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

Optimal Order Reduction of Probability Distributions by Maximizing Mutual Information

M. Vidyasagar

***Abstract*— In a companion paper [16], we defined a metric distance between two probability distributions ϕ, ψ defined on sets of different cardinality, called the variation of information metric d. In this paper we study of problem of finding an *optimal reduced-order approximation* in the variation of information metric. Let ϕ denote the probability distribution of high dimension that is to be approximated. It is shown first that any optimal approximation of ϕ must be an aggregation of ϕ. Then it is shown that any optimal *aggregation* of ϕ is one that has maximum entropy. Using these two results, we then formulate the problem of optimal order reduction as a nonstandard bin-packing problem with overstuffing. Unfortunately this problem is NP-hard. So a greedy algorithm is presented to solve this problem, and an upper bound on its performance is presented. The application of the greedy algorithm is illustrated via a large example.**

I. Introduction

In a companion paper [16], we defined a metric between two probability distributions, say ϕ, ψ, defined on sets of different cardinalities, called the variation of information metric d. If there is a way of measuring the 'distance' between a high-dimensional probability distribution and a lower-dimensional probability distribution, it is natural to study the problem of *optimal order reduction*. Specifically, suppose ϕ is an n-dimensional probability distribution, and $m < n$ is a specified integer. Then one can ask: What is the (or an) 'optimal' m-dimensional approximation ψ to ϕ in the sense of minimizing the variation of information metric $d(\phi, \psi)$? That is the question studied in this paper. The question is answered via several steps. First, it is shown that *any optimal approximation of ϕ must in fact be an aggregation of ϕ*. Then it is shown that an aggregation of ϕ is optimal if and only if it has maximum entropy amongst all aggregations. Thus the optimal order reduction problem is equivalent to maximizing entropy via aggregation. This problem is then reformulated as a nonstandard bin-packing problem with overstuffing. Unfortunately this problem is NP-hard. So we propose a greedy algorithm to construct a suboptimal solution, and also prove an upper bound on its performance. The approach is illustrated through a large example.

Cecil & Ida Green Endowed Chair, Erik Jonsson School of Engineering & Computer Science, University of Texas at Dallas, 800 W. Campbell Road, EC38, Richardson, TX 75080, USA; email: M.Vidyasagar@utdallas.edu. This research was supported by National Science Foundation Award #1001643.

II. Preliminaries

In this section we reprise some relevant results from [16].

A. Notation

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 $P \in [0,1]^{n \times m}$ is said to be **stochastic** if $P\mathbf{e}_m = \mathbf{e}_n$, that is, all row sums of P are equal to one. Note that P need not be a square matrix; but this definition is consistent with the more familiar usage for square matrices. The set of $n \times m$ stochastic matrices is denoted by $\mathbb{S}_{n \times m}$. If we take the degenerate case of $m = 1$, then the symbol $\mathbb{S}_n$ 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 $\mathbb{A}$ on a set of cardinality n.

Throughout, 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 $h'(r) = -(1 + \log r)$. Throughout, we use the symbol H to denote the Shannon entropy of a probability distribution. Thus if $\phi \in \mathbb{S}_n$, then

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

B. Reprise of Relevant Results from [16]

Suppose $\phi \in \mathbb{S}_n, \psi \in \mathbb{S}_m$. Then we can think of ϕ, ψ as probability distributions on some sets $\mathbb{A}, \mathbb{B}$ respectively where $|\mathbb{A}| = n, |\mathbb{B}| = m$. We can also think of ϕ, ψ as probability distributions of some random variables X, Y assuming values in $\mathbb{A}, \mathbb{B}$ respectively. Let θ denote the joint distribution of (X, Y), so that $\theta \in \mathcal{M}(\mathbb{A} \times \mathbb{B})$, and let $\theta_{\mathbb{A}}, \theta_{\mathbb{B}}$ denote its marginals on $\mathbb{A}, \mathbb{B}$ respectively. With this convention, we define the following quantities:

$$W(\phi, \psi) := \min_{\theta \in \mathcal{M}(\mathbb{A} \times \mathbb{B})} \{H(\theta) : \theta_{\mathbb{A}} = \phi, \theta_{\mathbb{B}} = \psi\}, \quad (1)$$

$$V(\phi, \psi) := W(\phi, \psi) - H(\phi). \quad (2)$$

Compare with [16], Equations (5) and (6). Then it is possible to define a metric between ϕ, ψ.

Definition 1: Suppose $\phi \in \mathbb{S}_n, \psi \in \mathbb{S}_m$ and let $W(\phi, \psi)$ be as in (6). Then

$$d(\phi, \psi) := V(\phi, \psi) + V(\psi, \phi) \quad (3)$$

is called the **variation of information metric** between ϕ and ψ.

The salient properties of the function d are brought out next.

Theorem 1: The functions d defined in (3) is a pseudo-metric.

The computation of the quantity d relies on the computation of $V(\boldsymbol{\phi}, \boldsymbol{\psi})$. This can be achieved by formulating an associated optimization problem. Given $\boldsymbol{\phi} \in \mathbb{S}_n$, define the function $J_{\boldsymbol{\phi}} : \mathbb{S}_{n\times m} \to \mathbb{R}_+$ by

$$J_{\boldsymbol{\phi}}(P) = \sum_{i=1}^{n} \phi_i H(\mathbf{p}_i), \tag{4}$$

where $\mathbf{p}_i$ is the i-th row of P. Then

$$V(\boldsymbol{\phi}, \boldsymbol{\psi}) = \min_{P \in \mathbb{S}_{n\times m}} J_{\boldsymbol{\phi}}(P) \text{ s.t. } \boldsymbol{\phi} P = \boldsymbol{\psi}. \tag{5}$$

The solution of minimizing $J_{\boldsymbol{\phi}}(P)$ with respect to P is the main topic of [16].

III. All Optimal Reduced-Order Approximations are Aggregations

Once we have a way of quantifying the distance between probability distributions having different dimensions, it is natural to examine the problem of approximating a distribution $\boldsymbol{\phi} \in \mathbb{S}_n$ by another $\boldsymbol{\psi} \in \mathbb{S}_m$ where $m \ll n$, such that the distance between them is as small as possible. This may be referred to as the ‘order reduction’ problem. This is precisely the problem studied in the present paper, namely: Given a distribution $\boldsymbol{\phi} \in \mathbb{S}_n$, and an integer $m < n$ (perhaps $m \ll n$), find a $\boldsymbol{\psi} \in \mathbb{S}_m$ such that $d(\boldsymbol{\phi}, \boldsymbol{\psi})$ is as small as possible.

Given $\boldsymbol{\phi} \in \mathbb{S}_n$, let us refer to $\boldsymbol{\phi}^{(a)}$ as an **aggregation** of $\boldsymbol{\phi}$ if it can be obtained by aggregating the components of $\boldsymbol{\phi}$. In other words, $\boldsymbol{\phi}^{(a)}$ is an aggregation of $\boldsymbol{\phi}$ if there exists a partition of $\{1, \ldots, n\}$ into m pairwise disjoint sets $I_1, \ldots, I_m$ such that

$$\phi_j^{(a)} = \sum_{i \in I_j} \phi_i, j = 1, \ldots, m.$$

An equivalent way of saying the same thing is the following: Suppose $m < n$. Then $\boldsymbol{\psi} \in \mathbb{S}_m$ is an aggregation of $\boldsymbol{\phi} \in \mathbb{S}_n$ if and only if there exists a matrix $P \in \mathbb{S}_{n\times m}$ such that $\boldsymbol{\psi} = \boldsymbol{\phi} P$, and in addition, $p_{ij} = 0$ or 1 for all i, j. Note that the two conditions $P \in \mathbb{S}_{n\times m}$ and p_{ij} equals 0 or 1 ensure that every row of P consists of a degenerate probability distribution, with a solitary component equal to 1 and the rest equal to zero.

In this section, it is shown that, given a distribution $\boldsymbol{\phi} \in \mathbb{S}_n$, an optimal approximation of $\boldsymbol{\phi}$ in $\mathbb{S}_m, m < n$, in terms of the variation of information metric, *must be an aggregation of* $\boldsymbol{\phi}$. Unfortunately the proof of this intuitive result is very long.

Theorem 2: Suppose $\boldsymbol{\phi} \in \mathbb{S}_n, \boldsymbol{\psi} \in \mathbb{S}_m, m < n$, and that $\boldsymbol{\psi}$ is not an aggregation of $\boldsymbol{\phi}$. Then there exists a $\boldsymbol{\psi}' \in \mathbb{S}_m$ such that $d(\boldsymbol{\phi}, \boldsymbol{\psi}') < d(\boldsymbol{\phi}, \boldsymbol{\psi})$.

The proof of the theorem makes use of a couple of preliminary lemmas.

Lemma 1: Suppose $\boldsymbol{\mu} \in \mathbb{R}_+^m, \boldsymbol{\mu} \neq \mathbf{0}$. Then

$$\sum_{j=1}^{m} h(\mu_j) = cH((1/c)\boldsymbol{\mu})) + h(c), \tag{6}$$

where $c = \boldsymbol{\mu}\mathbf{e}_m$ is a normalizing constant.

Proof: We have that

$$\begin{aligned}
\sum_{j=1}^{m} h(\mu_j) &= \sum_{j=1}^{m} \mu_j \log \frac{1}{\mu_j} \\
&= \sum_{j=1}^{m} \mu_j \log \frac{c}{\mu_j} - \left(\sum_{j=1}^{m} \mu_j \right) \log c \\
&= c \sum_{j=1}^{m} \frac{\mu_j}{c} \log \frac{c}{\mu_j} - c \log c \\
&= cH((1/c)\boldsymbol{\mu})) + h(c).
\end{aligned}$$

This completes the proof.

Lemma 2: Suppose $c_1, c_2, b > 0$ with $c_1 + c_2 + b = 1$. For each $\lambda \in [0, 1]$, define $\boldsymbol{\psi}(\lambda) \in \mathbb{S}_2$ by

$$\boldsymbol{\psi}(\lambda) = [\ c_1 + \lambda b \quad c_2 + (1 - \lambda) b\ ],$$

and $G : [0, 1] \to \mathbb{R}$ by

$$G(\lambda) = bH([\lambda\ \ 1 - \lambda]) - H(\boldsymbol{\psi}(\lambda)).$$

Then

$$G(\lambda) > \min\{G(0), G(1)\} = \min\{-H(\boldsymbol{\psi}(0)), -H(\boldsymbol{\psi}(1))\}. \tag{7}$$

Proof: This follows from elementary calculus. Recall that for the function $h(r) = r \log(1/r)$, we have that $h'(r) = -(1 + \log r)$ for all $r > 0$. Now expand $G(\lambda)$ as

$$G(\lambda) = b(h(\lambda) + h(1-\lambda)) - h(c_1 + \lambda b) - h(c_2 + (1-\lambda) b).$$

Then it follows that

$$\begin{aligned}
G'(\lambda) &= -b(1 + \log \lambda) + b(1 + \log(1 - \lambda)) \\
&+ b(1 + \log(c_1 + \lambda b) - b(1 + \log(c_2 + (1 - \lambda) b) \\
&= b \log \left[\frac{(c_1 + \lambda b) \cdot (1 - \lambda)}{(c_2 + (1 - \lambda) b) \cdot \lambda} \right] \\
&= b \log \left[\frac{c_1 - c_1 \lambda + b\lambda - b\lambda^2}{c_2 \lambda + b\lambda - b\lambda^2} \right].
\end{aligned}$$

From the above, it is clear that $G'(\lambda) = 0$ when

$$c_1 - c_1\lambda + b\lambda - b\lambda^2 = c_2\lambda + b\lambda - b\lambda^2,$$

or

$$c_1 - c_1\lambda = c_2\lambda, \lambda = \frac{c_1}{c_1 + c_2} =: \lambda^*.$$

Now if $\lambda > \lambda^*$, then

$$c_1 < (c_1 + c_2)\lambda, \text{ or } c_1 - c_1\lambda < c_2\lambda.$$

So the numerator in the fraction above is smaller than the denominator, and as a result $G'(\lambda) < 0$ if $\lambda > \lambda^*$. Similar reasoning shows that $G'(\lambda) > 0$ if $\lambda < \lambda^*$. So $G(\lambda)$ attains its maximum when $\lambda = \lambda^*$, and decreases on either side of

λ^*. So in particular, if $\lambda < \lambda^*$, then $G(\lambda) > G(0)$, whereas if $\lambda > \lambda^*$, then $G(\lambda) > G(1)$. In either case, (7) is satisfied.

Proof (of Theorem 2): Now we give a proof of Theorem 2. Suppose $\boldsymbol{\phi} \in \mathbb{S}_n, \boldsymbol{\psi} \in \mathbb{S}_m, m < n$, and $\boldsymbol{\psi}$ is *not* an aggregation of $\boldsymbol{\phi}$. Choose $P \in \mathbb{S}_{n\times m}$ such that $\boldsymbol{\phi}P = \boldsymbol{\psi}$ and $J_{\boldsymbol{\phi}}(P) = V(\boldsymbol{\phi}, \boldsymbol{\psi})$. Since $\boldsymbol{\psi}$ is not an aggregation of $\boldsymbol{\phi}$, at least one row of P contains at least two nonzero (i.e., positive) elements. Let k be such a row, and without loss of generality permute the components of $\boldsymbol{\psi}$ in such a way that $p_{k1} > 0, p_{k2} > 0$. To show that $\boldsymbol{\psi}$ cannot be an optimal approximation of $\boldsymbol{\phi}$ in the d metric, we will construct another distribution $\boldsymbol{\psi}' \in \mathbb{S}_m$ that matches $\boldsymbol{\psi}$ from component 3 onwards. We will do this by perturbing *only* the two elements p_{k1}, p_{k2} in such a way that $p'_{k1} + p'_{k2} = p_{k1} + p_{k2}$, and defining $\boldsymbol{\psi}' = \boldsymbol{\phi}P'$. This means that many of the quantities are common to $\boldsymbol{\psi}$ and $\boldsymbol{\psi}'$, so in the various equations below, we will just write ‘constant’ or ‘const.’ to avoid notational clutter.

From the manner in which P was chosen, it follows that

$$\begin{aligned} V(\boldsymbol{\phi}, \boldsymbol{\psi}) &= \sum_{i=1}^{n} \phi_i H(\mathbf{p}_i) \\ &= \phi_k H(\mathbf{p}_k) + \text{const} \\ &= \phi_k[h(p_{k1}) + h(p_{k2})] + \text{const}, \end{aligned}$$

$$\begin{aligned} V(\boldsymbol{\psi}, \boldsymbol{\phi}) &= V(\boldsymbol{\phi}, \boldsymbol{\psi}) + H(\boldsymbol{\phi}) - H(\boldsymbol{\psi}) \\ &= \phi_k[h(p_{k1}) + h(p_{k2})] - [h(\psi_1) + h(\psi_2)] \\ &+ \text{const}. \end{aligned}$$

Note that the ‘constant’ in the two equations need not be the same. Our use of the phrase ‘constant’ means only that all the ignored summations remain unchanged when we replace $\boldsymbol{\psi}$ by $\boldsymbol{\psi}'$. Proceeding further, let us write

$$\psi_1 = \sum_{i=1}^{n} \phi_i p_{i1} = \sum_{i \neq k} \phi_i p_{i1} + \phi_k p_{k1} =: d_1 + \phi_k p_{k1}.$$

Similarly,

$$\psi_2 = \sum_{i=1}^{n} \phi_i p_{i2} = \sum_{i \neq k} \phi_i p_{i2} + \phi_k p_{k2} =: d_2 + \phi_k p_{k2}.$$

With these definitions, we can write

$$\begin{aligned} V(\boldsymbol{\psi}, \boldsymbol{\phi}) &= \phi_k[h(p_{k1}) + h(p_{k2})] \\ &- [h(d_1 + \phi_k p_{k1}) + h(d_2 + \phi_k p_{k2})] \\ &+ \text{const}. \end{aligned} \tag{8}$$

This looks similar to the function $G(\lambda)$ in Lemma 2, *except that* neither $p_{k1} + p_{k2}$ nor $d_1 + d_2 + \phi_k$ necessarily add up to one. So we proceed as in the proof of Lemma 2 and apply the correction terms from Lemma 1 wherever necessary. Let us define

$$\lambda = \frac{p_{k1}}{p_{k1} + p_{k2}}, 1 - \lambda = \frac{p_{k2}}{p_{k1} + p_{k2}},$$
$$\beta = p_{k1} + p_{k2}, \alpha = d_1 + d_2 + \beta\phi_k,$$

and note that

$$\psi_1 + \psi_2 = d_1 + d_2 + \beta\phi_k = \alpha.$$

With these definitions, and making repeated use of Lemma 1, we get

$$\phi_k[h(p_{k1}) + h(p_{k2})] = \beta\phi_k H([\lambda \ \ 1 - \lambda]) + \phi_k h(\beta),$$
$$h(\psi_1) + h(\psi_2) = \alpha H(\boldsymbol{\gamma}) + h(\alpha),$$

where

$$\boldsymbol{\gamma} = \left[\begin{array}{cc} \dfrac{d_1}{\alpha} + \lambda \dfrac{\beta\phi_k}{\alpha} & \dfrac{d_2}{\alpha} + (1-\lambda)\dfrac{\beta\phi_k}{\alpha} \end{array} \right] \in \mathbb{S}_2.$$

Hence

$$\begin{aligned} V(\boldsymbol{\psi}, \boldsymbol{\phi}) &= \alpha \left[\frac{\beta\phi_k}{\alpha} H([\lambda \ \ 1-\lambda]) + H(\boldsymbol{\gamma}) \right] \\ &+ \phi_k h(\beta) - h(\alpha) + \text{const}. \end{aligned}$$

Now the quantity inside the brackets is like $G(\lambda)$ in Lemma 2, with

$$c_1 = \frac{d_1}{\alpha}, c_2 = \frac{d_2}{\alpha}, b = \frac{\beta\phi_k}{\alpha}.$$

And these three numbers *do* add up to one. So we know from Lemma 2 that the quantity inside the brackets can be made smaller by choosing $\lambda = 0$ or 1. The choice $\lambda = 0$ causes $p_{k1}, p_{k2}, \psi_1, \psi_2$ to be replaced by

$$[p'_{k1} \ \ p'_{k2}] = [0 \ \ p_{k1} + p_{k2}],$$
$$[\psi'_1 \ \ \psi'_2] = [d_1 \ \ d_2 + \phi_k(p_{k1} + p_{k2})],$$

while the choice $\lambda = 1$ causes $p_{k1}, p_{k2}, \psi_1, \psi_2$ to be replaced by

$$[p'_{k1} \ \ p'_{k2}] = [p_{k1} + p_{k2} \ \ 0],$$
$$[\psi'_1 \ \ \psi'_2] = [d_1 + \phi_k(p_{k1} + p_{k2}) \ \ d_2].$$

In either case the numbers β, α remain the same, whence the correction term $\phi_k h(\beta) - h(\alpha)$ also remains the same. So decreasing the quantity inside the brackets reduces $V(\boldsymbol{\psi}, \boldsymbol{\phi})$. So the conclusion is that there exists a $P' \in \mathbb{S}_{n\times m}$ such that, with $\boldsymbol{\psi}' = \boldsymbol{\phi}P'$, we have

$$\begin{aligned} J_{\boldsymbol{\phi}}(P') + H(\boldsymbol{\phi}) - H(\boldsymbol{\psi}') &= \phi_k[h(p'_{k1}) + h(p'_{k2})] \\ &- [h(\psi'_1) + h(\psi'_2)] + \text{const} \\ &< \phi_k[h(p_{k1}) + h(p_{k2})] \\ &- [h(\psi_1) + h(\psi_2)] + \text{const} \\ &= V(\boldsymbol{\psi}, \boldsymbol{\phi}). \end{aligned}$$

Now, since $V(\boldsymbol{\phi}, \boldsymbol{\psi}')$ is the *minimum* of the quantity $J_{\boldsymbol{\phi}}(Q)$ over all $Q \in \mathbb{S}_{n\times m}$ such that $\boldsymbol{\phi}Q = \boldsymbol{\psi}'$, we conclude from the above that

$$V(\boldsymbol{\phi}, \boldsymbol{\psi}') + H(\boldsymbol{\phi}) - H(\boldsymbol{\psi}') < V(\boldsymbol{\psi}, \boldsymbol{\phi}),$$

or equivalently that

$$V(\boldsymbol{\psi}', \boldsymbol{\phi}) < V(\boldsymbol{\psi}, \boldsymbol{\phi}).$$

Similarly, we can compare $V(\boldsymbol{\phi}, \boldsymbol{\psi}')$ and $V(\boldsymbol{\phi}, \boldsymbol{\psi})$. Since $p'_{k1} + p'_{k2} = p_{k1} + p_{k2}$, and one of p'_{k1}, p'_{k2} is zero, it is obvious that

$$\phi_k[h(p'_{k1}) + h(p'_{k2})] < \phi_k[h(p_{k1}) + h(p_{k2})].$$

Hence

$$J_{\boldsymbol{\phi}}(P') < J_{\boldsymbol{\phi}}(P) = V(\boldsymbol{\phi}, \boldsymbol{\psi}).$$

As a consequence, we have as before that

$$\begin{aligned} V(\boldsymbol{\phi}, \boldsymbol{\psi}') &= \min_{Q \in \mathbb{S}_{n \times m}} J_{\boldsymbol{\phi}}(Q) \text{ s.t. } \boldsymbol{\phi} Q = \boldsymbol{\psi}' \\ &\leq J_{\boldsymbol{\phi}}(P') < V(\boldsymbol{\phi}, \boldsymbol{\psi}). \end{aligned}$$

Combining both inequalities leads to the desired conclusion, namely that

$$d(\boldsymbol{\phi}, \boldsymbol{\psi}') < d(\boldsymbol{\phi}, \boldsymbol{\psi}).$$

This completes the proof.

IV. Finding an Optimal Aggregation: A Reformulation

Now that we know that any reduced-order approximation in the variation of information metric must be an aggregation, the next logical step is to characterize the distance between a distribution and its aggregations, and choose one aggregation that is closest to the original distribution. That is the objective of this section. We first show that minimizing the distance between a given distribution $\boldsymbol{\phi}$ and its aggregation $\boldsymbol{\phi}^{(a)}$ is equivalent to maximizing the entropy of $\boldsymbol{\phi}^{(a)}$. Since there are $O(m^n)$ aggregations, finding one with maximum entropy turns out to be NP-hard (not surprisingly). So we reformulate the problem as a bin-packing problem with over-stuffing, and propose a greedy algorithm. A worst-case performance bound for the greedy algorithm is derived.

Now we can ask: What is the best possible aggregation $\boldsymbol{\phi}^{(a)}$ that is 'closest' to $\boldsymbol{\phi}$? If $\boldsymbol{\phi}^{(a)}$ is an aggregation of $\boldsymbol{\phi}$, it is obvious that $V(\boldsymbol{\phi}, \boldsymbol{\phi}^{(a)}) = 0$. By (4) it follows that $V(\boldsymbol{\phi}^{(a)}, \boldsymbol{\phi}) = H(\boldsymbol{\phi}) - H(\boldsymbol{\phi}^{(a)})$. Applying the definitions of d shows that

$$d(\boldsymbol{\phi}, \boldsymbol{\phi}^{(a)}) = H(\boldsymbol{\phi}) - H(\boldsymbol{\phi}^{(a)}).$$

Since $H(\boldsymbol{\phi})$ is a part of the data, minimizing $d(\boldsymbol{\phi}, \boldsymbol{\phi}^{(a)})$ requires us to *maximize* the entropy $H(\boldsymbol{\phi}^{(a)}$. This leads to the

The Optimal Aggregation Problem to Maximize Entropy: Given $\boldsymbol{\phi} \in \mathbb{S}_n$ and an integer $m < n$, find an aggregation of $\boldsymbol{\phi}$ into $\mathbb{S}_m$ with maximum entropy.

Note that if $m = 2$, the aggregation $\boldsymbol{\phi}^{(a)}$ has maximum entropy if and only if it is closest to the uniform vector $\mathbf{u}_2$ in the total variation metric. In turn, finding an aggregation such that $\rho(\boldsymbol{\phi}^{(a)}, \mathbf{u}_2)$ is minimized (where ρ denotes the total variation metric) is equivalent to a bin-packing problem with overstuffing where both bin sizes are equal, and this problem is NP-hard. Hence it is plausible that the above problem is also NP-hard. Moreover, a natural suboptimal algorithm is also not readily available, unless we reformulate the problem, which is the next step.

We observe that amongst all distributions in $\mathbb{S}_m$, the uniform distribution has the maximum entropy. Thus we attempt to aggregate $\boldsymbol{\phi}$ in such a way that every component of $\boldsymbol{\phi}^{(a)}$ is as close as possible to $1/m$. That problem is a special case of aggregating $\boldsymbol{\phi}$ in such a way that every component of $\boldsymbol{\phi}^{(a)}$ is as close as possible to the corresponding component of a given distribution $\boldsymbol{\psi} \in \mathbb{S}_m$, which need not be the uniform distribution. This more general problem is formulated in [14], as a follow up to earlier work in [6], [7], [8], and can be stated as follows.

Optimal Aggregation in Total Variation Metric to a Desired Distribution: Given $\boldsymbol{\phi} \in \mathbb{S}_n$, $\boldsymbol{\psi} \in \mathbb{S}_m$, find an aggregation $\boldsymbol{\phi}^{(a)}$ of $\boldsymbol{\phi}$ such that the total variation metric $\rho(\boldsymbol{\phi}^{(a)}, \boldsymbol{\psi})$ is as small as possible, where the total variation metric ρ between $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{S}_m$ is defined as:

$$\begin{aligned} \rho(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2} \sum_{j \in \mathbb{B}} |\alpha_j - \beta_j| &= \sum_{j \in \mathbb{B}} \max\{\alpha_j - \beta_j, 0\} \\ &= \sum_{j \in \mathbb{B}} \max\{\beta_j - \alpha_j, 0\}. \end{aligned}$$

V. A Greedy Algorithm for Optimal Aggregation

As shown below, the optimal aggregation problem can be formulated as a non-standard bin-packing problem. Specifically, the problem of optimal aggregation in the total variation metric can be thought of as a bin-packing problem with the longer probability distribution $\phi_1, \ldots, \phi_n$ as the 'list' to be packed, and the shorter distribution $\psi_1, \ldots, \psi_m$ as the capacity of the 'bins', while minimizing the unutilized capacity. This problem differs from the conventional bin-packing problem in at least three respects:

- In the standard bin-packing problem, all bins have the same capacity, whereas here they need not.
- In the standard bin-packing problem, if a list item does not fit any bin, then a new bin is created; here the number of bins is fixed.
- Since $\sum_i \phi_i = \sum_j \psi_j = 1$, if all list items *have* to be put into the available bins, then some bins need to be 'overstuffed', that is, have their capacity exceeded.

Fortunately, thanks to the propensity of the research community to study every possible variation of a problem, this very situation has been studied in [18]. We don't use their results directly; rather, we adapt their method of proof to the situation at hand. First we adapt the LS algorithm to the situation where the number of bins is fixed but overstuffing is allowed.

1) Sort the elements of $\boldsymbol{\phi}, \boldsymbol{\psi}$ into descending order of magnitude.
2) Set i (the round counter) to 0, and set the initial bin capacities as $c_j = \psi_j$ for $j = 1, \ldots, m$.
3) Increment the counter i by one until $i = n$. Include the element ϕ_i into the bin with the greatest capacity c_j, and then replace c_j by $c_j - \phi_i$. If $c_j - \phi_i < 0$ then put no more elements in bin j. End when $i = n$.

Theorem 3: For the LS algorithm described above, we have

$$\rho(\boldsymbol{\phi}^{(a)}, \boldsymbol{\psi}) \leq 0.25 m \phi_{\max}. \tag{9}$$

where $\boldsymbol{\phi}^{(a)}$ is the aggregation produced by the algorithm, and $\phi_{\max} = \max_i \{\phi_i\}$.

Proof: The steps in the proof follow the corresponding steps in [18]. Once the greedy algorithm is completed, let us denote the resulting aggregation $\boldsymbol{\phi}^{(a)}$ by $\boldsymbol{\alpha}$ to reduce clutter. Let us refer to bin j as 'heavy' if $\alpha_j > \psi_j$, and 'light' if $\alpha_j \leq \psi_j$. Suppose there are k heavy bins. Without loss of generality, renumber the bins such that the first k bins are heavy and the rest are light. Let $e_1, \ldots, e_k$ denote the excess and $s_{k+1}, \ldots, s_m$ denote the slack. In other words,

$$e_j = \alpha_j - \psi_j, j = 1, \ldots, k,$$

and

$$s_j = \psi_j - \alpha_j, j = k+1, \ldots, m.$$

For $j = 1, \ldots, k$, let r_j denote its excess capacity *just before* the last item was placed into it (making it heavy). Then two things are obvious. First, $r_j + e_j$ equals the last component of $\boldsymbol{\phi}$ that was placed into this bin, and as a result $r_j + e_j \leq \phi_{\max}$. Second, the nature of the LS algorithm implies that r_j is at least equal to the capacity of all the other bins at the time this item was placed into bin j. Since bin capacity can only decrease as the algorithm is run, in particular this implies that

$$r_j \geq s_{k+1}, \ldots, s_m, j = 1, \ldots, k.$$

Therefore

$$\frac{1}{k} \sum_{j=1}^{k} r_j \geq \min_{j=1,\ldots,k} r_j \geq \max_{j=k+1,\ldots,m} s_j \geq \frac{1}{m-k} \sum_{j=k+1}^{m} s_j.$$

Rearranging gives

$$(m-k) \sum_{j=1}^{k} r_j \geq k \sum_{j=k+1}^{m} s_j.$$

Since both $\boldsymbol{\phi}, \boldsymbol{\psi}$ are unit vectors, it follows that

$$\sum_{j=1}^{k} e_j = \sum_{j=k+1}^{m} s_j.$$

Therefore

$$(m-k) \sum_{j=1}^{k} (r_j + e_j) \geq m \sum_{j=k+1}^{m} s_j.$$

Note that the right side is precisely $m\rho(\boldsymbol{\alpha}, \boldsymbol{\psi})$. Hence

$$\begin{aligned} \rho(\boldsymbol{\alpha}, \boldsymbol{\psi}) &\leq \frac{m-k}{m} \sum_{j=1}^{k} (r_j + e_j) \\ &\leq \frac{k(m-k)}{m} \phi_{\max} \leq \frac{m\phi_{\max}}{4}, \end{aligned} \tag{10}$$

which follows from the obvious observation that $k(m-k) \leq m^2/4$ no matter what k is.

It is quite easy to show that the performance of the algorithm is bounded by $0.5 m \phi_{\max}$. This is because no bin can be overstuffed by more than $\phi_{\max}$, and no bin can have unutilized capacity more than $\phi_{\max}$. Since the totals of over- and under-capacity have to balance out, the bound $0.5 m \phi_{\max}$ follows. Thus the real essence of the theorem is to gain an extra factor of 0.5.

The specific result of [18] bounds the *total weight* of all the bins (call it A) and shows that $A_{\text{LS}} \leq 1.25 A_{\text{opt}}$, where LS is the on-line list-scheduling algorithm. Moreover, they also require an extra assumption that $\phi_{\max} \leq \psi_{\min}$, something that is not needed here. It is easy to verify that the weight of an algorithm equals $1 + \rho(\boldsymbol{\phi}^{(a)}, \boldsymbol{\psi})$ achieved by that algorithm. Hence a direct application of the results of [18] would imply that

$$\rho_{\text{LS}}(\boldsymbol{\phi}^{(a)}, \boldsymbol{\psi}) \leq 1.25 \rho_{\text{opt}}(\boldsymbol{\phi}^{(a)}, \boldsymbol{\psi}) + 0.25.$$

Because of the additive constant of 0.25, this bound is less useful than the bound (9) given by Theorem 3.

The above analysis works also when the bins are nonuniform in size. Moreover, as pointed out in [18], the LS (in this case BFD) algorithm actually works better when the bin sizes are widely disparate. However, the problem of optimal aggregation to maximize entropy is a conventional bin-packing problem with equal bin sizes of $1/m$ and over-stuffing permitted. The bound given in Theorem 3 holds in this case as well.

VI. Example of Aggregation Using the Greedy Algorithm

Example 1: To illustrate the above algorithm, we solve a 40×10 problem.[1] First two uniformly distributed random vectors $\mathbf{x} \in [0,1]^{40}, \mathbf{y} \in [0,1]^{10}$ were generated using the `rand` command of `Matlab`. Then these were stretched out via the transformation

$$\phi_i = \exp(x_i)/s_1, \psi_j = \exp(y_j)/s_2,$$

where s_1, s_2 are scaling constants to make the sums come out equal to one. Then *only the smaller vector* is sorted in descending order. The results are shown below. For display purposes the resulting $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ are shown as a matrices, though in reality both are row vectors.

$$\boldsymbol{\phi} = \begin{bmatrix} 0.0304 & 0.0333 & 0.0153 & 0.0335 & 0.0253 \\ 0.0148 & 0.0178 & 0.0232 & 0.0350 & 0.0353 \\ 0.0157 & 0.0355 & 0.0350 & 0.0219 & 0.0299 \\ 0.0155 & 0.0205 & 0.0336 & 0.0297 & 0.0351 \\ 0.0259 & 0.0139 & 0.0314 & 0.0342 & 0.0265 \\ 0.0287 & 0.0283 & 0.0199 & 0.0259 & 0.0160 \\ 0.0273 & 0.0139 & 0.0177 & 0.0141 & 0.0148 \\ 0.0307 & 0.0270 & 0.0185 & 0.0348 & 0.0139 \end{bmatrix},$$

$$\boldsymbol{\psi} = \begin{bmatrix} 0.1241 & 0.1205 & 0.1192 & 0.1139 & 0.1069 \\ 0.0914 & 0.0875 & 0.0869 & 0.0821 & 0.0675 \end{bmatrix}.$$

[1]The diary of the example is available upon request from the author.

Applying the best fit algorithm for aggregation *without sorting* ϕ results in the following grouping and aggregation (shown as a matrix for convenience):

$$I_1 = \{1, 16, 23, 33\}, I_2 = \{2, 18, 31\}, I_3 = \{3, 12, 24, 40\},$$

$$I_4 = \{4, 19, 30, 36\}, I_5 = \{5, 15, 27, 38\}, I_6\{6, 11, 17, 25, 34\},$$

$$I_7 = \{7, 13, 26, 37\}, I_8 = \{8, 14, 22, 28, 35\},$$

$$I_9 = \{9, 20, 32, 39\}, I_{10} = \{10, 21, 29\},$$

$$\phi^{(a)} = \begin{bmatrix} 0.0951 & 0.0942 & 0.0989 & 0.1099 & 0.1020 \\ 0.0917 & 0.1085 & 0.0938 & 0.1188 & 0.0871 \end{bmatrix}.$$

We have that $H(\phi^{(a)} = 2.2934$, quite close to the theoretical maximum of 2.3026.

In contrast, if we first sort ϕ before applying the best fit algorithm, the following grouping results:

$$I_1 = \{1, 20, 21, 39\}, I_2 = \{2, 19, 22, 40\}, I_3 = \{3, 18, 23, 38\},$$

$$I_4 = \{4, 17, 24, 37\}, I_5 = \{5, 16, 25, 36\}, I_6 = \{6, 15, 29, 31\},$$

$$I_7 = \{7, 14, 27, 34\}, I_8 = \{8, 13, 26, 35\},$$

$$I_9 = \{9, 12, 28, 33\}, I_{10} = \{10, 11, 30, 32\}.$$

The resulting aggregation is

$$\phi^{(a)} = \begin{bmatrix} 0.1019 & 0.1021 & 0.1016 & 0.1007 & 0.1004 \\ 0.0982 & 0.0994 & 0.0993 & 0.0982 & 0.0982 \end{bmatrix},$$

which is much closer to being uniform than the earlier aggregation.

VII. Conclusions

In this paper we have studied the problem of finding an optimal lower-order approximation to a higher-order probability distribution, where the metric distance between the original and reduced-order distributions is the variation of information metric introduced in a companion paper [16]. It is first shown that every optimal reduced-order approximation must in fact be an *aggregation* of the original distribution. Thus the optimal order reduction problem is shown to be equivalent to finding an aggregation that has maximum entropy. Since this problem is NP-hard, we have reformulated it as a problem of bin-packing with over-stuffing, which is also NP-hard. However, for the latter problem we are able to give a greedy algorithm and also to prove an upper bound on its performance. The approach has been illustrated by a fairly large example of finding a 10-th order approximation to a 40-th order distribution.

Note that a preprint that combines both [16] as well as the present paper can be found at [15].

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 [9].

[3] Edward G. Coffman, Jr., János Csirik and Joseph Y.-T. Leung, "Variants of classical one-dimensional bin packing", Chapter 33 in [9].

[4] Edward G. Coffman, Jr., János Csirik and Joseph Y.-T. Leung, "Variable-sized bin packing and bin covering", Chapter 34 in [9].

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

[6] 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.

[7] 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.

[8] 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*.

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

[10] 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.

[11] 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.

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

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

[14] 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.

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

[16] M. Vidyasagar, "A metric between probability distributions on finite sets of different cardinalities," to be presented at CDC 2011.

[17] Wikipedia page on mutual information, found at http://en.wikipedia.org/wiki/Mutual_information

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

[19] 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.