A Novel Approach to Distributed Quantization via Multivariate Information Bottleneck Method

Shayan Hassanpour, Dirk Wübben, and Armin Dekorsy
Department of Communications Engineering
University of Bremen, 28359 Bremen, Germany
Email: {hassanpour, wuebben, dekorsy}@ant.uni-bremen.de

***Abstract*—Consider following setup: A number of observations from a data source shall be compressed jointly prior to a forward transmission via several rate-limited links to a central processing unit. To design the respective quantizers, here, *Mutual Information* is chosen as the fidelity criterion and the broad-ranging structure of *Multivariate Information Bottleneck* is then aptly tailored to that purpose. This, indeed, not only yields a novel design approach for the considered distributed scenario but also paves the way towards perceiving the chance of leveraging this flexible conceptual frame in a vast variety of applications regarding digital data transmission. Explicitly, it immediately enables addressing various extensions of the presumed arrangement, incorporating the parallel construction of intertwined compression systems for several correlated sources.**

I. Introduction

The joint compression of multiple observations from a given source is considered. This frequently appearing distributed setup is, indeed, the underlying scenario in a variety of applications, i.a., decentralized inference sensor networks wherein a certain number of measured (sensed) values must be quantized ahead of transmission to the fusion center [1], cooperative relaying schemes with *Quantize-and-Forward* strategy [2], and last but not least, *Cloud-based Radio-Access Networks* with rate-limited fronthaul links to the central processor in the cloud [3].

Most studies in the available literature on this setup follow the *Rate-Distortion* philosophy and propose some algorithmic approaches for the quantization design problem w.r.t. a specific distortion measure, e.g., the *Mean-Squared-Error (MSE)* [4], the *Ali-Silvey* distance [5], or the *Fisher Information* [6]. Contrary to the previous investigations, here we employ the novel design paradigm of the *Multivariate Information Bottleneck (MIB)* [7]. MIB is an immediate extension of the preliminary idea of the *Information Bottleneck (IB)* [8] that has emerged originally in the *Machine Learning* context as a novel, information-theoretic approach towards *Clustering* which is a fundamental task in the sub-branch of *Unsupervised Learning* [9].

To put it in a nutshell, the IB method is a variational principle aiming for compressing a *Random Variable (RV)* in a fashion that it retains most of the information content w.r.t. another relevant variable and, interestingly, this preservation capability can be controlled through twiddling a trade-off parameter. To attain an overall picture on the IB method and several related algorithmic approaches, interested readers are referred to [10]–[12]. There exist a number of intriguing aspects which support the idea of deploying this framework for communication applications as well. Concerning a totally connected example, in case of noisy source coding, following the IB philosophy, a purely statistical design structure is achieved which directly engages the actual source into its formulation. Besides, a major special instance of this principle boils down to designing quantizers that maximize the end-to-end data transmission rate for a given input statistics, something sought in (almost) all communication schemes. In fact, the IB paradigm has already found its path into various aspects of modern transmission systems from construction of polar codes [13] to advanced discrete (channel) decoding concepts [14] with relatively low complexity and yet quite promising performance.

MIB is a generic principle that not only enables considering the cases for which the compression shall be relevant w.r.t. multiple variables but also allows for simultaneous construction of several systems of clusters. To make that happen, it utilizes the concept of *Multi-Information*, a natural extension of the pairwise concept of *Mutual Information*, over two *Bayesian Networks (BNs)*. The first network stipulates the imposed constraints, i.e., statistical independencies among the involved RVs, and identifies the set of compression variables. The second one, specifies the relations that shall be retained. The general principle is then formulated as a trade-off between the multi-information each network carries. The fascinating feature of this mathematical establishment is that the optimal solution and subsequently the relevant algorithms are derived formally, i.e., irrespective of particular choices of BNs. This, indeed, brings about a lot of flexibility into play and turns the MIB into a comprehensive framework that can be suitably applied to address a wide range of applications, especially, more sophisticated situations wherein multiple RVs are involved.

To vividly demonstrate the usability of exploiting MIB, within this work we consider the predescribed distributed quantization setup and tailor the general framework of MIB to that matter. An asymptotic case of this *Variational Principle* then aims for maximizing the mutual information between the given source and the random vector comprising all the compressed variables. This scenario has been recently investigated in [1] and as shown, it engenders a set of quantizers which perform quite comparably to the ones exclusively designed for the estimation and detection purposes. That can be reckoned as another cogent argument for MIB deployment. Indeed, it will be shown that our suggested algorithm not only outperforms the proposed approach in [1], but also broadens the scope of the underlying problem through establishing a fundamental trade-off between the acquired level of compression on the one hand and the amount of achievable relevant information preservation on the other.

978-1-7281-0962-6/19/$31.00 ©2019 IEEE

II. Multivariate Information Bottleneck

A. Preliminary IB Method

The original IB setup [8] considers the quantization of a given RV, a_2, into the compression variable, z, such that it is highly informative w.r.t. a relevant variable, a_1. As a straightforward translation to the context of *Noisy Source Coding (NSC)*, one can think of a_2 as the noisy observation of the source, a_1. The aim is then to have a compressed representation, z, of the observation, a_2, that still preserves most of its information content w.r.t. the source, a_1. It is presumed that the joint distribution $p(\mathsf{a}_1,\mathsf{a}_2)$ is given and, further, $\mathsf{a}_1 \leftrightarrow \mathsf{a}_2 \leftrightarrow \mathsf{z}$ institutes a Markov chain. The IB method then establishes a fundamental trade-off between the *compactness* and *informativity* of its outcome in a symmetric fashion, employing mutual information [15] terms to quantify each aspect. On the one hand, $I(\mathsf{a}_2;\mathsf{z})$ is considered as the term gauging the compactness of the outcome. Clearly, lower values of this quantity signify acuter compression and vice versa. A more formal interpretation relates $I(\mathsf{a}_2;\mathsf{z})$ to maximal number of bits that can be reliably transmitted over quantizer block, exploiting the *Asymptotic Equipartition Property* [15]. On the other hand, $I(\mathsf{a}_1;\mathsf{z})$ is chosen as the indicator of information preservation.

The quantizer design problem is then mathematically stated as finding the mapping $p(\mathsf{z}|\mathsf{a}_2)$ that minimizes the IB functional $\mathcal{L}_{\text{IB}} = I(\mathsf{a}_2;\mathsf{z}) - \beta I(\mathsf{a}_1;\mathsf{z})$, in which β denotes a non-negative trade-off parameter. Applying the *Variational Calculus*, a formal characterization of the optimal solution to the pertinent design problem is derived in [8] for each pair $(a_2, z) \in \mathcal{A}_2 \times \mathcal{Z}$ as

$$p(z|a_2) = \frac{p(z)}{\psi(a_2,\beta)} \exp\big(-\beta d(z,a_2)\big), \tag{1}$$

wherein $\psi(a_2,\beta)$ is a normalization function assuring a valid distribution and the *Relevant Distortion*, $d(z,a_2)$, is given as

$$d(z,a_2) = D_{\text{KL}}\big(p(\mathsf{a}_1|a_2)\|p(\mathsf{a}_1|z)\big), \tag{2}$$

with $D_{\text{KL}}(\cdot\|\cdot)$ denoting the *Kullback-Leibler* divergence [15]. Further, an iterative algorithm is also given in [8] that exerts the *Fixed-Point Iteration* method [16] on the optimal solution (1).

B. Structural Extension to Multivariate Setup

A highly generalized version of the previous arrangement is then to have a number of compression variables, $\mathsf{z}_j : 1 \le j \le J$, each quantizing a certain subset, $\mathbf{y}_j$, of the set of input RVs, $\mathbf{a} = \{\mathsf{a}_i \,|i\}$, while preserving information about another arbitrary subset, $\mathbf{x}_j$, of elements in $\mathbf{a}$. In occasions of dealing with multiple RVs, the concept of *Multi-information* [17] will be the counterpart of the pairwise concept of mutual information. It is defined as

$$\mathcal{I}\big(p(\mathbf{a})\big) = \sum_{\boldsymbol{a}} p(\boldsymbol{a}) \log \frac{p(\boldsymbol{a})}{\prod\limits_i p(a_i)}, \tag{3}$$

which captures the average amount of bits that can be secured by the joint vs. independent compression of elements in $\mathbf{a}$.

The Bayesian Network, $\mathcal{G}$, is a powerful tool to describe the statistical relations among the RVs in $\mathbf{a}$. It is a directed acyclic graph that considers the entries of $\mathbf{a}$ as its nodes and encodes the proper factorization of $p(\mathbf{a})$ with its edges in a sense that it applies $p(\mathbf{a}) = \prod_i p(\mathsf{a}_i|\mathbf{P}^{\mathcal{G}}_{\mathsf{a}_i})$, with $\mathbf{P}^{\mathcal{G}}_{\mathsf{a}_i}$ denoting the parent nodes of a_i in $\mathcal{G}$. In that case, the multi-information (3) can be calculated as the sum of *local* mutual information terms between each variable a_i and its parents $\mathbf{P}^{\mathcal{G}}_{\mathsf{a}_i}$. Note that for an arbitrary distribution $q(\mathbf{a})$ that may not be correctly factorable as $\mathcal{G}$ suggests, this sum of local mutual information terms is still well defined. This leads to the definition of *Multi-Information in* $q(\mathbf{a})$ *w.r.t. the BN*, $\mathcal{G}$, [7] as

$$\mathcal{I}^{\mathcal{G}}\big(q(\mathbf{a})\big) = \sum_i I(\mathsf{a}_i;\mathbf{P}^{\mathcal{G}}_{\mathsf{a}_i}), \tag{4}$$

where every local mutual information term is calculated using the marginal distributions of $q(\mathbf{a})$. In general, $\mathcal{I}\big(q(\mathbf{a})\big) > \mathcal{I}^{\mathcal{G}}\big(q(\mathbf{a})\big)$ and their gap measures how close is $q(\mathbf{a})$ to the class of distributions being correctly factorable as suggested by the structure of $\mathcal{G}$.

Then, Slonim et al. in [7] make use of two BNs, $\mathcal{G}_{\text{in}}$ and $\mathcal{G}_{\text{out}}$, to establish the MIB variational principle as a trade-off between the multi-information term each network carries. The structure of input BN, $\mathcal{G}_{\text{in}}$, determines the *Solution Space* and also signifies "*what quantizes what*". Basically, the statistical relations among input RVs get projected in the construction of $\mathcal{G}_{\text{in}}$. Furthermore, the compression variables, $\mathsf{z}_j : 1 \le j \le J$, appearing as the leaves in $\mathcal{G}_{\text{in}}$, are set to be the children of $\mathbf{y}_j$, i.e., the RVs they have to represent compactly. Hence, via BN conventions (see, e.g., [18]), given its parents, $\mathbf{y}_j$, each compression variable, z_j, is assumed to be independent of other nodes. The multi-information, $\mathcal{I}^{\mathcal{G}_{\text{in}}}$, will then be a suitable measure for indication of the *compactness* of outcome, as based on (4) and similar to the preliminary IB setup, it contains input-output mutual information terms $I(\mathsf{z}_j;\mathbf{y}_j)$ for all the involved quantizers. The output BN, $\mathcal{G}_{\text{out}}$, on the other hand, specifies "*what is informative w.r.t. what*" and is built in a fashion that each compression variable, z_j, is set to be the parent of its relevant RVs, i.e., $\mathbf{x}_j$. By doing so, the multi-information, $\mathcal{I}^{\mathcal{G}_{\text{out}}}$, becomes a natural gauge regarding the *informativity* of outcome as it sums up all the relevant mutual information terms $I(\mathsf{z}_j;\mathbf{x}_j)$. Analogous to the original IB setup, the trade-off between both aspects can then be formalized as minimizing the MIB functional, $\mathcal{L}_{\text{MIB}} = \mathcal{I}^{\mathcal{G}_{\text{in}}} - \beta \mathcal{I}^{\mathcal{G}_{\text{out}}}$, with β playing the same role as before. This minimization is carried out over the complete set of mappings $\{p(\mathsf{z}_j|\mathbf{y}_j)\ |j\}$ from subsets of $\mathbf{a}$ entries which are intended to be quantized, i.e., $\mathbf{y}_j$, and their pertinent compact versions, z_j.

To more tangibly understand the above description, one may use the original IB setup as a simple yet illustrative example. For that, the respective input/output BNs are depicted in Fig. 1. The input BN, $\mathcal{G}_{\text{in}}$, stipulates that the overall joint distribution shall be factorable as $p(\mathsf{a}_1,\mathsf{a}_2,\mathsf{z}) = p(\mathsf{a}_1)p(\mathsf{a}_2|\mathsf{a}_1)p(\mathsf{z}|\mathsf{a}_2)$. This indicates the presumed Markov chain $\mathsf{a}_1 \leftrightarrow \mathsf{a}_2 \leftrightarrow \mathsf{z}$ in the original IB setup. Further, noting both BNs, it is realized that z must be a compressed representation of $\mathsf{y} = \mathsf{a}_2$ such that it is informative w.r.t. $\mathsf{x} = \mathsf{a}_1$. Applying (4), it holds $\mathcal{I}^{\mathcal{G}_{\text{in}}} = I(\mathsf{a}_2;\mathsf{a}_1) + I(\mathsf{a}_2;\mathsf{z})$ and $\mathcal{I}^{\mathcal{G}_{\text{out}}} = I(\mathsf{a}_1;\mathsf{z})$. Since $I(\mathsf{a}_2;\mathsf{a}_1)$ is a fixed term (joint distribution of all input RVs are assumed to be fixed), it can be dropped and then the MIB functional, $\mathcal{L}_{\text{MIB}}$, equals the IB functional, $\mathcal{L}_{\text{IB}}$.

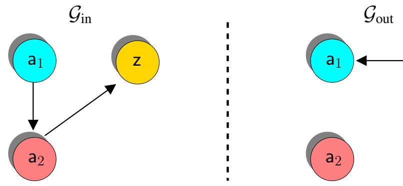


Fig. 1. Input BN, $\mathcal{G}_{\text{in}}$, and output BN, $\mathcal{G}_{\text{out}}$, of the original IB setup

C. Optimal Solution & an Iterative Design Algorithm

For a given trade-off parameter, β, and input statistics, $p(\mathbf{a})$, a formal optimal solution (yielding a stationary point of $\mathcal{L}_{\text{MIB}}$) regarding any of the present mappings $p(\mathsf{z}_j|\mathbf{y}_j)$ for $1\leq j\leq J$ between the compression variable, z_j, and its parents in $\mathcal{G}_{\text{in}}$ denoted by $\mathbf{y}_j=\mathbf{P}_{\mathsf{z}_j}^{\mathcal{G}_{\text{in}}}$ is derived for each $(z_j,\boldsymbol{y}_j)\in\mathcal{Z}_j\times\boldsymbol{\mathcal{Y}}_j$ as [7]

$$p(z_j|\boldsymbol{y}_j)=\frac{p(z_j)}{\psi_{\mathsf{z}_j}(\boldsymbol{y}_j,\beta)}\exp\bigl(-\beta d(z_j,\boldsymbol{y}_j)\bigr)\ . \tag{5}$$

$\psi_{\mathsf{z}_j}(\boldsymbol{y}_j,\beta)$ is a partition function that assures a valid conditional distribution and the *Multivariate Relevant Distortion (MRD)*, $d(z_j,\boldsymbol{y}_j)$, is calculated as

$$\begin{aligned}d(z_j,\boldsymbol{y}_j)=&\sum_{i:\mathsf{z}_j\in\mathbf{v}_{\mathsf{a}_i}}\mathbb{E}_{p(\cdot|\boldsymbol{y}_j)}\Bigl\{D_{\text{KL}}\bigl(p(\mathsf{a}_i|\boldsymbol{v}_{\mathsf{a}_i}^{-j},\boldsymbol{y}_j)\|p(\mathsf{a}_i|\boldsymbol{v}_{\mathsf{a}_i}^{-j},z_j)\bigr)\Bigr\}\\&+\sum_{\ell:\mathsf{z}_j\in\mathbf{v}_{\mathsf{z}_\ell}}\mathbb{E}_{p(\cdot|\boldsymbol{y}_j)}\Bigl\{D_{\text{KL}}\bigl(p(\mathsf{z}_\ell|\boldsymbol{v}_{\mathsf{z}_\ell}^{-j},\boldsymbol{y}_j)\|p(\mathsf{z}_\ell|\boldsymbol{v}_{\mathsf{z}_\ell}^{-j},z_j)\bigr)\Bigr\}\\&+D_{\text{KL}}\bigl(p(\mathbf{v}_{\mathsf{z}_j}|\boldsymbol{y}_j)\|p(\mathbf{v}_{\mathsf{z}_j}|z_j)\bigr),\end{aligned} \tag{6}$$

with $\mathbf{v}_{\mathsf{a}_i}=\mathbf{P}_{\mathsf{a}_i}^{\mathcal{G}_{\text{out}}}$, $\mathbf{v}_{\mathsf{z}_\ell}=\mathbf{P}_{\mathsf{z}_\ell}^{\mathcal{G}_{\text{out}}}$, denoting sets of parent nodes of a_i and z_ℓ in $\mathcal{G}_{\text{out}}$, meaning the RVs that have to be informative about a_i and z_ℓ, respectively, and $\mathbf{v}_{\mathsf{a}_i}^{-j}=\mathbf{v}_{\mathsf{a}_i}\backslash\{\mathsf{z}_j\}$, $\mathbf{v}_{\mathsf{z}_\ell}^{-j}=\mathbf{v}_{\mathsf{z}_\ell}\backslash\{\mathsf{z}_j\}$. Moreover, by definition

$$\begin{aligned}&\mathbb{E}_{p(\cdot|\boldsymbol{y}_j)}\Bigl\{D_{\text{KL}}\bigl(p(\mathsf{b}|\boldsymbol{r},\boldsymbol{y}_j)\|p(\mathsf{b}|\boldsymbol{r},z_j)\bigr)\Bigr\}\\&=\sum_{\boldsymbol{r}}p(\boldsymbol{r}|\boldsymbol{y}_j)D_{\text{KL}}\bigl(p(\mathsf{b}|\boldsymbol{r},\boldsymbol{y}_j)\|p(\mathsf{b}|\boldsymbol{r},z_j)\bigr),\end{aligned} \tag{7}$$

where b and $\mathbf{r}$ denote a RV and a set of RVs (a random vector), respectively. It should be noted that the first summand in (6) concerns all input RVs, a_i, where z_j must preserve information about while its second summand contributes in cases where z_j must be informative w.r.t. some other compression variables, z_ℓ, as well. Eventually, the third summand in (6) comes into play when information shall be maintained by at least one of the other compression variables w.r.t. z_j itself. From the form of (5) it is directly inferred that for a given $\boldsymbol{y}_j$, the lower the value of $d(z_j,\boldsymbol{y}_j)$, the higher the probability of assigning $\boldsymbol{y}_j$ to the cluster $z_j\in\mathcal{Z}_j$. Principally, the better z_j represents $\boldsymbol{y}_j$, the lower become the respective KL divergences in (6) and, consequently, the larger gets the probability of allotting $\boldsymbol{y}_j$ to z_j. It is also noteworthy that for the given input/output BNs in Fig. 1, the MRD in (6) reduces to the provided relevant distortion in (2).

Since minimizing the MIB functional, $\mathcal{L}_{\text{MIB}}$, w.r.t. the set of all involved mappings, $\{p(\mathsf{z}_j|\mathbf{y}_j)\ |j\}$, for a particular input statistics, $p(\mathbf{a})$, is not a convex optimization task in general [7], attaining the globally optimal solution is quite demanding. Therefore, following a pragmatic approach, one shall resort to some heuristics which aim for addressing the design problem efficiently at the cost of converging to local optima. Based on the assumption of either having a fixed or varying set of output levels, $\{|\mathcal{Z}_j|\ |j\}$, the authors in [7] have adapted the *Partitional* and *Hierarchical Clustering* concepts [19] to the MIB paradigm and proposed four heuristics to practically address its underlying optimization problem. Here, we solely discuss a generally soft (stochastic) clustering procedure known as the *Multivariate iterative IB (MultiIB)* algorithm which is, indeed, the immediate generalization of the presented routine in [8] for preliminary IB setup. In the asymptotic case of letting $\beta\to\infty$, this leads to a partitional approach. As its name suggests, the MultiIB is an iterative routine which aims for obtaining the set of required mappings, $\{p(\mathsf{z}_j|\mathbf{y}_j)\ |j\}$, by direct use of (5). Note, that (5) has an implicit form as $p(z_j)$ and $d(z_j,\boldsymbol{y}_j)$ on its right hand side depend on $\{p(\mathsf{z}_j|\mathbf{y}_j)\ |j\}$. The principal idea behind the MultiIB is then to commence with a random (still valid) initialization of the mappings, $\{p^{(0)}(\mathsf{z}_j|\mathbf{y}_j)\ |j\}$, and perform the update steps (till convergence/fulfillment of a stopping criterion) for every pair $(\boldsymbol{y}_j,z_j)\in\boldsymbol{\mathcal{Y}}_j\times\mathcal{Z}_j$ via

Alg. 1 Multivariate iterative IB (MultiIB)

Input: $p(\mathbf{a})$, $\mathcal{G}_{\text{in}}$, $\mathcal{G}_{\text{out}}$, β, $|\mathcal{Z}_j|$, convergence parameter $\varepsilon>0$
Output: Generally soft partition z_j of $\boldsymbol{\mathcal{Y}}_j$ into $|\mathcal{Z}_j|$ bins $\forall j=1:J$
Initialization: $m=0$, random mappings $\{p^{(m)}(\mathsf{z}_j|\mathbf{y}_j)\ |j\}$
while True **do**
  for $j=1:J$ **do**
  - $p^{(m)}(z_j)\leftarrow\sum_{\boldsymbol{y}_j}p^{(m)}(z_j|\boldsymbol{y}_j)p(\boldsymbol{y}_j)\quad\forall z_j\in\mathcal{Z}_j$
  - find the mth update for all distributions in $d(\mathsf{z}_j,\mathbf{y}_j)$ via marginalizing w.r.t. $p^{(m)}(\mathbf{a},\mathbf{z})=p(\mathbf{a})\prod_{j'=1}^{J}p^{(m)}(\mathsf{z}_{j'}|\mathbf{y}_{j'})$
  - $p^{(m+1)}(z_j|\boldsymbol{y}_j)\leftarrow\frac{p^{(m)}(z_j)}{\psi_{\mathsf{z}_j}^{(m+1)}(\boldsymbol{y}_j,\beta)}\exp\bigl(-\beta d^{(m)}(z_j,\boldsymbol{y}_j)\bigr)$
  - $p^{(m+1)}(\mathsf{z}_\ell|\mathbf{y}_\ell)\leftarrow p^{(m)}(\mathsf{z}_\ell|\mathbf{y}_\ell)\quad\forall\ell=1:J,\ \ell\neq j$
  - $m\leftarrow m+1$

  end for
  if $\forall j,\forall\boldsymbol{y}_j: D_{\text{JS}}^{\{\frac{1}{2},\frac{1}{2}\}}\bigl(p^{(m)}(\mathsf{z}_j|\boldsymbol{y}_j)\|p^{(m-J)}(\mathsf{z}_j|\boldsymbol{y}_j)\bigr)\leq\varepsilon$ **then**
    Break
  end if
end while

$$p^{(m+1)}(z_j|\boldsymbol{y}_j)=\frac{p^{(m)}(z_j)}{\psi_{\mathsf{z}_j}^{(m+1)}(\boldsymbol{y}_j,\beta)}\exp\bigl(-\beta d^{(m)}(z_j,\boldsymbol{y}_j)\bigr), \tag{8}$$

wherein m denotes the running index. The quantizer output probability, $p^{(m)}(z_j)$, and the respective MRD, $d^{(m)}(z_j,\boldsymbol{y}_j)$, are calculated employing $\{p^{(m)}(\mathsf{z}_j|\mathbf{y}_j)\ |j\}$ and the conditional independencies imposed by the structure of $\mathcal{G}_{\text{in}}$. Updates are performed *asynchronously*, meaning when a RV, z_j, is chosen the update will be executed merely for this variable and for every $1\leq\ell\leq J$ and $\ell\neq j$, $p^{(m+1)}(\mathsf{z}_\ell|\mathbf{y}_\ell)=p^{(m)}(\mathsf{z}_\ell|\mathbf{y}_\ell)$. To avoid getting trapped in bad local optima, this procedure is repeated several times (with different initialization) and the best outcome is retained. The pertinent pseudo-code of the MultiIB routine is presented in Alg. 1 where $D_{\text{JS}}^{\{\cdot,\cdot\}}(\cdot\|\cdot)$ stated in the termination criterion part denotes the *Jensen-Shannon (JS)* divergence [7].

III. MIB-Based Distributed Quantization

A. System Model & Problem Formulation

In this part, we focus on the predescribed distributed scenario known as the *Chief Executive Officer (CEO)* setup [20] and aptly tailor the general paradigm of MIB to that matter. The presented discussion for this concrete case study better clarifies the concise and rather abstract presentation of MIB in the previous section. Consider the presumed system model that is illustrated in Fig. 2.

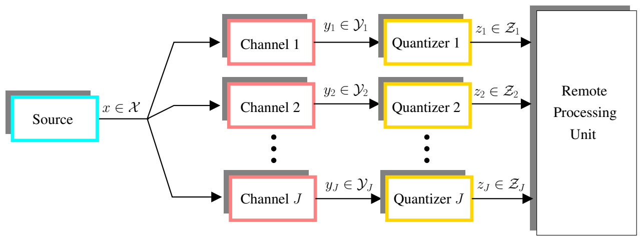


Fig. 2. Presumed system model for distributed quantization

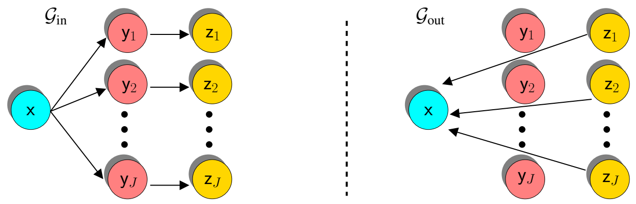


Fig. 3. Chosen input/output BNs for distributed quantization

A data source, x, is transmitted over J noisy channels. The access channels' output variables, y_j, for $1 \leq j \leq J$, then have to be compressed into the variables, z_j, for $1 \leq j \leq J$, before getting transmitted over several rate-limited fronthaul links to a central unit for further processing. To perform this task, we propose utilizing the MIB framework which leads to a joint yet local quantization of individual noisy observations y_j.

The very first step towards achieving this end is then to specify the corresponding input and output BNs. Fig. 3 illustrates them. Please note that to clearly distinguish between the source and the access channels' output variables which both are the input RVs from perspective of the MIB for this setup, we chose two different letters for pertinent denotations. As the joint probability of involved RVs, $p(\mathsf{x}, \mathbf{y}, \mathbf{z})$, must be consistent with the structure of $\mathcal{G}_{\text{in}}$, the solution space consists of distributions conforming to the layout $p(\mathsf{x}, \mathbf{y}, \mathbf{z}) = p(\mathsf{x}) \prod_j p(\mathsf{y}_j|\mathsf{x}) \prod_j p(\mathsf{z}_j|\mathsf{y}_j)$, implying the Markov relation of $\mathsf{x} \leftrightarrow \mathbf{y} \leftrightarrow \mathbf{z}$. Assuming a given input statistics, $p(\mathsf{x}, \mathbf{y})$, the free parameters are then the mappings $p(\mathsf{z}_j|\mathsf{y}_j)$ for $1 \leq j \leq J$. The compression rate that is aimed to be minimized, is the multi-information, $\mathcal{I}^{\mathcal{G}_{\text{in}}}$, calculated as

$$\mathcal{I}^{\mathcal{G}_{\text{in}}} = \sum_j I(\mathsf{x}; \mathsf{y}_j) + \sum_j I(\mathsf{y}_j; \mathsf{z}_j) \ . \tag{9}$$

Considering $\mathcal{G}_{\text{out}}$, the relevant information term that is aimed to be maximized, is the multi-information, $\mathcal{I}^{\mathcal{G}_{\text{out}}}$, being equal to

$$\mathcal{I}^{\mathcal{G}_{\text{out}}} = I(\mathsf{x}; \mathsf{z}_1, \cdots, \mathsf{z}_J) \ . \tag{10}$$

Consequently, the MIB functional is derived as

$$\mathcal{L}_{\text{MIB}}^{\text{Dist.}} = \sum_j I(\mathsf{y}_j; \mathsf{z}_j) - \beta I(\mathsf{x}; \mathsf{z}_1, \cdots, \mathsf{z}_J), \tag{11}$$

where the first summation in (9) has been dropped since it is a constant term given by the input statistics, $p(\mathsf{x}, \mathbf{y})$. The design optimization problem is then formulated as

$$Q^* = [p^*(\mathsf{z}_1|\mathsf{y}_1), \cdots, p^*(\mathsf{z}_J|\mathsf{y}_J)] = \underset{Q}{\text{argmin}} \ \mathcal{L}_{\text{MIB}}^{\text{Dist.}}, \tag{12}$$

subject to a fixed cardinality of the output levels, $\{|\mathcal{Z}_j| \ |j\}$. It shall be also noted that in the extreme case of letting $\beta \to \infty$, the design formulation in (12) boils down to

$$Q^* = \underset{Q}{\text{argmax}} \ I(\mathsf{x}; \mathsf{z}_1, \cdots, \mathsf{z}_J), \tag{13}$$

to derive which, the effective compression rate, i.e., the first term in (11) is not considered anymore and the minimization is substituted by the maximization through dropping the minus sign. Please note that even in this case, although the focus is solely on the preservation of relevant information, the effective compression rate is not allowed to grow arbitrarily large and, indeed, will be upper-bounded by $\sum_j \log_2 |\mathcal{Z}_j|$ bits. For each pair $(y_j, z_j) \in \mathcal{Y}_j \times \mathcal{Z}_j$ the optimal solution regarding the present quantizers in (12) is then given as

$$p(z_j|y_j) = \frac{p(z_j)}{\psi_{\mathsf{z}_j}(y_j, \beta)} \exp\left(-\beta d(z_j, y_j)\right), \tag{14}$$

with the respective MRD, $d(z_j, y_j)$, being equal to

$$\begin{aligned} d(z_j, y_j) &= \mathbb{E}_{p(\cdot|y_j)} \Big\{ D_{\text{KL}}\big(p(\mathsf{x}|\boldsymbol{v}_{\mathsf{x}}^{-j}, y_j) \| p(\mathsf{x}|\boldsymbol{v}_{\mathsf{x}}^{-j}, z_j)\big) \Big\} \\ &= \sum_{\boldsymbol{v}_{\mathsf{x}}^{-j}} p(\boldsymbol{v}_{\mathsf{x}}^{-j}|y_j) D_{\text{KL}}\big(p(\mathsf{x}|\boldsymbol{v}_{\mathsf{x}}^{-j}, y_j) \| p(\mathsf{x}|\boldsymbol{v}_{\mathsf{x}}^{-j}, z_j)\big), \end{aligned} \tag{15}$$

where $\boldsymbol{v}_{\mathsf{x}}^{-j} = \{z_1, \cdots, z_{j-1}, z_{j+1}, \cdots, z_J\}$. Comparing (15) with (6), one may note that the second and the third summands in (6) do not appear and the first summation is only w.r.t. the source, x.

B. Distributed Design Algorithm

Employing the MultiIB as an algorithmic approach towards addressing (12) for a given input statistics, $p(\mathsf{x}, \mathbf{y})$, a trade-off parameter, β, and a set of quantizers' output levels, $\{|\mathcal{Z}_j| \ |j\}$, one has to start with a group of random (yet valid) mappings, $\{p^{(0)}(\mathsf{z}_j|\mathsf{y}_j) \ |j\}$, and then for every particular branch, j, the respective quantizer mapping update is

$$p^{(m+1)}(z_j|y_j) = \frac{p^{(m)}(z_j)}{\psi_{\mathsf{z}_j}^{(m+1)}(y_j, \beta)} \exp\left(-\beta d^{(m)}(z_j, y_j)\right), \tag{16}$$

where

$$p^{(m)}(z_j) = \sum_{x \in \mathcal{X}} \sum_{y_j \in \mathcal{Y}_j} p(x) \, p(y_j|x) \, p^{(m)}(z_j|y_j) \ . \tag{17}$$

To calculate the corresponding MRD, $d^{(m)}(z_j, y_j)$, in (16) three conditional probabilities have to be derived (with $\boldsymbol{z} = \{z_j \ |j\}$)

$$p^{(m)}(\mathsf{x}|\boldsymbol{v}_{\mathsf{x}}^{-j}, z_j) = p^{(m)}(\mathsf{x}|\boldsymbol{z}) = \frac{p^{(m)}(\mathsf{x}, \boldsymbol{z})}{\sum_{x' \in \mathcal{X}} p^{(m)}(x', \boldsymbol{z})}, \tag{18}$$

with

$$p^{(m)}(\mathsf{x}, \boldsymbol{z}) = p(\mathsf{x}) \prod_j p^{(m)}(z_j|\mathsf{x}) \tag{19a}$$

$$p^{(m)}(z_j|\mathsf{x}) = \sum_{y_j \in \mathcal{Y}_j} p^{(m)}(z_j|y_j) \, p(y_j|\mathsf{x}), \tag{19b}$$

in which (19a) results from the fact that given the source, x, all the compression variables, z_j, for $1 \leq j \leq J$, are independent and (19b) is due to the presumed Markov chain per branch. Further, it holds

$$p^{(m)}(\boldsymbol{v}_{\mathsf{x}}^{-j}|y_j) = \frac{\sum_{x \in \mathcal{X}} p(x, y_j) \prod_{j' \neq j} p^{(m)}(z_{j'}|x)}{p(y_j)}, \tag{20}$$

and, finally,

$$p^{(m)}(\mathsf{x}|\boldsymbol{v}_{\mathsf{x}}^{-j}, y_j) = \frac{p(\mathsf{x}, y_j) \prod_{j' \neq j} p^{(m)}(z_{j'}|\mathsf{x})}{\sum_{x' \in \mathcal{X}} p(x', y_j) \prod_{j' \neq j} p^{(m)}(z_{j'}|x')} \ . \tag{21}$$

Performing the update process iteratively is, indeed, nothing but applying the *Multivariate Fixed-Point Iteration* method [16] in an asynchronous fashion (i.e., the update for z_j encompasses the implications of recent updates from all of its preceding compression variables, z_ℓ, for $1 \leq \ell \leq j-1$, a similar idea as the *Gauss-Seidel* method now applied to a nonlinear system) over the set of all mappings and their respective optimal solutions.

For finite values of the trade-off parameter, β, this yields a set of *stochastic* mappings while for case of letting $\beta \to \infty$, the normalization function, $\psi_{\mathsf{z}_j}(y_j, \beta)$, for each realization y_j concentrates all of the probability mass into only one cluster and therefore induces the quantizers to become *hard*. To justify this behavior, one shall note that the objective functional in (13) is *separately* convex w.r.t. any of the mappings $p(\mathsf{z}_j|\mathsf{y}_j)$ for $1 \leq j \leq J$. This is due to the fact that, for a given $p(\mathsf{x})$, $I(\mathsf{x};\mathbf{z})$ becomes convex w.r.t. $p(\mathbf{z}|\mathsf{x}) = \prod_j p(\mathsf{z}_j|\mathsf{x})$ [15] and as fixing the quantizer mappings $p(\mathsf{z}_j|\mathsf{y}_j)$ for all $j \neq \ell$ directly corresponds to fixing the pertinent distributions $p(\mathsf{z}_j|\mathsf{x})$, the relation among $p(\mathbf{z}|\mathsf{x})$ and $p(\mathsf{z}_\ell|\mathsf{x})$ becomes *affine* which preserves convexity. As it holds $p(\mathsf{z}_\ell|\mathsf{x}) = \sum_{y_\ell} p(\mathsf{z}_\ell|y_\ell) p(y_\ell|\mathsf{x})$ that once again defines an affine relation between $p(\mathsf{z}_\ell|\mathsf{x})$ and $p(\mathsf{z}_\ell|\mathsf{y}_\ell)$, the claim is proven. Recalling that in each processing step of MultiIB only one mapping is altered (due to asynchronous update procedure) and since the solution space of this active mapping is a closed convex polytope, $\mathcal{S}$, engendered by the Cartesian product of its constituent probability simplices [10], the objective functional in (13) obtains its maximum over one of the *extreme points* of $\mathcal{S}$ (convex maximization [21]) which correspond to its vertices, implying a hard mapping result at the end.

C. Supplementary Mathematical Discussion

To provide better insights, one shall note that the underlying design optimization (12) can be reformulated as maximizing the end-to-end transmission rate, $I(\mathsf{x};\mathbf{z})$, with a side-constraint in the form of an upper-bound on the effective compression rate, i.e., sum of the individual compression rates of different branches

$$Q^* = \underset{Q:\, \sum_j I(\mathsf{y}_j;\mathsf{z}_j) \leq R}{\operatorname{argmax}} \quad I(\mathsf{x};\mathbf{z}), \tag{22}$$

wherein each certain R value corresponds to a certain β value. Attaining the required β for a particular R is then usually done by (repeatedly) performing the bisection method over a proper initial interval of β values, running the MultiIB, calculating the resultant effective compression rate and finally modifying the current interval accordingly (up to a certain precision).

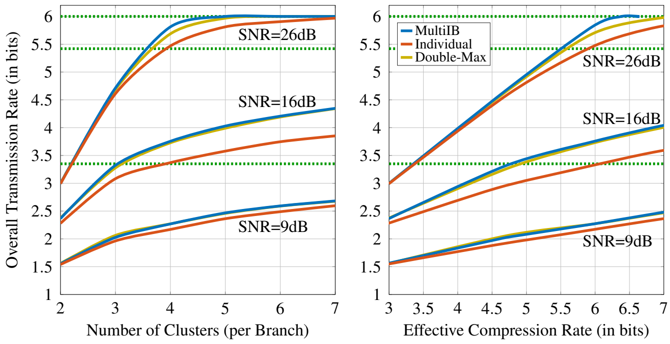
Interesting is the fact that the derived optimal solution in (14) with the corresponding MRD presented in (15) is also valid for the case in which a more stringent constraint set is demanded, i.e., maximizing the overall transmission rate, $I(\mathsf{x};\mathbf{z})$, subject to a set of constraints in a form of an upper-bound on each compression rate of different branches individually

$$Q^* = \underset{Q:\, \forall j\ I(\mathsf{y}_j;\mathsf{z}_j) \leq R_j}{\operatorname{argmax}} \quad I(\mathsf{x};\mathbf{z})\ . \tag{23}$$

To realize this, one shall recall that the stated optimal solution per branch is obtained by taking the functional derivative of the MIB functional given in (11) w.r.t. each of the mappings $p(\mathsf{z}_j|\mathsf{y}_j)$, when fixing others. Multiplying (11) by $-\lambda = -\frac{1}{\beta}$, the $\mathcal{L}_{\text{MIB}}^{\text{Dist.}}$ can be reformulated as $\mathcal{L}_{\text{MIB}}^{\text{Dist.},(1)} = I(\mathsf{x};\mathbf{z}) - \lambda \sum_j I(\mathsf{y}_j;\mathsf{z}_j)$, which gets simplified to $\mathcal{L}_{\text{MIB}}^{\text{Dist.},(2)} = I(\mathsf{x};\mathbf{z}) - \lambda I(\mathsf{y}_\ell;\mathsf{z}_\ell)$, when taking the required functional derivative w.r.t. $p(\mathsf{z}_\ell|\mathsf{y}_\ell)$ and fixing $p(\mathsf{z}_j|\mathsf{y}_j)$ for all $j \neq \ell$. The corresponding functional for (23) is obtained as $I(\mathsf{x};\mathbf{z}) - \sum_j \lambda_j I(\mathsf{y}_j;\mathsf{z}_j)$, wherein each λ_j is, indeed, the pertinent *Lagrange Multiplier* for the jth constraint. Taking its functional derivative w.r.t. $p(\mathsf{z}_\ell|\mathsf{y}_\ell)$ is then basically the same as taking the relevant functional derivative from the simplified version of $\mathcal{L}_{\text{MIB}}^{\text{Dist.},(2)}$ up to the trivial substitution of $\lambda \leftarrow \lambda_\ell$, that is already provided in (14)–(15). All in all, it is inferred that one can directly address (23) by a straightforward extension of the original MultiIB wherein a vector-valued $\boldsymbol{\beta}$ input comprising all the individual β_j for $1 \leq j \leq J$ replaces the scalar-valued β input and, subsequently, the quantizer mapping for jth branch is updated w.r.t. its own β_j. The corresponding procedure to obtain the apposite vector $\boldsymbol{\beta}$ is then to alternate between all branches and to perform bisections over a fine grid per branch, till fulfillment of all individual constraints.

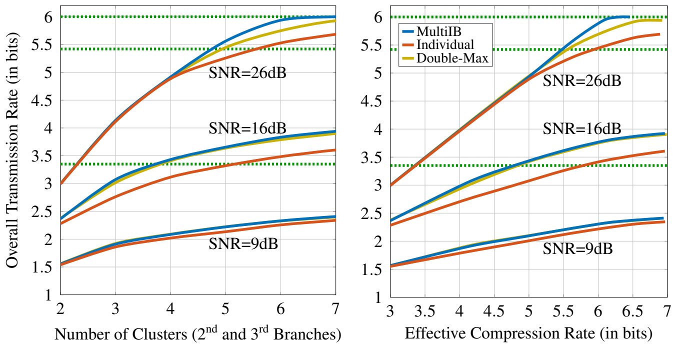
IV. Simulation Results

In this part, we investigate the performance of our proposed treatment (joint yet local quantization of multiple observations) and compare it with the approach of fully-independent local quantization per branch to verify achievable gains. Specifically, we consider an equiprobable standard 64-QAM ($\sigma_{\mathsf{x}}^2 = 42$) source signaling over a number of Additive White Gaussian Noise (AWGN) channels. To simulate this, we generate 1000 samples per branch. The outputs of these channels are then compressed (utilizing the MultiIB) such that the overall transmission rate, $I(\mathsf{x};\mathbf{z})$, is maximized (i.e., assuming $\beta \to \infty$). In case of the individual (and fully-independent) quantization per observation, the so-called *iterative Information Bottleneck* algorithm [8] is executed to compress each channel output such that it becomes highly informative w.r.t. the given source (maximize $I(\mathsf{x};\mathsf{z}_j)$ for $1 \leq j \leq J$). Further, since (as already mentioned) these routines are initialized randomly, for the sake of a fair comparison, we use the same starting points for both approaches and to avoid getting stuck into bad local optima, we repeat each method 100 times and retain the best outcome. We consider both symmetric and asymmetric setups with $J = 3$ branches. For the symmetric setup, all present model parameters are set to be the same for different branches while in case of the asymmetric arrangement, we fix the output cardinality of the first branch to $|\mathcal{Z}_1| = 2$ and then vary the output cardinalities of the other two branches that are set to be the same. Figs. 4 and 5 illustrate the obtained results, respectively. Explicitly, the overall transmission rates, $I(\mathsf{x};\mathbf{z})$, vs. the number of output levels (per branch) are depicted on the left while the resultant compression/informativity trade-offs are represented on the right. Principally, a quite similar behavior is observed in both configurations except for the restricting effects of the first branch (in asymmetric case) that necessitates higher output levels (compared to the symmetric configuration) for other branches to attain the same overall transmission rate. For relatively high values of the *Signal-to-Noise Ratio (SNR)* ($\frac{\sigma_{\mathsf{x}}^2}{\sigma_{\mathsf{n}}^2}$ per branch), the available mutual information, i.e., $I(\mathsf{x};\mathbf{y})$,



(a) $I(\mathrm{x};\mathbf{z})$ vs. $|\mathcal{Z}_j|$ (b) $I(\mathrm{x};\mathbf{z})$ vs. $\sum_j I(\mathrm{y}_j;\mathrm{z}_j)$

Fig. 4. Overall transmission rate vs. a) number of clusters, b) effective compression rate, 64-QAM signaling ($\sigma_{\mathrm{x}}^2=42$), symmetric setup with $J=3$ AWGN channels, (···) available mutual information $I(\mathrm{x};\mathbf{y})$, $\varepsilon=10^{-15}$



(a) $I(\mathrm{x};\mathbf{z})$ vs. $|\mathcal{Z}_j|$ (b) $I(\mathrm{x};\mathbf{z})$ vs. $\sum_j I(\mathrm{y}_j;\mathrm{z}_j)$

Fig. 5. Overall transmission rate vs. a) number of clusters, b) effective compression rate, 64-QAM signaling ($\sigma_{\mathrm{x}}^2=42$), asymmetric setup with $J=3$ AWGN channels, $|\mathcal{Z}_1|=2$, (···) available mutual information $I(\mathrm{x};\mathbf{y})$, $\varepsilon=10^{-15}$

reaches its upper-bound given by the source entropy [15], $H(\mathrm{x})$, being equal to 6 bits for equiprobable signaling. This amount of information can be entirely supported by relatively small output level cardinalities when applying the MultiIB, whereas the fully individual treatment of different branches requires larger numbers of output levels. By decreasing the SNR, the available mutual information decreases as well and, besides, it is ostensible that for its full support, regardless of the chosen approach, the required amounts for output cardinalities become substantially larger. One may note that irrespective of specific choices of model parameters, i.e., the noise variance and the number of output levels, our proposed approach outperforms the non-cooperative method. This, substantiates the fact that exploiting cooperation among different branches brings about some performance gain at the expense of requiring a more complex treatment compared to the non-cooperative approach. Finally, the performance comparison of our proposed method with the Double-Max [1], a SotA routine which formulates the quantizer design problem (per branch) as a double (alternating) maximization (hence the name), reveals better or at least the same results. Moreover, it should be noted that the setup in [1] solely considers the extreme instance of full informativity, meaning to have an asymptotically large trade-off parameter, β. Thus, also in this sense, our approach expands the horizon of problem through enabling a complete sweep over the entire range of valid β values.

V. Summary

We considered a certain distributed quantization setup which frequently appears in multiple applications. For that, we were the first to successfully apply the wide-ranging design framework of the *Multivariate Information Bottleneck* and to envision the potentialities of such a generic conceptual paradigm to cover a rich family of applications in communications context. This, particularly, enables addressing various extensions of presumed distributed arrangement including the simultaneous construction of intertwined compress models of multiple correlated sources.

Acknowledgment

This was partly funded by German ministry of education and research (BMBF) under grant 16KIS0720 (TACNET 4.0).

References

[1] S. Movaghati, M. Ardakani, “Distributed Channel-Aware Quantization Based on Maximum Mutual Information,” *International Journal of Distributed Sensor Networks*, vol. 12, no. 5, May 2016.
[2] G. C. Zeitler, “Low-Precision Quantizer Design for Communication Problems,” Ph.D. dissertation, TU Munich, Germany, 2012.
[3] S.-H. Park, O. Simeone, O. Sahin, S. S. Shitz, “Fronthaul Compression for Cloud Radio Access Networks: Signal Processing Advances Inspired by Network Information Theory,” *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 69–79, Nov. 2014.
[4] J. A. Gubner, “Distributed Estimation and Quantization,” *IEEE Trans. on Information Theory*, vol. 39, no. 4, pp. 1456–1459, July 1993.
[5] M. Longo, T. Lookabaugh, R. Gray, “Quantization for Decentralized Hypothesis Testing under Communication Constraints,” *IEEE Trans. on Information Theory*, vol. 36, no. 2, pp. 241–255, Mar. 1990.
[6] W.-M. Lam and A. R. Reibman, “Design of Quantizers for Decentralized Estimation Systems,” *IEEE Transactions on Communications*, vol. 41, no. 11, pp. 1602–1605, Nov. 1993.
[7] N. Slonim, N. Friedman, N. Tishby, “Multivariate Information Bottleneck,” *Neural Computation*, vol. 18, no. 8, pp. 1739–1789, Aug. 2006.
[8] N. Tishby, F. Pereira, W. Bialek, “The Information Bottleneck Method,” in *37th Annual Allerton Conference on Communication, Control, and Computing*, pp. 368–377, Monticello, IL, USA, Sep. 1999.
[9] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
[10] S. Hassanpour, D. Wübben, A. Dekorsy, and B. M. Kurkoski, “On the Relation Between the Asymptotic Performance of Different Algorithms for Information Bottleneck Framework,” in *IEEE Int. Conference on Communications (ICC)*, Paris, France, May 2017.
[11] S. Hassanpour, D. Wübben, and A. Dekorsy, “On the Equivalence of Double Maxima and KL-Means for Information Bottleneck-Based Source Coding,” in *IEEE Wireless Communications and Networking Conference (WCNC)*, Barcelona, Spain, Apr. 2018.
[12] —–, “A Graph-Based Message Passing Approach for Noisy Source Coding via Information Bottleneck Principle,” in *IEEE Global Communications Conference (GLOBECOM)*, Abu Dhabi, UAE, Dec. 2018.
[13] I. Tal and A. Vardy, “How to Construct Polar Codes,” *IEEE Trans. on Information Theory*, vol. 59, no. 10, pp. 6562–6582, Oct. 2013.
[14] J. Lewandowsky and G. Bauch, “Information-Optimum LDPC Decoders Based on the Information Bottleneck Method,” *IEEE Access*, vol. 6, pp. 4054–4071, Feb. 2018.
[15] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, 2006.
[16] J. H. Mathews, K. D. Fink *et al.*, *Numerical Methods Using MATLAB*, Upper Saddle River, NJ: Pearson Prentice Hall, 2004.
[17] M. Studenỳ and J. Vejnarová, “The Multiinformation Function as a Tool for Measuring Stochastic Dependence,” in *Learning in Graphical Models*, pp. 261–297, Springer, Dordrecht, 1998.
[18] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Elsevier, 2014.
[19] P.-N. Tan, M. Steinbach, and V. Kumar, “Data Mining Cluster Analysis: Basic Concepts and Algorithms,” *Introduction to Data Mining*, 2013.
[20] T. Berger, Z. Zhang, and H. Viswanathan, “The CEO Problem,” *IEEE Trans. on Information Theory*, vol. 42, no. 3, pp. 887–902, May 1996.
[21] R. Horst, P. M. Pardalos, and N. Van Thoai, *Introduction to Global Optimization*, 2nd ed, Springer Science & Business Media, 2000.