Efficient Approximate Minimum Entropy Coupling of Multiple Probability Distributions

Cheuk Ting Li

Department of Information Engineering

The Chinese University of Hong Kong

Email: ctli@ie.cuhk.edu.hk

Abstract

Given a collection of probability distributions $p_1, \ldots, p_m$, the minimum entropy coupling is the coupling $X_1, \ldots, X_m$ ($X_i \sim p_i$) with the smallest entropy $H(X_1, \ldots, X_m)$. While this problem is known to be NP-hard, we present an efficient algorithm for computing a coupling with entropy within 2 bits from the optimal value. More precisely, we construct a coupling with entropy within 2 bits from the entropy of the greatest lower bound of $p_1, \ldots, p_m$ with respect to majorization. This construction is also valid when the collection of distributions is infinite, and when the supports of the distributions are infinite. Potential applications of our results include random number generation, entropic causal inference, and functional representation of random variables.

Index Terms

Entropy minimization, coupling, random number generation, alias method, functional representation.

I. INTRODUCTION

The problem of finding the minimum entropy coupling of two discrete probability distributions p, q, i.e., finding a pair of jointly distributed random variables X, Y such that X has marginal distribution p, Y has marginal distribution q, and the joint entropy $H(X, Y)$ is minimized, has been studied by Vidyasagar [1], Painsky, Rosset and Feder [2], [3], Kovačević, Stanojević and Šenk [4], Kocaoglu, Dimakis, Vishwanath and Hassibi [5], [6], Cicalese, Gargano and Vaccaro [7], [8], Yu and Tan [9], and Rossi [10]. Also see [11], [12], [13] for related problems. While it is shown in [1], [4] that this problem is NP-hard, a polynomial time approximation algorithm (within 1 bit from the optimum) is given in [8] (also see [6], [10]).

This problem can be generalized to the coupling of m probability distributions $p_1, \ldots, p_m$ (i.e., constructing random variables $X_1, \ldots, X_m$ with marginals $X_i \sim p_i$), where [8] gives an algorithm for constructing a coupling with entropy $H(X_1, \ldots, X_m)$ within $\lceil \log m \rceil$ bits from the optimum (also see [6] for another algorithm). More precisely, [8] gives a coupling with entropy at most $H(\bigwedge_i p_i) + \lceil \log m \rceil$ bits, where $\bigwedge_i p_i$ denotes the greatest lower bound of $p_1, \ldots, p_m$ with respect to majorization of probability vectors [14]. Since any coupling of $p_1, \ldots, p_m$ has entropy at least $H(\bigwedge_i p_i)$ [8], this gives a construction within $\lceil \log m \rceil$ bits from the optimum.

In this paper, we improve this result by constructing a coupling of $p_1, \ldots, p_m$ with entropy at most

$$H\left(\bigwedge_i p_i\right) + 2 - 2^{2-m}, \tag{1}$$

which is at most 2 bits from the optimum. A more general bound in terms of Rényi entropy [15] can also be obtained. See Corollary 9 and Theorem 11. Compared to the $\lceil \log m \rceil$ gap in [8], the gap $2 - 2^{2-m} \leq 2$ in our result does not scale with m. Also note that the gap becomes 1 when $m = 2$, the same gap as in [8], [10] for the coupling of two distributions. We describe an algorithm (Algorithm 3) for computing a coupling achieving (1) with time complexity $O(m^2 n + mn \log n)$, where we assume the pmf's p_i are over a finite set $\mathcal{X}$ with $|\mathcal{X}| = n$. If we allow an error at most ϵ (i.e., changing each p_i by at most ϵ in total variation distance), we can reduce the time complexity to $O(mn \log(1/\epsilon) + mn \log n)$ (see Remark 14).

Moreover, (1) continues to hold when the collection of pmf's to be coupled is infinite, or even uncountable (in this case, $m = \infty$ and $2^{2-m} = 0$). The bound in (1) also applies to the case where the supports of the distributions are infinite. These cases are not handled in [5], [6], [8], [10].

Below are some potential applications of a low entropy coupling of a collection of distributions.

A. Random Number Generation

It was shown by Knuth and Yao [16] that a discrete random variable X can be generated using an expected number of fair coin flips no more than $H(X) + 2$, indicating that the entropy $H(X)$ is a measure of the amount of resources (coin flips) needed to generate the random number X (also see [17], [18]). The entropy of a coupling of a collection of distributions S can be regarded as the amount of resources needed to allow generation of any distribution in S. More precisely, consider the setting where there is a random number generator device that can output a random number to the user (who does not have access to random sources other than the generator). The user wants to generate $X \sim p$ for a distribution $p \in S$ of the user's

choice (p is not fixed a priori). If the generator is versatile enough to generate any distribution p at the user's request, then the minimum amount of entropy used by the generator is $H(p)$. Nevertheless, the generator may not be programmable or configurable. If we assume the generator is only capable of generating a random number Z following a fixed distribution (that depends on the design of the generator, but cannot depend on the user's choice of p), then the user has to apply a mapping g_p (depending on the choice of p) to obtain the final random number $X = g_p(Z) \sim p$. This induces a coupling $\{g_p(Z)\}_{p \in S}$ of the distributions in S. Therefore, the minimum entropy coupling of S corresponds to the distribution of Z that has the minimum entropy needed to accomplish this task.

Existing hardware random number generators are capable of generating a uniformly random integer within a range of integers. While we can generate from any discrete distribution by repeated and interactive usages of such generator (e.g. by [16]), such interactive communication between the generator and the user may not be feasible depending on the situation (e.g. delay in generating the random number and communication). The minimum entropy coupling allows us to design the generator according to S (with possibly non-uniform output Z) so that we only need to use the generator once per random number $X \sim p$ obtained by the user.

We will see in the following sections that our construction is similar to the alias method for random number generation by Walker [19]. While the alias method only works for discrete distributions with finite support, and requires an amount of entropy approximately $\log k$ (where k is the size of the support), our construction works for any discrete distribution (with finite or infinite support), and requires an amount of entropy close to the theoretical minimum (which can be much smaller than $\log k$ depending on the collection of distributions S).

A related setting is channel simulation (see [20], [21], [22], [23] for the asymptotic case, and [24], [25], [26] for the one-shot case), where the encoder observes a distribution $p \in S$ in a collection of distributions S and transmits a message M to the decoder (who knows S but does not know p a priori), so as to allow the decoder to generate $X \sim p$. The aforementioned random number generation setting corresponds to the one-shot channel simulation setting where the encoder does not have local randomness, the communication M from the encoder to the decoder is unlimited, and our goal is to minimize the amount of local randomness at the decoder in order to generate $X \sim p$ (we require the distribution of the local randomness to be fixed).

B. Functional Representation and Entropic Causal Inference

The functional representation lemma [27] states that for any pair of random variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$, there exists a random variable Z independent of X such that $Y = g(X, Z)$ is a function of (X, Z). See [28], [29] for applications of this lemma in information theory. Since $Y_x := g(x, Z) \sim p_{Y|X=x}$, $\{Y_x\}_{x \in \mathcal{X}}$ is a coupling of the conditional distributions $p_{Y|X=x}$, and hence the problem of finding a functional representation with the smallest $H(Z)$ is equivalent to the minimum entropy coupling problem (see [5], [6]).

Shannon [30, Fig. 1] considers a channel to be a function mapping the input signal and noise source to the received signal. Letting the input signal and the received signal be X and Y respectively, the minimum $H(Z)$ in the functional representation would be the minimum entropy of the noise source of the channel. Note that this measure is an inherent property of the channel, and does not depend on the input distribution p_X as long as $p_X(x) > 0$ for all $x \in \mathcal{X}$ (since the minimum $H(Z)$ is the minimum entropy of a coupling of $\{p_{Y|X=x}\}_{x \in \mathcal{X}}$ which does not depend on p_X).

Kocaoglu, Dimakis, Vishwanath and Hassibi [5], [6] consider the problem of identifying the causal direction between X and Y, based on the assumption that the correct causal direction gives a small $H(Z)$. More precisely, the *entropic causal inference* method declares that $X \to Y$ is the correct direction if $Y = g(X, Z)$ can be achieved with a smaller $H(X) + H(Z)$ compared to the smallest $H(Y) + H(\tilde{Z})$ satisfying $X = \tilde{g}(Y, \tilde{Z})$. They have proposed algorithms for minimizing $H(Z)$, or equivalently, minimizing the entropy of the coupling of $p_{Y|X=x}$ (also see [8] for another algorithm). Nevertheless, these algorithms only work when $\mathcal{X}$ is finite (or the number of distributions to couple is finite). The method in this paper works regardless of whether X is a discrete or continuous random variable (though Y must be discrete). By (1), the minimum of $H(Z)$ is closely approximated by $H(\bigwedge_{x \in \mathcal{X}} p_{Y|X=x})$ (within 2 bits), and hence replacing $H(Z)$ by $H(\bigwedge_{x \in \mathcal{X}} p_{Y|X=x})$ (which can be computed in $O(|\mathcal{X}||\mathcal{Y}| \log |\mathcal{Y}|)$ time if $|\mathcal{X}|, |\mathcal{Y}| < \infty$) in the entropic causal inference method provides a close approximation that can be computed efficiently (compared to the exact minimization of $H(Z)$ which is NP-hard [1], [4]). If the function g is also needed, then it can be computed in $O(|\mathcal{X}|^2|\mathcal{Y}| + |\mathcal{X}||\mathcal{Y}| \log |\mathcal{Y}|)$ time using Algorithm 3.

The problem of minimizing $H(Y|Z)$ (instead of $H(Z)$) was studied by Li and El Gamal [31]. The strong functional representation lemma [31] states that for any pair of random variables (X, Y), there exists a random variable Z independent of X such that Y is a function of (X, Z), and $H(Y|Z) \le I(X; Y) + \log(I(X; Y) + 1) + 4$ (also see [24], [32]). The lemma is applied to show several one-shot variable-length lossy source coding results, and a short proof of the asymptotic achievability in the Gelfand-Pinsker theorem [33]. It is also used in [34] to prove a result on minimax remote prediction with a communication constraint. The Poisson functional representation given in [31] (which induces a coupling of $p_{Y|X=x}$) is also used in [35] to prove various results in multi-user information theory. In this paper, we concern the minimization of $H(Z)$ instead of $H(Y|Z)$ (while [31] gives a cardinality bound $|\mathcal{Z}| \le |\mathcal{X}|(|\mathcal{Y}|-1)+2$ in addition to the bound on $H(Y|Z)$, this is not the main objective there).

C. Other Uses of Coupling of Collections of Distributions

It has been shown that for any collection of distributions $p_1, \ldots, p_m$, it is possible to find a coupling $X_1, \ldots, X_m$ such that

$$\mathbf{P}(X_i \neq X_j) \leq 2d_{\mathrm{TV}}(p_i, p_j) \tag{2}$$

for any i, j, where d_{TV} is the total variation distance. This was shown in [36] for uniform distributions, [37] for discrete distributions, and [38], [39] for general distributions. This result was used in locality sensitive hashing [40] and randomized rounding algorithms [37], [41]. While a coupling achieving (2) is likely to have low entropy (since many values of X_i are the same), a low entropy coupling does not necessarily have a low $\mathbf{P}(X_i \neq X_j)$ (since whether $X_i \neq X_j$ is irrelevant in the calculation of entropy). We also remark that the connection between entropy and total variation distance has been studied in [42] using coupling.

In the study of Markov chains, it is often useful to represent the Markov chain $X_1, X_2, \ldots$ using the functional representation $X_n = g(X_{n-1}, Z_n)$, where $Z_n \overset{iid}{\sim} p_Z$. In the coupling from the past algorithm for sampling from the stationary distribution of a Markov chain [43], [44], the function g is designed so that $g(x, z)$ are likely to be equal for different values of x. This representation is also referred as innovation representation in [3]. Since the minimum entropy of Z_n is the entropy of the minimum entropy coupling of $\{p_{X_n|X_{n-1}=x}\}_x$, we can apply the coupling achieving (1) in this paper to generate $X_1, X_2, \ldots$ using a small entropy rate of $Z_1, Z_2, \ldots$.

Other works on the coupling of a collection of distributions include Wasserstein barycenter [45] and multi-marginal optimal transport [46], [47], [48], [39].

Notations

Throughout this paper, we assume that $\log$ is to base 2 and the entropy H is in bits. We write $\mathbb{N} = \{1, 2, 3, \ldots\}$, $[n] = \{1, \ldots, n\}$. For a statement E, we write $\mathbf{1}\{E\}$ for the indicator function where $\mathbf{1}\{E\} = 1$ if E holds, $\mathbf{1}\{E\} = 0$ otherwise.

A right stochastic matrix is a square matrix with nonnegative entries where each row sums to 1. The support of a probability mass function (pmf) p is written as $\mathrm{supp}(p)$. For a pmf p over $[n]$, its probability vector is denoted as $\vec{p} \in \mathbb{R}^n$ (a row vector). For a pmf p over the set $\mathcal{X}$, and a pmf q over the set $\mathcal{Y}$, the product pmf $p \times q$ is a pmf over $\mathcal{X} \times \mathcal{Y}$ with $(p \times q)(x, y) := p(x)q(y)$. The pmf of the Bernoulli distribution is denoted as $\mathrm{Bern}_\gamma(x) := \mathbf{1}\{x = 0\}(1 - \gamma) + \mathbf{1}\{x = 1\}\gamma$. The pmf of the geometric distribution over $\mathbb{N}$ is denoted as $\mathrm{Geom}_\gamma(x) := \gamma(1 - \gamma)^{x-1}$. The pmf of the capped geometric distribution over $[k]$ is denoted as

$$\mathrm{CGeom}_{\gamma,k}(x) := \begin{cases} \gamma(1-\gamma)^{x-1} & \text{if } x < k \\ (1-\gamma)^{k-1} & \text{if } x = k \\ 0 & \text{if } x > k. \end{cases} \tag{3}$$

The Rényi entropy [15] of a pmf p is defined as

$$H_\alpha(p) := \frac{1}{1 - \alpha} \log \sum_{x \in \mathrm{supp}(p)} (p(x))^\alpha$$

for $\alpha \in \mathbb{R}_{\geq 0} \backslash \{1\}$. When $\alpha = 1$, $H_\alpha(p) := H(p)$ is the Shannon entropy. When $\alpha = \infty$, $H_\alpha(p) := -\log \max_x p(x)$.

II. COUPLING AND MAJORIZATION

We first define a coupling of a set of distributions.

Definition 1. For a set of pmf's S, we say that an indexed set of random variables $\{X_p\}_{p \in S}$ is a coupling of S, written as $\{X_p\}_{p \in S} \in \Gamma(S)$, if X_p has marginal distribution p for any $p \in S$. We say that a pmf q is an *underlying distribution of a coupling* of S, written as $q \in \tilde{\Gamma}(S)$, if there exists $\{X_p\}_{p \in S} \in \Gamma(S)$ and random variable $Z \sim q$ such that X_p is a function of Z for all $p \in S$.[1]

Define the *minimum Rényi entropy of couplings* of a set of pmf's S as

$$H_\alpha^*(S) := \inf_{q \in \tilde{\Gamma}(S)} H_\alpha(q) \tag{4}$$

for $\alpha \in \mathbb{R}_{\geq 0} \cup \{\infty\}$. We write $H^*(S) = H_1^*(S)$. It is straightforward to show that when $S = \{p_1, \ldots, p_m\}$ is finite, then $H_\alpha^*(S) = \inf_{\{X_{p_i}\}_{i \in [m]} \in \Gamma(S)} H_\alpha(X_{p_1}, \ldots, X_{p_m})$ (to show a one-to-one correspondence between $\tilde{\Gamma}(S)$ and $\Gamma(S)$, simply take $q \in \tilde{\Gamma}(S)$ to be the joint pmf of $\{X_{p_i}\}_{i \in [m]}$). Nevertheless, we define $H_\alpha^*(S)$ in (4) for general S using $\tilde{\Gamma}(S)$ instead of $\Gamma(S)$, in order to avoid working with the joint entropy of an infinite collection of random variables when $|S| = \infty$.

The goal of this paper is to find or approximate $H_\alpha^*(S)$. We present the concept of aggregation in [1], [13].

---

[1]Technically, to make the set $\tilde{\Gamma}(S)$ well-defined, we can restrict q to be a pmf over $\mathbb{N}$, which will not cause any loss of generality since the support of a pmf is always countable.

Definition 2. For two pmf's p, q, we say p is an *aggregation* of q, written as $q \sqsubseteq p$, if there exists a function $g : \operatorname{supp}(q) \to \operatorname{supp}(p)$ (called the *aggregation map*) such that p is the pmf of $g(X)$, where $X \sim q$.

It is clear that "$\sqsubseteq$" is a transitive relation. If p, q are pmf's over $[n]$, then $q \sqsubseteq p$ if and only if there exists a right stochastic matrix M with $\{0, 1\}$ entries such that the probability vectors satisfy $\vec{p} = \vec{q}M$. Note that $q \in \tilde{\Gamma}(S)$ if and only if $q \sqsubseteq p$ for any $p \in S$. Therefore, a coupling of S can be specified using an underlying distribution $q \in \tilde{\Gamma}(S)$ and the set of aggregation maps $\{g_p\}_{p \in S}$, where g_p is the aggregation map for $q \sqsubseteq p$.

We will then show that "$\sqsubseteq$" is "closed under pointwise limit" in the following sense:

Proposition 3. *Let q be a pmf over a countable set $\mathcal{X}$, and $p, p_1, p_2, \ldots$ be pmf's over a countable set $\mathcal{Y}$, such that $p(y) = \lim_{i \to \infty} p_i(y)$ for any $y \in \mathcal{Y}$, and $q \sqsubseteq p_i$ for any $i \geq 1$, then we have $q \sqsubseteq p$.*

Proof: Without loss of generality, assume $\mathcal{X} = \mathcal{Y} = \mathbb{N}$, and $q(1) \geq q(2) \geq \cdots$. Let $g_1, g_2, \ldots$ be functions from $\mathbb{N}$ to $\mathbb{N}$ such that $g_i(X) \sim p_i$, where $X \sim q$. Consider whether $g_i(1) = 1$. There exists an increasing sequence $i_1, i_2, \ldots$ such that $\mathbf{1}\{g_{i_j}(1) = 1\}$ are the same for all j (since there are only two possibilities of $\mathbf{1}\{g_{i_j}(1) = 1\} \in \{0, 1\}$). Let that value of $\mathbf{1}\{g_{i_j}(1) = 1\}$ be $b_{1,1}$. There exists an increasing subsequence $i'_1, i'_2, \ldots$ of $i_1, i_2, \ldots$ such that for any $x, y \leq 2$, $\mathbf{1}\{g_{i'_j}(x) = y\}$ are the same for all j (since there are only 2^4 possibilities of $\{\mathbf{1}\{g_{i'_j}(x) = y\}\}_{x,y \leq 2}$). Let those values of $\mathbf{1}\{g_{i'_j}(x) = y\}$ be $b_{x,y}$ for $x, y \leq 2$. Repeat this procedure to define $b_{x,y}$ for any $x, y \in \mathbb{N}$.

Define $g : \operatorname{supp}(q) \to \mathbb{N}$ by $g(x) = y$ if $b_{x,y} = 1$. We now check that g is well-defined and $g(X) \sim p$. It is clear from the definition that there does not exist x and $y \neq y'$ such that $b_{x,y} = b_{x,y'} = 1$ (consider the $\max\{x, y, y'\}$-th iteration of the above procedure). Fix any $a, y \in \mathbb{N}$ and $\epsilon > 0$, and consider the $\max\{a, y\}$-th iteration of the above procedure that fixes $b_{x,y}$ for $x \leq a$. There exists an increasing sequence $i_1, i_2, \ldots$ such that $\mathbf{1}\{g_{i_j}(x) = y\} = b_{x,y}$ for all j and $x \leq a$. By $p(y) = \lim_{i \to \infty} p_i(y)$, there exists j such that $|p(y) - p_{i_j}(y)| \leq \epsilon$. Since $p_{i_j}(y) = \sum_x \mathbf{1}\{g_{i_j}(x) = y\}q(x)$, we have

$$\left| p(y) - \sum_{x \leq a} b_{x,y}q(x) \right| \leq \epsilon + \sum_{x > a} q(x).$$

Taking $a \to \infty$ and $\epsilon \to 0$, we have $\sum_x b_{x,y}q(x) = p(y)$. Since $\sum_x \sum_y b_{x,y}q(x) = \sum_y p(y) = 1$, for any $x \in \operatorname{supp}(q)$ (where $q(x) > 0$), there exists at least one (and thus exactly one) y such that $b_{x,y} = 1$. The result follows. $\blacksquare$

It is demonstrated in [8] that majorization is a useful tool in the study of coupling. Here we present the concept of majorization which allows infinite sequences or pmf's with infinite support (e.g. see [14]):

Definition 4. For two pmf's p, q, we say q is *majorized by* p, written as $q \preceq p$, if

$$\max_{B \subseteq \operatorname{supp}(q):\, |B| \leq k} q(B) \leq \max_{A \subseteq \operatorname{supp}(p):\, |A| \leq k} p(A)$$

for any $k \in \mathbb{N}$, where we write $p(A) := \sum_{x \in A} p(x)$. In other words, the sum of the k largest $q(x)$'s is not greater than the sum of the k largest $p(x)$'s.

It is clear that "$\preceq$" is a transitive relation. It is shown in [13] that $q \sqsubseteq p$ implies $q \preceq p$. If p, q are pmf's over $[n]$, then it has been shown that $q \preceq p$ if and only if there exists a doubly stochastic matrix (i.e., square matrix with nonnegative entries where each row and column sums to 1) M such that the probability vectors satisfy $\vec{p} = \vec{q}M$ (e.g. see [14]). Also, if p, q are pmf's over $[n]$ sorted in descending order (i.e., $p(1) \geq p(2) \geq \cdots \geq p(n)$ and likewise for q), then $q \preceq p$ if and only if there exists a lower triangular right stochastic matrix M such that $\vec{p} = \vec{q}M$. This property will be strenghtened in Lemma 10.

Also note that Rényi entropy is Schur concave [14], i.e., we have $H_\alpha(q) \geq H_\alpha(p)$ if $q \preceq p$. We then prove a useful property of majorization and aggregation:

Proposition 5. *Let X be a random variable with pmf p_X, and Y be a random variable with conditional pmf $p_{Y|X=x}$, and $p_{X,Y}$ be the joint pmf of (X, Y). Define $\tilde{X}, p_{\tilde{X}}, \tilde{Y}, p_{\tilde{Y}|\tilde{X}=x}, p_{\tilde{X},\tilde{Y}}$ similarly. We have:*

- *If $p_X = p_{\tilde{X}}$ and $p_{Y|X=x} \sqsubseteq p_{\tilde{Y}|\tilde{X}=x}$ for all x, then $p_{X,Y} \sqsubseteq p_{\tilde{X},\tilde{Y}}$.*
- *If $p_X = p_{\tilde{X}}$ and $p_{Y|X=x} \preceq p_{\tilde{Y}|\tilde{X}=x}$ for all x, then $p_{X,Y} \preceq p_{\tilde{X},\tilde{Y}}$.*

Proof: Assume $p_X = p_{\tilde{X}}$ and $p_{Y|X=x} \sqsubseteq p_{\tilde{Y}|\tilde{X}=x}$ for all x. There exists functions g_x for $x \in \operatorname{supp}(p_X)$ such that $(X, g_X(Y)) \stackrel{d}{=} (\tilde{X}, \tilde{Y})$. Hence, $p_{X,Y} \sqsubseteq p_{\tilde{X},\tilde{Y}}$ with the aggregation map $(x, y) \mapsto (x, g_x(y))$.

Assume $p_X = p_{\tilde{X}}$ and $p_{Y|X=x} \preceq p_{\tilde{Y}|\tilde{X}=x}$ for all x. Fix any $A \subseteq \operatorname{supp}(p_{X,Y})$. For any x, let B_x attains the maximum in $\max_{B \subseteq \operatorname{supp}(p_{\tilde{Y}|\tilde{X}=x}):\, |B| \leq |\{y:\, (x,y) \in A\}|} q(B)$. Since $p_{Y|X=x} \preceq p_{\tilde{Y}|\tilde{X}=x}$, we have

$$p_{X,Y}(A) = \sum_x p_X(x) \sum_{y:\, (x,y) \in A} p_{Y|X=x}(y)$$

$$\leq \sum_x p_X(x) p_{\tilde{Y}|\tilde{X}=x}(B_x)$$

$$= p_{\tilde{X},\tilde{Y}}\left(\{(x, y):\, y \in B_x\}\right).$$

Since $|\{(x, y) : y \in B_x\}| \leq |A|$, we have $p_{X,Y} \preceq p_{\tilde{X}, \tilde{Y}}$. ∎

We then present the definition of the greatest lower bound (see e.g. [49] for the finite case):

Definition 6. For a set of pmf's S where

$$\lim_{k \to \infty} \inf_{p \in S} \max_{A \subseteq \text{supp}(p): |A| \leq k} p(A) = 1, \tag{5}$$

define its *greatest lower bound with respect to majorization*, written as $q = \bigwedge S$, as a pmf q over $\mathbb{N}$ given by

$$q(k) := \inf_{p \in S} \max_{A \subseteq \text{supp}(p): |A| \leq k} p(A) - \inf_{p \in S} \max_{A \subseteq \text{supp}(p): |A| \leq k-1} p(A).$$

If (5) is not satisfied, then $\bigwedge S$ is undefined. (Note that (5) is always satisfied when S is finite.)

If S contains pmf's over the set $\mathcal{X}$, and $|S| = m < \infty$, $|\mathcal{X}| = n < \infty$, then it is clear that $\bigwedge S$ can be computed in $O(mn \log n)$ time (by sorting the pmf's and computing partial sums). We give some properties of the greatest lower bound. Note that the case $|S| = 2$ has been shown in [49]. While it is straightforward to generalize it to $|S| > 2$ and $|S| = \infty$, we state these properties for the sake of completeness.

Proposition 7. *For a set of pmf's S, if $q = \bigwedge S$ exists, then*
 1) $q(1) \geq q(2) \geq q(3) \geq \cdots$.
 2) $q \preceq p$ for any $p \in S$.
 3) *For any $\tilde{q}$ such that $\tilde{q} \preceq p$ for any $p \in S$, we have $\tilde{q} \preceq q$.*

Proof: Note that $q(1) \geq q(2) \geq \cdots$ is equivalent to the concavity of $\inf_{p \in S} \max_{A \subseteq \text{supp}(p): |A| \leq k} p(A)$ in k, which holds because the infimum of concave functions is concave. We have $\sum_{i=1}^{k} q(i) = \inf_{p \in S} \max_{A \subseteq \text{supp}(p): |A| \leq k} p(A)$, and hence $q \preceq p$ for any $p \in S$. For any $\tilde{q}$ such that $\tilde{q} \preceq p$ for any $p \in S$, we have $\max_{A \subseteq \text{supp}(\tilde{q}): |A| \leq k} \tilde{q}(A) \leq \inf_{p \in S} \max_{A \subseteq \text{supp}(p): |A| \leq k} p(A) = \sum_{i=1}^{k} q(i)$, and hence $\tilde{q} \preceq q$. ∎

As a result of these properties, if $\bigwedge S$ exists, for any $q \in \tilde{\Gamma}(S)$, we have $q \preceq \bigwedge S$, and hence $H_\alpha(q) \geq H_\alpha(\bigwedge S)$. Therefore, $H_\alpha^*(S) \geq H_\alpha(\bigwedge S)$.

III. Coupling by Geometric Splitting

We now present the main result in this paper, which shows that if the pmf's p, q satisfy $q \preceq p$, then after splitting each mass $q(y)$ into a sequence of masses $q(y)/2$, $q(y)/4$, $q(y)/8$,... (or equivalently, consider the joint pmf of (Y, Z) where $Y \sim q$ is independent of $Z \sim \text{Geom}_{1/2}$), then p will be an aggregation of the resultant pmf $q \times \text{Geom}_{1/2}$ ("$\times$" denote the independent product of two pmf's, i.e., it is the pmf of (Y, Z) mentioned before; refer to the notation section for the definition), which we call the *geometric splitting* of q.

Theorem 8. *If $q \preceq p$, then*

$$q \times \text{Geom}_{1/2} \sqsubseteq p.$$

A direct result of this theorem is the following explicit formula of an underlying distribution of a coupling.

Corollary 9. *For a set of pmf's S, if $\bigwedge S$ exists, then*

$$\left(\bigwedge S \right) \times \text{Geom}_{1/2} \in \tilde{\Gamma}(S).$$

As a result, the minimum Rényi entropy of couplings of S satisfies

$$H_\alpha\left(\bigwedge S \right) \leq H_\alpha^*(S) \leq H_\alpha\left(\bigwedge S \right) + H_\alpha(\text{Geom}_{1/2}),$$

where

$$H_\alpha(\text{Geom}_{1/2}) = \begin{cases} \infty & \text{if } \alpha = 0 \\ 2 & \text{if } \alpha = 1 \\ 1 & \text{if } \alpha = \infty \\ \frac{-\alpha - \log(1 - 2^{-\alpha})}{1 - \alpha} & \text{otherwise} \end{cases}$$

is the Rényi entropy of $\text{Geom}_{1/2}$.

Another way to state Theorem 8 is that for any pmf q, we have $q \times \text{Geom}_{1/2} \in \tilde{\Gamma}(\{p \text{ pmf over } \mathbb{N} : q \preceq p\})$.

Before we prove Theorem 8, we present a lemma similar to the alias method [19], and is a special case of the algorithm in [8]. We include a proof of the claim for the sake of completeness, and describe a linear time algorithm (Algorithm 1) which is considerably simpler than that in [8].

Lemma 10. *For any pmf's p, q over $[n]$ such that $q \preceq p$, $p(1) \geq p(2) \geq \cdots \geq p(n)$ and $q(1) \geq q(2) \geq \cdots \geq q(n)$, there exists $a_x \in [x-1]$ and $0 \leq r_x \leq q(x)$ for $x = 2, \ldots, n$ such that*

$$p(x) = q(x) - r_x + \sum_{y:\, a_y = x} r_y \tag{6}$$

for any $x = 1, \ldots, n$ (we let $r_1 = 0$). Moreover, a_x, r_x can be computed in $O(n)$ time (see Algorithm 1).

Lemma 10 can be stated in the following more compact form using matrices. For any pmf's p, q over $[n]$ sorted in descending order such that $q \preceq p$, there exists a lower triangular right stochastic matrix M where each row has at most one positive off-diagonal entry, and the probability vectors satisfy $\vec{p} = \vec{q} M$. Its equivalence to Lemma 10 can be shown by letting $M_{x, a_x} = r_x / q(x)$ and $M_{x,x} = 1 - r_x / q(x)$ for $x \in [n]$ (all other entries of M are zeros).

Proof of Lemma 10: Let $B_x := \{y \in [n] : a_y = x\}$. We give B_x and r_x recursively. Take $B_n = \emptyset$, $r_n = q(n) - p(n)$ ($r_n \geq 0$ since $q \preceq p$). Assume $B_{x+1}, \ldots, B_n$ and $r_{x+1}, \ldots, r_n$ are defined and satisfies that $B_{x+1}, \ldots, B_n$ are disjoint, $B_{x'} \subseteq \{x'+1, \ldots, n\}$ and

$$p(x') = q(x') - r_{x'} + \sum_{y \in B_{x'}} r_y \tag{7}$$

for all $x' > x$. We now define B_x, r_x. Take

$$B_x = \{t, \ldots, n\} \setminus \bigcup_{y=x+1}^{n} B_y,$$

where $t \in \{x+1, \ldots, n+1\}$ such that $\sum_{y \in B_x} r_y \in [p(x) - q(x),\, p(x)]$. Such t exists since $r_y \leq q(y) \leq q(x)$ for $y > x$, and

$$\sum_{y \in \{x+1, \ldots, n\} \setminus \bigcup_{y'=x+1}^{n} B_{y'}} r_y$$

$$= \sum_{y=x+1}^{n} r_y - \sum_{y'=x+1}^{n} \sum_{y \in B_{y'}} r_y$$

$$\overset{(a)}{=} \sum_{y=x+1}^{n} r_y - \sum_{y=x+1}^{n} (p(y) - q(y) + r_y)$$

$$= \sum_{y=x+1}^{n} (q(y) - p(y))$$

$$= p(x) - q(x) + \sum_{y=x}^{n} (q(y) - p(y))$$

$$\geq p(x) - q(x)$$

since $q \preceq p$, where (a) is by (7), and hence when t decreases from $n+1$ to $x+1$, $\sum_{y \in B_x} r_y$ increases from 0 to $\geq p(x) - q(x)$, with step size at most $q(x)$, and thus there exists t such that $\sum_{y \in B_x} r_y \in [p(x) - q(x),\, p(x)]$. In practice, to find B_x, we only need to scan the elements in $\bar{B}_x := [n] \setminus \bigcup_{y=x+1}^{n} B_y$ in decreasing order, and add elements from $\bar{B}_x$ to B_x until $\sum_{y \in B_x} r_y \geq p(x) - q(x)$. We then take

$$r_x = q(x) - p(x) + \sum_{y \in B_x} r_y.$$

Therefore, we have defined B_x, r_x (and hence a_x) recursively.

For the running time complexity, note that since $\bar{B}_x$ is decreasing as x decreases, only the elements in $\bar{B}_x \setminus \bar{B}_{x-1}$ are relevant to the computation of B_x, r_x. Since each $y \in [n]$ can only be removed from $\bar{B}_x$ once (i.e., $y \in \bar{B}_x \setminus \bar{B}_{x-1}$ for at most one x), the overall time complexity is $O(n)$. Also note that the B_x produced by this method must be contiguous segments of integers, and $\bar{B}_x$ must be in the form $\{1, \ldots, |\bar{B}_x|\}$, which allows simpler implementations (e.g. we only need to store $b_x := |\bar{B}_x|$ instead of $\bar{B}_x$). Refer to Algorithm 1 (which we call the *majorized alias* algorithm) for the precise description. ∎

	y	1	2	3	4	5
	$p(y)$	0.37	0.36	0.25	0.02	0
	$q(y)$	0.3	0.3	0.2	0.1	0.1
Step $x = 5$	r_y					0.1
	a_y					
Step $x = 4$	r_y				0.08	0.1
	a_y					
Step $x = 3$	r_y			0.05	0.08	0.1
	a_y					3
Step $x = 2$	r_y		0.02	0.05	0.08	0.1
	a_y				2	3
Step $x = 1$	r_y	0	0.02	0.05	0.08	0.1
	a_y		1	1	2	3

Table I

ALGORITHM 1 APPLIED ON $\vec{p} = [0.37, 0.36, 0.25, 0.02, 0]$, $\vec{q} = [0.3, 0.3, 0.2, 0.1, 0.1]$.

Algorithm 1 MAJORIZEDALIAS(p, q)

Input: pmf's p, q over $[n]$ such that $q \preceq p$, $p(1) \geq \cdots \geq p(n)$ and $q(1) \geq \cdots \geq q(n)$
Output: a_x, r_x for $x \in \{2, \ldots, n\}$

$b \leftarrow n$
$a_x \leftarrow 1$ for $x = 2, \ldots, n$
for $x \leftarrow n, n-1, \ldots, 2$ **do**
    $r_x \leftarrow q(x) - p(x)$
    while $r_x < 0$ **do**
        $r_x \leftarrow r_x + r_b$
        $a_b \leftarrow x$
        $b \leftarrow b - 1$
    end while
end for
    return $\{a_x\}, \{r_x\}$

Algorithm 1 has time complexity $O(n)$ since the block inside the while loop is executed at most n times (b decreases each time it is executed). We remark that Algorithm 1 reduces to the alias method [19] when q is the uniform distribution. The alias method is an efficient algorithm that can generate a random number following the distribution p over $[n]$, using a uniformly random integer in $[n]$ and a uniformly random real number in $[0, 1]$. The alias method requires an $O(n)$ (or $O(n \log n)$ if p is unsorted and needs to be sorted first) precomputation time to compute a_x and r_x satisfying (6) (where q is the uniform distribution over $[n]$). After the precomputation, each sample of $x \sim p$ can be generated in constant time by first generating $y \sim q$ independent of $z \sim \text{Unif}[0, 1]$, and then outputting $x = y$ if $z \geq r_y / q(y)$, $x = a_y$ if $z < r_y / q(y)$. While the alias method focuses only on the case where q is uniform (which guarantees $q \preceq p$), Algorithm 1 generalizes it to any q satisfying $q \preceq p$.

Table I shows Algorithm 1 applied on $\vec{p} = [0.37, 0.36, 0.25, 0.02, 0]$, $\vec{q} = [0.3, 0.3, 0.2, 0.1, 0.1]$. The values of $\{r_y\}, \{a_y\}$ for each iteration $x = 5, 4, 3, 2, 1$ in the algorithm are given. The red cells are cells with positions y in the interval $y \in [x..b]$, which are unfinished cells with r_y (the excess amount) computed, but a_y is not computed yet, i.e., it is not known where the excess amount r_y will be allocated (while Algorithm 1 initializes a_y to 1, here we assume a_y is initialized to be undefined for the sake of clarity). The green cells are cells y in the interval $y \in [b+1 .. n]$, which are finished cells with r_y, a_y computed. At each iteration x, we keep allocating the the excess amount of the right-most red cell to the current cell x (and change the right-most red cell to green), until $q(x)$ plus the total excess amount allocated to the current cell is at least $p(x)$. The amount in excess ($q(x)$ plus the total excess amount allocated to the current cell minus $p(x)$) is written to the r_x of the current cell.

We now give a sketch of the proof of Theorem 8. Let $q \preceq p$ with $p(1) \geq p(2) \geq \cdots$ and $q(1) \geq q(2) \geq \cdots$. Assume they have finite support for now, and consider them as probability vectors $\vec{p}, \vec{q} \in \mathbb{R}^n$. To show $q \times \mathrm{Geom}_{1/2} \sqsubseteq p$, it is equivalent to show that there exist right stochastic matrices $M_1, M_2, \ldots$ with $\{0,1\}$ entries such that $\vec{p} = \sum_{i=1}^{\infty} 2^{-i} \vec{q} M_i$. By Lemma 10, we have $\vec{p} = \vec{q} M$ for a stochastic matrix M where each row has at most one positive off-diagonal entry. For row x with off-diagonal entry $M_{x, a_x} = r_x/q(x)$, consider the binary representation of $r_x/q(x)$, and put a "1" at the position (x, a_x) of M_j if the j-th digit after the decimal point of the binary representation is "1" (otherwise put a "1" at the position (x, x)) for $j = 1, 2, \ldots$. This ensures that $M = \sum_{i=1}^{\infty} 2^{-i} M_i$, and hence the requirement is satisfied. The following is the complete proof for the case where the support size may be infinite.

Proof: Without loss of generality, assume p, q are pmf's over $\mathbb{N}$ with $p(1) \geq p(2) \geq \cdots$ and $q(1) \geq q(2) \geq \cdots$. Define pmf q_l by

$$q_l(x) = \begin{cases} q(x) & \text{if } x < l \\ \sum_{y=l}^{\infty} q(y) & \text{if } x = l \\ 0 & \text{if } x > l. \end{cases}$$

Define p_l similarly. Since $p \sqsubseteq p_l$, we have $q \preceq p \preceq p_l$. Fix l and let $n > l$ be large enough that $\sum_{y=n}^{\infty} q(y) \leq q(l)$, and hence the l largest entries of q_n are the same as the l largest entries of q. Since p_l has at most l nonzero entries, whether $q \preceq p_l$ holds only depend on the l largest entries of q. Hence, we have $q_n \preceq p_l$. By Lemma 10, there exists $a_x \in [n] \setminus \{x\}$ (we no longer have $a_x < x$ since we have to sort $q_n(x)$ in descending order before applying the lemma) and $r_x \in [0, q_n(x)]$ for $x = 1, \ldots, n$ such that

$$p_l(x) = q_n(x) - r_x + \sum_{y \in [n]:\, a_y = x} r_y$$

for any $x = 1, \ldots, n$. Define a mapping $g : \mathrm{supp}(q_n) \times \mathbb{N} \to [n]$ by

$$g(x, i) = \begin{cases} x & \text{if } 2^i r_x / q_n(x) \bmod 2 < 1 \\ a_x & \text{if } 2^i r_x / q_n(x) \bmod 2 \geq 1, \end{cases}$$

where $a \bmod b := a - b \lfloor a/b \rfloor$. Since $\sum_{i=1}^{\infty} 2^{-i} \mathbf{1}\{2^i r_x / q_n(x) \bmod 2 \geq 1\} = r_x / q_n(x)$ is the binary representation of $r_x / q_n(x)$, we have $\mathbf{P}(g(x, Z) = a_x) = r_x / q_n(x)$ and $\mathbf{P}(g(x, Z) = x) = 1 - r_x / q_n(x)$, where $Z \sim \mathrm{Geom}_{1/2}$. Let $X \sim q_n$ independent of Z, we have

$$\mathbf{P}(g(X, Z) = x)$$
$$= q_n(x) \mathbf{P}(g(x, Z) = x) + \sum_{y \in [n]:\, a_y = x} q_n(y) \mathbf{P}(g(y, Z) = a_y)$$
$$= q_n(x) - r_x + \sum_{y \in [n]:\, a_y = x} r_y$$
$$= p_l(x),$$

and hence $q_n \times \mathrm{Geom}_{1/2} \sqsubseteq p_l$. Since $q \sqsubseteq q_n$, we have $q \times \mathrm{Geom}_{1/2} \sqsubseteq q_n \times \mathrm{Geom}_{1/2} \sqsubseteq p_l$ by Proposition 5. Since $p_l(x) \to p(x)$ as $l \to \infty$ for any $x \in \mathbb{N}$, by Proposition 3, we have $q \times \mathrm{Geom}_{1/2} \sqsubseteq p$. ∎

Note that $(\bigwedge S) \times \mathrm{Geom}_{1/2}$ in Theorem 8 has an infinite support size or cardinality. If S is finite and the pmf's in S are over a set $\mathcal{X}$ which is finite, then we can reduce the cardinality to $|S|(|\mathcal{X}| - 1) + 1$ (without increasing its Rényi entropy), as given in the following theorem.

Theorem 11. *For a finite set of pmf's S with $|S| = m$, where the pmf's in S are over a finite set $\mathcal{X}$ with $|\mathcal{X}| = n$, there exists a pmf $q \in \tilde{\Gamma}(S)$ with $|\mathrm{supp}(q)| \leq m(n - 1) + 1$ and*

$$\left(\bigwedge S \right) \times \mathrm{CGeom}_{1/2, m} \preceq q,$$

where $\mathrm{CGeom}_{1/2, m}$ is the capped geometric distribution defined in (3). *As a result, the minimum Rényi entropy of couplings of S satisfies*

$$H_\alpha \left(\bigwedge S \right) \leq H_\alpha^*(S) \leq H_\alpha \left(\bigwedge S \right) + H_\alpha(\mathrm{CGeom}_{1/2, m}).$$

Note that $H(\mathrm{CGeom}_{1/2, m}) = 2 - 2^{2-m}$. Moreover, q and the aggregation maps for $q \sqsubseteq p$ for all $p \in S$ can be computed in $O(m^2 n + mn \log n)$ time (see Algorithm 3).

We remark that the cardinality bound $|\mathrm{supp}(q)| \leq m(n - 1) + 1$ is the same as that in [2], [5]. Therefore, the coupling in Theorem 11 gives a small Rényi entropy, without penalty on the cardinality.

To prove Theorem 11, we first show a lemma about coupling Bernoulli distributions.

Lemma 12. *For a finite set of pmf's S with $|S| = m$, where the pmf's in S are over $\{0,1\}$, there exists a pmf $q \in \tilde{\Gamma}(S)$ with $|\mathrm{supp}(q)| \le m + 1$ and $\mathrm{CGeom}_{1/2,m+1} \preceq q$. Moreover, q and the aggregation maps for $q \sqsubseteq p$ for all $p \in S$ can be computed in $O(m^2)$ time (see Algorithm 2).*

Proof: We prove the lemma by induction on m. The lemma is true when $m = 1$ since $\mathrm{CGeom}_{1/2,2} \preceq p$ for any pmf p over $\{0,1\}$. We now prove the lemma for m, assuming that the lemma is true for any smaller m. Let $S = \{p_1, \ldots, p_m\}$. Without loss of generality, assume p_1 attains the minimum of $\max\{p_i(0), p_i(1)\}$ for $i = 1, \ldots, m$. Let $\gamma := \max\{p_1(0), p_1(1)\}$. If $\gamma = 1$, all distributions in S are degenerate, and the lemma clearly holds, and hence we can assume $\gamma < 1$. Let $\tilde{p}_2, \ldots, \tilde{p}_m$ be pmf's over $\{0,1\}$ defined as

$$\tilde{p}_i(0) = \frac{p_i(0) - \gamma\mathbf{1}\{p_i(0) \ge p_i(1)\}}{1 - \gamma},$$

$$\tilde{p}_i(1) = \frac{p_i(1) - \gamma\mathbf{1}\{p_i(0) < p_i(1)\}}{1 - \gamma},$$

for $i = 2, \ldots, m$. Invoke the induction hypothesis to obtain a pmf $\tilde{q} \in \tilde{\Gamma}(\{\tilde{p}_2, \ldots, \tilde{p}_m\})$ over $[m]$ satisfying $\mathrm{CGeom}_{1/2,m} \preceq \tilde{q}$. Let q be a pmf over $[m+1]$ with $q(x) = (1-\gamma)\tilde{q}(x)$ for $x \le m$, and $q(m+1) = \gamma$. Since $\gamma \ge 1/2$, we have $\mathrm{CGeom}_{1/2,m+1} \preceq q$. It is left to show that $q \sqsubseteq p_i$ for $i = 2, \ldots, m$. For p_i, without loss of generality assume $p_i(0) \ge p_i(1)$. Since $\tilde{q} \sqsubseteq \tilde{p}_i$, there exists $A \subseteq [m]$ such that $\tilde{p}_i(1) = \sum_{x \in A} \tilde{q}(x)$. We have

$$p_i(1) = (1 - \gamma)\tilde{p}_i(1) = \sum_{x \in A} q(x),$$

and hence $q \sqsubseteq p_i$. Refer to Algorithm 2 (which we call the *Bernoulli splitting* algorithm) for the precise description of the algorithm. ∎

Algorithm 2 $\mathrm{BERNOULLISPLITTING}(\rho_1, \ldots, \rho_m)$

Input: $\rho_1, \ldots, \rho_m \in [0,1]$ (let $\rho_i = p_i(1)$)
Output: $\{q_x\}_{x \in [k]}$, $\{g_{i,x}\}_{i \in [m], x \in [k]}$ (where $k \le m + 1$)
   (let $q_x = q(x)$, $g_{i,x} = g_i(x) \in \{0,1\}$ for the aggregation mapping for $q \sqsubseteq p_i$)

$g_{i,x} \leftarrow 0$ for $i \in [m], x \in [k]$
$c \leftarrow 1$
$k \leftarrow 0$
while $c > 0$ **do**
   $\gamma \leftarrow \min_i \max\{\rho_i, c - \rho_i\}$
   $k \leftarrow k + 1$
   $q_k \leftarrow \gamma$
   for $i \leftarrow 1, \ldots, m$ **do**
      if $\rho_i \ge c/2$ **then**
         $g_{i,k} \leftarrow 1$
         $\rho_i \leftarrow \rho_i - \gamma$
      end if
   end for
   $c \leftarrow c - \gamma$
end while
   return $\{q_x\}, \{g_{i,x}\}$

Algorithm 2 has time complexity $O(m^2)$, since after each iteration of the while loop, the number of i's where $\rho_i \in \{0, c\}$ increases by at least one (letting $i^* = \mathrm{argmin}_i \max\{\rho_i, c - \rho_i\}$, then $\rho_{i^*} \in \{0, c\}$ after the iteration), and hence the number of iterations of the while loop is upper bounded by $m + 1$.

Figure 1 shows Algorithm 2 applied on $\{\rho_i\}_i = (0.175, 0.35, 0.6, 0.925)$. The graphs on top are ρ_i at each iteration of the while loop, and the graphs on the bottom show q_x and $g_{i,x}$ at each iteration. We can consider the problem as finding a set of sticks with lengths q_x which sum to c (initially $c = 1$), such that every ρ_i (we require $0 \le \rho_i \le c$) is the sum of the lengths of a subset of sticks. We use the following greedy approach. If the length of the longest stick is γ, then every ρ_i must satisfy either $\rho_i \ge \gamma$ if we use the stick to form ρ_i, or $\rho_i \le c - \gamma$ if we do not use the stick to form ρ_i (the regions of inadmissible ρ_i are shaded in gray on the graphs on top). Therefore, the longest possible length of the longest stick is $q_1 = \gamma = \min_i \max\{\rho_i, c - \rho_i\}$. For every ρ_i where $\rho_i \ge \gamma$ (or equivalently $\rho_i \ge c/2$), we set $g_{i,1} = 1$, meaning that we use
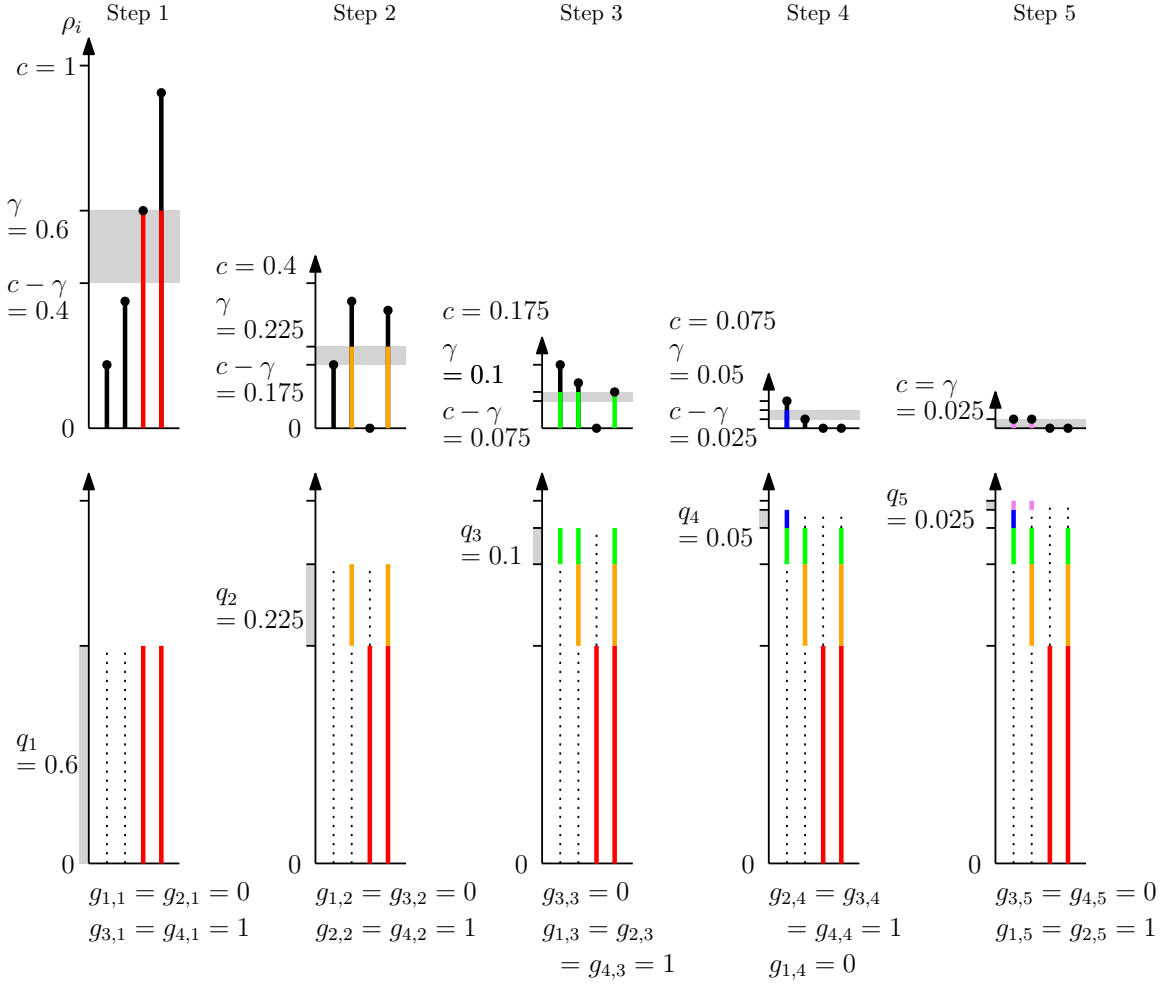
Figure 1. Algorithm 2 applied on $\{\rho_i\}_i = (0.175, 0.35, 0.6, 0.925)$.

the first stick to form ρ_i, and then set $\rho_i \leftarrow \rho_i - \gamma$ (the remainder of the length to be fulfilled by other sticks). We set $g_{i,1} = 0$ for the rest of ρ_i. Now the remaining total length of sticks become $c \leftarrow c - \gamma$. Repeat this process until $\rho_i = 0$ for all i.

Remark 13. Note that $\sum_{x>L} q(x) \leq 2^{-L}$ in Lemma 12 and Algorithm 2. Therefore, stopping Algorithm 2 after L steps reduces the time complexity to $O(mL)$, and incurs an error (in total variation distance) upper bounded by 2^{-L}, i.e., it computes a coupling of $\{\mathrm{Bern}(\tilde{\rho}_i)\}_{i \in [m]}$ instead of $\{\mathrm{Bern}(\rho_i)\}_{i \in [m]}$, where $|\tilde{\rho}_i - \rho_i| \leq 2^{-L}$. One can also replace "while $c > 0$" in Algorithm 2 to "while $c > \epsilon$" (and adjust q_x so they sum to 1) to set the desired error level.

We now prove Theorem 11.

Proof: Assume $p_1, \ldots, p_m$ are pmf's over $[n]$ with $p_i(1) \geq \cdots \geq p_i(n)$. Let $S = \{p_1, \ldots, p_m\}$, $\bar{q} := \bigwedge S$. By Lemma 10, let $a_{i,x} \in [x-1]$ and $r_{i,x} \in [0, \bar{q}(x)]$ for $x = 2, \ldots, n$, $i = 1, \ldots, m$ such that

$$p_i(x) = \bar{q}(x) - r_{i,x} + \sum_{x': a_{i,x'}=x} r_{i,x'} \tag{8}$$

for any $x = 1, \ldots, n$, $i = 1, \ldots, m$ (let $r_{i,1} = 0$).

Fix any $2 \leq x \leq n$. Since

$$\sum_{z=1}^{x-1} \bar{q}(z) = \min_{i \in [m]} \sum_{z=1}^{x-1} p_i(z),$$

there exists j such that $\sum_{z=1}^{x-1} \bar{q}(z) = \sum_{z=1}^{x-1} p_j(z)$, and hence by (8) and $a_{j,x} < x$,

$$0 = \sum_{z=1}^{x-1} (p_j(z) - \bar{q}(z))$$

$$= \sum_{z=1}^{x-1} \left(\bar{q}(z) - r_{j,z} + \sum_{x':\, a_{j,x'}=z} r_{j,x'} - \bar{q}(z) \right)$$

$$= -\sum_{z=1}^{x-1} r_{j,z} + \sum_{z=1}^{x-1} \sum_{x':\, a_{j,x'}=z} r_{j,x'}$$

$$= \sum_{z=1}^{x-1} \sum_{x' \geq x:\, a_{j,x'}=z} r_{j,x'}$$

$$\geq r_{j,x},$$

which means there exists j such that $r_{j,x} = 0$. Applying Lemma 12 on $\mathrm{Bern}_{r_{i,x}/\bar{q}(x)}$ for $i \neq j$, let $\tilde{q}_x \in \tilde{\Gamma}(\{\mathrm{Bern}_{r_{i,x}/\bar{q}(x)}\}_{i \in [m] \setminus \{j\}})$ be a pmf over $[m]$ with $\mathrm{CGeom}_{1/2,m} \preceq \tilde{q}_x$. We have $\tilde{q}_x \sqsubseteq \mathrm{Bern}_{r_{i,x}/\bar{q}(x)}$ for any $i \in [m]$ (this trivially holds when $i = j$).

For $x = 1$, let $\tilde{q}_1(1) = 1$. Since $r_{i,1} = 0$, we have $\tilde{q}_1 \in \tilde{\Gamma}(\{\mathrm{Bern}_{r_{i,1}/\bar{q}(1)}\}_{i \in [m] \setminus \{j\}})$ and $\mathrm{CGeom}_{1/2,m} \preceq \tilde{q}_1$.

Let q be a pmf over $[n] \times [m]$ defined by $q(x,y) := \bar{q}(x)\tilde{q}_x(y)$. For any $i \in [m]$, since $\tilde{q}_x \sqsubseteq \mathrm{Bern}_{r_{i,x}/\bar{q}(x)}$, there exists $A_x \subseteq [m]$ such that $\sum_{y \in A_x} \tilde{q}_x(y) = r_{i,x}/\bar{q}(x)$. Hence,

$$p_i(x) = \bar{q}(x) - r_{i,x} + \sum_{x' \in [n]:\, a_{i,x'}=x} r_{i,x'}$$

$$= \bar{q}(x) - \bar{q}(x) \sum_{y \in A_x} \tilde{q}_x(y) + \sum_{x' \in [n]:\, a_{i,x'}=x} \bar{q}(x') \sum_{y \in A_{x'}} \tilde{q}_{x'}(y)$$

$$= \sum_{y \in [m+1] \setminus A_x} q(x,y) + \sum_{x' \in [n]:\, a_{i,x'}=x} \sum_{y \in A_{x'}} q(x',y),$$

and thus

$$g_i(x,y) := \begin{cases} x & \text{if } y \notin A_x \\ a_{i,x} & \text{if } y \in A_x \end{cases}$$

is an aggregation map for $q \sqsubseteq p_i$. Since $|\mathrm{supp}(\tilde{q}_1)| = 1$, we have $|\mathrm{supp}(q)| \leq m(n-1) + 1$. We have $\bar{q} \times \mathrm{CGeom}_{1/2,m} \preceq q$ by Proposition 5. Refer to Algorithm 3 for the precise description of the algorithm. ∎

Algorithm 3 COMPUTECOUPLING$(p_1, \ldots, p_m)$

Input: pmf's $p_1, \ldots, p_m$ over $[n]$ with $p_i(1) \geq \cdots \geq p_i(n)$
Output: $\{q_x\}_{x \in [k]}$, $\{g_{i,x}\}_{i \in [m], x \in [k]}$ (where $k \leq m(n-1) + 1$)
    (let $q_x = q(x)$ for pmf q over $[k]$,
    $g_{i,x} = g_i(x) \in [n]$ for the aggregation mapping for $q \sqsubseteq p_i$)

$\bar{q} \leftarrow \bigwedge_{i \in [m]} p_i$
for $i \leftarrow 1, \ldots, m$ **do**
    $\{a_{i,x}\}_{x=2,\ldots,n}, \{r_{i,x}\}_{x=2,\ldots,n} \leftarrow$ MAJORIZEDALIAS$(p_i, \bar{q})$
    $r_{i,1} \leftarrow 0$
end for
$k \leftarrow 0$
for $x \leftarrow 1, \ldots, n$ **do**
    $\{\tilde{q}_y\}_{y \in [\tilde{k}]}, \{\tilde{g}_{i,y}\}_{i \in [m], y \in [\tilde{k}]} \leftarrow$ BERNOULLISPLITTING$(\{r_{i,x}/\bar{q}(x)\}_{i \in [m]})$
    $q_{k+y} \leftarrow \bar{q}(x)\tilde{q}_y$ for $y \in [\tilde{k}]$
    $g_{i,k+y} \leftarrow \mathbf{1}\{\tilde{g}_{i,y} = 0\}x + \mathbf{1}\{\tilde{g}_{i,y} = 1\}a_{i,x}$ for $i \in [m]$, $y \in [\tilde{k}]$
    $k \leftarrow k + \tilde{k}$
end for
    return $\{q_x\}, \{g_{i,x}\}$

Remark 14. If we perform the modification in Remark 13 (stopping Algorithm 2 after L steps), it would reduce the time complexity of Algorithm 3 to $O(mnL + mn \log n)$, the support bound to $|\mathrm{supp}(q)| \leq (L+1)(n-1) + 1$, but incur an error (in total variation distance on each $p \in S$) upper bounded by 2^{-L}, i.e., it computes a coupling of $\{\tilde{p}_i\}_{i \in [n]}$ instead of $\{p_i\}_{i \in [n]}$, where $d_{\mathrm{TV}}(p_i, \tilde{p}_i) \leq 2^{-L}$. In practical implementations, setting $L \approx 60$ will make the error negligible compared to floating-point error. Therefore, the practical running time complexity of Algorithm 3 is close to $O(mn \log n)$.

IV. Acknowledgement

References

[1] 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, no. 10, pp. 2464–2477, 2012.

[2] A. Painsky, S. Rosset, and M. Feder, "Memoryless representation of Markov processes," in *2013 IEEE International Symposium on Information Theory*. IEEE, 2013, pp. 2294–298.

[3] ——, "Innovation representation of stochastic processes with application to causal inference," *IEEE Transactions on Information Theory*, 2019.

[4] M. Kovačević, I. Stanojević, and V. Šenk, "On the entropy of couplings," *Information and Computation*, vol. 242, pp. 369–382, 2015.

[5] M. Kocaoglu, A. G. Dimakis, S. Vishwanath, and B. Hassibi, "Entropic causal inference," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[6] ——, "Entropic causality and greedy minimum entropy coupling," in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 1465–1469.

[7] F. Cicalese, L. Gargano, and U. Vaccaro, "How to find a joint probability distribution of minimum entropy (almost) given the marginals," in *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 2173–2177.

[8] ——, "Minimum-entropy couplings and their applications," *IEEE Transactions on Information Theory*, vol. 65, no. 6, pp. 3436–3451, 2019.

[9] L. Yu and V. Y. Tan, "Asymptotic coupling and its applications in information theory," *IEEE Transactions on Information Theory*, vol. 65, no. 3, pp. 1321–1344, 2018.

[10] M. Rossi, "Greedy additive approximation algorithms for minimum-entropy coupling problem," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 1127–1131.

[11] T. Roughgarden and M. Kearns, "Marginals-to-models reducibility," in *Advances in Neural Information Processing Systems*, 2013, pp. 1043–1051.

[12] Y. Han, O. Ordentlich, and O. Shayevitz, "Mutual information bounds via adjacency events," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 6068–6080, 2016.

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

[14] A. W. Marshall, I. Olkin, and B. C. Arnold, *Inequalities: theory of majorization and its applications*. Springer, 1979, vol. 143.

[15] A. Rényi, "On measures of entropy and information," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

[16] D. E. Knuth and A. C. Yao, "The complexity of nonuniform random number generation," *Algorithms and Complexity: New Directions and Recent Results*, pp. 357–428, 1976.

[17] J. R. Roche, "Efficient generation of random variables from biased coins," in *Proc. IEEE Int. Symp. Inf. Theory (papers in summary form only received)*, Jun 1991, pp. 169–169.

[18] T. S. Han and M. Hoshi, "Interval algorithm for random number generation," *IEEE Trans. Inf. Theory*, vol. 43, no. 2, pp. 599–611, Mar 1997.

[19] A. J. Walker, "An efficient method for generating discrete random variables with general distributions," *ACM Transactions on Mathematical Software (TOMS)*, vol. 3, no. 3, pp. 253–256, 1977.

[20] C. H. Bennett, P. W. Shor, J. Smolin, and A. V. Thapliyal, "Entanglement-assisted capacity of a quantum channel and the reverse Shannon theorem," *IEEE Trans. Inf. Theory*, vol. 48, no. 10, pp. 2637–2655, 2002.

[21] A. Winter, "Compression of sources of probability distributions and density operators," *arXiv preprint quant-ph/0208131*, 2002.

[22] P. Cuff, "Distributed channel synthesis," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7071–7096, Nov 2013.

[23] C. H. Bennett, I. Devetak, A. W. Harrow, P. W. Shor, and A. Winter, "The quantum reverse Shannon theorem and resource tradeoffs for simulating quantum channels," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2926–2959, May 2014.

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

[25] G. R. Kumar, C. T. Li, and A. El Gamal, "Exact common information," in *Proc. IEEE Int. Symp. Inf. Theory*, June 2014, pp. 161–165.

[26] C. T. Li and A. El Gamal, "A universal coding scheme for remote generation of continuous random variables," *IEEE Transactions on Information Theory*, vol. 64, no. 4, pp. 2583–2592, April 2018.

[27] A. El Gamal and Y.-H. Kim, *Network information theory*. Cambridge University Press, 2011.

[28] B. Hajek and M. Pursley, "Evaluation of an achievable rate region for the broadcast channel," *IEEE Trans. Inf. Theory*, vol. 25, no. 1, pp. 36–46, Jan 1979.

[29] F. Willems and E. van der Meulen, "The discrete memoryless multiple-access channel with cribbing encoders," *IEEE Trans. Inf. Theory*, vol. 31, no. 3, pp. 313–327, May 1985.

[30] C. E. Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[31] C. T. Li and A. El Gamal, "Strong functional representation lemma and applications to coding theorems," *IEEE Transactions on Information Theory*, vol. 64, no. 11, pp. 6967–6978, Nov 2018.

[32] M. Braverman and A. Garg, "Public vs private coin in bounded-round information," in *International Colloquium on Automata, Languages, and Programming*. Springer, 2014, pp. 502–513.

[33] S. I. Gel'fand and M. S. Pinsker, "Coding for channel with random parameters," *Probl. Contr. and Inf. Theory*, vol. 9, no. 1, pp. 19–31, 1980.

[34] C. T. Li, X. Wu, A. Ozgur, and A. El Gamal, "Minimax learning for remote prediction," in *2018 IEEE International Symposium on Information Theory (ISIT)*, June 2018, pp. 541–545.

[35] C. T. Li and V. Anantharam, "A unified framework for one-shot achievability via the Poisson matching lemma," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 942–946.

[36] A. Z. Broder, "On the resemblance and containment of documents," in *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*. IEEE, 1997, pp. 21–29.

[37] J. Kleinberg and E. Tardos, "Approximation algorithms for classification problems with pairwise relationships: Metric labeling and Markov random fields," *Journal of the ACM (JACM)*, vol. 49, no. 5, pp. 616–639, 2002.

[38] O. Angel and Y. Spinka, "Pairwise optimal coupling of multiple random variables," *arXiv preprint arXiv:1903.00632*, 2019.

[39] C. T. Li and V. Anantharam, "Pairwise multi-marginal optimal transport and embedding for earth mover's distance," *arXiv preprint arXiv:1908.01388*, 2019.

[40] M. S. Charikar, "Similarity estimation techniques from rounding algorithms," in *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*. ACM, 2002, pp. 380–388.

[41] B. Barak, M. Hardt, I. Haviv, A. Rao, O. Regev, and D. Steurer, "Rounding parallel repetitions of unique games," in *2008 49th Annual IEEE Symposium on Foundations of Computer Science*. IEEE, 2008, pp. 374–383.

[42] I. Sason, "Entropy bounds for discrete random variables via maximal coupling," *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 7118–7131, 2013.

[43] J. G. Propp and D. B. Wilson, "Exact sampling with coupled Markov chains and applications to statistical mechanics," *Random Structures & Algorithms*, vol. 9, no. 1-2, pp. 223–252, 1996.

[44] J. Propp and D. Wilson, "Coupling from the past: a user's guide," *Microsurveys in Discrete Probability*, vol. 41, pp. 181–192, 1998.

[45] M. Agueh and G. Carlier, "Barycenters in the Wasserstein space," *SIAM Journal on Mathematical Analysis*, vol. 43, no. 2, pp. 904–924, 2011.

[46] H. G. Kellerer, "Duality theorems for marginal problems," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 67, no. 4, pp. 399–432, 1984.

[47] W. Gangbo and A. Święch, "Optimal maps for the multidimensional Monge-Kantorovich problem," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 51, no. 1, pp. 23–45, 1998.

[48] B. Pass, "Uniqueness and Monge solutions in the multimarginal optimal transportation problem," *SIAM Journal on Mathematical Analysis*, vol. 43, no. 6, pp. 2758–2775, 2011.

[49] F. Cicalese and U. Vaccaro, "Supermodularity and subadditivity properties of the entropy on the majorization lattice," *IEEE Transactions on Information Theory*, vol. 48, no. 4, pp. 933–938, 2002.