

# Mixing Coefficients Between Discrete and Real Random Variables: Computation and Properties

Mehmet Eren Ahsen and M. Vidyasagar

## Abstract

In this paper we study the problem of estimating the mixing coefficients between two random variables. Three different mixing coefficients are studied, namely alpha-mixing, beta-mixing and phi-mixing coefficients. The random variables can either assume values in a finite set or the set of real numbers. We derive upper and lower bounds for both the alpha-mixing and the phi-mixing coefficients. Moreover, in case the marginal distributions of the two random variables are uniform, an exact expression is given for the phi-mixing coefficient. This situation arises when empirically generated samples are binned using percentile binning. We also prove analogs of the data-processing inequality from information theory for each of the three kinds of mixing coefficients. Then we move on to real-valued random variables, and show that by using percentile binning and allowing the number of bins to increase more slowly than the number of samples, we can generate empirical estimates that are consistent, i.e., converge to the true values as the number of samples approaches infinity.

## I. INTRODUCTION

The notion of independence of random variables is central to probability theory. In [7, p. 8], Kolmogorov says:

“Indeed, as we have already seen, the theory of probability can be regarded from the mathematical point of view as a special application of the general theory of additive set functions.

Department of Bioengineering, University of Texas at Dallas, 800 W. Campbell Road, Richardson, TX 75080; Emails: {Ahsen, M.Vidyasagar}@utdallas.edu. This work is supported by National Science Foundation Award #1001643.

and

“Historically, the independence of experiments and random variables represents the very mathematical concept that has given the theory of probability its peculiar stamp.”

In effect, Kolmogorov is saying that, if the notion of independence is removed, then probability theory reduces to just measure theory.

Independence is a binary concept: Either two random variables are independent, or they are not. It is therefore worthwhile to replace the concept of independence with a more nuanced measure that *quantifies* the extent to which given random variables are dependent. In the case of stationary stochastic processes, there are various notions of ‘mixing’, corresponding to long term asymptotic independence. These notions can be readily adapted to define various mixing coefficients between two random variables. Several such definitions are presented in [4, p. 3], out of which three are of interest to us, namely the  $\alpha$ -,  $\beta$ - and  $\phi$ -mixing coefficients. While the definitions themselves are well-known, there is very little work on actually *computing* (or at least estimating) these mixing coefficients in a given situation. The  $\beta$ -mixing coefficient is easy to compute but this is not the case for the  $\alpha$ - and the  $\phi$ -mixing coefficients.

Against this background, the present paper makes the following specific contributions:

- 1) For discrete random variables, simple upper and lower bounds are derived for both the  $\alpha$ - and the  $\phi$ -mixing coefficients.
- 2) In the special case where the discrete random variables have uniform marginal distributions, a closed-form formula is given for the  $\phi$ -mixing coefficient. This situation arises when two real-valued random variables are sampled, and the sampled values are discretized using percentile binning, that is, the end points of the grids are chosen such that the marginals are (nearly) uniform. It is well-known in the statistics literature that this kind of ‘data-dependent partitioning’, also referred as ‘partitioning into statistically equivalent blocks’, offers better performance than using a fixed partitioning for discretization; see the introduction of [9].
- 3) We study the case where  $X, Y, Z$  are discrete random variables, and  $X, Z$  are conditionally independent given  $Y$ , or equivalently,  $X \rightarrow Y \rightarrow Z$  is a short Markov chain. In this case a well-known inequality from information theory [3, p. 34] states that

$$I(X, Z) \leq \min\{I(X, Y), I(Y, Z)\}, \quad (1)$$

where  $I(\cdot, \cdot)$  denotes the mutual information. This inequality is usually referred to as the

‘data processing inequality (DPI)’. We state and prove analogs of the DPI for each of the  $\alpha$ -,  $\beta$ - and  $\phi$ -mixing coefficients.

- 4) Suppose  $X, Y$  are real-valued random variables whose joint distribution has a density with respect to the Lebesgue measure, and that  $\{(x_1, y_1), \dots, (x_l, y_l)\}$  are independent samples of  $(X, Y)$ . If we compute the empirical joint distribution of  $(X, Y)$  from these samples, then the Glivenko-Cantelli Lemma states that the empirical joint distribution converges with probability one to the true joint distribution; in other words, the empirical distribution gives a *consistent* estimate. However, it is shown here that if the empirical distribution is used to estimate the mixing coefficients, then with probability one both the estimated  $\beta$ -mixing coefficient and the estimated  $\phi$ -mixing coefficient approach one as  $l \rightarrow \infty$ , irrespective of what the true value might be. Thus a quantity derived from a consistent estimator need not itself be consistent.
- 5) On the other hand, if we bin the  $l$  samples into  $k_l$  bins such that the quantized versions of  $X$  and  $Y$  have nearly uniform distributions, and choose  $k_l$  in such a way that  $k_l \rightarrow \infty$  and  $k_l/l \rightarrow 0$  as  $l \rightarrow \infty$ , and a few technical conditions are satisfied, then the empirically estimated  $\alpha$ -,  $\beta$ - and  $\phi$ -mixing coefficients converge to their true values as  $l \rightarrow \infty$ , with probability one.

The problems of efficiently computing mixing coefficients and proving analogs of the data processing inequality are not just of academic interest. Recent work on reverse-engineering genome-wide interaction networks from gene expression data is based on using the  $\phi$ -mixing coefficient as a measure of the interaction between two genes; see [11]. If there are  $n$  genes in the study, this approach requires the computation of  $n^2$   $\phi$ -mixing coefficients. So for a typical genome-wide study involving 20,000 genes, it becomes necessary to compute 400 million  $\phi$ -mixing coefficients. Hence efficient computation is mandatory in order to have a practically viable implementation. The approach suggested in [13], [11] is to compute all pair-wise  $\phi$ -mixing coefficients, start with a complete directed graph on  $n$  nodes, and then to use the analog of the data processing inequality for the  $\phi$ -mixing coefficient to prune the network. The results presented in this paper provide the analytical justification for the approach in [11].

## II. DEFINITIONS OF MIXING COEFFICIENTS

The notion of mixing originated in an attempt to establish the law of large numbers for stationary stochastic processes that are not i.i.d. General definitions of the  $\alpha$ -,  $\beta$ - and  $\phi$ -mixing coefficients of a stationary stochastic process can be found, among other places, in [12, pp. 34-35]. The  $\alpha$ -mixing coefficient was introduced by Rosenblatt [10]. According to Doukhan [4, p. 5], Kolmogorov introduced the  $\beta$ -mixing coefficient, but it appeared in print for the first time in a paper published by some other authors. The  $\phi$ -mixing coefficient was introduced by Ibragimov [6].

Essentially, all notions of mixing try to quantify the idea that, in a stationary stochastic process of the form  $\{X_t\}_{t=-\infty}^{\infty}$ , the random variables  $X_t$  and  $X_\tau$  become more and more independent as  $|t - \tau|$  approaches infinity, in other words, there is an asymptotic long-term near-independence. However, these very general notions can be simplified and readily adapted to define mixing coefficients between a pair of random variables  $X$  and  $Y$ .<sup>1</sup> Though they can be defined for arbitrary random variables, in the interests of avoiding a lot of technicalities we restrict our attention in this paper to just two practically important cases: real-valued and discrete random variables. We first define mixing coefficients between real-valued random variables, and then between discrete random variables.

*Definition 1:* Suppose  $X$  and  $Y$  are real-valued random variables. Let  $\mathcal{B}$  denote the Borel  $\sigma$ -algebra of subsets of  $\mathbb{R}$ . Then we define

$$\alpha(X, Y) := \sup_{S, T \in \mathcal{B}} |\Pr\{X \in S \& Y \in T\} - \Pr\{X \in S\} \cdot \Pr\{Y \in T\}|. \quad (2)$$

$$\begin{aligned} \phi(X|Y) &:= \sup_{S, T \in \mathcal{B}} |\Pr\{X \in S|Y \in T\} - \Pr\{X \in S\}| \\ &= \sup_{S, T \in \mathcal{B}} \left| \frac{\Pr\{X \in S \& Y \in T\}}{\Pr\{Y \in T\}} - \Pr\{X \in S\} \right|. \end{aligned} \quad (3)$$

In applying the above definition, in case  $\Pr\{Y \in T\} = 0$ , we use the standard convention that

$$\Pr\{X \in S|Y \in T\} = \Pr\{X \in S\}.$$

<sup>1</sup>Strictly speaking, mixing is a property not of the random variables  $X$  and  $Y$ , but rather of the  $\sigma$ -algebras generated by  $X$  and  $Y$ . This is how they are defined in [4].

Note that the  $\alpha$ -mixing coefficient is symmetric:  $\alpha(X, Y) = \alpha(Y, X)$ . However, in general  $\phi(X|Y) \neq \phi(Y|X)$ .

The third coefficient, called the  $\beta$ -mixing coefficient, has a somewhat more elaborate definition, at least in the general case. Let  $\theta$  denote the probability measure of the joint random variable  $(X, Y)$ , and let  $\mu, \nu$  denote the marginal measures of  $X$  and  $Y$  respectively. Note that  $\theta$  is a measure on  $\mathbb{R}^2$  while  $\mu, \nu$  are measures on  $\mathbb{R}$ . If  $X$  and  $Y$  were independent, then  $\theta$  would equal  $\mu \times \nu$ , the product measure. With this in mind, we define

$$\beta(X, Y) = \rho(\theta, \mu \times \nu), \quad (4)$$

where  $\rho$  denotes the total variation distance between two measures. The  $\beta$ -mixing coefficient is also symmetric.

Next we deal with discrete random variables, and for this purpose we introduce some notation that is used throughout the remainder of the paper. The most important notational change is that, since probability distributions on finite sets are vectors, we use bold-face Greek letters to denote them, whereas we use normal Greek letters to denote measures on  $\mathbb{R}$  or  $\mathbb{R}^2$ . For each integer  $n$ , let  $\mathbb{S}_n$  denote the  $n$ -dimensional simplex. Thus

$$\mathbb{S}_n := \{\mathbf{v} \in \mathbb{R}^n : v_i \geq 0 \forall i, \sum_{i=1}^n v_i = 1\}.$$

If  $\mathbb{A} = \{a_1, \dots, a_n\}$  and  $\boldsymbol{\mu} \in \mathbb{S}_n$ , then  $\boldsymbol{\mu}$  defines a measure  $P_{\boldsymbol{\mu}}$  on the set  $\mathbb{A}$  according to

$$P_{\boldsymbol{\mu}}(S) = \sum_{i=1}^n \mu_i I_S(a_i),$$

where  $I_S(\cdot)$  denotes the indicator function of  $S$ . To avoid more notation, we will write  $\boldsymbol{\mu}(S)$  instead of the more precise  $P_{\boldsymbol{\mu}}(S)$ .

Suppose  $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{S}_n$  are probability distributions on a set  $\mathbb{A}$  of cardinality  $n$ . Then the **total variation distance** between  $\boldsymbol{\mu}$  and  $\boldsymbol{\nu}$  is defined as

$$\rho(\boldsymbol{\mu}, \boldsymbol{\nu}) := \max_{S \subseteq \mathbb{A}} |\boldsymbol{\mu}(S) - \boldsymbol{\nu}(S)|.$$

It is easy to give several equivalent closed-form formulas for the total variation distance.

$$\rho(\boldsymbol{\mu}, \boldsymbol{\nu}) = 0.5 \|\boldsymbol{\mu} - \boldsymbol{\nu}\|_1 = \sum_{i=1}^n (\mu_i - \nu_i)_+ = - \sum_{i=1}^n (\mu_i - \nu_i)_-,$$

where as usual  $(\cdot)_+$  and  $(\cdot)_-$  denote the nonnegative and the nonpositive parts of a number:

$$(x)_+ = \max\{x, 0\}, (x)_- = \min\{x, 0\}.$$

Now suppose  $\mathbb{A}, \mathbb{B}$  denotes sets of cardinality  $n, m$  respectively, and that  $\boldsymbol{\mu} \in \mathbb{S}_n, \boldsymbol{\nu} \in \mathbb{S}_m$ . Then the distribution  $\boldsymbol{\psi} \in \mathbb{S}_{nm}$  defined by  $\psi_{ij} = \mu_i \nu_j$  is called the **product distribution** on  $\mathbb{A} \times \mathbb{B}$ . In the other direction, if  $\boldsymbol{\theta} \in \mathbb{S}_{nm}$  is a distribution on  $\mathbb{A} \times \mathbb{B}$ , then  $\boldsymbol{\theta}_{\mathbb{A}} \in \mathbb{S}_n, \boldsymbol{\theta}_{\mathbb{B}} \in \mathbb{S}_m$  defined respectively by

$$(\boldsymbol{\theta}_{\mathbb{A}})_i := \sum_{j=1}^m \theta_{ij}, (\boldsymbol{\theta}_{\mathbb{B}})_j := \sum_{i=1}^n \theta_{ij}$$

are called the **marginal distributions** of  $\boldsymbol{\theta}$  on  $\mathbb{A}$  and  $\mathbb{B}$  respectively.

The earlier definitions of mixing coefficients become quite explicit in the case where  $X, Y$  are discrete random variables assuming values in the finite sets  $\mathbb{A}, \mathbb{B}$  of cardinalities  $n, m$  respectively. In this case it does not matter whether the ranges of  $X, Y$  are finite subsets of  $\mathbb{R}$  or some abstract finite sets. Definition 1 can now be restated in this context. Note that, since  $\mathbb{A}, \mathbb{B}$  are finite sets, the associated  $\sigma$ -algebras are just the power sets, that is, the collection of all subsets.

*Definition 2:* With the above notation, we define

$$\alpha(X, Y) := \max_{S \subseteq \mathbb{A}, T \subseteq \mathbb{B}} |\boldsymbol{\theta}(S \times T) - \boldsymbol{\mu}(S)\boldsymbol{\nu}(T)|, \quad (5)$$

$$\beta(X, Y) := \rho(\boldsymbol{\theta}, \boldsymbol{\mu} \times \boldsymbol{\nu}), \quad (6)$$

$$\phi(X|Y) := \max_{S \subseteq \mathbb{A}, T \subseteq \mathbb{B}} \left| \frac{\boldsymbol{\theta}(S \times T)}{\boldsymbol{\nu}(T)} - \boldsymbol{\mu}(S) \right|. \quad (7)$$

Whether  $X, Y$  are real-valued or discrete random variables, the mixing coefficients satisfy the following inequalities:

$$0 \leq \alpha(X, Y) \leq \beta(X, Y) \leq \min\{\phi(X|Y), \phi(Y|X)\} \leq \max\{\phi(X|Y), \phi(Y|X)\} \leq 1.$$

Also, the following statements are equivalent:

- 1)  $X$  and  $Y$  are independent random variables.
- 2)  $\alpha(X, Y) = 0$ .
- 3)  $\beta(X, Y) = 0$ .
- 4)  $\phi(X|Y) = 0$ .
- 5)  $\phi(Y|X) = 0$ .

### III. COMPUTATION OF MIXING COEFFICIENTS FOR DISCRETE RANDOM VARIABLES

In this section, we present explicit upper and lower bounds for the  $\alpha$ - and  $\phi$ -mixing coefficients, as well as an exact formula for the  $\phi$ -mixing coefficient in the case where one of the random variables has a uniform marginal distribution. This situation arises when a real-valued random variable is quantized using percentile binning.

From the definitions, it is obvious that  $\beta(X, Y)$  can be readily computed in closed form. As before, let us define  $\boldsymbol{\psi} = \boldsymbol{\mu} \times \boldsymbol{\mu}$  to be the product distribution of the two marginals, and define

$$\gamma_{ij} := \theta_{ij} - \psi_{ij}, \Gamma := [\gamma_{ij}] \in [-1, 1]^{n \times m}.$$

Then it is obvious that

$$\beta(X, Y) := \rho(\boldsymbol{\theta}, \boldsymbol{\psi}) = 0.5 \sum_{i=1}^n \sum_{j=1}^m |\gamma_{ij}| = \sum_{i=1}^n \sum_{j=1}^m (\gamma_{ij})_+ = - \sum_{i=1}^n \sum_{j=1}^m (\gamma_{ij})_-.$$

On the other hand, computing  $\alpha(X, Y)$  or  $\phi(X|Y)$  directly from Definition 2 would require  $2^{n+m}$  computations, since  $S, T$  must be allowed to vary over all subsets of  $\mathbb{A}, \mathbb{B}$  respectively. It is shown later that the number of computations can be brought down to  $O(2^m)$  but this is still exponential. Thus the objectives of the present section are to derive explicit upper and lower bounds for these mixing coefficients, and also to derive an exact formula for  $\phi(X|Y)$  in case  $\nu$  is the uniform distribution.

For this purpose we recall the definition of the matrix induced norm. For indices  $i$  and  $j$ , let  $\boldsymbol{\gamma}^i, \boldsymbol{\gamma}_j$  denote respectively the  $i$ -th row and  $j$ -th column of the matrix  $\Gamma$ . The quantity

$$\|\Gamma\|_{i1} := \max_{1 \leq j \leq m} \sum_{i=1}^n |\gamma_{ij}| = \max_{1 \leq j \leq m} \|\boldsymbol{\gamma}_j\|_1$$

is called the  $\ell_1$ -**induced matrix norm** of  $\Gamma$ . It is well-known that

$$\|\Gamma\|_{i1} = \max_{\|\mathbf{v}\|_1 \leq 1} \|\Gamma \mathbf{v}\|_1 = \max_{\mathbf{v} \neq \mathbf{0}} \frac{\|\Gamma \mathbf{v}\|_1}{\|\mathbf{v}\|_1}.$$

With this notation we are ready to state the main results of this section.

*Theorem 1:* We have that

$$0.5 \|\Gamma\|_{i1} \leq \alpha(X, Y) \leq 0.25m \|\Gamma\|_{i1}. \quad (8)$$

*Theorem 2:* We have that

$$\frac{0.5 \|\Gamma\|_{i1}}{\max_j \nu_j} \leq \phi(X|Y) \leq \frac{0.5 \|\Gamma\|_{i1}}{\min_j \nu_j}. \quad (9)$$

In particular, if  $\nu$  is the uniform distribution on  $\mathbb{B}$ , then

$$\phi(X|Y) = 0.5m\|\Gamma\|_{i1}. \quad (10)$$

In proving Theorems 1 and 2, the first step is to get rid of the absolute value signs in the definitions of the  $\alpha$ - and  $\phi$ -mixing coefficients.

*Theorem 3:* It is the case that

$$\alpha(X, Y) = \max_{S \subseteq \mathbb{A}, T \subseteq \mathbb{B}} [\theta(S \times T) - \mu(S)\nu(T)], \quad (11)$$

$$\phi(X|Y) = \max_{S \subseteq \mathbb{A}, T \subseteq \mathbb{B}} \left[ \frac{\theta(S \times T)}{\nu(T)} - \mu(S) \right]. \quad (12)$$

**Proof:** Define

$$\mathcal{R}_\alpha := \{\theta(S \times T) - \mu(S)\nu(T), S \subseteq \mathbb{A}, T \subseteq \mathbb{B}\}.$$

Then  $\mathcal{R}_\alpha$  is a subset of the real line consisting of at most  $2^{n+m}$  elements. Now it is claimed that the set  $\mathcal{R}_\alpha$  is symmetric; that is,  $x \in \mathcal{R}_\alpha$  implies that  $-x \in \mathcal{R}_\alpha$ . If this claim can be established, then (11) follows readily. So suppose  $x \in \mathcal{R}_\alpha$ , and choose  $S \subseteq \mathbb{A}, T \subseteq \mathbb{B}$  such that

$$\theta(S \times T) - \mu(S)\nu(T) = x.$$

Let  $S^c$  denote the complement of  $S$  in  $\mathbb{A}$ . Then, using the facts that

$$\mu(S^c) = 1 - \mu(S),$$

$$\theta(S^c \times T) = \theta(\mathbb{A} \times T) - \theta(S \times T) = \nu(T) - \theta(S \times T),$$

it is easy to verify that

$$\theta(S^c \times T) - \mu(S^c)\nu(T) = -x.$$

So  $\mathcal{R}_\alpha$  is symmetric and (11) follows. By analogous reasoning, the set

$$\mathcal{R}_\phi := \left\{ \frac{\theta(S \times T)}{\nu(T)} - \mu(S) : S \subseteq \mathbb{A}, T \subseteq \mathbb{B} \right\}$$

is also symmetric, which establishes (12). ■

To facilitate the proofs of Theorems 1 and 2, we introduce a map from the power set of  $\mathbb{A}$  into  $\{0, 1\}^n$ . For a subset  $S \subseteq \mathbb{A}$ , we define  $\mathbf{h}(S) \in \{0, 1\}^n$  by

$$h_i(S) = \begin{cases} 1, & \text{if } a_i \in S, \\ 0, & \text{if } a_i \notin S. \end{cases}$$



The map  $\mathbf{h} : 2^{\mathbb{B}} \rightarrow \{0, 1\}^m$  is defined analogously. With these definitions, it is obvious that, for  $S \subseteq \mathbb{A}, T \subseteq \mathbb{B}$ , we have

$$\boldsymbol{\mu}(S) = [\mathbf{h}(S)]^t \boldsymbol{\mu} = \boldsymbol{\mu}^t \mathbf{h}(S), \boldsymbol{\nu}(T) = [\mathbf{h}(T)]^t \boldsymbol{\nu} = \boldsymbol{\nu}^t \mathbf{h}(T),$$

$$\boldsymbol{\theta}(S \times T) = [\mathbf{h}(S)]^t \Theta \mathbf{h}(T),$$

where  $\Theta = [\theta_{ij}]$ . By replacing  $\mathbf{h}(S)$  and  $\mathbf{h}(T)$  by arbitrary binary vectors  $\mathbf{a} \in \{0, 1\}^n, \mathbf{b} \in \{0, 1\}^m$ , it readily follows from (11) and (12) that

$$\alpha(X, Y) = \max_{\mathbf{a} \in \{0, 1\}^n, \mathbf{b} \in \{0, 1\}^m} \mathbf{a}^t \Gamma \mathbf{b}, \quad (13)$$

$$\phi(X|Y) = \max_{\mathbf{a} \in \{0, 1\}^n, \mathbf{b} \in \{0, 1\}^m} \frac{\mathbf{a}^t \Gamma \mathbf{b}}{\boldsymbol{\nu}^t \mathbf{b}}. \quad (14)$$

Now we are in a position to prove Theorems 1 and 2.

**Proof of Theorem 1:** It is obvious that

$$\max_{\mathbf{a} \in \{0, 1\}^n, \mathbf{b} \in \{0, 1\}^m} \mathbf{a}^t \Gamma \mathbf{b} = \max_{\mathbf{b} \in \{0, 1\}^m} \max_{\mathbf{a} \in \{0, 1\}^n} \mathbf{a}^t \Gamma \mathbf{b}.$$

Now, for fixed  $\mathbf{b} \in \{0, 1\}^m$ , it is obvious that

$$\max_{\mathbf{a} \in \{0, 1\}^n} \mathbf{a}^t \Gamma \mathbf{b} = \sum_{i=1}^n (\gamma^i \mathbf{b})_+,$$

corresponding to the choice

$$a_i = \begin{cases} 1, & \text{if } \gamma^i \mathbf{b} \geq 0, \\ 0, & \text{if } \gamma^i \mathbf{b} < 0. \end{cases}$$

Therefore

$$\alpha(X, Y) = \max_{\mathbf{b} \in \{0, 1\}^m} \sum_{i=1}^n (\gamma^i \mathbf{b})_+. \quad (15)$$

Next, let  $\mathbf{e}$  denote a column vector consisting of all ones, with the subscript denoting its dimension, and observe that

$$\boldsymbol{\mu}^t = \mathbf{e}_n^t \Theta = \mathbf{e}_n^t \Psi \Rightarrow \mathbf{e}_n^t \Gamma = \mathbf{0}_n, \text{ similarly } \Gamma \mathbf{e}_m = \mathbf{0}_m.$$

Therefore, for any vector  $\mathbf{v} \in \mathbb{R}^m$ , it follows that

$$\begin{aligned}
\mathbf{e}_n^t \Gamma \mathbf{v} = 0 &\Rightarrow \sum_{i=1}^n \gamma^i \mathbf{v} = 0 \\
&\Rightarrow \sum_{i=1}^n (\gamma^i \mathbf{v})_+ + \sum_{i=1}^n (\gamma^i \mathbf{v})_- = 0 \\
&\Rightarrow \sum_{i=1}^n (\gamma^i \mathbf{v})_+ = - \sum_{i=1}^n (\gamma^i \mathbf{v})_- \\
&\Rightarrow \sum_{i=1}^n (\gamma^i \mathbf{v})_+ = 0.5 \sum_{i=1}^n |\gamma^i \mathbf{v}| = 0.5 \|\Gamma \mathbf{v}\|_1.
\end{aligned} \tag{16}$$

So in particular it follows that

$$\alpha(X, Y) = \max_{\mathbf{b} \in \{0,1\}^m} 0.5 \|\Gamma \mathbf{b}\|_1 \tag{17}$$

To prove the lower bound in (8), choose an index  $j_0 \in \{1, \dots, m\}$  such that  $\|\gamma_{j_0}\|_1 = \|\Gamma\|_{i1}$ . Then choose  $\mathbf{b}_0 \in \{0, 1\}^m$  to be the binary vector with  $b_{j_0} = 1$  and  $b_j = 0$  for all  $j \neq j_0$ . Now it follows from (16) that

$$\begin{aligned}
\sum_{i=1}^n (\gamma^i \mathbf{b}_0)_+ &= 0.5 \sum_{i=1}^n |\gamma^i \mathbf{b}_0| \\
&= 0.5 \sum_{i=1}^n |\gamma_{i,j_0}| = 0.5 \|\gamma_{j_0}\|_1 = 0.5 \|\Gamma\|_{i1}.
\end{aligned}$$

Hence the maximum over all  $\mathbf{b} \in \{0, 1\}^m$  is at least equal to this much.

To prove the upper bound in (8), observe from the definition that

$$\|\Gamma\|_{i1} = \max_{1 \leq j \leq m} \|\gamma_j\|_1.$$

Now for any  $\mathbf{b} \in \{0, 1\}^m$ , we have

$$\Gamma \mathbf{b} = \sum_{j=1}^m \gamma_j b_j.$$

Therefore

$$0.5 \|\Gamma \mathbf{b}\|_1 = 0.5 \left\| \sum_{j=1}^m \gamma_j b_j \right\|_1 \leq 0.5 \left[ \sum_{j=1}^m b_j \|\gamma_j\|_1 \right]. \tag{18}$$

The proof is completed by showing that an optimal  $\mathbf{b}$  can be chosen with no more than  $m/2$  nonzero entries. Choose a  $\mathbf{b}^* \in \{0, 1\}^m$  that achieves the maximum in (17). If  $\mathbf{b}^*$  has  $m/2$  or fewer nonzero entries, we are done, because we can substitute into (18) and conclude that

$$0.5 \|\Gamma \mathbf{b}^*\|_1 \leq 0.25m \max_{1 \leq j \leq m} \|\gamma_j\|_1 = 0.25m \|\Gamma\|_{i1}.$$

If  $\mathbf{b}^*$  has more than  $m/2$  nonzero entries, define  $\bar{\mathbf{b}} = \mathbf{e}_m - \mathbf{b}^*$ , and note that  $\bar{\mathbf{b}}$  has fewer than  $m/2$  nonzero entries. Also, since  $\Gamma \mathbf{e}_m = \mathbf{0}_m$ , it follows that  $\Gamma \bar{\mathbf{b}} = -\Gamma \mathbf{b}^*$ . So by earlier reasoning

$$\|\Gamma \bar{\mathbf{b}}\|_1 = \|\Gamma \mathbf{b}^*\|_1$$

and the bound follows. ■

**Proof of Theorem 2:** By reasoning analogous to that in the proof of Theorem 1, we arrive at

$$\begin{aligned} \phi(X|Y) &= \max_{\mathbf{a} \in \{0,1\}^n, \mathbf{b} \in \{0,1\}^m} \frac{\mathbf{a}^t \Gamma \mathbf{b}}{\boldsymbol{\nu}^t \mathbf{b}} \\ &= \max_{\mathbf{b} \in \{0,1\}^m} \max_{\mathbf{a} \in \{0,1\}^n} \frac{\mathbf{a}^t \Gamma \mathbf{b}}{\boldsymbol{\nu}^t \mathbf{b}} \\ &= \max_{\mathbf{b} \in \{0,1\}^m} \sum_{i=1}^n \left( \frac{\gamma^i \mathbf{b}}{\boldsymbol{\nu}^t \mathbf{b}} \right)_+. \end{aligned} \quad (19)$$

To prove the lower bound, choose an index  $j_0$  such that  $\|\gamma_{j_0}\|_1 = \|\Gamma\|_{i1}$ , and choose  $\mathbf{b}_0 \in \{0, 1\}^m$  such that  $b_{j_0} = 1$  and  $b_j = 0$  for all  $j \neq j_0$ . Then

$$\begin{aligned} \sum_{i=1}^n \left( \frac{\gamma^i \mathbf{b}_0}{\boldsymbol{\nu}^t \mathbf{b}_0} \right)_+ &= \frac{1}{\nu_{j_0}} \sum_{i=1}^n (\gamma_{i,j_0})_+ \\ &= \frac{0.5}{\nu_{j_0}} \sum_{i=1}^n |\gamma_{i,j_0}| \\ &= \frac{0.5 \|\Gamma\|_{i1}}{\nu_{j_0}} \\ &\geq \frac{0.5 \|\Gamma\|_{i1}}{\max_j \nu_j}. \end{aligned}$$

To prove the upper bound, note that for all  $\mathbf{b} \in \{0, 1\}^m$ , we have

$$\sum_{i=1}^n \frac{(\gamma^i \mathbf{b})_+}{\boldsymbol{\nu}^t \mathbf{b}} = 0.5 \sum_{i=1}^n \frac{|\gamma^i \mathbf{b}|}{\boldsymbol{\nu}^t \mathbf{b}} = 0.5 \frac{\|\Gamma \mathbf{b}\|_1}{\boldsymbol{\nu}^t \mathbf{b}}.$$

Now we change the variable of optimization from  $\mathbf{b}$  to  $\mathbf{v} := \text{Diag}(\boldsymbol{\nu})\mathbf{b}$ , and use the fact that

the induced matrix norm  $\|\cdot\|_{i1}$  is submultiplicative. This leads to

$$\begin{aligned}
\phi(X|Y) &= 0.5 \max_{\mathbf{b} \in \{0,1\}^m} \frac{\|\Gamma \mathbf{b}\|_1}{\boldsymbol{\nu} \mathbf{b}} \\
&\leq 0.5 \max_{\mathbf{b} \in \mathbb{R}^m} \frac{\|\Gamma \mathbf{b}\|_1}{|\boldsymbol{\nu} \mathbf{b}|} \\
&= 0.5 \max_{\mathbf{v} \in \mathbb{R}^m} \frac{\|\Gamma [\text{Diag}(\boldsymbol{\nu})]^{-1} \mathbf{v}\|_1}{\|\mathbf{v}\|_1} \\
&= 0.5 \|\Gamma [\text{Diag}(\boldsymbol{\nu})]^{-1}\|_{i1} \\
&\leq 0.5 \|\Gamma\|_{i1} \cdot \|[\text{Diag}(\boldsymbol{\nu})]^{-1}\|_{i1} \\
&= \frac{0.5 \|\Gamma\|_{i1}}{\min_j \nu_j}.
\end{aligned}$$

Finally, if  $\boldsymbol{\nu}$  is the uniform distribution, then  $\min_j \nu_j = \max_j \nu_j = 1/m$ . So the two inequalities in (9) become equalities. ■

For later use, we collect (15) and (19) and state them a separate theorem.

*Theorem 4:* With all notation as above, we have

$$\alpha(X, Y) = \max_{T \subseteq \mathbb{B}} \sum_{i=1}^n [\Pr\{X = i \& Y \in T\} - \Pr\{X = i\} \Pr\{Y \in T\}]_+. \quad (20)$$

and

$$\phi(X|Y) = \max_{T \subseteq \mathbb{B}} \sum_{i=1}^n [\Pr\{X = i|Y \in T\} - \Pr\{X = i\}]_+. \quad (21)$$

#### IV. DATA PROCESSING-TYPE INEQUALITIES FOR MIXING COEFFICIENTS

In this section we study the case where two random variables are conditionally independent given a third, and prove inequalities of the data processing-type for the associated mixing coefficients. The nomenclature ‘data processing-type’ is motivated by the well-known data processing inequality in information theory.

*Definition 3:* Suppose  $X, Y, Z$  are discrete random variables assuming values in finite sets  $\mathbb{A}, \mathbb{B}, \mathbb{C}$  respectively. Then  $X, Z$  are said to be conditionally independent given  $Y$  if

$$\Pr\{X = i \& Z = k | Y = j\} = \Pr\{X = i | Y = j\} \Pr\{Z = k | Y = j\}, \quad \forall i \in \mathbb{A}, j \in \mathbb{B}, k \in \mathbb{C}. \quad (22)$$

If  $X, Z$  are conditionally independent given  $Y$ , we denote this by  $(X \perp Z) | Y$ . Some authors also write this as ‘ $X \rightarrow Y \rightarrow Z$  is a short Markov chain’, ignoring the fact that the three random variables can belong to quite distinct sets. In this case, it makes no difference whether we write

$X \rightarrow Y \rightarrow Z$  or  $Z \rightarrow Y \rightarrow X$ , because it is obvious from (22) that conditional independence is a symmetric relationship. Thus

$$(X \perp Z)|Y \Leftrightarrow (Z \perp X)|Y.$$

Also, from the definition, it follows readily that if  $(X \perp Z)|Y$ , then

$$\Pr\{X \in S \& Z \in U | Y = j\} = \Pr\{X \in S | Y = j\} \Pr\{Z \in U | Y = j\}, \quad \forall S \subseteq \mathbb{A}, j \in \mathbb{B}, U \subseteq \mathbb{C}. \quad (23)$$

However, in general, it is *not true* that

$$\Pr\{X \in S \& Z \in U | Y \in T\} = \Pr\{X \in S | Y \in T\} \Pr\{Z \in U | Y \in T\}, \quad \forall S \subseteq \mathbb{A}, T \subseteq \mathbb{B}, U \subseteq \mathbb{C}.$$

In fact, by setting  $T = \mathbb{B}$ , it would follow from the above relationship that  $X$  and  $Z$  are independent, which is a stronger requirement than conditional independence.

Given two random variables  $X, Y$  with joint distribution  $\theta$  and marginal distributions  $\mu, \nu$  of  $X, Y$  respectively, the quantity

$$H(\mu) := - \sum_{i=1}^n \mu_i \log \mu_i$$

is called the **entropy** of  $\mu$ , with analogous definitions for  $H(\nu)$  and  $H(\theta)$ ; and the quantity

$$I(X, Y) = H(\mu) + H(\nu) - H(\theta)$$

is called the **mutual information** between  $X$  and  $Y$ . It is clear that  $I(X, Y) = I(Y, X)$ . The following well-known inequality, referred to as the **data-processing inequality**, is the motivation for the contents of this section; see [3, p. 34]. Suppose  $(X \perp Z)|Y$ . Then

$$I(X, Z) \leq \min\{I(X, Y), I(Y, Z)\}. \quad (24)$$

*Theorem 5:* Suppose  $(X \perp Z)|Y$ . Then

$$\alpha(X, Z) \leq \min\{\alpha(X, Y), \alpha(Y, Z)\}. \quad (25)$$

*Theorem 6:* Suppose  $(X \perp Z)|Y$ . Then

$$\beta(X, Z) \leq \min\{\beta(X, Y), \beta(Y, Z)\}. \quad (26)$$

*Theorem 7:* Suppose  $(X \perp Z)|Y$ . Then

$$\phi(X|Z) \leq \min\{\phi(X|Y), \phi(Y|Z)\}, \quad (27)$$

$$\phi(Z|X) \leq \min\{\phi(Z|Y), \phi(Y|X)\}. \quad (28)$$

**Proof of Theorem 5:** Let  $S \subseteq \mathbb{A}, U \subseteq \mathbb{C}$  be arbitrary, and define

$$r_\alpha(S, U) := \Pr\{X \in S \& Z \in U\} - \Pr\{X \in S\} \Pr\{Z \in U\}.$$

Then

$$\begin{aligned} r_\alpha(S, U) &= \sum_{j=1}^m [\Pr\{X \in S \& Y = j \& Z \in U\} - \Pr\{X \in S \& Y = j\} \Pr\{Z \in U\}] \\ &= \sum_{j=1}^m [\Pr\{X \in S|Y = j\} \Pr\{Z \in U|Y = j\} \Pr\{Y = j\} \\ &\quad - \Pr\{X \in S|Y = j\} \Pr\{Y = j\} \Pr\{Z \in U\}] \\ &= \sum_{j=1}^m \Pr\{X \in S|Y = j\} [\Pr\{Z \in U \& Y = j\} - \Pr\{Y = j\} \Pr\{Z \in U\}] \\ &\leq \sum_{j=1}^m \Pr\{X \in S|Y = j\} [\Pr\{Z \in U \& Y = j\} - \Pr\{Y = j\} \Pr\{Z \in U\}]_+ \\ &\leq \sum_{j=1}^m [\Pr\{Z \in U \& Y = j\} - \Pr\{Y = j\} \Pr\{Z \in U\}]_+ \\ &\leq \max_{U \subseteq \mathbb{C}} \sum_{j=1}^m [\Pr\{Z \in U \& Y = j\} - \Pr\{Y = j\} \Pr\{Z \in U\}]_+ \\ &= \alpha(Y, Z). \end{aligned}$$

Since  $S$  and  $U$  are arbitrary, this implies that  $\alpha(X, Z) \leq \alpha(Y, Z)$  whenever  $X \rightarrow Y \rightarrow Z$  is a short Markov chain. Since  $X \rightarrow Y \rightarrow Z$  is the same as  $Z \rightarrow Y \rightarrow X$ , it also follows that  $\alpha(Z, X) \leq \alpha(Y, X)$ . Finally, since  $\alpha$  is symmetric, the desired conclusion (25) follows. ■

**Proof of Theorem 6:** Suppose that  $\mathbb{A}, \mathbb{B}, \mathbb{C}$  have cardinalities  $n, m, l$  respectively. (The symbols  $n, m$  have been introduced earlier and now  $l$  is introduced.) Let  $\delta$  denote the joint distribution of  $(X, Y, Z)$ ,  $\zeta$  the joint distribution of  $(X, Z)$ ,  $\eta$  the joint distribution of  $(Y, Z)$ , and as before,  $\theta$  the joint distribution of  $(X, Y)$ . Let  $\xi$  the marginal distribution of  $Z$ , and as before, let  $\mu, \nu$  denote the marginal distributions of  $X$  and  $Y$ . Finally, define

$$c_{jk} = \frac{\eta_{jk}}{\nu_j} = \Pr\{Z = k|Y = j\}.$$

As can be easily verified, the fact that  $(X \perp Z)|Y$  (or (22)) is equivalent to

$$\delta_{ijk} = \frac{\theta_{ij}\eta_{jk}}{\nu_j} = \theta_{ij}c_{jk}, \quad \forall i, j, k.$$

Also note the following identities:

$$\sum_{i=1}^n \theta_{ij} = \nu_j, \sum_{j=1}^m \theta_{ij} = \mu_i, \sum_{j=1}^m \delta_{ijk} = \zeta_{ik}, \forall i, j, k.$$

Now it follows from the various definitions that

$$\begin{aligned} \beta(X, Z) &= \sum_{i=1}^n \sum_{k=1}^l (\zeta_{ik} - \mu_i \xi_k)_+ \\ &= \sum_{i=1}^n \sum_{k=1}^l \left( \sum_{j=1}^m (\delta_{ijk} - \theta_{ij} \xi_k) \right)_+ \\ &\leq \sum_{i=1}^n \sum_{k=1}^l \sum_{j=1}^m (\delta_{ijk} - \theta_{ij} \xi_k)_+ \\ &= \sum_{i=1}^n \sum_{k=1}^l \sum_{j=1}^m (\theta_{ij} c_{jk} - \theta_{ij} \xi_k)_+ \\ &= \sum_{k=1}^l \sum_{j=1}^m \left[ \sum_{i=1}^n \theta_{ij} \right] (c_{jk} - \xi_k)_+ \\ &= \sum_{k=1}^l \sum_{j=1}^m (\nu_j c_{jk} - \nu_j \xi_k)_+ \\ &= \sum_{k=1}^l \sum_{j=1}^m (\eta_{jk} - \nu_j \xi_k)_+ \\ &= \beta(Y, Z). \end{aligned}$$

Now the symmetry of  $\beta(\cdot, \cdot)$  serves to show that  $\beta(X, Z) \leq \beta(X, Y)$ . Putting both inequalities together leads to the desired conclusion.  $\blacksquare$

**Proof of Theorem 7:** Suppose  $(X \perp Z)|Y$ . Since the  $\phi$ -mixing coefficient is not symmetric, it is necessary to prove two distinct inequalities, namely: (i)  $\phi(X|Z) \leq \phi(X|Y)$ , and (ii)  $\phi(X|Z) \leq \phi(Y|Z)$ .

**Proof that  $\phi(X|Z) \leq \phi(X|Y)$ :** For  $S \subseteq \mathbb{A}$ , define

$$r_\phi(S) := \max_{T \subseteq \mathbb{B}} \Pr\{X \in S | Y \in T\},$$

and observe that

$$\phi(X|Y) = \max_{S \subseteq \mathbb{A}} [r_\phi(S) - \mu(S)].$$

For a given  $S \subseteq \mathbb{A}$ , choose  $T^* = T^*(S) \subseteq \mathbb{B}$  such that

$$\Pr\{X \in S | Y \in T^*\} = r_\phi(S).$$

Suppose  $U \subseteq \mathbb{C}$  is arbitrary. Then

$$\begin{aligned} \Pr\{X \in S \& Z \in U\} &= \sum_{j=1}^m \Pr\{X \in S \& Y = j \& Z \in U\} \\ &= \sum_{j=1}^m \Pr\{X \in S | Y = j\} \Pr\{Z \in U | Y = j\} \Pr\{Y = j\} \\ &= \sum_{j=1}^m \Pr\{X \in S | Y = j\} \Pr\{Z \in U \& Y = j\} \\ &\leq r_\phi(S) \sum_{j=1}^m \Pr\{Z \in U \& Y = j\} \\ &= r_\phi(S) \Pr\{Z \in U\}. \end{aligned}$$

Dividing both sides by  $\Pr\{Z \in U\}$  leads to

$$\Pr\{X \in S | Z \in U\} \leq r_\phi(S),$$

$$\Pr\{X \in S | Z \in U\} - \mu(S) \leq r_\phi(S) - \mu(S) \leq \phi(X|Y).$$

**Proof that  $\phi(X|Z) \leq \phi(Y|Z)$ :** Let us define

$$c(S, U) := \Pr\{X \in S | Z \in U\} - \mu(S),$$



and reason as follows:

$$\begin{aligned}
c(S, U) &= \Pr\{X \in S|Z \in U\} - \Pr\{X \in S\} \\
&= \sum_{j=1}^m [\Pr\{X \in S \& Y = j|Z \in U\} - \Pr\{X \in S \& Y = j\}] \\
&= \sum_{j=1}^m [\Pr\{X \in S|Y = j \& Z \in U\} \Pr\{Y = j|Z \in U\} - \Pr\{X \in S|Y = j\} \Pr\{Y = j\}] \\
&= \sum_{j=1}^m \Pr\{X \in S|Y = j\} [\Pr\{Y = j|Z \in U\} - \Pr\{Y = j\}] \\
&\leq \sum_{j=1}^m \Pr\{X \in S|Y = j\} [\Pr\{Y = j|Z \in U\} - \Pr\{Y = j\}]_+ \\
&\leq \sum_{j=1}^m [\Pr\{Y = j|Z \in U\} - \Pr\{Y = j\}]_+ \\
&\leq \max_{U \subseteq \mathbb{C}} \sum_{j=1}^m [\Pr\{Y = j \& Z \in U\} - \Pr\{Y = j\}]_+ \\
&= \phi(Y|Z).
\end{aligned}$$

Since the right side is independent of both  $S$  and  $U$ , the desired conclusion follows.  $\blacksquare$

## V. INCONSISTENCY OF AN ESTIMATOR FOR MIXING COEFFICIENTS

Suppose  $X, Y$  are real-valued random variables with some unknown joint distribution, and suppose we are given an infinite sequence of independent samples  $\{(x_i, y_i), i = 1, 2, \dots\}$ . The question studied in this section and the next is whether it is possible to construct empirical estimates of the various mixing coefficients that converge to the true values as the number of samples approaches infinity.

Let

$$\Phi_{X,Y}(a, b) = \Pr\{X \leq a \& Y \leq b\}$$

denote the true but unknown joint distribution function of  $X$  and  $Y$ , and let  $\Phi_X(\cdot), \Phi_Y(\cdot)$  denote the true but unknown marginal distribution functions of  $X, Y$  respectively. Using the samples  $\{(x_i, y_i), i = 1, 2, \dots\}$ , we can construct three ‘stair-case functions’ that are empirical estimates of  $\Phi_X, \Phi_Y$  and  $\Phi_{X,Y}$  based on the first  $l$  samples, as follows:

$$\hat{\Phi}_X(a; l) := \frac{1}{l} \sum_{i=1}^l I_{\{x_i \leq a\}}, \quad (29)$$

$$\hat{\Phi}_Y(b; l) := \frac{1}{l} \sum_{i=1}^l I_{\{y_i \leq b\}}, \quad (30)$$

$$\hat{\Phi}_{X,Y}(a, b; l) := \frac{1}{l} \sum_{i=1}^l I_{\{x_i \leq a \& y_i \leq b\}}, \quad (31)$$

where as usual  $I$  denotes the indicator function. Thus  $\hat{\Phi}_X(a; l)$  counts the fraction of the first  $l$  samples that are less than or equal to  $a$ , and so on. With this construction, a well-known result called the Glivenko-Cantelli lemma (see [5], [2] or [8, p. 20]) states that the empirical estimates converge uniformly and almost surely to their true functions as the number of samples  $l \rightarrow \infty$ . Thus  $\hat{\Phi}_{X,Y}$  is a consistent estimator of the true joint distribution. Thus one might be tempted to think that an empirical estimate of any (or all) of the three mixing coefficients based on  $\hat{\Phi}_{X,Y}$  will also converge to the true value as  $l \rightarrow \infty$ . The objective of this brief section is to show that this is not so. Hence estimates of mixing coefficients derived from a consistent estimator of the joint distribution need themselves be consistent.

*Theorem 8:* Suppose  $\hat{\Phi}_{X,Y}$  is defined by (31), and that  $x_i \neq x_j$  and  $y_i \neq y_j$  whenever  $i \neq j$ . Let  $\hat{\beta}_l$  denote the  $\beta$ -mixing coefficient associated with the joint distribution  $\hat{\Phi}_{X,Y}(\cdot, \cdot; l)$ . Then  $\hat{\beta}_l = (l - 1)/l$ .

**Proof:** Fix the integer  $l$  in what follows. Note that the empirical distribution  $\hat{\Phi}_{X,Y}(\cdot, \cdot; l)$  depends only the totality of the  $l$  samples, and not the order in which they are generated. Without loss of generality, we can replace the samples  $(x_1, \dots, x_n)$  by their ‘order statistics’, that is, the same samples arranged in increasing order, and do the same for the  $y_i$ . Thus the assumption is that  $x_1 < x_2 < \dots < x_l$  and similarly  $y_1 < y_2 < \dots < y_l$ . With this convention, the empirical samples will be of the form  $\{(x_1, y_{\pi(1)}), \dots, (x_l, y_{\pi(l)})\}$  for some permutation  $\pi$  of  $\{1, \dots, l\}$ . Therefore the probability measure associated with the empirical distribution  $\hat{\Phi}$  is purely atomic, with jumps of magnitude  $1/l$  at the points  $\{(x_1, y_{\pi(1)}), \dots, (x_l, y_{\pi(l)})\}$ . So we can simplify matters by replacing the real line on the  $X$ -axis by the finite set  $\{x_1, \dots, x_l\}$ , and the real line on the  $Y$ -axis by the finite set  $\{y_1, \dots, y_l\}$ . With this redefinition, the joint distribution  $\theta$  assigns a weight of  $1/l$  to each of the points  $(x_i, y_{\pi(i)})$  and a weight of zero to all other points  $(x_i, y_j)$  whenever  $j \neq \pi(i)$ , while the marginal measures  $\mu, \nu$  of  $X$  and  $Y$  will be uniform on the respective finite sets. Thus the product measure  $\mu \times \nu$  assigns a weight of  $1/l^2$  to each of the  $l^2$  grid points  $(x_i, y_j)$ . From this, it is easy to see that

$$\hat{\beta}_l = \rho(\theta, \mu \times \nu) = (l - 1)/l.$$

This is the desired conclusion. ■

*Corollary 1:* Suppose the true but unknown distribution  $\Phi_{X,Y}$  has density with respect to the Lebesgue measure. Then  $\hat{\beta}_l \rightarrow 1, \hat{\phi}_l \rightarrow 1$  almost surely as  $l \rightarrow \infty$ .

**Proof:** If the true distribution has a density, then it is nonatomic, which means that with probability one, samples will be pairwise distinct. It now follows from Theorem 8 that

$$\hat{\phi}_l \geq \hat{\beta}_l = \frac{l-1}{l} \rightarrow 1 \text{ as } l \rightarrow \infty.$$

This is the desired conclusion. ■

## VI. CONSISTENT ESTIMATORS FOR MIXING COEFFICIENTS

The objective of the present section is to show that a simple modification of the ‘naive’ algorithm proposed in Section V does indeed lead to consistent estimates, provided appropriate technical conditions are satisfied.

The basic idea behind the estimators is quite simple. Suppose that one is given samples  $\{(x_i, y_i), i \geq 1\}$  generated independently and at random from an unknown joint probability measure  $\theta \in \mathcal{M}(\mathbb{R}^2)$ . Given  $l$  samples, choose an integer  $k_l$  of bins. Divide the real line into  $k_l$  intervals such that each bin contains  $\lfloor l/k_l \rfloor$  or  $\lfloor l/k_l \rfloor + 1$  samples for both  $X$  and  $Y$ . In other words, carry out percentile binning of both random variables. One way to do this (but the proof is not dependent on how precisely this is done) is as follows: Define  $m_l = \lfloor l/k_l \rfloor, r = l - k_l m_l$ , and place  $m_l + 1$  samples in the first  $r$  bins and  $m_l$  samples in the next  $m_l - r$  bins. This gives a way of discretizing the real line for both  $X$  and  $Y$  such that the discretized random variables have nearly uniform marginals. With this binning, compute the corresponding joint distribution, and the associated empirical estimates of the mixing coefficients. The various theorems below show that, subject to some regularity conditions, the empirical estimates produced by this scheme do indeed converge to their right values with probability one as  $l \rightarrow \infty$ , *provided that*  $m_l \rightarrow \infty$ , or equivalently,  $k_l/l \rightarrow 0$ , as  $l \rightarrow \infty$ . In other words, in order for this theorem to apply, the number of bins must increase more slowly than the number of samples, so that the number of samples per bin must approach infinity. In contrast, in Theorem 8, we have effectively chosen  $k_l = l$  so that each bin contains precisely one sample, which explains why that approximation scheme does not work.

To state the various theorems, we introduce a little bit of notation, and refer the reader to [1] for all concepts from measure theory that are not explicitly defined here. Let  $\mathcal{M}(\mathbb{R}), \mathcal{M}(\mathbb{R}^2)$

denote the set of all measures on  $\mathbb{R}$  or  $\mathbb{R}^2$  equipped with the Borel  $\sigma$ -algebra. Recall that if  $\theta, \eta \in \mathcal{M}(\mathbb{R})$  or  $\mathcal{M}(\mathbb{R}^2)$ , then  $\theta$  is said to be **absolutely continuous** with respect to  $\eta$ , denoted by  $\theta \ll \eta$ , if for every measurable set  $E$ ,  $\eta(E) = 0 \Rightarrow \theta(E) = 0$ .

Next, let  $\theta$  denote the joint probability measure of  $(X, Y)$ , and let  $\mu, \nu$  denote the marginal measures. Thus, for every measurable<sup>2</sup> subset  $S \subseteq \mathbb{R}$ , the measure  $\mu(S)$  is defined as  $\theta(S \times \mathbb{R})$  and similarly for all  $T \subseteq \mathbb{R}$ , the measure  $\nu(T)$  is defined as  $\theta(\mathbb{R} \times T)$ . Now the key assumption made here is that *the joint measure  $\theta$  is absolutely continuous with respect to the product measure  $\mu \times \nu$* . In the case of discrete random variables, this assumption is automatically satisfied. Suppose that for some pair of indices  $i, j$ , it is the case that  $\mu_i \cdot \nu_j = 0$ . Then either  $\mu_i = 0$  or  $\nu_j = 0$ . If  $\mu_i = 0$ , then it follows from the identity  $\sum_{j'} \theta_{ij'} = \mu_i$  that  $\theta_{ij'} = 0$  for all  $j'$ , and in particular  $\theta_{ij} = 0$ . Similarly if  $\nu_j = 0$ , then it follows from the identity  $\sum_{i'} \theta_{i'j} = \nu_j$  that  $\theta_{i'j} = 0$  for all  $i'$ , and in particular  $\theta_{ij} = 0$ . In either case it follows that  $\theta_{ij} = 0$ , so that  $\theta \ll \mu \times \nu$ . However, in the case of real random variables, this need not be so. For example, replace  $\mathbb{R} \times \mathbb{R}$  by the unit square, and let  $\theta$  be the diagonal measure. Then both marginals  $\mu, \nu$  are the uniform measures on the unit interval, and the product  $\mu \times \nu$  is the uniform measure on the unit square – and  $\theta$  is singular with respect to the uniform measure.

Next we introduce symbols for the various densities. Since  $\theta \ll \mu \times \nu$ , it follows that  $\theta$  has a Radon-Nikodym derivative with respect to  $\mu \times \nu$ , which is denoted by  $f(\cdot, \cdot)$ . So for any sets  $S, T \subseteq \mathbb{R}$ , it follows that

$$\theta(S \times T) = \int_S \int_T f(x, y) d\nu(y) d\mu(x) = \int_T \int_S f(x, y) d\mu(x) d\nu(y).$$

For any  $T \subseteq \mathbb{R}$  with  $\nu(T) > 0$ , the conditional probability  $\Pr\{X \in S | Y \in T\}$  is given by

$$\begin{aligned} \Pr\{X \in S | Y \in T\} &= \frac{\Pr\{X \in S \& Y \in T\}}{\Pr\{Y \in T\}} = \frac{\theta(S \times T)}{\nu(T)} \\ &= \int_S \left[ \int_T \frac{f(x, y)}{\nu(T)} d\nu(y) \right] d\mu(x). \end{aligned}$$

*Theorem 9:* With the above notation and conditions, the empirically estimated  $\beta$ -mixing coefficient  $\hat{\beta}_l$  converges almost surely to the true value  $\beta$  as  $l \rightarrow \infty$ , provided that  $k_l \rightarrow \infty$  and  $k_l/l \rightarrow 0$  as  $l \rightarrow \infty$ .

<sup>2</sup>Hereafter we drop this adjective; it is assumed that all sets that are encountered are measurable.

*Theorem 10:* Suppose that the density  $f(\cdot, \cdot)$  belongs to  $L_\infty(\mathbb{R}^2)$ , and that  $k_l/l \rightarrow 0$  as  $l \rightarrow \infty$ . Then the empirically estimated  $\alpha$ -mixing coefficient  $\hat{\alpha}_l$  converges almost surely to the true value  $\alpha$  as  $l \rightarrow \infty$ , and the empirically estimated  $\phi$ -mixing coefficient  $\hat{\phi}_l$  converges almost surely to the true value  $\phi$  as  $l \rightarrow \infty$ .

Note that the density  $f(\cdot, \cdot) \in L_1(\mathbb{R}^2, \mu \times \nu)$ . So the sequence of empirical estimates  $\hat{\beta}_l$  converges to its true value without any additional technical conditions. The sequences of empirical estimates  $\hat{\alpha}_l$  and  $\hat{\phi}_l$  converge to their true values provided the density  $f$  is bounded almost everywhere. This condition is intended to ensure that conditional densities do not ‘blow up’. In the case of discrete variables, we have already seen that the condition  $\theta \ll \mu \times \nu$  holds automatically, which means that the ‘density’  $f_{ij}\theta_{ij}/(\mu_i\nu_j)$  is always well-defined. Since there are only finitely many values of  $i$  and  $j$ , this ratio is also bounded. However, in the case of real-valued random variables, this condition needs to be imposed explicitly.

The proofs of these two theorems are based on arguments in [9], [14]. In the proof of Theorem 9, we can use those arguments as they are, whereas in the proof of Theorem 10, we need to adapt them. To facilitate the discussion, we first reprise the relevant results from [9], [14].

*Definition 4:* Let  $(\Omega, \mathcal{F})$  be a measurable space, and let  $Q$  be a probability measure on  $(\Omega, \mathcal{F})$ . Suppose  $\{I_1, \dots, I_L\}$  is a finite partition of  $\Omega$ , and that  $\{I_1^{(m)}, \dots, I_L^{(m)}\}$  is a sequence of partitions of  $\Omega$ . Then  $\{I_1^{(m)}, \dots, I_L^{(m)}\}$  is said to **converge to**  $\{I_1, I_L\}$  with respect to  $Q$  if, for every probability measure  $P$  on  $(\Omega, \mathcal{F})$  such that  $P \ll Q$ , it is the case that

$$P(I_i^{(m)}) \rightarrow P(I_i) \text{ as } m \rightarrow \infty.$$

See [14, Definition 1].

*Theorem 11:* Suppose  $Q$  is a probability measure on  $(\mathbb{R}, \mathcal{B})$  that is absolutely continuous with respect to the Lebesgue measure,  $L$  is a fixed integer, and that  $\{I_1, \dots, I_L\}$  is an equiprobable partitioning of  $\mathbb{R}$ . In other words, choose numbers

$$-\infty = a_0 < a_1 < \dots < a_{L-1} < a_L = +\infty$$

such that the semi-open intervals  $I_i = (a_{i-1}, a_i]$  satisfy

$$Q(I_i) = 1/L, i = 1, \dots, L.$$

Suppose  $\{y_1, \dots, y_m\}$  are i.i.d. samples generated in accordance with  $Q$ , and that  $m = l_m T$  with  $l_m \in \mathbb{N}$ , an integer. Let  $\{I_1^{(m)}, \dots, I_L^{(m)}\}$  denote the empirical equiprobable partitioning

associated with the samples  $\{y_1, \dots, y_m\}$ . Then  $\{I_1^{(m)}, \dots, I_L^{(m)}\}$  converges to  $\{I_1, \dots, I_L\}$  with respect to  $Q$  as  $m \rightarrow \infty$ .

**Proof:** See [14, Lemma 1].

*Theorem 12:* Let  $(\Omega, \mathcal{F})$  be a measurable space, and let  $Q$  be a probability measure on  $(\Omega, \mathcal{F})$ . Suppose  $\{I_1^{(m)}, \dots, I_L^{(m)}\}$  is a sequence of partitions of  $\Omega$  that converges with respect to  $Q$  to another partition  $\{I_1, I_L\}$  as  $m \rightarrow \infty$ . Suppose  $\{x_1, \dots, x_n\}$  are i.i.d. samples generated in accordance with a probability measure  $P \ll Q$ , and let  $P_n$  the empirical measure generated by these samples. Then

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} P_n(I_i^{(m)}) = P(I_i), \text{ a.s. } \forall i.$$

**Proof:** See [14, Lemma 2].

Before proceeding to the proofs of the two theorems, we express the three mixing coefficients in terms of the densities. As stated in (4), we have that

$$\beta(X, Y) = 0.5 \int_{\mathbb{R}} \int_{\mathbb{R}} |f(x, y) - 1| d\mu(x) d\nu(y). \quad (32)$$

Here we take advantage of the fact that the ‘density’ of  $\mu$  with respect to itself is one, and similarly for  $\nu$ . Next, as in Theorem 3, we can drop the absolute value signs in the definitions of  $\alpha(X, Y)$  and of  $\phi(X|Y)$ . Therefore the various mixing coefficients can be expressed as follows:

$$\alpha(X, Y) = \sup_T \sup_S \int_S \int_T [f(x, y) - 1] d\nu(y) d\mu(x), \quad (33)$$

$$\phi(X, Y) = \sup_T \sup_S \int_S \left[ \int_T \frac{f(x, y)}{\nu(T)} d\nu(y) - 1 \right] d\mu(x). \quad (34)$$

Now, for each fixed set  $T$ , let us define signed measures  $\kappa_T$  and  $\delta_T$  as follows:

$$\begin{aligned} \kappa_T(x) &= \int_T [f(x, y) - 1] d\nu(y), \\ \delta_T(x) &= \int_T \frac{f(x, y)}{\nu(T)} d\nu(y) - 1, \end{aligned}$$

and associated support sets

$$A_+(T) = \{x \in \mathbb{R} : \kappa_T(x) \geq 0\}, B_+(T) = \{x \in \mathbb{R} : \delta_T(x) \geq 0\}.$$

Then it is easy to see that, for each fixed set  $T$ , the supremum in (33) is achieved by the choice  $S = A_+(T)$  while the supremum in (34) is achieved by the choice  $S = B_+(T)$ . Therefore

$$\alpha(X, Y) = \sup_T \int_{A_+(T)} \kappa_T(x) d\mu(x) = \sup_T \int_{\mathbb{R}} [\kappa_T(x)]_+ d\mu(x), \quad (35)$$

$$\phi(X|Y) = \sup_T \int_{B_+(T)} \delta_T(x) d\mu(x) = \sup_T \int_{\mathbb{R}} [\delta_T(x)]_+ d\mu(x). \quad (36)$$

These formulas are the continuous analogs of (20) and (21) respectively.

**Proof of Theorem 9:** For a fixed integer  $L \geq 2$ , choose real numbers

$$-\infty = a_0 < a_1 < \dots < a_{L-1} < a_L = +\infty,$$

$$-\infty = b_0 < b_1 < \dots < b_{L-1} < b_L = +\infty$$

such that the semi-open intervals  $I_i = (a_{i-1}, a_i]$ ,  $J_i = (b_{i-1}, b_i]$  satisfy

$$\mu(I_i) = 1/L, \nu(J_i) = 1/L, i = 1, \dots, L.$$

Now define the equiprobable partition of  $\mathbb{R}^2$  consisting of the  $L \times L$  grid  $\{I_i \times J_j, i, j = 1, \dots, L\}$ . Next, based on the  $l$ -length empirical sample  $\{(x_1, y_1), \dots, (x_l, y_l)\}$ , construct empirical marginal distributions  $\hat{\mu}$  for  $X$  and  $\hat{\nu}$  for  $Y$ . Based on these empirical marginals, divide both the  $X$ -axis  $\mathbb{R}$  and  $Y$ -axis  $\mathbb{R}$  into  $L$  bins each having nearly equal fractions of the  $l$  samples in each bin. This gives an empirical  $L \times L$  partitioning of  $\mathbb{R}^2$ , which is denoted by  $\{I_i^{(L)} \times J_j^{(L)}, i, j = 1, \dots, L\}$ . Using this grid, compute the associated empirical joint distribution  $\hat{\theta}_l$  on  $\mathbb{R}^2$ . Then the proof of [14, Lemma 1] can be adapted to show that the empirical partition  $\{I_i^{(L)} \times J_j^{(L)}, i, j = 1, \dots, L\}$  converges to the true partition  $\{I_i \times J_j, i, j = 1, \dots, L\}$  as  $l \rightarrow \infty$ , with respect to the product measure  $\mu \times \nu$ . The only detail that differs from [14] is the computation of the so-called ‘growth function’. Given a set  $A \subseteq \mathbb{R}^2$  of cardinality  $m$ , the number of different ways in which this set can be partitioned by a rectangular grid of dimension  $L \times L$  is called the growth function, denoted by  $\Delta_m$ . It is shown in [14, Eq. (15)] that when the partition consists of  $L$  intervals and the set being partitioned is  $\mathbb{R}$ , then  $\Delta_m$  is given by the combinatorial parameter

$$\Delta_m = \binom{m+L}{L} = \frac{(m+L)!}{m!L!}.$$

It is also shown in [14, Eq. (21)] that

$$\frac{1}{m} \log \left[ \binom{m+L}{L} \right] \leq 2mh(1/L),$$

where  $h(\cdot)$  is defined by

$$h(x) = -x \log x - (1-x) \log(1-x), \forall x \in (0, 1).$$

When  $\mathbb{R}$  is replaced by  $\mathbb{R}^2$  and a set of  $L$  intervals is replaced by a grid of  $L^2$  rectangles, it is easy to see that the growth function is *no larger than*

$$\Delta_m \leq \left[ \binom{m+L}{L} \right]^2.$$

Therefore

$$\frac{\log \Delta_m}{m} \leq 4mh(1/L).$$

In any case, since  $L$ , the number of grid elements, approaches  $\infty$  as  $l \rightarrow \infty$ , it follows that the growth condition proposed in [9] is satisfied. Therefore the empirical partition converges to the true partition as  $l \rightarrow \infty$ .

Next, let  $\{I_i \times J_j, i, j = 1, \dots, L\}$  denote, as before, the true equiprobable  $L \times L$  gridding of  $\mathbb{R}^2$ . Suppose that, after  $l$  samples  $(x_r, y_r), r = 1, \dots, l$  have been drawn, the data is put into  $k_l$  bins. Then the expression (32) defining the true  $\beta$ -mixing coefficient can be rewritten as

$$\beta(X, Y) = 0.5 \sum_{i=1}^{k_l} \sum_{j=1}^{k_l} \int_{I_i} \int_{J_j} |f(x, y) - 1| d\mu(x) d\nu(y).$$

Now suppose  $l$  is an exact multiple of  $k_l$ . Then the empirical estimate based on the  $k_l \times k_l$  empirical grid can be written as

$$\hat{\beta}_l = 0.5 \sum_{i=1}^{k_l} \sum_{j=1}^{k_l} |C_{ij} - 1| k_l^{-2},$$

where  $C_{ij}$  denotes the number of samples  $(x_r, y_r)$  in the  $ij$ -th cell of the *empirical* (not true) equiprobable grid. If  $l$  is not an exact multiple of  $k_l$ , then some bins will have  $\lfloor l/k_l \rfloor$  elements while other bins will have  $\lfloor l/k_l \rfloor + 1$  elements. As a result, the term  $k_l^{-2}$  gets replaced by  $(s_i t_j)/l^2$  where  $s_i$  is the number of samples in  $I_i^{(l)}$  and  $t_j$  is the number of samples in  $J_j^{(l)}$ . Now, just as in [14, Eq. (36) *et seq.*], the error  $|\hat{\beta}_l - \beta(X, Y)|$  can be bounded by the sum of two errors, the first of which is caused by the fact that the empirical equiprobable grid is not the same as the true equiprobable grid (the term  $e_1$  of [14]), and the second is the error caused by approximating an integral by a finite sum over the true equiprobable grid (the term  $e_2$  of [14]). Out of these, the first error term goes to zero as  $l \rightarrow \infty$  because, if  $k_l/l \rightarrow 0$  so that each bin contains increasingly many samples, the empirical equiprobable grid converges to the true equiprobable grid. The second error terms goes to zero because the integrand in (32) belongs to  $L_1(\mathbb{R}^2, \mu \times \nu)$ , as shown in [14, Eq. (37)]. ■



**Proof of Theorem 10:** The main source of difficulty here is that, whereas the expression for  $\beta(X, Y)$  involves just a single integral, the expressions for  $\alpha(X, Y)$  and for  $\phi(X, Y)$  involve the supremum over all sets  $T \subseteq \mathbb{R}$ . Thus, in order to show that the empirical estimates converge to the true values, we must show not only that empirical estimates of integrals of the form  $\int_{\mathbb{R}}[\kappa_T]_+ d\mu(x)$  and  $\int_{\mathbb{R}}[\delta_T]_+ d\mu(x)$  converge to their correct values for each fixed set  $T$ , but also that the convergence is in some sense uniform with respect to  $T$ . This is where we use the boundedness of the density  $f(\cdot, \cdot)$ . The details are fairly routine modifications of arguments in [14]. Specifically, (switching notation to that of [14]), suppose that in their Equation (27), we have not just one measure  $\mu$ , but rather a family of measures  $\mu_T$ , indexed by  $T$ , and suppose there exists a finite constant  $c$  such that for every set  $S$  we have  $\mu_T(S) \leq cQ(S)$ . Then it follows from Equation (27) *et seq.* of [14] that

$$\mu_T((a_i \wedge a_i^m, a_i \vee a_i^m]) \leq cQ((a_i \wedge a_i^m, a_i \vee a_i^m]), \forall T.$$

Therefore

$$\lim_{m \rightarrow \infty} \sup_T \mu_T((a_i \wedge a_i^m, a_i \vee a_i^m]) = 0.$$

With this modification, the rest of the proof in [14] can be mimicked to show the following: In the interests of brevity, define

$$r_T = \int_{\mathbb{R}} [\kappa_T]_+ d\mu(x)$$

and let  $\hat{r}_{T,l}$  denote its empirical approximation. Then, using the above modification of the argument in [14], it follows that

$$\lim_{l \rightarrow \infty} \sup_T |r_T - \hat{r}_{T,l}| = 0.$$

As a consequence,

$$\lim_{l \rightarrow \infty} \sup_T \hat{r}_{T,l} = \sup_T r_T = \alpha(X, Y).$$

The proof for the  $\phi$ -mixing coefficient is entirely similar. ■

## VII. CONCLUDING REMARKS

In this paper we have studied the problem of estimating the mixing coefficients between two random variables. Three different mixing coefficients were studied, namely  $\alpha$ -mixing,  $\beta$ -mixing and  $\phi$ -mixing coefficients. The random variables can either assume values in a finite set or

the set of real numbers. We derived upper and lower bounds for both the  $\alpha$ -mixing and the  $\phi$ -mixing coefficients. Moreover, in case the marginal distributions of the two random variables are uniform, an exact expression was given for the  $\phi$ -mixing coefficient. This situation arises when empirically generated samples are binned using percentile binning. We also proved analogs of the data-processing inequality from information theory for each of the three kinds of mixing coefficients. Then we moved on to real-valued random variables, and showed that, even though the empirically estimated joint distribution converges to the true joint distribution, estimates of the  $\beta$ - or  $\phi$ -mixing coefficients based on the empirical joint distribution converges to 1 under mild conditions. However, by using percentile binning and allowing the number of bins to increase more slowly than the number of samples, we can generate empirical estimates that converge to the true values.

In general both the  $\alpha$ - and the  $\phi$ -mixing coefficient are solutions of integer programming problems, as shown in (17) and (19) respectively. So it would be interesting to explore whether the computation of these mixing coefficients is NP-hard. Also, it is clear from these same two equations that it is possible to construct a convex programming relaxation of (17) and a nonlinear programming relaxation of (19). It would be interesting to analyze how close, if at all, the solutions of these relaxed problems are to those of the original integer programming problems.

The later parts of the paper [14] contain some proposals on how to speed up the convergence of the empirical estimates of the Kullback-Leibler divergence between two unknown measures. It might be worthwhile to explore whether similar speed-ups can be found for the algorithms proposed here for estimating mixing coefficients from empirical data.

## REFERENCES

- [1] Sterling K. Berberian, *Measure and Integration*, Chelsea, New York, 1970.
- [2] F. P. Cantelli, “Sulla determinazione empirica delle legge di probabilità”, *Giornali dell’Istituto Italia degli Attuari*, 4, 421-424, 1933.
- [3] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, (Second Edition) John Wiley, New York, 2006.
- [4] Paul Doukhan, *Mixing: Properties and Examples*, Springer-Verlag, Heidelberg, 1994.
- [5] V. I. Glivenko, “Sulla determinazione empirica delle legge di probabilità”, *Giornali dell’Istituto Italia degli Attuari*, 4, 92-99, 1933.
- [6] I. A. Ibragimov, “Some limit theorems for stationary processes”, *Theory of Probability and its Applications*, 7, 349-382, 1962.

- [7] A. N. Kolmogorov, *Foundations of Probability*, (Second English Edition), Chelsea, New York, 1956.
- [8] Michel Loève, *Probability Theory I*, (4th Edition), Springer-Verlag, Heidelberg, 1977.
- [9] Bela Lugosi and Andrew Nobel, “Consistency of data-driven histogram methods for density estimation and classification”, *The Annals of Statistics*, 24(2), 687-706, 1996.
- [10] M. Rosenblatt, “A central limit theorem and a strong mixing condition”, *Proc. Nat’l. Acad. Sci.*, 42(1), 43-47. January 1956.
- [11] Nitin Kumar Singh et al., “Reverse engineering gene interaction networks using the phi mixing coefficient”, under preparation; preliminary version *arXiv*, 1288.4066.
- [12] M. Vidyasagar, *Learning and Generalization: With Applications to Neural Networks and Control Systems*, Springer-Verlag, London, 2003.
- [13] M. Vidyasagar, “Probabilistic methods in cancer biology”, *European Journal of Control*, 17(5-6), 483-511, September-December 2011.
- [14] Qing Wang, Sanjeev R. Kulkarni and Sergio Verdú, “Divergence estimation of continuous distributions based on data-dependent partitions”, *IEEE Transactions on Information Theory*, 51(9), 3064-3074, September 2005.