2011 50th IEEE Conference on Decision and Control and
European Control Conference (CDC-ECC)
Orlando, FL, USA, December 12-15, 2011

A Metric Between Probability Distributions on Finite Sets of Different Cardinalities

M. Vidyasagar, Fellow IEEE

***Abstract*— With increasing use of digital control it is natural to view control inputs and outputs as stochastic processes assuming values over finite alphabets rather than in a Euclidean space. As control over networks becomes increasingly common, data compression by reducing the size of the input and output alphabets without losing the fidelity of representation becomes relevant. This requires us to define a notion of distance between two stochastic processes assuming values in distinct sets, possibly of different cardinalities. If the two processes are i.i.d., then the problem becomes one of defining a metric between two probability distributions over distinct finite sets of possibly different cardinalities. This is the problem addressed in the present paper. A metric is defined in terms of a joint distribution on the product of the two sets, which has the two given distributions as its marginals, and has minimum entropy. Computing the metric exactly turns out to be NP-hard. Therefore an efficient greedy algorithm is presented for finding an upper bound on the distance.**

I. Introduction

Suppose we view a control system as an input-output map where the input signal is a sequence $\{u_t\}$ assuming values in some finite set U, while the output signal is a sequence $\{y_t\}$ assuming values in another finite set Y. In this setting, the problem of order reduction is quite different in nature from the traditional order reduction problem, where the emphasis is on reducing the dimension of the (Euclidean) state space. If the system has some element of randomness in it, we should view $\{(u_t, y_t)\}$ as a stochastic process assuming values in the set $U \times Y$.[1] For the purposes of controller design, it would be worthwhile to know whether the finely quantized inputs and outputs can be replaced by a coarser quantization without losing too much accuracy in the representation. Such considerations become particularly germane in the problem of control over networks, whereby the plant and controller may be connected only through a noisy channel. This type of order reduction would require approximating the original stochastic process by another one assuming values in a set of smaller cardinality $U' \times Y'$. The approximation can be quantified by defining a metric distance between two stochastic processes assuming values in distinct sets (of different cardinalities). So far as the author is aware, no such metric is available in the literature. The closest the author has been able to find is a paper by Ornstein [15] in the inaugural issue of the *Annals of Probability*, in which he defines a metric distance between two stochastic processes assuming values in a *common finite set*.

Our analysis is based on information theory. The use of information-theoretic methods in the controls community has a long history, going back at least to [16] if not much earlier. In this paper, we define a metric distance between two distributions on distinct finite sets by maximizing their mutual information. It turns out that actually computing the metric distance between two probability distributions is an NP-hard problem as it can be reduced to a nonstandard bin-packing problem. Therefore we develop efficient greedy algorithm. Specifically, we can compute an upper bound on the distance in $O((n + m^2) \log m)$ operations where n and m are the cardinalities of the two sets with $n \geq m$.

Cecil & Ida Green Chair, Erik Jonsson School of Engineering & Computer Science, The University of Texas at Dallas, 800 W. Campbell Road, Richardson, TX 75080; email: M.Vidyasagar@utdallas.edu. This research was supported by National Science Foundation Award #1001643.

[1] By this we mean that at each instant of time t, (u_t, y_t) belongs to the set $U \times Y$.

II. The Variation of Information Metric

A. Concepts from Information Theory

Throughout the paper, we shall use the symbols $\mathbb{A}, \mathbb{B}, \mathbb{C}$ for finite sets of cardinality n, m, l respectively. The symbols X, Y, Z denote random variables assuming values in $\mathbb{A}, \mathbb{B}, \mathbb{C}$ respectively. The symbols $\boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{\xi}$ denote probability distributions on the sets $\mathbb{A}, \mathbb{B}, \mathbb{C}$ respectively. Though the elements of these sets could be any abstract entities, to avoid notational clutter we shall write $\mathbb{A} = \{1, \ldots, n\}$ instead of the more precise $\mathbb{A} = \{a_1, \ldots, a_n\}$ etc. Let $\mathbf{e}$ denote the column vector of all one's, and the subscript denote its dimension. Thus $\mathbf{e}_n$ is a column vector of n one's. A matrix $P \in [0,1]^{m \times n}$ is said to be **stochastic** if $P\mathbf{e}_n = \mathbf{e}_m$, that is, for each row, the sum of all columns equals one. The set of $m \times n$ stochastic matrices is denoted by $\mathbb{S}_{m \times n}$. If we take the degenerate case of $m = 1$, then the symbol $\mathbb{S}_n = \mathbb{S}_{1 \times n}$ denotes the set of nonnegative (row) vectors that add up to one. Clearly $\mathbb{S}_n$ can be identified with the set $\mathcal{M}(\mathbb{A})$ of all probability distributions on $\mathbb{A}$.

Suppose X, Y are random variables assuming values in $\mathbb{A}, \mathbb{B}$ respectively, and let $\boldsymbol{\theta} \in \mathcal{M}(\mathbb{A} \times \mathbb{B})$ denote their joint distribution. For each index i between 1 and n, let $\mathbf{p}_i$ denote the conditional distribution of Y given that $X = i$. That is

$$p_{ij} = \frac{\theta_{ij}}{\sum_{j'=1}^{m} \theta_{ij'}}.$$

Note that the matrix $P = [p_{ij}]$ belongs to $\mathbb{S}_{n \times m}$, and the i-th row of P, denoted by $\mathbf{p}_i$, belongs to $\mathbb{S}_m$ for each i. If we represent the joint distribution of X and Y by an $n \times m$ matrix $\Theta = [\theta_{ij}]$ where $\theta_{ij} = \Pr\{X = i \& Y = j\}$, then we

can write

$$P = [\mathrm{Diag}(\boldsymbol{\phi})]^{-1}\Theta, \tag{1}$$

where $\mathrm{Diag}(\boldsymbol{\phi})$ represents the $n \times n$ diagonal matrix with $\phi_1, \ldots, \phi_n$ as the diagonal elements. Suppose we now define $Q \in \mathbb{S}_{m\times n}$ by

$$q_{ji} = \Pr\{X = i | Y = j\}.$$

Then it is easy to see that the following identities hold:

$$\Theta = \mathrm{Diag}(\boldsymbol{\phi})P = Q^T\mathrm{Diag}(\boldsymbol{\psi}), \tag{2}$$

$$Q = [\mathrm{Diag}(\boldsymbol{\psi})]^{-1}P^T\mathrm{Diag}(\boldsymbol{\phi}). \tag{3}$$

Now we introduce various concepts from information theory. All the concepts introduced below are discussed in [6, Chapter 2]. The function $h : [0, 1] \to \mathbb{R}_+$ is defined by $h(r) = -r \log r$, with the standard convention that $h(0) = 0$. Note that h is continuously differentiable except at $r = 0$, and that $h'(r) = -(1 + \log r)$. The symbol H denotes the Shannon entropy of a probability distribution. Thus if $\boldsymbol{\phi} \in \mathbb{S}_n$, then

$$H(\boldsymbol{\phi}) = -\sum_{i=1}^{n} \phi_i \log \phi_i = \sum_{i=1}^{n} h(\phi_i).$$

We define the **conditional entropy** of Y given X as

$$\begin{aligned} H(Y|X) &= \sum_{i=1}^{n} \phi_i H(\mathbf{p}_i) = \sum_{i=1}^{n} \phi_i \sum_{j=1}^{m} h(p_{ij}) \\ &= -\sum_{i=1}^{n} \phi_i \sum_{j=1}^{m} p_{ij} \log p_{ij}, \end{aligned}$$

where $\mathbf{p}_i$ denotes the i-th row of the matrix P. With this definition the identities

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \tag{4}$$

hold. The **mutual information** between X and Y is defined as

$$\begin{aligned} I(X, Y) &= H(X) + H(Y) - H(X, Y) \\ &= H(Y) - H(Y|X) = H(X) - H(X|Y). \end{aligned}$$

B. Setting Up the Problem

Suppose X, Y are random variables assuming values in the sets $\mathbb{A}, \mathbb{B}$ respectively, with distributions $\boldsymbol{\phi}, \boldsymbol{\psi}$ respectively. We ask: *What is the maximum possible mutual information between X and Y?* Clearly this is equivalent to asking the question: What is a (or the) distribution $\boldsymbol{\theta}$ on $\mathbb{A} \times \mathbb{B}$ that has minimum entropy, while satisfying the boundary conditions $\boldsymbol{\theta}_{\mathbb{A}} = \boldsymbol{\phi}, \boldsymbol{\theta}_{\mathbb{B}} = \boldsymbol{\psi}$?

Definition 1: Given sets $\mathbb{A}, \mathbb{B}$ with $|\mathbb{A}| = n, |\mathbb{B}| = m$, and given $\boldsymbol{\phi} \in \mathbb{S}_n, \boldsymbol{\psi} \in \mathbb{S}_m$, define

$$W(\boldsymbol{\phi}, \boldsymbol{\psi}) := \min_{\boldsymbol{\theta} \in \mathcal{M}(\mathbb{A}\times\mathbb{B})} \{H(\boldsymbol{\theta}) : \boldsymbol{\theta}_{\mathbb{A}} = \boldsymbol{\phi}, \boldsymbol{\theta}_{\mathbb{B}} = \boldsymbol{\psi}\}, \tag{5}$$

$$V(\boldsymbol{\phi}, \boldsymbol{\psi}) := W(\boldsymbol{\phi}, \boldsymbol{\psi}) - H(\boldsymbol{\phi}). \tag{6}$$

It is obvious that

$$W(\boldsymbol{\psi}, \boldsymbol{\phi}) = W(\boldsymbol{\phi}, \boldsymbol{\psi}), V(\boldsymbol{\psi}, \boldsymbol{\phi}) = V(\boldsymbol{\phi}, \boldsymbol{\psi}) + H(\boldsymbol{\phi}) - H(\boldsymbol{\psi}), \tag{7}$$

where the second identity follows from (4).

C. The Variation of Information Metric

We begin by defining a metric between *random variables*, and then move on to distributions.

Definition 2: Given two random variables X, Y, the **variation of information** between them is defined as

$$v(X, Y) = H(X|Y) + H(Y|X). \tag{8}$$

This measure is introduced in [13], [14] where it is referred to as the 'variation of information' metric between random variables. So we retain the same nomenclature, though our metric is between probability distributions.

Theorem 1: The function $v(\cdot, \cdot)$ satisfies the axioms of a pseudometric. Thus v has the properties that for all random variables X, Y, Z, we have $v(X, Y) \geq 0$, $v(X, Y) = v(Y, X)$, and $v(X, Y) \leq v(X, Z) + v(Y, Z)$.

Proof: It is obvious that $v(X, Y) \geq 0$, and it follows from (8) that $v(X, Y) = v(Y, X)$. To show that $v(\cdot, \cdot)$ satisfies the triangle inequality, we make use of the easily-proved inequality

$$\begin{aligned} H(X|Y) &\leq H(X, Z|Y) = H(Z|Y) + H(X|Z, Y) \\ &\leq H(Z|Y) + H(X|Z). \end{aligned} \tag{9}$$

To prove the triangle inequality, invoke the one-sided triangle inequality (9) and observe that

$$\begin{aligned} v(X, Y) &= H(X|Y) + H(Y|X) \\ &\leq H(X|Z) + H(Z|Y) + H(Y|Z) + H(Z|X) \\ &= v(X, Z) + v(Y, Z). \end{aligned}$$

This completes the proof. ■

Now we turn the above pseudometric *between random variables* into a pseudometric *between probability distributions*.

Definition 3: Given two probability distributions $\boldsymbol{\phi} \in \mathbb{S}_n, \boldsymbol{\psi} \in \mathbb{S}_m$, the **variation of information metric** between them is defined as

$$d(\boldsymbol{\phi}, \boldsymbol{\psi}) = V(\boldsymbol{\phi}, \boldsymbol{\psi}) + V(\boldsymbol{\psi}, \boldsymbol{\phi}). \tag{10}$$

Theorem 2: The function d defined in (10) is a pseudometric in that it is nonnegative, symmetric and satisfies the triangle inequality.

Proof: It is obvious that d is nonnegative and symmetric; so it only remains to prove the triangle inequality. To prove this, we first establish a small technical point. Suppose $\boldsymbol{\eta} \in \mathcal{M}(\mathbb{A}\times\mathbb{C}), \boldsymbol{\zeta} \in \mathcal{M}(\mathbb{B}\times\mathbb{C})$ and that $\boldsymbol{\eta}_{\mathbb{C}} = \boldsymbol{\zeta}_{\mathbb{C}} = \boldsymbol{\xi}$. Then it is always possible to find a distribution $\boldsymbol{\nu} \in \mathcal{M}(\mathbb{A}\times\mathbb{B}\times\mathbb{C})$ such that $\boldsymbol{\nu}_{\mathbb{A}\times\mathbb{C}} = \boldsymbol{\eta}$ and $\boldsymbol{\nu}_{\mathbb{B}\times\mathbb{C}} = \boldsymbol{\zeta}$. In words, the claim is that, given two joint distributions, one of X and Z, and another of Y and Z, both of them having the same marginal distribution for Z, it is possible to find a joint distribution for all three variables X, Y, Z such that the marginal distributions of (X, Y) and of (Y, Z) match the two given joint distributions. To establish the claim, we construct $\boldsymbol{\nu}$ by making X and Y conditionally independent given Z, or equivalently, by making $X \to Z \to Y$ into a very short Markov chain. Accordingly, let

$$\nu_{ijk} = \frac{\eta_{ik}\zeta_{jk}}{\xi_k}.$$

It is routine to verify that $\boldsymbol{\nu}$ has the required properties, using the identities

$$\xi_k = \sum_{i \in \mathbb{A}} \eta_{ik} = \sum_{j \in \mathbb{B}} \zeta_{jk}.$$

Now we return to the proof that d satisfies the triangle inequality. Given three different probability distributions $\boldsymbol{\phi} \in \mathcal{M}(\mathbb{A})$, $\boldsymbol{\psi} \in \mathcal{M}(\mathbb{B})$, $\boldsymbol{\xi} \in \mathcal{M}(\mathbb{C})$, let us choose distributions $\boldsymbol{\theta} \in \mathcal{M}(\mathbb{A} \times \mathbb{B})$, $\boldsymbol{\eta} \in \mathcal{M}(\mathbb{A} \times \mathbb{C})$ and $\boldsymbol{\zeta} \in \mathcal{M}(\mathbb{B} \times \mathbb{C})$ such that

$$\boldsymbol{\theta}_{\mathbb{A}} = \boldsymbol{\phi}, \boldsymbol{\theta}_{\mathbb{B}} = \boldsymbol{\psi}, H(\boldsymbol{\theta}) = W(\boldsymbol{\phi}, \boldsymbol{\psi}), \tag{11}$$

$$\boldsymbol{\eta}_{\mathbb{A}} = \boldsymbol{\phi}, \boldsymbol{\eta}_{\mathbb{C}} = \boldsymbol{\xi}, H(\boldsymbol{\eta}) = W(\boldsymbol{\phi}, \boldsymbol{\xi}), \tag{12}$$

$$\boldsymbol{\zeta}_{\mathbb{B}} = \boldsymbol{\psi}, \boldsymbol{\zeta}_{\mathbb{C}} = \boldsymbol{\xi}, H(\boldsymbol{\zeta}) = W(\boldsymbol{\psi}, \boldsymbol{\xi}). \tag{13}$$

Now choose $\boldsymbol{\nu}$ to be any distribution on $\mathbb{A} \times \mathbb{B} \times \mathbb{C}$ such that

$$\boldsymbol{\nu}_{\mathbb{A} \times \mathbb{C}} = \boldsymbol{\eta}, \ \boldsymbol{\nu}_{\mathbb{B} \times \mathbb{C}} = \boldsymbol{\zeta}. \tag{14}$$

Let X, Y, Z be three random variables with the joint distribution $\boldsymbol{\nu}$. Then the triangle inequality for the quantity v shows that

$$v(X, Y) \leq v(X, Z) + v(Y, Z).$$

The manner in which $\boldsymbol{\eta}$ and $\boldsymbol{\zeta}$ were chosen shows that

$$v(X, Z) = d(\boldsymbol{\phi}, \boldsymbol{\xi}), \ v(Y, Z) = d(\boldsymbol{\psi}, \boldsymbol{\xi}).$$

However, an analogous statement about $v(X, Y)$ may not be true. So we note instead that $d(\boldsymbol{\phi}, \boldsymbol{\psi})$ is the *minimum* of $v(X, Y)$ whenever X and Y have distributions $\boldsymbol{\phi}, \boldsymbol{\xi}$ respectively. Hence

$$d(\boldsymbol{\phi}, \boldsymbol{\psi}) \leq v(X, Y) \leq v(X, Z) + v(Y, Z) = d(\boldsymbol{\phi}, \boldsymbol{\xi}) + d(\boldsymbol{\psi}, \boldsymbol{\xi}),$$

which is the desired conclusion. ■

III. Computing the Metric

A. Problem Formulation and Elementary Properties

Now that we have defined the metric, the next step is to compute it. Note that if we compute $V(\boldsymbol{\phi}, \boldsymbol{\psi})$, then $V(\boldsymbol{\psi}, \boldsymbol{\phi})$ is automatically determined by (7). Also, minimizing the conditional entropy maximizes the mutual information, so we refer to this approach as MMI. For reasons that will become later, we assume that $n \geq m$. Clearly there is no loss of generality in doing this. The next step is to reparametrize the problem, by changing the variable of optimization from the joint distribution $\boldsymbol{\theta} \in \mathbb{S}_{nm}$ to the matrix of conditional probabilities $P \in \mathbb{S}_{n \times m}$. Thus the boundary conditions $\boldsymbol{\theta}_{\mathbb{A}} = \boldsymbol{\phi}, \boldsymbol{\theta}_{\mathbb{B}} = \boldsymbol{\psi}$ get replaced by $\boldsymbol{\phi} P = \boldsymbol{\psi}$. Also, it is clear that, for a particular choice of P, the conditional entropy $H(Y|X)$ is given by

$$J_{\boldsymbol{\phi}}(P) = \sum_{i=1}^{n} \phi_i H(\mathbf{p}_i), \tag{15}$$

where $\mathbf{p}_i$ is the i-th row of P. Moreover, it follows from (4) that if P and Q are related by (2), then

$$J_{\boldsymbol{\psi}}(Q) = J_{\boldsymbol{\phi}}(P) + H(\boldsymbol{\phi}) - H(\boldsymbol{\psi}). \tag{16}$$

Finally, it is easy to see that, given $\boldsymbol{\phi} \in \mathbb{S}_n, \boldsymbol{\psi} \in \mathbb{S}_m$, the quantity V defined in (6) can also be defined equivalently as

$$V(\boldsymbol{\phi}, \boldsymbol{\psi}) = \min_{P \in \mathbb{S}_{n \times m}} J_{\boldsymbol{\phi}}(P). \tag{17}$$

MMI Problem: Given $\boldsymbol{\phi} \in \mathbb{S}_n, \boldsymbol{\psi} \in \mathbb{S}_m$, find a $P \in \mathbb{S}_{n \times m}$ that minimizes $J_{\boldsymbol{\phi}}(P)$ subject to the boundary condition $\boldsymbol{\phi} P = \boldsymbol{\psi}$.

It is clear that the feasible region for this problem

$$\mathcal{F} := \{P \in \mathbb{S}_{n \times m} : \boldsymbol{\phi} P = \boldsymbol{\psi}\} \tag{18}$$

is a polyhedral convex set. Recall that an element of a convex set is said to be an **extreme point** if it cannot be expressed as a nontrivial convex combination of two other points belonging to the set.

Theorem 3: Suppose all elements of $\boldsymbol{\phi}$ are strictly positive. Then the solution to the optimization problem in (16) occurs at an extreme point of $\mathcal{F}$. Thus if P achieves the minimum of $J_{\boldsymbol{\phi}}(\cdot)$, then at least one element of P is zero.

The proof is omitted as it is obvious.

B. A Principle of Optimality

We now state a 'principle of optimality' for this problem. Suppose $\boldsymbol{\phi} \in \mathbb{S}_n, \boldsymbol{\psi} \in \mathbb{S}_m$ are specified, and that $\phi_i > 0$ for all i. Suppose $\mathbb{A} = \{1, \ldots, n\}$, and let $\mathbb{A}'$ be a nonempty proper subset of $\mathbb{A}$. For notational convenience, suppose $\mathbb{A}' = \{1, \ldots, k\}$ where $k < n$. For $\boldsymbol{\phi} \in \mathbb{S}_n, P \in \mathbb{S}_{n \times m}$, define

$$\boldsymbol{\phi}' := [\phi_1 \ldots \phi_k], P' := \begin{bmatrix} \mathbf{p}_1 \\ \hline \vdots \\ \hline \mathbf{p}_k \end{bmatrix},$$

and note that $P' \in \mathbb{S}_{k \times m}$, though in general $\boldsymbol{\phi}'$ need not belong to $\mathbb{S}_k$. After this elaborate build-up we can now state the principle of optimality.

Theorem 4: With all notation as above, suppose $\phi_i > 0 \ \forall i$, and suppose that P^* minimizes $J_{\boldsymbol{\phi}}(P)$ subject to the constraint that $\boldsymbol{\phi} P = \boldsymbol{\psi}$. Define $c = \boldsymbol{\phi}' \mathbf{e}_k > 0$, and $\boldsymbol{\psi}' = \sum_{i=1}^{k} \phi_i \mathbf{p}_i^* = \boldsymbol{\phi}' (P^*)'$. Observe that $(1/c)\boldsymbol{\phi}' \in \mathbb{S}_k, (1/c)\boldsymbol{\psi}' \in \mathbb{S}_m$. Then $(P^*)'$ minimizes $J_{\boldsymbol{\phi}'}(P')$ over $\mathbb{S}_{k \times m}$ subject to the constraint that $(1/c)\boldsymbol{\phi}' P' = (1/c)\boldsymbol{\psi}'$.

Proof: Note that $(P^*)'$ is also a stochastic matrix in that $(P^*)' \mathbf{e}_m = \mathbf{e}_k$. Hence

$$\boldsymbol{\psi}' \mathbf{e}_m = \boldsymbol{\phi}' (P^*)' \mathbf{e}_m = \boldsymbol{\phi}' \mathbf{e}_k = c > 0,$$

because every component of $\boldsymbol{\phi}$ is positive. Hence $\boldsymbol{\psi}'$ is certainly not the zero vector, even though some components of $\boldsymbol{\psi}'$ could be zero. Thus $(1/c)\boldsymbol{\phi}' \in \mathbb{S}_k, (1/c)\boldsymbol{\psi}' \in \mathbb{S}_m$, and the minimization problem under study is similar to the larger problem. To prove the claim, suppose by way of contradiction that there exists another matrix $Q' \in \mathbb{S}_{k \times m}$ that satisfies $\boldsymbol{\phi}' Q' = \boldsymbol{\psi}'$ such that

$$J_{\boldsymbol{\phi}'}(Q') = \sum_{i=1}^{k} \phi_i H(\mathbf{q}_i) < J_{\boldsymbol{\phi}'}((P^*)') = \sum_{i=1}^{k} \phi_i H(\mathbf{p}_i^*).$$

Define in an analogous fashion

$$(P^*)'' = \left[\begin{array}{c} \mathbf{p}_{k+1} \\ \hline \vdots \\ \hline \mathbf{p}_n \end{array}\right], Q = \left[\begin{array}{c} Q' \\ (P^*)'' \end{array}\right],$$

and note that, since P^* is feasible for the original problem, we have that

$$\sum_{i=k+1}^{n} \phi_i \mathbf{p}_i^* = \boldsymbol{\phi} P^* - \boldsymbol{\phi}'(P^*)' = \boldsymbol{\psi} - \boldsymbol{\psi}'.$$

Now

$$\begin{aligned} J_{\boldsymbol{\phi}}(Q) &= \sum_{i=1}^{k} \phi_i H(\mathbf{q}_i) + \sum_{i=k+1}^{n} \phi_i H(\mathbf{p}_i^*) \\ &< \sum_{i=1}^{k} \phi_i H(\mathbf{p}_i^*) + \sum_{i=k+1}^{n} \phi_i H(\mathbf{p}_i^*) = J_{\boldsymbol{\phi}}(P^*), \end{aligned}$$

while

$$\boldsymbol{\phi} Q = \boldsymbol{\phi}' Q' + \sum_{i=k+1}^{n} \phi_i \mathbf{p}_i^* = \boldsymbol{\psi}' + \boldsymbol{\psi} - \boldsymbol{\psi}' = \boldsymbol{\psi}.$$

Hence Q is feasible for the original problem and has a lower objective function, which is a contradiction. Hence $(P^*)'$ is a minimizer of the reduced-size problem. ■

IV. Solution to the MMI Problem in the $n \times 2$ Case

A. The 2×2 Case

In this subsection we give an explicit closed-form expression for $V(\boldsymbol{\phi}, \boldsymbol{\psi})$ when $n = m = 2$ and $\boldsymbol{\phi}, \boldsymbol{\psi} \in \mathbb{S}_2$. Without loss of generality, assume that $\boldsymbol{\phi} \neq \boldsymbol{\psi}$ because $V(\boldsymbol{\phi}, \boldsymbol{\psi}) = 0$ if $\boldsymbol{\phi} = \boldsymbol{\psi}$. Also, again without loss of generality, rearrange the elements of $\boldsymbol{\phi}, \boldsymbol{\psi}$ such that both vectors are in strictly increasing order.[2] Then we can distinguish between two cases, namely (a) $0 < \psi_1 < \phi_1 < \phi_2 < \psi_2$, and (b) $0 < \phi_1 < \psi_1 < \psi_2 < \phi_2$.

Theorem 5: Suppose $n = m = 2$ and $\boldsymbol{\phi}, \boldsymbol{\psi} \in \mathbb{S}_2$. Suppose further that we have either case (a) or case (b) above. If $0 < \psi_1 < \phi_1 < \phi_2 < \psi_2$, then

$$\begin{aligned} V(\boldsymbol{\phi}, \boldsymbol{\psi}) &= -\psi_1 \log(\psi_1/\phi_1) \\ &- (\phi_1 - \psi_1) \log(1 - \psi_1/\phi_1). \end{aligned} \tag{19}$$

If $0 < \phi_1 < \psi_1 < \psi_2 < \phi_2$ then

$$\begin{aligned} V(\boldsymbol{\phi}, \boldsymbol{\psi}) &= -(\phi_2 - \psi_2) \log(1 - \psi_2/\phi_2) \\ &- \psi_2 \log(\psi_2/\phi_2). \end{aligned} \tag{20}$$

Proof: From Theorem 3, we know that any optimal choice of $P \in \mathbb{S}_{2\times 2}$ must be an extreme point of the feasible region. Thus at least one component of P must be zero. The constraints that P is stochastic and that $\boldsymbol{\phi} P = \boldsymbol{\psi}$ lead to the following four possible extreme points of the feasible region.

$$P_{11} = \left[\begin{array}{cc} 0 & 1 \\ \frac{\psi_1}{\phi_2} & 1 - \frac{\psi_1}{\phi_2} \end{array}\right], P_{12} = \left[\begin{array}{cc} 1 & 0 \\ 1 - \frac{\psi_2}{\phi_2} & \frac{\psi_2}{\phi_2} \end{array}\right],$$

$$P_{21} = \left[\begin{array}{cc} \frac{\psi_1}{\phi_1} & 1 - \frac{\psi_1}{\phi_1} \\ 0 & 1 \end{array}\right], P_{22} = \left[\begin{array}{cc} 1 - \frac{\psi_2}{\phi_1} & \frac{\psi_2}{\phi_1} \\ 1 & 0 \end{array}\right].$$

Moreover, since $\psi_2 > \phi_2$ and $\psi_2 > \phi_1$, it follows that P_{12} and P_{22} are infeasible, and the only possibilities are P_{11} and P_{21}. So all we need to do is to compute $J_{\boldsymbol{\phi}}(P_{11}), J_{\boldsymbol{\phi}}(P_{21})$, and pick the one that is smaller. This is an exercise in calculus and is omitted. The other case follows by symmetry. ■

B. The $n \times 2$ Case

We begin with a notion that is encountered again several times in the paper.

Definition 4: Given $\boldsymbol{\phi} \in \mathbb{S}_n, \boldsymbol{\psi} \in \mathbb{S}_m$ with $n > m$, $\boldsymbol{\psi}$ is said to be an **aggregation** of $\boldsymbol{\phi}$ if there exists a partition of $\mathbb{A}$ into m sets $I_1, \ldots, I_m$ such that $\sum_{i \in I_j} \phi_i = \psi_j$ for $j = 1, \ldots, m$.

Next we introduce the bin-packing problem with overstuffing and variable bin capacities as follows: Given $\boldsymbol{\phi} \in \mathbb{S}_n, \boldsymbol{\psi} \in \mathbb{S}_m$, find a partition of $\mathbb{A}$ into m sets $I_1, \ldots, I_m$ such that the total mismatch

$$MI = \sum_{j \in \mathbb{B}} \left| \psi_j - \sum_{i \in I_j} \phi_i \right|$$

is as small as possible. Unfortunately, this problem is also NP-hard [7]. Even determining whether a given $\boldsymbol{\psi}$ is an aggregation of a given $\boldsymbol{\phi}$ or not is also NP-hard. The bin packing with overstuffing is discussed in [7], [3], [4] among other papers.

With this background, we now present a partial solution to the problem of computing $V(\boldsymbol{\phi}, \boldsymbol{\psi})$ when $m = 2$ in terms of the bin-packing problem with overstuffing with two bins. If $\boldsymbol{\psi}$ is an aggregation of $\boldsymbol{\phi}$, then obviously $V(\boldsymbol{\phi}, \boldsymbol{\psi}) = 0$. Otherwise, let ψ_1, ψ_2 denote the capacity of the two bins, and let $\phi_1, \ldots, \phi_n$ denote the list to be packed. Without loss of generality, assume that the ϕ_i are in decreasing order of magnitude. Let $\mathcal{N}_1, \mathcal{N}_2$ denote an optimal partition of $\mathcal{N} = \{1, \ldots, n\}$ and let c denote the minimum unutilized capacity. Again, without loss of generality, assume that bin 1 is underutilized and that bin 2 is overstuffed. This means that

$$\psi_2 - \sum_{i \in \mathcal{N}_2} \phi_i = -\psi_1 + \sum_{i \in \mathcal{N}_1} \phi_i = c. \tag{21}$$

Theorem 6: Suppose $\boldsymbol{\psi}$ is not an aggregation of $\boldsymbol{\phi}$, and solve the bin-packing problem as above. If $n \in \mathcal{N}_2$, then an optimal choice of P that minimizes $J_{\boldsymbol{\phi}}(P)$ subject to $\boldsymbol{\phi} P = \boldsymbol{\psi}$ is given by

$$\begin{aligned} \mathbf{p}_i = [1 \;\; 0] \; \forall i \in \mathcal{N}_1, \mathbf{p}_i = [0 \;\; 1] \; \forall i \in \mathcal{N}_2 \setminus \{n\}, \\ \mathbf{p}_n = [\; c/\phi_n \quad (\phi_n - c)/\phi_n \;]. \end{aligned} \tag{22}$$

[2] To avoid unnecessary pedantry, we assume that lots of strict inequalities hold. The modifications needed to handle the case where some of the inequalities are not strict are easy and are left to the reader.

Moreover

$$V(\boldsymbol{\phi}, \boldsymbol{\psi}) = \phi_n H(\mathbf{p}_n) = f_c(\phi_n),$$

where the function f is defined as

$$f_u(\phi) := \phi[h(u/\phi) + h(1 - (u/\phi))]. \tag{23}$$

Proof: From the principle of optimality, we know that if a matrix P is optimal for the $n \times 2$ problem, then every 2×2 submatrix is optimal for its respective problem, and thus has at most one strictly positive row. Taken together this shows that any optimal choice of P has at most one strictly positive row, while the rest are either $[1 \;\; 0]$ or $[0 \;\; 1]$. Accordingly, define P as above, and let R be another matrix that has exactly one strictly positive row such that $\boldsymbol{\phi}R = \boldsymbol{\psi}$. All we need to do is to show that $J_{\boldsymbol{\phi}}(R) \geq J_{\boldsymbol{\phi}}(P)$. For this purpose, suppose the k-th row of R is strictly positive, and define

$$I_1 = \{i : \mathbf{r}_i = [1 \;\; 0]\}, I_2 = \{i : \mathbf{r}_i = [0 \;\; 1]\},$$

while $\mathbf{r}_k$ is strictly positive. Then $\boldsymbol{\phi}R = \boldsymbol{\psi}$ implies that

$$u_1 := \psi_1 - \sum_{i \in I_1} \phi_i > 0, u_2 := \psi_2 - \sum_{i \in I_2} \phi_i > 0, u_1 + u_2 = \phi_k,$$

$$\mathbf{r}_k = [\;\; u_1/\phi_k \quad u_2/\phi_k \;\;], J_{\boldsymbol{\phi}}(R) = \phi_k H(\mathbf{r}_k) = f_{u_1}(\phi_k),$$

where the function f is defined in (23), and we use the fact that $u_2 = \phi_k - u_1$. The fact that c is the optimal unutilized capacity implies that $c \leq \min\{u_1, u_2\}$, so that $c \leq \min\{u_1, u_2\} \leq \max\{u_1, u_2\} \leq \phi_k - c$. In turn this implies that

$$\begin{aligned} H(\mathbf{r}_k) &= H([\;\; u_1/\phi_k \quad u_2/\phi_k \;\;]) \\ &\geq H([\;\; c/\phi_k \quad (\phi_k - c)/\phi_k \;\;]). \end{aligned}$$

So we now conclude that

$$\begin{aligned} J_{\boldsymbol{\phi}}(R) &= \phi_k H(\mathbf{r}_k) \geq \phi_k H([\;\; c/\phi_k \quad (\phi_k - c)/\phi_k \;\;]) \\ &= f_c(\phi_k) \geq f_c(\phi_n) = J_{\boldsymbol{\phi}}(P) \end{aligned}$$

because $\phi_n \leq \phi_k$ and $f_c(\cdot)$ is a strictly increasing function. ■

V. Solution to the MMI Problem in the $n \times m$ Case

A. Greedy Algorithm for the MMI Problem

In general, determining whether $\boldsymbol{\psi}$ is an aggregation of $\boldsymbol{\phi}$, or finding the optimal bin allocations allowing overstuffing, are both NP-hard problems [3], [4]. It follows that computing $V(\boldsymbol{\phi}, \boldsymbol{\psi})$, or equivalently, computing the maximum mutual information, is also NP-hard when $m = 2$. It is therefore plausible that the problem of computing $V(\boldsymbol{\phi}, \boldsymbol{\psi})$ continues to be NP-hard if $3 \leq m \leq n$. But we do not explore this issue further. Instead, we borrow a standard greedy algorithm for bin-packing with overstuffing from the computer science literature [21], known as 'best fit,' and adapt it to the current situation. We begin by arranging the elements of $\boldsymbol{\psi}$ in descending order. In general it is *not* necessary to sort the elements of $\boldsymbol{\phi}$.

Given $\boldsymbol{\phi} \in \mathbb{S}_n, \boldsymbol{\psi} \in \mathbb{S}_m$ with $m < n$, proceed as follows:

1) Set $s = 1$, where s is the round counter. Define $n_s = n, m_s = m$, $\boldsymbol{\phi}_s = \boldsymbol{\phi}, \boldsymbol{\psi}_s = \boldsymbol{\psi}$.
2) Place each element of $\boldsymbol{\phi}$ in the bin with the largest unused capacity. If a particular component $(\boldsymbol{\phi}_s)_i$ does not fit into any bin, assign the index i to an overflow index set K_s.
3) When all elements of $\boldsymbol{\phi}_s$ have been processed, let $I_1^{(s)}, \ldots, I_{m_s}^{(s)}$ be the indices from $\{1, \ldots, n_s\}$ that have been assigned to the various bins, and let K_s denote the set of indices that cannot be assigned to any bin. If $|K_s| > 1$ go to Step 4; otherwise go to Step 5.
4) Define $\alpha_1^{(s)}, \ldots, \alpha_{m_s}^{(s)}$ to be the unutilized capacities of the m_s bins, and define $\boldsymbol{\alpha}^{(s)} = [\alpha_1^{(s)} \ldots \alpha_{m_s}^{(s)}]$. Then the total unutilized capacity $c_s := \boldsymbol{\alpha}^{(s)} \mathbf{e}_{m_s}$ satisfies

$$c_s = \sum_{j=1}^{m_s} \alpha_j^{(s)} = \sum_{i \in K_s} (\boldsymbol{\phi}_s)_i. \tag{24}$$

Since each $(\boldsymbol{\phi}_s)_i, i \in K_s$ does not fit into any bin, it is clear that $(\boldsymbol{\phi}_s)_i > \alpha_j^{(s)}, \; \forall i, j$. In turn this implies that $|K_s| < m_s$. Next, set $n_{s+1} = m_s, m_{s+1} = |K_s|$, and define

$$\boldsymbol{\phi}_{s+1} = \frac{1}{c_s} \boldsymbol{\alpha}^{(s)} \in \mathbb{S}_{n_{s+1}}, \boldsymbol{\psi}_{s+1} = \frac{1}{c_s} [(\boldsymbol{\phi}_s)_i] \in \mathbb{S}_{m_{s+1}}.$$

Increment the counter and go to Step 2.
5) When this step is reached, $|K_s|$ is either zero or one. If $|K_s| = 0$, then it means that $\boldsymbol{\psi}_s$ is a perfect aggregation of $\boldsymbol{\phi}_s$. So define $V_s = 0$ and proceed as below. If $|K_s| = 1$, then only one element of $\boldsymbol{\phi}_s$, call it $(\boldsymbol{\phi}_s)_k$, cannot be packed into any bin, and this component must equal c_s. So let

$$\mathbf{v}_s = \frac{1}{c_s} \boldsymbol{\alpha}^{(s)} \in \mathbb{S}_{m_s}, V_s = c_s H(\mathbf{v}_s),$$
$$U_s = V_s + H(\boldsymbol{\phi}_s) - H(\boldsymbol{\psi}_s).$$

Define $P_s \in \mathbb{S}_{n_s \times m_s}$ by

$$\mathbf{p}_i = \mathbf{b}_j \text{ if } i \in I_j^{(s)}, \mathbf{p}_k = \mathbf{v}_s,$$

where $\mathbf{b}_j$ is the j-th unit vector with m_s components. Then V_s is the minimum value of $J_{\boldsymbol{\phi}_s}(\cdot)$, and P_s achieves that minimum. Next, define $Q_s \in \mathbb{S}_{m_s \times n_s}$ by

$$Q_s = [\mathrm{diag}(\boldsymbol{\psi}_s)]^{-1} P_s^T \mathrm{Diag}(\boldsymbol{\phi}_s).$$

Then it follows from (16) that Q_s minimizes $J_{\boldsymbol{\psi}_s}(\cdot)$, and that U_s is the value of that minimum.
6) In this step, we invert all of the above steps by transposing Q_{s+1}, applying the transformation in (2), and embedding the resulting matrix into P_s. We also correct the cost function using (16). Decrement the counter s and recall that $m_s = n_{s+1}$. Recall the unutilized capacity c_s defined in (24) which has been found during the forward iteration, and define

$$V_s = c_s U_{s+1}, U_s = V_s + H(\boldsymbol{\phi}_s) - H(\boldsymbol{\psi}_s).$$

Define $P_s \in \mathbb{S}_{n_s \times m_s}$ by

$$\mathbf{p}_i = \mathbf{b}_j \text{ if } i \in I_j^{(s)}, \mathbf{p}_i = i\text{-th row of } Q_{s+1}.$$

If $s = 1$, halt; otherwise repeat the step.

B. Computational Complexity

The computational complexity of algorithm is easy to bound. The first step is to sort the elements of $\boldsymbol{\psi}$, which has complexity $O(m^2)$ if we insist on an exact answer or $O(m \log m)$ if we use a randomized algorithm like quick sort. We use the latter bound here. In each step of the best fit algorithm, the bin in which the current element of $\boldsymbol{\phi}$ has been placed has maximum capacity *before* placing, but necessarily *after* placing. So it needs to moved into the right place. Since the rest of the bins are still in descending order of capacity, this can be achieved in $O(\log m)$ steps using a bisection search. And this has to be done n times. So once $\boldsymbol{\psi}$ is sorted, one run of the best fit algorithm has complexity $O(n \log m)$, which dominates the complexity $O(m \log m)$ of sorting $\boldsymbol{\psi}$, since $m \leq n$. Since the size of the problem decreases at each round, at worst we may have to run the best fit algorithm $m-1$ times. Moreover, after the first round, the size of the problem is not any larger than $m \times (m-1)$. So the overall complexity of the greedy algorithm is no worse than $O(n \log m) + mO(m \log m) = O((n+m^2) \log m)$. The fact that the complexity is only linear in n is heartening.

In [18], the application of the greedy algorithm is illustrated on a large 40×10 example that needs to go through three rounds.

VI. Conclusions

In this paper we have studied the problem of defining a metric distance between two probability distributions over distinct finite sets of possibly different cardinalities. Along the way, we have formulated the problem of constructing a joint distribution on the product of the two sets, which has the two given distributions as its marginals, in such a way that the joint distribution has minimum entropy. While the problem of maximizing mutual information is occasionally discussed in the literature, this specific problem does not appear to have been studied earlier. This problem turns out to be NP-hard, so we reformulated the problem as a bin-packing problem with overstuffing, and adapt the best fit algorithm for bin-packing, leading to an upper bound on the distance between the two given distributions. The complexity of this algorithm is $O((n+m^2) \log m)$, where n is the larger of the two cardinalities and m is the smaller.

Applications of the metric to the problem of order reduction are presented in a companion paper [19]. A full length version that combines both papers and is under review for journal publication can be found at [18].

References

[1] Rudi Cilibrasi and Paul M. B. Vitányi, "Clustering by comparison", *IEEE Trans. Info. Thy.*, 51(4), 1523-1545, April 2005.

[2] Edward G. Coffman, Jr. and János Csirik, "Performance guarantees for one-dimensional bin packing", Chapter 32 in [11].

[3] Edward G. Coffman, Jr., János Csirik and Joseph Y.-T. Leung, "Variants of classical one-dimensional bin packing", Chapter 33 in [11].

[4] Edward G. Coffman, Jr., János Csirik and Joseph Y.-T. Leung, "Variable-sized bin packing and bin covering", Chapter 34 in [11].

[5] E. G. Coffman, Jr. and George S. Lueker, "Approximation algorithms for extensible bin packing," *Proc. SODA*, 586-588, January 2001.

[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, (Second Edition), Wiley, New York, 2006.

[7] Paolo Dell'Olmo, Hans Kellerer, Maria Grazia Speranza and Zsolt Tuza, "A 13/12 approximation algorithm for bin packing with extendable bins," *Information Processing Letters*, 65, 229- 233, 1998.

[8] Kun Deng, Prashant G. Mehta and Sean P. Meyn, 'Optimal Kullback-Leibler aggregation via the spectral theory of Markov chains", *Proc. Amer. Control Conf.*, St. Louis, MO, 731-736, 2009.

[9] Kun Deng, Prashant G. Mehta and Sean P. Meyn, "A simulation-based method for aggregating Markov chains", *Proc. IEEE Conf. on Decision and Control*, Shanghai, China, 4710-4716, 2009.

[10] Kun Deng, Prashant G. Mehta and Sean P. Meyn, "Optimal Kullback-Leibler aggregation via the spectral theory of Markov chains", to appear in *IEEE Trans. Auto. Control*.

[11] Teofilo González (Editor), *Handbook of Approximation Algorithms and Metaheuristics*, Chapman and Hall CRC, London, 2007.

[12] Ming Li, Xin Chen, Xin Li, Bin Ma and Paul M. B. Vitányi, "The similarity metric", *IEEE Trans. Info. Thy.*, 50(12), 3250-3264, Dec. 2004.

[13] Marina Meila, "Comparing clusterings by the variation of information", in *Learning Theory and Kernel Machines: 16th Annual Conference on Learning and 7th Kernel Workshop*, Bernard Schölkopf, Manfred Warmuth and Manfred K. Warmuth (Editors), pp. 173-187, 2003.

[14] Marina Meila, "Comparing clusterings – an information-based distance", *J. Multivariate Anal.*, 98(5), 873-895, 2007.

[15] Donald S. Ornstein, "An application of ergodic theory to probability theory", *The Annals of Probability*, 1(1), 43-65, 1973.

[16] J. C. Spall and S. D. Hull, "Least-informative Bayesian prior distributions for finite samples based on information theory," *IEEE Trans. Auto. Control*, 35(5), 580-583, May 1990.

[17] M. Vidyasagar, "Kullback-Leibler Divergence Rate Between Probability Distributions on Sets of Different Cardinalities", *Proc. IEEE Conf. on Decision and Control*, Atlanta, GA, 947-953, 2010.

[18] M. Vidyasagar, "Metrics between probability distributions on finite sets of different cardinalities by maximizing mutual information (MMI)," *arxiv:1104.4521v2.pdf*.

[19] M. Vidyasagar, "Optimal order reduction of probability distributions by maximizing mutual information," to be presented at CDC 2011.

[20] Deshi Yu and Guochuan Zhang, "On-line extensible bin packing with unequal bin sizes", *Lecture Notes in Computer Science*, Vol. 2909, 235-247, 2004.

[21] Minyi Yue, "A simple proof of the inequality $FFD(L) \leq (11/9)OPT(L) + 1$, $\forall L$ for the FFD bin-packing algorithm", *Acta Mathematicae Applicatae Sinica*, 7(4), 321-331, Oct. 1991.