

where

$$\begin{align} v_\alpha(t) &:= \frac{\partial^2 f_\alpha(t)}{\partial\alpha^2} \tag{A92}\\ &= 2\log(\alpha+t) - 2\log(\alpha+1), \quad t>0. \tag{A93}\end{align}$$

The function $v_\alpha\colon (0,\infty)\to\mathbb{R}$ is concave, and $v_\alpha(1)=0$. By referring to the f-divergence $D_{-v_\alpha}(\cdot\|\cdot)$, it follows from (A91)–(A93) that

$$\begin{align} \frac{\partial^2}{\partial\alpha^2}\{D_{f_\alpha}(P\|Q)\} &= -D_{-v_\alpha}(P\|Q) \tag{A94}\\ &= -2\sum_{x\in\mathcal{X}} Q(x)\left[\log(\alpha+1) - \log\left(\alpha+\frac{P(x)}{Q(x)}\right)\right] \tag{A95}\\ &= -2\sum_{x\in\mathcal{X}} Q(x)\log\left(\frac{(\alpha+1)Q(x)}{\alpha Q(x)+P(x)}\right) \tag{A96}\\ &= -2\,D\!\left(Q\,\|\,\frac{\alpha Q+P}{\alpha+1}\right) \leq 0, \tag{A97}\end{align}$$

which gives (57), so $D_{f_\alpha}(\cdot\|\cdot)$ is concave in α for $\alpha \geq \mathrm{e}^{-\frac{3}{2}}$. Differentiation of both sides of (A93) with respect to α gives that

$$\frac{\partial^3 f_\alpha(t)}{\partial\alpha^3} = 2\left(\frac{1}{\alpha+t} - \frac{1}{\alpha+1}\right)\log \mathrm{e}, \tag{A98}$$

which implies that

$$\frac{\partial^3}{\partial\alpha^3}\{D_{f_\alpha}(P\|Q)\} = 2\log \mathrm{e} \sum_{x\in\mathcal{X}} Q(x)\left(\frac{1}{\alpha+\frac{P(x)}{Q(x)}} - \frac{1}{\alpha+1}\right) \tag{A99}$$

$$= \frac{2\log \mathrm{e}}{\alpha+1}\left[\sum_{x\in\mathcal{X}} \frac{Q^2(x)}{\frac{\alpha Q(x)+P(x)}{\alpha+1}} - 1\right] \tag{A100}$$

$$= \frac{2\log \mathrm{e}}{\alpha+1}\cdot\chi^2\Big(Q \,\|\, \frac{\alpha Q+P}{\alpha+1}\Big) \geq 0. \tag{A101}$$

This gives (58), and it completes the proof of Item (a).

We next prove Item (b). From Item (a), the result in (59) holds for $n = 1, 2, 3$. We provide in the following a proof of (59) for all $n \geq 3$. In view of (A98), it can be verified that for $n \geq 3$,

$$\frac{\partial^n f_\alpha(t)}{\partial\alpha^n} = 2(-1)^{n-1}(n-3)!\left[\frac{1}{(\alpha+t)^{n-2}} - \frac{1}{(\alpha+1)^{n-2}}\right]\log \mathrm{e}, \tag{A102}$$

which, from (A82), implies that

$$(-1)^{n-1}\frac{\partial^n}{\partial\alpha^n}\{D_{f_\alpha}(P\|Q)\} = \sum_{x\in\mathcal{X}} Q(x)\, g_{\alpha,n}\left(\frac{P(x)}{Q(x)}\right) \tag{A103}$$

with

$$g_{\alpha,n}(t) := (-1)^{n-1}\frac{\partial^n f_\alpha(t)}{\partial\alpha^n} \tag{A104}$$

$$= 2(n-3)!\left[\frac{1}{(\alpha+t)^{n-2}} - \frac{1}{(\alpha+1)^{n-2}}\right]\log \mathrm{e}, \quad t>0. \tag{A105}$$

The function $g_{\alpha,n}\colon (0,\infty)\to\mathbb{R}$ is convex for $n\geq 3$, with $g_{\alpha,n}(1)=0$. By referring to the f-divergence $D_{g_{\alpha,n}}(\cdot\|\cdot)$, its non-negativity and (A103) imply that for all $n \geq 3$

$$(-1)^{n-1}\frac{\partial^n}{\partial\alpha^n}\{D_{f_\alpha}(P\|Q)\} = D_{g_{\alpha,n}}(P\|Q) \geq 0. \tag{A106}$$

Furthermore, we get the following explicit formula for n-th partial derivative of $D_{f_\alpha}(P\|Q)$ with respect to α for $n \geq 3$:

$$\frac{\partial^n}{\partial\alpha^n}\{D_{f_\alpha}(P\|Q)\} = (-1)^{n-1}\sum_{x\in\mathcal{X}} Q(x)\, g_{\alpha,n}\left(\frac{P(x)}{Q(x)}\right) \tag{A107}$$

$$= \frac{2(-1)^{n-1}(n-3)!\,\log \mathrm{e}}{(\alpha+1)^{n-2}}\left[\sum_{x\in\mathcal{X}}\left\{Q(x)\left(\frac{\alpha+1}{\alpha+\frac{P(x)}{Q(x)}}\right)^{n-2}\right\} - 1\right] \tag{A108}$$

$$= \frac{2(-1)^{n-1}(n-3)!\,\log \mathrm{e}}{(\alpha+1)^{n-2}}\left[\sum_{x\in\mathcal{X}}\frac{Q^{n-1}(x)}{\left(\frac{\alpha Q(x)+P(x)}{\alpha+1}\right)^{n-2}} - 1\right] \tag{A109}$$

$$= \frac{2(-1)^{n-1}(n-3)!\,\log \mathrm{e}}{(\alpha+1)^{n-2}}\left[\exp\Big((n-2)\, D_{n-1}\Big(Q \,\|\, \tfrac{\alpha Q+P}{\alpha+1}\Big)\Big) - 1\right] \tag{A110}$$

where (A107) holds due to (A103); (A108) follows from (A104), and (A110) is satisfied by the definition of the Rényi divergence [40] which is given by

$$D_\beta(P\|Q) := \frac{1}{\beta-1} \log\left(\sum_{x\in\mathcal{X}} P^\beta(x)\, Q^{1-\beta}(x)\right), \quad \forall\, \beta \in (0,1)\cup(1,\infty) \tag{A111}$$

with $D_1(P\|Q) := D(P\|Q)$ by continuous extension of $D_\beta(\cdot\|\cdot)$ at $\beta = 1$. For $n = 3$, the right side of (A110) is simplified to the right side of (58); this holds due to the identity

$$D_2(P\|Q) = \log\bigl(1 + \chi^2(P\|Q)\bigr). \tag{A112}$$

To prove Item (c), from (55), for all $t \geq 0$

$$f'_\alpha(t) = 2(\alpha+t)\log(\alpha+t) + (\alpha+t)\log \mathrm{e}, \tag{A113}$$

$$f''_\alpha(t) = 2\log(\alpha+t) + 3\log \mathrm{e}, \tag{A114}$$

$$f^{(3)}_\alpha(t) = \tfrac{2\log \mathrm{e}}{\alpha+t}, \tag{A115}$$

which implies by a Taylor series expansion of $f_\alpha(\cdot)$ that

$$f_\alpha(t) = f_\alpha(1) + f'_\alpha(1)(t-1) + \tfrac{1}{2} f''_\alpha(1)(t-1)^2 + \tfrac{1}{6} f^{(3)}_\alpha(\xi)(t-1)^3, \quad \forall\, t \geq 0 \tag{A116}$$

where ξ in the right side of (A116) is an intermediate value between 1 and t. Hence, for $t \geq 0$,

$$\begin{aligned} f_\alpha(t) &\geq f'_\alpha(1)(t-1) + \tfrac{1}{2} f''_\alpha(1)(t-1)^2 + \tfrac{1}{6} f^{(3)}_\alpha(0)(t-1)^3\, 1\{t \in [0,1]\} && \text{(A117)}\\ &\geq f'_\alpha(1)(t-1) + \bigl(\tfrac{1}{2} f''_\alpha(1) - \tfrac{1}{6} f^{(3)}_\alpha(0)\bigr)(t-1)^2 && \text{(A118)}\\ &= f'_\alpha(1)(t-1) + k(\alpha)\,(t-1)^2 && \text{(A119)} \end{aligned}$$

where (A117) follows from (A116) since $f_\alpha(1) = 0$ and $f^{(3)}_\alpha(\cdot)$ is monotonically decreasing and positive (see (A115)); $1\{t \in [0,1]\}$ in the right side of (A117) denotes the indicator function which is equal to 1 if the relation $t \in [0,1]$ holds, and it is otherwise equal to zero; (A118) holds since $(t-1)^3\, 1\{t \in [0,1]\} \geq -(t-1)^2$ for all $t \geq 0$, and $f^{(3)}_\alpha(0) > 0$; finally, (A119) follows by substituting (A114) and (A115) into the right side of (A118), which gives the equality

$$\tfrac{1}{2} f''_\alpha(1) - \tfrac{1}{6} f^{(3)}_\alpha(0) = k(\alpha) \tag{A120}$$

with $k(\cdot)$ as defined in (63). Since the first term in the right side of (A119) does not affect an f-divergence (as it is equal to $c\,(t-1)$ for $t \geq 0$ and some constant c), and for an arbitrary positive constant $k > 0$ and $g(t) := (t-1)^2$ for $t \geq 0$, we get $D_{kg}(P\|Q) = k\,\chi^2(P\|Q)$, inequality (61) follows from (A117) and (A119). To that end, note that $k = k(\alpha)$ defined in (63) is monotonically increasing in α, and therefore $k(\alpha) \geq k(\mathrm{e}^{-\frac{3}{2}}) > 0.2075$ for all $\alpha \geq \mathrm{e}^{-\frac{3}{2}}$. Due to the inequality (see, e.g., ([64], Theorem 5), followed by refined versions in ([62], Theorem 20) and ([65], Theorem 9))

$$D(P\|Q) \leq \log\bigl(1 + \chi^2(P\|Q)\bigr), \tag{A121}$$

the looser lower bound on $D_{f_\alpha}(P\|Q)$ in the right side of (62), expressed as a function of the relative entropy $D(P\|Q)$, follows from (61). Hence, if P and Q are not identical, then (64) follows from (61) since $\chi^2(P\|Q) > 0$ and $\lim\limits_{\alpha\to\infty} k(\alpha) = \infty$.

We next prove Item (d). The Taylor series expansion of $f_\alpha(\cdot)$ implies that, for all $t \geq 0$,

$$f_\alpha(t) = f_\alpha(1) + f'_\alpha(1)(t-1) + \tfrac{1}{2} f''_\alpha(1)(t-1)^2 + \tfrac{1}{6} f^{(3)}_\alpha(1)(t-1)^3 + \tfrac{1}{24} f^{(4)}_\alpha(\xi)(t-1)^4 \tag{A122}$$

where ξ in the right side of (A122) is an intermediate value between 1 and t. Consequently, since $f_\alpha^{(4)}(\xi) = -\frac{2\log e}{(\alpha+\xi)^2} < 0$ and $f_\alpha(1) = 0$, it follows from (A122) that, for all $t \geq 0$,

$$\begin{aligned} f_\alpha(t) &\leq f_\alpha'(1)(t-1) + \tfrac{1}{2} f_\alpha''(1)(t-1)^2 + \tfrac{1}{6} f_\alpha^{(3)}(1)(t-1)^3 && \text{(A123)} \\ &= f_\alpha'(1)(t-1) + \tfrac{1}{2} f_\alpha''(1)(t-1)^2 + \tfrac{1}{6} f_\alpha^{(3)}(1)\left[t^3 - 3(t-1)^2 - 3(t-1) - 1\right] && \text{(A124)} \\ &= \left[f_\alpha'(1) - \tfrac{1}{2} f_\alpha^{(3)}(1)\right](t-1) + \tfrac{1}{2}\left[f_\alpha''(1) - f_\alpha^{(3)}(1)\right](t-1)^2 + \tfrac{1}{6} f_\alpha^{(3)}(1)\,(t^3-1). && \text{(A125)} \end{aligned}$$

Based on (A123)–(A125), it follows that

$$\begin{aligned} D_{f_\alpha}(P\|Q) &\leq \tfrac{1}{2}\left[f_\alpha''(1) - f_\alpha^{(3)}(1)\right]\chi^2(P\|Q) + \tfrac{1}{6} f_\alpha^{(3)}(1) \sum_{x\in\mathcal{X}} \left\{ Q(x) \left[\left(\frac{P(x)}{Q(x)} \right)^3 - 1 \right] \right\} \\ &= \tfrac{1}{2}\left[f_\alpha''(1) - f_\alpha^{(3)}(1)\right]\chi^2(P\|Q) + \tfrac{1}{6} f_\alpha^{(3)}(1) \left(-1 + \sum_{x\in\mathcal{X}} \frac{P^3(x)}{Q^2(x)} \right) && \text{(A126)} \\ &= \tfrac{1}{2}\left[f_\alpha''(1) - f_\alpha^{(3)}(1)\right]\chi^2(P\|Q) + \tfrac{1}{6} f_\alpha^{(3)}(1) \Big[\exp\big(2D_3(P\|Q)\big) - 1\Big], && \text{(A127)} \end{aligned}$$

where (A127) holds due to (A111) (with $\beta = 3$). Substituting (A114) and (A115) into the right side of (A127) gives (65).

We next prove Item (e). Let P and Q be probability mass functions such that $D_3(P\|Q) < \infty$, and let $\varepsilon > 0$ be arbitrarily small. Since the Rényi divergence $D_\alpha(P\|Q)$ is monotonically non-decreasing in $\alpha > 0$ (see ([66], Theorem 3)), it follows that $D_2(P\|Q) < \infty$, and therefore also

$$\chi^2(P\|Q) = \exp\big(D_2(P\|Q)\big) - 1 < \infty. \tag{A128}$$

In view of (61), there exists $\alpha_1 := \alpha_1(P, Q, \varepsilon)$ such that for all $\alpha > \alpha_1$

$$D_{f_\alpha}(P\|Q) > \left(\log(\alpha+1) + \tfrac{3}{2}\log e\right)\chi^2(P\|Q) - \varepsilon, \tag{A129}$$

and, from (65), there exists $\alpha_2 := \alpha_2(P, Q, \varepsilon)$ such that for all $\alpha > \alpha_2$

$$D_{f_\alpha}(P\|Q) < \left(\log(\alpha+1) + \tfrac{3}{2}\log e\right)\chi^2(P\|Q) + \varepsilon. \tag{A130}$$

Letting $\alpha^* := \max\{\alpha_1, \alpha_2\}$ gives the result in (66) for all $\alpha > \alpha^*$.

Item (f) of Theorem 5 is a direct consequence of ([45], Lemma 4), which relies on ([67], Theorem 3). Let $g(t) := (t-1)^2$ for $t \geq 0$ (hence, $D_g(\cdot\|\cdot)$ is the χ^2 divergence). If a sequence $\{P_n\}$ converges to a probability measure Q in the sense that the condition in (67) is satisfied, and $P_n \ll Q$ for all sufficiently large n, then ([45], Lemma 4) yields

$$\lim_{n\to\infty} \frac{D_{f_\alpha}(P_n\|Q)}{\chi^2(P_n\|Q)} = \tfrac{1}{2} f_\alpha''(1), \tag{A131}$$

which gives (68) from (A114) and (A131).

We next prove Item (g). Inequality (69) is trivial. Inequality (70) is obtained as follows:

$$\begin{aligned} D_{f_\alpha}(P\|Q) - D_{f_\beta}(P\|Q) &= \int_\beta^\alpha \frac{\partial}{\partial u}\{D_{f_u}(P\|Q)\}\,\mathrm{d}u && \text{(A132)}\\ &= \int_\beta^\alpha 2(u+1)\,D\Big(\tfrac{uQ+P}{u+1}\,\|\,Q\Big)\,\mathrm{d}u && \text{(A133)}\\ &\geq \int_\beta^\alpha 2(u+1)\,\mathrm{d}u \cdot D\Big(\tfrac{\alpha Q+P}{\alpha+1}\,\|\,Q\Big) && \text{(A134)}\\ &= \big[(\alpha+1)^2 - (\beta+1)^2\big]\,D\Big(\tfrac{\alpha Q+P}{\alpha+1}\,\|\,Q\Big) && \text{(A135)}\\ &= (\alpha-\beta)(\alpha+\beta+2)\,D\Big(\tfrac{\alpha Q+P}{\alpha+1}\,\|\,Q\Big) && \text{(A136)} \end{aligned}$$

where (A133) follows from (56), and (A134) holds since the function $I\colon [0,\infty) \to [0,\infty)$ given by

$$I(u) := D\Big(\tfrac{uQ+P}{u+1}\,\|\,Q\Big), \quad u \geq 0 \tag{A137}$$

is monotonically decreasing in u (note that by increasing the value of the non-negative variable u, the probability mass function $\frac{uQ+P}{u+1}$ gets closer to Q). This gives (70).

For proving inequality (71), we obtain two upper bounds on $D_{f_\alpha}(P\|Q) - D_{f_\beta}(P\|Q)$ with $\alpha > \beta \geq \mathrm{e}^{-\frac{3}{2}}$. For the derivation of the first bound, we rely on (A83). From (A84) and (A85),

$$r_\alpha(t) = 2t\log t - s_\alpha(t), \quad t \geq 0 \tag{A138}$$

where $s_\alpha\colon (0,\infty) \to \mathbb{R}$ is given by

$$s_\alpha(t) := 2t\log t - 2(\alpha+t)\log(\alpha+t) + (1-t)\log \mathrm{e} + 2(\alpha+1)\log(\alpha+1), \quad t \geq 0, \tag{A139}$$

with the convention that $0\log 0 = 0$ (by a continuous extension of $t\log t$ at $t=0$). Since $s_\alpha(1) = 0$, and

$$s''_\alpha(t) = \frac{2\alpha}{t(\alpha+t)} > 0, \quad \forall\, t > 0, \tag{A140}$$

which implies that $s_\alpha(\cdot)$ is convex on $(0,\infty)$, we get

$$\begin{aligned} \frac{\partial}{\partial \alpha}\{D_{f_\alpha}(P\|Q)\} &= D_{r_\alpha}(P\|Q) && \text{(A141)}\\ &= 2D(P\|Q) - D_{s_\alpha}(P\|Q) && \text{(A142)}\\ &\leq 2D(P\|Q) && \text{(A143)} \end{aligned}$$

where (A141) holds due to (A83) (recall the convexity of $r_\alpha\colon (0,\infty) \to \mathbb{R}$ with $r_\alpha(1) = 0$); (A142) holds due to (A138) and since $r(t) := t\log t$ for $t > 0$ implies that $D_r(P\|Q) = D(P\|Q)$; finally, (A143) follows from the non-negativity of the f-divergence $D_{s_\alpha}(\cdot\|\cdot)$. Consequently, integration over the interval $[\beta,\alpha]$ ($\alpha > \beta$) on the left side of (A141) and the right side of (A143) gives

$$D_{f_\alpha}(P\|Q) - D_{f_\beta}(P\|Q) \leq 2(\alpha-\beta)\,D(P\|Q). \tag{A144}$$

Note that the same reasoning of (A132)–(A136) also implies that

$$D_{f_\alpha}(P\|Q) - D_{f_\beta}(P\|Q) \leq (\alpha-\beta)(\alpha+\beta+2)\,D\Big(\tfrac{\beta Q+P}{\beta+1}\,\|\,Q\Big), \tag{A145}$$

which gives a second upper bound on the left side of (A145). Taking the minimal value among the two upper bounds in the right sides of (A144) and (A145) gives (71) (see Remark A4).

We finally prove Item (h). From (55) and (A81), the function $f_\alpha \colon [0,\infty) \to \mathbb{R}$ is convex for $\alpha \geq \mathrm{e}^{-\frac{3}{2}}$ with $f_\alpha(1) = 0$, $f_\alpha(0) = \alpha^2 \log \alpha - (\alpha+1)^2 \log(\alpha+1) \in \mathbb{R}$, and it is also differentiable at 1. It is left to prove that the function $g_\alpha \colon (0,\infty) \to \mathbb{R}$, defined as $g_\alpha(t) := \frac{f_\alpha(t) - f_\alpha(0)}{t}$ for $t > 0$, is convex. From (55), the function g_α is given explicitly by

$$g_\alpha(t) = \frac{(\alpha+t)^2 \log(\alpha+t) - \alpha^2 \log \alpha}{t}, \quad t > 0, \tag{A146}$$

and its second derivative is given by

$$g''_\alpha(t) = \frac{w_\alpha(t)}{t^3}, \quad t > 0, \tag{A147}$$

with

$$w_\alpha(t) := 2\alpha^2 \log\left(1 + \frac{t}{\alpha}\right) + t(t - 2\alpha) \log \mathrm{e}, \quad t \geq 0. \tag{A148}$$

Since $w_\alpha(0) = 0$, and

$$w'_\alpha(t) = \frac{2t^2 \log \mathrm{e}}{\alpha + t} > 0, \quad \forall\, t > 0, \tag{A149}$$

it follows that $w_\alpha(t) > 0$ for all $t > 0$; hence, from (A147), $g''_\alpha(t) > 0$ for $t \in (0,\infty)$, which yields the convexity of the function $g_\alpha(\cdot)$ on $(0,\infty)$ for all $\alpha \geq 0$. This shows that, for every $\alpha \geq \mathrm{e}^{-\frac{3}{2}}$, the function $f_\alpha \colon [0,\infty) \to \mathbb{R}$ satisfies all the required conditions in Theorems 3 and 4. We proceed to calculate the function $\kappa_\alpha \colon [0,1) \times (1,\infty) \to \mathbb{R}$ in (51), which corresponds to $f := f_\alpha$, i.e., (see (72)),

$$\kappa_\alpha(\xi_1, \xi_2) = \sup_{t \in (\xi_1,1) \cup (1,\xi_2)} z_\alpha(t), \tag{A150}$$

with

$$z_\alpha(t) := \begin{cases} \dfrac{f_\alpha(t) + f'_\alpha(1)\,(1-t)}{(t-1)^2}, & t \in [0,1) \cup (1,\infty), \\ \frac{3}{2} \log \mathrm{e} + \log(\alpha+1), & t = 1, \end{cases} \tag{A151}$$

where the definition of $z_\alpha(1)$ is obtained by continuous extension of the function $z_\alpha(\cdot)$ at $t = 1$ (recall that the function $f_\alpha(\cdot)$ is given in (55)). Differentiation shows that

$$\frac{\partial\, z_\alpha(t)}{\partial t} = \frac{v_\alpha(t)}{(t-1)^4}, \quad t \in [0,1) \cup (1,\infty), \tag{A152}$$

where, for $t \geq 0$,

$$v_\alpha(t) := (2\alpha + t + 1)(t-1)^2 \log \mathrm{e} - 2(\alpha+1)(\alpha+t)(t-1) \log \frac{\alpha+t}{\alpha+1}, \tag{A153}$$

and

$$v'_\alpha(t) = (t-1)^2 \log \mathrm{e} + 2(\alpha+t)(t-1) \log \mathrm{e} - 2(\alpha+1)(2t + \alpha - 1) \log \frac{\alpha+t}{\alpha+1}, \tag{A154}$$

$$v''_\alpha(t) = 6(t-1) \log \mathrm{e} + \frac{2(\alpha+1)^2 \log \mathrm{e}}{\alpha+t} - 4(\alpha+1) \log \frac{\alpha+t}{\alpha+1}, \tag{A155}$$

$$v^{(3)}_\alpha(t) = \frac{2(t-1)(3t + 4\alpha + 1)}{(\alpha+t)^2}. \tag{A156}$$

From (A156), it follows that $v_\alpha^{(3)}(t) < 0$ if $t \in [0,1)$, $v_\alpha^{(3)}(1) = 0$, and $v_\alpha^{(3)}(t) > 0$ if $t \in (1,\infty)$. Since $v_\alpha''(\cdot)$ is therefore monotonically decreasing on $[0,1]$ and it is monotonically increasing on $[1,\infty)$, (A155) implies that

$$v_\alpha''(t) \geq v_\alpha''(1) = 2(\alpha+1)\log \mathrm{e} > 0, \quad \forall\, t \geq 0. \tag{A157}$$

Since $v_\alpha'(1) = 0$ (see (A154)), and $v_\alpha'(\cdot)$ is monotonically increasing on $[0,\infty)$, it follows that $v_\alpha'(t) < 0$ for all $t \in [0,1)$ and $v_\alpha'(t) > 0$ for all $t > 1$. This implies that $v_\alpha(t) \geq v_\alpha(1) = 0$ for all $t \geq 0$ (see (A153)); hence, from (A152), the function $z_\alpha(\cdot)$ is monotonically increasing on $[0,\infty)$, and it is continuous over this interval (see (A151)). It therefore follows from (A150) that

$$\kappa_\alpha(\xi_1, \xi_2) = z_\alpha(\xi_2), \tag{A158}$$

for every $\xi_1 \in [0,1)$ and $\xi_2 \in (1,\infty)$ (independently of ξ_1), which proves (73).

Remark A4. *None of the upper bounds in the right sides of* (A144) *and* (A145) *supersedes the other. For example, if P and Q correspond to Bernoulli(p) and Bernoulli(q), respectively, and $(\alpha,\beta,p,q) = (2,1,\frac{1}{5},\frac{2}{5})$, then the right sides of* (A144) *and* (A145) *are, respectively, equal to* $0.264\log\mathrm{e}$ *and* $0.156\log\mathrm{e}$. *If on the other hand $(\alpha,\beta,p,q) = (10,1,\frac{1}{5},\frac{2}{5})$, then the right sides of* (A144) *and* (A145) *are, respectively, equal to* $2.377\log\mathrm{e}$ *and* $3.646\log\mathrm{e}$.

Appendix E. Proof of Theorem 6

By assumption, $P \prec Q$ where the probability mass functions P and Q are defined on the set $\mathcal{A} := \{1,\ldots,n\}$. The majorization relation $P \prec Q$ is equivalent to the existence of a doubly-stochastic transformation $W_{Y|X}\colon \mathcal{A} \to \mathcal{A}$ such that (see Proposition 4)

$$Q \to W_{Y|X} \to P. \tag{A159}$$

(See, e.g., ([32], Theorem 2.1.10) or ([30], Theorem 2.B.2) or ([31], pp. 195–204)). Define

$$\mathcal{X} = \mathcal{Y} := \mathcal{A}, \quad P_X := Q, \quad Q_X := U_n. \tag{A160}$$

The probability mass functions given by

$$P_Y := P, \quad Q_Y := U_n \tag{A161}$$

satisfy, respectively, (20) and (21). The first one is obvious from (A159)–(A161); equality (21) holds due to the fact that $W_{Y|X}\colon \mathcal{A} \to \mathcal{A}$ is a doubly stochastic transformation, which implies that for all $y \in \mathcal{A}$

$$\sum_{x\in\mathcal{A}} Q_X(x) P_{Y|X}(y|x) = \frac{1}{n} \sum_{x\in\mathcal{A}} P_{Y|X}(y|x) \tag{A162}$$

$$= \frac{1}{n} = Q_Y(y). \tag{A163}$$

Since (by assumption) P_X and Q_X are supported on $\mathcal{A}$, relations (20) and (21) hold in the setting of (A159)–(A161), and $f\colon (0,\infty) \to \mathbb{R}$ is (by assumption) convex and twice differentiable, it is possible to apply the bounds in Theorem 1 (b) and (d). To that end, from (18), (19), (A160) and (A161),

$$\xi_1 = \min_{x\in\mathcal{A}} \frac{Q(x)}{\frac{1}{n}} = n q_{\min}, \tag{A164}$$

$$\xi_2 = \max_{x\in\mathcal{A}} \frac{Q(x)}{\frac{1}{n}} = n q_{\max}, \tag{A165}$$

which, from (24), (25), (32), (A160), (A161) and (A164), give that

$$\begin{aligned} & e_f(nq_{\min}, nq_{\max}) \left[\chi^2(Q\|U_n) - \chi^2(P\|U_n)\right] \\ & \geq D_f(Q\|U_n) - D_f(P\|U_n) \qquad \text{(A166)} \\ & \geq c_f(nq_{\min}, nq_{\max}) \left[\chi^2(Q\|U_n) - \chi^2(P\|U_n)\right] \qquad \text{(A167)} \\ & \geq 0. \qquad \text{(A168)} \end{aligned}$$

The difference of the χ^2 divergences in the left side of (A166) and the right side of (A167) satisfies

$$\begin{aligned} \chi^2(Q\|U_n) - \chi^2(P\|U_n) &= \sum_{x\in\mathcal{A}} \frac{Q^2(x)}{\frac{1}{n}} - \sum_{x\in\mathcal{A}} \frac{P^2(x)}{\frac{1}{n}} \\ &= n\big(\|Q\|_2^2 - \|P\|_2^2\big), \qquad \text{(A169)} \end{aligned}$$

and the substitution of (A169) into the bounds in (A166) and (A167) give the result in (74) and (75).

Let $f(t) = (t-1)^2$ for $t > 0$, which yields from (26) and (31) that $c_f(\cdot,\cdot) = e_f(\cdot,\cdot) = 1$. Since $D_f(\cdot\|\cdot) = \chi^2(\cdot\|\cdot)$, it follows from (A169) that the upper and lower bounds in the left side of (74) and the right side of (75), respectively, coincide for the χ^2-divergence; this therefore yields the tightness of these bounds in this special case.

We next prove (76). The following lower bound on the second-order Rényi entropy (a.k.a. the collision entropy) holds (see ([34], (25)–(27))):

$$H_2(Q) := -\log\big(\|Q\|_2^2\big) \geq \log \frac{4n\rho}{(1+\rho)^2}, \qquad \text{(A170)}$$

where $\frac{q_{\max}}{q_{\min}} \leq \rho$. This gives

$$\|Q\|_2^2 = \exp\big(-H_2(Q)\big) \leq \frac{(1+\rho)^2}{4n\rho}. \qquad \text{(A171)}$$

By Cauchy-Schwartz inequality $\|P\|_2^2 \geq \frac{1}{n}$ which, together with (A171), give

$$\|Q\|_2^2 - \|P\|_2^2 \leq \frac{(\rho-1)^2}{4n\rho}. \qquad \text{(A172)}$$

In view of the Schur-concavity of the Rényi entropy (see ([30], Theorem 13.F.3.a.)), the assumption $P \prec Q$ implies that

$$H_2(P) \geq H_2(Q), \qquad \text{(A173)}$$

and an exponentiation of both sides of (A173) (see the left-side equality in (A170)) gives

$$\|Q\|_2^2 \geq \|P\|_2^2. \qquad \text{(A174)}$$

Combining (A172) and (A173) gives (76).

Appendix F. Proof of Theorem 7

We prove Item (a), showing that the set $\mathcal{P}_n(\rho)$ (with $\rho \geq 1$) is non-empty, convex and compact. Note that $\mathcal{P}_n(1) = \{U_n\}$ is a singleton, so the claim is trivial for $\rho = 1$.

Let $\rho > 1$. The non-emptiness of $\mathcal{P}_n(\rho)$ is trivial since $U_n \in \mathcal{P}_n(\rho)$. To prove the convexity of $\mathcal{P}_n(\rho)$, let $P_1, P_2 \in \mathcal{P}_n(\rho)$, and let $p_{\max}^{(1)}$, $p_{\max}^{(2)}$, $p_{\min}^{(1)}$ and $p_{\min}^{(2)}$ be the (positive) maximal and minimal probability masses of P_1 and P_2, respectively. Then, $\frac{p_{\max}^{(1)}}{p_{\min}^{(1)}} \leq \rho$ and $\frac{p_{\max}^{(2)}}{p_{\min}^{(2)}} \leq \rho$ yield

$$\frac{\lambda p_{\max}^{(1)} + (1-\lambda) p_{\max}^{(2)}}{\lambda p_{\min}^{(1)} + (1-\lambda) p_{\min}^{(2)}} \leq \rho, \quad \forall \lambda \in [0,1]. \tag{A175}$$

For every $\lambda \in [0,1]$,

$$\min_{1 \leq i \leq n} \bigl\{ \lambda P_1(i) + (1-\lambda) P_2(i) \bigr\} \geq \lambda\, p_{\min}^{(1)} + (1-\lambda)\, p_{\min}^{(2)}, \tag{A176}$$

$$\max_{1 \leq i \leq n} \bigl\{ \lambda P_1(i) + (1-\lambda) P_2(i) \bigr\} \leq \lambda\, p_{\max}^{(1)} + (1-\lambda)\, p_{\max}^{(2)}. \tag{A177}$$

Combining (A175)–(A177) implies that

$$\frac{\max\limits_{1 \leq i \leq n} \bigl\{ \lambda P_1(i) + (1-\lambda) P_2(i) \bigr\}}{\min\limits_{1 \leq i \leq n} \bigl\{ \lambda P_1(i) + (1-\lambda) P_2(i) \bigr\}} \leq \rho, \tag{A178}$$

so $\lambda P_1 + (1-\lambda) P_2 \in \mathcal{P}_n(\rho)$ for all $\lambda \in [0,1]$. This proves the convexity of $\mathcal{P}_n(\rho)$.

The set of probability mass functions $\mathcal{P}_n(\rho)$ is clearly bounded; for showing its compactness, it is left to show that $\mathcal{P}_n(\rho)$ is closed. Let $\rho > 1$, and let $\{P^{(m)}\}_{m=1}^{\infty}$ be a sequence of probability mass functions in $\mathcal{P}_n(\rho)$ which pointwise converges to P over the finite set $\mathcal{A}_n$. It is required to show that $P \in \mathcal{P}_n(\rho) \subseteq \mathcal{P}_n$. As a limit of probability mass functions, $P \in \mathcal{P}_n$, and since by assumption $P^{(m)} \in \mathcal{P}_n(\rho)$ for all $m \in \mathbb{N}$, it follows that

$$(n-1)\rho p_{\min}^{(m)} + p_{\min}^{(m)} \geq (n-1) p_{\max}^{(m)} + p_{\min}^{(m)} \geq 1,$$

which yields $p_{\min}^{(m)} \geq \frac{1}{(n-1)\rho+1}$ for all m. Since $p_{\max}^{(m)} \leq \rho p_{\min}^{(m)}$ for every m, it follows that also for the limiting probability mass function P we have $p_{\min} \geq \frac{1}{(n-1)\rho+1} > 0$, and $p_{\max} \leq \rho p_{\min}$. This proves that $P \in \mathcal{P}_n(\rho)$, and therefore $\mathcal{P}_n(\rho)$ is a closed set.

An alternative proof for Item (a) relies on the observation that, for $\rho \geq 1$,

$$\mathcal{P}_n(\rho) = \mathcal{P}_n \bigcap \left\{ \bigcap_{i \neq j} \{P : \, P(i) - \rho P(j) \leq 0\} \right\}, \tag{A179}$$

which yields the convexity and compactness of the set $\mathcal{P}_n(\rho)$ for all $\rho \geq 1$.

The result in Item (b) holds in view of Item (a), and due to the convexity and continuity of $D_f(P\|Q)$ in $(P,Q) \in \mathcal{P}_n(\rho) \times \mathcal{P}_n(\rho)$ (where $p_{\min}, q_{\min} \geq \frac{1}{(n-1)\rho+1} > 0$). This implication is justified by the statement that a convex and continuous function over a non-empty convex and compact set attains its supremum over this set (see, e.g., ([68], Theorem 7.42) or ([59], Theorem 10.1 and Corollary 32.3.2)).

We next prove Item (c). If $Q \in \mathcal{P}_n(\rho)$, then $\frac{1}{1+(n-1)\rho} \leq q_{\min} \leq \frac{1}{n}$ where the lower bound on $q_{\min}$ is attained when Q is the probability mass function with $n-1$ masses equal to $\rho q_{\min}$ and a single smaller mass equal to $q_{\min}$, and the upper bound is attained when Q is the equiprobable distribution. For an arbitrary $Q \in \mathcal{P}_n(\rho)$, let $q_{\min} := \beta$ where β can get any value in the interval $\Gamma_n(\rho)$ defined in (79). By ([34], Lemma 1), $Q \prec Q_\beta$ and $Q_\beta \in \mathcal{P}_n(\rho)$ where Q_β is given in (80). The Schur-convexity of $D_f(\cdot\|U_n)$ (see ([38], Lemma 1)) and the identity $D_f(U_n\|\cdot) = D_{f^*}(\cdot\|U_n)$ give that

$$D_f(Q\|U_n) \leq D_f(Q_\beta\|U_n), \quad D_f(U_n\|Q) \leq D_f(U_n\|Q_\beta) \tag{A180}$$

for all $Q \in \mathcal{P}_n(\rho)$ with $q_{\min} = \beta \in \Gamma_n(\rho)$; equalities hold in (A180) if $Q = Q_\beta \in \mathcal{P}_n(\rho)$. The maximization of $D_f(Q\|U_n)$ and $D_f(U_n\|Q)$ over all probability mass functions $Q \in \mathcal{P}_n(\rho)$ can be therefore simplified to the maximization of $D_f(Q_\beta\|U_n)$ and $D_f(U_n\|Q_\beta)$, respectively, over the parameter β which lies in the interval $\Gamma_n(\rho)$ in (79). This proves (82) and (83).

We next prove Item (e), and then prove Item (d). In view of Item (c), the maximum of $D_f(Q\|U_n)$ over all the probability mass functions $Q \in \mathcal{P}_n(\rho)$ is attained by $Q = Q_\beta$ with $\beta \in \Gamma_n(\rho)$ (see (79)–(81)). From (80), Q_β can be expressed as the n-length probability vector

$$Q_\beta = (\underbrace{\rho\beta, \ldots, \rho\beta}_{i_\beta}, 1 - (n + i_\beta\rho - i_\beta - 1)\beta, \underbrace{\beta, \ldots, \beta}_{n-i_\beta-1}). \tag{A181}$$

The influence of the $(i_\beta + 1)$-th entry of the probability vector in (A181) on $D_f(Q_\beta\|U_n)$ tends to zero as we let $n \to \infty$. This holds since the entries of the vector in (A181) are written in decreasing order, which implies that for all $\beta \in \Gamma_n(\rho)$ (with $\rho \geq 1$)

$$n\big[1 - (n + i_\beta\rho - i_\beta - 1)\big] \in [n\beta, n\rho\beta] \subseteq \left[\tfrac{n}{(n-1)\rho+1}, \rho\right] \subseteq \left[\tfrac{1}{\rho}, \rho\right]; \tag{A182}$$

from (A182) and the convexity of f on $(0, \infty)$ (so, f attains its finite maximum on every closed sub-interval of $(0, \infty)$), it follows that

$$\begin{aligned} &\Big|\big[1 - (n + i_\beta\rho - i_\beta - 1)\beta\big]\, f\big(n\big[1 - (n + i_\beta\rho - i_\beta - 1)\big]\big)\Big| \\ &\leq \Big|\big[1 - (n + i_\beta\rho - i_\beta - 1)\beta\big]\Big| \max_{u \in \left[\frac{1}{\rho}, \rho\right]} |f(u)| \\ &\leq \frac{\rho}{n} \max_{u \in \left[\frac{1}{\rho}, \rho\right]} |f(u)| \xrightarrow[n \to \infty]{} 0. \end{aligned} \tag{A183}$$

In view of (A181) and (A183), by letting $n \to \infty$, the maximization of $D_f(Q_\beta\|U_n)$ over $\beta \in \Gamma_n(\rho)$ can be replaced by a maximization of $D_f(\widetilde{Q}_m\|U_n)$ where

$$\widetilde{Q}_m := (\underbrace{\rho\beta, \ldots, \rho\beta}_{m}, \underbrace{\beta, \ldots, \beta}_{n-m}) \in \mathcal{P}_n(\rho) \tag{A184}$$

with the free parameter $m \in \{0, \ldots, n\}$, and with $\beta := \frac{1}{n+(\rho-1)m}$ (the value of β is determined so that the total mass of $\widetilde{Q}_m$ is 1). Hence, we get

$$\lim_{n\to\infty} \max_{\beta \in \Gamma_n(\rho)} D_f(Q_\beta\|U_n) = \lim_{n\to\infty} \max_{m \in \{0,\ldots,n\}} D_f(\widetilde{Q}_m\|U_n). \tag{A185}$$

The f-divergence in the right side of (A185) satisfies

$$\begin{aligned} D_f(\widetilde{Q}_m\|U_n) &= \frac{1}{n}\sum_{i=1}^{n} f\big(n\,\widetilde{Q}_m(i)\big) && \text{(A186)} \\ &= \frac{m}{n}\, f\left(\frac{\rho n}{n + (\rho-1)m}\right) + \left(1 - \frac{m}{n}\right) f\left(\frac{n}{n + (\rho-1)m}\right) && \text{(A187)} \\ &= g_f^{(\rho)}\left(\frac{m}{n}\right), && \text{(A188)} \end{aligned}$$

where (A188) holds by the definition of the function $g_f^{(\rho)}(\cdot)$ in (84). It therefore follows that

$$\begin{aligned} &\lim_{n\to\infty} u_f(n,\rho) \\ &= \lim_{n\to\infty} \max_{m\in\{0,\ldots,n\}} g_f^{(\rho)}\left(\frac{m}{n}\right) && \text{(A189)} \\ &= \max_{x\in[0,1]} g_f^{(\rho)}(x) && \text{(A190)} \end{aligned}$$

where (A189) holds by combining (82) and (A185)–(A188); (A190) holds by the continuity of the function $g_f^{(\rho)}(\cdot)$ on $[0,1]$, which follows from (84) and the continuity of the convex function f on $[\frac{1}{\rho},\rho]$ for $\rho \geq 1$ (recall that a convex function is continuous on every closed sub-interval of its domain of region, and by assumption f is convex on $(0,\infty)$). This proves (87), by the definition of $g_f^{(\rho)}(\cdot)$ in (84).

Equality (88) follows from (87) by replacing $g_f^{(\rho)}(\cdot)$ with $g_{f^*}^{(\rho)}(\cdot)$, with $f^*\colon (0,\infty) \to \mathbb{R}$ as given in (29); this replacement is justified by the equality $D_f(U_n\|Q) = D_{f^*}(Q\|U_n)$.

Once Item (e) is proved, we return to prove Item (d). To that end, it is first shown that

$$u_f(n,\rho) \leq u_f(2n,\rho), \tag{A191}$$
$$v_f(n,\rho) \leq v_f(2n,\rho), \tag{A192}$$

for all $\rho \geq 1$ and integers $n \geq 2$, with the functions u_f and v_f, respectively, defined in (77) and (78). Since $D_f(P\|Q) = D_{f^*}(Q\|P)$ for all $P, Q \in \mathcal{P}_n$, (77) and (78) give that

$$v_f(n,\rho) = u_{f^*}(n,\rho), \tag{A193}$$

so the monotonicity property in (A192) follows from (A191) by replacing f with f^*. To prove (A191), let $Q^* \in \mathcal{P}_n(\rho)$ be a probability mass function which attains the maximum at the right side of (77), and let P^* be the probability mass function supported on $\mathcal{A}_{2n} = \{1,\ldots,2n\}$, and defined as follows:

$$P^*(i) = \begin{cases} \frac{1}{2}Q^*(i), & \text{if } i \in \{1,\ldots,n\}, \\ \frac{1}{2}Q^*(i-n), & \text{if } i \in \{n+1,\ldots,2n\}. \end{cases} \tag{A194}$$

Since by assumption $Q^* \in \mathcal{P}_n(\rho)$, (A194) implies that $P^* \in \mathcal{P}_{2n}(\rho)$. It therefore follows that

$$\begin{aligned} u_f(2n,\rho) &= \max_{Q\in\mathcal{P}_{2n}(\rho)} D_f(Q\|U_{2n}) && \text{(A195)} \\ &\geq D_f(P^*\|U_{2n}) && \text{(A196)} \\ &= \frac{1}{2n}\left[\sum_{i=1}^{n} f\big(2nP^*(i)\big) + \sum_{i=n+1}^{2n} f\big(2nP^*(i)\big)\right] && \text{(A197)} \\ &= \frac{1}{n}\sum_{i=1}^{n} f\big(nQ^*(i)\big) && \text{(A198)} \\ &= D_f(Q^*\|U_n) && \text{(A199)} \\ &= \max_{Q\in\mathcal{P}_n(\rho)} D_f(Q\|U_{2n}) && \text{(A200)} \\ &= u_f(n,\rho) && \text{(A201)} \end{aligned}$$

where (A195) and (A201) hold due to (77); (A196) holds since $P^* \in \mathcal{P}_{2n}(\rho)$; finally, (A198) holds due to (A194), which implies that the two sums in the right side of (A197) are identical, and they equal to the sum in the right side of (A198). This gives (A191), and likewise also (A192) (see (A193)).

$$\begin{aligned} u_f(n,\rho) &\leq \lim_{k\to\infty} u_f(2^k n,\rho) && \text{(A202)} \\ &= \lim_{n'\to\infty} u_f(n',\rho) && \text{(A203)} \\ &= \max_{x\in[0,1]} g_f^{(\rho)}(x) && \text{(A204)} \end{aligned}$$

where (A202) holds since, due to (A191), the sequence $\{u_f(2^k n,\rho)\}_{k=0}^{\infty}$ is monotonically increasing, which implies that the first term of this sequence is less than or equal to its limit. Equality (A203) holds since the limit in its right side exists (in view of the above proof of (87)), so its limit coincides with the limit of every subsequence; (A204) holds due to (A189) and (A190). A replacement of f with f^* gives, from (A193), that

$$v_f(n,\rho) \leq \max_{x\in[0,1]} g_{f^*}^{(\rho)}(x). \tag{A205}$$

Combining (A202)–(A205) gives the right-side inequalities in (85) and (86). The left-side inequality in (85) follows by combining (77), (A184) and (A186)–(A188), which gives

$$\begin{aligned} u_f(n,\rho) &= \max_{Q\in\mathcal{P}_n(\rho)} D_f(Q\|U_n) && \text{(A206)} \\ &\geq \max_{m\in\{0,\ldots,n\}} D_f(\widetilde{Q}_m\|U_n) && \text{(A207)} \\ &= \max_{m\in\{0,\ldots,n\}} g_f^{(\rho)}\left(\frac{m}{n}\right). && \text{(A208)} \end{aligned}$$

Likewise, in view of (A193), the left-side inequality in (86) follows from the left-side inequality in (85) by replacing f with f^*.

We next prove Item (f), providing an upper bound on the convergence rate of the limit in (87); an analogous result can be obtained for the convergence rate to the limit in (88) by replacing f with f^* in (29). To prove (89), in view of Items (d) and (e), we get that for every integer $n \geq 2$

$$\begin{aligned} 0 &\leq \lim_{n'\to\infty}\left\{u_f(n',\rho)\right\} - u_f(n,\rho) && \text{(A209)} \\ &\leq \max_{x\in[0,1]} g_f^{(\rho)}(x) - \max_{m\in\{0,\ldots,n\}} g_f^{(\rho)}\left(\frac{m}{n}\right) && \text{(A210)} \\ &= \max_{x\in[0,1]} g_f^{(\rho)}(x) - \max_{m\in\{0,\ldots,n-1\}} g_f^{(\rho)}\left(\frac{m}{n}\right) && \text{(A211)} \\ &= \max_{m\in\{0,\ldots,n-1\}}\left\{\max_{x\in\left[\frac{m}{n},\frac{m+1}{n}\right]} g_f^{(\rho)}(x)\right\} - \max_{m\in\{0,\ldots,n-1\}} g_f^{(\rho)}\left(\frac{m}{n}\right) && \text{(A212)} \\ &\leq \max_{m\in\{0,\ldots,n-1\}}\left\{\max_{x\in\left[\frac{m}{n},\frac{m+1}{n}\right]}\left\{g_f^{(\rho)}(x) - g_f^{(\rho)}\left(\frac{m}{n}\right)\right\}\right\} && \text{(A213)} \end{aligned}$$

where (A209) holds due to monotonicity property in (A191), and also due to the existence of the limit of $\{u_f(n',\rho)\}_{n'\in\mathbb{N}}$; (A210) holds due to (85); (A211) holds since the function $g_f^{(\rho)}\colon [0,1]\to\mathbb{R}$ (as it is defined in (84)) satisfies $g_f^{(\rho)}(1) = g_f^{(\rho)}(0) = 0$ (recall that by assumption $f(1) = 0$); (A212) holds since $[0,1] = \bigcup_{m=1}^{n-1}\left[\frac{m}{n},\frac{m+1}{n}\right]$, so the maximization of $g_f^{(\rho)}(\cdot)$ over the interval $[0,1]$ is the maximum over the maximal values over the sub-intervals $\left[\frac{m}{n},\frac{m+1}{n}\right]$ for $m\in\{0,\ldots,n-1\}$; finally, (A213) holds since the maximum of a sum of functions is less than or equal to the sum of the maxima of these functions. If the function $g_f^{(\rho)} \colon [0,1] \to \mathbb{R}$ is differentiable on $(0,1)$, and its derivative is upper bounded by $K_f(\rho) \geq 0$, then by the mean value theorem of Lagrange, for every $m \in \{0, \ldots, n-1\}$,

$$g_f^{(\rho)}(x) - g_f^{(\rho)}\left(\frac{m}{n}\right) \leq \frac{K_f(\rho)}{n}, \quad \forall\, x \in \left[\frac{m}{n}, \frac{m+1}{n}\right]. \tag{A214}$$

Combining (A209)–(A214) gives (89).

We next prove Item (g). By definition, it readily follows that $\mathcal{P}_n(\rho_1) \subseteq \mathcal{P}_n(\rho_2)$ if $1 \leq \rho_1 < \rho_2$. By the definition in (77), for a fixed integer $n \geq 2$, it follows that the function $u_f(n, \cdot)$ is monotonically increasing on $[1, \infty)$. The limit in the left side of (90) therefore exists. Since $D_f(Q \| U_n)$ is convex in Q, its maximum over the convex set of probability mass functions $Q \in \mathcal{P}_n$ is obtained at one of the vertices of the simplex $\mathcal{P}_n$. Hence, a maximum of $D_f(Q \| U_n)$ over this set is attained at $Q^* = (q_1^*, \ldots, q_n^*)$ with $q_i^* = 1$ for some $i \in \{1, \ldots, n\}$, and $q_j^* = 0$ for $j \neq i$. In the latter case,

$$D_f(Q^* \| U_n) = \frac{1}{n} \sum_{k=1}^{n} f(nq_k^*) = \frac{1}{n} \big[(n-1) f(0) + f(n) \big]. \tag{A215}$$

Note that $Q^* \notin \bigcup\limits_{\rho \geq 1} \mathcal{P}_n(\rho)$ (since the union of $\{\mathcal{P}_n(\rho)\}$, for all $\rho \geq 1$, includes all the probability mass functions in $\mathcal{P}_n$ which are *supported* on $\mathcal{A}_n = \{1, \ldots, n\}$, so $Q^* \in \mathcal{P}_n$ is not an element of this union); hence, it follows that

$$\lim_{\rho \to \infty} u_f(n, \rho) \leq \left(1 - \frac{1}{n}\right) f(0) + \frac{f(n)}{n}. \tag{A216}$$

On the other hand, for every $\rho \geq 1$,

$$\begin{align} u_f(n, \rho) &\geq g_f^{(\rho)}\left(\frac{1}{n}\right) \tag{A217} \\ &= \frac{1}{n} f\left(\frac{\rho n}{n + \rho - 1}\right) + \left(1 - \frac{1}{n}\right) f\left(\frac{n}{n + \rho - 1}\right) \tag{A218} \end{align}$$

where (A217) holds due to the left-side inequality of (85), and (A218) is due to (84). Combining (A217) and (A218), and the continuity of f at zero (by the continuous extension of the convex function f at zero), yields (by letting $\rho \to \infty$)

$$\lim_{\rho \to \infty} u_f(n, \rho) \geq \left(1 - \frac{1}{n}\right) f(0) + \frac{f(n)}{n}. \tag{A219}$$

Combining (A216) and (A219) gives (90) for every integer $n \geq 2$. In order to get an upper bound on the convergence rate in (90), suppose that $f(0) < \infty$, f is differentiable on $(0, n)$, and $K_n := \sup\limits_{t \in (0,n)} |f'(t)| < \infty$. For every $\rho \geq 1$, we get

$$\begin{align} 0 &\leq \lim_{\rho' \to \infty} \left\{ u_f(n, \rho') \right\} - u_f(n, \rho) \tag{A220} \\ &\leq \frac{1}{n} \left[f(n) - f\left(\frac{\rho n}{n + \rho - 1}\right) \right] + \left(1 - \frac{1}{n}\right) \left[f(0) - f\left(\frac{n}{n + \rho - 1}\right) \right] \tag{A221} \\ &\leq \frac{K_n}{n} \left(n - \frac{\rho n}{n + \rho - 1} \right) + \left(1 - \frac{1}{n}\right) \frac{K_n\, n}{n + \rho - 1} \tag{A222} \\ &= \frac{2K_n\, (n-1)}{n + \rho - 1}, \tag{A223} \end{align}$$

where (A220) holds since the sets $\{\mathcal{P}_n(\rho)\}_{\rho\geq 1}$ are monotonically increasing in ρ; (A221) follows from (A216)–(A218); (A222) holds by the assumption that $|f'(t)| \leq K_n$ for all $t \in (0, n)$, by the mean value theorem of Lagrange, and since $0 < \frac{n}{n+\rho-1} \leq \frac{\rho n}{n+\rho-1} \leq n$ for all $\rho \geq 1$ and $n \in \mathbb{N}$. This proves (91).

We next prove Item (h). Setting $P := U_n$ yields $P \prec Q$ for every probability mass function Q which is supported on $\{1, \ldots, n\}$. Since $q_{\min} + (n-1)q_{\max} \geq 1$ and $(n-1)q_{\min} + q_{\max} \leq 1$, and since by assumption $\frac{q_{\max}}{q_{\min}} \leq \rho$, it follows that

$$[nq_{\min}, nq_{\max}] \subseteq \left[\frac{n}{1+(n-1)\rho}, \frac{\rho n}{n-1+\rho}\right] \subseteq \left[\frac{1}{\rho}, \rho\right]. \tag{A224}$$

Combining the assumption in (92) with (A224) implies that

$$m \leq f''(t) \leq M, \quad \forall\, t \in [nq_{\min}, nq_{\max}]. \tag{A225}$$

Hence, (26), (31) and (A225) yield

$$\tfrac{1}{2}\, m \leq c_f(nq_{\min}, nq_{\max}) \leq e_f(nq_{\min}, nq_{\max}) \leq \tfrac{1}{2}\, M. \tag{A226}$$

The lower bound on $D_f(Q\|U_n)$ in the left side of (94) follows from a combination of (75), the left-side inequality in (A226), and $\|P\|_2^2 = \frac{1}{n}$. Similarly, the upper bound on $D_f(Q\|U_n)$ in the right side of (95) follows from a combination of (74), the right-side inequality in (A226), and the equality $\|P\|_2^2 = \frac{1}{n}$. The looser upper bound on $D_f(Q\|U_n)$ in the right side of (96), expressed as a function of M and ρ, follows by combining (74), (76), and the right-side inequality in (A226).

The tightness of the lower bound in the left side of (94) and the upper bound in the right side of (95) for the χ^2 divergence is clear from the fact that $M = m = 2$ if $f(t) = (t-1)^2$ for all $t > 0$; in this case, $\chi^2(Q\|U_n) = n\|Q\|_2^2 - 1$.

To prove Item (i), suppose that the second derivative of f is upper bounded on $(0, \infty)$ with $f''(t) \leq M_f \in (0, \infty)$ for all $t > 0$, and there is a need to assert that $D_f(Q\|U_n) \leq d$ for an arbitrary $d > 0$. Condition (97) follows from (96) by solving the inequality $\frac{M_f\,(\rho-1)^2}{8\rho} \leq d$, with the variable $\rho \geq 1$, for given $d > 0$ and $M_f > 0$ (note that M_f does not depend on ρ).

Appendix G. Proof of Theorem 8

The proof of Theorem 8 relies on Theorem 6. For $\alpha \in (0,1) \cup (1, \infty)$, let $u_\alpha \colon (0, \infty) \to \mathbb{R}$ be the non-negative and convex function given by (see, e.g., ([8], (2.1)) or ([16], (17)))

$$u_\alpha(t) := \frac{t^\alpha - \alpha(t-1) - 1}{\alpha(\alpha-1)}, \quad t > 0, \tag{A227}$$

and let $u_1 \colon (0, \infty) \to \mathbb{R}$ be the convex function given by

$$u_1(t) := \lim_{\alpha \to 1} u_\alpha(t) = t \log_e t + 1 - t, \quad t > 0. \tag{A228}$$

Let P and Q be probability mass functions which are supported on a finite set; without loss of generality, let their support be given by $\mathcal{A}_n := \{1, \ldots, n\}$. Then, for $\alpha \in (0,1) \cup (1, \infty)$,

$$\begin{aligned}
& D_{u_\alpha}(Q\|U_n) - D_{u_\alpha}(P\|U_n) \\
&= \frac{1}{n}\sum_{i=1}^{n} u_\alpha\big(nQ(i)\big) - \frac{1}{n}\sum_{i=1}^{n} u_\alpha\big(nP(i)\big) \\
&= \frac{n^{\alpha-1}}{\alpha(\alpha-1)}\left[\sum_{i=1}^{n} Q^\alpha(i) - \sum_{i=1}^{n} P^\alpha(i)\right] \\
&= \frac{n^{\alpha-1}\big[S_\alpha(P) - S_\alpha(Q)\big]}{\alpha},
\end{aligned} \tag{A229}$$

where

$$S_\alpha(P) := \begin{cases} \dfrac{1}{1-\alpha}\left(\displaystyle\sum_{i=1}^{n} P^\alpha(i) - 1\right), & \alpha \in (0,1)\cup(1,\infty), \\ -\displaystyle\sum_{i=1}^{n} P(i)\,\log_{\mathrm{e}} P(i), & \alpha = 1. \end{cases} \tag{A230}$$

designates the order-α Tsallis entropy of a probability mass P defined on the set $\mathcal{A}_n$. Equality (A229) also holds for $\alpha = 1$ by continuous extension.

In view of (26) and (31), since $u''_\alpha(t) = t^{\alpha-2}$ for all $t > 0$, it follows that

$$c_{u_\alpha}(nq_{\min}, nq_{\max}) = \begin{cases} \frac{1}{2}\, n^{\alpha-2}\, q_{\max}^{\alpha-2}, & \text{if } \alpha \in (0,2], \\ \frac{1}{2}\, n^{\alpha-2}\, q_{\min}^{\alpha-2}, & \text{if } \alpha \in (2,\infty), \end{cases} \tag{A231}$$

and

$$e_{u_\alpha}(nq_{\min}, nq_{\max}) = \begin{cases} \frac{1}{2}\, n^{\alpha-2}\, q_{\min}^{\alpha-2}, & \text{if } \alpha \in (0,2], \\ \frac{1}{2}\, n^{\alpha-2}\, q_{\max}^{\alpha-2}, & \text{if } \alpha \in (2,\infty). \end{cases} \tag{A232}$$

The combination of (74) and (75) under the assumption that P and Q are supported on $\mathcal{A}_n$ and $P \prec Q$, together with (A229), (A231) and (A232) gives (100)–(102). Furthermore, the left and right-side inequalities in (100) hold with equality if $c_{u_\alpha}(\cdot,\cdot)$ in (A231) and $e_{u_\alpha}(\cdot,\cdot)$ in (A232) coincide, which implies that the upper and lower bounds in (74) and (75) are tight in that case. Comparing $c_{u_\alpha}(\cdot,\cdot)$ in (A231) and $e_{u_\alpha}(\cdot,\cdot)$ in (A232) shows that they coincide if $\alpha = 2$.

To prove Item (b) of Theorem 8, let P_ε and Q_ε be probability mass functions supported on $\mathcal{A} = \{0,1\}$ where $P_\varepsilon(0) = \frac{1}{2} + \varepsilon$, $Q_\varepsilon(0) = \frac{1}{2} + \beta\varepsilon$, and $\beta > 1$ and $0 < \varepsilon < \frac{1}{2\beta}$. This yields $P_\varepsilon \prec Q_\varepsilon$. The result in (103) is proved by showing that, for all $\alpha > 0$,

$$\lim_{\varepsilon\to 0^+} \frac{S_\alpha(P_\varepsilon) - S_\alpha(Q_\varepsilon)}{L(\alpha, P_\varepsilon, Q_\varepsilon)} = 1, \tag{A233}$$

$$\lim_{\varepsilon\to 0^+} \frac{S_\alpha(P_\varepsilon) - S_\alpha(Q_\varepsilon)}{U(\alpha, P_\varepsilon, Q_\varepsilon)} = 1, \tag{A234}$$

which shows that the infimum and supremum in (103) can be even restricted to the binary alphabet setting. For every $\alpha \in (0,1)\cup(1,\infty)$,

$$\begin{aligned}
S_\alpha(P_\varepsilon) - S_\alpha(Q_\varepsilon) &= \frac{1}{1-\alpha}\left(\sum_i P_\varepsilon^\alpha(i) - \sum_i Q_\varepsilon^\alpha(i)\right) \\
&= \frac{1}{1-\alpha}\left[\left(\tfrac{1}{2}+\varepsilon\right)^\alpha + \left(\tfrac{1}{2}-\varepsilon\right)^\alpha - \left(\tfrac{1}{2}+\beta\varepsilon\right)^\alpha - \left(\tfrac{1}{2}-\beta\varepsilon\right)^\alpha\right] \\
&= \alpha 2^{2-\alpha}(\beta^2-1)\varepsilon^2 + O(\varepsilon^4), \qquad \text{(A235)}
\end{aligned}$$

where (A235) follows from a Taylor series expansion around $\varepsilon = 0$, and the passage in the limit where $\alpha \to 1$ shows that (A235) also holds at $\alpha = 1$ (due to the continuous extension of the order-α Tsallis entropy at $\alpha = 1$). This implies that (A235) holds for all $\alpha > 0$. We now calculate the lower and upper bounds on $S_\alpha(P_\varepsilon) - S_\alpha(Q_\varepsilon)$ in (101) and (102), respectively.

- For $\alpha \in (0, 2]$,

$$\begin{aligned}
L(\alpha, P_\varepsilon, Q_\varepsilon) &= \tfrac{1}{2}\,\alpha q_{\max}^{\alpha-2}\left(\|Q_\varepsilon\|_2^2 - \|P_\varepsilon\|_2^2\right) \\
&= \tfrac{1}{2}\,\alpha\left(\tfrac{1}{2}+\beta\varepsilon\right)^{\alpha-2}\left[\left(\tfrac{1}{2}+\beta\varepsilon\right)^2 + \left(\tfrac{1}{2}-\beta\varepsilon\right)^2 - \left(\tfrac{1}{2}+\varepsilon\right)^2 - \left(\tfrac{1}{2}-\varepsilon\right)^2\right] \\
&= \alpha 2^{2-\alpha}(\beta^2-1)(1+2\beta\varepsilon)^{\alpha-2}. \qquad \text{(A236)}
\end{aligned}$$

- For $\alpha \in (2, \infty)$,

$$\begin{aligned}
L(\alpha, P_\varepsilon, Q_\varepsilon) &= \tfrac{1}{2}\,\alpha q_{\min}^{\alpha-2}\left(\|Q_\varepsilon\|_2^2 - \|P_\varepsilon\|_2^2\right) \\
&= \alpha 2^{2-\alpha}(\beta^2-1)(1-2\beta\varepsilon)^{\alpha-2}. \qquad \text{(A237)}
\end{aligned}$$

- Similarly, for $\alpha \in (0, 2]$,

$$\begin{aligned}
U(\alpha, P_\varepsilon, Q_\varepsilon) &= \tfrac{1}{2}\,\alpha q_{\min}^{\alpha-2}\left(\|Q_\varepsilon\|_2^2 - \|P_\varepsilon\|_2^2\right) \\
&= \alpha 2^{2-\alpha}(\beta^2-1)(1-2\beta\varepsilon)^{\alpha-2}, \qquad \text{(A238)}
\end{aligned}$$

and, for $\alpha \in (2, \infty)$,

$$\begin{aligned}
U(\alpha, P_\varepsilon, Q_\varepsilon) &= \tfrac{1}{2}\,\alpha q_{\max}^{\alpha-2}\left(\|Q_\varepsilon\|_2^2 - \|P_\varepsilon\|_2^2\right) \\
&= \alpha 2^{2-\alpha}(\beta^2-1)(1+2\beta\varepsilon)^{\alpha-2}. \qquad \text{(A239)}
\end{aligned}$$

The combination of (A235)–(A237) yields (A233); similarly, the combination of (A235), (A238) and (A239) yields (A234).

Appendix H. Proof of Theorem 9 and Corollary 1

Appendix H.1. Proof of Theorem 9

The proof of the convexity property of $\Delta(\cdot, \rho)$ in (149), with $\rho > 1$, over the real line $\mathbb{R}$ relies on ([69], Theorem 2.1) which states that if W is a non-negative random variable, then

$$\lambda_\alpha := \begin{cases} \dfrac{\left(\mathbb{E}[W^\alpha] - \mathbb{E}^\alpha[W]\right)\,\log \mathrm{e}}{\alpha(\alpha-1)}, & \alpha \neq 0, 1 \\ \log\left(\mathbb{E}[W]\right) - \mathbb{E}[\log W], & \alpha = 0 \\ \mathbb{E}[W \log W] - \mathbb{E}[W]\,\log\left(\mathbb{E}[W]\right), & \alpha = 1 \end{cases} \qquad \text{(A240)}$$

is log-convex in $\alpha \in \mathbb{R}$. This property has been used to derive f-divergence inequalities (see, e.g., ([62], Theorem 20), [65,69]).

Let $Q \ll P$, and let $W := \frac{\mathrm{d}Q}{\mathrm{d}P}$ be the Radon-Nikodym derivative (W is a non-negative random variable). Let the expectations in the right side of (A240) be taken with respect to P. In view of the above statement from ([69], Theorem 2.1), this gives the log-convexity of $D_{\mathrm{A}}^{(\alpha)}(Q\|P)$ in $\alpha \in \mathbb{R}$. Since log-convexity yields convexity, it follows that $D_{\mathrm{A}}^{(\alpha)}(Q\|P)$ is convex in α over the real line. Let $P := U_n$, and let $Q \in \mathcal{P}_n(\rho)$; since $Q \ll P$, it follows that $D_{\mathrm{A}}^{(\alpha)}(Q\|U_n)$ is convex in $\alpha \in \mathbb{R}$. The pointwise maximum of a set of convex functions is a convex function, which implies that $\max\limits_{Q \in \mathcal{P}_n(\rho)} D_{\mathrm{A}}^{(\alpha)}(Q\|U_n)$ is convex in $\alpha \in \mathbb{R}$ for every integer $n \geq 2$. Since the pointwise limit of a convergent sequence of convex functions is convex, it follows that $\lim\limits_{n\to\infty} \max\limits_{Q \in \mathcal{P}_n(\rho)} D_{\mathrm{A}}^{(\alpha)}(Q\|U_n)$ is convex in α. This, by definition, is equal to $\Delta(\alpha, \rho)$ (see (146)), which proves the convexity of this function in $\alpha \in \mathbb{R}$. From (149), for all $\rho > 1$,

$$\begin{aligned}
\Delta(1+\alpha, \rho) &= \frac{1}{(\alpha+1)\alpha} \left[\frac{(-\alpha)^{\alpha} \left(\rho^{1+\alpha} - 1\right)^{1+\alpha} \left(\rho - \rho^{1+\alpha}\right)^{-\alpha}}{(\rho-1)(1+\alpha)^{1+\alpha}} - 1 \right] \\
&= \frac{1}{(-\alpha)(-\alpha-1)} \left[\frac{(1+\alpha)^{-\alpha-1} \left(\rho^{\alpha}(\rho - \rho^{-\alpha})\right)^{1+\alpha} \left(\rho^{1+\alpha}(\rho^{-\alpha} - 1)\right)^{-\alpha}}{(\rho-1)(-\alpha)^{-\alpha}} - 1 \right] \\
&= \frac{1}{(-\alpha)(-\alpha-1)} \left[\frac{(1+\alpha)^{-\alpha-1} \left(\rho - \rho^{-\alpha}\right)^{1+\alpha} \left(\rho^{-\alpha} - 1\right)^{-\alpha}}{(\rho-1)(-\alpha)^{-\alpha}} - 1 \right] \\
&= \Delta(-\alpha, \rho),
\end{aligned} \tag{A241}$$

which proves the symmetry property of $\Delta(\alpha, \rho)$ around $\alpha = \frac{1}{2}$ for all $\rho > 1$. The convexity in α over the real line, and the symmetry around $\alpha = \frac{1}{2}$ implies that $\Delta(\alpha, \rho)$ gets its global minimum at $\alpha = \frac{1}{2}$, which is equal to $\frac{4(\sqrt[4]{\rho}-1)^2}{\sqrt{\rho}+1}$ for all $\rho > 1$.

Inequalities (162) and (163) follow from ([8], Proposition 2.7); this proposition implies that, for every integer $n \geq 2$ and for all probability mass functions Q defined on $\mathcal{A}_n := \{1, \ldots, n\}$,

$$\alpha \, D_{\mathrm{A}}^{(\alpha)}(Q\|U_n) \leq \beta \, D_{\mathrm{A}}^{(\beta)}(Q\|U_n), \qquad 0 < \alpha \leq \beta < \infty, \tag{A242}$$

$$(1-\beta) \, D_{\mathrm{A}}^{(1-\beta)}(Q\|U_n) \leq (1-\alpha) \, D_{\mathrm{A}}^{(1-\alpha)}(Q\|U_n), \quad -\infty < \alpha \leq \beta < 1. \tag{A243}$$

Inequalities (162) and (163) follow, respectively, by maximizing both sides of (A242) or (A243) over $Q \in \mathcal{P}_n(\rho)$, and letting n tend to infinity.

For every $\alpha \in \mathbb{R}$, the function $\Delta(\alpha, \rho)$ is monotonically increasing in $\rho \in (1, \infty)$ since (by definition) the set of probability mass functions $\{\mathcal{P}_n(\rho)\}_{\rho \geq 1}$ is monotonically increasing (i.e., $\mathcal{P}_n(\rho_1) \subseteq \mathcal{P}_n(\rho_2)$ if $1 \leq \rho_1 < \rho_2 < \infty$), and therefore the maximum of $D_{\mathrm{A}}^{(\alpha)}(Q\|U_n)$ over $Q \in \mathcal{P}_n(\rho)$ is a monotonically increasing function of $\rho \in [1, \infty)$; the limit of this maximum, as we let $n \to \infty$, is equal to $\Delta(\alpha, \rho)$ in (149) for all $\rho > 1$, which is therefore monotonically increasing in ρ over the interval $(1, \infty)$. The continuity of $\Delta(\alpha, \rho)$ in both α and ρ is due to its expression in (149) with its continuous extension at $\alpha = 0$ and $\alpha = 1$ in (150). Since $\mathcal{P}_n(1) = \{U_n\}$, it follows from the continuity of $\Delta(\alpha, \rho)$ that

$$\lim_{\rho \to 1^+} \Delta(\alpha, \rho) = D_{\mathrm{A}}^{(\alpha)}(U_n\|U_n) = 0.$$

Appendix H.2. Proof of Corollary 1

For all $\alpha \in \mathbb{R}$ and $\rho > 1$,

$$\begin{aligned}
&\lim_{n\to\infty} \max_{Q \in \mathcal{P}_n(\rho)} D_{\mathrm{A}}^{(\alpha)}(U_n \| Q) \\
&= \lim_{n\to\infty} \max_{Q \in \mathcal{P}_n(\rho)} D_{\mathrm{A}}^{(1-\alpha)}(Q \| U_n) && \text{(A244)} \\
&= \Delta(1-\alpha, \rho) && \text{(A245)} \\
&= \Delta(\alpha, \rho), && \text{(A246)}
\end{aligned}$$

where (A244) holds due to the symmetry property in ([8], p. 36), which states that

$$D_{\mathrm{A}}^{(\alpha)}(P \| Q) = D_{\mathrm{A}}^{(1-\alpha)}(Q \| P), \tag{A247}$$

for every $\alpha \in \mathbb{R}$ and probability mass functions P and Q; (A245) is due to (146); finally, (A246) holds due to the symmetry property of $\Delta(\cdot, \rho)$ around $\frac{1}{2}$ in Theorem 9 (a).

Appendix I. Proof of (171)

In view of (154) and (155), it follows that the condition in (170) is satisfied if and only if $\rho \le \rho^*$ where $\rho^* \in (1, \infty)$ is the solution of the equation

$$\frac{\rho^* \log \rho^*}{\rho^* - 1} - \log\left(\frac{\mathrm{e}\rho^* \log_{\mathrm{e}} \rho^*}{\rho^* - 1}\right) = d \log \mathrm{e}. \tag{A248}$$

with a fixed $d > 0$. The substitution

$$x := \frac{\rho^* \log_{\mathrm{e}} \rho^*}{\rho^* - 1} \tag{A249}$$

leads to the equation

$$x - \log_{\mathrm{e}} x = d + 1. \tag{A250}$$

Negation and exponentiation of both sides of (A250) gives

$$(-x)\mathrm{e}^{-x} = -\mathrm{e}^{-d-1}. \tag{A251}$$

Since $\rho^* > 1$ implies by (A249) that $x > 1$, the proper solution for x is given by

$$x = -W_{-1}\big(-\mathrm{e}^{-d-1}\big), \quad d > 0, \tag{A252}$$

where W_{-1} denotes the secondary real branch of the Lambert W function [37]; otherwise, the replacement of W_{-1} in the right side of (A252) with the principal real branch W_0 yields $x \in (0, 1)$.

We next proceed to solve ρ^* as a function of x. From (A249), letting $u := \frac{1}{\rho^*}$ gives the equation $u = \mathrm{e}^{(u-1)x}$, which is equivalent to

$$\begin{aligned}
(-ux)\mathrm{e}^{-ux} &= -x\mathrm{e}^{-x} && \text{(A253)} \\
&= -\mathrm{e}^{-d-1}, && \text{(A254)}
\end{aligned}$$

where (A254) follows from (A252) and by the definition of the Lambert W function (i.e., $t = W(u)$ if and only if $t\mathrm{e}^t = u$). The solutions of (A253) are given by

$$-ux = W_{-1}\big(-\mathrm{e}^{-d-1}\big), \tag{A255}$$

and

$$-ux = W_{0}\big(-\mathrm{e}^{-d-1}\big), \tag{A256}$$

which (from (A252)) correspond, respectively, to $u = 1$ and

$$u = \frac{W_0\big(-\mathrm{e}^{-d-1}\big)}{W_{-1}\big(-\mathrm{e}^{-d-1}\big)} \in (0,1). \tag{A257}$$

Since $\rho^* \in (1,\infty)$ is equal to $\frac{1}{u}$, the reciprocal of the right side of (A257) gives the proper solution for ρ^* (denoted by $\rho^{(1)}_{\max}(d)$ in (171)).

Appendix J. Proof of (176), (177) and (180)

We first derive the upper bound on $\Phi(\alpha,\rho)$ in (176) for $\alpha \geq \mathrm{e}^{-\frac{3}{2}}$ and $\rho \geq 1$. For every $Q \in \mathcal{P}_n(\rho)$, with an integer $n \geq 2$,

$$\begin{aligned} D_{f_\alpha}(Q\|U_n) \leq{}& \Big[\log(\alpha+1) + \tfrac{3}{2}\log\mathrm{e} - \frac{\log\mathrm{e}}{\alpha+1}\Big]\,\chi^2(Q\|U_n) \\ &+ \frac{\log\mathrm{e}}{3(\alpha+1)}\Big[\exp\big(2D_3(Q\|U_n)\big)-1\Big] \qquad \text{(A258)} \\ \leq{}& \Big[\log(\alpha+1) + \tfrac{3}{2}\log\mathrm{e} - \frac{\log\mathrm{e}}{\alpha+1}\Big]\,\frac{(\rho-1)^2}{4\rho} \\ &+ \frac{\log\mathrm{e}}{3(\alpha+1)}\Big[\exp\big(2D_3(Q\|U_n)\big)-1\Big] \qquad \text{(A259)} \end{aligned}$$

where (A258) follows from (65), and (A258) holds due to (159). By upper bounding the second term in the right side of (A259), for all $Q \in \mathcal{P}_n(\rho)$,

$$\begin{aligned} D_3(Q\|U_n) &= \tfrac{1}{2}\log\big(1+6D_{\mathrm{A}}^{(3)}(Q\|U_n)\big) && \text{(A260)} \\ &\leq \tfrac{1}{2}\log\big(1+6\Delta(3,\rho)\big) && \text{(A261)} \\ &= \tfrac{1}{2}\log\left(\frac{4(\rho^3-1)^3}{27(\rho-1)(\rho-\rho^3)^2}\right) && \text{(A262)} \\ &= \tfrac{1}{2}\log\left(\frac{4(\rho^2+\rho+1)^3}{27\rho^2(\rho+1)^2}\right) && \text{(A263)} \end{aligned}$$

where (A260) holds by setting $\alpha = 3$ in (156); (A261) follows from (135), (138) and (145); (A262) holds by setting $\alpha = 3$ in (149); finally, (A263) follows from the factorizations

$$(\rho^3-1)^3 = (\rho-1)^3(\rho^2+\rho+1)^3, \quad (\rho-1)(\rho-\rho^3)^2 = (\rho-1)^3\rho^2(\rho+1)^2.$$

Substituting the bound in the right side of (A263) into the second term of the bound on the right side of (A259) implies that, for all $Q \in \mathcal{P}_n(\rho)$,

$$\begin{aligned} D_{f_\alpha}(Q\|U_n) &\le \left[\log(\alpha+1)+\tfrac{3}{2}\log\mathrm{e}-\frac{\log\mathrm{e}}{\alpha+1}\right]\frac{(\rho-1)^2}{4\rho} \\ &\quad+\frac{\log\mathrm{e}}{3(\alpha+1)}\left[\frac{4(\rho^2+\rho+1)^3}{27\rho^2(\rho+1)^2}-1\right] \qquad \text{(A264)} \\ &=\left[\log(\alpha+1)+\tfrac{3}{2}\log\mathrm{e}-\frac{\log\mathrm{e}}{\alpha+1}\right]\frac{(\rho-1)^2}{4\rho} \\ &\quad+\frac{\log\mathrm{e}}{81(\alpha+1)}\left(\frac{(\rho-1)(2\rho+1)(\rho+2)}{\rho(\rho+1)}\right)^2, \qquad \text{(A265)} \end{aligned}$$

which therefore gives (176) by maximizing the left side of (A264) over $Q \in \mathcal{P}_n(\rho)$, and letting n tend to infinity (see (174)).

We next derive the upper bound in (177). The second derivative of the convex function $f_\alpha\colon (0,\infty) \to \mathbb{R}$ in (55) is upper bounded over the interval $\left[\frac{1}{\rho},\rho\right]$ by the positive constant $M = 2\log(\alpha+\rho)+3\log\mathrm{e}$. From (96), it follows that for all $Q \in \mathcal{P}_n(\rho)$ (with $\rho \ge 1$ and an integer $n \ge 2$) and $\alpha \ge \mathrm{e}^{-\frac{3}{2}}$,

$$D_{f_\alpha}(Q\|U_n) \le \left[\log(\alpha+\rho)+\tfrac{3}{2}\log\mathrm{e}\right]\frac{(\rho-1)^2}{4\rho}, \tag{A266}$$

which, from (174), yields (177).

We finally derive the upper bound in (180) by loosening the bound in (176). The upper bound in the right side of (176) can be rewritten as

$$\begin{aligned} \Phi(\alpha,\rho) &\le \left[\tfrac{1}{4}\,\log(\alpha+1)+\tfrac{3}{8}\,\log\mathrm{e}\right]\frac{(\rho-1)^2}{\rho} \\ &\quad+\frac{\log\mathrm{e}}{\alpha+1}\left[\frac{1}{81}\left(2+\frac{2}{\rho}+\frac{1}{1+\rho}\right)^2-\frac{1}{4\rho}\right](\rho-1)^2. \qquad \text{(A267)} \end{aligned}$$

For all $\rho \ge 1$,

$$\frac{1}{81}\left(2+\frac{2}{\rho}+\frac{1}{1+\rho}\right)^2-\frac{1}{4\rho} \le \frac{4}{81}, \tag{A268}$$

which can be verified by showing that the left side of (A268) is monotonically increasing in ρ over the interval $[1,\infty)$, and it tends to $\frac{4}{81}$ as we let $\rho \to \infty$. Furthermore, for all $\rho \ge 1$,

$$\frac{(\rho-1)^2}{\rho} \le \min\{\rho-1,(\rho-1)^2\}. \tag{A269}$$

In view of inequalities (A268) and (A269), one gets (180) from (A267) (where the latter is an equivalent form of (176)).

Appendix K. Proof of Theorem 10

We start by proving Item (a). In view of the variational representation of f-divergences (see ([70], Theorem 2.1), and ([71], Lemma 1)), if $f\colon (0,\infty) \to \mathbb{R}$ is convex with $f(1) = 0$, and P and Q are probability measures defined on a set $\mathcal{A}$, then

$$D_f(P\|Q) = \sup_{g\colon \mathcal{A}\to\mathbb{R}}\Big(\mathbb{E}\big[g(X)\big]-\mathbb{E}\big[\overline{f}\big(g(Y)\big)\big]\Big), \tag{A270}$$

where $X \sim P$ and $Y \sim Q$, and the supremum is taken over all measurable functions g under which the expectations are finite.

Let $P \in \mathcal{P}_n(\rho)$, with $\rho > 1$, and let $Q := U_n$; these probability mass functions are defined on the set $\mathcal{A}_n := \{1, \ldots, n\}$, and it follows that

$$u_f(n,\rho) \geq D_f(P \| U_n) \tag{A271}$$

$$\geq \mathbb{E}\big[g(X)\big] - \frac{1}{n} \sum_{i=1}^{n} \overline{f}\big(g(i)\big), \tag{A272}$$

where (A271) holds by the definition in (77); (A272) holds due to (A270) with $X \sim P$, and Y being an equiprobable random variable over $\mathcal{A}_n$. This gives (187).

We next prove Item (b). As above, let $f \colon (0, \infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$. Let $\beta^* \in \Gamma_n(\rho)$ be a maximizer of the right side of (82). Then,

$$u_f(n,\rho) = D_f(Q_{\beta^*} \| U_n) \tag{A273}$$

$$= \frac{1}{n} \sum_{i=1}^{n} f\big(nQ_{\beta^*}(i)\big). \tag{A274}$$

Let $\varepsilon > 0$ be selected arbitrarily. We have $\overline{(\overline{f})} \equiv f$ (i.e., repeating twice the convex conjugate operation (see (186)) on a convex function f, returns f itself). From the convexity of f, it therefore follows that, for all $t > 0$, there exists $x \in \mathbb{R}$ such that

$$f(t) \leq tx - \overline{f}(x) + \varepsilon. \tag{A275}$$

Let

$$t_i := nQ_{\beta^*}(i), \quad \forall\, i \in \mathcal{A}_n, \tag{A276}$$

let $x := x_i(\varepsilon) \in \mathbb{R}$ be selected to satisfy (A275) with $t := t_i$, and let the function $g_\varepsilon \colon \mathcal{A}_n \to \mathbb{R}$ be defined as

$$g_\varepsilon(i) = x_i(\varepsilon), \quad \forall\, i \in \mathcal{A}_n. \tag{A277}$$

Consequently, it follows from (A275)–(A277) that for all such i

$$f\big(nQ_{\beta^*}(i)\big) \leq nQ_{\beta^*}(i)\, g_\varepsilon(i) - \overline{f}\big(g_\varepsilon(i)\big) + \varepsilon. \tag{A278}$$

Let $P := Q_{\beta^*} \in \mathcal{P}_n(\rho)$ (see (80)), and $X \sim P$. Then,

$$u_f(n,\rho) = \frac{1}{n} \sum_{i=1}^{n} f\big(nQ_{\beta^*}(i)\big) \tag{A279}$$

$$\leq \sum_{i=1}^{n} Q_{\beta^*}(i)\, g_\varepsilon(i) - \frac{1}{n} \sum_{i=1}^{n} \overline{f}\big(g_\varepsilon(i)\big) + \varepsilon \tag{A280}$$

$$= \mathbb{E}\big[g_\varepsilon(X)\big] - \frac{1}{n} \sum_{i=1}^{n} \overline{f}\big(g_\varepsilon(i)\big) + \varepsilon \tag{A281}$$

where (A279) holds due to (A273) and (A274); (A280) follows from (A278); (A281) holds since by assumption $P_X = Q_{\beta^*}$. This gives (188).

Appendix L. Proof of Theorem 11

For $y \in \mathcal{Y}$, let the L-size list of the decoder be given by $\mathcal{L}(y) = \{x_1(y), \dots, x_L(y)\}$ with $L < M$. Then, the (average) list decoding error probability is given by

$$P_{\mathcal{L}} = \mathbb{E}\big[P_{\mathcal{L}}(Y)\big] \tag{A282}$$

where the conditional list decoding error probability, given that $Y = y \in \mathcal{Y}$, is equal to

$$P_{\mathcal{L}}(y) = 1 - \sum_{\ell=1}^{L} P_{X|Y}\big(x_\ell(y)\,|\,y\big). \tag{A283}$$

For every $y \in \mathcal{Y}$,

$$\begin{aligned} & D_f\big(P_{X|Y}(\cdot|y)\,\|\,U_M\big) \\ & \geq D_f\left(\left[\sum_{\ell=1}^{L} P_{X|Y}\big(x_\ell(y)\,|\,y\big),\, 1 - \sum_{\ell=1}^{L} P_{X|Y}\big(x_\ell(y)\,|\,y\big)\right] \,\Big\|\, \left[\frac{L}{M},\, 1 - \frac{L}{M}\right]\right) && \text{(A284)} \\ & = D_f\left(\big[1 - P_{\mathcal{L}}(y),\, P_{\mathcal{L}}(y)\big] \,\Big\|\, \left[\frac{L}{M},\, 1 - \frac{L}{M}\right]\right), && \text{(A285)} \end{aligned}$$

where (A284) holds by the data-processing inequality for f-divergences, and since for every $y \in \mathcal{Y}$

$$\sum_{\ell=1}^{L} U_M\big(x_\ell(y)\big) = \sum_{\ell=1}^{L} \frac{1}{M} = \frac{L}{M}; \tag{A286}$$

(A285) is due to (A283). Hence, it follows that

$$\begin{aligned} & \mathbb{E}\Big[D_f\big(P_{X|Y}(\cdot|Y)\,\|\,U_M\big)\Big] \\ & \geq \mathbb{E}\left[D_f\left(\big[1 - P_{\mathcal{L}}(Y),\, P_{\mathcal{L}}(Y)\big] \,\Big\|\, \left[\frac{L}{M},\, 1 - \frac{L}{M}\right]\right)\right] && \text{(A287)} \\ & = \frac{L}{M}\,\mathbb{E}\left[f\left(\frac{M(1 - P_{\mathcal{L}}(Y))}{L}\right)\right] + \left(1 - \frac{L}{M}\right)\mathbb{E}\left[f\left(\frac{M P_{\mathcal{L}}(Y)}{M - L}\right)\right] && \text{(A288)} \\ & \geq \frac{L}{M}\, f\left(\frac{M\,\mathbb{E}[1 - P_{\mathcal{L}}(Y)]}{L}\right) + \left(1 - \frac{L}{M}\right) f\left(\frac{M\,\mathbb{E}[P_{\mathcal{L}}(Y)]}{M - L}\right) && \text{(A289)} \\ & = \frac{L}{M}\, f\left(\frac{M(1 - P_{\mathcal{L}})}{L}\right) + \left(1 - \frac{L}{M}\right) f\left(\frac{M P_{\mathcal{L}}}{M - L}\right), && \text{(A290)} \end{aligned}$$

where (A287) holds by taking expectations in (A284) and (A285) with respect to Y; (A288) holds by the definition of f-divergence, and the linearity of expectation operator; (A289) follows from the convexity of f and Jensen's inequality; finally, (A290) holds by (A282).

Appendix M. Proof of Corollary 3

Let $\alpha \in (0,1) \cup (1,\infty)$, and let $y \in \mathcal{Y}$. The proof starts by applying Theorem 11 in the setting where $Y = y$ is deterministic, and the convex function $f\colon (0,\infty) \to \mathbb{R}$ is given by $f := u_\alpha$ in (139), i.e.,

$$f(t) = \frac{t^\alpha - \alpha(t-1) - 1}{\alpha(\alpha - 1)}, \quad t \geq 0. \tag{A291}$$

In this setting, (192) is specialized to

$$D_f\left(P_{X|Y}(\cdot|y)\,\|\,U_M\right) \geq \frac{L}{M}\, f\left(\frac{M\,(1-P_{\mathcal{L}}(y))}{L}\right) + \left(1-\frac{L}{M}\right) f\left(\frac{MP_{\mathcal{L}}(y)}{M-L}\right), \tag{A292}$$

where $P_{\mathcal{L}}(y)$ is the conditional list decoding error probability given that $Y = y$. Substituting (A291) into the right side of (A292) gives

$$\begin{aligned}
&\frac{L}{M}\, f\left(\frac{M\,(1-P_{\mathcal{L}}(y))}{L}\right) + \left(1-\frac{L}{M}\right) f\left(\frac{MP_{\mathcal{L}}(y)}{M-L}\right) \\
&= \frac{1}{\alpha(\alpha-1)}\left[P_{\mathcal{L}}^{\alpha}(y)\left(1-\frac{L}{M}\right)^{1-\alpha} + (1-P_{\mathcal{L}}(y))^{\alpha}\left(\frac{L}{M}\right)^{1-\alpha} - 1\right] &\text{(A293)}\\
&= \frac{1}{\alpha(\alpha-1)}\left[\exp\left((\alpha-1)\, d_\alpha\left(P_{\mathcal{L}}(y)\,\|\,1-\frac{L}{M}\right)\right) - 1\right], &\text{(A294)}
\end{aligned}$$

where (A294) follows from (203). Substituting (A291) into the left side of (A292) gives

$$\begin{aligned}
&D_f\left(P_{X|Y}(\cdot|y)\,\|\,U_M\right) \\
&= \frac{1}{M\alpha(\alpha-1)} \sum_{x\in\mathcal{X}} \left[\left(MP_{X|Y}(x|y)\right)^{\alpha} - \alpha\left(MP_{X|Y}(x|y)-1\right) - 1\right] &\text{(A295)}\\
&= \frac{1}{M\alpha(\alpha-1)}\left[M^{\alpha}\sum_{x\in\mathcal{X}} P_{X|Y}^{\alpha}(x|y) - \alpha \underbrace{\sum_{x\in\mathcal{X}}\left(MP_{X|Y}(x|y)-1\right)}_{=0\ \ (|\mathcal{X}|=M)} - M\right] &\text{(A296)}\\
&= \frac{1}{\alpha(\alpha-1)}\left[M^{\alpha-1}\sum_{x\in\mathcal{X}} P_{X|Y}^{\alpha}(x|y) - 1\right] &\text{(A297)}\\
&= \frac{1}{\alpha(\alpha-1)}\left[\exp\left((\alpha-1)\,\left[\log M - H_\alpha(X|Y=y)\right]\right) - 1\right]. &\text{(A298)}
\end{aligned}$$

Substituting (A294) and (A298) into the right and left sides of (A292), and rearranging terms while relying on the monotonicity property of an exponential function gives

$$H_\alpha(X|Y=y) \leq \log M - d_\alpha\left(P_{\mathcal{L}}(y)\,\|\,1-\frac{L}{M}\right). \tag{A299}$$

We next obtain an upper bound on the Arimoto-Rényi conditional entropy.

$$\begin{aligned}
&H_\alpha(X|Y) \\
&= \frac{\alpha}{1-\alpha}\,\log \int_{\mathcal{Y}} \mathrm{d}P_Y(y)\, \exp\left(\frac{1-\alpha}{\alpha}\, H_\alpha(X|Y=y)\right) &\text{(A300)}\\
&\leq \frac{\alpha}{1-\alpha}\,\log \int_{\mathcal{Y}} \mathrm{d}P_Y(y)\, \exp\left(\frac{1-\alpha}{\alpha}\left[\log M - d_\alpha\left(P_{\mathcal{L}}(y)\,\|\,1-\frac{L}{M}\right)\right]\right) &\text{(A301)}\\
&= \log M + \frac{\alpha}{1-\alpha}\,\log \int_{\mathcal{Y}} \mathrm{d}P_Y(y)\left[P_{\mathcal{L}}^{\alpha}(y)\left(1-\frac{L}{M}\right)^{1-\alpha} + (1-P_{\mathcal{L}}(y))^{\alpha}\left(\frac{L}{M}\right)^{1-\alpha}\right]^{\frac{1}{\alpha}} &\text{(A302)}
\end{aligned}$$

where (A300) holds due to (202); (A301) follows from (A299), and (A302) follows from (203). By ([42], Lemma 1), it follows that the integrand in the right side of (A302) is convex in $P_{\mathcal{L}}(y)$ if $\alpha > 1$;

furthermore, it is concave in $P_{\mathcal{L}}(y)$ if $\alpha \in (0,1)$. Invoking Jensen's inequality therefore yields (see (A282))

$$H_\alpha(X|Y) \leq \log M + \frac{\alpha}{1-\alpha} \log \left(\left[P_{\mathcal{L}}^{\alpha} \left(1 - \frac{L}{M}\right)^{1-\alpha} + (1-P_{\mathcal{L}})^{\alpha} \left(\frac{L}{M}\right)^{1-\alpha} \right]^{\frac{1}{\alpha}} \right) \tag{A303}$$

$$= \log M - \frac{1}{\alpha-1} \log \left(P_{\mathcal{L}}^{\alpha} \left(1 - \frac{L}{M}\right)^{1-\alpha} + (1-P_{\mathcal{L}})^{\alpha} \left(\frac{L}{M}\right)^{1-\alpha} \right) \tag{A304}$$

$$= \log M - d_\alpha \left(P_{\mathcal{L}} \,\|\, 1 - \frac{L}{M} \right), \tag{A305}$$

where (A303) follows from Jensen's inequality, and (A305) follows from (203). This proves (205) and (206) for all $\alpha \in (0,1) \cup (1,\infty)$. The necessary and sufficient condition for (205) to hold with equality, as given in (207), follows from the proof of (A292) (see (A284)–(A286)), and from the use of Jensen's inequality in (A303).

Appendix N. Proof of Theorem 12

The proof of Theorem 12 relies on Theorem 1, and the proof of Theorem 11.

Let $\mathcal{Z} = \{0,1\}$ and, without any loss of generality, let $\mathcal{X} = \{1,\ldots,M\}$. For every $y \in \mathcal{Y}$, define a deterministic transformation from $\mathcal{X}$ to $\mathcal{Z}$ such that every $x \in \mathcal{L}(y)$ is mapped to $z = 0$, and every $x \notin \mathcal{L}(y)$ is mapped to $z = 1$. This corresponds to a conditional probability mass function, for every $y \in \mathcal{Y}$, where $W^{(y)}_{Z|X}(z|x) = 1$ if $x \in \mathcal{L}(y)$ and $z = 0$, or if $x \notin \mathcal{L}(y)$ and $z = 1$; otherwise, $W^{(y)}_{Z|X}(z|x) = 0$. Let $\mathcal{L}(y) := \{x_1(y), \ldots, x_L(y)\}$ with $L < M$. Then, for every $y \in \mathcal{Y}$, a conditional probability mass function $P_{X|Y}(\cdot|y)$ implies that

$$P^{(y)}_Z(z) := \sum_{x \in \mathcal{X}} P_{X|Y}(x|y)\, W^{(y)}_{Z|X}(z|x), \quad \forall\, z \in \{0,1\}, \tag{A306}$$

satisfies (see (A283))

$$P^{(y)}_Z(0) = \sum_{\ell=1}^{L} P_{X|Y}(x_\ell(y)|y) = 1 - P_{\mathcal{L}}(y), \tag{A307}$$

$$P^{(y)}_Z(1) = P_{\mathcal{L}}(y). \tag{A308}$$

Under the deterministic transformation $W^{(y)}_{Z|X}$ as above, the equiprobable distribution $Q^{(y)}_X = U_M$ (independently of $y \in \mathcal{Y}$) is mapped to a Bernoulli distribution over the two-elements set $\mathcal{Z}$ where

$$Q^{(y)}_Z = \left[\frac{L}{M}, 1 - \frac{L}{M} \right], \quad \forall\, y \in \mathcal{Y}. \tag{A309}$$

Given $Y = y \in \mathcal{Y}$, applying Theorem 1 with the transformation $W^{(y)}_{Z|X}$ as above gives that

$$\begin{aligned} & D_f\big(P_{X|Y}(\cdot|y) \,\|\, U_M\big) \\ & \geq D_f\big(P^{(y)}_Z \| Q^{(y)}_Z\big) + c_f\big(\xi_1(y), \xi_2(y)\big) \left[\chi^2\big(P_{X|Y}(\cdot|y) \,\|\, U_M\big) - \chi^2\big(P^{(y)}_Z \| Q^{(y)}_Z\big) \right] \end{aligned} \tag{A310}$$

where, from (18) and (19),

$$\xi_1(y) = \min_{x\in\mathcal{X}} \frac{P_{X|Y}(x|y)}{U_M(x)} = M \min_{x\in\mathcal{X}} P_{X|Y}(x|y), \tag{A311}$$

$$\xi_2(y) = \max_{x\in\mathcal{X}} \frac{P_{X|Y}(x|y)}{U_M(x)} = M \max_{x\in\mathcal{X}} P_{X|Y}(x|y). \tag{A312}$$

Since, from (212), (213), (A311) and (A312),

$$\inf_{y\in\mathcal{Y}} \xi_1(y) = M \inf_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_{X|Y}(x|y) = \xi_1^*, \tag{A313}$$

$$\sup_{y\in\mathcal{Y}} \xi_2(y) = M \sup_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_{X|Y}(x|y) = \xi_2^*, \tag{A314}$$

it follows from the definition of $c_f(\cdot,\cdot)$ in (26) that for every $y\in\mathcal{Y}$

$$\begin{aligned} c_f\big(\xi_1(y),\xi_2(y)\big) &\geq c_f(\xi_1^*,\xi_2^*) && \text{(A315)}\\ &= \tfrac{1}{2} \inf_{t\in\mathcal{I}(\xi_1^*,\xi_2^*)} f''(t) && \text{(A316)}\\ &\geq \tfrac{1}{2}\, m_f && \text{(A317)} \end{aligned}$$

where the last inequality holds by the assumption in (211). Combining (A310) and (A315)–(A317) yields

$$D_f(P_{X|Y}(\cdot|y)\,\|\,U_M) \geq D_f(P_Z^{(y)}\|Q_Z^{(y)}) + \tfrac{1}{2}m_f\Big[\chi^2(P_{X|Y}(\cdot|y)\,\|\,U_M) - \chi^2(P_Z^{(y)}\|Q_Z^{(y)})\Big], \tag{A318}$$

for every $y\in\mathcal{Y}$. Hence,

$$\mathbb{E}[D_f(P_{X|Y}(\cdot|Y)\,\|\,U_M)] \geq \mathbb{E}[D_f(P_Z^{(Y)}\|Q_Z^{(Y)})] + \tfrac{1}{2}m_f\,\mathbb{E}\Big[\chi^2(P_{X|Y}(\cdot|Y)\,\|\,U_M) - \chi^2(P_Z^{(Y)}\|Q_Z^{(Y)})\Big] \tag{A319}$$

where (A319) holds by taking expectations with respect to Y on both sides of (A318).

Referring to the first term in the right side of (A319) gives

$$\begin{aligned} \mathbb{E}[D_f(P_Z^{(Y)}\|Q_Z^{(Y)})] &= \mathbb{E}\left[D_f\left([1-P_{\mathcal{L}}(Y),\, P_{\mathcal{L}}(Y)]\,\|\,\left[\frac{L}{M},\, 1-\frac{L}{M}\right]\right)\right] && \text{(A320)}\\ &\geq \frac{L}{M}\, f\!\left(\frac{M(1-P_{\mathcal{L}})}{L}\right) + \left(1-\frac{L}{M}\right) f\!\left(\frac{MP_{\mathcal{L}}}{M-L}\right), && \text{(A321)} \end{aligned}$$

where (A320) follows from (A307)–(A309), and (A321) holds due to (A288)–(A290).

Referring to the second term in the right side of (A319) gives

$$\begin{aligned} &\mathbb{E}\Big[\chi^2(P_{X|Y}(\cdot|Y)\,\|\,U_M) - \chi^2(P_Z^{(Y)}\|Q_Z^{(Y)})\Big]\\ &= \mathbb{E}\left[\chi^2(P_{X|Y}(\cdot|Y)\,\|\,U_M) - \chi^2\left([1-P_{\mathcal{L}}(Y),\, P_{\mathcal{L}}(Y)]\,\|\,\left[\frac{L}{M},\, 1-\frac{L}{M}\right]\right)\right] && \text{(A322)}\\ &= \mathbb{E}\left[M\sum_{x\in\mathcal{X}} P^2_{X|Y}(x|Y) - \frac{M\big(1-P_{\mathcal{L}}(Y)\big)^2}{L} - \frac{MP^2_{\mathcal{L}}(Y)}{M-L}\right] && \text{(A323)}\\ &= M\,\mathbb{E}\left[\sum_{x\in\mathcal{X}} P^2_{X|Y}(x|Y)\right] - \frac{M}{L} + \frac{2M}{L}\cdot\mathbb{E}[P_{\mathcal{L}}(Y)] - \left(\frac{M}{L}+\frac{M}{M-L}\right)\mathbb{E}[P^2_{\mathcal{L}}(Y)] && \text{(A324)}\\ &= M\,\mathbb{E}\left[\sum_{x\in\mathcal{X}} P^2_{X|Y}(x|Y)\right] - \frac{M(1-2P_{\mathcal{L}})}{L} - \frac{M^2\,\mathbb{E}[P^2_{\mathcal{L}}(Y)]}{L(M-L)}, && \text{(A325)} \end{aligned}$$

where (A322) follows from (A306)–(A309); (A323) follows from (A16)–(A18); (A325) is due to (A282). Furthermore, we get (since $P_{\mathcal{L}}(Y) \in [0,1]$)

$$\mathbb{E}\big[P^2_{\mathcal{L}}(Y)\big] \leq \mathbb{E}\big[P_{\mathcal{L}}(Y)\big] = P_{\mathcal{L}}, \tag{A326}$$

$$\mathbb{E}\big[P^2_{\mathcal{L}}(Y)\big] \geq \mathbb{E}^2\big[P_{\mathcal{L}}(Y)\big] = P^2_{\mathcal{L}}, \tag{A327}$$

and

$$\begin{aligned}\mathbb{E}\left[\sum_{x\in\mathcal{X}} P^2_{X|Y}(x|Y)\right] &= \int_{\mathcal{Y}} \mathrm{d}P_Y(y) \sum_{x\in\mathcal{X}} P^2_{X|Y}(x|y) &\text{(A328)}\\ &= \int_{\mathcal{X}\times\mathcal{Y}} \mathrm{d}P_{XY}(x,y)\, P(x|y) &\text{(A329)}\\ &= \mathbb{E}\big[P_{X|Y}(X|Y)\big]. &\text{(A330)}\end{aligned}$$

Combining (A322)–(A330) gives

$$\begin{aligned}&M\left(\mathbb{E}\big[P_{X|Y}(X|Y)\big] - \frac{1-P_{\mathcal{L}}}{L} - \frac{P_{\mathcal{L}}}{M-L}\right)^{+}\\ &\leq \mathbb{E}\left[\chi^2(P_{X|Y}(\cdot|Y)\,\|\,U_M) - \chi^2(P_Z^{(Y)}\|Q_Z^{(Y)})\right] &\text{(A331)}\\ &\leq M\left(\mathbb{E}\big[P_{X|Y}(X|Y)\big] - \frac{(1-P_{\mathcal{L}})^2}{L} - \frac{P^2_{\mathcal{L}}}{M-L}\right), &\text{(A332)}\end{aligned}$$

which provides tight upper and lower bounds on $\mathbb{E}\left[\chi^2(P_{X|Y}(\cdot|Y)\,\|\,U_M) - \chi^2(P_Z^{(Y)}\|Q_Z^{(Y)})\right]$ if $P_{\mathcal{L}}$ is small. Note that the lower bound on the left side of (A331) is non-negative since, by the data-processing inequality for the χ^2 divergence, the right side of (A331) should be non-negative (see (A306)–(A309)). Finally, combining (A319)–(A332) yields (214), which proves Item (a).

For proving Item (b), the upper bound on the left side of (A326) is tightened. If the list decoder selects the L most probable elements from $\mathcal{X}$ given the value of $Y \in \mathcal{Y}$, then $P_{\mathcal{L}}(y) \leq 1 - \frac{L}{M}$ for every $y \in \mathcal{Y}$. Hence, the bound in (A326) is replaced by the tighter bound

$$\mathbb{E}\big[P^2_{\mathcal{L}}(Y)\big] \leq \left(1 - \frac{L}{M}\right) P_{\mathcal{L}}. \tag{A333}$$

Combining (A322)–(A325), (A328)–(A330) and (A333) gives the following improved lower bound in the left side of (A331):

$$M\left(\mathbb{E}\big[P_{X|Y}(X|Y)\big] - \frac{1-P_{\mathcal{L}}}{L}\right)^{+} \leq \mathbb{E}\left[\chi^2(P_{X|Y}(\cdot|Y)\,\|\,U_M) - \chi^2(P_Z^{(Y)}\|Q_Z^{(Y)})\right]. \tag{A334}$$

It is next shown that the operation $(\cdot)^+$ in the left side of (A334) is redundant. From (A282) and (A283),

$$\begin{aligned}P_{\mathcal{L}} &= 1 - \sum_{\ell=1}^{L} \mathbb{E}\big[P_{X|Y}(x_\ell(Y)\,|\,Y)\big] &\text{(A335)}\\ &= 1 - \sum_{\ell=1}^{L} \int_{\mathcal{Y}} \mathrm{d}P_Y(y)\, P_{X|Y}(x_\ell(y)\,|\,y) &\text{(A336)}\\ &= 1 - \int_{\mathcal{Y}} \mathrm{d}P_Y(y) \sum_{\ell=1}^{L} P_{X|Y}(x_\ell(y)\,|\,y), &\text{(A337)}\end{aligned}$$

which then implies that

$$P_{\mathcal{L}} \geq 1 - L \int_{\mathcal{Y}} \mathrm{d}P_Y(y) \sum_{\ell=1}^{L} P^2_{X|Y}(x_\ell(y)\,|\,y) \tag{A338}$$

$$\geq 1 - L \int_{\mathcal{Y}} \mathrm{d}P_Y(y) \sum_{x \in \mathcal{X}} P^2_{X|Y}(x|y) \tag{A339}$$

$$\geq 1 - L \int_{\mathcal{X} \times \mathcal{Y}} \mathrm{d}P_{XY}(x,y)\; P_{X|Y}(x|y) \tag{A340}$$

$$= 1 - L\, \mathbb{E}\big[P_{X|Y}(X|Y)\big], \tag{A341}$$

where (A338) is due the Cauchy-Schwarz inequality applied to the right side of (A337), and (A339) holds since $\mathcal{L}(y) \subseteq \mathcal{X}$ for all $y \in \mathcal{Y}$. From (A335)–(A341), $\mathbb{E}\big[P_{X|Y}(X|Y)\big] \geq \frac{1-P_{\mathcal{L}}}{L}$, which implies that the operation $(\cdot)^+$ in the left side of (A334) is indeed redundant. Similarly to the proof of (214) (see (A319)–(A321)), (A334) yields (215) while ignoring the operation $(\cdot)^+$ in the left side of (A334).

Appendix O. Proof of Theorem 13

For every $y \in \mathcal{Y}$, let the M elements of $\mathcal{X}$ be sorted in decreasing order according to the conditional probabilities $P_{X|Y}(\cdot|y)$. Let $x_\ell(y)$ be the ℓ-th most probable element in $\mathcal{X}$ given $Y = y$, i.e.,

$$P_{X|Y}(x_1(y)\,|y) \geq P_{X|Y}(x_2(y)\,|y) \geq \ldots \geq P_{X|Y}(x_M(y)\,|y). \tag{A342}$$

The conditional list decoding error probability, given $Y = y$, satisfies

$$P_{\mathcal{L}}(y) \geq 1 - \sum_{\ell=1}^{|\mathcal{L}(y)|} P_{X|Y}(x_\ell(y)\,|y) \tag{A343}$$

$$:= P_{\mathcal{L}}^{(\mathrm{opt})}(y), \tag{A344}$$

and the (average) list decoding error probability satisfies $P_{\mathcal{L}} \geq P_{\mathcal{L}}^{(\mathrm{opt})}$. Let U_M denote the equiprobable distribution on $\mathcal{X}$, and let $g_\gamma \colon [0,\infty) \to \mathbb{R}$ be given by $g_\gamma(t) := (t-\gamma)^+$ with $\gamma \geq 1$, where $u^+ := \max\{u, 0\}$ for $u \in \mathbb{R}$. The function $g_\gamma(\cdot)$ is convex, and $g_\gamma(1) = 0$ for $\gamma \geq 1$; the f-divergence $D_{g_\gamma}(\cdot\|\cdot)$ is named as the E_γ divergence (see, e.g., [54]), i.e.,

$$E_\gamma(P\|Q) := D_{g_\gamma}(P\|Q), \quad \forall\, \gamma \geq 1, \tag{A345}$$

for all probability measures P and Q. For every $y \in \mathcal{Y}$,

$$E_\gamma\big(P_{X|Y}(\cdot|y)\,\|\,U_M\big) \geq E_\gamma\left([1 - P_{\mathcal{L}}^{(\mathrm{opt})}(y),\, P_{\mathcal{L}}^{(\mathrm{opt})}(y)] \,\|\, \left[\frac{|\mathcal{L}(y)|}{M},\, 1 - \frac{|\mathcal{L}(y)|}{M}\right]\right) \tag{A346}$$

$$= \frac{|\mathcal{L}(y)|}{M} \cdot g_\gamma\left(\frac{M\big(1 - P_{\mathcal{L}}^{(\mathrm{opt})}(y)\big)}{|\mathcal{L}(y)|}\right) + \left(1 - \frac{|\mathcal{L}(y)|}{M}\right) g_\gamma\left(\frac{M\, P_{\mathcal{L}}^{(\mathrm{opt})}(y)}{M - |\mathcal{L}(y)|}\right), \tag{A347}$$

where (A346) holds due to the data-processing inequality for f-divergences, and because of (A344); (A347) holds due to (A345). Furthermore, in view of (A342) and (A344), it follows that $\frac{M\, P_{\mathcal{L}}^{(\mathrm{opt})}(y)}{M - |\mathcal{L}(y)|} \leq 1$ for all $y \in \mathcal{Y}$; by the definition of g_γ, it follows that

$$g_\gamma\left(\frac{M\, P_{\mathcal{L}}^{(\mathrm{opt})}(y)}{M - |\mathcal{L}(y)|}\right) = 0, \quad \forall\, \gamma \geq 1. \tag{A348}$$

Substituting (A348) into the right side of (A347) gives that, for all $y \in \mathcal{Y}$,

$$E_\gamma\big(P_{X|Y}(\cdot|y) \,\|\, U_M\big) \geq \frac{|\mathcal{L}(y)|}{M} \cdot g_\gamma\left(\frac{M\big(1-P_{\mathcal{L}}^{(\mathrm{opt})}(y)\big)}{|\mathcal{L}(y)|}\right) \tag{A349}$$

$$= \left(1 - P_{\mathcal{L}}^{(\mathrm{opt})}(y) - \frac{\gamma\,|\mathcal{L}(y)|}{M}\right)^+. \tag{A350}$$

Taking expectations with respect to Y in (A349) and (A350), and applying Jensen's inequality to the convex function $f(u) := (u)^+$, for $u \in \mathbb{R}$, gives

$$\mathbb{E}\Big[E_\gamma\big(P_{X|Y}(\cdot|Y) \,\|\, U_M\big)\Big] \geq \mathbb{E}\left[\left(1 - P_{\mathcal{L}}^{(\mathrm{opt})}(Y) - \frac{\gamma\,|\mathcal{L}(Y)|}{M}\right)^+\right] \tag{A351}$$

$$\geq \left(1 - \mathbb{E}\big[P_{\mathcal{L}}^{(\mathrm{opt})}(Y)\big] - \frac{\gamma\,\mathbb{E}\big[|\mathcal{L}(Y)|\big]}{M}\right)^+ \tag{A352}$$

$$= \left(1 - P_{\mathcal{L}}^{(\mathrm{opt})} - \frac{\gamma\,\mathbb{E}\big[|\mathcal{L}(Y)|\big]}{M}\right)^+ \tag{A353}$$

$$\geq 1 - P_{\mathcal{L}}^{(\mathrm{opt})} - \frac{\gamma\,\mathbb{E}\big[|\mathcal{L}(Y)|\big]}{M}. \tag{A354}$$

On the other hand, the left side of (A351) is equal to

$$\mathbb{E}\Big[E_\gamma\big(P_{X|Y}(\cdot|Y) \,\|\, U_M\big)\Big]$$

$$= \mathbb{E}\left[\frac{1}{M}\sum_{x\in\mathcal{X}}\big(MP_{X|Y}(x|Y) - \gamma\big)^+\right] \tag{A355}$$

$$= \mathbb{E}\left[\sum_{x\in\mathcal{X}}\left(P_{X|Y}(x|Y) - \frac{\gamma}{M}\right)^+\right] \tag{A356}$$

$$= \tfrac{1}{2}\,\mathbb{E}\left[\sum_{x\in\mathcal{X}}\left\{\left|P_{X|Y}(x|Y) - \frac{\gamma}{M}\right| + P_{X|Y}(x|Y) - \frac{\gamma}{M}\right\}\right] \tag{A357}$$

$$= \tfrac{1}{2}\,\mathbb{E}\left[\sum_{x\in\mathcal{X}}\left|P_{X|Y}(x|Y) - \frac{\gamma}{M}\right|\right] + \tfrac{1}{2}(1-\gamma), \tag{A358}$$

where (A355) is due to (A345), and since $U_M(x) = \frac{1}{M}$ for all $x \in \mathcal{X}$; (A356) and (A357) hold, respectively, by the simple identities $(cu)^+ = c\,u^+$, and $u^+ = \frac{1}{2}(|u| + u)$ for $c \geq 0$ and $u \in \mathbb{R}$; finally, (A358) holds since

$$\sum_{x\in\mathcal{X}}\left(P_{X|Y}(x|y) - \frac{\gamma}{M}\right) = -\gamma + \sum_{x\in\mathcal{X}} P_{X|Y}(x|y) = 1 - \gamma,$$

for all $y \in \mathcal{Y}$. Substituting (A355)–(A358) and rearranging terms gives that

$$P_{\mathcal{L}} \geq P_{\mathcal{L}}^{(\mathrm{opt})} \geq \frac{1+\gamma}{2} - \frac{\gamma\,\mathbb{E}\big[|\mathcal{L}(Y)|\big]}{M} - \tfrac{1}{2}\,\mathbb{E}\left[\sum_{x\in\mathcal{X}}\left|P_{X|Y}(x|Y) - \frac{\gamma}{M}\right|\right], \tag{A359}$$

which is the lower bound on the list decoding error probability in (222).

We next proceed to prove the sufficient conditions for equality in (222). First, if for all $y \in \mathcal{Y}$, the list decoder selects the $|\mathcal{L}(y)|$ most probable elements in $\mathcal{X}$ given that $Y = y$, then equality holds in (A359). In this case, for all $y \in \mathcal{Y}$, $\mathcal{L}(y) := \{x_1(y), \ldots, x_{|\mathcal{L}(y)|}\}$ where $x_\ell(y)$ denotes the ℓ-th most probable element in $\mathcal{X}$, given $Y = y$, with ties in probabilities which are resolved arbitrarily (see (A342)). Let $\gamma \geq 1$. If, for every $y \in \mathcal{Y}$, $P_{X|Y}\big(x_\ell(y)\,|y\big)$ is fixed for all $\ell \in \{1, \dots, |\mathcal{L}(y)|\}$ and $P_{X|Y}\big(x_\ell(y)\,|y\big)$ is fixed for all $\ell \in \{|\mathcal{L}(y)|+1, \dots, M\}$, then equality holds in (A346) (and therefore equalities also hold in (A349) and (A351)). For all $y \in \mathcal{Y}$, let the common values of the conditional probabilities $P_{X|Y}(\cdot|y)$ over each of these two sets, respectively, be equal to $\alpha(y)$ and $\beta(y)$. Then,

$$\alpha(y)\,|\mathcal{L}(y)| + \beta(y)\,\big(M - |\mathcal{L}(y)|\big) = \sum_{x \in \mathcal{X}} P_{X|Y}(x|y) = 1, \tag{A360}$$

which gives the condition in (223). Furthermore, if for all $y \in \mathcal{Y}$, $1 - P_{\mathcal{L}}^{(\mathrm{opt})}(y) - \frac{\gamma\,|\mathcal{L}(y)|}{M} \geq 0$, then the operation $(\cdot)^+$ in the right side of (A351) is redundant, which causes (A352) to hold with equality as an expectation of a linear function; furthermore, also (A354) holds with equality in this case (since an expectation of a non-negative and bounded function is non-negative and finite). By (223) and (A344), it follows that $P_{\mathcal{L}}^{(\mathrm{opt})}(y) = 1 - \alpha(y)\,|\mathcal{L}(y)|$ for all $y \in \mathcal{Y}$, and therefore the satisfiability of (224) implies that equalities hold in (A352) and (A354). Overall, under the above condition, it therefore follows that (222) holds with equality. To verify it explicitly, under conditions (223) and (224) which have been derived as above, the right side of (222) satisfies

$$\begin{aligned}
&\frac{1+\gamma}{2} - \frac{\gamma \mathbb{E}[|\mathcal{L}(Y)|]}{M} - \tfrac{1}{2}\,\mathbb{E}\left[\sum_{x \in \mathcal{X}} \left| P_{X|Y}(x|Y) - \frac{\gamma}{M} \right|\right] \\
&= \frac{1+\gamma}{2} - \frac{\gamma \mathbb{E}[|\mathcal{L}(Y)|]}{M} \\
&\quad - \tfrac{1}{2}\,\mathbb{E}\left[\left(\alpha(Y) - \frac{\gamma}{M}\right)|\mathcal{L}(Y)| + \left(\frac{\gamma}{M} - \frac{1 - \alpha(Y)\,|\mathcal{L}(Y)|}{M - |\mathcal{L}(Y)|}\right)\big(M - |\mathcal{L}(Y)|\big)\right] && \text{(A361)} \\
&= 1 - \mathbb{E}\big[\alpha(Y)\,|\mathcal{L}(Y)|\big] && \text{(A362)} \\
&= \mathbb{E}\left[1 - \sum_{\ell=1}^{|\mathcal{L}(Y)|} P_{X|Y}\big(x_\ell(Y)\,|Y\big)\right] && \text{(A363)} \\
&= P_{\mathcal{L}}, && \text{(A364)}
\end{aligned}$$

where (A361) holds since, under (224), it follows that $0 \leq \frac{1-\alpha(Y)\,|\mathcal{L}(Y)|}{M-|\mathcal{L}(Y)|} \leq \frac{1}{M} \leq \frac{\gamma}{M}$ for all $\gamma \geq 1$; (A362) holds by straightforward algebra, where γ is canceled out; (A363) holds by the condition in (223); finally, (A364) holds by (A282), (A283) and (A342). This indeed explicitly verifies that the conditions in Theorem 13 yield an equality in (222).

Appendix P. Proofs of Theorems Related to Tunstall Trees

Appendix P.1. Proof of Theorem 14

Theorem 14 (a) follows from (226) (see ([38], Corollary 1)).

By ([72], Lemma 6), the ratio of the maximal to minimal positive masses of P_ℓ is upper bounded by the reciprocal of the minimal probability mass of the source symbols. Theorem 14 (b) is therefore obtained from Theorem 7 (c). Theorem 14 (c) consequently holds due to Theorem 7 (d); the bound in the right side of (233), which holds for every number of leaves n in the Tunstall tree, is equal to the limit of the upper bound in the right side of (232) when we let $n \to \infty$.

Theorem 14 (d) relies on ([16], Theorem 11) and the definition in (231), providing an integral representation of an f-divergence in (234) under the conditions in Item (d).

Appendix P.2. Proof of Theorem 15

In view of ([33], Theorem 4), if the fixed length of the codewords of the Tunstall code is equal to m, then the compression rate R of the code satisfies

$$R \leq \frac{\lceil \log_{|\mathcal{X}|} n \rceil \, H(P)}{\log_{|\mathcal{X}|} n - \left[\frac{\rho \log \rho}{\rho - 1} - \log\left(\frac{e\rho \log_e \rho}{\rho - 1} \right) \right] \frac{1}{\log |\mathcal{X}|}}, \tag{A365}$$

where $H(P)$ denotes the Shannon entropy of the memoryless and stationary discrete source, $\rho := \frac{1}{p_{\min}}$, n is the number of leaves in Tunstall tree, and the logarithms with an unspecified base can be taken on an arbitrary base in the right side of (A365). By the setting in Theorem 15, the construction of the Tunstall tree satisfies $n \leq |\mathcal{X}|^m < n + (D-1)$. Hence, if $D = 2$, then $\log_{|\mathcal{X}|} n = m$; if $D > 2$, then $\lceil \log_{|\mathcal{X}|} n \rceil = m$ (since the length of the codewords is m), and $\log_{|\mathcal{X}|} n > m + \log_{|\mathcal{X}|} \left(1 - \frac{D-1}{|\mathcal{X}|^m} \right)$. Combining this with (A365) yields

$$R \leq \begin{cases} \dfrac{mH(P)}{m + \left\{ \log\left(1 - \frac{D-1}{|\mathcal{X}|^m} \right) - \left[\frac{\rho \log \rho}{\rho - 1} - \log\left(\frac{e\rho \log_e \rho}{\rho - 1} \right) \right] \right\} \frac{1}{\log |\mathcal{X}|}}, & \text{if } D > 2, \\[2ex] \dfrac{mH(P)}{m - \left[\frac{\rho \log \rho}{\rho - 1} - \log\left(\frac{e\rho \log_e \rho}{\rho - 1} \right) \right] \frac{1}{\log |\mathcal{X}|}}, & \text{if } D = 2. \end{cases} \tag{A366}$$

In order to assert that $R \leq (1+\varepsilon)\, H(P)$, it is requested that the right side of (A366) does not exceed $(1+\varepsilon)\, H(P)$. This gives

$$\frac{\rho \log \rho}{\rho - 1} - \log\left(\frac{e\rho \log_e \rho}{\rho - 1} \right) \leq d \log e, \tag{A367}$$

where d is given in (235). In view of the part in Section 3.3.2 with respect to the exemplification of Theorem 7 for the relative entropy, and the related analysis in Appendix I, the condition in (A367) is equivalent to $\rho \leq \rho_{\max}^{(1)}(d)$ where $\rho_{\max}^{(1)}(d)$ is defined in (171). Since $p_{\min} = \frac{1}{\rho}$, it leads to the sufficient condition in (236) for the requested compression rate R of the Tunstall code.

References

1. Ali, S.M.; Silvey, S.D. A general class of coefficients of divergence of one distribution from another. *J. R. Stat. Soc.* **1966**, *28*, 131–142. [CrossRef]
2. Csiszár, I. Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Bewis der Ergodizität von Markhoffschen Ketten. *Publ. Math. Inst. Hungar. Acad. Sci.* **1963**, *8*, 85–108. (In German)
3. Csiszár, I. A note on Jensen's inequality. *Studia Scientiarum Mathematicarum Hungarica* **1966**, *1*, 185–188.
4. Csiszár, I. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica* **1967**, *2*, 299–318.
5. Csiszár, I. On topological properties of f-divergences. *Studia Scientiarum Mathematicarum Hungarica* **1967**, *2*, 329–339.
6. Csiszár, I. A class of measures of informativity of observation channels. *Periodica Mathematicarum Hungarica* **1972**, *2*, 191–213. [CrossRef]
7. Morimoto, T. Markov processes and the H-theorem. *J. Phys. Soc. Jpn.* **1963**, *18*, 328–331. [CrossRef]
8. Liese F.; Vajda, I. Convex Statistical Distances. In *Teubner-Texte Zur Mathematik*; Springer: Leipzig, Germany, 1987; Volume 95.
9. Pardo, L. *Statistical Inference Based on Divergence Measures*; Chapman and Hall/CRC, Taylor & Francis Group: Boca Raton, FL, USA, 2006.
10. Pardo M.C.; Vajda, I. About distances of discrete distributions satisfying the data processing theorem of information theory. *IEEE Trans. Inf. Theory* **1997**, *43*, 1288–1293. [CrossRef]

11. Stummer W.; Vajda, I. On divergences of finite measures and their applicability in statistics and information theory. *Statistics* **2010**, *44*, 169–187. [CrossRef]
12. Vajda, I. *Theory of Statistical Inference and Information*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1989.
13. Ziv, J.; Zakai, M. On functionals satisfying a data-processing theorem. *IEEE Trans. Inf. Theory* **1973**, *19*, 275–283. [CrossRef]
14. Zakai, M.; Ziv, J. A generalization of the rate-distortion theory and applications. In *Information Theory—New Trends and Open Problems*; Longo, G., Ed.; Springer: Berlin/Heidelberg, Germany, 1975; pp. 87–123.
15. Merhav, N. Data processing theorems and the second law of thermodynamics. *IEEE Trans. Inf. Theory* **2011**, *57*, 4926–4939. [CrossRef]
16. Liese, F.; Vajda, I. On divergences and informations in statistics and information theory. *IEEE Trans. Inf. Theory* **2006**, *52*, 4394–4412. [CrossRef]
17. Csiszár, I.; Körner, J. *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2011.
18. Ahlswede R.; Gács, P. Spreading of sets in product spaces and hypercontraction of the Markov operator. *Ann. Probab.* **1976**, *4*, 925–939. [CrossRef]
19. Calmon, F.P.; Polyanskiy, Y.; Wu, Y. Strong data processing inequalities for input constrained additive noise channels. *IEEE Trans. Inf. Theory* **2018**, *64*, 1879–1892. [CrossRef]
20. Cohen, J.E.; Iwasa, Y.; Rautu, Gh.; Ruskai, M.B.; Seneta E.; Zbăganu, Gh. Relative entropy under mappings by stochastic matrices. *Linear Algebra Appl.* **1993**, *179*, 211–235. [CrossRef]
21. Cohen, J.E.; Kemperman, J.H.B.; Zbăganu, G. *Comparison of Stochastic Matrices with Applications in Information Theory, Statistics, Economics and Population Sciences*; Birkhäuser: Boston, MA, USA, 1998.
22. Makur, A.; Polyanskiy, Y. Comparison of channels: Criteria for domination by a symmetric channel. *IEEE Trans. Inf. Theory* **2018**, *64*, 5704–5725. [CrossRef]
23. Polyanskiy, Y.; Wu, Y. Dissipation of information in channels with input constraints. *IEEE Trans. Inf. Theory* **2016**, *62*, 35–55. [CrossRef]
24. Raginsky, M. Strong data processing inequalities and Φ-Sobolev inequalities for discrete channels. *IEEE Trans. Inf. Theory* **2016**, *62*, 3355–3389. [CrossRef]
25. Polyanskiy, Y.; Wu, Y. Strong data processing inequalities for channels and Bayesian networks. In *Convexity and Concentration*; Carlen, E., Madiman M., Werner, E.M., Eds.; Springer: Berlin/Heidelberg, Germany, 2017; Volume 161, pp. 211–249.
26. Makur, A.; Zheng, L. Linear bounds between contraction coefficients for f-divergences. *arXiv* **2018**, arxiv:1510.01844.v4.
27. Pearson, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1900**, *50*, 157–175. [CrossRef]
28. Neyman, J. Contribution to the theory of the χ^2 test. In Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 13–18 August 1945 and 27–29 January 1946; University of California Press: Berkeley, CA, USA, 1949; pp. 239–273.
29. Sarmanov, O.V. Maximum correlation coefficient (non-symmetric case). *In Selected Translations in Mathematical Statistics and Probability*; American Mathematical Society: Providence, RI, USA, 1962; Volume. 2, pp. 207–210.
30. Marshall, A.W.; Olkin, I.; Arnold, B.C. *Inequalities: Theory of Majorization and Its Applications*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2011.
31. Steele, J.M. *The Cauchy-Schwarz Master Class*; Cambridge University Press: Cambridge, UK, 2004.
32. Bhatia, R. *Matrix Analysis*; Springer: Berlin/Heidelberg, Germany, 1997.
33. Cicalese, F.; Gargano, L.; Vaccaro, U. Bounds on the entropy of a function of a random variable and their applications. *IEEE Trans. Inf. Theory* **2018**, *64*, 2220–2230. [CrossRef]
34. Sason, I. Tight bounds on the Rényi entropy via majorization with applications to guessing and compression. *Entropy* **2018**, **20**, 896. [CrossRef]
35. Ho, S.W.; Verdú, S. On the interplay between conditional entropy and error probability. *IEEE Trans. Inf. Theory* **2010**, *56*, 5930–5942. [CrossRef]

36. Ho, S.W.; Verdú, S. Convexity/concavity of the Rényi entropy and α-mutual information. In Proceedings of the 2015 IEEE International Symposium on Information Theory, Hong Kong, China, 14–19 June 2015; pp. 745–749.
37. Corless, R.M.; Gonnet, G.H.; Hare, D.E.G.; Jeffrey, D.J.; Knuth, D.E. On the Lambert W function. *Adv. Comput. Math.* **1996**, *5*, 329–359. [CrossRef]
38. Cicalese, F.; Gargano L.; Vaccaro, U. A note on approximation of uniform distributions from variable-to-fixed length codes. *IEEE Trans. Inf. Theory* **2006**, *52*, 3772–3777. [CrossRef]
39. Tsallis, C. Possible generalization of the Boltzmann-Gibbs statistics. *J. Stat. Phys.* **1988**, *52*, 479–487. [CrossRef]
40. Rényi, A. On measures of entropy and information. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 20 June–30 July 1960; University of California Press: Berkeley, CA, USA, 1961; Volume 1, pp. 547–561.
41. Cicalese, F.; Gargano L.; Vaccaro, U. Minimum-entropy couplings and their applications. *IEEE Trans. Inf. Theory* **2019**, *65*, 3436–3451. [CrossRef]
42. Sason, I.; Verdú, S. Arimoto-Rényi conditional entropy and Bayesian M-ary hypothesis testing. *IEEE Trans. Inf. Theory* **2018**, *64*, 4–25. [CrossRef]
43. Amari, S.; Nagaoka, H. *Methods of Information Geometry*; Oxford University Press: New York, NJ, USA, 2000.
44. Cichocki, A.; Amari, S.I. Families of Alpha- Beta- and Gamma- divergences: Flexible and robust measures of similarities. *Entropy* **2010**, *12*, 1532–1568. [CrossRef]
45. Sason, I. On f-divergences: Integral representations, local behavior, and inequalities. *Entropy* **2018**, *20*, 383. [CrossRef]
46. Fano, R.M. *Class Notes for Course 6.574: Transmission of Information*; MIT: Cambridge, MA, USA, 1952.
47. Ahlswede, R.; Gács, P.; Körner, J. Bounds on conditional probabilities with applications in multi-user communication. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **1976**, *34*, 157–177; Correction in **1977**, *39*, 353–354. [CrossRef]
48. Raginsky, M.; Sason, I. Concentration of measure inequalities in information theory, communications and coding: Third edition. In *Foundations and Trends (FnT) in Communications and Information Theory*; NOW Publishers: Delft, The Netherlands, 2019; pp. 1–266.
49. Chen, X.; Guntuboyina, A.; Zhang, Y. On Bayes risk lower bounds. *J. Mach. Learn. Res.* **2016**, *17*, 7687–7744.
50. Guntuboyina, A. Lower bounds for the minimax risk using f-divergences, and applications. *IEEE Trans. Inf. Theory* **2011**, *57*, 2386–2399. [CrossRef]
51. Kim, Y.H.; Sutivong, A.; Cover, T.M. State amplification. *IEEE Trans. Inf. Theory* **2008**, *54*, 1850–1859. [CrossRef]
52. Arimoto, S. Information measures and capacity of order α for discrete memoryless channels. In *Topics in Information Theory—2nd Colloquium*; Csiszár, I., Elias, P., Eds.; Colloquia Mathematica Societatis Janós Bolyai; Elsevier: Amsterdam, The Netherlands, 1977; Volume 16, pp. 41–52.
53. Ahlswede, R.; Körner, J. Source coding with side information and a converse for degraded broadcast channels. *IEEE Trans. Inf. Theory* **1975**, *21*, 629–637. [CrossRef]
54. Liu, J.; Cuff, P.; Verdú, S. E_γ resolvability. *IEEE Trans. Inf. Theory* **2017**, *63*, 2629–2658.
55. Brémaud, P. *Discrete Probability Models and Methods: Probability on Graphs and Trees, Markov Chains and Random Fields, Entropy and Coding*; Springer: Basel, Switzerland, 2017.
56. Tunstall, B.K. Synthesis of Noiseless Compression Codes. Ph.D. Thesis, Georgia Institute of Technology, Atlanta, GA, USA, 1967.
57. DeGroot, M.H. Uncertainty, information and sequential experiments. *Ann. Math. Stat.* **1962**, *33*, 404–419. [CrossRef]
58. Roberts, A.W.; Varberg, D.E. *Convex Functions*; Academic Press: Cambridge, MA, USA, 1973.
59. Rockafellar, R.T. *Convex Analysis*; Princeton University Press: Princeton, NJ, USA, 1996.
60. Collet, J.F. An exact expression for the gap in the data processing inequality for f-divergences. *IEEE Trans. Inf. Theory* **2019**, *65*, 4387–4391. [CrossRef]
61. Bregman, L.M. The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. Math. Phys.* **1967**, *7*, 200–217. [CrossRef]
62. Sason, I.; Verdú, S. f-divergence inequalities. *IEEE Trans. Inf. Theory* **2016**, *62*, 5973–6006. [CrossRef]
63. Gilardoni, G.L. On Pinsker's and Vajda's type inequalities for Csiszár's f-divergences. *IEEE Trans. Inf. Theory* **2010**, *56*, 5377–5386. [CrossRef]

64. Gibbs, A.L.; Su, F.E. On choosing and bounding probability metrics. *Int. Stat. Rev.* **2002**, *70*, 419–435. [CrossRef]
65. Simic, S. Second and third order moment inequalities for probability distributions. *Acta Math. Hung.* **2018**, *155*, 518–532. [CrossRef]
66. Van Erven, T.; Harremoës, P. Rényi divergence and Kullback–Leibler divergence. *IEEE Trans. Inf. Theory* **2014**, *60*, 3797–3820. [CrossRef]
67. Pardo, M.C.; Vajda, I. On asymptotic properties of information-theoretic divergences. *IEEE Trans. Inf. Theory* **2003**, *49*, 1860–1868. [CrossRef]
68. Beck, A. *Introduction to Nonlinear Optimization: Theory, Algorithms and Applications with Matlab*; SIAM-Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2014.
69. Simic, S. On logarithmic convexity for differences of power means. *J. Inequalities Appl.* **2008**, *2007*, 037359. [CrossRef]
70. Keziou, A. Dual representation of φ-divergences and applications. *C. R. Math.* **2003**, *336*, 857–862. [CrossRef]
71. Nguyen, X.; Wainwright, M.J.; Jordan, M.I. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Trans. Inf. Theory* **2010**, *56*, 5847–5861. [CrossRef]
72. Jelineck, F.; Schneider, K.S. On variable-length-to-block coding. *IEEE Trans. Inf. Theory* **1972**, *18*, 765–774. [CrossRef]