Shared Information for a Markov Chain on a Tree

Sagnik Bhattacharya and Prakash Narayan[†]

2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834365

***Abstract*—Shared information is a measure of mutual dependence among $m \geq 2$ jointly distributed discrete random variables. For a Markov chain on a tree with a given joint distribution, we give a new proof of an explicit characterization of shared information. When the joint distribution is not known, we exploit the special form of this characterization to provide a multiarmed bandit algorithm for estimating shared information, and analyze its error performance.**

***Index Terms*—Shared information, Markov chain on a tree, correlated bandits.**

I. Introduction

Let $X_1, \ldots, X_m$, $m \geq 2$ be random variables (rvs) with finite alphabets $\mathcal{X}_1, \ldots, \mathcal{X}_m$, respectively, and joint probability mass function (pmf) $P_{X_1 \cdots X_m}$. The *shared information* $\mathrm{SI}(X_1, \ldots, X_m)$ of the rvs $X_1, \ldots, X_m$ is a measure of mutual dependence among them, and for $m = 2$, $\mathrm{SI}(X_1, X_2)$ particularizes to mutual information $\mathrm{I}(X_1 \wedge X_2)$. Consider m terminals, with Terminal i having privileged access to independent and identically distributed (i.i.d.) repetitions of X_i, $i = 1, \ldots, m$. Shared information $\mathrm{SI}(X_1, \ldots, X_m)$ has the operational meaning of being the largest rate of *shared common randomness* that the m terminals can generate in a decentralized manner upon cooperating among themselves by means of interactive, publicly broadcast and noise-free communication[1]. Shared information measures the maximum rate of common randomness that is (nearly) independent of the open communication used to generate it.

The (Kullback-Leibler) divergence-based expression for $\mathrm{SI}(X_1, \ldots, X_m)$ was discovered in [15, Example 4], where it was derived as an upper bound for a single-letter formula for the "secret key capacity of a source model" with m terminals, a concept defined by the operational meaning above. The upper bound was shown to be tight for $m = 2$ and 3. Subsequently, in a significant advance [6], [11], [8], tightness of the upper bound was established for arbitrary m, thereby imbuing $\mathrm{SI}(X_1, \ldots, X_m)$ with the operational significance of being the mentioned maximum rate of shared secret common randomness. The potential for shared information to serve as a natural measure of mutual dependence of $m \geq 2$ rvs, in the manner of mutual information for $m = 2$ rvs, was suggested in [24]; see also [25].

A comprehensive study of shared information [9], where it is termed "multivariate mutual information," examines the role of secret key capacity as a measure of mutual dependence among multiple rvs and derives important properties including structural features of an underlying optimization along with connections to the theory of submodular functions.

In addition to constituting secret key capacity for a multiterminal source model ([15], [6], [11]), shared information also affords operational meaning for: maximal packing of edge-disjoint spanning trees in a multigraph ([27], [26]; see also [7], [14], [9] for variant models); optimum querying exponent for resolving common randomness [31]; strong converse for multiterminal secret key capacity [31], [32]; and also undirected network coding [8], data clustering [10], among others.

As argued in [9], shared information also possesses several attributes of measures of dependence among $m \geq 2$ rvs proposed earlier, including Watanabe's total correlation [33] and Han's dual total correlation [19] (both mentioned in Section II). For $m = 2$ rvs, measures of common information due to Gács-Körner [17], Wyner [34] and Tyagi [30] have operational meanings; extensions to $m > 2$ rvs merit further study (see, however, [22]).

For a given joint pmf $P_{X_1 \cdots X_m}$ of the rvs $X_1, \ldots, X_m$, an explicit characterization of $\mathrm{SI}(X_1, \ldots, X_m)$ can be challenging (see Definition 1 below); exact formulas are available for special cases (cf. e.g., [15], [27], [9]). An efficient algorithm for calculating $\mathrm{SI}(X_1, \ldots, X_m)$ is given in [9].

Our focus in this paper is on a Markov chain on a tree (MCT) [18]. Tree-structured probabilistic graphical models are appealing owing to desirable statistical properties that enable, for instance, efficient algorithms for exact inference [21], [29]; decoding [23], [21]; sampling [16]; and structure learning [12]. We take the tree structure of our model to be known; algorithms exist already for learning tree structure from data samples [12], [13]. We exploit the special form of $P_{X_1 \cdots X_m}$ in the setting of an MCT to obtain a simple characterization for shared information. When the joint pmf $P_{X_1 \cdots X_m}$ is not known but the tree structure is, the said characterization facilitates an estimation of shared information.

In the setting of an MCT [18], our contributions are two-fold. First, we derive an explicit characterization of shared information for an MCT with a given joint pmf $P_{X_1 \cdots X_m}$ by means of a direct approach that exploits tree structure and Markovity of the pmf. A characterization of shared information had been sketched already in [15]; our new proof does not seek recourse to a secret key interpretation of shared information, unlike in [15]. Also our proof differs in a material way from that in prior work [10] with a similar objective. Second, when $P_{X_1 \cdots X_m}$ is not known, with the mentioned characterization serving as a linchpin, we provide an approach for estimating shared information for an MCT. Formulated as a correlated

[†]S. Bhattacharya and P. Narayan are with the Department of Electrical and Computer Engineering and the Institute for Systems Research, University of Maryland, College Park, MD 20742, USA. E-mail: {sagnikb, prakash}@umd.edu. This work was supported by the U.S. National Science Foundation under Grant CCF1910497.

[1]Our preferred nomenclature of shared information is justified by its operational meaning.

bandits problem [4], this approach seeks to identify the best arm-pair across which mutual information is minimal. Using a uniform sampling of arms, redolent of sampling mechanisms in [3], we provide an upper bound for the probability of estimation error and associated sample complexity. Our uniform sampling algorithm is similar to that in [2], [4]; however, our modified analysis takes into account estimator bias, a feature that is not common in known bandit algorithms. Also, this approach can accommodate more refined bandit algorithms as also alternatives to the probability of error criterion such as regret [5].

Section II contains the preliminaries. An explicit characterization of shared information for an MCT with a given $P_{X_1\cdots X_m}$ is provided in Section III. Section IV describes our approach for estimating shared information when $P_{X_1\cdots X_m}$ is not known.

II. Preliminaries

Let $X_1,\ldots,X_m$, $m\geq 2$, be rvs with finite alphabets $\mathcal{X}_1,\ldots,\mathcal{X}_m$, respectively, and joint pmf $P_{X_1\cdots X_m}$. For $A\subseteq\mathcal{M}=\{1,\ldots,m\}$, we write $X_A=(X_i,i\in A)$. Let $\pi=(\pi_1,\ldots,\pi_k)$ denote a k-partition of $\mathcal{M}$, $2\leq k\leq m$. All logarithms and exponentiations are with respect to the base 2, except when indicated otherwise.

Definition 1 (Shared information)**.** *The* shared information *of* $X_1,\ldots,X_m$ *is defined as*

$$\mathrm{SI}(X_{\mathcal{M}}) = \min_{2\leq k\leq m}\ \min_{\pi=(\pi_u,u=1,\cdots,k)}\ \frac{1}{k-1}D(P_{X_{\mathcal{M}}}\parallel\prod_{u=1}^{k}P_{X_{\pi_u}}).$$

Given a partition π of $\mathcal{M}$ with $2\leq|\pi|\leq m$ atoms, it will be convenient to denote

$$\mathcal{I}(\pi)=\frac{1}{|\pi|-1}D(P_{X_{\mathcal{M}}}\parallel\prod_{u=1}^{|\pi|}P_{X_{\pi_u}})$$

so that $\mathrm{SI}(X_{\mathcal{M}})=\min_{2\leq|\pi|\leq m}\mathcal{I}(\pi)$.

Example 1. *For* $\mathcal{M}=\{1,2\}$*, we have*

$$\mathrm{SI}(X_1,X_2)=\textit{mutual information}\ \mathrm{I}(X_1\wedge X_2)$$

and for $\mathcal{M}=\{1,2,3\}$*, it is checked readily that* $\mathrm{SI}(X_1,X_2,X_3)$ *is the minimum of* $\mathrm{I}(X_1\wedge X_2,X_3)$*,* $I(X_2\wedge X_1,X_3)$*,* $\mathrm{I}(X_3\wedge X_1,X_2)$ *and*

$$\frac{1}{2}\left[\mathrm{H}(X_1)+\mathrm{H}(X_2)+\mathrm{H}(X_3)-\mathrm{H}(X_1,X_2,X_3)\right].$$

Shared information possesses several properties befitting a measure of mutual dependence among multiple rvs. Clearly $\mathrm{SI}(X_{\mathcal{M}})\geq 0$, and equality holds iff $P_{X_{\mathcal{M}}}=P_{X_A}P_{X_{A^c}}$ for some $A\subsetneq\mathcal{M}$; the latter follows from [15, Theorem 5] and [6], [11], [8]. When $X_1,\ldots,X_m$ are bijections of each other, i.e., $\mathrm{H}(X_i\,|\,X_j)=0$, $1\leq i\neq j\leq m$, then $\mathrm{SI}(X_{\mathcal{M}})=\mathrm{H}(X_1)$, as expected [9].

Next, the secret key capacity interpretation of $\mathrm{SI}(X_{\mathcal{M}})$ [15], [6], [11], [8], [25] implies that upon grouping the rvs $X_1,\ldots,X_m$ into teams represented by the atoms of any k-partition $\pi=(\pi_1,\ldots,\pi_k)$ of $\mathcal{M}$, $2\leq k\leq m$, the resulting shared information of the teamed rvs can be only larger, i.e.,

$$\mathrm{SI}(X_{\pi_1},\ldots,X_{\pi_k})\geq\mathrm{SI}(X_1,\ldots,X_m). \tag{1}$$

Suppose that $\pi^*=(\pi_1^*,\ldots,\pi_l^*)$, $l\geq 2$, attains $\mathrm{SI}(X_{\mathcal{M}})>0$ (not necessarily uniquely) in Definition 1, i.e,

$$\mathrm{SI}(X_{\mathcal{M}})=\frac{1}{l-1}D(P_{X_{\mathcal{M}}}\parallel\prod_{u=1}^{l}P_{X_{\pi_u^*}}). \tag{2}$$

A simple but useful observation based on Definition 1, (1) and (2) is that upon agglomerating the rvs in each atom of an optimum partition $\pi^*=(\pi_1^*,\ldots,\pi_l^*)$, the resulting shared information of the teams, $\mathrm{SI}(X_{\pi_1^*},\ldots,X_{\pi_l^*})$, equals the shared information $\mathrm{SI}(X_{\mathcal{M}})$ of the (unteamed) rvs $X_1,\ldots,X_m$, and cannot be increased in the manner of (1) by further coalitions formed out of $X_{\pi_1^*},\ldots,X_{\pi_l^*}$. This property has benefited information-clustering applications (cf. e.g., [9], [10]).

Shared information satisfies the data processing inequality [9]. For $X_{\mathcal{M}}=(X_1,\ldots,X_m)$, consider $X'_{\mathcal{M}}=(X'_1,\ldots,X'_m)$ where for a fixed $1\leq j\leq m$, $X'_i=X_i$ for $i\in\mathcal{M}\backslash\{j\}$ and X'_j is obtained as the output of a stochastic matrix $W:\mathcal{X}_j\to\mathcal{X}_j$ with input X_j. Then, $\mathrm{SI}(X'_{\mathcal{M}})\leq\mathrm{SI}(X_{\mathcal{M}})$.

It is worth comparing $\mathrm{SI}(X_{\mathcal{M}})$ with two well-known measures of correlation among $X_1,\ldots,X_m$, $m\geq 2$, of a similar vein. Watanabe's *total correlation* [33] is defined by

$$\mathcal{C}(X_{\mathcal{M}})=D(P_{X_{\mathcal{M}}}\parallel\prod_{i=1}^{m}P_{X_i})=\sum_{i=1}^{m-1}\mathrm{I}(X_{i+1}\wedge X_1,\ldots,X_i)$$

and Han's *dual total correlation* [19] by

$$\mathcal{D}(X_{\mathcal{M}})=\sum_{i=1}^{m}\mathrm{H}(X_{\mathcal{M}\backslash\{i\}})-(m-1)\,\mathrm{H}(X_{\mathcal{M}}).$$

These measures satisfy

$$\mathrm{SI}(X_{\mathcal{M}})\leq\frac{1}{m-1}\,\mathcal{C}(X_{\mathcal{M}}),\quad\mathrm{SI}(X_{\mathcal{M}})\leq\mathcal{D}(X_{\mathcal{M}}).$$

When $\mathcal{M}=\{1,2\}$,

$$\mathrm{SI}(X_1,X_2)=\mathcal{C}(X_1,X_2)=\mathcal{D}(X_1,X_2)=\mathrm{I}(X_1\wedge X_2).$$

Our focus is on shared information for a Markov chain on a tree.

Definition 2 (Markov Chain on a Tree)**.** *Let* $\mathcal{G}=(\mathcal{M},\mathcal{E})$ *be a tree with vertex set* $\mathcal{M}=\{1,\ldots,m\}$*,* $m\geq 2$*, i.e., a* connected *graph containing no circuits. For* (i,j) *in the edge set* $\mathcal{E}$*, let* $\mathcal{B}(i\leftarrow j)$ *denote the set of all vertices connected with* j *by a path containing the edge* (i,j)*. The rvs* $X_1,\ldots,X_m$ *form a* Markov chain on a tree *(MCT)* $\mathcal{G}$ *if for every* $(i,j)\in\mathcal{E}$*, the conditional pmf of* X_j *given* $X_{\mathcal{B}(i\leftarrow j)}=\{X_l:l\in\mathcal{B}(i\leftarrow j)\}$ *depends only on* X_i*. Specifically,* X_j *is conditionally independent of* $X_{\mathcal{B}(i\leftarrow j)\backslash\{i\}}$ *when conditioned on* X_i*. Thus,* $P_{X_{\mathcal{M}}}$ *is such that for each* $(i,j)\in\mathcal{E}$*,*

$$P_{X_j\,|\,\mathcal{B}(i\leftarrow j)}=P_{X_j\,|\,X_i}. \tag{3}$$

When $\mathcal{G}$ is a chain, an MCT reduces to a standard Markov chain.

As will be seen below, estimation of $\mathrm{SI}(X_{\mathcal{M}})$ for an MCT will entail estimating $\mathrm{I}(X_i \wedge X_j)$, $(i,j) \in \mathcal{E}$. We close this section with pertinent tools that will be used to this end.

Let $(X_t, Y_t)_{t=1}^n$ be $n \geq 1$ i.i.d. repetitions of rvs (X, Y) with (unknown) pmf P_{XY} of *assumed full support* on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are finite sets. For $(\mathbf{x}, \mathbf{y})$ in $\mathcal{X}^n \times \mathcal{Y}^n$, let $P_{\mathbf{xy}}^{(n)}$ represent their joint type on $\mathcal{X} \times \mathcal{Y}$. An estimate of $\mathrm{I}(X \wedge Y) = \mathrm{I}_{P_{XY}}(X \wedge Y)$ on the basis of $(\mathbf{x}, \mathbf{y})$ in $\mathcal{X}^n \times \mathcal{Y}^n$ is provided by the empirical mutual information or the 'plug-in' estimator

$$\hat{\mathrm{I}}^{(n)}(\mathbf{x} \wedge \mathbf{y}) = \mathrm{H}(P_{\mathbf{x}}^{(n)}) + \mathrm{H}(P_{\mathbf{y}}^{(n)}) - \mathrm{H}(P_{\mathbf{xy}}^{(n)}) \tag{4}$$

where $P_{\mathbf{x}}^{(n)}$ and $P_{\mathbf{x}}^{(n)}$ are the (marginal) types of $\mathbf{x}$ and $\mathbf{y}$, respectively [1].

Lemma 1 (Bias of empirical mutual information estimator)**.** *The bias*

$$\mathrm{Bias}(\hat{\mathrm{I}}^{(n)}(\mathbf{X} \wedge \mathbf{Y})) \triangleq \mathbb{E}_{P_{XY}}\left[\hat{\mathrm{I}}^{(n)}(\mathbf{X} \wedge \mathbf{Y})\right] - \mathrm{I}(X \wedge Y)$$

satisfies

$$-\log\left(1 + \frac{|\mathcal{X}|-1}{n}\right)\left(1 + \frac{|\mathcal{Y}|-1}{n}\right) \leq \mathrm{Bias}(\hat{\mathrm{I}}^{(n)}(\mathbf{X} \wedge \mathbf{Y})) \leq \log\left(1 + \frac{|\mathcal{X}|\,|\mathcal{Y}|-1}{n}\right).$$

Proof. The proof follows immediately from [28, Proposition 1]. □

A concentration bound for the estimator $\hat{\mathrm{I}}^{(n)}$ above using techniques from [1], is given by[2]

Lemma 2. *For $\epsilon > 0$ and every $n \geq 1$,*

$$P_{XY}\left(\left|\hat{\mathrm{I}}^{(n)}(\mathbf{X} \wedge \mathbf{Y}) - \mathbb{E}_{P_{XY}}\left[\hat{\mathrm{I}}^{(n)}(\mathbf{X} \wedge \mathbf{Y})\right]\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{18\log^2 n}\right).$$

Proof. The empirical mutual information $\hat{\mathrm{I}}^{(n)} : \mathcal{X}^n \times \mathcal{Y}^n \to \mathbb{R}^+ \cup \{0\}$ satisfies the bounded differences property, namely

$$\max_{\substack{(\mathbf{x},\mathbf{y}) \in \mathcal{X}^n \times \mathcal{Y}^n \\ (x_i', y_i') \in \mathcal{X} \times \mathcal{Y}}} \left|\hat{\mathrm{I}}^{(n)}(\mathbf{x} \wedge \mathbf{y}) - \hat{\mathrm{I}}^{(n)}((x_1^{i-1}, x_i', x_{i+1}^n) \wedge (y_1^{i-1}, y_i', y_{i+1}^n))\right| \leq \frac{6\log n}{n}$$

for $1 \leq i \leq n$, where for $l < k$, $x_l^k = (x_l, x_{l+1}, \ldots, x_k)$. The previous bound is obtained upon noting that the left-side is bounded above by six terms via the triangle inequality, the first of which is

$$\left|P_{\mathbf{xy}}^{(n)}(x_i, y_i)\log P_{\mathbf{xy}}^{(n)}(x_i, y_i) - \left(P_{\mathbf{xy}}^{(n)}(x_i, y_i) - \frac{1}{n}\right)\log\left(P_{\mathbf{xy}}^{(n)}(x_i, y_i) - \frac{1}{n}\right)\right|.$$

[2]In Lemma 2 and Theorem 4, exponentiation is with respect to e.

Each of these terms is $\leq \log n/n$, using the inequality [1]

$$\left|\frac{j+1}{n}\log\frac{j+1}{n} - \frac{j}{n}\log\frac{j}{n}\right| \leq \frac{\log n}{n}, \qquad 0 \leq j < n.$$

The claim of the lemma then follows by a standard application of McDiarmid's Bounded Differences Inequality [5, Lemma A.7]. □

III. Shared Information for a Markov Chain on a Tree

Our first main result is a new proof of an explicit characterization of $\mathrm{SI}(X_{\mathcal{M}})$ for an MCT. While the upper bound for $\mathrm{SI}(X_{\mathcal{M}})$ is akin to that involving secret key capacity in [15], the proof of the lower bound uses an altogether new approach based on the structure of a "good" partition π in Definition 1.

Theorem 3. *Let $\mathcal{G} = (\mathcal{M}, \mathcal{E})$ be an MCT with pmf $P_{X_{\mathcal{M}}}$ in* (3)*. Then*

$$\mathrm{SI}(X_{\mathcal{M}}) = \min_{(i,j)\in\mathcal{E}} \mathrm{I}(X_i \wedge X_j). \tag{5}$$

Proof. As shown in [15],

$$\mathrm{SI}(X_{\mathcal{M}}) \leq \min_{(i,j)\in\mathcal{E}} \mathrm{I}(X_i \wedge X_j) \tag{6}$$

and is seen as follows. For each $(i,j) \in \mathcal{E}$, consider a partition of $\mathcal{M}$ with $k = 2$ atoms, viz. $\pi = \pi((i,j)) = (\pi_1, \pi_2)$ where $\pi_1 = \mathcal{B}(i \leftarrow j)$, $\pi_2 = \mathcal{B}(j \leftarrow i)$. Then,

$$\mathrm{I}(X_{\pi_1} \wedge X_{\pi_2}) = \mathrm{I}(X_{\mathcal{B}(i\leftarrow j)} \wedge X_{\mathcal{B}(j\leftarrow i)}) = \mathrm{I}(X_i \wedge X_j) \tag{7}$$

by the Markov property (3). Hence,

$$\mathrm{SI}(X_{\mathcal{M}}) \leq \mathrm{I}(X_{\pi_1} \wedge X_{\pi_2}) = \mathrm{I}(X_i \wedge X_j), \qquad (i,j) \in \mathcal{E}$$

leading to (6).

Next, we show that

$$\mathrm{SI}(X_{\mathcal{M}}) \geq \min_{(i,j)\in\mathcal{E}} \mathrm{I}(X_i \wedge X_j). \tag{8}$$

This is done in two steps. First, we show that for any k-partition π of $\mathcal{M}$, $k \geq 2$, with (individually) connected atoms, $\mathcal{I}(\pi) \geq \min_{(i,j)\in\mathcal{E}} \mathrm{I}(X_i \wedge X_j)$. Second, an argument is sketched to show that for any k-partition $\pi = (\pi_1, \ldots, \pi_k)$ containing disconnected atoms, there exists a k'-partition $\pi' = (\pi', \ldots, \pi_{k'}')$, possibly with $k' \neq k$, and with fewer disconnected atoms such that $\mathcal{I}(\pi') \leq \mathcal{I}(\pi)$.

Step 1: Let $\pi = (\pi_1, \ldots, \pi_k)$, $k \geq 2$, be a k-partition such that each atom π_i is a connected set. Each such atom is connected directly to another atom by exactly one edge in $\mathcal{E}$ (owing to the absence of circuits in $\mathcal{G}$). Let $\mathcal{E}' \subseteq \mathcal{E}$ denote the collection of such edges. It follows that the atoms $\pi_1, \ldots, \pi_k$, taken as vertices, together with the edges in $\mathcal{E}'$, constitute a tree. Furthermore, it follows from Definition 2 that if X_{π_u} and X_{π_v} are connected by the edge $(u,v) \in \mathcal{E}'$,

$$\mathrm{I}(X_{\pi_u} \wedge X_{\pi_v}) = \mathrm{I}(X_u \wedge X_v), \qquad (u,v) \in \mathcal{E}'. \tag{9}$$

Now, let $\pi_1, \ldots, \pi_k$ be an enumeration of the atoms, obtained from a breadth-first search run on the agglomerated MCT with π_1 as the root vertex. Then,

$$\begin{aligned}\mathcal{I}(\pi) &= \frac{1}{k-1} D(P_{X_{\mathcal{M}}} \parallel \prod_{u=1}^{k} P_{X_{\pi_u}}) \\ &= \frac{1}{k-1} \sum_{u=2}^{k} \mathrm{I}(X_{\pi_u} \wedge X_{\pi_1}, \ldots, X_{\pi_{u-1}}) \\ &= \frac{1}{k-1} \sum_{u=2}^{k} \mathrm{I}(X_{\pi_u} \wedge X_{\text{parent}(\pi_u)}) \\ &\geq \min_{(u,v)\in\mathcal{E}'} \mathrm{I}(X_u \wedge X_v) \geq \min_{(i,j)\in\mathcal{E}} \mathrm{I}(X_i \wedge X_j),\end{aligned}$$

where the last equality follows from the Markov property of the agglomerated MCT and the first inequality is by (9).

Step 2: Consider first the case $k = 2$. Let $(\bar{i}, \bar{j})$ be the (not necessarily unique) minimizer in the right-side of (5). Take any 2-partition $\pi = (\pi_1, \pi_2)$ with possibly disconnected atoms, where $\pi_1 = \cup_{\rho=1}^{r} C_\rho$ and $\pi_2 = \cup_{\sigma=1}^{s} D_\sigma$ are unions of disjoint components. Noting that some C_ρ and D_σ must be connected by some edge (i, j) in $\mathcal{E}$, we have

$$\begin{aligned}\mathcal{I}(\pi) = \mathrm{I}(X_{\pi_1} \wedge X_{\pi_2}) \geq \mathrm{I}(X_{C_\rho} \wedge X_{D_\sigma}) &\geq \mathrm{I}(X_i \wedge X_j) \\ &\geq \mathrm{I}(X_{\bar{i}} \wedge X_{\bar{j}})\end{aligned}$$

where the lower bound is attained by the 2-partition with connected atoms $(\mathcal{B}(\bar{i} \leftarrow \bar{j}), \mathcal{B}(\bar{j} \leftarrow \bar{i}))$ as in (7).

Next, consider a k-partition $\pi = (\pi_1, \ldots, \pi_k)$, $k \geq 3$, and suppose that the atom π_1 is not connected. Without loss of generality, assume π_1 to be the (disjoint) union of connected components $A_1, \ldots, A_t$, $t \geq 2$ (which, at an extreme, can be the individual vertices constituting π_1). In the same vein, each A_l, $l = 1, \ldots, t$, can be taken to be maximally connected in π_1, i.e., A_l is connected and has the attribute that addition to A_l of a vertex in $\pi_1 \setminus A_l$ will render it disconnected. In general, any connected component of π_1 that is not maximally connected can be enlarged to absorb vertices outside it in π_1 that do not render it disconnected.

Take any A_l, say $A_l = A_{\bar{l}}$, and consider all its boundary edges, namely those edges for which one vertex is in $A_{\bar{l}}$ and the other outside it. As $A_{\bar{l}}$ is maximally connected in π_1, for each boundary edge the outside vertex cannot belong to π_1 and so must lie in $\mathcal{M} \setminus \pi_1$. Also, every such outside vertex associated with $A_{\bar{l}}$ must be the root of a subtree and, like $A_{\bar{l}}$, every A_l, $l \neq \bar{l}$, too, must be a subset of one such subtree linked to $A_{\bar{l}}$ – owing to connectedness within $A_{\bar{l}}$. Furthermore, since $A_1, \ldots, A_t$ are connected, and only through the subtrees rooted in $\mathcal{M} \setminus \pi_1$, there must exist at least one A_l such that all $A_{l'}$s, $l' \neq l$, are subsets of one subtree linked to A_l. In other words, denoting this A_l as A, we note that A has the property that

$$\pi_1 \setminus A = \bigcup_{\substack{l\in\{1,\cdots,t\}: \\ A_l \neq A}} A_l$$

is contained entirely in a subtree rooted at an outside vertex associated with A and lying in $\mathcal{M} \setminus \pi_1$. Let this vertex be $j \in \mathcal{M} \setminus \pi_1$, and let $\pi_u \in \pi$ be the atom that contains j. Since vertex j separates A from $\pi_1 \setminus A$, so does π_u. By (3), it follows that

$$A \multimap \pi_u \multimap \pi_1 \setminus A$$

whereby

$$\mathrm{I}(X_A \wedge X_{\pi_1 \setminus A}) \leq \mathrm{I}(X_{\pi_u} \wedge X_{\pi_1 \setminus A}) \leq \mathrm{I}(X_{\pi_u} \wedge X_{\pi_1}). \quad (10)$$

Next, consider the $(k-1)$-partition π' and the $(k+1)$-partition π'' of $\mathcal{M}$, defined by

$$\pi' = \left(\pi_1 \cup \pi_u, \{\pi_v\}_{v\neq 1, v\neq u}\right), \quad (11)$$

$$\pi'' = \left(\pi_1 \setminus A, A, \pi_u, \{\pi_v\}_{v\neq 1, v\neq u}\right). \quad (12)$$

We claim that

$$\mathcal{I}(\pi) \geq \min\left\{\mathcal{I}(\pi'), \mathcal{I}(\pi'')\right\}. \quad (13)$$

Referring to (11) and (12), we can infer from the claim (13) that for a given k-partition π with a disconnected atom π_1 as above, merging a disconnected atom with another atom (as in (11)) or breaking it to create a connected atom (as in (12)), lead to partitions π' or π'', of which at least one has $\mathcal{I}$-value not more than that of π. This argument is repeated until a partition with connected atoms is reached.

It remains to show (13). Suppose (13) were not true, i.e.,

$$\mathcal{I}(\pi) < \min\left\{\mathcal{I}(\pi'), \mathcal{I}(\pi'')\right\}.$$

Then,

$$\begin{aligned}\mathcal{I}(\pi) < \mathcal{I}(\pi') &\Leftrightarrow (k-2)\mathcal{I}(\pi) < (k-2)\mathcal{I}(\pi') \\ &\Leftrightarrow \mathrm{I}(X_{\pi_u} \wedge X_{\pi_1}) < \mathcal{I}(\pi), \quad (14)\end{aligned}$$

and similarly,

$$\begin{aligned}\mathcal{I}(\pi) < \mathcal{I}(\pi'') &\Leftrightarrow k\mathcal{I}(\pi) < k\mathcal{I}(\pi'') \\ &\Leftrightarrow \mathcal{I}(\pi) < \mathrm{I}(X_{\pi_1 \setminus A} \wedge X_A) \quad (15)\end{aligned}$$

where the second equivalences in (14) and (15) are obtained by straightforward manipulation. By (14) and (15),

$$\mathrm{I}(X_{\pi_u} \wedge X_{\pi_1}) < \mathrm{I}(X_{\pi_1 \setminus A} \wedge X_A)$$

which contradicts (10). Hence, (13) is true. □

IV. Estimating Shared Information for an MCT

Consider the estimation of $\mathrm{SI}(X_{\mathcal{M}})$ when the pmf of the rv $X_{\mathcal{M}} = (X_1, \ldots, X_m)$ is unknown to an "agent" who, however, knows the tree $\mathcal{G} = (\mathcal{M}, \mathcal{E})$. We assume in this section that $\mathcal{X}_1 = \cdots = \mathcal{X}_m = \mathcal{X}$, say, and further that the minimizing edge $(\bar{i}, \bar{j})$ in the right side of (5) is unique. By Theorem 3, $\mathrm{SI}(X_{\mathcal{M}})$ equals the minimum mutual information across an edge in the tree $\mathcal{G}$. Treating the determination of this edge as a correlated bandits problem of best-arm-pair identification, we provide an algorithm to pinpoint it, and analyze its error performance and associated sample complexity. *The estimate of shared information is taken to be the mutual information across the best arm-pair.* This estimation procedure is motivated by the special form of $\mathrm{SI}(X_{\mathcal{M}})$ in Theorem 3.

In the parlance of banditry, the environment has m arms, one arm corresponding to each vertex in $\mathcal{G} = (\mathcal{M}, \mathcal{E})$. The agent can pull, in any step, two arms that are connected by an edge in $\mathcal{E}$. Each action of the agent is specified by the pair (i, j), $1 \le i < j \le m$, $(i, j) \in \mathcal{E}$, with associated reward being the realizations $(X_i = x_i, X_j = x_j)$. The agent is allowed to pull a total of N pairs of arms, say. By means of these actions, the agent seeks to form estimates of all two-dimensional marginal pmfs $P_{X_iX_j}$ for (i, j) as above, and subsequently identify $(\bar{i}, \bar{j}) \in \mathcal{E}$. Let $X_{\mathcal{M}}^N$ denote N i.i.d. repetitions of $X_{\mathcal{M}} = (X_1, \ldots, X_m)$. Specifically, the agent must produce an estimate $\hat{e}_N = \hat{e}_N(X_{\mathcal{M}}^N) \in \mathcal{E}$ of $(\bar{i}, \bar{j}) \in \mathcal{E}$ at the end of N steps so as to minimize the error probability $P(\hat{e}_N \neq (\bar{i}, \bar{j}))$; and an estimate of SI as that of the mutual information across $\hat{e}_N$.

Denote $\Delta_{ij} = \mathrm{I}(X_i \wedge X_j) - \mathrm{I}(X_{\bar{i}}, X_{\bar{j}})$, $(i, j) \in \mathcal{E}$, and

$$\Delta_1 = \min_{\substack{(i,j)\in\mathcal{E} \\ (i,j)\neq(\bar{i},\bar{j})}} \mathrm{I}(X_i \wedge X_j) - \mathrm{I}(X_{\bar{i}} \wedge X_{\bar{j}}),$$

where the latter is the difference between the second-lowest and lowest mutual information across edges in $\mathcal{E}$. Note that $\Delta_1 > 0$ by the assumed uniqueness of the minimizing edge $(\bar{i}, \bar{j})$.

The estimation scheme below uses uniform sampling with pairs of rvs corresponding to edges of the tree being sampled equally often. Suppose that the agent samples a pair of arms corresponding to an edge $n \ge 1$ times; owing to uniform sampling, $N = |\mathcal{E}|\, n$. Let $x_{\mathcal{M}}^N$ represent a realization of $X_{\mathcal{M}}^N$. For each $(i, j) \in \mathcal{E}$, the agent computes the empirical mutual information estimate $\hat{\mathrm{I}}^{(n)}(\mathbf{x}_i \wedge \mathbf{x}_j)$ of $\mathrm{I}(X_i \wedge X_j)$ (see (4)). Note that the sampling of arm-pairs occurs over different steps. Define $\hat{e}_N(X_{\mathcal{M}}^N) = \arg_{(i,j)\in\mathcal{E}} \min \hat{\mathrm{I}}^{(n)}(\mathbf{x}_i \wedge \mathbf{y}_j)$. We take as our estimate of SI to be the mutual information estimate $\hat{\mathrm{I}}^{(n)}$ across the edge $\hat{e}_N$. Our second main result is an upper bound for the probability that $\hat{e}_N \neq (\bar{i}, \bar{j})$.

Theorem 4 (Probability of estimation error for uniform sampling)**.** *For uniform sampling, the probability of error in identifying the optimal pair of arms is*

$$P_{X_{\mathcal{M}}}\left(\hat{e}_N(X_{\mathcal{M}}^N) \neq (\bar{i}, \bar{j})\right) \le 4\,|\mathcal{E}| \exp\left(\frac{-(N/\,|\mathcal{E}|)\Delta_1^2}{648 \log^2(N/\,|\mathcal{E}|)}\right)$$

if

$$N > |\mathcal{E}| \max\left\{\frac{|\mathcal{X}|^2 - 1}{2^{\Delta_1/3} - 1}, \frac{|\mathcal{X}| - 1}{2^{\Delta_1/6} - 1}\right\}. \tag{16}$$

Proof. By Lemma 1, the bias of the estimate $\hat{\mathrm{I}}^{(n)}(X_i \wedge X_j)$ is bounded above as

$$\begin{aligned}
&\left|\,\mathrm{Bias}(\hat{\mathrm{I}}^{(n)}(X_i \wedge X_j))\right| \\
&\quad \le \max\left\{\log\left(1 + \frac{|\mathcal{X}|^2 - 1}{n}\right), 2\log\left(1 + \frac{|\mathcal{X}| - 1}{n}\right)\right\} \\
&\quad \le \frac{\Delta_1}{3} \le \frac{\Delta_{ij}}{3} \qquad \text{for } (i, j) \neq (\bar{i}, \bar{j})
\end{aligned} \tag{17}$$

where the second inequality follows from (16).

The idea underlying the theorem uses the concentration bound for $\hat{\mathrm{I}}^{(n)}$ in Lemma 2 together with the bound for bias in (17) to show that with large probability, the estimates $\hat{\mathrm{I}}^{(n)}(X_i \wedge X_j)$ and $\hat{\mathrm{I}}^{(n)}(X_{\bar{i}} \wedge X_{\bar{j}})$, $(i, j) \neq (\bar{i}, \bar{j})$, are separated so as to enable the corresponding edges to be distinguished. To simplify notation below, we denote $\hat{\mathrm{I}}_{ij}^{(n)} = \hat{\mathrm{I}}^{(n)}(X_i \wedge X_j)$ and $\mathrm{I}_{ij} = \mathrm{I}(X_i \wedge X_j)$, $(i, j) \in \mathcal{E}$.

For each $(i, j) \in \mathcal{E}$, consider the event

$$\mathcal{T}_{ij} = \left\{ \mathrm{I}_{\bar{i}\bar{j}} - \frac{\Delta_{ij}}{2} < \hat{\mathrm{I}}_{\bar{i}\bar{j}}^{(n)} < \mathrm{I}_{\bar{i}\bar{j}} + \frac{\Delta_{ij}}{2}, \right. \\ \left. \mathrm{I}_{ij} - \frac{\Delta_{ij}}{2} < \hat{\mathrm{I}}_{ij}^{(n)} < \mathrm{I}_{ij} + \frac{\Delta_{ij}}{2} \right\}$$

that the estimates $\hat{\mathrm{I}}_{\bar{i}\bar{j}}^{(n)}$ and $\hat{\mathrm{I}}_{ij}^{(n)}$ are both close to the respective true values. Then, with $P = P_{X_{\mathcal{M}}}$, we have

$$\begin{aligned}
&P\left(\hat{e}_N(X_{\mathcal{M}}^N) \neq (\bar{i}, \bar{j})\right) \\
&\qquad = P\left(\hat{\mathrm{I}}_{\bar{i}\bar{j}}^{(n)} \ge \hat{\mathrm{I}}_{ij}^{(n)} \text{for some } (i, j) \neq (\bar{i}, \bar{j})\right) \\
&\qquad \le \sum_{(i,j)\neq(\bar{i},\bar{j})} P\left(\hat{\mathrm{I}}_{\bar{i}\bar{j}}^{(n)} \ge \hat{\mathrm{I}}_{ij}^{(n)}\right) \\
&\qquad \le \sum_{(i,j)\neq(\bar{i},\bar{j})} P(\mathcal{T}_{ij}^c).
\end{aligned}$$

Using the bound for bias in (17) and Lemma 2 with $\epsilon = \Delta_{ij}/6$, it follows that

$$P(\mathcal{T}_{ij}^c) \le 4 \exp\left(\frac{-n\Delta_1^2}{648 \log^2 n}\right).$$

Thus,

$$P\left(\hat{e}_N(X_{\mathcal{M}}^N) \neq (\bar{i}, \bar{j})\right) \le 4\,|\mathcal{E}| \exp\left(\frac{-(N/\,|\mathcal{E}|)\Delta_1^2}{648 \log^2(N/\,|\mathcal{E}|)}\right). \square$$

V. Closing Remarks

The proof of Theorem 3 implies that for *any* partition π with disconnected atoms, there is a partition with connected atoms that has $\mathcal{I}$-value less than or equal to that of π. This structural property is stronger than that needed for Theorem 3.

In Section IV, we have resorted to a simple uniform sampling strategy for the sake of simplicity. As in [2] and [4], it is expected that a successive rejects algorithm would yield a better sample complexity for our estimator, and could be improved further by using more refined estimators. For instance, known estimators for entropic quantities with lower bias, e.g., jack-knifed estimators [28] and polynomial approximation-based estimators [20], can be expected to yield better error performance. By Theorem 4, since the probability of estimation error decays as $\exp(-O(N/\log^2 N))$. It remains open if this dependence on N can be bettered.

Acknowledgement

SB thanks Priyanka Kaswan for helpful discussions regarding the proof of Theorem 4.

REFERENCES

[1] A. Antos and I. Kontoyiannis, "Convergence properties of functional estimates for discrete distributions," *Random Structures & Algorithms*, vol. 19, no. 3-4, 2001.

[2] J.-y. Audibert, S. Bubeck, and R. Munos, "Best arm identification in multi-armed bandits," in *Proceedings of the Twenty-Third Annual Conference on Learning Theory*, 2010.

[3] V. P. Boda and P. Narayan, "Universal sampling rate distortion," *IEEE Transactions on Information Theory*, vol. 64, no. 12, Dec. 2018.

[4] V. P. Boda and L. A. Prashanth, "Correlated bandits or: How to minimize mean-squared error online," in *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, PMLR, Jun. 2019.

[5] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[6] C. Chan, "On tightness of mutual dependence upper-bound for secret-key capacity of multiple terminals," *ArXiv*, vol. abs/0805.3200, 2008.

[7] C. Chan, "Linear perfect secret key agreement," in *2011 IEEE Information Theory Workshop*, 2011.

[8] C. Chan, "The hidden flow of information," *2011 IEEE International Symposium on Information Theory Proceedings*, 2011.

[9] C. Chan, A. Al-Bashabsheh, J. B. Ebrahimi, T. Kaced, and T. Liu, "Multivariate mutual information inspired by secret-key agreement," *Proceedings of the IEEE*, vol. 103, no. 10, 2015.

[10] C. Chan, A. Al-Bashabsheh, Q. Zhou, T. Kaced, and T. Liu, "Info-clustering: A mathematical theory for data clustering," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, 2016.

[11] C. Chan and L. Zheng, "Mutual dependence for secret key agreement," in *2010 44th Annual Conference on Information Sciences and Systems (CISS)*, 2010.

[12] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Transactions on Information Theory*, 1968.

[13] C. Chow and T. Wagner, "Consistency of an estimate of tree-dependent probability distributions (corresp.)," *IEEE Transactions on Information Theory*, vol. 19, 1973.

[14] T. A. Courtade and T. R. Halford, "Coded cooperative data exchange for a secret key," *IEEE Transactions on Information Theory*, vol. 62, 2016.

[15] I. Csiszár and P. Narayan, "Secrecy capacities for multiple terminals," *IEEE Transactions on Information Theory*, vol. 50, no. 12, Dec. 2004.

[16] W. Feng, N. K. Vishnoi, and Y. Yin, "Dynamic sampling from graphical models," *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, 2019.

[17] P. Gács and J. Körner, "Common information is far less than mutual information," *Problems of Control and Information Theory*, vol. 2, Jan. 1973.

[18] H.-O. Georgii, *Gibbs Measures and Phase Transitions*. De Gruyter, 2011.

[19] T. S. Han, "Nonnegative entropy measures of multivariate symmetric correlations," *Information and Control*, vol. 36, no. 2, 1978.

[20] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Minimax estimation of functionals of discrete distributions," *IEEE Transactions on Information Theory*, vol. 61, no. 5, 2015.

[21] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.

[22] W. Liu, G. Xu, and B. Chen, "The common information of n dependent random variables," *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2010.

[23] D. J. C. MacKay, *Information Theory, Inference & Learning Algorithms*. USA: Cambridge University Press, 2002.

[24] P. Narayan, "Omniscience and secrecy," Plenary Talk, *IEEE International Symposium on Information Theory*, Cambridge, MA, 2012.

[25] P. Narayan and H. Tyagi, "Multiterminal secrecy by public discussion," *Foundations and Trends in Communications and Information Theory*, vol. 13, no. 2-3, 2016.

[26] S. Nitinawarat and P. Narayan, "Perfect omniscience, perfect secrecy, and Steiner tree packing," *IEEE Trans. Inf. Theory*, vol. 56, 2010.

[27] S. Nitinawarat, C. Ye, A. Barg, P. Narayan, and A. Reznik, "Secret key generation for a pairwise independent network model," *IEEE Transactions on Information Theory*, vol. 56, no. 12, Dec. 2010.

[28] L. Paninski, "Estimation of entropy and mutual information," *Neural Comput.*, vol. 15, no. 6, Jun. 2003.

[29] J. Pearl, "Reverend Bayes on inference engines: A distributed hierarchical approach," in *Proceedings of the Second AAAI Conference on Artificial Intelligence*, ser. AAAI'82, AAAI Press, 1982.

[30] H. Tyagi, "Common information and secret key capacity," *IEEE Transactions on Information Theory*, vol. 59, 2013.

[31] H. Tyagi and P. Narayan, "How many queries will resolve common randomness?" *IEEE Trans. Inf. Theory*, vol. 59, no. 9, 2013.

[32] H. Tyagi and S. Watanabe, "Converses for secret key agreement and secure computing," *IEEE Transactions on Information Theory*, vol. 61, 2015.

[33] S. Watanabe, "Information theoretical analysis of multivariate correlation," *IBM Journal of Research and Development*, vol. 4, no. 1, 1960.

[34] A. Wyner, "The common information of two dependent random variables," *IEEE Transactions on Information Theory*, vol. 21, no. 2, 1975.