Published in: IEEE Transactions on Information Theory journal.
2018 IEEE.

Minimum–Entropy Couplings and their Applications

Ferdinando Cicalese, Luisa Gargano, *Member, IEEE* and Ugo Vaccaro, *Senior Member, IEEE*

***Abstract*—Given two discrete random variables X and Y, with probability distributions $\mathbf{p} = (p_1, \ldots, p_n)$ and $\mathbf{q} = (q_1, \ldots, q_m)$, respectively, denote by $\mathcal{C}(\mathbf{p}, \mathbf{q})$ the set of all couplings of $\mathbf{p}$ and $\mathbf{q}$, that is, the set of all bivariate probability distributions that have $\mathbf{p}$ and $\mathbf{q}$ as marginals. In this paper, we study the problem of finding a joint probability distribution in $\mathcal{C}(\mathbf{p}, \mathbf{q})$ of *minimum entropy* (equivalently, a coupling that *maximizes* the mutual information between X and Y), and we discuss several situations where the need for this kind of optimization naturally arises. Since the optimization problem is known to be NP-hard, we give an efficient algorithm to find a joint probability distribution in $\mathcal{C}(\mathbf{p}, \mathbf{q})$ with entropy exceeding the minimum possible at most by 1 bit, thus providing an approximation algorithm with an additive gap of at most 1 bit. Leveraging on this algorithm, we extend our result to the problem of finding a minimum–entropy joint distribution of arbitrary $k \geq 2$ discrete random variables $X_1, \ldots, X_k$, consistent with the known k marginal distributions of the individual random variables $X_1, \ldots, X_k$. In this case, our algorithm has an additive gap of at most $\log k$ from optimum. We also discuss several related applications of our findings and extensions of our results to entropies different from the Shannon entropy.**

***Index Terms*—Entropy minimization, Mutual Information maximization, coupling, majorization.**

I. Introduction and Motivations

Inferring an unknown joint distribution of two random variables (r.v.), when only their marginals are given, is an old problem in the area of probabilistic inference. The problem goes back at least to Hoeffding [27] and Frechet [18], who studied the question of identifying the extremal joint distribution of r.v. X and Y that maximizes (resp., minimizes) their correlation, given the marginal distributions of X and Y. We refer the reader to [3], [11], [13], [37] for a (partial) account of the vast literature in the area and the many applications in the pure and applied sciences.

In this paper, we consider the following case of the problem described above. Let X and Y be two discrete r.v., distributed according to $\mathbf{p} = (p_1, \ldots, p_n)$ and $\mathbf{q} = (q_1, \ldots, q_m)$, respectively. We seek a *minimum-entropy* joint probability distribution of X and Y, whose marginals are equal to $\mathbf{p}$ and $\mathbf{q}$.

We discuss below a few scenarios where this problem naturally arises.

A. Entropic Causal Inference

In papers [31] and [32], the authors consider the important question of identifying the correct causal direction between two arbitrary r.v.'s X and Y, that is, they want to discover whether it is the case that X causes Y or it is Y that causes X. In general, X causes Y if there exists an exogenous r.v. E (independent of X) and a deterministic function f such that $Y = f(X,E)$. In order to identify the correct causal direction (i.e., either from X to Y or from Y to X), the authors of [31] and [32] make the reasonable postulate that the entropy of the exogenous r.v. E is small in the *true* causal direction, and empirically validate this assumption. Additionally, they prove the important fact that the problem of finding the exogenous variable E with minimum entropy is equivalent to the problem of finding the minimum-entropy joint distribution of properly defined random variables, *given* (i.e., by fixing) their marginal distributions (see Theorem 3 of [32]). This is exactly the problem we consider in this paper. The authors of [32] observe that the latter optimization problem is NP-hard (due to results of [34] and [54]) and propose a greedy *approximation* algorithm to find the minimum-entropy joint distribution, given the marginals. For this greedy algorithm, the authors prove that it always finds a local minimum and that the local minimum is within an additive guaranteed gap from the unknown global optimum. The authors of [32] observe that this additive guaranteed gap can be as large as $\log n$ (here n is the cardinality of the support of each involved random variable). Similar results are contained in [41], and references therein.

In this paper, we design a different greedy algorithm, and we prove that it returns a correct joint probability distribution (i.e., with the prescribed marginals) with entropy exceeding the *minimum possible by at most 1 bit*. Subsequently, in Section V, we extend our algorithm to the case of more than two random variables. More precisely, we consider the problem of finding a minimum–entropy joint distribution of arbitrary $k \geq 2$ discrete random variables $X_1, \ldots, X_k$, consistent with the known k marginal distributions of $X_1, \ldots, X_k$. In this case, our algorithm has an additive guaranteed gap of at most $\log k$.

B. On the functional representation of correlated random variables

Let X and Y be two arbitrary random variables with joint distribution $p(x, y)$. The functional representation lemma [16, p. 626] states that there exists a random variable Z independent of X, and a deterministic function f, such that the pair of r.v.'s $(X, f(X, Z))$ is distributed like (X, Y), that is, they have the same joint distribution $p(x, y)$. This lemma has been applied to establish several results in network information theory (see [16] and references therein). In several applications, it is important to find a r.v. Z such that the conditional entropy $H(Y|Z)$ is close to its natural lower bound, that is, it is close to $I(X;Y)$. Recently, a very strong result to that respect

F. Cicalese is with the Dipartimento di Informatica, Università di Verona, Verona, Italy (email: ferdinando.cicalese@univr.it), L. Gargano is with the Dipartimento di Informatica, Università di Salerno, Fisciano (SA), Italy (email: lgargano@unisa.it), and U. Vaccaro is with the Dipartimento di Informatica, Università di Salerno, Fisciano (SA), Italy (email: uvaccaro@unisa.it). This paper was presented in part at the 2017 IEEE International Symposium on Information Theory.

was proved in [36], showing that one can indeed find a r.v. Z such that $H(Y|Z) \leq I(X;Y) + \log(I(X;Y)+1) + 4$ bits. Among the papers that have used (versions of) the functional representation lemma, papers [4] and [26] have considered the problem of one-shot channel simulation with unlimited common randomness.[1] The bounds on $H(Y|Z)$ are essentially used to set a limit on the amount of bits exchanged among parties involved in the simulation. However, there is here another resource which is reasonable to bound: the amount of randomness used in the protocol, i.e.,, the amount of randomness needed to generate the auxiliary r.v. Z. Indeed, randomness *is not* free, and several clever (but expensive) methods have been devised to produce it, based on physical systems like Geiger-Muller tubes, chaotic laser, etc.; therefore it seems reasonable to require that the entropy of auxiliary r.v. Z be *minimum*[2]. On the basis of the (already mentioned) important result proved in [31], showing the equivalence between the problem of finding the minimum entropy auxiliary variable Z such that $(X, f(X,Z)) = (X,Y)$, and the problem of finding the minimum-entropy joint distribution of properly defined random variables (given their marginal distributions), it follows that the results of our paper offer a solution to the question of seeking a minimum entropy r.v. Z such that the pair of r.v. $(X;Y)$ can be simulated as $(X, f(X,Z))$. The exact statement of our result is given in Corollary 2 of Section V.

C. Metric for dimension reduction

Another work that considers the problem of finding the minimum-entropy joint distribution of two r.v. X and Y, given the marginals of X and Y, is the paper [54]. There, the author introduces a pseudo-metric $\mathtt{D}(\cdot,\cdot)$ among discrete probability distributions in the following way: given arbitrary $\mathbf{p} = (p_1, \ldots, p_n)$ and $\mathbf{q} = (q_1, \ldots, q_m)$, the distance $\mathtt{D}(\mathbf{p}, \mathbf{q})$ among $\mathbf{p}$ and $\mathbf{q}$ is defined as the quantity

$$\mathtt{D}(\mathbf{p}, \mathbf{q}) = 2W(\mathbf{p}, \mathbf{q}) - H(\mathbf{p}) - H(\mathbf{q}),$$

where $W(\mathbf{p}, \mathbf{q})$ is the *minimum* entropy of a bivariate probability distribution that has $\mathbf{p}$ and $\mathbf{q}$ as marginals, and H denotes the Shannon entropy. This metric is applied in [54] to the problem of dimension-reduction of stochastic processes. The author of [54] observes that the problem of computing $W(\mathbf{p}, \mathbf{q})$ is NP-hard (see also [34]) and proposes another different greedy algorithm for its computation, based on some analogy with the problem of Bin Packing with overstuffing. No performance guarantee is given in [54] for the proposed algorithm. Our result directly implies that we can compute the value of the pseudometric $\mathtt{D}(\mathbf{p}, \mathbf{q})$, for *arbitrary* $\mathbf{p}$ and $\mathbf{q}$, with an additive gap of at most 1 bit.[3]

[1]The situation is also somewhat reminiscent of the important "reverse Shannon Theorem" of [2], where one wants to simulate an arbitrary noisy channel with a noiseless one, plus some additional source of randomness (see [2] for formal definitions).

[2]This requirement can be made more formal by invoking the fundamental result of Knuth and Yao [30], stating that the minimum average number of unbiased random bits necessary to generate an arbitrary discrete r.v. Z is sandwiched between $H(Z)$ and $H(Z)+2$.

[3]We remark that in [8], [9] we considered the different problem of computing the probability distributions $\mathbf{q}^*$ that *minimizes* $\mathtt{D}(\mathbf{p}, \mathbf{q})$, given $\mathbf{p}$.

D. Contingency tables and transportation polytopes

In the field of Combinatorial Optimization, the set $\mathcal{C}(\mathbf{p}, \mathbf{q})$ of all couplings of given $\mathbf{p} = (p_1, \ldots, p_n)$ and $\mathbf{q} = (q_1, \ldots, q_m)$ is known as the transportation polytope P defined by $\mathbf{p}$ and $\mathbf{q}$. The fact that in our case $\mathbf{p}$ and $\mathbf{q}$ are probability distributions (i.e., their components are non-negative and sum up to 1) is immaterial, since one can always normalize. A similar concepts is known in Statistics under the name of contingency tables [15]. Polytopes $\mathcal{C}(\mathbf{p}, \mathbf{q})$ are called transportation polytopes because they model the transportation of goods from n supply locations (with the i-th location supplying a quantity of p_i) to m demand locations (with the j-th location demanding a quantity of q_j). The feasible points $M_{i,j}$ of an element $\mathbf{M} = [M_{i,j}] \in \mathcal{C}(\mathbf{p}, \mathbf{q})$ model the scenario where a quantity of $M_{i,j}$ of goods is transported from the i-th supply location to the j-th demand location. Many hard combinatorial optimization problems become solvable on the transportation polytope because of its rich and well studied combinatorial structure. We refer to the survey paper [14] for an account of the vast literature on the topic. The problem we consider in this paper can be equivalently stated as the one of finding a minimum-entropy element of the polytope $\mathcal{C}(\mathbf{p}, \mathbf{q})$. To see that this is not a simple translation of a problem from one language into another, we point out a recent important trend in the area of combinatorial optimization, that is, the one that seeks sparse solutions to optimization problems. More precisely, researchers aim at finding algorithms that trade the optimality of a solution to a given problem with the sparseness of the solution (e.g., the number of variables with non-zero values in the solution, but other measure of sparseness can be employed). We address the reader to [1], [49], and references therein, for motivations and a review of this line of research. Our problem of finding a minimum entropy element in the transportation polytope $\mathcal{C}(\mathbf{p}, \mathbf{q})$ fits in the above perspective. This interpretation is possible not only because entropy can be often interpreted as a reasonable measure of sparseness (see [28]) but also because our algorithm produces an elements of $\mathcal{C}(\mathbf{p}, \mathbf{q})$ whose number of non zero elements is, in the worst case, at most twice the minimum possible. We remark that finding a matrix $\mathbf{M} \in \mathcal{C}(\mathbf{p}, \mathbf{q})$ with the minimum number of non-zero entries is NP-hard in general [31].

E. Additional relations and applications

We believe that the problem of finding a minimum entropy joint distribution, with fixed marginal, is indeed a basic one. In this section we will briefly illustrate a few other scenarios where the problem matters.

Let us write the joint entropy of two r.v. X and Y, distributed according to $\mathbf{p}$ and $\mathbf{q}$, respectively, as $H(X,Y) = H(X) + H(Y) - I(X;Y)$, where $I(X;Y)$ is the mutual information between X and Y. Then, one sees that our original problem can be equivalently stated as the determination of a joint probability distribution of X and Y (having given marginals $\mathbf{p}$ and $\mathbf{q}$) that *maximizes* the mutual information $I(X;Y)$. In [34], this maximal mutual information is interpreted, in agreement with Renyi's axioms for a *bona fide* dependence measure [44], as a measure of the *largest possible*

correlations between two r.v. X and Y. One can see the soundness of this interpretation also in the following way. Let $q(x,y)$ be an arbitrary joint distribution of r.v. X and Y, with marginals equal to $\mathbf{p} = \{p(x)\}$ and $\mathbf{q} = \{q(y)\}$, respectively. Then, it is well known that the mutual information $I(X;Y)$ can be written as the relative entropy (divergence) between the joint distribution $q(x,y)$ and the joint distribution $r(x,y) = p(x)q(y)$ (i.e., a joint distribution that would make X and Y *independent*). Therefore, our problem of maximizing $I(X;Y)$ is equivalent to the one of finding a joint distribution of X and Y that is the *farthest* (in the sense of relative entropy) from $r(x,y) = p(x)q(y)$, that is, finding the joint distribution of X and Y that makes them "most dependent" (or correlated) as possible. Another way to see the question is to realize that we are seeking a joint distribution that minimizes the conditional entropies $H(X|Y)$ and $H(Y|X)$, that represent sensible measures of the strength of the dependence between X and Y. Since the problem of its exact computation is NP-hard, our result implies an approximation algorithm for it. We would like to remark that there are indeed situations in which measuring the "potential" correlation between variables (as opposed to their actual correlation) can be useful. For instance, the authors of [35] introduces a measure that, in their words, "*provides a score for the strength of the influence protein X has on protein Y. In many physiological conditions, only a small fraction of the cells have activated protein X in response to stimuli, and these active populations have little influence on the mutual information metric*". Since other standard measures of correlation would also fail, using a measure of *potential* correlation in that context could be useful.

Another situation where the need to maximize the mutual information between two r.v. (with fixed probability distributions) naturally arises, is in the area of medical imaging [43], [55].

In [46], the authors asks several algorithmic problems of this vein: given some fixed marginal distributions, find a joint probability distribution, with marginals *close* to the given ones, and satisfying some additional properties dictated by certain applications scenarios. Among the several problems considered in [46], the authors mention the problem of finding a minimum entropy joint distribution, with the given marginals, as an interesting open problem.

In the recent paper [57], the authors study the problem of finding a joint probability distribution of two random variables with fixed marginals, that minimizes a given function f of the joint distribution. The authors study this problem in the *asymptotic setting* (i.e., for product marginal distributions). A strictly related problem was also studied in [52].

In [21], the authors study the problem of finding good upper and lower bounds on the mutual information $I(X;Y)$ of two r.v.'s X and Y when the only available knowledge consists of the marginals of X and Y, and the pair of values (x,y) for which the unknown joint distribution of X and Y (consistent with the given marginals) assign a non-zero probability. It is clear that our maximization problem gives an upper bound on $I(X;Y)$ when the available knowledge consists of the marginals of X and Y, and *nothing else*.

Other papers considering problems somewhat related to ours are [19], [29], [38], [40], [47], [56], and [58].

F. Structure of the paper

The rest of the paper is organized as follows. In Section II we present the mathematical tools and the auxiliary results that are needed to prove our results. In Section III we present our algorithm to find a joint probability distribution of two input r.v.'s X and Y, with given marginal distributions, whose entropy is at most one bit away from the joint distribution of minimal entropy. We also present a worked out example to illustrate the behaviour of the algorithm in an intuitive way. The formal proofs of the correctness of the algorithm are spelled out in the successive Section III-B. In Section V we extend the algorithm presented in Section III to an arbitrary number of $k \geq 2$ input random variables. The entropy of the joint distribution produced by our algorithm is at most $\log k$ bits away from the minimum-entropy joint distribution of the k r.v.'s.

Throughout this paper, the logarithms are on base 2 unless specified otherwise.

II. Preliminary Results

To prove our results, we use ideas and techniques from majorization theory [39], a mathematical framework that has been proved to be very much useful in information theory (e.g., see [5], [6], [7], [8], [17], [23], [24], [48] and references therein). In this section we recall the notions and results that are relevant to our context.

Definition 1. *Given two probability distributions* $\mathbf{a} = (a_1, \ldots, a_n)$ *and* $\mathbf{b} = (b_1, \ldots, b_n)$ *with* $a_1 \geq \ldots \geq a_n \geq 0$ *and* $b_1 \geq \ldots \geq b_n \geq 0$, $\sum_{i=1}^n a_i = \sum_{i=1}^n b_i = 1$, *we say that* $\mathbf{a}$ *is* majorized *by* $\mathbf{b}$, *and write* $\mathbf{a} \preceq \mathbf{b}$, *if and only if* $\sum_{k=1}^i a_k \leq \sum_{k=1}^i b_k, \quad \text{for all} \;\; i = 1, \ldots, n.$

We assume that all the probability distributions we deal with have been ordered in non-increasing order. This assumption does not affect our results, since the quantities we compute (i.e., entropies) are invariant with respect to permutations of the components of the involved probability distributions. We also use the majorization relationship between vectors of unequal lengths, by properly padding the shorter one with the appropriate number of 0's at the end. The majorization relation $\preceq$ is a partial ordering on the $(n-1)$-dimensional simplex

$$\mathcal{P}_n = \{(p_1, \ldots, p_n) \;:\; \sum_{i=1}^n p_i = 1,\; p_1 \geq \ldots \geq p_n \geq 0\}$$

of all ordered probability vectors of n elements, that is, for each $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{P}_n$ it holds that

1) $\mathbf{x} \preceq \mathbf{x}$;
2) $\mathbf{x} \preceq \mathbf{y}$ and $\mathbf{y} \preceq \mathbf{z}$ implies $\mathbf{x} \preceq \mathbf{z}$;
3) $\mathbf{x} \preceq \mathbf{y}$ and $\mathbf{y} \preceq \mathbf{x}$ implies $\mathbf{x} = \mathbf{y}$.

It turns out that the partially ordered set $(\mathcal{P}_n, \preceq)$ is indeed a *lattice* [5],[4] i.e., for all $\mathbf{x}, \mathbf{y} \in \mathcal{P}_n$ there exists a unique *least upper bound* $\mathbf{x} \vee \mathbf{y}$ and a unique *greatest lower bound* $\mathbf{x} \wedge \mathbf{y}$.

[4]The same result was independently rediscovered in [12], see also [22] for a different proof.

We recall that the least upper bound $\mathbf{x} \vee \mathbf{y}$ is the vector in $\mathcal{P}_n$ such that

$$\mathbf{x} \preceq \mathbf{x} \vee \mathbf{y}, \ \mathbf{y} \preceq \mathbf{x} \vee \mathbf{y},$$

and for all $\mathbf{z} \in \mathcal{P}_n$ for which $\mathbf{x} \preceq \mathbf{z}, \ \mathbf{y} \preceq \mathbf{z}$ it holds that

$$\mathbf{x} \vee \mathbf{y} \preceq \mathbf{z}.$$

Analogously, the greatest lower bound $\mathbf{x} \wedge \mathbf{y}$ is the vector in $\mathcal{P}_n$ such that

$$\mathbf{x} \wedge \mathbf{y} \preceq \mathbf{x}, \ \mathbf{x} \wedge \mathbf{y} \preceq \mathbf{y},$$

and for all $\mathbf{z} \in \mathcal{P}_n$ for which $\mathbf{z} \preceq \mathbf{x}, \ \mathbf{z} \preceq \mathbf{y}$ it holds that

$$\mathbf{z} \preceq \mathbf{x} \wedge \mathbf{y}.$$

In the paper [5] the authors also gave a simple and efficient algorithm to explicitly compute $\mathbf{x} \vee \mathbf{y}$ and $\mathbf{x} \wedge \mathbf{y}$, given arbitrary vectors $\mathbf{x}, \mathbf{y} \in \mathcal{P}_n$. Due to the important role it will play in our main result, we recall how to compute the greatest lower bound.

Fact 1. [5] *Let $\mathbf{x} = (x_1, \ldots, x_n), \mathbf{y} = (y_1, \ldots, y_n) \in \mathcal{P}_n$ and let $\mathbf{z} = (z_1, \ldots, z_n) = \mathbf{x} \wedge \mathbf{y}$. Then, $z_1 = \min\{p_1, q_1\}$ and for each $i = 2, \ldots, n$, it holds that*

$$z_i = \min\Big\{\sum_{j=1}^{i} p_j, \sum_{j=1}^{i} q_j\Big\} - \sum_{j=1}^{i-1} z_j.$$

Equivalently, we have

$$\sum_{k=1}^{i} z_k = \min\Big\{\sum_{k=1}^{i} p_k, \sum_{k=1}^{i} q_k\Big\}.$$

Moreover, using $\sum_k z_k = \sum_k p_k = \sum_k q_k = 1$, we also have that for each $i = 1, \ldots, n$, it holds that

$$\sum_{k=i}^{n} z_k = \max\Big\{\sum_{k=i}^{n} p_k, \sum_{k=i}^{n} q_k\Big\}. \tag{1}$$

We also remind the important Schur-concavity property of the entropy function [39]:

For any $\mathbf{x}, \mathbf{y} \in \mathcal{P}_n$, $\mathbf{x} \preceq \mathbf{y}$ implies that $H(\mathbf{x}) \geq H(\mathbf{y})$, with equality if and only if $\mathbf{x} = \mathbf{y}$.

A notable strengthening of above fact has been proved in [24]. There, the authors prove that $\mathbf{x} \preceq \mathbf{y}$ implies

$$H(\mathbf{x}) \geq H(\mathbf{y}) + D(\mathbf{y}||\mathbf{x}), \tag{2}$$

where $D(\mathbf{y}||\mathbf{x})$ is the relative entropy between $\mathbf{x}$ and $\mathbf{y}$.

We also need the concept of *aggregation* (see [54] and [8]), and a result from [8], whose proof is repeated here to make the paper self-contained. Given $\mathbf{p} = (p_1, \ldots, p_n) \in \mathcal{P}_n$ and an integer $2 \leq m < n$, we say that $\mathbf{q} = (q_1, \ldots, q_m) \in \mathcal{P}_m$ is an *aggregation* of $\mathbf{p}$ if there is a partition of $\{1, \ldots, n\}$ into disjoint sets $I_1, \ldots, I_m$ such that $q_j = \sum_{i \in I_j} p_i$, for $j = 1, \ldots m$.

Lemma 1. [8] *Let $\mathbf{q} \in \mathcal{P}_m$ be* any *aggregation of $\mathbf{p} \in \mathcal{P}_n$. Then it holds that $\mathbf{p} \preceq \mathbf{q}$.*

Proof: We shall prove by induction on i that $\sum_{k=1}^{i} q_k \geq \sum_{k=1}^{i} p_k$. Because $\mathbf{q}$ is an aggregation of $\mathbf{p}$, we know that there exists $I_j \subseteq \{1, \ldots, n\}$ such that $1 \in I_j$. This implies that $q_1 \geq q_j \geq p_1$. Let us suppose that $\sum_{k=1}^{i-1} q_k \geq \sum_{k=1}^{i-1} p_k$. If there exist indices $j \geq i$ and $\ell \leq i$ such that $\ell \in I_j$, then $q_i \geq q_j \geq p_\ell \geq p_i$, implying $\sum_{k=1}^{i} q_k \geq \sum_{k=1}^{i} p_k$. Should it be otherwise, for each $j \geq i$ and $\ell \leq i$ it holds that $\ell \notin I_j$. Therefore, $\{1, \ldots, i\} \subseteq I_1 \cup \ldots \cup I_{i-1}$. This immediately gives $\sum_{k=1}^{i-1} q_k \geq \sum_{k=1}^{i} p_k$, from which we get $\sum_{k=1}^{i} q_k \geq \sum_{k=1}^{i} p_k$. ∎

Let us now discuss some first consequences of the above framework. Given two discrete random variables X and Y, with probability distributions $\mathbf{p} = (p_1, \ldots, p_n)$ and $\mathbf{q} = (q_1, \ldots, q_m)$, respectively, denote by $\mathcal{C}(\mathbf{p}, \mathbf{q})$ the set of all joint distributions of X and Y that have $\mathbf{p}$ and $\mathbf{q}$ as marginals. In the literature, elements of $\mathcal{C}(\mathbf{p}, \mathbf{q})$ are often called *couplings* of $\mathbf{p}$ and $\mathbf{q}$, and play an important role in many information theoretic problems, e.g, see [47]. For our purposes, each element in $\mathcal{C}(\mathbf{p}, \mathbf{q})$ can be seen as an $n \times m$ matrix $M = [m_{ij}] \in \mathbb{R}^{n \times m}$ such that its row-sums give the elements of $\mathbf{p} = (p_1, \ldots, p_n)$ and its column-sums give the elements of $\mathbf{q} = (q_1, \ldots, q_m)$, that is,

$$\mathcal{C}(\mathbf{p}, \mathbf{q}) = \Big\{\mathbf{M} = [m_{ij}] : \sum_j m_{ij} = p_i, \sum_i m_{ij} = q_j\Big\}. \tag{3}$$

Now, for any $\mathbf{M} \in \mathcal{C}(\mathbf{p}, \mathbf{q})$, let us write its elements in a $1 \times mn$ vector $\mathbf{m} \in \mathcal{P}_{mn}$, with its components ordered in non-increasing fashion. From (3) we obtain that both $\mathbf{p}$ and $\mathbf{q}$ are aggregations of *each* $\mathbf{m} \in \mathcal{P}_{mn}$ obtained from some $\mathbf{M} \in \mathcal{C}(\mathbf{p}, \mathbf{q})$. By Lemma 1, we get that[5]

$$\mathbf{m} \preceq \mathbf{p} \quad \text{and} \quad \mathbf{m} \preceq \mathbf{q}. \tag{4}$$

Recalling the definition and properties of the greatest lower bound of two vectors in $\mathcal{P}_{mn}$, we also obtain

$$\mathbf{m} \preceq \mathbf{p} \wedge \mathbf{q}. \tag{5}$$

From (5), and the Schur-concavity of the Shannon entropy, we also obtain that

$$H(\mathbf{m}) \geq H(\mathbf{p} \wedge \mathbf{q}).$$

Since, obviously, the entropy of $H(\mathbf{m})$ is equal to the entropy $H(\mathbf{M})$, where $\mathbf{M}$ is the matrix in $\mathcal{C}(\mathbf{p}, \mathbf{q})$ from which the vector $\mathbf{m}$ was obtained, we get the following result (a key one for our purposes).

Lemma 2. For any $\mathbf{p}$ *and* $\mathbf{q}$*, and* $\mathbf{M} \in \mathcal{C}(\mathbf{p}, \mathbf{q})$*, it holds that*

$$H(\mathbf{M}) \geq H(\mathbf{p} \wedge \mathbf{q}). \tag{6}$$

Lemma 2 obviously implies that the minimum-entropy coupling of $\mathbf{p}$ and $\mathbf{q}$ that we are seeking satisfies the inequality

$$\min_{\mathbf{N} \in \mathcal{C}(\mathbf{p}, \mathbf{q})} H(\mathbf{N}) \geq H(\mathbf{p} \wedge \mathbf{q}),$$

and it is one of the key results towards our algorithm to find an element $\mathbf{M} \in \mathcal{C}(\mathbf{p}, \mathbf{q})$ with entropy at most 1 bit larger than the entropy of the minimum entropy coupling, i.e., $H(\mathbf{M}) \leq OPT + 1$, where $OPT = \min_{\mathbf{N} \in \mathcal{C}(\mathbf{p}, \mathbf{q})} H(\mathbf{N})$.

[5]Recall that we use the majorization relationship between vectors of unequal lengths, by properly padding the shorter one with the appropriate number of 0's at the end. This trick does not affect our subsequent results, since we use the customary assumption that $0 \log 0 = 0$.

A. An interlude

Before describing our algorithm and its analysis, let us illustrate some consequences of Lemma 2 not directly aimed towards proving our main results, but nevertheless of some interest.

It is well known that for any joint distribution of the two r.v. X and Y it holds that

$$H(X,Y) \geq \max\{H(X), H(Y)\}. \tag{7}$$

Since $H(XY) = H(\mathbf{M})$, for some $\mathbf{M} \in \mathcal{C}(\mathbf{p}, \mathbf{q})$ (where $\mathbf{p}$ and $\mathbf{q}$ are the marginal distributions of X and Y, respectively), we can formulate the bound (7) in the following *equivalent* way:

for any $\mathbf{M} \in \mathcal{C}(\mathbf{p}, \mathbf{q})$ it holds that

$$H(\mathbf{M}) \geq \max\{H(\mathbf{p}), H(\mathbf{q})\}.$$

Lemma 2 allows us to strengthen the lower bound (7). Indeed, by the definition of the greatest lower bound $\mathbf{p} \wedge \mathbf{q}$ of probability distributions $\mathbf{p}$ and $\mathbf{q}$, it holds that $\mathbf{p} \wedge \mathbf{q} \preceq \mathbf{p}$ and $\mathbf{p} \wedge \mathbf{q} \preceq \mathbf{q}$, and therefore, by the Schur-concavity of the entropy function and Lemma 2 we get the improved lower bound

$$H(XY) = H(\mathbf{M}) \geq H(\mathbf{p} \wedge \mathbf{q}) \geq \max\{H(\mathbf{p}), H(\mathbf{q})\}. \tag{8}$$

Inequality (8) also allows us to improve on the classical upper bound on the mutual information given by $I(X;Y) \leq \min\{H(X), H(Y)\}$, since (8) implies

$$I(X;Y) \leq H(\mathbf{p}) + H(\mathbf{q}) - H(\mathbf{p} \wedge \mathbf{q}) \leq \min\{H(X), H(Y)\}. \tag{9}$$

The new bounds (8) and (9) are *strictly* better than the usual ones, whenever $\mathbf{p} \not\preceq \mathbf{q}$ and $\mathbf{q} \not\preceq \mathbf{p}$. Technically, one could improve them even more, by using the inequality $H(\mathbf{x}) \geq H(\mathbf{y}) + D(\mathbf{y}||\mathbf{x})$, whenever $\mathbf{x} \preceq \mathbf{y}$ [24]. However, in this paper we just need what we can get from the inequality $H(\mathbf{x}) \geq H(\mathbf{y})$, if $\mathbf{x} \preceq \mathbf{y}$ holds.

Inequalities (8) and (9) could be useful also in other contexts, when one needs to bound the joint entropy (or the mutual information) of two r.v.'s X and Y, and the only available knowledge is given by the marginal distributions of X and Y (and not their joint distribution). Let X and Y be two r.v.'s, where X is distributed according to $\mathbf{p}$ and Y according to $\mathbf{q}$, and let $H(X|Y)$ be the conditional entropy of X given Y. From (9) we get

$$\begin{aligned} H(X|Y) &= H(X) - I(X;Y) \\ &\geq H(X) - H(\mathbf{p}) - H(\mathbf{q}) + H(\mathbf{p} \wedge \mathbf{q}) \\ &= H(\mathbf{p} \wedge \mathbf{q}) - H(\mathbf{q}). \end{aligned}$$

The inequality $H(X|Y) \geq H(\mathbf{p} \wedge \mathbf{q}) - H(\mathbf{q})$ gives a lower bound on $H(X|Y)$ that does not depend on the joint distribution of X and Y. In particular, it also implies that if the probability distributions $\mathbf{p}$ and $\mathbf{q}$ of X and Y are such that $\mathbf{q} \not\preceq \mathbf{p}$, then the conditional entropy $H(X|Y)$ cannot be zero, *no matter* what the joint distribution of X and Y is. By the Fano inequality, one gets a lower bound of the error probability $\Pr\{X \neq Y\}$ that depends only on the "structure" of the probability distributions $\mathbf{p}$ and $\mathbf{q}$ of X and Y and not on the joint distribution of X and Y. Admittedly, this lower bound is weak, but the only fact that one could derive one that is independent from the joint distribution of X and Y seems novel and interesting to us.

Another possible application of the framework of Section II concerns the problem of sumset estimates for Shannon entropy [33], [50]. There, one wants to find upper and lower bounds on the entropy of $H(X+Y), H(X-Y)$ (and similar expressions), in terms of the individual entropies $H(X), H(Y)$. As an example, one could somewhat improve the trivial estimate $H(X) + H(Y) \geq H(X+Y)$ in the following way. Let us consider $X+Y$ and observe that the probability mass function of $X+Y$ is an aggregation of the pmf of the joint random variable (X,Y). Then, by Lemma 1 and formula (2), one immediately gets the inequality

$$\begin{aligned} H(X) + H(Y) \geq H(X,Y) \geq H(X+Y) \\ + D(X+Y||(X,Y)) \geq H(X+Y), \end{aligned} \tag{10}$$

where the last inequality is strict unless the pmf of $X+Y$ is equal to that of (X,Y). Similar improvements can be obtained for other expressions like $X-Y$. More in general, one has the following inequality that holds for any determinist function f and discrete r.v. Z:

$$H(Z) \geq H(f(Z)) + D(f(Z)||Z), \tag{11}$$

where one recovers (10) when $Z = (X,Y)$ and $f(X,Y) = X+Y$.

III. An Algorithm to Approximate $OPT = \min_{\mathbf{N} \in \mathcal{C}(\mathbf{p},\mathbf{q})} H(\mathbf{N})$.

In this section we present our main result, that is, an algorithm that from the input distributions $\mathbf{p}$ and $\mathbf{q}$, constructs a coupling $\mathbf{M} \in \mathcal{C}(\mathbf{p}, \mathbf{q})$ such that

$$H(\mathbf{M}) \leq H(\mathbf{p} \wedge \mathbf{q}) + 1. \tag{12}$$

Lemma 2 will imply our desired result, that is

$$H(\mathbf{M}) \leq \min_{\mathbf{N} \in \mathcal{C}(\mathbf{p},\mathbf{q})} H(\mathbf{N}) + 1.$$

The following lemma is technical in nature, but it turns out to be a very useful tool of our main algorithm.

Lemma 3. *Let $A[1 \dots k]$ be an array of k non-negative real numbers and z a positive real number such that $z \geq A[i]$ for each $i = 1, \dots, k$. For any $x \geq 0$ such that $x \leq z + \sum_{i=1}^{k} A[i]$ there exists a subset $I \subseteq \{1, \dots k\}$ and $0 \leq z^{(d)} \leq z$ such that*

$$z^{(d)} + \sum_{i \in I} A[i] = x.$$

Moreover, I and $z^{(d)}$ can be computed in linear time.

Proof:

If $\sum_{i=1}^{k} A[i] < x$, the desired result is given by setting $I = \{1, \dots, k\}$ and $z^{(d)} = x - \sum_{i=1}^{k} A[i]$ which is a positive number not larger than z, from the assumption that $z + \sum_{i=1}^{k} A[i] \geq x$. Note that the condition can be checked in linear time.

Let us now assume that $\sum_{i=1}^{k} A[i] \geq x$. Let j be the minimum index such that $\sum_{i=1}^{j} A[j] \geq x$. Then setting

$I = \{1, \ldots, j-1\}$ (if $j = 1$, we set $I = \emptyset$) and—using the assumption that $z \geq A[j]$—$z^{(d)} = x - \sum_{i=1}^{j} A[j]$ we have the desired result. Note that also in this case the index j can be found in linear time. ∎

As said before, Lemma 3 is an important technical tool of our main algorithm. Therefore, in **Algorithm 2** we give an efficient way to compute the value $z^{(d)}$ and the set of indices I.

Algorithm 1 The Min Entropy Joint Distribution Algorithm

MIN-ENTROPY-JOINT-DISTR($\mathbf{p}, \mathbf{q}$)
Input: prob. distributions $\mathbf{p} = (p_1, \ldots, p_n)$ and $\mathbf{q} = (q_1, \ldots, q_n)$
Output: An $n \times n$ matrix $\mathbf{M} = [m_{i\,j}]$ s.t. $\sum_j m_{i\,j} = p_i$ and $\sum_i m_{i\,j} = q_j$.
1: **for** $i = 1, \ldots, n$ and $j = 1, \ldots, n$ **set** $m_{i\,j} \leftarrow 0$
2: **if** $\mathbf{p} \neq \mathbf{q}$, let $i = \max\{j \mid p_j \neq q_j\}$; **if** $p_i < q_i$ **swap** $\mathbf{p} \leftrightarrow \mathbf{q}$
3: $\mathbf{z} = (z_1, \ldots, z_n) \leftarrow \mathbf{p} \wedge \mathbf{q}$
4: **for** $i = 1, \ldots, n$ **set** $m_{i\,i} \leftarrow z_i$
5: $i \leftarrow n$
6: **while** $i \geq 1$ **do**
7: **if** $\sum_{k=i}^{n} m_{k\,i} > q_i$ **then**
8: $(z_i^{(d)}, z_i^{(r)}, I) \leftarrow$ LEMMA3$(z_i, q_i, [m_{1\,i}, m_{2\,i}, \ldots, m_{n\,i}])$
9: $m_{i\,i} \leftarrow z_i^{(d)}$, $m_{i,i-1} \leftarrow z_i^{(r)}$
10: **for each** $k \notin I \cup \{i\}$ **do**
11: $m_{k\,i-1} \leftarrow m_{k\,i}$
12: $m_{k\,i} \leftarrow 0$
13: **if** $\sum_{k=i}^{n} m_{i\,k} > p_i$ **then**
14: $(z_i^{(d)}, z_i^{(r)}, I) \leftarrow$ LEMMA3$(z_i, p_i, [m_{i\,1}, m_{i\,2}, \ldots, m_{i\,n}])$
15: $m_{i\,i} \leftarrow z_i^{(d)}$, $m_{i-1,i} \leftarrow z_i^{(r)}$
16: **for each** $k \notin I \cup \{i\}$ **do**
17: $m_{i-1\,k} \leftarrow m_{i\,k}$
18: $m_{i\,k} \leftarrow 0$
19: $i \leftarrow i - 1$

Algorithm 2 The procedure implementing Lemma 3

LEMMA3($z, x, A[i \ldots j]$)
Input: reals $z > 0$, $x \geq 0$, and $A[i \ldots j] \geq 0$ s.t. $\sum_k A[k] + x \geq z$
Output: $z^{(d)}, z^{(r)} \geq 0$, and $I \subseteq \{i, i+1, \ldots, j\}$ s.t. $z^{(d)} + z^{(r)} = z$, and $z^{(d)} + \sum_{\ell \in I} A[\ell] = x$.
1: $k \leftarrow i$, $I \leftarrow \emptyset$, $sum \leftarrow 0$
2: **while** $k \leq j$ **and** $sum + A[k] < x$ **do**
3: $I \leftarrow I \cup \{k\}$, $sum \leftarrow sum + A[k]$, $k \leftarrow k + 1$
4: $z^{(d)} \leftarrow x - sum$, $z^{(r)} \leftarrow z - z^{(d)}$
5: **return** $(z^{(d)}, z^{(r)}, I)$

By padding the probability distributions with the appropriate number of 0's, we can assume that both $\mathbf{p}, \mathbf{q} \in \mathcal{P}_n$. We are now ready to present our main algorithm. The pseudocode is given in **Algorithm 1**. Since the description of **Algorithm 1** might look complicated, we see fit to illustrate and comment its behavior with a worked out example. The reader is advised to go through the content of Section III-A before reading the formal proofs of Section III-B.

*A. How **Algorithm 1** works: An informal description of its functioning and a numerical example*

At any point during the execution of the algorithm, we say that $\mathbf{q}$ is i-satisfied if the sum of the entries on columns $i, i+1, \ldots, n$ of the matrix the algorithm is constructing, is equal to $q_i + q_{i+1} + \cdots + q_n$ Analogously, we say that $\mathbf{p}$ is i-satisfied if the sum of the entries on rows $i, i+1, \ldots, n$ is equal to $p_i + p_{i+1} + \cdots + p_n$. Clearly, a matrix $\mathbf{M} \in \mathcal{C}(\mathbf{p}, \mathbf{q})$ if and only if it holds that both $\mathbf{p}$ and $\mathbf{q}$ are i-satisfied for each $i = 1, \ldots, n$.

Let $\mathbf{z}$ be the vector defined in **Fact 1**, and $\mathbf{M_z}$ be a matrix defined by setting $\mathbf{M_z}[i, i] = z_i$ and setting all the other entries to zero. The basic observation is that for the matrix $\mathbf{M_z}$ either $\mathbf{p}$ or $\mathbf{q}$ is i-satisfied, for each $i = 1, \ldots, n$, (but not necessarily both). In addition, every constraint which is not satisfied, coincides with an overflow, i.e., if for instance for some i we have that for $\mathbf{M_z}$ defined above $\mathbf{p}$ is not i-satisfied, it is necessarily the case that the sum of rows $i, i+1, \ldots, n$ of $\mathbf{M_z}$ is strictly greater than $p_i + p_{i+1} + \cdots + p_n$.

We can understand our algorithm as working on how to modify $\mathbf{M_z}$ in order to achieve i-satisfiability for both $\mathbf{p}$ and $\mathbf{q}$ for each $i = 1, \ldots, n$, by splitting *in at most two parts* each diagonal element. The algorithm processes the vector $\mathbf{z}$ from the smallest component z_n to the largest z_1. For $i = n, \ldots, 1$ it keeps z_i in the diagonal entry $\mathbf{M}[i, i]$ as long as both $\mathbf{p}$ and $\mathbf{q}$ are i-satisfied.

When, e.g., $\mathbf{q}$ is not i-satisfied, it must be necessarily overflowed, i.e., the sum of the components on the i-th column is larger than q_i. Then, the algorithm's action is equivalent to removing the surplus from the i-th column and place it onto the column $i-1$ so that $\mathbf{q}$ becomes i-satisfied and $\mathbf{p}$ remains i-satisfied, as the mass moved is still on the same rows.

This operation can be accomplished using Lemma 3, i.e., by selecting a subset of the non-zero components on column i together with $0 < z_i' < z_i$ so that their sum is equal to q_i. Keep this mass on column i and move the remaining components and the left over of z_i to column $i-1$. In this process only z_i gets split.

Analogously, when $\mathbf{p}$ is not i-satisfied, it must be necessarily overflowed, i.e., the sum of the components on the i-th row is larger than p_i. Then, the algorithm's action is equivalent to removing the surplus from the i-th row and place it onto the row $i-1$ so that $\mathbf{p}$ becomes i-satisfied and $\mathbf{q}$ remains i-satisfied, as the mass moved is still on the same columns.

This operation is again accomplished using Lemma 3: select a subset of the non-zero components on row i together with $0 < z_i' < z_i$ so that their sum is equal to p_i. Keep this mass on row i and move the remaining components and the left over of z_i to row $i-1$. Again in this process only z_i gets split.

Let us consider the following example: Let $n = 6$ and $\mathbf{p} = (0.4, 0.3, 0.15, 0.08, 0.04, 0.03)$ and $\mathbf{q} = (0.44, 0.18, 0.18, 0.15, 0.03, 0.02)$, be the two probability distributions for which we are seeking a coupling of minimum entropy. We have $\mathbf{z} = \mathbf{p} \wedge \mathbf{q} = (0.4, 0.22, 0.18, 0.13, 0.04, 0.03)$.

In the first iteration, we process the entry $(6, 6)$ containing z_6 (indicated in bold, below). In the matrix $\mathbf{M}_z$ (below) we have that $\mathbf{p}$ is 6-satisfied but q_6 is overflowed. Therefore, we split z_6 into $0.2 = q_6$ and 0.1 and leave the former as entry $m_{6\,6}$ and the make the latter be entry $m_{6\,5}$, obtaining the matrix $\mathbf{M}^{(6)}$ on the right. The underlined values represent the mass that has been moved from one column to the next one

on the left.

$$\mathrm{M}_{\mathbf{z}} = \begin{pmatrix} 0.4 & & & & & 0 \\ & 0.22 & & & & 0 \\ & & 0.18 & & & 0 \\ & & & 0.13 & & 0 \\ & & & & 0.04 & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{0.03} \end{pmatrix}$$

$$\mathrm{M}^{(\mathbf{6})} = \begin{pmatrix} 0.4 & & & & & 0 \\ & 0.22 & & & & 0 \\ & & 0.18 & & & 0 \\ & & & 0.13 & & 0 \\ & & & & 0.04 & 0 \\ 0 & 0 & 0 & 0 & \underline{0.01} & 0.02 \end{pmatrix}$$

Then, we process entry $(5,5)$ containing z_5 (indicated in bold, below). In $\mathbf{M}^{(6)}$ we now have that $\mathbf{p}$ is 5-satisfied but q_5 is overflowed. Therefore, we apply Lemma 3 to column 5, in order to find a split of z_5 and some of the other components of column 5 whose total sum is equal to q_5 and we move the remaining mass to column 4. Splitting z_5 into $0.2 + 0.2$ we obtain the matrix $\mathbf{M}^{(5)}$ on the right. The underlined values represent the mass that has been moved from one column to the next one on the left.

$$\mathrm{M}^{(\mathbf{6})} = \begin{pmatrix} 0.4 & & & & & 0 \\ & 0.22 & & & & 0 \\ & & 0.18 & & & 0 \\ & & & 0.13 & & 0 \\ & & & & \mathbf{0.04} & 0 \\ 0 & 0 & 0 & 0 & \underline{0.01} & 0.02 \end{pmatrix}$$

$$\mathrm{M}^{(\mathbf{5})} = \begin{pmatrix} 0.4 & & & & 0 & 0 \\ & 0.22 & & & 0 & 0 \\ & & 0.18 & & 0 & 0 \\ & & & 0.13 & 0 & 0 \\ 0 & 0 & 0 & \underline{0.02} & 0.02 & 0 \\ 0 & 0 & 0 & 0 & 0.01 & 0.02 \end{pmatrix}$$

Then, we process entry $(4,4)$ containing z_4 (indicated in bold, below). In $\mathbf{M}^{(5)}$ we now have that $\mathbf{q}$ is 4-satisfied but p_4 is overflowed. Therefore, we apply Lemma 3 to row 4, in order to find a split of z_4 such that one part is equal to p_4 and we move the remaining mass to row 3. Splitting z_4 into $0.8 + 0.5$ we obtain the matrix $\mathbf{M}^{(4)}$ on the right. The underlined values represent the mass that has been moved.

$$\mathrm{M}^{(\mathbf{5})} = \begin{pmatrix} 0.4 & & & & 0 & 0 \\ & 0.22 & & & 0 & 0 \\ & & 0.18 & & 0 & 0 \\ & & & \mathbf{0.13} & 0 & 0 \\ 0 & 0 & 0 & \underline{0.02} & 0.02 & 0 \\ 0 & 0 & 0 & 0 & 0.01 & 0.02 \end{pmatrix}$$

$$\mathrm{M}^{(\mathbf{4})} = \begin{pmatrix} 0.4 & & & 0 & 0 & 0 \\ & 0.22 & & 0 & 0 & 0 \\ & & 0.18 & \underline{0.05} & 0 & 0 \\ 0 & 0 & 0 & 0.08 & 0 & 0 \\ 0 & 0 & 0 & 0.02 & 0.02 & 0 \\ 0 & 0 & 0 & 0 & 0.01 & 0.02 \end{pmatrix}$$

Then, we process entry $(3,3)$ containing z_3 (indicated in bold, below). In $\mathbf{M}^{(4)}$ we now have that $\mathbf{q}$ is 3-satisfied but p_3 is overflowed. Therefore, we apply Lemma 3 to row 4, in order to find a split of z_3 and some of the other components of row 3 whose total sum is equal to p_3 and we move the remaining mass to row 2. If we split z_3 into $0.15 + 0.03$ we obtain the matrix $\mathbf{M}^{(3)}$ on the right. The underlined values represent the mass that has been moved from one row to the next one above.

$$\mathrm{M}^{(\mathbf{4})} = \begin{pmatrix} 0.4 & & & 0 & 0 & 0 \\ & 0.22 & & 0 & 0 & 0 \\ & & \mathbf{0.18} & \underline{0.05} & 0 & 0 \\ 0 & 0 & 0 & 0.08 & 0 & 0 \\ 0 & 0 & 0 & 0.02 & 0.02 & 0 \\ 0 & 0 & 0 & 0 & 0.01 & 0.02 \end{pmatrix}$$

$$\mathrm{M}^{(\mathbf{3})} = \begin{pmatrix} 0.4 & & 0 & 0 & 0 & 0 \\ & 0.22 & \underline{0.03} & \underline{0.05} & 0 & 0 \\ 0 & 0 & 0.15 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.08 & 0 & 0 \\ 0 & 0 & 0 & 0.02 & 0.02 & 0 \\ 0 & 0 & 0 & 0 & 0.01 & 0.02 \end{pmatrix}$$

Then, we process entry $(2,2)$ containing z_2 (indicated in bold, below). In $\mathbf{M}^{(3)}$ we now have that $\mathbf{p}$ is 2-satisfied but q_2 is overflowed. Therefore, we apply Lemma 3 to column 2, in order to find a split of z_2 and some of the other components of column 2 whose total sum is equal to q_2 and we move the remaining mass to column 1. If we split z_2 into $0.18 + 0.04$ we obtain the matrix $\mathbf{M}^{(2)}$ on the right. The underlined values represent the mass that has been moved from one column to the next one on the left.

$$\mathrm{M}^{(\mathbf{3})} = \begin{pmatrix} 0.4 & & 0 & 0 & 0 & 0 \\ & \mathbf{0.22} & \underline{0.03} & \underline{0.05} & 0 & 0 \\ 0 & 0 & 0.15 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.08 & 0 & 0 \\ 0 & 0 & 0 & 0.02 & 0.02 & 0 \\ 0 & 0 & 0 & 0 & 0.01 & 0.02 \end{pmatrix}$$

$$\mathrm{M}^{(\mathbf{2})} = \begin{pmatrix} 0.4 & 0 & 0 & 0 & 0 & 0 \\ \underline{0.04} & 0.18 & 0.03 & 0.05 & 0 & 0 \\ 0 & 0 & 0.15 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.08 & 0 & 0 \\ 0 & 0 & 0 & 0.02 & 0.02 & 0 \\ 0 & 0 & 0 & 0 & 0.01 & 0.02 \end{pmatrix}$$

Finally, we process entry $(1,1)$ containing z_1 (indicated in bold, below). In $\mathbf{M}^{(2)}$ we have that both $\mathbf{p}$ and $\mathbf{q}$ are 1-satisfied. Therefore we get the unmodified matrix $\mathbf{M}^{(1)}$ on the right which is our joint distribution. Notice that each component of $\mathbf{z}$ has been split at most into two parts. In particular only when z_i is processed the first time it might get split, while the other components (obtained by the previous subdivision of some other components of $\mathbf{z}$) might be relocated but not chunked again.

$$\mathrm{M}^{(\mathbf{2})} = \begin{pmatrix} \mathbf{0.4} & 0 & 0 & 0 & 0 & 0 \\ \underline{0.04} & 0.18 & 0.03 & 0.05 & 0 & 0 \\ 0 & 0 & 0.15 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.08 & 0 & 0 \\ 0 & 0 & 0 & 0.02 & 0.02 & 0 \\ 0 & 0 & 0 & 0 & 0.01 & 0.02 \end{pmatrix}$$

$$\mathrm{M}^{(\mathbf{1})} = \begin{pmatrix} 0.4 & 0 & 0 & 0 & 0 & 0 \\ 0.04 & 0.18 & 0.03 & 0.05 & 0 & 0 \\ 0 & 0 & 0.15 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.08 & 0 & 0 \\ 0 & 0 & 0 & 0.02 & 0.02 & 0 \\ 0 & 0 & 0 & 0 & 0.01 & 0.02 \end{pmatrix} = \mathrm{M}$$

*B. The proof of correctness of **Algorithm 1***

The following theorem shows the correctness of **Algorithm 1**. In particular, the equalities in (13), for the case $i = 1$, imply that the matrix built by the algorithm is a coupling of $\mathbf{p}$ and $\mathbf{q}$.

Theorem 1. *For each $i = n, n-1, \ldots, 1$ at the beginning of iteration i of the main* **while** *loop the following conditions hold*

1) for each $i' \leq i$ we have

$$\sum_{\ell=i'}^{n}\sum_{k=1}^{n} m_{\ell\,k} = \sum_{\ell=i'}^{n}\sum_{k=1}^{n} m_{k\,\ell} = \sum_{\ell=i'}^{n} z_\ell,$$

2) exactly one of the following holds

a) $\sum_{k=1}^{n} m_{k\,i} = q_i$ and $\sum_{k=1}^{n} m_{i\,ki} = p_i$
b) $\sum_{k=1}^{n} m_{k\,i} > q_i$ and $\sum_{k=1}^{n} m_{i\,ki} = p_i$
c) $\sum_{k=1}^{n} m_{k\,i} = q_i$ and $\sum_{k=1}^{n} m_{i\,ki} > p_i$ hence, at most one of the **if** *conditions is true.*

Moreover at the end of the iteration i, for each $i' \geq i$ it holds that

$$\sum_{k=1}^{n} m_{k\,i'} = q_{i'} \qquad \textit{and} \qquad \sum_{k=1}^{n} m_{i'\,k} = p_{i'}. \tag{13}$$

Proof: We prove the statement by reverse induction. For $i = n$, due to the initialization in lines 1 and 4 we have that for each $i' \leq n$ the only non-zero entry in row i' and in column i' is $m_{i'\,i'} = z_{i'}$ and 1 holds.

By definition we have $m_{n\,n} = z_n = \max\{p_n, q_n\} = p_n$, since, by the initialisation in line 2 we can assume $p_n \geq q_n$. Therefore, either $m_{n\,n} = p_n = q_n$ and 2a) holds; or $m_{n\,n} = p_n > q_n$ and 2b) holds. Thus, 2) holds.

Finally, if 2a) holds during the iteration n then no modification of the matrix entries is performed and at the end of the iteration equation (13) holds. Otherwise, as already observed, because of the initialization in line 2, we have $z_n = p_n > q_n$. Then, as a result of the call to procedure LEMMA3($z_n, q_n, A = \emptyset$) we will have $z_n^{(d)} = q_n$, $z_n^{(r)} = p_n - q_n$ and the modifications to the matrix entries $m_{n,n} = z_n^{(d)} = q_n$ and $m_{n,n-1} = z_n^{(r)} = p_n - q_n$, from which (13) holds at the end of the iteration as desired. This settles the induction base.

Let us now assume that the claims hold for iteration $i+1$ and prove it for iteration i.

1. By induction hypothesis, at the beginning of iteration $i+1$ for each $i' \leq i+1$, hence in particular for each $i' \leq i$ it holds that

$$\sum_{\ell=i'}^{n}\sum_{k=1}^{n} m_{\ell\,k} = \sum_{\ell=i'}^{n}\sum_{k=1}^{n} m_{k\,\ell} = \sum_{\ell=i'}^{n} z_\ell. \tag{14}$$

During iteration $i+1$ the only possible changes to entries of the matrix are either in rows i and $i+1$ (when the **if** at line 7 is satisfied) or in columns i and $i+1$ (when the **if** at line 13 is satisfied). Moreover, such modifications do not change the total probability mass in rows i and $i+1$ and the total probability mass in column i and $i+1$, i.e., the sums $\sum_{k=1}^{n}(m_{k\,i} + m_{k\,i+1})$ and $\sum_{k=1}^{n}(m_{i\,k} + m_{i+1\,k})$ remain unchanged during iteration $i+1$. It follows that at the beginning of iteration i equality (14) still holds for each $i' \leq i$. This settles the inductive steps for property 1.

2. By induction hypothesis from 1 with $i' = i < i+1$ we have

$$\sum_{\ell=i}^{n}\sum_{k=1}^{n} m_{\ell\,k} = \sum_{\ell=i}^{n}\sum_{k=1}^{n} m_{k\,\ell} = \sum_{\ell=i}^{n} z_\ell. \tag{15}$$

By induction hypothesis, we also have that for each $\ell = i+1, \ldots, n$,

$$\sum_{k=1}^{n} m_{\ell\,k} = p_\ell \quad \text{and} \quad \sum_{k=1}^{n} m_{k\,\ell} = q_\ell \tag{16}$$

From equations (15)-(16) together with (1) we get

$$\max\{\sum_{\ell=i}^{n} p_\ell, \sum_{\ell=i}^{n} q_\ell\} = \sum_{\ell=i}^{n} z_\ell = \sum_{\ell=i+1}^{n} p_\ell + \sum_{k=1}^{n} m_{i\,\ell} \tag{17}$$

and

$$\max\{\sum_{\ell=i}^{n} p_\ell, \sum_{\ell=i}^{n} q_\ell\} = \sum_{\ell=i}^{n} z_\ell = \sum_{\ell=i+1}^{n} q_\ell + \sum_{k=1}^{n} m_{\ell\,i}. \tag{18}$$

Therefore, (a) if $\sum_{\ell=i}^{n} z_\ell = \sum_{\ell=i}^{n} q_\ell = \sum_{\ell=i}^{n} p_\ell$ then from (17) we have $p_i = \sum_{k=1}^{n} m_{i\,k}$ and from (18) we have $q_i = \sum_{k=1}^{n} m_{k\,i}$.

(b) If $\sum_{\ell=i}^{n} z_\ell = \sum_{\ell=i}^{n} p_\ell > \sum_{\ell=i}^{n} q_\ell$ then from (17) we have $p_i = \sum_{k=1}^{n} m_{i\,k}$ and from (18) we have $q_i < \sum_{k=1}^{n} m_{k\,i}$.

(c) If $\sum_{\ell=i}^{n} z_\ell = \sum_{\ell=i}^{n} q_\ell > \sum_{\ell=i}^{n} p_\ell$ then from (17) we have $p_i < \sum_{k=1}^{n} m_{i\,k}$ and from (18) we have $q_i = \sum_{k=1}^{n} m_{k\,i}$.

Exactly one of these three cases is possible, which proves the induction step for 2).

Let us now prove the induction step for (13).

Assume first that during the iteration i case 2a) applies. From the previous point, this means that $\sum_{\ell=i}^{n} z_\ell = \sum_{\ell=i}^{n} q_\ell = \sum_{\ell=i}^{n} p_\ell$. Then, none of the two **if** compounds (lines 7-12 and lines 13-18) are executed and no matrix entry is changed in this iteration. As a result at the end of the iteration we have that, for each $i' > i$ the formula (13) holds by induction hypothesis. Moreover it also holds for $i' = i$ since, from 1) and 2a) we have

$$\sum_{k=1}^{n} m_{k\,i} \;=\; \sum_{\ell=i}^{n} z_\ell - \sum_{\ell=i+1}^{n}\sum_{k=1}^{n} m_{k\,\ell}$$

$$= \sum_{\ell=i}^{n} q_\ell - \sum_{\ell=i+1}^{n} q_\ell = q_i$$

and analogously

$$\begin{aligned}\sum_{k=1}^{n} m_{i\,k} &= \sum_{\ell=i}^{n} z_\ell - \sum_{\ell=i+1}^{n}\sum_{k=1}^{n} m_{\ell\,k} \\ &= \sum_{\ell=i}^{n} p_\ell - \sum_{\ell=i+1}^{n} p_\ell = p_i\end{aligned}$$

Assume now that during the iteration i case 2b) applies (the case 2c) can be dealt with symmetrically). Then, the **if** compound in lines 7-12 is executed. As a result, values $z_i^{(d)}, z_i^{(r)}$ and set $I \subseteq [n]$ are computed such that $z_i^{(d)} + z_i^{(r)} = z_i$ and $z_i^{(d)} + \sum_{k\in I} m_{k\,i} = q_i$. Before the assignments in line 9 and the execution of the **for** loop, we had that

$$\begin{aligned}\sum_{k=1}^{n} m_{k\,i} &= m_{i\,i} + \sum_{k\neq i} m_{k\,i} = z_i + \sum_{k\in I} m_{k\,i} + \sum_{k\notin I\cup\{i\}} m_{k\,i} \\ &= z_i^{(d)} + \sum_{k\in I} m_{k\,i} + z_i^{(r)} + \sum_{k\notin I\cup\{i\}} m_{k\,i}.\end{aligned}$$

After the assignments in line 9 and the execution of the **for** loop, the mass in the last two terms is moved to column $i-1$, but without changing the row. Therefore the row sums do not change and the column sums of column i and $i-1$ change so that $\sum_{k=1}^{n} m_{k\,i} = z_i^{(d)} + \sum_{k\in I} m_{k\,i} = q_i$ as desired.

Finally, it is possible that during the iteration i case 2c) applies. The analysis for this case is analogous to the one for the previous case, from which it can be easily obtained by symmetry swapping the roles of rows and columns, taking into account that we have to consider the result of the operations executed within the **if** compound in lines 13-18.

The proof is complete. ∎

C. The guaranteed additive gap of *Algorithm 1*

We are now ready to formally prove our main result.

Theorem 2. *For any $\mathbf{p}, \mathbf{q} \in \mathcal{P}_n$, Algorithm 1 outputs* in polynomial time *an $\mathbf{M} \in \mathcal{C}(\mathbf{p}, \mathbf{q})$ such that*

$$H(\mathbf{M}) \leq H(\mathbf{p} \wedge \mathbf{q}) + 1. \tag{19}$$

Proof: It is not hard to see that the values of all non-zero entry of the matrix $\mathbf{M}$ are initially set in line 4 and then in lines 9 and 15—in fact, the assignments in lines 11-12 and 17-18 have the effect of shifting by one column to the left or by one row up values that had been fixed at some point earlier in lines 9 and 15. Therefore, all the final non-zero entries of $\mathbf{M}$ can be partitioned into n pairs $z_j^{(r)}, z_j^{(d)}$ with $z_j^{(r)} + z_j^{(d)} = z_j$ for $j = 1, \ldots, n$. By using the standard assumption $0 \log \frac{1}{0} = 0$ and applying Jensen inequality we have

$$\begin{aligned}H(\mathbf{M}) &= \sum_{j=1}^{n} z_j^{(r)} \log \frac{1}{z_j^{(r)}} + z_j^{(d)} \log \frac{1}{z_j^{(d)}} \\ &= \sum_{j=1}^{n} \left[z_j \left(\frac{z_j^{(r)}}{z_j} \log \frac{1}{z_j^{(r)}} + \frac{z_j^{(d)}}{z_j} \log \frac{1}{z_j^{(d)}} \right) \right] \\ &\leq \sum_{j=1}^{n} z_j \log \frac{2}{z_j} = H(\mathbf{z}) + 1\end{aligned}$$

which concludes the proof of the bound on the additive gap guaranteed by Algorithm 1. Moreover, one can see that Algorithm 1 can be implemented so to run in $O(n^2)$ time. For the time complexity of the algorithm we observe the following easily verifiable fact:

- the initialization in line 1 takes $O(n^2)$;
- the condition in line 2 can be easily verified in $O(n)$ which is also the complexity of swapping $\mathbf{p}$ with $\mathbf{q}$, if needed;
- the vector $z = \mathbf{p} \wedge \mathbf{q}$ can be computed in $O(n)$, once the suffix sums $\sum_{j=k}^{n} p_j$, $\sum_{j=k}^{n} q_j$ $(k = 1, \ldots, n)$ have been precomputed (also doable in $O(n)$);
- the main **while** loop is executed n times and all the operations executed in an iteration are easily upper bounded by $O(n)$. The most expensive are the calls to the procedure **Lemma3**, and the **for**-loops. All these take $O(n)$. Therefore, the overall running time of the **while** loop is also $O(n^2)$.

Therefore we can conclude that the time complexity of Algorithm 1 is $O(n^2)$, hence polynomial time. ∎

D. Improving the time complexity

We note that the time complexity of Algorithm 1 can be improved if we build the coupling $\mathbf{M}$ in sparse form, i.e., as the set of values $\{(\mathbf{M}[i,j], (i,j)) \mid \mathbf{M}[i,j] \neq 0\}$ containing only the non-zero entries of $\mathbf{M}$ together with their coordinates.

We can keep the moved masses, i.e., the pieces $z_i^{(r)}$ that are iteratively moved from one column to the previous one, in line 11 (respectively, from one row to the previous one, in line 17) in a priority queue $\mathcal{Q}$. For each such element we store in the priority queue its row index (resp. column index) and its mass. With a standard implementation of a priority queue, we can then efficiently find the value of the minimum mass stored in constant time $O(1)$ (we refer to this operation as MIN($\mathcal{Q}$) and extract the minimum mass in time logarithmic in the number of elements stored in the priority queue (we refer to this operation as EXTRACTMIN($\mathcal{Q}$))[10]. Accordingly, procedure Lemma 3 amounts to iteratively extract from the priority queue the smallest mass as long as the queue is not empty and the sum of the masses extracted do not overcomes $\min\{p_i, q_i\}$. Whenever we split z_i, we insert $z_i^{(r)}$ into the priority queue (this operation can also be implemented to require time logarithmic in the size of the queue; we refer to it as INSERT($\mathcal{Q}, (z_i^{(r)}, i)$)).

At any time, the priority queue will contain $O(n)$ elements. Therefore, each insertion (INSERT) and extraction (EXTRACTMIN) from the priority queue takes $O(\log n)$ time. Moreover, since each element enters the queue at most once, the overall time of *all* insertion and extraction operations is upper bounded by $O(n \log n)$. The remaining part of the algorithm takes $O(n)$, apart from the possible initial sorting of the two distribution, adding another $O(n \log n)$ term. Therefore, the resulting implementation has complexity $O(n \log n)$.

We report in appendix a pseudocode of such implementation, where, for the sake of a clearer description, we use two

priority queues, $\mathcal{Q}^{(row)}, \mathcal{Q}^{(col)}$, storing masses moved among rows and masses moved among columns respectively.

IV. Extending the Results to other Entropy Measures

Our approach to prove entropic inequalities via majorization theory seems quite powerful. Indeed, it allows us to extend our results to different kind of entropies, with no additional effort. As an example, let us consider the order α Rényi entropy [45] of a probability distribution $\mathbf{p} = (p_1, \ldots, p_n)$, defined as

$$H_\alpha(\mathbf{p}) = \frac{1}{1-\alpha} \log \sum_{i=1}^{n} p_i^\alpha, \tag{20}$$

where $\alpha \in (0,1) \cup (\infty)$. It is well known that the Rényi entropy is Schur-concave, for all the values of the parameter α [25]. Therefore, we immediately have the analogous of Lemma 2 of Section II.

Lemma 4. For any $\mathbf{p}$ *and* $\mathbf{q}$*, for any* $\mathbf{M} \in \mathcal{C}(\mathbf{p}, \mathbf{q})$ *and* $\alpha \in (0,1) \cup (\infty)$*, it holds that*

$$H_\alpha(\mathbf{M}) \geq H_\alpha(\mathbf{p} \wedge \mathbf{q}). \tag{21}$$

We now prove the analogous of Theorem 2, that is

Theorem 3. *For any* $\mathbf{p}, \mathbf{q} \in \mathcal{P}_n$*, Algorithm 1 outputs* in polynomial time *an* $\mathbf{M} \in \mathcal{C}(\mathbf{p}, \mathbf{q})$ *such that*

$$H_\alpha(\mathbf{M}) \leq H_\alpha(\mathbf{p} \wedge \mathbf{q}) + 1 \leq \min_{\mathbf{N} \in \mathcal{C}(\mathbf{p}, \mathbf{q})} H_\alpha(\mathbf{N}) + 1. \tag{22}$$

Proof: Let $\mathbf{M}$ be the matrix constructed by our Algorithm 1, and let $\alpha \in (0,1)$. Proceedings as in Theorem 2 (and with the same notations), we have:

$$\begin{aligned}
H_\alpha(\mathbf{M}) &= \frac{1}{1-\alpha} \log \sum_{j=1}^{n} \left[\left(z_j^{(r)}\right)^\alpha + \left(z_j^{(d)}\right)^\alpha \right] \\
&= \frac{1}{1-\alpha} \log \sum_{j=1}^{n} 2 \left[\frac{1}{2} \left(z_j^{(r)}\right)^\alpha + \frac{1}{2} \left(z_j^{(d)}\right)^\alpha \right] \\
&\leq \frac{1}{1-\alpha} \log \sum_{j=1}^{n} 2 \left(\frac{1}{2} z_j^{(r)} + \frac{1}{2} z_j^{(d)} \right)^\alpha \\
&\qquad \text{(by the Jensen inequality applied to } x^\alpha) \\
&= \frac{1}{1-\alpha} \log \sum_{j=1}^{n} 2 \left(\frac{z_j}{2} \right)^\alpha \quad (\text{since } z_j^{(r)} + z_j^{(d)} = z_j) \\
&= H_\alpha(\mathbf{z}) + 1 = H_\alpha(\mathbf{p} \wedge \mathbf{q}) + 1.
\end{aligned}$$

The proof for the case $\alpha \in (1, \infty)$ is the same, by noticing that for $\alpha > 1$ the Jensen inequality goes into the opposite direction and that $1/(1-\alpha) < 0$. ∎

One can extend our results also to other entropies, like the Tsallis entropy [51], using its Schur-concavity property proved in [20]. The mathematical details can be easily worked out by the motivated reader.

V. An Extension to Multivariate Distributions

In this section we will show how the algorithm Min-Entropy-Joint-Distr can be used to guarantee an additive gap at most $\log k$ for the problem of minimizing the entropy of a joint distribution, with marginals equal to k given input distributions, for any $k \geq 2$.

In what follows, for the ease of the description, we shall assume that $k = 2^\kappa$ for some integer $\kappa \geq 1$, i.e., k is a power of 2. A pictorial perspective on the algorithm's behaviour is to imagine that the input distributions are in the leaves of a complete binary tree with $k = 2^\kappa$ leaves. Each internal node ν of the tree contains the joint distribution of the distributions in the leaves of the subtree rooted at ν. Such a distribution is computed by applying the algorithm Min-Entropy-Joint-Distr to the distributions in the children of ν.

The algorithm builds such a tree starting from the leaves. Thus, the joint distribution of all the input distributions will be given by the distribution computed at the root of the tree.

Let us first see how in the case of four input distributions, our algorithms does indeed guarantee that the final matrix is a joint distribution with marginals equal to the input distributions. An high-level pictorial way to describe how our algorithm operates is given in Figure 1.

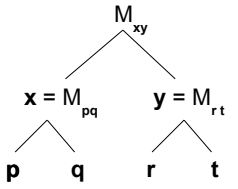

Fig. 1. The binary tree representing the process of producing the joint distribution for input probability vectors $\mathbf{p}, \mathbf{q}, \mathbf{r}, \mathbf{t}$.

Given probability distributions $\mathbf{p}$ and $\mathbf{q}$, of dimension $n_\mathbf{p}$ and $n_\mathbf{q}$, respectively, the algorithm Min-Entropy-Joint-Distr described in Section III produces a matrix $M_{\mathbf{p}\,\mathbf{q}}$ such that for fixed $\overline{i}$ and $\overline{j}$ it holds that

$$\sum_j M_{\mathbf{p}\,\mathbf{q}}[\overline{i}, j] = p_{\overline{i}} \qquad \text{and} \qquad \sum_i M_{\mathbf{p}\,\mathbf{q}}[i, \overline{j}] = q_{\overline{j}}. \tag{23}$$

We also have that, for each $i_1 \neq i_2$ the set of entries of $M_{\mathbf{p}\,\mathbf{q}}$ whose sum is p_{i_1} is disjoint from the set of entries of $M_{\mathbf{p}\,\mathbf{q}}$ whose sum is p_{i_2}. Analogously, for each $j_1 \neq j_2$ the set of entries of $M_{\mathbf{p}\,\mathbf{q}}$ whose sum is q_{j_1} is disjoint from the set of entries of $M_{\mathbf{p}\,\mathbf{q}}$ whose sum is q_{j_2}.

Let us define the probability distribution $\mathbf{x} = (x_1, x_2, \ldots)$ whose components are all and only the non-zero entries of $M_{\mathbf{p}\,\mathbf{q}}$ sorted in non-increasing order. Let $n_\mathbf{x}$ denote the number of components of $\mathbf{x}$. For each $a = 1, \ldots n_\mathbf{x}$ let us fix a one-one mapping $a \leftrightarrow (i, j)$ recording the fact that $x_a = M_{\mathbf{p}\,\mathbf{q}}[i, j]$.

Consider now another pair of distributions $\mathbf{r}, \mathbf{t}$, of dimension $n_\mathbf{r}$ and $n_\mathbf{t}$, respectively. Applying Algorithm Min-Entropy-Joint-Distr to $\mathbf{r}, \mathbf{t}$ we obtain a matrix $M_{\mathbf{r}\,\mathbf{t}}$ such that for fixed $\overline{i'}$ and $\overline{j'}$ it holds that

$$\sum_{j'} M_{\mathbf{r}\,\mathbf{t}}[\overline{i'}, j'] = p_{\overline{i'}} \qquad \text{and} \qquad \sum_{i'} M_{\mathbf{r}\,\mathbf{t}}[i', \overline{j'}] = q_{\overline{j'}}.$$

As before, for each $i'_1 \neq i'_2$ the set of entries of $M_{\mathbf{rt}}$ whose sum is $p_{i'_1}$ is disjoint from the set of entries of $M_{\mathbf{rt}}$ whose sum is $p_{i'_2}$. Also, for each $j'_1 \neq j'_2$ the set of entries of $M_{\mathbf{rt}}$ whose sum is $q_{j'_1}$ is disjoint from the set of entries of $M_{\mathbf{rt}}$ whose sum is $q_{j'_2}$.

Let $\mathbf{y} = (y_1, y_2, \ldots)$ be the probability distribution whose components are all and only the non-zero entries of $M_{\mathbf{rt}}$ sorted in non-increasing order. Let $n_{\mathbf{y}}$ denote the number of components of $\mathbf{y}$. For each $b = 1, \ldots n_{\mathbf{y}}$ let us fix a one-one mapping $b \leftrightarrow (i', j')$ recording the fact that $y_b = M_{\mathbf{rt}}[i', j']$.

If we now apply algorithm MIN-ENTROPY-JOINT-DISTR on the distributions $\mathbf{x}$ and $\mathbf{y}$ we get a matrix $M_{\mathbf{xy}}$ such that for fixed $\overline{a}$ and $\overline{b}$ it holds that

$$\sum_b M_{\mathbf{xy}}[\overline{a}, b] = x_{\overline{a}} \qquad \text{and} \qquad \sum_a M_{\mathbf{xy}}[a, \overline{b}] = y_{\overline{b}}. \tag{24}$$

Let us now define a new 4-dimensional array $\tilde{M}[i, j, i', j']$ by stipulating that for each $i \in [n_{\mathbf{p}}]$, $j \in [n_{\mathbf{q}}]$, $i' \in [n_{\mathbf{r}}]$, $j' \in [n_{\mathbf{t}}]$, the following equalities hold

$$\tilde{M}[i, j, i', j'] = \begin{cases} M_{\mathbf{xy}}[a, b] & \text{if there exist } a, b \text{ s.t.} \\ & \quad a \leftrightarrow (i, j), b \leftrightarrow (i', j') \\ 0 & \text{otherwise.} \end{cases}$$

Then, applying the properties above, for each $\overline{i} \in [n_{\mathbf{p}}]$, we have that

$$\sum_{j \in [n_{\mathbf{q}}]} \sum_{i' \in [n_{\mathbf{r}}]} \sum_{j' \in [n_{\mathbf{t}}]} \tilde{M}[\overline{i}, j, i', j'] \tag{25}$$

$$= \sum_{\substack{(\overline{i}, j) | \\ \exists a, a \leftrightarrow (\overline{i}, j)}} \sum_{\substack{(i', j') | \\ \exists b, b \leftrightarrow (i', j')}} \tilde{M}[\overline{i}, j, i', j'] \tag{26}$$

$$= \sum_{a | \exists j, a \leftrightarrow (\overline{i}, j)} \sum_{b \in [n_{\mathbf{y}}]} M_{\mathbf{xy}}[a, b] \tag{27}$$

$$= \sum_{a | \exists j, a \leftrightarrow (\overline{i}, j)} x_a \tag{28}$$

$$= \sum_{j | M_{\mathbf{pq}}[\overline{i}, j] \neq 0} M_{\mathbf{pq}}[\overline{i}, j] = \sum_{j \in [n_{\mathbf{q}}]} M_{\mathbf{pq}}[\overline{i}, j] = p_{\overline{i}}, \tag{29}$$

where the equality in (26) follows by restricting the sum over the non-zero entries of $\tilde{M}$; (27) follows by the definition of $\tilde{M}$; (28) follows by (24); the first part of (29) follows by the fact that the components of $\mathbf{x}$ coincide with non zero entries of $M_{\mathbf{pq}}$; the first equality in (29) follows since we are adding to the previous term only entries $M_{\mathbf{pq}}[\overline{i}, j] = 0$; finally the last equality follows from (23).

Proceeding in the same way we can show that for each $\overline{j}, \overline{i'}, \overline{j'}$ we have

$$\sum_{i \in [n_{\mathbf{p}}]} \sum_{i' \in [n_{\mathbf{r}}]} \sum_{j' \in [n_{\mathbf{t}}]} \tilde{M}[i, \overline{j}, i', j'] = q_{\overline{j}} \tag{30}$$

$$\sum_{i \in [n_{\mathbf{p}}]} \sum_{j \in [n_{\mathbf{q}}]} \sum_{j' \in [n_{\mathbf{r}}]} \tilde{M}[i, j, \overline{i'}, j'] = r_{\overline{i'}} \tag{31}$$

$$\sum_{i \in [n_{\mathbf{p}}]} \sum_{j \in [n_{\mathbf{q}}]} \sum_{i' \in [n_{\mathbf{r}}]} \tilde{M}[i, j, i', \overline{j'}] = t_{\overline{j}}, \tag{32}$$

hence concluding that $\tilde{M}$ is a joint distribution with marginals equal to $\mathbf{p}, \mathbf{q}, \mathbf{r}, \mathbf{t}$, as desired.

Completing the argument. We can now inductively extend the above argument to the case of more distributions. Assume that we have produced an array $\mathbf{M}_{\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(r)}}$ which is a joint distribution with marginals equal to $\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(r)}$, of dimension $n_1, \ldots, n_r$ respectively. Analogously, let us assume that we have produced $\mathbf{M}_{\mathbf{q}^{(1)}, \ldots, \mathbf{q}^{(s)}}$ which is a joint distribution with marginals equal to $\mathbf{q}^{(1)}, \ldots, \mathbf{q}^{(s)}$, of dimension $m_1, \ldots, m_s$ respectively. This means that for each $\ell = 1, \ldots, r$ and $1 \leq i \leq n_\ell$ and for each $\ell' = 1, \ldots, r$ and $1 \leq j \leq n_{\ell'}$ we have that

$$\sum_{i_1, \ldots i_{\ell-1}, i_{\ell+1}, \ldots, r} \mathbf{M}_{\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(r)}}[i_1, \ldots, i_{\ell-1}, i, i_{\ell+1}, \ldots, i_r] = p_i^{(\ell)}$$

$$\sum_{j_1, \ldots j_{\ell'-1}, i_{\ell'+1}, \ldots, s} \mathbf{M}_{\mathbf{q}^{(1)}, \ldots, \mathbf{q}^{(r)}}[j_1, \ldots, j_{\ell'-1}, j, j_{\ell'+1}, \ldots, j_s] = q_j^{(\ell')}.$$

Proceeding as before, let us define the probability distribution $\mathbf{x}$ whose components are all and only the non-zero entries of $M_{\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(r)}}$ sorted in non-increasing order. Let $n_{\mathbf{x}}$ denote the number of components of $\mathbf{x}$. For each $a = 1, \ldots n_{\mathbf{x}}$ let us fix a one-one mapping $a \leftrightarrow (i_1, \ldots, i_r)$ recording the fact that $x_a = M_{\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(r)}}[i_1, \ldots, i_r]$.

Let $\mathbf{y}$ be the probability distribution whose components are all and only the non-zero entries of $M_{\mathbf{q}^{(1)}, \ldots, \mathbf{q}^{(s)}}$ sorted in non-increasing order. Let $n_{\mathbf{y}}$ denote the number of components of $\mathbf{y}$. For each $b = 1, \ldots n_{\mathbf{y}}$ let us fix a one-one mapping $b \leftrightarrow (j_1, \ldots, j_s)$ recording the fact that $y_b = M_{\mathbf{q}^{(1)}, \ldots, \mathbf{q}^{(s)}}[j_1, \ldots, j_s]$.

Applying algorithm MIN-ENTROPY-JOINT-DISTR on the distributions $\mathbf{x}$ and $\mathbf{y}$ we get a matrix $M_{\mathbf{xy}}$ such that for fixed $\overline{k}$ and $\overline{\ell}$

$$\sum_b M_{\mathbf{xy}}[\overline{a}, b] = x_{\overline{a}} \qquad \text{and} \qquad \sum_a M_{\mathbf{xy}}[a, \overline{b}] = y_{\overline{b}}. \tag{33}$$

Therefore, we can define a new $r + s$-dimensional array $\tilde{M}[i_1, \ldots, i_r, j_1, \ldots, j_s]$ by stipulating that for each $i_1, \ldots, i_r$ such that $i_\ell \in [n_k]$ for $\ell = 1, \ldots, r$ and for any $j_1, \ldots, j_s$ such that $j_{\ell'} \in [m_{\ell'}]$ for $\ell' = 1, \ldots, s$,

$$\tilde{M}[i_1, \ldots, i_r, j_1, \ldots, j_s] = \begin{cases} M_{\mathbf{xy}}[a, b] & \text{if there are } a, b \text{ s.t.} \\ & \quad a \leftrightarrow (i_1, \ldots, i_r), \\ & \quad\quad b \leftrightarrow (j_1, \ldots, j_s) \\ 0 & \text{otherwise.} \end{cases}$$

It is not hard to see that proceeding like in (26)-(29) one can show that $\tilde{M}$ is indeed a joint distribution with marginals equal to $\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(r)}, \mathbf{q}(1), \ldots, \mathbf{q}^{(s)}$.

A. The pseudocode: Algorithm 3

Algorithm 3 shows the pseudocode for our procedure. We denote by $m^{(i-j)}$ the non-zero components of the distribution that our algorithm builds as joint distribution of $\mathbf{p}^{(i)}, \mathbf{p}^{(i+1)}, \ldots, \mathbf{p}^{(j)}$.

The vector $Ind^{(i-j)}$ is used to record for each component $\mathbf{m}^{(i-j)}[w]$ the indices of the component of the joint probability distribution of $\mathbf{p}^{(i)}, \dots, \mathbf{p}^{(j)}$ which coincides with $\mathbf{m}^{(i-j)}[w]$. With respect to the description above $Ind^{(i-j)}$ is used to record the one-one mapping between the elements of $\mathbf{m}^{(i-j)}$ and the non-zero elements of the joint distribution of $\mathbf{p}^{(i)}, \mathbf{p}^{(i+1)}, \dots, \mathbf{p}^{(j)}$. Therefore, in accordance to the above arguments, after the execution of line 17, for $w = 1, \dots, |\mathbf{m}^{(i-j)}|$, if, e.g., $Ind^{(i-j)}[w] = \langle s_i[w], s_{i+1}[w], \dots, s_j[w] \rangle$ it means that setting $M^{(i-j)}[s_i[w], s_{i+1}[w], \dots, s_j[w]] \leftarrow \mathbf{m}^{(i-j)}[w]$ and setting the remaining components of $M^{(i-j)}$ to zero, the array $M^{(i-j)}$ is a joint distribution matrix for $\mathbf{p}^{(i)}, \dots, \mathbf{p}^{(j)}$ whose non-zero components are equal to the components of $\mathbf{m}^{(i-j)}$. Hence, in particular, we have that $H(M^{(i-j)}) = H(\mathbf{m}^{(i-j)})$.

Note that the algorithm explicitly uses this correspondence only for the final array $M^{(1-k)}$ representing the joint distribution of all input distributions. Based on the above discussion the correctness of the algorithm can be easily verified.

B. The additive gap guaranteed by K-MIN-ENTROPY-JOINT-DISTRIBUTION

In this section we will prove that the entropy of the joint distribution output by the algorithm guarantees an additive gap at most $\log k$.

We will prepare some definitions and lemmas which will be key tools for proving the approximation guarantee of our algorithm. The proof of these technical lemmas is deferred to the next section.

Let us define the following:

Definition 2. *For any $\mathbf{p} = (p_1, \dots, p_n) \in \mathcal{P}_n$ we denote by $\mathsf{half}(\mathbf{p})$ the distribution $(\frac{p_1}{2}, \frac{p_1}{2}, \frac{p_2}{2}, \frac{p_2}{2}, \dots, \frac{p_n}{2}, \frac{p_n}{2})$ obtained by splitting each component of $\mathbf{p}$ into two identical halves.*

For any $i \geq 2$, let us also define $\mathsf{half}^{(i)}(\mathbf{p}) = \mathsf{half}(\mathsf{half}^{(i-1)}(\mathbf{p}))$, where $\mathsf{half}^{(1)}(\mathbf{p}) = \mathsf{half}(\mathbf{p})$ and $\mathsf{half}^{(0)}(\mathbf{p}) = \mathbf{p}$.

We will employ the following two technical lemmas whose proofs are in the next section.

Lemma 5. *For any $\mathbf{p} \preceq \mathbf{q}$ we have also $\mathsf{half}(\mathbf{p}) \preceq \mathsf{half}(\mathbf{q})$*

Lemma 6. *For any pair of distributions $\mathbf{p}, \mathbf{q} \in \mathcal{P}_n$. and any $i \geq 0$, It holds that*

$$\mathsf{half}^{(i)}(\mathbf{p} \wedge \mathbf{q}) \preceq \mathsf{half}^{(i)}(\mathbf{p}) \wedge \mathsf{half}^{(i)}(\mathbf{q}).$$

Theorem 4. *For each $\ell = 0, 1, \dots \kappa$ and $s = 0, 1, 2, \dots, k/2^\ell - 1$ let $i = i(\ell, s) = s \cdot 2^\ell + 1$ and $j = j(\ell, s) = (s+1) \cdot 2^\ell = i + 2^\ell - 1$. Then, we have*

$$\mathsf{half}^{(\ell)}(\mathbf{p}^{(i)} \wedge \mathbf{p}^{(i+1)} \wedge \cdots \wedge \mathbf{p}^{(j)}) \preceq \mathbf{m}^{(i-j)}.$$

Proof: The proof is by induction on ℓ. The base case follows by definition of the operator $\mathsf{half}^{(\ell)}$ and the fact that the algorithm sets $\mathbf{m}^{(i-i)} = \mathbf{p}^{(i)}$, for each i hence in particular $\mathbf{m}^{(i-i)} = \mathbf{p}^{(i)} = \mathsf{half}^{(0)}(\mathbf{p}^{(i)})$, which proves the desired inequality.

Algorithm 3 The Min Entropy Joint Distribution Algorithm for $k > 2$ distributions

K-MIN-ENTROPY-JOINT-DISTRIBUTION($\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \dots, \mathbf{p}^{(k)}$)
Input: prob. distributions $\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \dots, \mathbf{p}^{(k)}$, with $k = 2^\kappa$
Output: A k-dimensional array $\mathbf{M} = [m_{i_1, i_2, \dots, i_k}]$ s.t. $\sum_{i_1, \dots, i_{j-1}, i_{j+1}, \dots, i_k} m_{i_1, \dots, i_{j-1}, t, i_{j+1}, \dots, i_k} = p_t^{(j)}$ for each $j = 1, \dots, k$ and each t.

1: **for** $i = 1$ **to** k **do**
2: **for** $j = 1$ **to** n **do**
3: **set** $\mathbf{m}^{(i-i)}[j] = \mathbf{p}_j^{(i)}$ and $Ind^{(i-i)}[j] = \langle j \rangle$ {$Ind^{(i-i)}[j]$ is a vector of indices}
4: **for** $i = 1, \dots, k$ **permute the components of** $\mathbf{m}^{(i-i)}$ and $Ind^{(i-i)}$ using the permutation that sorts $\mathbf{m}^{(i-i)}$ in non-increasing order
5: **for** $\ell = 1$ **to** κ **do**
6: $i \leftarrow 1$, $j \leftarrow 2^\ell$
7: **while** $j \leq k$ **do**
8: $j_1 \leftarrow i + 2^{\ell-1} - 1$, $j_2 = j_1 + 1$
9: $M \leftarrow$ MIN-ENTROPY-JOINT-DISTR($\mathbf{m}^{(i-j_1)}, \mathbf{m}^{(j_2-j)}$)
10: $w \leftarrow 1$
11: **for** $s = 1$ **to** $|\mathbf{m}^{(i-j_1)}|$ **do**
12: **for** $t = 1$ **to** $|\mathbf{m}^{(j_2-j)}|$ **do**
13: **if** $M[s, t] \neq 0$ **then**
14: $\mathbf{m}^{(i-j)}[w] \leftarrow M[s, t]$
15: $Ind^{(i-j)}[w] \leftarrow Ind^{(i-j_1)}[s] \odot Ind^{(i-j_1)}[t]$ {$\odot$ denotes the concatenation of vectors}
16: $w \leftarrow w + 1$
17: **permute the components of** $\mathbf{m}^{(i-j)}$ and $Ind^{(i-j)}$ using the permutation that sorts $\mathbf{m}^{(i-j)}$ in non-increasing order
18: $i \leftarrow j + 1$, $j \leftarrow i + 2^\ell - 1$
19: **set** $M[i_1, i_2, \dots, i_k] = 0$ for each $i_1, i_2, \dots, i_k$.
20: **for** $j = 1$ **to** $|\mathbf{m}^{(1-k)}|$ **do**
21: $M[Ind^{(1-k)}[j]] \leftarrow \mathbf{m}^{(1-k)}[j]$
22: **return** M

We now prove the induction step. Let $\ell > 0$. It is enough to consider only the case $s = 0$, since the other cases are perfectly analogous.

Therefore, $i = 1$ and $j = 2^\ell$. Using the notation employed in the pseudocode, let $j_1 = 2^{\ell-1}$, $j_2 = 2^{\ell-1} + 1$. By induction hypothesis we can assume that

$$\mathsf{half}^{(\ell-1)}(\mathbf{p}^{(i)} \wedge \mathbf{p}^{(i+1)} \wedge \cdots \wedge \mathbf{p}^{(j_1)}) \preceq \mathbf{m}^{(i-j_1)} \tag{34}$$

$$\mathsf{half}^{(\ell-1)}(\mathbf{p}^{(j_2)} \wedge \mathbf{p}^{(j_2+1)} \wedge \cdots \wedge \mathbf{p}^{(j)}) \preceq \mathbf{m}^{(j_2-j)}. \tag{35}$$

It follows that

$$\mathsf{half}^{(\ell)}\left(\bigwedge_{\iota=i}^{j} \mathbf{p}^{(\iota)}\right) =$$

$$= \mathsf{half}^{(\ell)}\left(\left(\bigwedge_{\iota=i}^{j_1} \mathbf{p}^{(\iota)}\right) \wedge \left(\bigwedge_{\iota=j_2}^{j} \mathbf{p}^{(\iota)}\right)\right) \tag{36}$$

$$= \mathsf{half}\left(\mathsf{half}^{(\ell-1)}\left(\left(\bigwedge_{\iota=i}^{j_1} \mathbf{p}^{(\iota)}\right) \wedge \left(\bigwedge_{\iota=j_2}^{j} \mathbf{p}^{(\iota)}\right)\right)\right) \tag{37}$$

$$\preceq \mathsf{half}\left(\mathsf{half}^{(\ell-1)}\left(\bigwedge_{\iota=i}^{j_1} \mathbf{p}^{(\iota)}\right) \wedge \mathsf{half}^{(\ell-1)}\left(\bigwedge_{\iota=j_2}^{j} \mathbf{p}^{(\iota)}\right)\right) \tag{38}$$

$$\preceq \mathsf{half}\left(\mathbf{m}^{(i-j_1)} \wedge \mathbf{m}^{(j_2-j)}\right) \tag{39}$$

$$\preceq \quad \mathbf{m}^{(i-j)} \tag{40}$$

where

- (37) follows from (36) by the definition of the operator half;
- (38) follows from (37) by Lemma 6;
- (39) follows from (38) by the induction hypotheses (34)-(35) ;
- (40) follows from (39) by observing that the components of $\mathbf{m}^{(i-j)}$ coincide with the components of the array M output by algorithm MIN-ENTROPY-JOINT-DISTRIBUTION executed on the distributions $\mathbf{m}^{(i-j_1)}$ and $\mathbf{m}^{(j_2-j)}$. Let $\mathbf{z} = \mathbf{m}^{(i-j_1)} \wedge \mathbf{m}^{(j_2-j)}$ and $|\mathbf{z}|$ denote the number of components of $\mathbf{z}$. By the analysis presented in the previous section we have that we can partition the components of M (equivalently, the components of $\mathbf{m}^{(i-j)}$) into subsets $M_1, M_2, \ldots, M_{|\mathbf{z}|}$ such that
  - $1 \leq |M_i| \leq 2$
  - for each $i = 1, \ldots, |\mathbf{z}|$, it holds that $\sum_{x \in M_i} x = z_i$;

  Therefore—assuming, w.l.o.g., that the components of $\mathbf{m}^{(i-j)}$ are reordered such that those in M_i immediately precede those in M_{i+1}—we have $\mathsf{half}(\mathbf{z}) = \mathbf{m}^{(i-j)}P$ where $P = [p_{i\,j}]$ is a doubly stochastic matrix defined by

$$p_{i\,j} = \begin{cases} \frac{1}{2} & \text{if } (i \text{ is odd } \textbf{and } j \in \{i, i+1\}) \\ & \quad \textbf{or } (i \textbf{ is even and } j \in \{i, i-1\}); \\ 0 & otherwise \end{cases}$$

from which it follows that $\mathsf{half}(\mathbf{z}) \preceq \mathbf{m}^{(i-j)}$ yielding (40). ■

An immediate consequence of the last theorem is the following

Corollary 1. *For any k probability distributions $\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(k)}$ let M be the joint distribution, with marginals equal to $\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(k)}$, output by algorithm* K-MIN-ENTROPY-JOINT-DISTRIBUTION. *Then,*

$$H(M) \leq H(\mathbf{p}^{(1)} \wedge \mathbf{p}^{(2)} \wedge \cdots \mathbf{p}^{(k)}) + \lceil \log k \rceil$$

Proof: Let k be a power of 2. Otherwise, one can duplicate some of the probability distributions until there are $k' = 2^{\lceil \log k \rceil}$ of them. By Theorem 4 we have

$$\begin{aligned} &\mathsf{half}^{(\lceil \log k \rceil)}(\mathbf{p}^{(1)} \wedge \mathbf{p}^{(2)} \wedge \cdots \mathbf{p}^{(k)}) \\ &\qquad = \mathsf{half}^{(\log k')}(\mathbf{p}^{(1)} \wedge \mathbf{p}^{(2)} \wedge \cdots \mathbf{p}^{(k')}) \preceq \mathbf{m}^{(1-k)}. \end{aligned}$$

Therefore, by the Schur-concavity of the entropy we have

$$\begin{aligned} H(M) &= H(\mathbf{m}^{(1-k)}) \\ &\leq H(\mathsf{half}^{(\lceil \log k \rceil)}(\mathbf{p}^{(1)} \wedge \mathbf{p}^{(2)} \wedge \cdots \wedge \mathbf{p}^{(k)})) \\ &= H(\mathbf{p}^{(1)} \wedge \mathbf{p}^{(2)} \wedge \cdots \wedge \mathbf{p}^{(k)}) + \lceil \log k \rceil, \end{aligned}$$

where the last equality follows by the simple observation that for any probability distribution $\mathbf{x}$ and integer $i \geq 0$ we have $H(\mathsf{half}^{(i)}(\mathbf{x})) = H(\mathbf{x}) + i$. ■

We also have the following lower bound which, together with the previous corollary implies that our algorithm guarantees an additive gap of at most $\log k$ bits for the problem of computing the joint distribution of minimum entropy of k input distributions.

Lemma 7. *Fix k distributions $\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \cdots, \mathbf{p}^{(k)}$. For any M being a joint distribution with marginals $\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \cdots, \mathbf{p}^{(k)}$, it holds that*

$$H(M) \geq H(\mathbf{p}^{(1)} \wedge \mathbf{p}^{(2)} \wedge \cdots \wedge \mathbf{p}^{(k)}).$$

Proof: For each $i = 1, \ldots, k$, the distribution $\mathbf{p}^{(I)}$ is an aggregation of M, hence $M \preceq \mathbf{p}^{(i)}$.

By definition of the greatest lower bound operator $\wedge$ for any distribution $\mathbf{x}$ such that for each i it holds that $\mathbf{x} \prec \mathbf{p}^{(i)}$ we have $\mathbf{x} \preceq \mathbf{p}^{(1)} \wedge \mathbf{p}^{(2)} \wedge \cdots \mathbf{p}^{(k)}$. Therefore, in particular we have $M \preceq \mathbf{p}^{(1)} \wedge \mathbf{p}^{(2)} \wedge \cdots \mathbf{p}^{(k)}$, which, by the Schur concavity of the entropy gives the desired result. ■

Remark 1. *The time complexity of Algorithm 3 is dominated by the time to build the output matrix in line 19, which takes $O(n^k)$. However, if the output matrix is returned in sparse form and in line 9 the improved implementation of algorithm* MIN-ENTROPY-JOINT-DISTRIBUTION *is used (see section III-D and the appendix), the time for the overall construction is upper bounded by $\sum_{\ell=1}^{\log k} O(\frac{k}{2^\ell} 2^{\ell-1} n \log(2^{\ell-1} n)) = O(nk \log(nk))$. To see this, observe that the main* **for** *loop in line 5 is executed $O(\log k)$ times and in each iteration $\ell = 1, \ldots, \log k$ there are $\frac{k}{2^\ell}$ executions of* MIN-ENTROPY-JOINT-DISTRIBUTION *over distributions having $O(2^{\ell-1} n)$ non-zero entries and the algorithm employs the arrays Ind in order to perform the computation only considering the non-zero entries of these distributions.*

Summarising we have shown the following

Theorem 5. *Let $\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(m)} \in \mathcal{P}_n$. Let M^* be a joint distribution with marginals $\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(m)}$ of* minimum *entropy among all the joint distribution having marginals equal to $\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(m)}$. Let M be the joint distribution of $\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(m)}$ output by our algorithm. Then,*

$$H(M) \leq H(M^*) + \lceil \log(m) \rceil.$$

Hence, our (polynomial time) algorithm guarantees an additive gap of $\log(m)$.

Using Lemma 7 and Theorem 5 above, and Theorem 3 of [31], we obtain the following version of the functional representation lemma (please see the discussion in Section I-B of the present paper).

Corollary 2. *Let X and Y be two arbitrary random variables with joint distribution $p(x, y)$, where X takes values $x_1, \ldots, x_k$. Let $\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(k)}$ be the distribution of the conditioned r.v. $Y|X = x_1, \ldots, Y|X = x_k$, respectively. Then, for any r.v. Z independent from X for which there exist a function f such that $(X, Y) = (X, f(X, Z))$, it holds that*

$$H(Z) \geq H(\mathbf{p}^{(1)} \wedge \mathbf{p}^{(2)} \wedge \cdots \wedge \mathbf{p}^{(k)}).$$

Conversely, there exists a Z independent from X and a function f for which $(X, Y) = (X, f(X, Z))$ such that

$$H(Z) \leq H(\mathbf{p}^{(1)} \wedge \mathbf{p}^{(2)} \wedge \cdots \wedge \mathbf{p}^{(k)}) + \log k.$$

C. The proofs of the two technical lemmas about the operator half

Lemma 5. *For any* $\mathbf{p} \preceq \mathbf{q}$ *we have also* $\mathsf{half}(\mathbf{p}) \preceq \mathsf{half}(\mathbf{q})$.

Proof: It is easy to see that assuming $\mathbf{p}$ and $\mathbf{q}$ rearranged in order to have $p_1 \geq p_2 \geq \cdots \geq p_n$ and $q_1 \geq q_2 \geq \cdots \geq q_n$ we also have $\mathsf{half}(\mathbf{p})_1 \geq \mathsf{half}(\mathbf{p})_2 \geq \cdots \geq \mathsf{half}(\mathbf{p})_{2n}$ and $\mathsf{half}(\mathbf{q})_1 \geq \mathsf{half}(\mathbf{q})_2 \geq \cdots \geq \mathsf{half}(\mathbf{q})_{2n}$.

By assumption we also have that for each $j = 1, \ldots, n$ it holds that $\sum_{i=1}^{j} p_i \leq \sum_{i=1}^{j} p_i$.

Therefore, for each $j = 1, \ldots 2n$ it holds that

$$\begin{aligned}\sum_{i=1}^{j} \mathsf{half}(\mathbf{p})_i &= \frac{1}{2}\sum_{i=1}^{\lceil j/2 \rceil} p_i + \frac{1}{2}\sum_{i=1}^{\lfloor j/2 \rfloor} p_i \\ &\leq \frac{1}{2}\sum_{i=1}^{\lceil j/2 \rceil} q_i + \frac{1}{2}\sum_{i=1}^{\lfloor j/2 \rfloor} q_i = \sum_{i=1}^{j} \mathsf{half}(\mathbf{q})_i,\end{aligned}$$

proving that $\mathsf{half}(\mathbf{p}) \preceq \mathsf{half}(\mathbf{q})$. ∎

Fact 2. *For any pair of distributions* $\mathbf{p}, \mathbf{q} \in \mathcal{P}_n$. *It holds that*

$$\mathsf{half}(\mathbf{p} \wedge \mathbf{q}) \preceq \mathsf{half}(\mathbf{p}) \wedge \mathsf{half}(\mathbf{q}).$$

Proof: By Lemma 5 we have that

$$\mathsf{half}(\mathbf{p} \wedge \mathbf{q}) \preceq \mathsf{half}(\mathbf{p}) \qquad \text{and} \qquad \mathsf{half}(\mathbf{p} \wedge \mathbf{q}) \preceq \mathsf{half}(\mathbf{q})$$

Then, by the property of the operator $\wedge$ which gives the greatest lower bound we have the desired result. ∎

On the basis of Fact 2 we can extend the result to "powers" of the operator half and have our Lemma 6.

Lemma 6. *For any pair of distributions* $\mathbf{p}, \mathbf{q} \in \mathcal{P}_n$. *and any* $i \geq 0$, *It holds that*

$$\mathsf{half}^{(i)}(\mathbf{p} \wedge \mathbf{q}) \preceq \mathsf{half}^{(i)}(\mathbf{p}) \wedge \mathsf{half}^{(i)}(\mathbf{q}).$$

Proof: We argue by induction on i. The base case $i = 1$ is given by the previous Fact 2. Then, for any $i > 1$

$$\begin{aligned}\mathsf{half}^{(i)}(\mathbf{p} \wedge \mathbf{q}) &= \mathsf{half}(\mathsf{half}^{(i-1)}(\mathbf{p} \wedge \mathbf{q})) \\ &\preceq \mathsf{half}(\mathsf{half}^{(i-1)}(\mathbf{p}) \wedge \mathsf{half}^{(i-1)}(\mathbf{q})) \\ &\preceq \mathsf{half}(\mathsf{half}^{(i-1)}(\mathbf{p})) \wedge \mathsf{half}(\mathsf{half}^{(i-1)}(\mathbf{p}))\end{aligned}$$

from which the desired result immediately follows. The first $\preceq$-inequality follows by induction hypothesis and the second inequality by Fact 2. ∎

VI. Conclusions

In this paper we have studied the problem of finding a minimum entropy joint distribution with *fixed* marginals. We have pointed out that this problem naturally arises in a variety of situations: causal inference, one-shot channel simulation, metric computation for dimension reduction, optimization in the transportation polytope, and several others. Our main result consists in a polynomial time algorithm to find an $\mathbf{M} \in \mathcal{C}(\mathbf{p}, \mathbf{q})$ such that $H(\mathbf{M}) \leq OPT + 1$ bit, where $OPT = \min_{\mathbf{N} \in \mathcal{C}(\mathbf{p}, \mathbf{q})} H(\mathbf{N})$. We ave also shown that our approach (relying on majorization among probability distributions) allows us to easily extend our results to Rényi entropies of arbitrary positive orders (thus generalizing the result for the Shannon entropy where the latter is equal to the Rényi entropy of order 1).

There are many possible extensions of our work. Firstly, although our result for the minimum entropy bivariate joint distribution with fixed two marginals seems quite tight, it is very likely that a more direct approach (i.e., that does not rely on the iterative construction of Section V) could give better results for multivariate joint distributions. Another interesting problem would be to extend our results to the case in which one seeks a minimum entropy bivariate joint distribution with marginals "close" to given ones, for appropriate measures of closeness. Finally, a natural research problem is related to the scenario considered in Section I-B: Given arbitrary correlated r.v.'s X and Y, it would be interesting to find a r.v. Z, independent from X, such that the pair of r.v.'s $(X, f(X, Z))$ is distributed like (X, Y), for appropriate deterministic function f, for which *both* $H(Z)$ and $H(Y|Z)$ are close to their lower bounds.

Acknowledgments

The authors want to thank Executive Editor Professor I. Sason, Associate Editor Professor I. Kontoyiannis, and the anonymous referees for many useful comments and suggestions.

References

[1] S. Bahmani, *Algorithms for Sparsity-Constrained Optimization*, Springer 2014.

[2] C. H. Bennett, P. W. Shor, J. A. Smolin, and A. Thapliyal, "Entanglement-assisted capacity of a quantum channel and the reverse Shannon theorem", *IEEE Transactions on Information Theory*, vol. **48**, (2002) 2637-2655.

[3] V. Benes and J. Stepan (Eds.), *Distributions with given Marginals and Moment Problems*, Springer (1997).

[4] M. Braverman and A. Garg, "Public vs private coin in bounded-round information", in: Esparza J., Fraigniaud P., Husfeldt T., Koutsoupias E. (eds) *Automata, Languages, and Programming, ICALP 2014*. Lecture Notes in Computer Science, vol 8572. Springer, Berlin, Heidelberg (2014).

[5] F. Cicalese and U. Vaccaro, "Supermodularity and subadditivity properties of the entropy on the majorization lattice", *IEEE Transactions on Information Theory*, Vol. **48** (2002) 933–938.

[6] F. Cicalese and U. Vaccaro, "Bounding the average length of optimal source codes via majorization theory", *IEEE Transactions on Information Theory* , Vol. **50** (2004), 633–637.

[7] F. Cicalese, L. Gargano, and U. Vaccaro, "Information theoretic measures of distances and their econometric applications", *Proceedings of International Symposium on Information Theory* (ISIT 2013), 409–413.

[8] F. Cicalese, L. Gargano, and U. Vaccaro, "Approximating probability distributions with short vectors, via information theoretic distance measures", *Proceedings of International Symposium on Information Theory* (ISIT 2016), 1138–1142.

[9] F. Cicalese, L. Gargano, and U. Vaccaro, "Bounds on the entropy of a function of a random variable and their applications", *IEEE Transactions on Information Theory*, vol. **64**, (2018), 2220–2230.

[10] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein, *Introduction to Algorithms*, MIT Press, 2009.

[11] C.M. Cuadras, J. Fortiana, J.A. Rodriguez-Lallena (Eds.) *Distributions with Given Marginals and Statistical Modeling*. Springer (2002).

[12] P. Cuff, T. Cover, G. Kumar, and L. Zhao, "A lattice of gambles", *Proceedings of International Symposium on Information Theory* (ISIT 2011), 1762–1766.

[13] G. Dall'Aglio, S. Kotz, and G. Salinetti (Eds.), *Advances in Probability Distributions with Given Marginals*. Springer (1991).

[14] J.A. De Loera and E.D. Kim, "Combinatorics and geometry of transportation polytopes: an update." in: *Discrete geometry and algebraic combinatorics*, A. Barg and O. R. Musin (Eds.), vol. *625*, American Matyhematical Society, (2014), 37–76.

[15] A. Dobra and S. E. Fienberg, "Bounds for cell entries in contingency tables given marginal totals and decomposable graphs", in: *Proceedings of the National Academy of Sciences*, vol. **97**, (2000) 11185–11192.

[16] A. El Gamal and Y.-H. Kim, *Network Information Theory*, Cambridge Univesity Press, 2001.

[17] T. van Erven and P. Harremöes, "Rényi Divergence and majorization", In: *Proceedings of International Symposium on Information Theory* (ISIT 2010), pp. 1335-1339.

[18] M. Frechet, "Sur les tableaux de correlation dont le marges sont donnees", *Ann. Univ. Lyon Sci. Sect. A*, vol. **14**, (1951), 53–77.

[19] T. Fritz and R, Chaves, "Entropic inequalities and marginal problems", *IEEE Transactions on Information Theory*, Vol. **59**, No. 2, (2013) 803–817.

[20] S. Furuichi, K. Yanagi, and K. Kuriyama, "Fundamental properties of Tsallis relative entropy," *Journal of Mathematical Physics*, vol. **45**, no. 12, pp. 4868–4877, 2004

[21] Y. Han, O. Ordentlich, and O. Shayevitz, "Mutual information bounds via adjacency events", *IEEE Transactions on Information Theory*, Vol. **62**, (2016) 6068–6080.

[22] P. Harremöes, "A new look on majorization," in: *Proceedings of the International Symposium on Information Theory and Its Applications*, ISITA 2004, 1422-1425.

[23] S.W. Ho and R.W. Yeung, "The interplay between entropy and variational distance", *IEEE Transactions on Information Theory*, **56**, 5906–5929, 2010.

[24] S. W. Ho and S. Verdù, "On the interplay between conditional entropy and error probability", *IEEE Transactions on Information Theory*, **56**, 5930–5942, 2010.

[25] S. W. Ho and S. Verdù, "Convexity/concavity of Rényi entropy and α-mutual information", *Proceedings of International Symposium on Information Theory* (ISIT 2015), 745–749.

[26] P. Harsha, R. Jain, D. McAllester, and J. Radhakrishnan, "The communication complexity of correlation," *IEEE Trans. Info. Theory*, vol. **56**, no. 1, pp. 438–449, Jan 2010.

[27] W. Hoeffding, "Masstabinvariante Korrelationtheorie". *Schriften Math.*, Inst. Univ. Berlin, Vol. **5**, (1940) 181–233. English translation: Scale-invariant correlation theory. In: Fisher *et al.* (eds.) *The Collected Works of Wassily Hoeffding*, pp. 57–107, Springer-Verlag, (1999).

[28] N. Hurley and S. Rickard, "Comparing measures of sparsity", *IEEE Transactions on Information Theory*, vol. **55**, n. 10, 4723–4741, (2009)

[29] J.N. Kapur , G. Baciu and H.K. Kevesan, "The minmax information measure", *International Journal of Systems Science*, Vol. **26**, Issue 1, (1995) 1–12.

[30] D. Knuth and A. Yao, "The complexity of nonuniform random number generation," in: *Algorithms and Complexity, New Directions and Results*, J. F. Traub, Ed. New York: Academic, 1976, 357 – 428.

[31] M. Kocaoglu, A. G. Dimakis, S. Vishwanath and B. Hassibi, "Entropic causal inference", in: *Proceedings of 31-st AAAI Conference on Artificial Intelligence (AAAI-17)*, (2017), 1156 – 1162.

[32] M. Kocaoglu, A. G. Dimakis, S. Vishwanath and B. Hassibi, "Entropic causality and greedy minimum entropy coupling", in: *Proceedings of the 2017 International Symposium on Information Theory*, (2017), 1465 – 1469.

[33] I. Kontoyiannis and M. Madiman, "Sumset and inverse sumset inequalities for differential entropy and mutual information", *IEEE Transactions on Information Theory*, Vol. **60**, Issue 8, (2014) 4503–4514.

[34] M. Kovačević, I. Stanojević, and V. Senk, "On the entropy of couplings", *Information and Computation*, Vol. **242**, (2015) 369–382.

[35] S. Krishnaswamy, M.H. Spitzer, M. Mingueneau, S.C. Bendall, O. Litvin, E. Stone, D. Peer, and G. P. Nolan, "Conditional density-based analysis of T cell signaling in single-cell data", *Science*, vol. **346**, no. 6213, (2014).

[36] C. T. Li and A. El Gamal, "Strong functional representation lemma and applications to coding theorems", in: *Proceedings of the 2017 IEEE International Symposium on Information Theory*, (2017), 589–593.

[37] G.D. Lin, X. Dou, S. Kuriki and J.-S. Huang, "Recent developments on the construction of bivariate distributions with fixed marginals", *Journal of Statistical Distributions and Applications*, (2014) 1–14.

[38] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information", *IEEE Transactions on Medical Imaging*, Vol. **16**, No. 2, (1997), 187–198.

[39] A.W. Marshall, I. Olkin, and B.C. Arnold, *Inequalities: Theory of Majorization and Its Applications*, Springer, New York (2009).

[40] D.J. Miller and W.-H Liu, "On the recovery of joint distributions from limited information", *Journal of Econometrics*, Vol **107** (2002), 259–274.

[41] A. Painsky, S. Rosset and M. Feder, "Innovation representation of stochastic processes with application to causal inference", `arXiv:1811.10071 [cs.IT]`

[42] A. Perez and M. Studeny, "Comparison of two methods for approximation of probability distributions with prescribed marginals", *Kibernetika*, Vol. **43** (2007), No. 5, 591–618.

[43] J.P.W. Pluim, J.B.A. Maintz, and M.A. Viergever, "Mutual-information-based registration of medical images: A survey", *IEEE Transactions on Medical Imaging*, Vol. **22**, (2003), 986–1004.

[44] A. Rényi, "On measures of dependence", *Acta Math. Acad. Sci. Hung.*, vol. **10**, (1959), 441–451.

[45] A. Rényi, "On measures of entropy and information," in: *Fourth Berkeley Symp. on Mathematical Statistics and Probability*, 547–561, 1961.

[46] T. Roughgarden and M. Kearns, "Marginals-to-models reducibility", in: *Advances in Neural Information Processing Systems* (NIPS 2013), C.J.C. Burges *et al.* (Eds.), (2013), 1043–1051.

[47] I. Sason, "Entropy bounds for discrete random variables via maximal coupling", *IEEE Transactions on Information Theory*, Vol. **59**, (2013), 7118 – 7131.

[48] I. Sason, "Tight bounds on the Rényi entropy via majorization with applications to guessing and compression", *Entropy* 2018, **20**, 896.

[49] S. Shalev-Shwartz, N. Srebro, and T. Zhang, "Trading accuracy for sparsity in optimization problems with sparsity constraints", *SIAM J. Optim.*, vol. **20**, (2010), 2807–2832.

[50] T. Tao, "Sumset and inverse sumset theory for Shannon entropy", *Combinatorics, Probability and Computing*, vol. **19**, (2010), 603-639.

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

[52] Y. Steinberg and S. Verdú, "Simulation of random processes and rate-distortion theory", *IEEE Transactions on Information Theory*, vol. **42** (1) , (1996), 63 – 86.

[53] H.-Y. Xu, S.-H. Kuoa, G. Li, E.F.T. Legara, D. Zhao, C.P. Momterola, "Generalized cross entropy method for estimating joint distribution from incomplete information", *Physica A*, Vol. **453** (2016), 162–172.

[54] M. Vidyasagar, "A metric between probability distributions on finite sets of different cardinalities and applications to order reduction", *IEEE Transactions on Automatic Control*, Vol. **57** , 2464–2477, (2012).

[55] W.M. Wells III,, P. Viola, H. Atsumi, S. Nakajima and R. Kikinis, "Multi-modal volume registration by maximization of mutual information", *Medical Image Analysis*, Vol. **1** (1996), 35–51.

[56] W. Whitt, "Bivariate distributions with given marginals", *The Annals of Statistics*, Vol. **4**, No. 6, (1976), 1280–1289

[57] L. Yu and V. Y. F. Tan, "Asymptotic coupling and its applications in information theory", `arXiv:1712.06804 [cs.IT]`.

[58] L. Yuan and H.K. Kevasan, "Minimum entropy and information measure", *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. **28**, No. 3, (1998), 488–491.

APPENDIX

Algorithm 4 The Min Entropy Joint Distribution Algorithm outputting a sparse representation of $\mathbf{M}$

MIN-ENTROPY-JOINT-DISTRIBUTION-SPARSE($\mathbf{p}, \mathbf{q}$)
Input: prob. distributions $\mathbf{p} = (p_1, \ldots, p_n)$ and $\mathbf{q} = (q_1, \ldots, q_n)$
Output: A Coupling $\mathbf{M} = [m_{i\,j}]$ of $\mathbf{p}$ and $\mathbf{q}$ in sparse representation $\mathbf{L} = \{(m_{i\,j}, (i,j)) \mid m_{i\,j} \neq 0\}$

if $\mathbf{p} \neq \mathbf{q}$, let $i = \max\{j \mid p_j \neq q_j\}$; **if** $p_i < q_i$ **then swap** $\mathbf{p} \leftrightarrow \mathbf{q}$
$\mathbf{z} = (z_1, \ldots, z_n) \leftarrow \mathbf{p} \wedge \mathbf{q}$, $\mathbf{L} \leftarrow \emptyset$
CREATEPRIORITYQUEUE($\mathcal{Q}^{(row)}$), $qrowsum \leftarrow 0$
CREATEPRIORITYQUEUE($\mathcal{Q}^{(col)}$), $qcolsum \leftarrow 0$
for $i = n$ **downto** 1 **do**
  $z_i^{(d)} \leftarrow z_i$, $z_i^{(r)} \leftarrow 0$
  if $qcolsum + z_i > q_i$ **then**
    $(z_i^{(d)}, z_i^{(r)}, I, qcolsum) \leftarrow$ LEMMA3-SPARSE$(z_i, q_i, \mathcal{Q}^{(col)}, qcolsum)$
    for each $(m, \ell) \in I$ **do** $\mathbf{L} \leftarrow \mathbf{L} \cup \{(m, (\ell, i)\}$
    if $z_i^{(r)} > 0$ **then** INSERT$(\mathcal{Q}, (z_i^{(r)}, i))$; $qcolsum \leftarrow qcolsum + z_i^{(r)}$
  else $\{qcolsum + z_i = q_i\}$
    while $\mathcal{Q}^{(col)} \neq \emptyset$ **do**
      $(m, \ell) \leftarrow$ EXTRACTMIN$(\mathcal{Q}^{(col)})$, $qcolsum \leftarrow qcolsum - m$, $\mathbf{L} \leftarrow \mathbf{L} \cup \{(m, (\ell, i))\}$
  if $qrowsum + z_i > p_i$ **then**
    $(z_i^{(d)}, z_i^{(r)}, I, qrowsum) \leftarrow$ LEMMA3-SPARSE$(z_i, p_i, \mathcal{Q}^{(row)}, qrowsum)$
    for each $(m, \ell) \in I$ **do** $\mathbf{L} \leftarrow \mathbf{L} \cup \{(m, (i, \ell)\}$
    if $z_i^{(r)} > 0$ **then** INSERT$(\mathcal{Q}^{(row)}, (z_i^{(r)}, i))$; $qrowsum \leftarrow qrowsum + z_i^{(r)}$
  else $\{qrowsum + z_i = p_i\}$
    while $\mathcal{Q}^{(row)} \neq \emptyset$ **do**
      $(m, \ell) \leftarrow$ EXTRACTMIN$(\mathcal{Q}^{(row)})$, $qrowsum \leftarrow qrowsum - m$, $\mathbf{L} \leftarrow \mathbf{L} \cup \{(m, (i, \ell))\}$
  $\mathbf{L} \leftarrow \mathbf{L} \cup \{(z_i^{(d)}, (i, i))\}$;

Algorithm 5 The procedure implementing Lemma 3 for the sparse implementation

LEMMA3-SPARSE($z, x, \mathcal{Q}, qsum$)
Input: reals $z > 0$, $x \geq 0$, and priority queue $\mathcal{Q}$ s.t. $\left(\sum_{(m,\ell)\in\mathcal{Q}} m\right) = qsum$ and $qsum + x \geq z$
Output: $z^{(d)}, z^{(r)} \geq 0$, and $I \subseteq \mathcal{Q}$ s.t. $z^{(d)} + z^{(r)} = z$, and $z^{(d)} + \sum_{(m,\ell)\in I} m = x$.
$I \leftarrow \emptyset$, $sum \leftarrow 0$
while $\mathcal{Q} \neq \emptyset$ **and** $sum +$ MIN$(\mathcal{Q}) < x$ **do**
  $(m, \ell) \leftarrow$ EXTRACTMIN$(\mathcal{Q})$, $qsum \leftarrow qsum - m$
  $I \leftarrow I \cup \{(m, \ell)\}$, $sum \leftarrow sum + m$
$z^{(d)} \leftarrow x - sum$, $z^{(r)} \leftarrow z - z^{(d)}$
return $(z^{(d)}, z^{(r)}, I, qsum)$

Ferdinando Cicalese Ferdinando Cicalese received the Masters and PhD degrees in computer science from University of Salerno (Italy) in 1995 and 2001, respectively. From 2001 to 2014 he was first assistant professor and then associate professor at University of Salerno and from 2004 to 2009 he was research group leader at Bielefeld University (Germany). Since 2014, he has been with the Computer Science department at University of Verona (Italy). Dr. Cicalese is the recipient of the 2004 Sofja Kovalevskaja award from the Humboldt Foundation and the Germany Ministry of Education and Research. Dr. Cicaleses research interests are in the area of algorithms and complexity (with a special emphasis on combinatorial search algorithms, string matching and indexing, and decision tree construction optimization), information theory and fault tolerant error-correction codes. He is the author of more than 100 scientific publications and a monograph on fault tolerant search algorithms, published by Springer in 2011. Dr. Cicalese has also been guest editor for several international journals, including Theoretical Computer Science and Algorithmica, and has served as programme committee member, programme chair and organizer of several international conferences.

Luisa Gargano Luisa Gargano (M'16) received a Laurea degree (*cum laude*) in Computer Science in 1983 from the University of Salerno. She is Professor of Computer Science at the Department of Computer Science of the University of Salerno, that she joined in 1986. She has been visiting researcher at the Department of Computer Science of Columbia University (USA), at the University of Bielefeld (Germany), at the CNRS and INRIA at Sophia Antipolis (France), and at Simon Fraser University (Canada). Her research activity include communication problems in Networks, Parallel and Distributed Computing, Information Theory and Coding, applications of Information–Theoretical methods to Extremal Combinatorics, and Data Security. Her latest research interests lie in the areas of network algorithms, with a special emphasis in problems of routing in networks and social networks. She has co-authored more than 110 papers in international journals and refereed conferences.

Ugo Vaccaro Ugo Vaccaro (M'06, SM18) was born in Italy in 1958. He received the Laurea Degree (*cum laude*) in Computer Science from the University of Salerno in 1981. Since 1994 he is Professor of Computer Science at the University of Salerno, Italy. His current research interests include group testing, combinatorial and information theoretic methods for multiple access communication and testing, spread of influence in social network, and, especially, computational complexity issues in Shannon theory. In the past years he has done research in coding theory, search theory, parallel and distributed algorithms, and applications of information theoretic methods to data security and extremal combinatorics. In these areas he has published more than 150 papers in international journals and proceedings of conferences. He has been editor of the proceedings of several international conferences and member of the program committee of more than eighty international conferences. He has been Visiting Professor for extended periods at INRIA at Sophia Antipolis, CNRS Nice and University of Nice (France), Bielefeld University (Germany), Simon Fraser University (Canada), and the International Computer Science Institute at Berkeley (USA).