Measuring Dependencies of Order Statistics: An Information Theoretic Perspective

Alex Dytso*, Martina Cardone*, Cynthia Rush†
* New Jersey Institute of Technology, Newark, NJ 07102, USA Email: alex.dytso@njit.edu
* University of Minnesota, Minneapolis, MN 55404, USA, Email: mcardone@umn.edu
† Columbia University, New York, NY 10025, USA, Email: cynthia.rush@columbia.edu

***Abstract*—This work considers a random sample $X_1, X_2, ..., X_n$ drawn independently and identically distributed from some known parent distribution P_X with $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ being the order statistics of the sample. Under the assumption of an invertible cumulative distribution function associated with the parent distribution P_X, a distribution-free property is established showing that the f-divergence between the joint distribution of order statistics and the product distribution of order statistics does not depend on P_X. Moreover, it is shown that the mutual information between two subsets of order statistics also satisfies a distribution-free property; that is, it does not depend on P_X. Furthermore, the decoupling rates between $X_{(r)}$ and $X_{(m)}$ (i.e., rates at which the mutual information approaches zero) are characterized for various choices of (r, m). The work also considers discrete distributions, which do not satisfy the previously-stated invertibility assumption, and it is shown that no such distribution-free property holds: the mutual information between order statistics does depend on the parent distribution P_X. Upper bounds on the decoupling rates in the discrete setting are also established.**

I. Introduction

Consider a random sample $X_1, X_2, \ldots, X_n$ drawn independently and identically distributed (i.i.d.) from some known parent distribution P_X. Let the random variables $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ represent the order statistics of the sample. In this work, we are interested in studying the dependence between $X_{(\mathcal{I}_1)}$ and $X_{(\mathcal{I}_2)}$ where $\mathcal{I}_1$ and $\mathcal{I}_2$ are two arbitrary subsets of $\{1, \ldots, n\}$ and $X_{(\mathcal{I}_k)} = \{X_{(i)}\}_{i \in \mathcal{I}_k}$, for $k \in \{1, 2\}$. In particular, we choose to use the f-divergence and mutual information as measures of such dependence.

Our contributions and paper outline are as follows. In Section II, we consider the f-divergence and the mutual information of order statistics when the sample is drawn from a large family of parent distributions, namely, the set of all distributions having an invertible cumulative distribution function (cdf). Under this assumption, we show that the f-divergence between the joint distribution and the product distribution, as well as the mutual information between $X_{(\mathcal{I}_1)}$ and $X_{(\mathcal{I}_2)}$ do not depend on the parent distribution, for every finite n. We compute the exact value of the mutual information for the case $\mathcal{I}_1 = \{r\}$ and $\mathcal{I}_2 = \{m\}$, for integers $1 \leq r < m \leq n$. Furthermore, we characterize the rates of decoupling between $X_{(r)}$ and $X_{(m)}$ (i.e., rates at which the mutual information approaches zero) for various choices of (r, m). For example, we show that the minimum and maximum (i.e., $(r, m) = (1, n)$) decouple at a rate of $\frac{1}{n^2}$ while the median and maximum decouple at a rate of $\frac{1}{n}$. In Section III, we consider discrete distributions, which do not satisfy the invertibility assumption of Section II. In comparison to the results in Section II, we show that in the discrete setting, the mutual information between $X_{(r)}$ and $X_{(m)}$ does depend on the parent distribution. Nonetheless, we prove that the results in Section II can still be used as upper bounds on the decoupling rates in the discrete setting. Finally, to provide some comparisons, we compute the mutual information between $X_{(r)}$ and $X_{(m)}$ for the case when the parent distribution comes from the Bernoulli distribution.

The work of M. Cardone was supported in part by the U.S. National Science Foundation under Grant CCF-1849757.

Related Work. Order statistics have a broad range of applications including survival and reliability analysis, life testing, statistical quality control, filtering theory, signal processing, robustness and classification studies, radar target detection, and wireless communication. A comprehensive survey of applications of order statistics can be found in [1].

The distributional aspects of order statistics are well-studied in probability theory, and there exist several books on this theory [2]. Information measures of the distribution of order statistics have also received some attention. For example, the authors of [3] showed conditions under which the differential entropy of the order statistics characterizes the parent distribution. Other information measures that have been considered on the distribution of order statistics include the Rényi entropy [4], [5], the cumulative entropies [6], the Fisher information [7].

Distribution-free properties for information measures on order statistics have also been observed in the past. For instance, the authors of [5], for continuous distributions, have shown that the Rényi divergence between order statistics and their parent distribution does not depend on the underlying parent distribution. The authors of [8], for continuous distributions, have shown that the average entropy of the individual order statistics and the entropy of the parent distribution do not depend on the underlying parent distribution. The authors of [9], for continuous distributions, have shown that the mutual information between consecutive order statistics is independent of the parent distribution provided.

Notation. We use $[n]$ to denote the collection $\{1, 2, \ldots, n\}$. Logarithms are assumed to be in base e. The notation $\stackrel{D}{=}$ denotes equality in distribution. The harmonic number, denoted

as H_r, is defined as follows. For $r \in \mathbb{N}$,

$$H_r = \sum_{k=1}^{r} \frac{1}{k}. \tag{1}$$

We also define, for $r \in \mathbb{N}$,

$$T_r = \log(r!) - rH_r. \tag{2}$$

The Euler-Mascheroni constant is denoted by $\gamma \approx 0.5772$. Let $f : (0, \infty) \to \mathbb{R}$ be a convex function such that $f(0) = 1$. Then, for two probability distributions P and Q over a space Ω such that $P \ll Q$ (i.e., P is absolutely continuous with respect to Q), the f-divergence is defined as

$$D_f(P\|Q) = \int_{\Omega} f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}Q. \tag{3}$$

II. The Case of Distributions with Invertible CDFs

In this section, we consider a setting in which the cdf of a parent distribution is an invertible function (i.e., bijective function). Several classes of probability distributions satisfy this property. For example, all absolutely continuous distributions with a non-zero probability density function (pdf) – which are also those studied the most in conjunction with order statistics [10] – satisfy this property since the cdfs are strictly increasing and, hence, have an inverse. A non-example, however, is the set of discrete distributions with step functions for their cdfs, which do not have a proper inverse.

A. Distribution-Free Property for the f-divergence

We begin our study on the dependence structure of order statistics by showing that a large class of divergences, namely the f-divergence, have the following distribution-free property: if the cdf of the parent distribution is invertible, then the f-divergence between the joint distribution of order statistics and the product distribution of order statistics does not depend on the parent distribution.

Theorem 1. *Fix a subset $\mathcal{I} \subseteq [n]$ and assume that $X_1, \ldots, X_n$ i.i.d. $\sim P_X$, with P_X having an invertible cdf. Then,*

$$D_f\Big(P_{\{X_{(i)}\}_{i\in\mathcal{I}}}\Big\|\prod_{i\in\mathcal{I}} P_{X_{(i)}}\Big) = D_f\Big(P_{\{U_{(i)}\}_{i\in\mathcal{I}}}\Big\|\prod_{i\in\mathcal{I}} P_{U_{(i)}}\Big), \tag{4}$$

where $P_{\{X_{(i)}\}_{i\in\mathcal{I}}}$ and $\prod_{i\in\mathcal{I}} P_{X_{(i)}}$ are the joint distribution and the product distribution of the sequence $\{X_{(i)}\}_{i\in\mathcal{I}}$, respectively; $(U_{(1)}, \ldots, U_{(n)})$ are the order statistics associated with the sample $(U_1, \ldots, U_n)$ i.i.d. $\sim \mathcal{U}(0,1)$, where $\mathcal{U}(0,1)$ denotes the uniform distribution over $(0,1)$; and $P_{\{U_{(i)}\}_{i\in\mathcal{I}}}$ and $\prod_{i\in\mathcal{I}} P_{U_{(i)}}$ are the joint distribution and the product distribution of the sequence $\{U_{(i)}\}_{i\in\mathcal{I}}$, respectively.

Proof: Let F_X^{-1} be the inverse cdf of the parent distribution P_X. Recall that for $(U_1, \ldots, U_n)$ i.i.d. $\sim \mathcal{U}(0,1)$, we have $(X_1, \ldots, X_n) \overset{D}{=} (F_X^{-1}(U_1), \ldots, F_X^{-1}(U_n))$. Then, since $F_X^{-1}(\cdot)$ is order preserving (see [2, eq.(2.4.2)]), we have

$$X_{(\mathcal{I})} = \{X_{(i)}\}_{i\in\mathcal{I}} \overset{D}{=} \{F_X^{-1}(U_{(i)})\}_{i\in\mathcal{I}}. \tag{5}$$

Since F_X is a one-to-one mapping and the f-divergence is invariant under invertible transformations [11, Thm.14], we get

$$\begin{aligned} &D_f\Big(P_{\{X_{(i)}\}_{i\in\mathcal{I}}}\Big\|\prod_{i\in\mathcal{I}} P_{X_{(i)}}\Big) \\ &= D_f\Big(P_{\{F_X^{-1}(U_{(i)})\}_{i\in\mathcal{I}}}\Big\|\prod_{i\in\mathcal{I}} P_{F_X^{-1}(U_{(i)})}\Big) \\ &= D_f\Big(P_{\{U_{(i)}\}_{i\in\mathcal{I}}}\Big\|\prod_{i\in\mathcal{I}} P_{U_{(i)}}\Big). \end{aligned} \tag{6}$$

This concludes the proof of Theorem 1. ∎

We note that computing the f-divergence in (4) requires the knowledge of the joint distribution of $\{U_{(i)}\}_{i\in\mathcal{I}}$ for any subset $\mathcal{I}$. The joint pdf of this sequence can be readily computed and is given by the following expression [2]: let $\mathcal{I} = \{(i_1, i_2, \ldots, i_k) : 1 \le i_1, i_2, \ldots, i_k \le n\}$ where $|\mathcal{I}| = k$, then, $P_{\{U_{(i)}\}_{i\in\mathcal{I}}}$ is non-zero only if $-\infty < x_{(i_1)} < x_{(i_2)} < \ldots < x_{(i_k)} < \infty$, and, when this is true, its expression is

$$P_{\{U_{(i)}\}_{i\in\mathcal{I}}} = c_{\mathcal{I}} \prod_{t=1}^{k+1} \left[x_{(i_t)} - x_{(i_{t-1})}\right]^{i_t - i_{t-1}-1}, \tag{7}$$

where $x_{(i_0)} = x_{(i_{k+1})} = 0$, and, with $i_0 = 0$ and $i_{k+1} = n+1$,

$$c_{\mathcal{I}} = \frac{n!}{\prod_{t=1}^{k+1}(i_t - i_{t-1} - 1)!}.$$

The next result, the proof of which is in the Appendix, evaluates the Kullback-Leibler (KL) divergence, which is a special case of the f-divergence with $f(x) = x\log(x)$.

Proposition 1. *Under the assumptions of Theorem 1, where $\mathcal{I} \subseteq [n]$ with $|\mathcal{I}| = k$, we have that*

$$\begin{aligned} &D_{KL}\Big(P_{\{U_{(i)}\}_{i\in\mathcal{I}}}\Big\|\prod_{i\in\mathcal{I}} P_{U_{(i)}}\Big) \\ &= \sum_{t=2}^{k}(T_{i_t-1} - T_{i_t - i_{t-1}-1}) + \sum_{t=1}^{k-1} T_{n-i_t} - (k-1)T_n. \end{aligned} \tag{8}$$

In particular,

(Whole Sequence). *For $\mathcal{I} = [n]$, we have that*

$$D_{KL}\Big(P_{\{U_{(1)},\ldots,U_{(n)}\}}\Big\|\prod_{i=1}^{n} P_{U_{(i)}}\Big) = 2\sum_{t=2}^{n} T_{t-1} - (n-1)T_n.$$

(Min and Max). *For $\mathcal{I} = \{1, n\}$, we have that*

$$D_{KL}\Big(P_{\{U_{(1)},U_{(n)}\}}\Big\|P_{U_{(1)}}\,P_{U_{(n)}}\Big) = \log\left(\frac{n-1}{n}\right) + \frac{1}{n-1}.$$

Using the result of Proposition 1, we can study convergence rates of the KL divergence when $n \to \infty$. For example, when $\mathcal{I} = \{1, n\}$, we have that

$$\begin{aligned} &\lim_{n\to\infty} n^2 D_{\mathrm{KL}}\Big(P_{\{U_{(1)},U_{(n)}\}}\Big\|P_{U_{(1)}}\,P_{U_{(n)}}\Big) \\ &\quad = \lim_{n\to\infty} n^2\left[\log\left(\frac{n-1}{n}\right) + \frac{1}{n-1}\right] = \frac{1}{2}, \end{aligned}$$

where the last equality follows by using the Maclaurin series for the natural logarithm. Thus, when the KL divergence is considered, the joint and product distributions of the minimum and maximum converge at a rate equal to $1/n^2$.

B. Distribution-Free Property for the Mutual Information

Here we consider the mutual information measure, which we have recently shown to be a suitable base measure to quantify the level of informativeness that a subset of order statistics contains on the random sample $X_1, X_2, \ldots, X_n$ [12]. In particular, as a special case of the approach used for the proof of Theorem 1, we have the following result.

Corollary 1. *Assume that $X_1, \ldots, X_n$ i.i.d. $\sim P_X$, with P_X having an invertible cdf and fix two sets $\mathcal{I}_1, \mathcal{I}_2 \subseteq [n]$. Then,*

$$I(X_{(\mathcal{I}_1)}; X_{(\mathcal{I}_2)}) = I(U_{(\mathcal{I}_1)}; U_{(\mathcal{I}_2)}), \tag{9}$$

where $X_{(\mathcal{I}_k)} = \{X_{(i)}\}_{i \in \mathcal{I}_k}$ and $U_{(\mathcal{I}_k)} = \{U_{(i)}\}_{i \in \mathcal{I}_k}$ both for $k \in \{1, 2\}$. Consequently, $I(X_{(\mathcal{I}_1)}; X_{(\mathcal{I}_2)})$ is not a function of P_X, the parent distribution of X. Moreover, for $r < m$,

$$I(X_{(r)}; X_{(m)}) = T_{m-1} + T_{n-r} - T_{m-r-1} - T_n. \tag{10}$$

Proof: The proof of (9) follows along the same lines as the proof of Theorem 1 and relies on the invariance of the mutual information to one-to-one transformations.

To compute (10), recall that the mutual information can be written as a KL divergence, and then using the result in (8) of Proposition 1, we obtain

$$\begin{aligned} I(U_{(r)}; U_{(m)}) &= D_{\mathrm{KL}}(P_{U_{(r)}, U_{(m)}} \| P_{U_{(r)}} P_{U_{(m)}}) \\ &= T_{m-1} + T_{n-r} - T_{m-r-1} - T_n. \end{aligned}$$

This concludes the proof of Corollary 1. ∎

We notice that if $\mathcal{I}_1 \cap \mathcal{I}_2 \neq \varnothing$ then $I(X_{(\mathcal{I}_1)}; X_{(\mathcal{I}_2)}) = I(U_{(\mathcal{I}_1)}; U_{(\mathcal{I}_2)}) = \infty$. Moreover, since the mutual information is symmetric, the result also holds for $r > m$.

Remark 1. *For other measures of dependence of random variables, the distribution independence property of Corollary 1 does not necessarily hold. For example, as demonstrated in [13, Appendix B], the covariance of a pair of order statistics from a sample drawn according to an exponential distribution with rate λ is*

$$\mathrm{Cov}(X_{(1)}, X_{(2)}) = 1/(\lambda^2 n^2). \tag{11}$$

The result in (10) in Corollary 1 is stated in terms of factorials and harmonic numbers, as captured by the T's defined in (2). In the following lemma (see [13, Appendix C] for the proof), we provide an alternative formulation for these T's that will be helpful for the large sample size analysis.

Lemma 1. *For $k > 0$,*

$$T_k = k \log\left(\frac{2k}{2k+1}\right) + \frac{1}{2}\log(2\pi k) - (1+\gamma)k - e(k), \tag{12}$$

$$T_{k+1} - T_k = \log\left(\frac{2k+2}{2k+3}\right) - (1+\gamma) + \frac{1}{k+1} - c(k), \tag{13}$$

where

$$\frac{k}{24(k+1)^2} - \frac{1}{12k} \leq e(k) \leq \frac{1}{24k} - \frac{1}{12k+1}, \tag{14}$$

$$\frac{1}{24(k+2)^2} \leq c(k) \leq \frac{1}{24(k+1)^2}. \tag{15}$$

C. Large Sample Size Asymptotics of Mutual Information

Using Corollary 1 and the approximations in Lemma 1, we now study the rates of decoupling of the order statistics as the sample size grows. In particular, we have the next theorem, which is proved in [13, Appendix D].

Theorem 2. *Under the assumptions of Corollary 1:*

1) (r^{th} vs. Max). Fix some $r \geq 1$ independent of n. Then,

$$\lim_{n\to\infty} n^2 I(X_{(r)}; X_{(n)}) = r/2. \tag{16}$$

2) (r^{th} vs. m^{th}). Fix some $1 \leq r < m$ independent of n. Then,

$$\lim_{n\to\infty} I(X_{(r)}; X_{(m)}) = T_{m-1} - T_{m-r-1} + (1+\gamma)r. \tag{17}$$

3) (k-Step). Fix some $k \geq 1$. Then,

$$\begin{aligned} \lim_{n\to\infty} I(X_{(n-k)}; X_{(n)}) &= \log(k) - H_{k-1} + \gamma \\ &= \log\left(\frac{k}{k+\frac{1}{2}}\right) + \frac{1}{k} - c(k). \end{aligned} \tag{18}$$

4) ($\lfloor \alpha n \rfloor$ vs. $\lceil \beta n \rceil$). Fix $0 < \alpha < \beta < 1$, with (α, β) independent of n. Then,

$$\lim_{n\to\infty} I(X_{(\lfloor \alpha n \rfloor)}; X_{(\lceil \beta n \rceil)}) = \frac{1}{2}\log\left(\frac{\beta(1-\alpha)}{\beta - \alpha}\right). \tag{19}$$

5) ($\lfloor \alpha n \rfloor$ vs. Max). Fix some $0 < \alpha < 1$ with α independent of n. Then,

$$\lim_{n\to\infty} nI(X_{(\lfloor \alpha n \rfloor)}; X_{(n)}) = \frac{\alpha}{2(1-\alpha)}. \tag{20}$$

Some interesting special cases of the above results include the following:

$$\text{1 step: } \lim_{n\to\infty} I(X_{(n-1)}; X_{(n)}) = \gamma, \tag{21}$$

$$Q_3 \text{ vs. Max: } \lim_{n\to\infty} nI(X_{(\lfloor \frac{3n}{4} \rfloor)}; X_{(n)}) = \frac{3}{2}, \tag{22}$$

$$\text{Median vs. Max: } \lim_{n\to\infty} nI(X_{(\lfloor \frac{n}{2} \rfloor)}; X_{(n)}) = \frac{1}{2}, \tag{23}$$

$$Q_1 \text{ vs. Max: } \lim_{n\to\infty} nI(X_{(\lfloor \frac{n}{4} \rfloor)}; X_{(n)}) = \frac{1}{6}. \tag{24}$$

For comparison, Fig. 1 demonstrates how the median decouples from the maximum for finite values of n.

Remark 2. *We compare the rate of decoupling of the mutual information $I(U_{(r)}; U_{(m)})$ for integers $r < m$ to that of the covariance between $U_{(r)}$ and $U_{(m)}$ given by [2],*

$$\mathrm{Cov}(U_{(r)}, U_{(m)}) = \frac{r(n-m+1)}{(n+1)^2(n+2)}. \tag{25}$$

Note that, although this comparison is somewhat unfair as the covariance only captures correlation, it can still be used as a proxy for measuring independence. From Table I, we observe that the rates of decoupling of the mutual information and covariance are always different. Moreover, we also note a surprising behavior for Cases 2-4: although the mutual information does not decouple, the covariance goes to zero either at a rate $1/n^2$ (Cases 2-3) or at a rate $1/n$ (Case 4).

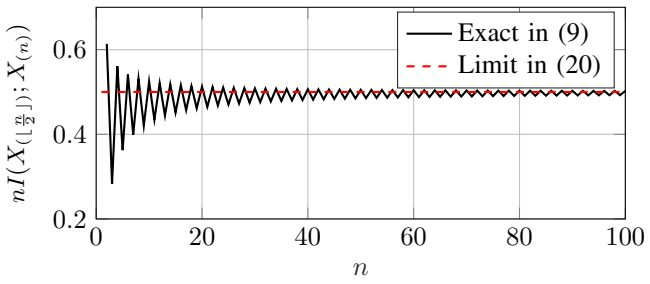

Fig. 1. Convergence of $nI(X_{(\lfloor \frac{n}{2} \rfloor)}; X_{(n)})$ (Median vs. Max) to (23).

Finally, by comparing Case 4 and Case 5, we observe that there is a phase transition at $\beta = 1$, i.e., in Case 4 ($0 < \beta < 1$) the mutual information does not decouple, whereas in Case 5 ($\beta = 1$) it decouples at a rate $1/n$.

	$\mathrm{Cov}(U_{(r)}, U_{(m)})$	$I(U_{(r)}; U_{(m)})$
Case 1: $m = n$	$1/n^3$	$1/n^2$
Case 2: (m, r) fixed	$1/n^2$	No decoupling
Case 3: $r = n - k$ and $m = n$	$1/n^2$	No decoupling
Case 4: $r = \lfloor \alpha n \rfloor$ and $m = \lceil \beta n \rceil$	$1/n$	No decoupling
Case 5: $r = \lfloor \alpha n \rfloor$ and $m = n$	$1/n^2$	$1/n$

TABLE I
RATES OF DECOUPLING FOR $\mathrm{Cov}(U_{(r)}, U_{(m)})$ AND $I(U_{(r)}; U_{(m)})$.

III. THE CASE OF DISCRETE DISTRIBUTIONS

In this section, we consider the case when the parent distribution is discrete. Historically, order statistics with a discrete distribution have received far less attention than those with a continuous distribution. However, recently, since discrete distributions naturally occur in several practical situations (e.g., image processing), discrete order statistics have started to receive more attention in the literature [14], [15].

The mutual information of discrete order statistics often behaves differently from that of continuous order statistics. For example, while order statistics $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ from a continuous distribution form a Markov chain [2], for the case of discrete distributions, the order statistics form a Markov chain if and only if the parent distribution has at most two points in its support [10].

The next theorem (see [13, Appendix E] for the proof) shows that, unlike in the case of continuous order statistics, when the parent distribution is discrete, the mutual information between $X_{(r)}$ and $X_{(m)}$ can indeed depend on the parent distribution.

Theorem 3. *Suppose that the parent distribution is Bernoulli with parameter $p \in (0, 1)$. Then, for $r \leq m$, we have that*

$$\begin{aligned} I(X_{(r)}; X_{(m)}) &= -P(B \geq m) \log \left(P(B \geq r)\right) \\ &+ (P(B \geq r) - P(B \geq m)) \log \left(\frac{P(B \geq r) - P(B \geq m)}{P(B \geq r)(1 - P(B \geq m))} \right) \\ &- \left(1 - P(B \geq r)\right) \log \left(1 - P(B \geq m)\right), \end{aligned} \tag{26}$$

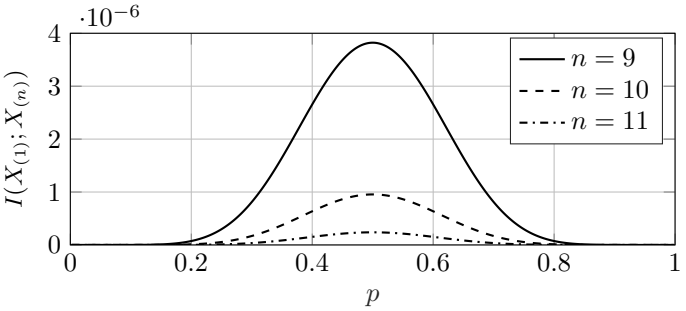

Fig. 2. $I(X_{(1)}; X_{(n)})$ in (27) versus $p \in (0, 1)$.

where B is $\mathrm{Binomial}(n, 1-p)$. *Consequently,*

$$\begin{aligned} &I(X_{(1)}; X_{(n)}) \\ &= -(1-p)^n \log \left(1 - p^n\right) - p^n \log \left(1 - (1-p)^n\right) \\ &+ (1 - p^n - (1-p)^n) \log \left(\frac{1 - p^n - (1-p)^n}{(1-p^n)(1-(1-p)^n)} \right). \end{aligned} \tag{27}$$

Theorem 3 provides an example of a discrete parent distribution (i.e., Bernoulli with parameter $p \in (0, 1)$) for which the mutual information *does* depend on the parent distribution (i.e., parameter p). For illustration, Fig. 2 shows the dependence of $I(X_{(1)}; X_{(n)})$ in (27) over $p \in (0, 1)$.

The fact that the mutual information does depend on the parent distribution prevents us from having universal results similar to those derived in Section II. However, the results in Section II can still be used as upper bounds as we show next.

Theorem 4. *Assume that $X_1, \ldots, X_n$ i.i.d. $\sim P_X$, with P_X having an arbitrary parent distribution. Then,*

- *(f-divergence). For $\mathcal{I} \subset [n]$,*

$$D_f\left(P_{\{X_{(i)}\}_{i \in \mathcal{I}}} \,\middle\|\, \prod_{i \in \mathcal{I}} P_{X_{(i)}}\right) \leq D_f\left(P_{\{U_{(i)}\}_{i \in \mathcal{I}}} \,\middle\|\, \prod_{i \in \mathcal{I}} P_{U_{(i)}}\right), \tag{28}$$

- *(Mutual Information). For $\mathcal{I}_1, \mathcal{I}_2 \subset [n]$,*

$$I(X_{(\mathcal{I}_1)}; X_{(\mathcal{I}_2)}) \leq I(U_{(\mathcal{I}_1)}; U_{(\mathcal{I}_2)}). \tag{29}$$

Proof: The proof relies on the data processing inequality. Due to space constraints we only show it for $I(X_{(r)}; X_{(m)})$.

We start by defining the quantile function as $F_X^{-1}(y) = \sup\{x : F_X(x) \leq y\}$. As discussed in the proof of Theorem 1, for an arbitrary parent distribution [2, eq.(1.1.3)],

$$(X_{(r)}, X_{(m)}) \stackrel{D}{=} (F_X^{-1}(U_{(r)}), F_X^{-1}(U_{(m)})), \tag{30}$$

thus $I(F_X^{-1}(U_{(r)}); F_X^{-1}(U_{(m)})) = I(X_{(r)}; X_{(m)})$. Moreover,

$$I(U_{(r)}; U_{(m)}) \geq I(F_X^{-1}(U_{(r)}); F_X^{-1}(U_{(m)})) = I(X_{(r)}; X_{(m)}),$$

where the inequality uses the data processing inequality for the mutual information since $U_{(r)} \to U_{(m)} \to F_X^{-1}(U_{(m)})$ and $F^{-1}(U_{(r)}) \to U_{(r)} \to F_X^{-1}(U_{(m)})$ are Markov chains. ■

The bound in (29) is appealing since all the results derived in Section II (e.g., Theorem 2) can be used to obtain upper bounds on $I(X_{(r)}; X_{(m)})$ for any arbitrary parent distribution. However, the bound in (29) can be very suboptimal. For example, as shown in Fig. 3, we have that

$\lim_{n\to\infty} I(X_{(n-1)};X_{(n)}) = 0$ for the Bernoulli case, while $\lim_{n\to\infty} I(U_{(n-1)};U_{(n)}) = \gamma$, computed in (21).

Remark 3. *Theorem 4 can be generalized to arbitrary non-overlapping subset of order statistics. Moreover, it can be generalized to f-divergences.*

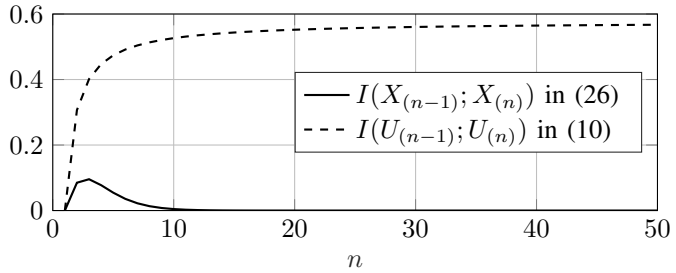

Fig. 3. $I(X_{(n-1)};X_{(n)})$ in (26) and $I(U_{(n-1)};U_{(n)})$ in (10) versus n.

APPENDIX

The computation for an arbitrary $\mathcal{I} \subseteq [n]$ with $|\mathcal{I}| = k$ proceeds as follows. First, notice

$$D_{\mathrm{KL}}\Big(P_{\{U_{(i)}\}_{i\in\mathcal{I}}} \Big\| \prod_{i\in\mathcal{I}} P_{U_{(i)}}\Big) \overset{(a)}{=} \mathbb{E}\left[\log\left(\frac{c_{\mathcal{I}}\prod_{t=1}^{k+1}\left[U_{(i_t)}-U_{(i_{t-1})}\right]^{i_t-i_{t-1}-1}}{\prod_{t=1}^{k} c_{i_t}\, U_{(i_t)}^{i_t-1}\,(1-U_{(i_t)})^{n-i_t}}\right)\right],$$

where (a) follows by using the expression of the joint distribution in (7) and $P_{U_{(r)}}(x) = c_r\, x^{r-1}\,(1-x)^{n-r}$, where $c_r = \frac{n!}{(r-1)!(n-r)!}$. Then, we simplify further as follows,

$$\begin{aligned}
&D_{\mathrm{KL}}\Big(P_{\{U_{(i)}\}_{i\in\mathcal{I}}} \Big\| \prod_{i\in\mathcal{I}} P_{U_{(i)}}\Big) \\
&= \log\Big(c_{\mathcal{I}}\prod_{t=1}^{k} c_{i_t}^{-1}\Big) - \sum_{t=1}^{k}(i_t-1)\mathbb{E}[\log(U_{(i_t)})] \\
&\quad - \sum_{t=1}^{k}(n-i_t)\mathbb{E}[\log(1-U_{(i_t)})] \\
&\quad + \sum_{t=1}^{k+1}(i_t-i_{t-1}-1)\mathbb{E}[\log(U_{(i_t)}-U_{(i_{t-1})})] \\
&= \log\Big(c_{\mathcal{I}}\prod_{t=1}^{k} c_{i_t}^{-1}\Big) - \sum_{t=1}^{k}(i_t-1)(\psi(i_t)-\psi(n+1)) \\
&\quad - \sum_{t=1}^{k}(n-i_t)(\psi(n+1-i_t)-\psi(n+1)) \\
&\quad + \sum_{t=1}^{k+1}(i_t-i_{t-1}-1)(\psi(i_t-i_{t-1})-\psi(n+1)), \qquad (31)
\end{aligned}$$

where we have used the fact that $U_{(m)} \sim \mathrm{Beta}(m, n+1-m)$ and $1-U_{(r)} \sim \mathrm{Beta}(n+1-r, r)$ with the difference $U_{(m)} - U_{(r)} \sim \mathrm{Beta}(m-r, n-(m-r)+1)$. Then, (see, e.g., [2]),

$$\begin{aligned}
\mathbb{E}\left[\log(U_{(m)})\right] &= \psi(m) - \psi(n+1), \\
\mathbb{E}\left[\log(1-U_{(r)})\right] &= \psi(n+1-r) - \psi(n+1), \\
\mathbb{E}\left[\log(U_{(m)}-U_{(r)})\right] &= \psi(m-r) - \psi(n+1),
\end{aligned}$$

where $\psi(\cdot)$ is the digamma function and where we use the convention that $U_{(0)} = 0$ and $U_{(n+1)} = 1$. Finally, collecting terms and using the result of (31), we have

$$\begin{aligned}
&D_{\mathrm{KL}}\Big(P_{\{U_{(i)}\}_{i\in\mathcal{I}}} \Big\| \prod_{i\in\mathcal{I}} P_{U_{(i)}}\Big) \\
&\overset{(a)}{=} \log\Big(c_{\mathcal{I}}\prod_{t=1}^{k} c_{i_t}^{-1}\Big) + \sum_{t=1}^{k+1}(i_t-i_{t-1}-1)\psi(i_t-i_{t-1}) \\
&\quad - \sum_{t=1}^{k}(i_t-1)\psi(i_t) - \sum_{t=1}^{k}(n-i_t)\psi(n+1-i_t) \\
&\quad + \psi(n+1)\,(k(n-1)-(n-k)) \\
&\overset{(b)}{=} -\sum_{t=1}^{k+1} T_{i_t-i_{t-1}-1} + \sum_{t=1}^{k} T_{i_t-1} + \sum_{t=1}^{k} T_{n-i_t} - (k-1)T_n, \qquad (32)
\end{aligned}$$

where the labeled equalities follow from: (a) the fact that $\sum_{t=1}^{k+1}(i_t - i_{t-1} - 1) = i_{k+1} - i_0 - (k+1) = n-k$, where we recall that $i_0 = 0$ and $i_{k+1} = n+1$; and (b) using the identity $\psi(n) = H_{n-1} - \gamma$, with γ being the Euler-Mascheroni constant, noting that the terms that multiply γ cancel out, and using the definition of T_n from (2). The result in (8) comes from canceling the terms $T_{i_1-i_0-1} = T_{i_1-1}$ and $T_{i_{k+1}-i_k-1} = T_{n-i_k}$ from (32). The special cases of $\mathcal{I} = [n]$ and $\mathcal{I} = \{1, n\}$ can be found in [13, Appendix A].

REFERENCES

[1] C. R. Rao and V. Govindaraju, *Handbook of Statistics*. Elsevier, 2006, vol. 17.

[2] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja, *A First Course in Order Statistics*. Siam, 1992, vol. 54.

[3] S. Baratpour, J. Ahmadi, and N. R. Arghami, "Some characterizations based on entropy of order statistics and record values," *Communications in Statistics-Theory and Methods*, vol. 36, no. 1, pp. 47–57, 2007.

[4] ——, "Characterizations based on Rényi entropy of order statistics and record values," *Journal of Statistical Planning and Inference*, vol. 138, no. 8, pp. 2544–2551, 2008.

[5] M. Abbasnejad and N. R. Arghami, "Renyi entropy properties of order statistics," *Communications in Statistics-Theory and Methods*, vol. 40, no. 1, pp. 40–52, 2010.

[6] N. Balakrishnan, F. Buono, and M. Longobardi, "On cumulative entropies in terms of moments of order statistics," *arXiv:2009.02029*, 2020.

[7] G. Zheng, N. Balakrishnan, and S. Park, "Fisher information in ordered data: A review," *Statistics and its Interface*, vol. 2, pp. 101–113, 2009.

[8] K. M. Wong and S. Chen, "The entropy of ordered sequences and order statistics," *IEEE Transactions on Information Theory*, vol. 36, no. 2, pp. 276–284, 1990.

[9] N. Ebrahimi, E. S. Soofi, and H. Zahedi, "Information properties of order statistics and spacings," *IEEE Transactions on Information Theory*, vol. 50, no. 1, pp. 177–183, 2004.

[10] H. N. Nagaraja, "On the non-Markovian structure of discrete order statistics," *Journal of Statistical Planning and Inference*, vol. 7, no. 1, pp. 29–33, 1982.

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

[12] A. Dytso, M. Cardone, and C. Rush, "The most informative order statistic and its application to image denoising," *arXiv:2101.11667*, 2021.

[13] ——, "Measuring dependencies of order statistics: An information theoretic perspective," *arXiv:2009.12337*, 2020.

[14] D. L. Evans, L. M. Leemis, and J. H. Drew, "The distribution of order statistics for discrete random variables with applications to bootstrapping," *INFORMS Journal on Comput.*, vol. 18, no. 1, pp. 19–30, 2006.

[15] A. Dembińska, "Discrete order statistics," *Wiley StatsRef: Statistics Reference Online*, 2014.