FACTORIZED MUTUAL INFORMATION MAXIMIZATION

Thomas Merkh and Guido Montúfar

We investigate the sets of joint probability distributions that maximize the average multi-information over a collection of margins. These functionals serve as proxies for maximizing the multi-information of a set of variables or the mutual information of two subsets of variables, at a lower computation and estimation complexity. We describe the maximizers and their relations to the maximizers of the multi-information and the mutual information.

Keywords: multi-information, mutual information, divergence maximization, marginal specification problem, transportation polytope

Classification: 94A17, 62B10

1. INTRODUCTION

The mutual information (MI) is a measure of mutual dependence between two random variables that plays a central role in information theory and machine learning. The multi-information is a generalization to composite systems with an arbitrary number of random variables. The problem of maximizing the multi-information was proposed and studied in [2], motivated by infomax principles. It was shown, in the setting of finite valued random variables, that the optimizers of the unconstrained maximization problem are attained within low dimensional exponential families of joint probability distributions. A characterization of the optimizers was obtained in [5]. Maximizing the multi-information can be regarded as a special instance of the more general problem of maximizing a divergence from an exponential family. Indeed, the multi-information of a joint probability distribution is equal to its Kullback-Leibler divergence from an independence model. The problem of maximizing the divergence from an exponential family has been advanced by Matúš [25, 26], Matúš and Ay [27], Matúš and Rauh [28], Rauh [36]. The divergence maximization problem can also be asked for more general classes of probability models, such as probabilistic graphical models with hidden variables and stochastic neural networks. Works in this direction include [34, 31] and the short overview [35].

Numerous machine learning applications involve optimizing the MI over a restricted set of joint probability distributions. Examples include the information bottleneck methods [48, 45, 1], the computation of positive information decompositions [10], the implementation of information theoretic regularizers in robotics and reinforcement learning [50, 33], the analysis of deep networks [18], and unsupervised representation learning in deep learning [20]. A parametrized set of distributions may or may not contain the unconstrained maximizers of a given functional. In some cases, the distributions are constrained by structures in addition to the chosen parametrization. For instance, in Markov decision processes the joint distributions between time consecutive states are governed not only by a parametrized policy model but also by the state transition mechanism; see, e.g., [32]. In order to optimize the mutual information in such cases, we usually need to resort to iterative parameter optimization techniques. Computing the parameter updates can quickly become intractable, even for a moderate number of variables, since the number of possible joint states grows exponentially in the number of variables. The estimation and computation of the mutual information is often a bottleneck in the implementation of algorithms that are based on it. For reference, the estimation of the MI from observations has been studied in [39, 24, 41, 19], and recently also using neural networks in [9].

We are interested in proxies that are easier to estimate from samples and easier to compute than a given functional, and which might provide a signal for optimizing it, possibly having the same optimizers, or a structured subset of the optimizers.

We consider measures defined as averages of the multi-information over subsets of all random variables in a composite system. This type of local quantities and averages have appeared in different contexts and under various names, including intricacies [13], mean information paths [7], and local stochastic interactions [3]. Based on previous investigations of the maximizers of multi-information [5, 2], we can expect that the maximizers of local multi-information averages will satisfy several properties. First, they may be contained within the closure of a low dimensional exponential family. Second, like many other instances of divergence maximizing distributions, they may exhibit reduced support. Last, if the measures are symmetric under the exchange of indices, the maximizing distributions are expected to display the same degree of symmetry. We characterize the sets of maximizers of various factorized measures and describe their relation to the maximizers of the multi-information and the maximizers of the mutual information of two subsets of variables.

As a motivating application, we have in mind the estimation and maximization of the mutual information between time consecutive sensor readings of a reinforcement learning agent acting in an environment. The mutual information is the 1 time step version of a notion called predictive information, which has been considered in the literature [11, 40, 4, 51, 15]. In [33], a factorized mutual information (here called SFMI) was used in place of the mutual information as an intrinsic reward signal to encourage more robust walking behaviors for a multi-legged robot. A similar approach was also used previously in [50] for a

chain of robots. Maximizing the mutual information between time consecutive measurements of a given variable should encourage behaviors that are diverse and, at the same time, predictable. This intuition comes from writing the mutual information between X, Y as $I(X, Y) = H(X) - H(X|Y)$, where H is the entropy. In this context, the mutual information serves as a regularizer that is added in order to ease the task-objective optimization problem and also to assign preferences to specific types of solutions. Some references to such regularizes, also known as intrinsic motivation, include [6, 30, 43, 14, 23, 11, 12].

This article is organized as follows. In Section 2 we review definitions and in Section 3 relevant existing results around the multi-information. In Section 4 we define notions of factorized multi and mutual information. The FMI of a set of random variables is defined as an average multi-information of subsets of variables. The SFMI of two random vectors is defined as an average mutual information of non-overlapping pairs with one variable from each vector. In Section 5 we study the maximizers of the FMI and characterize the cases when they coincide with the maximizers of the multi-information. In Section 6 we characterize the maximizers of the SFMI, which build a union of transportation polytopes that contain all maximizers of the multi-information and some maximizers of the mutual information. Section 7 describes partitions of sets of strings that we use for the characterization of SFMI maximizers. In Section 8 we offer a summary and discussion of our results.

2. MULTI-INFORMATION AND MUTUAL INFORMATION

We consider discrete probability distributions supported on a finite set $\mathcal{X}$. The set of all such distributions is a $(|\mathcal{X}| - 1)$-simplex denoted $\Delta_\mathcal{X}$. Let $D(p\|q)$ be the Kullback-Leibler divergence between the probability distributions p and q, which is defined as

$$D(p\|q) := \sum_{x \in \mathcal{X}} p(x) \log \left(\frac{p(x)}{q(x)} \right). \tag{1}$$

If $\mathcal{C}$ is a family of probability distributions, then let $D(p\|\mathcal{C})$ denote the infimum divergence between p and any distribution in $\mathcal{C}$,

$$D(p\|\mathcal{C}) := \inf_{q \in \mathcal{C}} D(p\|q). \tag{2}$$

We are concerned with joint probability distributions of n variables, so that $x = (x_1, \ldots, x_n)$ and $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$. We denote $\mathcal{F}$ the set of fully factorizable joint distributions, i.e., those $p \in \Delta_\mathcal{X}$ which can be written as

$$p(X_1, \ldots, X_n) = p_1(X_1) p_2(X_2) \cdots p_n(X_n), \quad p_i \in \Delta_{\mathcal{X}_i}, \ i = 1, \ldots, n. \tag{3}$$

The set $\mathcal{F}$ is also referred to as the independence model of $X_1, \ldots, X_n$. It is a $\sum_{i=1}^n (|\mathcal{X}_i| - 1)$-dimensional manifold in $\Delta_\mathcal{X}$. The multi-information of n random variables with joint distribution p is then defined as

$$I_p(X_1, X_2, \ldots, X_n) := D(p(X_1, X_2, \ldots, X_n)\|\mathcal{F}). \tag{4}$$

It has a natural interpretation as the distance of p from being independent. Following standard notation we drop the subscript p, and write simply $I(X_1, \ldots, X_n)$. The minimum Kullback-Leibler divergence from a joint distribution p to the set of factorizable distributions is attained uniquely by the product of its marginals,

$$\operatorname{arginf}_{q \in \mathcal{F}} D(p\|q) = p(X_1) \cdots p(X_n), \tag{5}$$

where $p(X_i = x_i) = \sum_{x_j: \, j \neq i} p(x_1, \ldots, x_n)$, $x_i \in \mathcal{X}_i$, for $i = 1, \ldots, n$. The product of marginals is the maximum likelihood estimate of a joint distribution as a product distribution.

The multi-information can be written equivalently in terms of entropies, as

$$I(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i) - H(X_1, \ldots, X_n). \tag{6}$$

Here $H(X) = -\sum_x p(x) \log p(x)$ denotes the entropy of X. If $|\mathcal{X}| = N$, then direct computation shows that the entropy is bounded above by $\log(N)$. Therefore, the maximum value of multi-information of n N-ary variables is bounded by $n \log(N)$. However, as noted in [5, Lemma 4.1], this bound is never attained, with a sharp bound being

$$I(X_1, X_2, \ldots, X_n) \leq (n-1) \log(N). \tag{7}$$

The set of all distributions which maximize multi-information for n variables will be denoted by $\mathcal{M}_n$, or $\mathcal{M}$ when n is understood. We note the chain rule of entropy,

$$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_1, \ldots, X_{i-1}). \tag{8}$$

From this, we see that maximizing the multi-information (6) can be interpreted as maximizing the marginal entropies while at the same time minimizing the conditional entropies. In turn, the maximizers of the multi-information are joint distributions where each individual variable is diverse (high entropy), but predictable from the values of the other variables (low conditional entropy).

The multi-information of two variables is called mutual information. We will consider the mutual information of two random vectors $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$, which is given by

$$MI(\mathbf{X}, \mathbf{Y}) := D(p(\mathbf{X}, \mathbf{Y}) \| p(\mathbf{X}) p(\mathbf{Y})), \tag{9}$$

where $p(\mathbf{X} = \mathbf{x}) = \sum_{\mathbf{y}} p(\mathbf{x}, \mathbf{y})$ and $p(\mathbf{Y} = \mathbf{y}) = \sum_{\mathbf{x}} p(\mathbf{x}, \mathbf{y})$ are the marginal distributions. Note that here the divergence is measured to the set of distributions that factorize as a product $p(\mathbf{X}) p(\mathbf{Y})$, which does not need to factorize further over the components of each random vector.

3. MAXIMIZERS OF THE MULTI-INFORMATION

In this section we collect relevant existing results on the multi-information that we will use in the following sections of this article.

Proposition 1. (Maximizers of the multi-information) [5, Corollary 4.10] Consider N-ary random variables $X_1, \ldots, X_n$. Then

$$\max_{p \in \Delta_{\mathcal{X}}} I_p(X_1, \ldots, X_n) = (n-1) \log(N),$$

and the maximizers are the uniform distributions on N-ary codes of length n, minimum Hamming distance n, and cardinality N. The number of such codes is $(N!)^{n-1}$.

Here a code refers simply to a subset of strings. The work [5] also characterizes the maximizers of the multi-information for variables with different state spaces, which is more technical. We will focus on the case where all variables are N-ary, i.e. $|\mathcal{X}_i| = N$ for all i. Such a collection of random variables are called *homogeneous*. We make a few remarks on the inhomogeneous setting in Appendix D.

Example 2. In the case of binary variables, $N = 2$, Proposition 1 states that the set of maximizers of the multi-information consists of all uniform distributions over pairs of binary vectors of Hamming distance n. Letting $\pi_i : \{0,1\} \to \{0,1\}$ be one-to-one maps for $i = 2, 3, \ldots, n$, the maximizers of the multi-information take the form

$$p = \frac{1}{2} \left(\delta_{(0\pi_2(0)\pi_3(0)\cdots\pi_n(0))} + \delta_{(1\pi_2(1)\pi_3(1)\cdots\pi_n(1))} \right),$$

where δ_x denotes the point measure supported on string x.

The maximizers of the mutual information are obtained from applying Proposition 1 to the special case of two random variables.

Corollary 3. For two N-ary variables X and Y, the maximum mutual information is $\max_{p \in \Delta_{\mathcal{X}}} MI_p(X, Y) = \log(N)$, and the maximizers are the uniform distributions on N-ary codes of length 2, minimum Hamming distance 2, and cardinality N. The number of such codes is $N!$.

We can apply Corollary 3 not only to pairs of variables, but also to pairs of random vectors, since random vectors can be regarded as random variables with states over a product space.

Corollary 4. For two N-ary random vectors $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$, the maximizers of $MI(\mathbf{X}, \mathbf{Y})$ are the uniform distributions over N-ary codes of length $2n$, minimum distance 2 when viewed as N^n-ary codes of length 2, and cardinality N^n. The number of such codes is $N^n!$.

Example 5. Consider four binary random variables (X_1, X_2, Y_1, Y_2); that is, $n = 2$ pairs of $N = 2$ valued variables. The $N!^{2n-1} = 8$ maximizers of the

multi-information of (X_1, X_2, Y_1, Y_2) are

$$\frac{1}{2}(\delta_{0000} + \delta_{1111}), \quad \frac{1}{2}(\delta_{0101} + \delta_{1010}),$$
$$\frac{1}{2}(\delta_{0001} + \delta_{1110}), \quad \frac{1}{2}(\delta_{0100} + \delta_{1011}),$$
$$\frac{1}{2}(\delta_{0010} + \delta_{1101}), \quad \frac{1}{2}(\delta_{0111} + \delta_{1000}),$$
$$\frac{1}{2}(\delta_{0011} + \delta_{1100}), \quad \frac{1}{2}(\delta_{0110} + \delta_{1001}).$$

The $N^n! = 24$ maximizers of the mutual information of $\mathbf{X} = (X_1, X_2)$ and $\mathbf{Y} = (Y_1, Y_2)$ are

$$\frac{1}{4}(\delta_{0010} + \delta_{0100} + \delta_{1001} + \delta_{1111}), \quad \frac{1}{4}(\delta_{0000} + \delta_{0110} + \delta_{1001} + \delta_{1111}),^{**}$$
$$\frac{1}{4}(\delta_{0001} + \delta_{0100} + \delta_{1010} + \delta_{1111}), \quad \frac{1}{4}(\delta_{0010} + \delta_{0100} + \delta_{1011} + \delta_{1101}),^{**}$$
$$\frac{1}{4}(\delta_{0001} + \delta_{0110} + \delta_{1000} + \delta_{1111}), \quad \frac{1}{4}(\delta_{0011} + \delta_{0101} + \delta_{1010} + \delta_{1100}),^{**}$$
$$\frac{1}{4}(\delta_{0000} + \delta_{0101} + \delta_{1011} + \delta_{1110}), \quad \frac{1}{4}(\delta_{0001} + \delta_{0111} + \delta_{1000} + \delta_{1110}),^{**}$$
$$\frac{1}{4}(\delta_{0000} + \delta_{0111} + \delta_{1001} + \delta_{1110}), \quad \frac{1}{4}(\delta_{0000} + \delta_{0101} + \delta_{1010} + \delta_{1111}),^{*}$$
$$\frac{1}{4}(\delta_{0011} + \delta_{0110} + \delta_{1000} + \delta_{1101}), \quad \frac{1}{4}(\delta_{0010} + \delta_{0111} + \delta_{1000} + \delta_{1101}),^{*}$$
$$\frac{1}{4}(\delta_{0000} + \delta_{0110} + \delta_{1011} + \delta_{1101}), \quad \frac{1}{4}(\delta_{0011} + \delta_{0110} + \delta_{1001} + \delta_{1100}),^{*}$$
$$\frac{1}{4}(\delta_{0000} + \delta_{0111} + \delta_{1010} + \delta_{1101}), \quad \frac{1}{4}(\delta_{0001} + \delta_{0100} + \delta_{1011} + \delta_{1110}),^{*}$$
$$\frac{1}{4}(\delta_{0011} + \delta_{0101} + \delta_{1000} + \delta_{1110}), \quad \frac{1}{4}(\delta_{0010} + \delta_{0101} + \delta_{1000} + \delta_{1111}),$$
$$\frac{1}{4}(\delta_{0001} + \delta_{0110} + \delta_{1011} + \delta_{1100}), \quad \frac{1}{4}(\delta_{0010} + \delta_{0101} + \delta_{1011} + \delta_{1100}),$$
$$\frac{1}{4}(\delta_{0001} + \delta_{0111} + \delta_{1010} + \delta_{1100}), \quad \frac{1}{4}(\delta_{0011} + \delta_{0100} + \delta_{1010} + \delta_{1101}),$$
$$\frac{1}{4}(\delta_{0010} + \delta_{0111} + \delta_{1001} + \delta_{1100}), \quad \frac{1}{4}(\delta_{0011} + \delta_{0100} + \delta_{1001} + \delta_{1110}).$$

Some maximizers of $MI(\mathbf{X}, \mathbf{Y})$ can be expressed as convex combinations of maximizers of $I(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$. In the current example, those marked by $*$ or $**$ can. The four distributions marked with $*$ have marginals $p(X_1, Y_1)$ and $p(X_2, Y_2)$ of maximum mutual information. The four distributions marked with $**$ have margins $p(X_1, Y_2)$ and $p(X_2, Y_1)$ of maximum mutual information. We will discuss them further in Example 19.

4. FACTORIZED MEASURES OF MULTI-INFORMATION

We propose alternatives to the multi-information based on averages over subsets of random variables. After presenting the definitions in this section, we investigate the maximizers in the next Sections 5 and 6.

Definition 6. For a family Λ of subsets of $\{1, \ldots, n\}$, the Λ-*factorized multi-information* of $(X_1, \ldots, X_n)$ is defined as

$$I_\Lambda(X_1, \ldots, X_n) := \frac{1}{|\Lambda|} \sum_{\lambda \in \Lambda} I((X_i)_{i \in \lambda}). \tag{10}$$

Here $I((X_i)_{i \in \lambda})$ is the multi-information of the marginal distribution of variables X_i, $i \in \lambda$. In particular, $I_{\{1,\ldots,n\}} \equiv I$ and $I_{\{i\}} \equiv 0$.
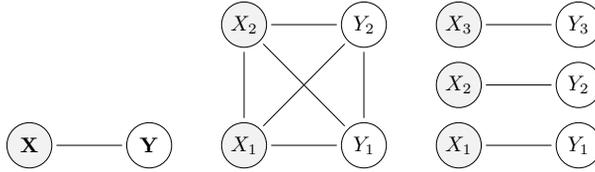
Fig. 1. Graphical illustration of MI, FMI, SFMI. Graphically, the I_Λ with sets $\lambda \in \Lambda$ of cardinality $|\lambda| > 2$ includes hyperedges, but the maximizing set can be described equivalently in terms of the edges contained in elements of Λ.

We will focus on two special cases. The first is the average mutual information between all pairs of variables. We call this the average mutual information, or the pairwise factorized multi-information (FMI).

Definition 7. (Factorized multi-information, FMI) For n random variables $X_1, \ldots, X_n$, the FMI is

$$FMI(X_1, \ldots, X_n) := \frac{1}{\binom{n}{2}} \sum_{i \neq j} MI(X_i, X_j). \tag{11}$$

The second case that we consider is the average mutual information between pairs of variables with one variable from each of two random vectors. We call this the separated average mutual information, or the separated factorized mutual information (SFMI).

Definition 8. (Separated factorized mutual information, SFMI) For two random vectors $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$, the SFMI is

$$SFMI(\mathbf{X}, \mathbf{Y}) := \frac{1}{n} \sum_{i=1}^{n} MI(X_i, Y_i). \tag{12}$$

In the reinforcement learning application mentioned in the introduction, (X_i, Y_i) corresponds to time consecutive readings of the ith sensor of the agent. The average over all sensors, $i = 1, \ldots, n$, is used as a proxy for $MI(\mathbf{X}, \mathbf{Y})$. We will investigate the properties of this choice.

An illustration of the proposed factorized measures is shown in Figure 1. In this figure, edges correspond to the pair marginals whose mutual-information is being added in the factorized measure. One natural question which arises is: how do the distributions with large multi-information compare to those with large factorized mutual information, under the various choices of margins? Since the factorized measures average interdependence between subsets of random variables, they can be less descriptive than measuring the joint interdependence of all of the variables. In the case where the joint probability distribution can be written as the product of pairwise interaction terms, the multi-information

naturally decomposes as the sum over mutual information terms [17]. When higher order interactions exist, the average of pairwise mutual information can serve as an approximation to the multi-information. In relation to this, we note that the exponential family with sufficient statistics computing Λ margins is, by definition, able to attain all feasible values of these margins, and in particular it is able to express all possible values of I_Λ.

The trade-off between computing averages of mutual information between pairs of variables and multi-information or mutual information between groups of variables is that the former can be more easily estimated, since it involves only pair margins and not the full joint distribution. The computation of the pairwise factorized multi-information scales with $n^2 \cdot N^2$, where $N = \max_i |\mathcal{X}_i|$, whereas the computation of multi-information scales with N^n.

5. MAXIMIZERS OF THE FACTORIZED MULTI-INFORMATION

The next lemma shows that every marginal of a joint distribution with maximum multi-information also has maximum multi-information. A related statement, for pair margins, appeared previously in [5, Remark 4.5 (c)].

Lemma 9. Consider N-ary variables $X_1, \ldots, X_n$. If a distribution maximizes the multi-information of $X_1, \ldots, X_n$, then for any subset of indices $\{j_1, \ldots, j_k\} \subseteq \{1, \ldots, n\}$ the corresponding marginal maximizes the multi-information of $X_{j_1}, \ldots, X_{j_k}$.

P r o o f . Consider a joint distribution $p(X_1, \ldots, X_n)$ which attains the maximum of the multi-information. By Proposition 1, this is a uniform distribution over an N-ary code of length n, minimum Hamming distance n, and cardinality N. In this case, the marginal $p(X_n)$ is the uniform distribution over $\mathcal{X}_n$. Moreover, the conditional distributions $p(X_1, \ldots, X_{n-1}|X_n = x_n)$, $x_n \in \mathcal{X}_n$, are N point distributions supported on strings of Hamming distance $n - 1$. The marginal distribution of $(X_1, \ldots, X_{n-1})$ can be written as

$$p(x_1, \ldots, x_{n-1}) = \sum_{x_n} p(x_1, \ldots, x_{n-1}, x_n)$$

$$= \sum_{x_n} p(x_1, \ldots, x_{n-1}|x_n)p(x_n)$$

$$= \frac{1}{|\mathcal{X}_n|} \sum_{x_n} p(x_1, \ldots, x_{n-1}|x_n),$$

which is a uniform distribution over N strings of length $n - 1$ and minimum Hamming distance $n - 1$, and hence has maximum multi-information. $\quad\square$

Lemma 9 can be used to show the following theorem, which characterizes the choices of Λ for which the maximizers of I_Λ coincide with those of I.

Definition 10. A family Λ of subsets of $\{1,\ldots,n\}$ is a *connected covering* if it contains a list of sets λ_t, $t = 1,\ldots,T$, with $\lambda_t \cap (\cup_{s=1}^{t-1}\lambda_s) \neq \emptyset$ for $t = 2,\ldots,T$, and $\cup_t \lambda_t = \{1,\ldots,n\}$.

Theorem 11. (Maximizers of FMI vs I) Consider N-ary variables $X_1,\ldots,X_n$ with $N > 1$, and a family Λ of subsets of $\{1,\ldots,n\}$. The set of distributions that maximize $I_\Lambda(X_1,\ldots,X_n) = \frac{1}{|\Lambda|}\sum_{\lambda\in\Lambda} I((X_i)_{i\in\lambda})$ coincides with the set of distributions that maximize $I(X_1,\ldots,X_n)$ if Λ is a connected covering of $\{1,\ldots,n\}$, and is a strict superset otherwise.

P r o o f. [Proof of Theorem 11] Denote $\mathcal{M}$ and $\mathcal{N}$ the sets of joint distributions that maximize the multi-information $I(X_1,\ldots,X_n)$ and the factorized multi-information $I_\Lambda(X_1,\ldots,X_n)$, respectively. We prove first that $\mathcal{M} \subseteq \mathcal{N}$. Since the I_Λ is a sum of bounded non-negative terms, if each term is independently maximized, then I_Λ will be maximized. Consider $p(X_1,\ldots,X_n) \in \mathcal{M}$. Lemma 9 shows that each marginal $p(X_\lambda)$, $\lambda \in \Lambda$, is a multi-information maximizer. Therefore, $p \in \mathcal{N}$ and $\mathcal{M} \subseteq \mathcal{N}$.

We proceed with the *if* statement. Assume that Λ is a connected covering of $\{1,\ldots,n\}$. We need to show that $\mathcal{N} \subseteq \mathcal{M}$. Suppose that $p \in \mathcal{N}$. Since Λ is a connected covering, there is a list of sets λ_t, $t = 1,\ldots,T$, with $\lambda_t \cap (\cup_s^{t-1}\lambda_s) \neq \emptyset$ for $t = 2,\ldots,T$, and $\cup_t\lambda_t = \{1,\ldots,n\}$. By the above arguments, each λ_t-marginal has maximum multi-information. This means that each marginal $p(X_{\lambda_t})$ is one of $N!^{|\lambda_t|-1}$ possible distributions (this is the number of N-ary codes of cardinality N, minimum distance $|\lambda_t|$, and length $|\lambda_t|$). Each marginal fixes the joint states of $|\lambda_t|$ variables to one of N possible choices. Marginals with overlapping variables need to be compatible on the overlap. For a given choice of the λ_s-marginals, $s = 1,\ldots,t-1$, there is exactly one compatible choice of the λ_t-marginal. Since each variable appears in at least one marginal and there is a sequence of overlapping marginals that cover all variables, the joint states of all variables are fixed to one of N possible strings which have distance n from each other. Therefore, $p \in \mathcal{M}$ and $\mathcal{N} \subseteq \mathcal{M}$.

It remains to show the *only if* statement. For this, we need to show that $\mathcal{M} \neq \mathcal{N}$ if Λ is not a connected covering. If Λ is not a connected covering, then one of the following is true: 1) there is an $i \in \{1,\ldots,n\}$ which is not contained in any $\lambda \in \Lambda$, or 2) Λ consists of two (or more) subfamilies of sets that are mutually disjoint.

In the first case, I_Λ is independent of the marginal distribution of X_i. On the other hand, for any maximizer of the multi-information the marginal of X_i is a uniform distribution. Hence, provided that $N > 1$ (which we always assume), $\mathcal{M} \neq \mathcal{N}$.

In the second case, let λ_1 and λ_2 be two disjoint subsets of $\{1,\ldots,n\}$ which contain all sets from the family Λ. If λ_1 or λ_2 only contains one point $i \in \{1,\ldots,n\}$, then I_Λ is independent of the marginal distribution of X_i (since $I(X_i) \equiv 0$), and we are in the same situation discussed in 1). Assume therefore

that λ_1 and λ_2 each contain at least two points. By definition, I_Λ only depends on the values of the marginal distributions p_1 of $(X_i)_{i \in \lambda_1}$ and p_2 of $(X_i)_{i \in \lambda_2}$. Each choice of the two marginals is realized by a set of joint distributions that is known as a 2-way transportation polytope. If the margins take positive value over m_1 and m_2 entries, respectively, then the dimension of this polytope is $(m_1 - 1)(m_2 - 1)$. The dimension of such 2-way transportation polytopes is known in the literature; see [49]. We will encounter a generalization later in this paper. For any $p \in \mathcal{M}$, Lemma 9 shows that the margins p_1 and p_2 are maximizers of the multi-information, and so, by Proposition 1, both have positive uniform value over N entries. Since $p \in \mathcal{N}$, $\mathcal{N}$ also contains a polytope of dimension $(N - 1)(N - 1) > 0$. On the other hand, $\mathcal{M}$ is a finite discrete set and has dimension zero. Therefore, $\mathcal{M} \neq \mathcal{N}$. This completes the proof. $\qquad \square$

Theorem 11 states that maximizing the multi-information of sufficiently many margins imposes enough restrictions on the joint distributions to ensure that they also have maximum multi-information. Moreover, it characterizes the minimal families of marginals that are sufficient.

In particular, since the set of pairs in $\{1, \ldots, n\}$ defines an overlapping covering, we have the following corollary.

Corollary 12. (Maximizers of the FMI) Consider N-ary variables $X_1, \ldots, X_n$ with $N > 1$. The maximizers of the FMI (11) agree with the maximizers of the multi-information (4).

In fact, Theorem 11 shows that specifying $n - 1$ pair margins suffices. In particular, the exponential family with sufficient statistics computing $n - 1$ pair margins contains in its closure all maximizers of the multi-information. This relates to the result from [5, Theorem 5.1], which shows that the closure of an exponential family of $n - 1$ pure pair interactions contains all maximizers of the multi-information.

Example 13. Consider the case of three binary variables, $\Lambda = \{\{1, 2\}, \{2, 3\}\}$, and the following two mutual information maximizing marginals:

$$p(X_1, X_2) = \frac{1}{2}(\delta_{00} + \delta_{11}), \tag{13}$$

$$p(X_2, X_3) = \frac{1}{2}(\delta_{00} + \delta_{11}). \tag{14}$$

These marginals fix the joint states of (X_1, X_2, X_3) in the following way. Write the joint distribution as a convex combination of point distributions,

$$p(X_1, X_2, X_3) = p_{000}\delta_{000} + p_{001}\delta_{001} + \cdots + p_{111}\delta_{111}. \tag{15}$$

Eqn. (13) imposes following linear constraints on the joint distribution:

$$\begin{aligned} p_{000} + p_{001} &= \tfrac{1}{2}, \quad p_{110} + p_{111} = \tfrac{1}{2}, \\ p_{010} + p_{011} &= 0, \quad p_{100} + p_{101} = 0. \end{aligned} \tag{16}$$

Eqn. (14) imposes following linear constraints:

$$p_{000} + p_{010} = \tfrac{1}{2}, \quad p_{101} + p_{111} = \tfrac{1}{2},$$
$$p_{001} + p_{011} = 0, \quad p_{100} + p_{110} = 0. \tag{17}$$

Since probabilities are non-negative, the last two constraints in each case imply $p_{010} = p_{011} = p_{100} = p_{101} = 0$ and $p_{001} = p_{011} = p_{100} = p_{110} = 0$. Therefore, it must be that $p_{000} = p_{111} = \tfrac{1}{2}$, showing that specifying two overlapping marginal distributions to be maximizers is sufficient to conclude that the joint distribution is also a multi-information maximizer.

When Λ is not a connected covering, the set of maximizers of I_Λ has dimension larger than zero. In this case, the exponential family with sufficient statistics computing Λ margins still contains in its closure one joint distribution that expresses each choice of the margins, and in particular it contains maximizers of I_Λ, but in general it does not contain all of them. We will discuss the setting of not connected coverings in more detail in the next Section 6. We conclude this section with the following remark.

Remark 14. We observe that, while there always exists a joint distribution each of whose marginals maximizes the multi-information, in general there are compatibility requirements among the marginals. For example, choosing $p(X_1, X_2) = \tfrac{1}{2}(\delta_{00} + \delta_{11})$ and $p(X_2, X_3) = \tfrac{1}{2}(\delta_{01} + \delta_{10})$ means that $p(X_1, X_3)$ must be exactly $\tfrac{1}{2}(\delta_{01} + \delta_{10})$.

Verifying that a certain specification of margins is indeed feasible, meaning that there exists a joint distribution with those margins, is known as the margin specification problem. Given a family Λ of subsets of $\{1, \ldots, n\}$, the set of margins $(p_\lambda)_{\lambda \in \Lambda}$ that are feasible, is known in the literature as the Λ-marginal polytope, which can be obtained as the convex hull of the vectors

$$F(x) = (F_{\lambda, \tilde{x}_\lambda}(x))_{\lambda \in \Lambda, \tilde{x}_\lambda \in \mathcal{X}_\lambda}, \quad x \in \mathcal{X}, \tag{18}$$

where for each λ and $\tilde{x}_\lambda$ there is a coordinate taking value one if $x_\lambda = \tilde{x}_\lambda$ and value zero otherwise. These vectors form the columns of a matrix of sufficient statistics, computing the Λ-margins, for the exponential family known as the Λ-interaction model.

For a given vector $(p_\lambda)_\lambda$ of λ-margins, $\lambda \in \Lambda$, the set of joint distributions that express these margins is known as a linear Λ-family at p (where p is a compatible joint distribution) [37], or also as the n-way transportation polytope $T(p_\lambda, \lambda \in \Lambda)$ [49]. This is obtained by intersecting the simplex of joint probability distributions with the orthogonal complement of the row span of F at p. We will encounter this type of structures in the next section. More details are provided in Appendix B.

6. MAXIMIZERS OF THE FACTORIZED MUTUAL INFORMATION

We have seen that the maximizers of any factorized multi-information measure with connected covering, like the ones shown in Figure 2, coincide with the
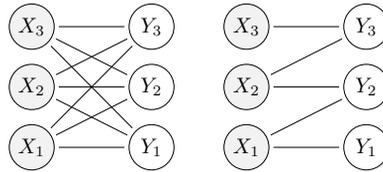
Fig. 2. At first sight, the factorized measures with these graphs might appear to be closely related to the mutual information $MI(\mathbf{X}, \mathbf{Y})$. However, by Theorem 11 the maximizers coincide with the maximizers of the multi-information. As illustrated by Example 5, they differ significantly from the maximizers of $MI(\mathbf{X}, \mathbf{Y})$.

maximizers of the multi-information. These distributions are very different from the maximizers of the mutual information of two random vectors of length $n \geq 2$. If we want to maximize $MI(\mathbf{X}, \mathbf{Y})$ by maximizing a factorized measure I_Λ, then we should choose a Λ that is not a connected covering.

The SFMI (12) only considers the mutual information of pairs (X_i, Y_i), $i = 1, \ldots, n$, which is not a connected covering. Theorem 11 shows that the SFMI maximizers are a strict superset of the multi-information maximizers:

Corollary 15. Consider N-ary variables $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ with $N > 1$. Let $n \geq 2$. The maximizers of $I(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ are a strict subset of the maximizers of $SFMI(\mathbf{X}, \mathbf{Y})$.

In the following we present a characterization of the set of maximizers of SFMI for homogeneous variables. We start with the next theorem, which describes the number of connected components, dimension, support sets.

Theorem 16. (Maximizers of SFMI vs I) Consider N-ary variables $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ with $N > 1$. The set of maximizers of the SFMI is a superset of the maximizers of the multi-information. It consists of the joint distributions with marginals $p(X_i, Y_i)$ maximizing $MI(X_i, Y_i)$, for $i = 1, \ldots, n$. This is the disjoint union of $N!^n$ transportation polytopes of dimension $N^n - 1 - n(N-1)$ and whose interiors consist of distributions with support of cardinality N^n.

The SFMI maximizing polytopes are the sets of joint probability distributions that are compatible with the specification of n pair margins as distributions with maximum mutual information. We provide some comments on multi-way transportation polytopes in Appendix B. Since the pair margins have disjoint variables, the polytopes can be regarded as orthogonal linear families of an independence model of n N^2-ary variables. But since the specified margins are uniform distributions with support of cardinality N, the polytopes have the structure of orthogonal linear families of an independence model of n N-ary variables at the uniform distribution. In this perspective, the elementary events

of the ith variable are pairs (x_i, y_i) forming an N-ary code of length 2, minimum Hamming distance 2, and cardinality N.

P r o o f. [Proof of Theorem 16] If we consider a subset λ of variables, fixing a marginal distribution $p(X_\lambda)$ with support $\tilde{\mathcal{X}}_\lambda \subseteq \mathcal{X}_\lambda$ defines a linear set in the joint probability simplex of dimension $|\tilde{\mathcal{X}}_\lambda|(|\mathcal{X}_{\lambda^c}| - 1)$. This is namely the set of distributions of the form $p(X_\lambda)q(X_{\lambda^c}|X_\lambda)$. The dimension count comes from the fact that for each $x_\lambda \in \tilde{\mathcal{X}}_\lambda$, the conditional distribution $q(X_{\lambda^c}|X_\lambda = x_\lambda)$ can be chosen independently as an arbitrary point in a $(|\mathcal{X}_{\lambda^c}| - 1)$ dimensional simplex.

For variables $(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$, the set of maximizers of the MI of a given pair (X_i, Y_i) corresponds to a set $\mathcal{S}_i = \cup_{j=1}^{N!}\mathcal{S}_{i,j}$ in the joint probability simplex which is the disjoint union of $N!$ linear spaces of dimension $N(N^{2n-2} - 1)$. Indeed, there are $N!$ maximizers of the marginal mutual information. Each of them corresponds to fixing a marginal distribution $p(X_i, Y_i)$ with support of cardinality N, which defines a set $\mathcal{S}_{i,j}$ in the joint probability simplex. In terms of the above discussion, $\mathcal{X}_\lambda = \mathcal{X}_i \times \mathcal{Y}_i$ and $|\tilde{\mathcal{X}}_\lambda| = N$. The set $\mathcal{S}$ of maximizers of the SFMI is the intersection of n such sets, $\mathcal{S} = \cap_{i=1}^n \mathcal{S}_i$, one corresponding to each pair marginal. Each marginal implies that $(N^2 - N)N^{n-2}$ entries of the joint distribution vanish. Overall, only N^n entries of the joint distribution can have nonzero value. Each marginal implies N linear constraints of the form $\frac{1}{N} = p(X_i = x_i, Y_i = y_i) = \sum_{x_{\hat{i}}, y_{\hat{i}}} p(x_1, \ldots, x_n, y_1, \ldots, y_n)$. Each marginal contributes $(N - 1)$ independent constraints, and have one shared constraint, namely that the sum of entries is 1. The number of independent constraints can be derived from the fact that $\mathbf{1}_{\lambda, x_\lambda}$, $\lambda \subseteq \Lambda$, $x_\lambda \in \prod_{i \in \lambda}(\mathcal{X}_i \setminus \{0\})$, including $\mathbf{1}_\emptyset \equiv 1$, is a basis of the set of functions $\sum_{\lambda \in \Lambda} f_\lambda(x)$; see [21]. We have N^n positive entries satisfying $n(N - 1) + 1$ independent linear equality constraints, which gives the desired dimension $N^n - n(N - 1) - 1$.

In terms of the number of pieces, all pieces defined by one marginal intersect with all pieces defined by any other marginal. Since each pair marginal defines $N!$ pieces, the total number of pieces is $N!^n$.

The fact that the set of SFMI maximizers is a superset of the set of FMI maximizers follows from the fact that it is defined by a subset of the affine constraints that define the FMI maximizers. $\square$

The next question is whether the transportation polytopes maximizing the SFMI contain some of the maximizers of the MI. Transportation polytopes are not very well understood in general. We obtain the following characterization.

Theorem 17. (Maximizers of SFMI vs MI) Consider N-ary variables $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ with $N > 1$. Each of the $N!^n$ transportation polytopes maximizing the SFMI contains $N!^{n-1}$ vertices which are maximizers of the multi-information. These vertices come in $(N - 1)!^{n-1}$ disjoint sets, each set consisting of N^{n-1} affinely independent vertices and having the same centroid, which is a maximizer of $MI(\mathbf{X}, \mathbf{Y})$.

To put this in perspective, recall that $I(X_1, Y_1, \ldots, X_n, Y_n)$ has $N!^{2n-1}$ maximizers, and $MI(\mathbf{X}, \mathbf{Y})$ has $N^n!$ maximizers. Moreover, the dimension of each of the transportation polytopes in question is $N^n - 1 - n(N - 1)$. The total number of vertices of a transportation polytope is not known in general. For the special type of transportation polytopes that we consider here, the theorem gives a lower bound $N!^{n-1}$.

P r o o f . [Proof of Theorem 17] The proof is by construction. Consider one of the $N!^n$ polytopes, $\mathcal{S} = \cap_{i=1}^n \mathcal{S}_{i,j_i}$ for a fixed choice of $j_i \in \{1, \ldots, N!\}$, for $i = 1, \ldots, n$. Let p_{i,j_i} be the corresponding (X_i, Y_i) margin, which is a uniform distribution on N strings of length two and minimum distance two, out of the N^2 possible strings of length two. Denote these N strings by $z_i^k = (x_i^k, y_i^k)$, $k = 1, \ldots, N$. In the following we will write the joint states of all variables as $((x_1, y_1), \ldots, (x_n, y_n))$.

We claim that $\mathcal{S}$ contains $\frac{1}{N} \sum_{k=1}^N \delta_{z_1^{\pi_1(k)} \ldots z_n^{\pi_n(k)}}$, for each choice of an ordering π_i of $\{1, \ldots, N\}$, for each $i = 1, \ldots, n$. Indeed, for each $i = 1, \ldots, n$ the ith margin is $\frac{1}{N} \sum_k \delta_{z_i^k} = p_{i,j_i}$.

Each of these points is a distinct maximizer of the multi-information of $(Z_1, \ldots, Z_n)$. Indeed, each of them is a uniform distribution on N strings of length n and minimum distance n (with each Z_i regarded as a single N valued variable). They are also maximizers of the multi-information of $(X_1, Y_1, \ldots, X_n, Y_n)$, since any distinct z_i^k and $z_i^{k'}$ from our list have, by definition, distance two when regarded as strings (x_i^k, y_i^k) and $(x_i^{k'}, y_i^{k'})$.

Summarizing, $\mathcal{S}$ contains $N!^{n-1}$ distinct maximizers of the multi-information. These points are also vertices of the polytope. Since they have support of cardinality N, they cannot be expressed as a non-trivial combination of maximizers of SFMI. Indeed, N is the minimum cardinality of the support of any maximizer of SFMI (because this is the minimum cardinality of any maximizer of the mutual information of any pair margin). Moreover, any maximizer of SFMI with support of cardinality N must be the uniform distribution over that support (because this is true for any maximizer of the MI of any pair margin).

Next we show that $\mathcal{S}$ contains $(N-1)!^{n-1}$ simplices which are each the convex hull of N^{n-1} distinct vertices of $\mathcal{S}$. Recall that $|\operatorname{supp}(\mathcal{S})| = \max\{|\operatorname{supp}(p)| \colon p \in \mathcal{S}\} = N^n$. Because $\mathcal{S}$ has $N!^{n-1}$ vertices which are multi-information maximizers, it follows that there are $N!^{n-1}$ codes of length $2n$, minimum distance $2n$, and cardinality N. These are the support sets of the vertices. One can find N^{n-1} codes of this form which are disjoint and cover all strings (see Proposition 22). For N^{n-1} disjoint codes, each of cardinality N, the union contains N^n distinct strings and covers $\operatorname{supp}(\mathcal{S})$. One can find $\frac{N!^{n-1}}{N^{n-1}} = (N - 1)!^{n-1}$ such collections of codes (see Proposition 23), which correspond to the different simplices.

Consider the vertices corresponding to a single collection of N^{n-1} vertex support sets. Any distribution written as a convex combination of these vertices is in $\mathcal{S}$. Since the support sets of these vertices are disjoint, it follows that a

strictly convex combination will have support on N^n vectors. Furthermore, the convex hull of these vertices can be seen as a simplex of dimension $N^{n-1} - 1$. At the center of the simplex where each vertex is given equal weight, one obtains the uniform distribution over the N^n vectors of supp$(\mathcal{S})$. Recall from the previous that the vectors in supp$(\mathcal{S})$ are N-ary vectors of length $2n$ with minimum distance 2, and therefore the uniform distribution over N^n such vectors is a maximizer of $MI(\mathbf{X}, \mathbf{Y})$. Since each of the $(N-1)!^{n-1}$ simplices have the same support of supp$(\mathcal{S})$, their centroids all coincide at a maximizer of $MI(\mathbf{X}, \mathbf{Y})$.

□

In particular, we note the following.

Corollary 18. Consider N-ary variables $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ with $N > 1$. The maximizers of the multi-information are vertices of the polytopes of maximizers of the SFMI.

We note that the MI is independent of the ordering of the variables within the vectors $\mathbf{X}$ and $\mathbf{Y}$. The SFMI measure that we have introduced corresponds to I_Λ where Λ is a specific pairing of variables in the two vectors. Each Λ consisting of n pairs that are a perfect matching between $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ defines a different SFMI measure. The situation is illustrated in Figure 3. We expect that for each of the $n!$ possible pairings, the corresponding SFMI measure will have among its maximizers, different subsets of maximizers of the MI that can be written as convex combinations of multi-information maximizers. At this point it remains an open question how the sets of MI maximizers that maximize different SFMI measures compare to each other, and how many MI maximizes total are maximizers of some SFMI measure.

We illustrate the discussion of this section in the next example.

Example 19. Consider $n = 2$ pairs of $N = 2$ valued variables (X_1, X_2, Y_1, Y_2). The maximizers of the SFMI (12) are $N!^n = 4$ polytopes of dimension $N^n - 1 - n(N-1) = 1$. With non-negative α, β such that $\alpha + \beta = 1$, the polytopes are

$$
\begin{aligned}
&\frac{\alpha}{2}(\delta_{0000} + \delta_{1111}) + \frac{\beta}{2}(\delta_{0101} + \delta_{1010}), \\
&\frac{\alpha}{2}(\delta_{0001} + \delta_{1110}) + \frac{\beta}{2}(\delta_{0100} + \delta_{1011}), \\
&\frac{\alpha}{2}(\delta_{0010} + \delta_{1101}) + \frac{\beta}{2}(\delta_{0111} + \delta_{1000}), \\
&\frac{\alpha}{2}(\delta_{0011} + \delta_{1100}) + \frac{\beta}{2}(\delta_{0110} + \delta_{1001}).
\end{aligned}
\tag{19}
$$

A detailed derivation is provided in Appendix C.2. When α, β are strictly positive, these distributions have support of cardinality 4 and do not maximize multi-information. However we see that the vertices, with $\beta = 0$ or $\alpha = 0$, recover all 8 maximizers of multi-information. Additionally, setting $\alpha = \beta = \frac{1}{2}$ recovers 4 out of the 24 maximizers of $MI(\mathbf{X}, \mathbf{Y})$. These are the four points marked with $*$ in Example 5. Notice that the SFMI maximizers with highest entropy are maximizers of $MI(\mathbf{X}, \mathbf{Y})$.
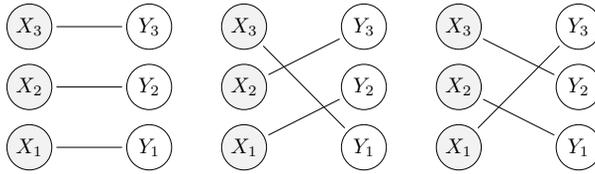
Fig. 3. Illustration of various types of SFMI measures.

Consider now an SFMI with a different assignment of Xs and Ys, $SFMI' = \frac{1}{2}\big(MI(X_1, Y_2) + MI(X_2, Y_1)\big)$. With non-negative α, β such that $\alpha + \beta = 1$, the polytopes of $SFMI'$ maximizers are

$$
\begin{aligned}
&\tfrac{\alpha}{2}(\delta_{0000} + \delta_{1111}) + \tfrac{\beta}{2}(\delta_{1001} + \delta_{0110}), \\
&\tfrac{\alpha}{2}(\delta_{0010} + \delta_{1101}) + \tfrac{\beta}{2}(\delta_{0100} + \delta_{1011}), \\
&\tfrac{\alpha}{2}(\delta_{0001} + \delta_{1110}) + \tfrac{\beta}{2}(\delta_{1000} + \delta_{0111}), \\
&\tfrac{\alpha}{2}(\delta_{0011} + \delta_{1100}) + \tfrac{\beta}{2}(\delta_{0101} + \delta_{1010}).
\end{aligned}
\tag{20}
$$

At the center of these polytopes where $\alpha = \beta$, the maximizers marked $**$ in Example 5 are found, demonstrating that distinct subsets of MI maximizers are contained in each of the $n!$ possible SFMI measures.

We provide more examples in Appendix C.2.

7. CODES AND PARTITIONS

In the proof of Theorem 17 we used certain properties of the set of N-ary strings of length n; specifically, we used that it can be partitioned into codes of minimum distance n and cardinality N, in a number of different ways. Here a code refers simply to a subset of strings. In this section we state and prove these facts.

Proposition 20. There are $N!^{n-1}$ N-ary codes of length n, minimum distance n, and cardinality N.

Proof. An N-ary code of length n, minimum distance n, and cardinality N corresponds to a perfect covering of a rectangular grid of size $N \times n$ by N strings of length n. In order to count the total number, consider the grid, and the possible choices for how to place the first string. The first coordinate is 1, for the second coordinate there are N choices, for the third coordinate there are N choices, etc. For the second string the first coordinate is 2, and there are $N - 1$ possible choices for each of the remaining coordinates. For the Nth string there is only one choice left. So in total we have $N^{n-1} \cdot (N-1)^{n-1} \cdots 1^{n-1} = N!^{n-1}$. $\square$

Proposition 21. The set of all N^2 edges of the full bipartite graph on $N \times N$ can be partitioned into N sets of N edges, each of which forms a perfect matching.

P r o o f . Denote the edges by $(u, v) \in \{0, \ldots, N-1\} \times \{0, \ldots, N-1\}$. Each $t = 1, \ldots, N$ defines a perfect matching $\{(u, u + t \mod (N)) \colon u \in \{0, \ldots, N-1\}\}$. Each of these matchings uses one distinct edge, out of the N edges that connect to the uth left node, for each u. Hence all of these matchings have disjoint edge sets, and they, together, contain all edges from the full bipartite graph. $\square$

Proposition 22. The set of all N^n N-ary strings of length n can be partitioned into N^{n-1} codes of minimum distance n and cardinality N.

P r o o f . We want to partition the set of all N^n strings into N^{n-1} sets, each of which gives a perfect covering of a grid of size $N \times n$. For any $(t_2, \ldots, t_n) \in \{0, \ldots, N-1\}^{n-1}$, define a code as $\{(u, u + t_2 \mod (N), \ldots, u + t_n \mod (N)) \colon u = 0, \ldots, N-1\}$. The proof of Proposition 21 discusses how this construction gives a perfect matching between all nodes in column i and column $i+1$, for all $i = 1, \ldots, n-1$. There are N^{n-1} of these assignments and they all build disjoint sets of strings. They, together, exhaust all N^n strings, by the same reasoning of Proposition 21. $\square$

Proposition 23. There are $N!^{n-1}/N^{n-1}$ ways to partition the set of all N^n N-ary strings of length n in the form described in Proposition 22.

P r o o f . Consider again each N-ary code of length n, minimum distance n, and cardinality N as a perfect covering of an $N \times n$ grid by strings. Consider the partition described in Proposition 22 of the set of N-ary strings into N^{n-1} such codes. Note that this partition is constructed by applying circular shifts, which define a subgroup of the symmetric group. Now we simply consider the cosets defined by the circular shifts. The cardinality of all cosets is equal. Two cosets are either equal or are disjoint. The number of distinct cosets is $N!/N = (N-1)!$. This gives a total of $(N-1)!^{n-1}$ partitions. $\square$

8. DISCUSSION

We formulated factorized versions of the mutual information and studied the sets of joint probability distributions that maximize them. As we observe, characterizing the maximizers of these measures translates to characterizing the sets of joint distributions that are compatible with a given specification of margins.

When a sufficient number of margins is included in the factorized measure, the maximizers correspond precisely to the maximizers of the multi-information. For this to be the case, it is necessary and sufficient that the set of margins builds a connected covering of all variables. In particular, there are families of pair margins that are sufficient, which reflects previous results showing that

the multi-information can be maximized over families of joint distributions that only include pair interactions.

When the considered margins have disjoint sets of variables, the constraints that they impose on the joint distributions no longer specify the joint distributions down to the maximizers of the multi-information. Instead, they specify families of transportation polytopes some of whose vertices are multi-information maximizers. As we showed, certain choices of margins (as in the SFMI) lead to polytopes that contain some of the maximizers of the mutual information between two subsets of variables.

MI vs I The set of maximizers of the multi-information $I(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ is a discrete set of cardinality $(N!)^{2n-1}$. The set of maximizers of the mutual information $MI(\mathbf{X}, \mathbf{Y})$ is a discrete set of cardinality $(N^n)!$. The two sets are equal if $n = 1$, and are disjoint otherwise. These are well known results, collected in Proposition 1 and Corollary 4.

FMI vs I and MI The set of maximizers of $I_\Lambda(X_1, \ldots, X_n)$ is equal to the set of maximizers of $I(X_1, \ldots, X_n)$ if the considered set Λ of margins builds a connected covering of all variables, and it is a strict superset otherwise. See Theorem 11. If Λ is a connected covering of all variables, then the maximizers of $I_\Lambda(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ are disjoint from maximizers of $MI(\mathbf{X}, \mathbf{Y})$, unless $n = 1$ in which case the two sets are equal. If Λ is not a connected covering, the relation is open in general. We studied in detail the special case of SFMI, summarized below.

SFMI vs I and MI SFMI is the special case of I_Λ with Λ being a perfect matching of X and Y variables. This is not a connected covering, except when $n = 1$. The set of SFMI maximizers consists of $N!^n$ disjoint polytopes of dimension $N^n - 1 - n(N - 1)$, each of which has the structure of a central n-way transportation polytope defined by 1-margins of size N. See Theorem 16. Each of the $N!^n$ polytopes maximizing SFMI contains $N!^{n-1}$ maximizers of $I(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ and one maximizer of $MI(\mathbf{X}, \mathbf{Y})$. See Theorem 17.

Factorized multi-information as an intrinsic reward

When the MI is superimposed with another objective function, we may obtain specific types of solutions which maximize both, or simply which tend to have a larger MI. The SFMI maximizers do not include all MI maximizers. An interpretation is that the SFMI maximizers also maximize the MI between specific pairs of variables. So for instance, the MI can be maximized if $H(Y_1|X_1) \neq 0$, if in exchange one has that $H(Y_2|X_1) = 0$, meaning that sensor 2 can be predicted from sensor 1. However, the maximizers of SFMI insist that $H(Y_1|X_1) = 0$. In this sense, the SFMI gives a more structured objective function than the MI.

Possible generalizations

We focused on two general types of margins. In the case of families that are not connected coverings, we considered the SFMI, which consists of a non overlapping covering by edges. Extensions of our analysis to other families of margins are possible. Any connected subfamily will then specify a maximizer of the multi-information over the covered variables. This will give rise to maximizing sets defined in terms of transportation polytopes with margins given by the maximal connected subfamilies. A natural question that arises is how to choose Λ in order to capture, as tightly as possible, a given subset of maximizers of $MI(\mathbf{X}, \mathbf{Y})$. We have focused on systems where all variables have the same number of states. The same problem can be studied in the case of inhomogeneous variables as well. One should be able to obtain a characterization as was done in [5] for the multi-information. The case of continuous variables is another generalization that would be interesting to consider in the future. This has not been discussed in as much detail in the literature as the discrete case. Another interesting analysis would pertain to local maximizers and criticality, in connection to the works [3] and [28].

REFERENCES

[1] A. Alemi, I. Fischer, J. Dillon, and K. Murphy: Deep variational information bottleneck. In: ICLR, 2017.

[2] N. Ay: An information-geometric approach to a theory of pragmatic structuring. Ann. Probab. *30* (2002), 1, 416–436. DOI:10.1214/aop/1020107773

[3] N. Ay: Locality of global stochastic interaction in directed acyclic networks. Neural Comput. *14* (2002), 12, 2959–2980. DOI:10.1162/089976602760805368

[4] N. Ay, N. Bertschinger, R. Der, F. Güttler, and E. Olbrich: Predictive information and explorative behavior of autonomous robots. Europ. Phys. J. B *63* (2008), 3, 329–339. DOI:10.1140/epjb/e2008-00175-0

[5] N. Ay and A. Knauf: Maximizing multi-information. Kybernetika *42* (2006), 5, 517–538.

[6] G. Baldassarre and M. Mirolli: Intrinsically motivated learning systems: an overview. In: Intrinsically motivated learning in natural and artificial systems, Springer 2013, pp. 1–14. DOI:10.1007/978-3-642-32375-1_1

[7] P. Baudot, M. Tapia, D. Bennequin, and J.-M. Goaillard: Topological information data analysis. Entropy *21* (2019), 9, 869. DOI:10.3390/e21090869

[8] R. Bekkerman, M. Sahami, and E. Learned-Miller: Combinatorial markov random fields. In: European Conference on Machine Learning, Springer 2006, pp. 30–41. DOI:10.1007/11871842_8

[9] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm: Mutual information neural estimation. In: Proc. 35th International Conference on Machine Learning (J. Dy and A. Krause, eds.), Vol. 80 of *Proceedings of Machine Learning Research*, pp. 531–540, Stockholm 2018. PMLR.

[10] N. Bertschinger, J. Rauh, E. Olbrich, J. Jost, and N. Ay: Quantifying unique information. Entropy *16* (2014), 4, 2161–2183. DOI:10.3390/e16042161

[11] W. Bialek, I. Nemenman, and N. Tishby: Predictability, complexity, and learning. Neural Comput. *13* (2001), 11, 2409–2463. DOI:10.1162/089976601753195969

[12] Y. Burda, H. Edwards, D. Pathak, A. Storkey, T. Darrell, and A. A. Efros: Large-scale study of curiosity-driven learning. In: ICLR, 2019.

[13] J. Buzzi and L. Zambotti: Approximate maximizers of intricacy functionals. Probab. Theory Related Fields *153* (2012), 3–4, 421–440. DOI:10.1007/s00440-011-0350-y

[14] N. Chentanez, A. G. Barto, and S. P. Singh: Intrinsically motivated reinforcement learning. In: Adv. Neural Inform. Process. Systems 2005, pp. 1281–1288. DOI:10.21236/ada440280

[15] J. P. Crutchfield and D. P. Feldman: Synchronizing to the environment: Information-theoretic constraints on agent learning. Adv. Complex Systems *4* (2001), 02n03, 251–264. DOI:10.1142/s021952590100019x

[16] J. de Loera: Transportation polytopes. `https://www.math.ucdavis.edu/~deloera/TALKS/20yearsafter.pdf`

[17] N. Friedman, O. Mosenzon, N. Slonim, and N. Tishby: Multivariate information bottleneck. In: Proc. Seventeenth conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc., 2001, pp. 152–161.

[18] M. Gabrié, A. Manoel, C. Luneau, j. Barbier, N. Macris, F. Krzakala, and L. Zdeborová: Entropy and mutual information in models of deep neural networks. In: Advances in Neural Information Processing Systems 31 (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), Curran Associates, Inc. 2018, pp. 1821–1831.

[19] S. Gao, G. Ver Steeg, and A. Galstyan: Efficient estimation of mutual information for strongly dependent variables. In: Artificial Intelligence and Statistics 2015, pp. 277–286.

[20] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019.

[21] S. Hosten and S. Sullivant: Gröbner bases and polyhedral geometry of reducible and cyclic models. J. Comb. Theory Ser. A *100* (2002), 2, 277–301. DOI:10.1006/jcta.2002.3301

[22] A. Jakulin and I. Bratko: Quantifying and visualizing attribute interactions: An approach based on entropy. 2003.

[23] A. S. Klyubin, D. Polani, and C. L. Nehaniv: Empowerment: A universal agent-centric measure of control. In: 2005 IEEE Congress on Evolutionary Computation, Vol. 1, IEEE 2005, pp. 128–135.

[24] A. Kraskov, H./ Stögbauer, and P. Grassberger: Estimating mutual information. Phys. Rev. E *69* (2004), 6, 066138. DOI:10.1103/physreve.69.066138

[25] F. Matúš: Maximization of information divergences from binary i.i.d. sequences. In: Proc. IPMU 2004 *2* (2004), pp. 1303–1306.

[26] F. Matúš: Divergence from factorizable distributions and matroid representations by partitions. IEEE Trans. Inf. Theor. *55* (2009), 12, 5375–5381. DOI:10.1109/tit.2009.2032806

[27] F. Matúš and N. Ay: On maximization of the information divergence from an exponential family. In: Proc. 6th Workshop on Uncertainty Processing: Oeconomica 2003, Hejnice 2003, pp. 199–204.

[28] F. Matúš and J. Rauh: Maximization of the information divergence from an exponential family and criticality. In: 2011 IEEE International Symposium on Information Theory Proceedings 2011, pp. 903–907. DOI:10.1109/isit.2011.6034269

[29] W. McGill: Multivariate information transmission. Trans. IRE Profess. Group Inform. Theory *4* (1054), 4, 93–111. DOI:10.1109/tit.1954.1057469

[30] S. Mohamed and D. J. Rezende: Variational information maximisation for intrinsically motivated reinforcement learning. In: Advances in Neural Information Processing Systems 2015, 2125–2133, 2015.

[31] G. Montúfar: Universal approximation depth and errors of narrow belief networks with discrete units. Neural Comput. *26* (2014), 7, 1386–1407. DOI:10.1162/neco_a_00601

[32] G. Montúfar, K. Ghazi-Zahedi, and N. Ay: A theory of cheap control in embodied systems. PLOS Comput. Biology *11* (2015), 9, 1–22. DOI:10.1371/journal.pcbi.1004427

[33] G. Montúfar, K. Ghazi-Zahedi, and N. Ay: Information theoretically aided reinforcement learning for embodied agents. arXiv preprint arXiv:1605.09735, 2016.

[34] G. Montúfar, J. Rauh, and N. Ay: Expressive power and approximation errors of restricted Boltzmann machines. In: Advances in Neural Information Processing Systems 2011, pp. 415–423.

[35] G. Montúfar, J. Rauh, and N. Ay: Maximal information divergence from statistical models defined by neural networks. In: Geometric Science of Information GSI 2013 (F. Nielsen and F. Barbaresco, eds.), Lecture Notes in Computer Science *3085* Springer 2013, pp. 759–766. DOI:10.1007/978-3-642-40020-9_85

[36] J. Rauh: Finding the maximizers of the information divergence from an exponential family. IEEE Trans. Inform. Theory *57* (2011), 6, 3236–3247. DOI:10.1109/tit.2011.2136230

[37] J. Rauh: Finding the Maximizers of the Information Divergence from an Exponential Family. PhD. Thesis, Universität Leipzig 2011.

[38] R. A. A. Ince, S. Panzeri, and S. R. Schultz. *Summary of Information Theoretic Quantities*, pages 1–6. Springer New York, New York, NY, 2013.

[39] M. S. Roulston: Estimating the errors on measured entropy and mutual information. Physica D: Nonlinear Phenomena *125* (1999), 3–4, 285–294. DOI:10.1016/s0167-2789(98)00269-3

[40] J. Schossau, C. Adami, and A. Hintze: Information-theoretic neuro-correlates boost evolution of cognitive systems. Entropy *18* (2015), 1, 6. DOI:10.3390/e18010006

[41] N. Slonim, G. S. Atwal, G. Tkacik, and W. Bialek: Estimating mutual information and multi–information in large networks. arXiv preprint cs/0502017, 2005.

[42] N. Slonim, N. Friedman, and N. Tishby: Multivariate information bottleneck. Neural Comput. *18* (2006), 8, 1739–1789. DOI:10.1162/neco.2006.18.8.1739

[43] S. Still and D. Precup: An information-theoretic approach to curiosity-driven reinforcement learning. Theory Biosci. *131* (2012), 3, 139–148. DOI:10.1007/s12064-011-0142-z

[44] The Sage Developers: SageMath, the Sage Mathematics Software System (Version 8.7), 2019. https://www.sagemath.org.

[45] N. Tishby, F. C. Pereira, and W. Bialek: The information bottleneck method. In: Proc. 37th Annual Allerton Conference on Communication, Control and Computing 1999, pp. 368–377.

[46] J. R. Vergara and P. A. Estévez: A review of feature selection methods based on mutual information. Neural Comput. Appl. *24* (2014), 1, 175–186. DOI:10.1007/s00521-013-1368-0

[47] S. Watanabe: Information theoretical analysis of multivariate correlation. IBM J. Res. Develop. *4* (1960), 1, 66–82. DOI:10.1147/rd.41.0066

[48] H. S. Witsenhausen and A. D. Wyner: A conditional entropy bound for a pair of discrete random variables. IEEE Trans. Inform. Theory *21* (1075), 5, 493–501. DOI:10.1109/tit.1975.1055437

[49] V. Yemelichev, M. Kovalev, and M. Kravtsov: Polytopes, Graphs and Optimisation. Cambridge University Press, 1984.

[50] K. Zahedi, N. Ay, and R. Der: Higher coordination with less control: A result of information maximization in the sensorimotor loop. Adaptive Behavior *18* (2010), 3–4, 338–355. DOI:10.1177/1059712310375314

[51] K. Zahedi, G. Martius, and N. Ay: Linear combination of one-step predictive information with an external reward in an episodic policy gradient setting: a critical analysis. Front. Psychol. (2013), 4, 801. DOI:10.3389/fpsyg.2013.00801

Appendix

A. RELATED MEASURES OF MULTIVARIATE CORRELATION

Multi-information as defined above was first introduced by McGill (1954) [29] under the name of *total correlation*, and later discussed in [47] under the same name. Over time as the relationship between total correlation and mutual information became clear, several authors have begun to call this quantity multi-information [5, 17, 42, 41, 8]. However, McGill's seminal work on information transmission also introduced an entirely separate information-theoretic quantity under the name multivariate information. This quantity's name shifted over time to also being called multi-information

and multivariate mutual information, as referenced in [46, 22]. This naming confusion has been the source of false equalities between the two quantities, such as in [38, Eqn. 8]. One fundamental difference in the two quantities is that multi-information as defined here is non-negative, and multivariate mutual information can take positive and negative values.

B. DISTRIBUTIONS WITH FIXED MARGINS

In various contexts one is interested in the sets joint distributions that are compatible with given values of some of the marginals. We encounter a version of this problem here. Concretely, we study whether certain collections of subsets of variables can each have maximum multi-information, and, if so, what is the set of joint distributions that are compatible with these specifications.

B.1. Transportation polytopes

An n-table is an $N_1 \times \cdots \times N_n$ array of nonnegative real numbers. For a set of indices $\lambda \subseteq \{1, \ldots, n\}$, the λ-margin of a table p is the $|\lambda|$-table obtained by summing the entries over all but the λ indices. A multi-index transportation polytope or contingency table is the set of all n-tables that satisfy a set of given margins. The polytope is called *central* if all entries of any of the considered margins are equal. An assignment polytope is the special case where all entries across all of the considered margins are equal (which implies $N_1 = \cdots = N_n = N$). The two-way assignment polytopes are also called Birkhoff polytopes or polytopes of doubly stochastic matrices (when the value of the margin entries is 1). The $N!$ vertices of the Birkhoff polytope are the permutation matrices of size $N \times N$.

Optimal transportation problems are often formulated as linear programs with feasible set given by a transportation polytope. Hence the number of vertices and edges has been subject of study. For a 2-way transportation polytope, it is known that the dimension is equal to $(N_1 - 1)(N_2 - 1)$ (if the entry sums of the two margins are equal). A matrix p in the polytope is a vertex if and only if the graph F_p with edges $\{(x_1, x_2) : p(x_1, x_2) > 0\}$ forms a spanning forest of the full bipartite graph with edges $\{(x_1, x_2) : 1 \leq x_1 \leq N_1, 1 \leq x_2 \leq N_2\}$. See [49]. Already for 3-way transportation polytopes, the number of vertices is not known in general [16].

The set of maximizers of the SFMI is a union of polytopes, each of which has the structure of an n-way assignment polytope defined by 1-margins of size N.

B.2. Marginal polytopes

Marginal distributions satisfy certain relations among each other. Consider the tuple of λ-marginals for all $\lambda \in \Lambda$ for some fixed $\Lambda \in 2^{\{1,\ldots,n\}}$. The set of all possible tuples is called the marginal polytope. If we consider all q-marginals among n variables with state spaces of cardinality N, the dimension of the marginal polytope is $\sum_{i=1}^{q} \binom{n}{i}(N - 1)^i$. In turn, a fixed feasible and generic choice of all q marginals cuts out a set (a transportation polytope) in the joint probability simplex of dimension $N^n - 1 - \sum_{i=1}^{q} \binom{n}{i}(N - 1)^i$. If the choice of the margins sits at the boundary of the marginal polytope, then the set that is cut out in the joint probability simplex will be smaller. Classifying the possible values of the dimension is an open problem in general. In our discussion of transportation polytopes maximizing the SFMI, we remove zeros and obtain uniform margins.

Marginalizing over $n - q$ variables can be viewed as a linear projection $\mathcal{L} : \Delta_n \to \Delta_q$. Each marginal distribution that we specify in Δ_q provides N^q linear constraints (including normalization). In our discussion with margins being maximizers of multi-information, N entries of the margin take value $1/N$ and the other $N^q - N$ entries take value 0. The constraints with value 0 imply that $(N^q - N) \cdot N^{n-q}$ entries of the joint distributions also take value 0. So each margin creates N linear constraints for $N \cdot N^{n-q}$ entries of the joint distribution, and in addition, it equates the remaining entries of the joint distribution to zero. If we forget about the inequality constraints, the number of linearly independent constraints that arise from the specification of the marginal distributions is equal to the dimension of the marginal polytope.

C. EXAMPLES OF THE SETS OF SFMI MAXIMIZERS

C.1. The dimension of the sets of SFMI maximizers

The linear sets of SFMI maximizers come from taking the intersection of n margin constraints, each of which defines a set $\mathcal{S}_{i,j}$. Each of these sets is defined by a number of linear equality constraints, which in general are not independent for multiple margins. When there are n pairs of binary variables, each $\mathcal{S}_{i,j}$ is supported on 2^{2n-1} vectors. Only half of the support vectors between any given $\mathcal{S}_{i,j}$ agree, meaning $\mathcal{S}_{i,j} \cap \mathcal{S}'_{i,j}$ has common support of size 2^{2n-2}, and $\mathcal{S}_{i,j} \cap \mathcal{S}'_{i,j} \cap \mathcal{S}''_{i,j}$ has support of size 2^{2n-3}. Once the intersection of n of these sets is performed, one is left with $2^{2n-n} = 2^n$ support vectors. Then, the linearly independent constraints arising from the $\mathcal{S}$'s can be subtracted off of 2^n. The result will be the dimension of the linear sets. The sets of maximizers also need to satisfy linear inequality constraints. In general, these can further reduce the dimension of the solution set.

Example 24. Consider $n = 1$ pair of $N = 2$ valued (binary) variables. The SFMI maximizing set is exactly the MI maximizers, i. e. a zero dimensional set. The set $\mathcal{S}_{1,1}$ has support on $2^1 = 2$ vectors, 00 and 11, and there are two constraints, $a = 1/2$ and $b = 1/2$ for $a\delta_{00} + b\delta_{11}$. Written as a matrix, it would be a rank 2 matrix of constraints. Normalization is satisfied by these constraints, and the result is a zero dimensional set.

Example 25. Consider $n = 2$ pairs of $N = 2$ valued variables. Then $\mathcal{S}_{1,1} \cap \mathcal{S}_{2,1}$ has 4 vectors in the support, and the constraints that arise are given by

$$
\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \tag{21}
$$

where $p = \frac{a}{2}\delta_{0000} + \frac{b}{2}\delta_{0101} + \frac{c}{2}\delta_{1010} + \frac{d}{2}\delta_{1111}$. This matrix is rank 3, meaning the dimension of the resulting linear set is $4 - 3 = 1$.

Example 26. When $n = 3$ pairs of $N = 2$ valued variables. Then the intersection

$\mathcal{S}_{1,1} \cap \mathcal{S}_{2,1} \cap \mathcal{S}_{3,1}$ has support on $2^3 = 8$ vectors, and the rank of the constraint matrix,

$$
\begin{pmatrix}
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\
1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\
1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & 1
\end{pmatrix}
\begin{pmatrix} a \\ b \\ c \\ d \\ e \\ f \\ g \\ h \end{pmatrix}
=
\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}
\tag{22}
$$

is 4. Therefore we expect the linear sets of maximizers to be of dimension of null space, $8 - 4 = 4$.

Example 27. Consider $n = 2$ pairs of $N = 3$ valued variables. Then $\mathcal{S}_{1,1} \cap \mathcal{S}_{2,1}$ is supported on $N^n = 9$ vectors. The constraint matrix looks like,

$$
\begin{pmatrix}
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1
\end{pmatrix}
\begin{pmatrix} a \\ b \\ c \\ d \\ e \\ f \\ g \\ h \\ u \end{pmatrix}
=
\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.
\tag{23}
$$

Here, the coefficients correspond to the joint distributions probabilities on $0000, 0101, 02020, 1010, 1111, 1212, 2020, 2121, 2222$ respectively. This matrix is rank 5 meaning the null space is 4 dimensional.

C.2. Transportation polytopes of SFMI maximizers

Example 28. Consider the case of $n = 2$ pairs of $N = 2$ valued variables. In the following we illustrate explicitly how the SFMI maximizing distributions noted in Example 19 are all of the SFMI maximizers. Since it is known that distributions maximizing SFMI must maximize each term of the sum individually, one may consider the set of SFMI maximizers as the intersection of n sets where each set contains all the distributions which maximize a single term of the sum, i.e. each maximize $D(p(x_i, y_i) \| \mathcal{F})$ for some i.

Consider the four binary variables X_1, X_2, Y_1, Y_2. The maximizers of $I(X_1, Y_1)$ are $p_1^1(X_1, Y_1) = \frac{1}{2}(\delta_{00} + \delta_{11})$ and $p_1^2(X_1, Y_1) = \frac{1}{2}(\delta_{01} + \delta_{10})$. The corresponding joint distributions have the form $p(X_1, X_2, Y_1, Y_2) = p_1^j(X_1, Y_1)q(X_2, Y_2 | X_1, Y_1)$. Therefore, one may define $\mathcal{S}_{1,1}$ to be the set of distributions which maximize $I(X_1, Y_1)$ and have marginal $p_1^1(X_1, Y_1)$. Likewise, $\mathcal{S}_{1,2}$ is the set of distributions which maximize $I(X_1, Y_1)$ and have marginal $p_1^2(X_1, Y_1)$. These can be written as

$$
\left\{ p \in \Delta_{15} \, : \, p = \frac{1}{2} \sum_{x_2, y_2} q(x_2, y_2 | 0, 0) \delta_{0 x_2 0 y_2} + \frac{1}{2} \sum_{x_2, y_2} q(x_2, y_2 | 1, 1) \delta_{1 x_2 1 y_2} \right\}
\tag{24}
$$

and

$$\left\{ p \in \Delta_{15} \; : \; p = \frac{1}{2} \sum_{x_2, y_2} q(x_2, y_2 | 0, 1) \delta_{0 x_2 1 y_2} + \frac{1}{2} \sum_{x_2, y_2} q(x_2, y_2 | 1, 0) \delta_{1 x_2 0 y_2} \right\} \quad (25)$$

respectively. Here, $q(x_2, y_2 | x_1 = 0, y_1 = 0)$ is the conditional probability of observing the joint state $(X_1 = 0, X_2 = x_2, Y_1 = 0, Y_2 = y_2)$. Similarly, the two sets $\mathcal{S}_{2,1}$ and $\mathcal{S}_{2,2}$ which maximize $I(X_2, Y_2)$ and have marginals $p_2^1(x_2, y_2) = \frac{1}{2}(\delta_{00} + \delta_{11})$ and $p_2^2(x_2, y_2) = \frac{1}{2}(\delta_{01} + \delta_{10})$ respectively can be written as

$$\left\{ p \in \Delta_{15} \; : \; p = \frac{1}{2} \sum_{x_1, y_1} q(x_1, y_1 | 0, 0) \delta_{x_1 0 y_1 0} + \frac{1}{2} \sum_{x_1, y_1} q(x_1, y_1 | 1, 1) \delta_{x_1 1 y_1 1} \right\} \quad (26)$$

and

$$\left\{ p \in \Delta_{15} \; : \; p = \frac{1}{2} \sum_{x_1, y_1} q(x_1, y_1 | 0, 1) \delta_{x_1 0 y_1 1} + \frac{1}{2} \sum_{x_1, y_1} q(x_1, y_1 | 1, 0) \delta_{x_1 1 y_1 0} \right\}. \quad (27)$$

There are 4 possibilities for how these sets may be intersected to yield maximizers of SFMI. These are, $\mathcal{S}_{1,1} \cap \mathcal{S}_{2,1}$, $\mathcal{S}_{1,1} \cap \mathcal{S}_{2,2}$, $\mathcal{S}_{1,2} \cap \mathcal{S}_{2,1}$, and $\mathcal{S}_{1,2} \cap \mathcal{S}_{2,2}$. It will be shown that the intersection of $\mathcal{S}_{1,1} \cap \mathcal{S}_{2,1}$ corresponds to the line of distributions in Eqn. (19), and the other intersections correspond in a similar fashion. To see this, one may begin by considering the binary codes in the intersection. These are all the binary codes which match in the 1st and 3rd entry, and match in the 2nd and 4th entry. Thus, the distributions in $\mathcal{S}_{1,1} \cap \mathcal{S}_{2,1}$ can only have support on $\{0000, 0101, 1010, 1111\}$. Thus, at most the intersection could be a 3 dimensional set, due to the normalization constraint. Next, consider the linear constraints build into $\mathcal{S}_{1,1}$. Letting

$$\begin{aligned}
a &= q(X_2 = 0, Y_2 = 0 | X_1 = 0, Y_1 = 0) \\
b &= q(X_2 = 0, Y_2 = 0 | X_1 = 1, Y_1 = 1) \\
c &= q(X_2 = 1, Y_2 = 1 | X_1 = 0, Y_1 = 0) \\
d &= q(X_2 = 1, Y_2 = 1 | X_1 = 1, Y_1 = 1),
\end{aligned} \quad (28)$$

the constraints can be written as $a + c = 1$ and $b + d = 1$. Similarly, the distribution in $\mathcal{S}_{2,1}$ must satisfy $a + b = 1$ and $c + d = 1$. These constraints require that $a = d$, and $b = c$. The distributions in $\mathcal{S}_{1,1} \cap \mathcal{S}_{2,1}$ must take the form

$$p = \frac{a}{2} \delta_{0000} + \frac{b}{2} \delta_{1010} + \frac{c}{2} \delta_{0101} + \frac{d}{2} \delta_{1111}. \quad (29)$$

Letting $\alpha = \frac{a}{2} = \frac{d}{2}$ and $\beta = \frac{b}{2} = \frac{c}{2}$, the distributions in $\mathcal{S}_{1,1} \cap \mathcal{S}_{2,1}$ can be written as in Eqn. (19),

$$p = \alpha(\delta_{0000} + \delta_{1111}) + \beta(\delta_{0101} + \delta_{1010}), \quad (30)$$

where $\alpha + \beta = \frac{1}{2}$. By considering the other intersections, $\mathcal{S}_{1,1} \cap \mathcal{S}_{2,2}$, $\mathcal{S}_{1,2} \cap \mathcal{S}_{2,1}$, $\mathcal{S}_{1,2} \cap \mathcal{S}_{2,2}$, the other 3 transportation polytopes in Example 19 are recovered.

Example 29. Consider the case of $n = 3$ pairs of $N = 2$ valued variables $(X_1, X_2, X_3, Y_1, Y_2, Y_3)$. For any $(i, j) \in \{1, 2, 3\} \times \{1, 2\}$, let $\mathcal{S}_{i,j}$ denote the set of joint distributions whose (X_i, Y_i) margin is the jth maximizer of the mutual information. Each $\mathcal{S}_{i,j}$ contains distributions supported on $2^{6-1} = 32$ binary vectors. The intersections $\mathcal{S}_{1,j_1} \cap \mathcal{S}_{2,j_2}$ contain distributions supported on $2^{6-2} = 16$ vectors, and

when intersecting $\mathcal{S}_{1,j_1} \cap \mathcal{S}_{2,j_2} \cap \mathcal{S}_{3,j_3}$, we obtain distributions supported on 8 binary vectors.

For concreteness, let $\mathcal{S}_{i,1}$ correspond to $p(X_i, Y_i) = \frac{1}{2}(\delta_{00} + \delta_{11})$. Then $\mathcal{S}_{1,1}$ is the set of joint distributions of the form

$$p = \frac{1}{2} \sum_{\substack{x_2,x_3 \\ y_2,y_3}} \Big(q(x_2,x_3,y_2,y_3|0,0)\delta_{0x_2x_30y_2y_3} + q(x_2,x_3,y_2,y_3|1,1)\delta_{1x_2x_31y_2y_3} \Big), \quad (31)$$

where the numbers $q(x_2,x_3,y_2,y_3|x_1,y_1)$ define an arbitrary conditional distribution. Each of these p is supported on strings $(x_1,x_2,x_3,y_1,y_2,y_3)$ that satisfy $x_1 = y_1$.

The distributions $p \in \mathcal{S}_{1,1} \cap \mathcal{S}_{2,1} \cap \mathcal{S}_{3,1}$ are supported on strings $(x_1,x_2,x_3,y_1,y_2,y_3)$ that satisfy $x_i = y_i$ for all $i = 1,2,3$. In turn, they have the form

$$p = \frac{a}{2}\delta_{000000} + \frac{b}{2}\delta_{001001} + \frac{c}{2}\delta_{010010} + \frac{d}{2}\delta_{011011} + \frac{e}{2}\delta_{100100} + \frac{f}{2}\delta_{101101} + \frac{g}{2}\delta_{110110} + \frac{h}{2}\delta_{111111}.$$

The coefficients correspond to conditional probability distributions in three different ways. One way is as the conditional probabilities of X_2, X_3, Y_2, Y_3 given X_1, Y_1, with

$$
\begin{aligned}
a &= q(0,0,0,0|0,0), & e &= q(0,0,0,0|1,1), \\
b &= q(0,1,0,1|0,0), & f &= q(0,1,0,1|1,1), \\
c &= q(1,0,1,0|0,0), & g &= q(1,0,1,0|1,1), \\
d &= q(1,1,1,1|0,0), & h &= q(1,1,1,1|1,1),
\end{aligned}
\quad (32)
$$

which corresponds to (31) and implies that $a + b + c + d = 1$ and $e + f + g + h = 1$. The other two ways result from conditioning on X_2, Y_2 and X_3, Y_3, respectively, and imply two more pairs of linear constraints. The six constraints form the linear system $M\mathbf{q} = \mathbf{1}$ with

$$
\begin{pmatrix}
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\
1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\
1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & 1
\end{pmatrix}
\begin{pmatrix}
a \\ b \\ c \\ d \\ e \\ f \\ g \\ h
\end{pmatrix}
=
\begin{pmatrix}
1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1
\end{pmatrix}.
\quad (33)
$$

The matrix M is simply the map computing the (X_i, Y_i)-margins, restricted to the support points (i.e., strings with $x_i = y_i$ for $i = 1,2,3$). It can be regarded as a sufficient statistics matrix of a binary independence model of n variables, including the statistic that computes the total mass. The matrix has rank 4 and the vector $\frac{1}{4}\mathbf{1} \in \mathbb{R}^8$ is a particular solution of the linear system. At this particular solution, one obtains a maximizer of $MI(\mathbf{X}, \mathbf{Y})$.

As discussed in Example 26, the space of solutions is 4 dimensional. To obtain all solutions, we consider the kernel of M, which is spanned by the rows of

$$
\begin{pmatrix}
1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\
1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\
1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\
1 & -1 & -1 & 1 & -1 & 1 & 1 & -1
\end{pmatrix}.
\quad (34)
$$

Any linear combination of these rows can be added to $\frac{1}{4}\mathbf{1} \in \mathbb{R}^8$ to obtain another vector that satisfies the marginal equality constraints. The intersection of the 4-dimensional affine space with the probability simplex is obtained by requiring that the solutions have non-negative entries. This results in a polytope. We compute the vertex representation using the free open-source mathematics software Sage [44], and obtain the 6 rows

$$
\begin{pmatrix}
0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1/2 & 0 & 0 & 1/2 & 0 & 1/2 & 1/2 & 0 \\
0 & 1/2 & 1/2 & 0 & 1/2 & 0 & 0 & 1/2
\end{pmatrix}. \tag{35}
$$

Based off of the vertex representation above, the distributions which are in $\mathcal{S} = \mathcal{S}_{1,1} \cap \mathcal{S}_{2,1} \cap \mathcal{S}_{3,1}$ are of the form

$$
\begin{aligned}
p = {} & \tfrac{\alpha_1}{2}(\delta_{000000} + \delta_{111111}) + \tfrac{\alpha_2}{2}(\delta_{100100} + \delta_{011011}) \\
& + \tfrac{\alpha_3}{2}(\delta_{010010} + \delta_{101101}) + \tfrac{\alpha_4}{2}(\delta_{001001} + \delta_{110110}) \\
& + \tfrac{\alpha_5}{4}(\delta_{000000} + \delta_{011011} + \delta_{101101} + \delta_{110110}) \\
& + \tfrac{\alpha_6}{4}(\delta_{001001} + \delta_{010010} + \delta_{100100} + \delta_{111111}),
\end{aligned}
$$

where $\sum_i \alpha_i = 1$ ensures normalization and $\alpha_i \geq 0$ for $i = 1, \dots, 6$. One can quickly check that $p(X_i, Y_i) = \frac{1}{2}(\delta_{00} + \delta_{11})$ for $i = 1, 2, 3$. The vertices $\alpha_1 = 1$, $\alpha_2 = 1$, $\alpha_3 = 1$, $\alpha_4 = 1$ are the only distributions with support of cardinality two and they are also the only maximizers of multi-information within $\mathcal{S}$. The choices $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1/4$ and $\alpha_5 = \alpha_6 = 1/2$ give the same distribution, which is also the only maximizer of $MI(\mathbf{X}, \mathbf{Y})$ within $\mathcal{S}$. This observation motivates Theorem 17 which states that each polytope of SFMI maximizers contains one or more simplices, and at the center of each simplex lies a $MI(\mathbf{X}, \mathbf{Y})$ maximizer. Furthermore, when restricted to the set of SFMI maximizers, the distributions with maximum entropy are $MI(\mathbf{X}, \mathbf{Y})$ maximizers.

We have discussed one of the $N!^n = 2!^3 = 8$ polytopes. The other 7 are similar. Each of them has 4 vertices that are maximizers of the multi-information and form a 3-simplex with a maximizer of $MI(\mathbf{X}, \mathbf{Y})$ at its center. The 8 centers are following vectors scaled by $1/8$, whereby grouped terms correspond to the vertices:

$(\delta_{000000} + \delta_{111111}) + (\delta_{001001} + \delta_{110110}) + (\delta_{010010} + \delta_{101101}) + (\delta_{011011} + \delta_{100100})$,

$(\delta_{000001} + \delta_{111110}) + (\delta_{001000} + \delta_{110111}) + (\delta_{010011} + \delta_{101100}) + (\delta_{011010} + \delta_{100101})$,

$(\delta_{000010} + \delta_{111101}) + (\delta_{010000} + \delta_{101111}) + (\delta_{001011} + \delta_{110100}) + (\delta_{011001} + \delta_{100110})$,

$(\delta_{000100} + \delta_{111011}) + (\delta_{100000} + \delta_{011111}) + (\delta_{001101} + \delta_{110010}) + (\delta_{101001} + \delta_{010110})$,

$(\delta_{000011} + \delta_{111100}) + (\delta_{001010} + \delta_{110101}) + (\delta_{010001} + \delta_{101110}) + (\delta_{011000} + \delta_{100111})$,

$(\delta_{000101} + \delta_{111010}) + (\delta_{001100} + \delta_{110011}) + (\delta_{100001} + \delta_{011110}) + (\delta_{101000} + \delta_{010111})$,

$(\delta_{000110} + \delta_{111001}) + (\delta_{010100} + \delta_{101011}) + (\delta_{100010} + \delta_{011101}) + (\delta_{110000} + \delta_{001111})$,

$(\delta_{000111} + \delta_{111000}) + (\delta_{001110} + \delta_{110001}) + (\delta_{010101} + \delta_{101010}) + (\delta_{011100} + \delta_{100011})$.

One can see that the simplices are disjoint. In fact, distributions from different simplices have disjoint supports, meaning that the simplices sit in disjoint faces of the joint probability simplex. These simplices having disjoint support only occurs when dealing with binary variables.

Example 30. Consider the case of $n = 2$ pairs of $N = 3$ valued variables. For each $i = 1, 2$ let $\mathcal{S}_{i,j_i}$ be the set of joint distributions which have marginal $p(X_i, Y_i)$ as one of the $N! = 3! = 6$ multi-information maximizers, so $j_i \in \{1, \ldots, 6\}$. For concreteness, let the correspondence be

$$\mathcal{S}_{i,1} : \quad p(X_i, Y_i) = \frac{1}{3}(\delta_{00} + \delta_{11} + \delta_{22}),$$

$$\mathcal{S}_{i,2} : \quad p(X_i, Y_i) = \frac{1}{3}(\delta_{00} + \delta_{12} + \delta_{21}),$$

$$\mathcal{S}_{i,3} : \quad p(X_i, Y_i) = \frac{1}{3}(\delta_{01} + \delta_{10} + \delta_{22}),$$

$$\mathcal{S}_{i,4} : \quad p(X_i, Y_i) = \frac{1}{3}(\delta_{01} + \delta_{12} + \delta_{20}),$$

$$\mathcal{S}_{i,5} : \quad p(X_i, Y_i) = \frac{1}{3}(\delta_{02} + \delta_{11} + \delta_{20}),$$

$$\mathcal{S}_{i,6} : \quad p(X_i, Y_i) = \frac{1}{3}(\delta_{02} + \delta_{10} + \delta_{22}).$$

There are $N!^n = 3!^2 = 36$ intersections of the form $\mathcal{S}_{1,j_1} \cap \mathcal{S}_{2,j_2}$, each corresponding to a transportation polytope. They together contain all $N!^{2n-1} = 3!^{2 \cdot 2 - 1} = 216$ maximizers of the multi-information. Consider the support sets of three to start:

$$\text{supp}(\mathcal{S}_{1,1} \cap \mathcal{S}_{2,1}) = \{0000, 0101, 0202, 1010, 1111, 1212, 2020, 2121, 2222\},$$
$$\text{supp}(\mathcal{S}_{1,1} \cap \mathcal{S}_{2,2}) = \{0000, 0102, 0201, 1010, 1112, 1211, 2020, 2122, 2221\},$$
$$\text{supp}(\mathcal{S}_{1,1} \cap \mathcal{S}_{2,3}) = \{0001, 0100, 0202, 1011, 1110, 1212, 2021, 2120, 2222\}.$$

The probabilities assigned to the joint states are not arbitrary, since they need to satisfy the margins. One particular subset of distributions in $\mathcal{S}_{1,1} \cap \mathcal{S}_{2,1}$ can be written as

$$p = \frac{\alpha_1}{3}(\delta_{0000} + \delta_{1111} + \delta_{2222}) + \frac{\alpha_2}{3}(\delta_{0101} + \delta_{1212} + \delta_{2020}) + \frac{\alpha_3}{3}(\delta_{0202} + \delta_{1010} + \delta_{2121}),$$

where $\alpha_1 + \alpha_2 + \alpha_3 = 1$. This is a simplex whose $N^{n-1} = 3^{2-1} = 3$ vertices maximize multi-information. Another one can be written as

$$p = \frac{\alpha_1}{3}(\delta_{0000} + \delta_{1212} + \delta_{2121}) + \frac{\alpha_2}{3}(\delta_{1111} + \delta_{2020} + \delta_{0202}) + \frac{\alpha_3}{3}(\delta_{2222} + \delta_{0101} + \delta_{1010}).$$

This is another simplex whose 3 vertices maximize multi-information. Both simplices have the same centroid, with $\alpha_1 = \alpha_2 = \alpha_3 = 1/3$, which is a maximizer of $MI(\mathbf{X}, \mathbf{Y})$.

The other transportation polytopes (intersections $\mathcal{S}_{1,j_1} \cap \mathcal{S}_{2,j_2}$) are similar. For example, the two simplices in $\mathcal{S}_{1,1} \cap \mathcal{S}_{2,2}$ are

$$\frac{\alpha_1}{3}(\delta_{0000} + \delta_{1211} + \delta_{2122}) + \frac{\alpha_2}{3}(\delta_{1112} + \delta_{2020} + \delta_{0201}) + \frac{\alpha_3}{3}(\delta_{1010} + \delta_{0102} + \delta_{2221}),$$
$$\frac{\alpha_1}{3}(\delta_{0000} + \delta_{1112} + \delta_{2221}) + \frac{\alpha_2}{3}(\delta_{1211} + \delta_{0102} + \delta_{2020}) + \frac{\alpha_3}{3}(\delta_{2122} + \delta_{1010} + \delta_{0201}),$$

and the two simplices in $\mathcal{S}_{1,1} \cap \mathcal{S}_{2,3}$ are

$$\frac{\alpha_1}{3}(\delta_{0001} + \delta_{1212} + \delta_{2120}) + \frac{\alpha_2}{3}(\delta_{2222} + \delta_{1011} + \delta_{0100}) + \frac{\alpha_3}{3}(\delta_{2021} + \delta_{1110} + \delta_{0202}),$$
$$\frac{\alpha_1}{3}(\delta_{0001} + \delta_{2222} + \delta_{1110}) + \frac{\alpha_2}{3}(\delta_{1212} + \delta_{2021} + \delta_{0100}) + \frac{\alpha_3}{3}(\delta_{2120} + \delta_{0202} + \delta_{1011}).$$

Observe that there are $N!^{n-1} = 3!^{2-1} = 6$ multi-information maximizers per transportation polytope, and there are $N!^n$ transportation polytopes total. This number is expected because all of the $N!^{2n-1}$ maximizers of multi-information are SFMI maximizers.

D. INHOMOGENEOUS VARIABLES

We consider two scenarios for inhomogenous variables which have relevance in reinforcement learning and representation learning, respectively.

D.1. Pairs of different sizes

The first setting is considering the SFMI maximizers for $(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ where $|\mathcal{X}_i| = |\mathcal{Y}_i|$, but in general $|\mathcal{X}_i| \neq |\mathcal{X}_j|$. Without loss of generality, we can assume that the variables are labeled such that $|\mathcal{X}_1| \leq |\mathcal{X}_2| \leq \cdots \leq |\mathcal{X}_n|$. The state spaces here for example represent the number of states a given robot sensor can take, where $\mathbf{X} = (X_1, \ldots, X_n)$ is the joint sensor reading at time $t-1$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ is the sensor reading at time t. Here, the maximizers of

$$SFMI(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{n} MI(X_i, Y_i) \tag{36}$$

generalize from Theorem 16 when $\mathrm{LCM}(|\mathcal{X}_1|, \ldots, |\mathcal{X}_{n-1}|) = |\mathcal{X}_n|$. If this is true, then the conditions for the analysis from Ay and Knauf [5, Theorem 4.4 and Remark 4.5] for inhomogeneous variables are met. This means that the maximizers of the SFMI will maximize each term in the sum individually, allowing us to construct the set of SFMI maximizers with the same technique as before. If we denote $|\mathcal{X}_i| = N_i$, then the results in Theorem 16 generalize as follows: The support set of the SFMI maximizers have cardinality $\prod_{i=1}^{n} N_i$. The ith marginal constraint contributes $N_i - 1$ unique constraints, and they all share the normalization constraint, meaning the dimension of the resulting set of maximizers will be $\prod_{i=1}^{n} N_i - 1 - \prod_{i=1}^{n}(N_i - 1)$. The SFMI maximizers will still contain the distributions which maximize each $MI(X_i, Y_i)$ for $i \in [n]$, and be a superset of the multi-information maximizers. The number of transportation polytopes will be $\prod_{i=1}^{n} N_i!$, and the properties previously described are all effectively the same with the replacement of N by N_i, and replacing N^n with $\prod_{i \in [n]} N_i$.

If $\mathrm{LCM}(|\mathcal{X}_1|, \ldots, |\mathcal{X}_{n-1}|) \neq |\mathcal{X}_n|$, the terms in the SFMI can not be maximized independently of one another. Characterizing the SFMI maximizers for this scenario is left for future work.

D.2. Single output variable

In the second inhomogeneous situation that we consider, we would like to maximize the Λ-FMI for $(X_1, \ldots, X_n, Y)$, where $|\mathcal{X}_i| = N$ for all $i \in [n]$ and $|\mathcal{Y}| \geq N$, and $\Lambda = \big\{\{i, n+1\} : i \in [n]\big\}$. The Λ-FMI where $\Lambda = \big\{\{i, n+1\} : i \in [n]\big\}$ has been used as an alternative objective function to maximizing the global mutual information in deep representation learning, where Y is the learned representation. For more details on this see [20]. Since all marginals specified by Λ contain Y, the maximizers of multi-information will maximize the Λ-FMI. This follows directly from [5, Remark 4.5(c)], as the multi-information maximizers have marginals $p(X_i, Y)$ which maximize mutual information for all $i \in [n]$. Since each term in the sum

$$\Lambda\text{-FMI} = \sum_{i=1}^{n} MI(X_i, Y) \tag{37}$$

is maximized, Λ-FMI is also maximized. In this case the maximizers of multi-information equal the maximizers of Λ-FMI.

Thomas Merkh, Department of Mathematics, UCLA, CA 90095. U. S. A.
  e-mail: tmerkh@ucla.edu

Guido Montúfar, Department of Mathematics and Department of Statistics, UCLA, CA 90095, U. S. A. and Max Planck Institute for Mathematics in the Sciences, 04103 Leipzig. Germany.
  e-mail: montufar@math.ucla.edu