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: {hasanpour, 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.e., 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.
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 \(p(a_1,a_2)\) is given and, further, \(a_1 \leftrightarrow a_2 \leftrightarrow z\) 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, \(I(a_2;z)\) 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 \(I(a_2;z)\) to maximum number of bits that can be reliably transmitted over quantizer block, exploiting the Asymptotic Equipartition Property [15]. On the other hand, \(I(a_1;z)\) is chosen as the indicator of information preservation.

The quantizer design problem is then mathematically stated as finding the mapping \(p(z|a_2)\) that minimizes the IB functional \(L_{IB} = I(a_2;z) - \beta I(a_1;z)\), in which \(\beta\) 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 \((a_2, z) \in A_2 \times Z\) as

\[
p(z|a_2) = \frac{p(z)}{ψ(α_2, β)} \exp(-βd(z, a_2)),
\]

where \(ψ(α_2, β)\) is a normalization function assuring a valid distribution and the Relevant Distortion, \(d(z, a_2)\), is given as

\[
d(z, a_2) = D_{KL}(p(a_1|z)\|p(a_1|z)),
\]

with \(D_{KL}(/\|/)\) 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, \(z_j: 1 ≤ j ≤ J\), each quantizing a certain subset, \(y_j\), of the set of input RVs, \(a = \{a_i\}_{i=1}^I\), while preserving information about another arbitrary subset, \(x_j\), of elements in \(a\). 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

\[
\mathcal{I}(p(a)) = \sum a p(a) \log \frac{p(a)}{\prod_i p(a_i)},
\]

which captures the average amount of bits that can be secured by the joint vs. independent compression of elements in \(a\).

The Bayesian Network, \(\mathcal{G}\), is a powerful tool to describe the statistical relations among the RVs in \(a\). It is a directed acyclic graph that considers the entries of \(a\) as its nodes and encodes the proper factorization of \(p(a)\) with its edges in a sense that it applies \(p(a) = \prod_i p(a_i | p^G(a_i))\), with \(p^G(a_i\|\) 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 \(p^G_i\). Note that for an arbitrary distribution \(q(a)\) 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 \(q(a)\) w.r.t. the BN, \(\mathcal{G}\), [7] as

\[
\mathcal{T}^G(q(a)) = \sum I(a_i, p^G_i),
\]

where every local mutual information term is calculated using the marginal distributions of \(q(a)\). In general, \(\mathcal{T}(q(a)) > \mathcal{T}^G(q(a))\) and their gap measures how close is \(q(a)\) 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, \(\mathcal{G}_m\) and \(\mathcal{G}_o\), to establish the MIB variational principle as a trade-off between the multi-information term each network carries. The structure of input BN, \(\mathcal{G}_m\), determines the Solution Space and also signifies "what quantizes what". Basically, the statistical relations among input RVs get projected in the construction of \(\mathcal{G}_m\). Furthermore, the compression variables, \(z_j: 1 ≤ j ≤ J\), appearing as the leaves in \(\mathcal{G}_m\), are set to be the children of \(y_j\), i.e., the RVs they have to represent compactly. Hence, via BN conventions (see, e.g., [18]), given its parents, \(y_j\), each compression variable, \(z_j\), is assumed to be independent of other nodes. The multi-information, \(\mathcal{T}^{G_m}\), 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 \(I(z_j;y_j)\) for all the involved quantizers. The output BN, \(\mathcal{G}_o\), 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., \(x_j\). By doing so, the multi-information, \(\mathcal{T}^{G_o}\), becomes a natural gauge regarding the informativity of outcome as it sums up all the relevant mutual information terms \(I(z_j;x_j)\).

Analogous to the original IB setup, the trade-off between both aspects can then be formalized as minimizing the MIB functional, \(L_{MIB} = \mathcal{T}^{G_o} - \beta \mathcal{T}^{G_m}\), with \(\beta\) playing the same role as before. This minimization is carried out over the complete set of mappings \(\{p(z_j|y_j)\}_{j}\) from subsets of \(a\) entries which are intended to be quantized, i.e., \(y_j\), 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, \(\mathcal{G}_m\), stipulates that the overall joint distribution shall be factorable as \(p(a_1, a_2, z) = p(a_1)p(a_2|a_1)p(z|a_2)\). This indicates the presumed Markov chain \(a_1 \leftrightarrow a_2 \leftrightarrow z\) in the original IB setup. Further, noting both BNs, it is realized that \(z\) may be a compressed representation of \(y = a_2\) such that it is informative w.r.t. \(a_1\). Applying (4), it holds \(\mathcal{T}^{G_o} = I(a_2) + I(a_2;z)\) and \(\mathcal{T}^{G_m} = I(a_1; z)\). Since \(I(a_2; a_1)\) is a fixed term (joint distribution of all input RVs are assumed to be fixed), it can be dropped and then the MIB functional, \(L_{MIB}\), equals the IB functional, \(L_{IB}\).
C. Optimal Solution & an Iterative Design Algorithm

For a given trade-off parameter, $\beta$, and input statistics, $p(a)$, a formal optimal solution (yielding a stationary point of $L_{\text{MB}}$) regarding any of the present mappings $p(z_j | y_j)$ for $1 \leq j \leq J$ between the compression variable, $z_j$, and its parents in $G_{\text{in}}$ denoted by $\mathbf{y}_j = \mathbf{P}^0_{z_j}$ is derived for each $(z_j, y_j) \in Z_J \times Y_J$ as [7]

$$p(z_j | y_j) = \frac{p(z_j)}{\psi_{z_j}(y_j, \beta)} \exp(-\beta d(z_j, y_j)).$$ (5)

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

$$d(z_j, y_j) = \sum_{a \in P_{z_j}} \mathbb{E}_{p(y_j)} \left\{ D_{\text{KL}}(p(a | v_{z_j}^a, y_j) || p(a | v_{z_j}^a, z_j)) \right\}
+ \sum_{\ell \in z_j} \mathbb{E}_{p(y_j)} \left\{ D_{\text{KL}}(p(z_j | v_{z_j}^\ell, y_j) || p(z_j | v_{z_j}^\ell, z_j)) \right\}
+ D_{\text{KL}}(p(v_{z_j} | y_j) || p(v_{z_j} | z_j)),$$ (6)

with $v_{z_j}^a = \mathbf{P}^f_{z_j}$, $v_{z_j}^\ell = \mathbf{P}^f_{z_j}$, denoting sets of parent nodes of $a_i$ and $z_\ell$ in $G_{\text{out}}$, meaning the RVs that have to be informative about $a_i$ and $z_\ell$, respectively, and $v_{z_j}^a = v_{z_j}^a \setminus \{z_j\}$, $v_{z_j}^\ell = v_{z_j}^\ell \setminus \{z_j\}$.

Moreover, by definition

$$\mathbb{E}_{p(y_j)} \left\{ D_{\text{KL}}(p(b | r, z_j) || p(b | r, z_j)) \right\}
= \sum_r p(r | y_j) D_{\text{KL}}(p(b | r, y_j) || p(b | r, z_j)).$$ (7)

where $b$ and $r$ 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, $z_j$, 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 $y_j$, the lower the value of $d(z_j, y_j)$, the higher the probability of assigning $y_j$ to the cluster $z_j \in Z_J$. Principally, the better $z_j$ represents $y_j$, the lower becomes the respective KL divergences in (6) and, consequently, the larger gets the probability of allotting $y_j$ 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 MB function, $L_{\text{MB}}$, w.r.t. the set of all involved mappings, $\{p(z_j | y_j)\}$, for a particular input statistics, $p(a)$, 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, $\{|Z_j|\}$, the authors in [7] have adapted the Partitional and Hierarchical Clustering concepts [19] to the MB 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, $\{p(z_j | y_j)\}$, by direct use of (5). Note, that (5) has an implicit form as $p(z_j)$ and $d(z_j, y_j)$ on its right hand side depend on $\{p(z_j | y_j)\}$. The principal idea behind the MultiIB is then to commence with a random (still valid) initialization of the mappings, $\{p^{(0)}(z_j | y_j)\}$, and perform the update steps (till convergence/fulfillment of a stopping criterion) for every pair $(y_j, z_j) \in Y_J \times Z_J$ via

$$p^{(m+1)}(z_j | y_j) = \frac{p^{(m)}(z_j)}{\psi_{z_j}^{(m+1)}(y_j, \beta)} \exp(-\beta d^{(m)}(z_j, y_j)),$$ (8)

wherein $m$ denotes the running index. The quantizer output probability, $p^{(m)}(z_j)$, and the respective MRD, $d^{(m)}(z_j, y_j)$, are calculated employing $\{p^{(m)}(z_j | y_j)\}$ and the conditional independencies imposed by the structure of $G_{\text{in}}$. 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 $\ell \neq j$, $p^{(m+1)}(z_j | y_j) = p^{(m)}(z_j | y_j)$. 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 $D_{\text{JS}}(\cdot || \cdot)$ stated in the termination criterion part denotes the Jensen-Shannon (JS) divergence [7].

### Alg. 1 Multivariate iterative IB (MultiIB)

**Input:** $p(a)$, $G_{\text{in}}$, $G_{\text{out}}$, $\beta$, $|Z_J|$, convergence parameter $\varepsilon > 0$

**Output:** Generally soft partition $z_j$ of $Y_J$ into $|Z_j|$ bins $\forall j = 1 : J$

**Initialization:** $m = 0$, random mappings $\{p^{(m)}(z_j | y_j)\}$

while True do

for $j = 1 : J$ do

- $p^{(m)}(z_j) \leftarrow \sum_{y_j} p^{(m)}(z_j | y_j) p(y_j), \forall y_j \in Z_J$
- find the nth update for all distributions in $d(z_j, y_j)$ via marginalizing w.r.t. $p^{(m)}(a, z_j) = p(a) \prod_{j=1}^{J} p^{(m)}(z_j | y_j)$

- $p^{(m+1)}(z_j | y_j) \leftarrow \frac{p^{(m)}(z_j)}{\psi_{z_j}^{(m+1)}(y_j, \beta)} \exp(-\beta d^{(m)}(z_j, y_j))$

- $p^{(m+1)}(z_j | y_j) \leftarrow p^{(m)}(z_j | y_j), \forall \ell = 1 : J, \ell \neq j$

- $m \leftarrow m + 1$

end for

if $\forall j, \forall y_j : D_{\text{JS}}(\frac{1}{2} || \frac{1}{2}) (|p^{(m)}(z_j | y_j)||p^{(m-j)}(z_j | y_j)) \leq \varepsilon$ then

Break

end if

end while

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.
quantization of individual noisy observations utilizing the MIB framework which leads to a joint yet local unit for further processing. To perform this task, we propose compressed into the variables, channels’ output variables, indeed, will be upper-bounded by solely on the preservation of relevant information, the effective sign. Please note that even in this case, although the focus is addressing (12) for a given input statistics, \( p(x, y, z) \), a trade-off parameter, \( \beta \), and a set of quantizers’ output levels, \( \{ |Z| \} \), one has to start with a group of random (yet valid) mappings, \( \{ p(0)(z_j | y_j) \} \), and then for every particular branch, \( j \), the respective quantizer mapping update is

\[
p^{(m+1)}(z_j | y_j) = \frac{p^{(m)}(z_j)}{p^{(m+1)}(y_j | z_j, \beta)} \exp \left( -\beta d(z_j, y_j) \right),
\]

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

\[
d(z_j, y_j) = \mathbb{E}_{p(\cdot | y_j)} \left\{ D_{\text{KL}} \left( p(x | v_x^{-j}, y_j) \| p(x | v_x^{-j}, z_j) \right) \right\} = \sum_{v_x^{-j}} p(v_x^{-j} | y_j) D_{\text{KL}} \left( p(x | v_x^{-j}, y_j) \| p(x | v_x^{-j}, z_j) \right)
\]

where \( v_x^{-j} = \{ z_1, \cdots, z_{j-1}, z_{j+1}, \cdots, z_J \} \). 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, \( p(x, y) \), the design optimization problem is then formulated as

\[
p^{(m+1)}(z_j | y_j) = \frac{p^{(m)}(z_j)}{p^{(m+1)}(y_j | z_j, \beta)} \exp \left( -\beta d^{(m)}(z_j, y_j) \right),
\]

where

\[
\mathbb{E}_{p(\cdot | y_j)} \left\{ D_{\text{KL}} \left( p(x | v_x^{-j}, y_j) \| p(x | v_x^{-j}, z_j) \right) \right\} = \sum_{v_x^{-j}} p(v_x^{-j} | y_j) D_{\text{KL}} \left( p(x | v_x^{-j}, y_j) \| p(x | v_x^{-j}, z_j) \right)
\]

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

\[
Q^* = \arg \max_{Q} \mathbb{E}_{p(\cdot | y_j)} \left\{ D_{\text{KL}} \left( p(x | v_x^{-j}, y_j) \| p(x | v_x^{-j}, z_j) \right) \right\}
\]

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

\[
Q^* = \arg \min_{Q} I(x; z_1, \cdots, z_J),
\]

despite 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 \( \sum_j \log_2 |Z_j| \) bits. For each pair \((y_j, z_j) \in Y_j \times Z_j\), the optimal solution regarding the present quantizers in (12) is then given as

\[
p(z_j | y_j) = \frac{p(z_j)}{\psi_j(y_j, \beta)} \exp \left( -\beta d(z_j, y_j) \right),
\]

where

\[
\mathbb{E}_{p(\cdot | y_j)} \left\{ D_{\text{KL}} \left( p(x | v_x^{-j}, y_j) \| p(x | v_x^{-j}, z_j) \right) \right\} = \sum_{v_x^{-j}} p(v_x^{-j} | y_j) D_{\text{KL}} \left( p(x | v_x^{-j}, y_j) \| p(x | v_x^{-j}, z_j) \right)
\]

where \( v_x^{-j} = \{ z_1, \cdots, z_{j-1}, z_{j+1}, \cdots, z_J \} \). 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, \( p(x, y) \), a trade-off parameter, \( \beta \), and a set of quantizers’ output levels, \( \{ |Z| \} \), one has to start with a group of random (yet valid) mappings, \( \{ p(0)(z_j | y_j) \} \), and then for every particular branch, \( j \), the respective quantizer mapping update is

\[
p^{(m+1)}(z_j | y_j) = \frac{p^{(m)}(z_j)}{p^{(m+1)}(y_j | z_j, \beta)} \exp \left( -\beta d^{(m)}(z_j, y_j) \right),
\]

where

\[
\mathbb{E}_{p(\cdot | y_j)} \left\{ D_{\text{KL}} \left( p(x | v_x^{-j}, y_j) \| p(x | v_x^{-j}, z_j) \right) \right\} = \sum_{v_x^{-j}} p(v_x^{-j} | y_j) D_{\text{KL}} \left( p(x | v_x^{-j}, y_j) \| p(x | v_x^{-j}, z_j) \right)
\]

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

\[
Q^* = \arg \min_{Q} I(x; z_1, \cdots, z_J),
\]

despite 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 \( \sum_j \log_2 |Z_j| \) bits. For each pair \((y_j, z_j) \in Y_j \times Z_j\), the optimal solution regarding the present quantizers in (12) is then given as

\[
p(z_j | y_j) = \frac{p(z_j)}{\psi_j(y_j, \beta)} \exp \left( -\beta d(z_j, y_j) \right),
\]
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, \( z_t \), for \( 1 \leq t < j - 1 \), 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, \( \beta \), this yields a set of stochastic mappings while for case of letting \( \beta \to \infty \), the normalization function, \( \psi_{z_j}(y_j, \beta) \), 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 \( p(z_j | y_j) \) for \( 1 \leq j \leq J \). This is due to the fact that, for a given \( p(x) \), \( I(x; z) \) becomes convex w.r.t. \( p(z|x) = \prod_j p(z_j | x) \) [15] and as fixing the quantizer mappings \( p(z_j | y_j) \) for all \( j \neq \ell \) directly corresponds to fixing the pertinent distributions \( p(z_j | x) \), the relation among \( p(z|x) \) and \( p(z_j | x) \) becomes affine which preserves convexity. As it holds \( p(z|x) = \sum_{y_i} p(z_j | y_j) p(y_i | x) \) that once again defines an affine relation between \( p(z_j | x) \) and \( p(z_j | y_j) \), 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, \( S \), 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 \( S \) (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, \( I(x; z) \), 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

\[
Q^* = \arg \max_{Q; \sum_j I(y_j; z_j) \leq R} I(x; z), \tag{22}
\]

wherein each certain \( R \) value corresponds to a certain \( \beta \) value. Attaining the required \( \beta \) for a particular \( R \) is then usually done by (repeatedly) performing the bisection method over a proper initial interval of \( \beta \) 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, \( I(x; z) \), subject to a set of constraints in a form of an upper-bound on each compression rate of different branches individually

\[
Q^* = \arg \max_{Q; \forall j I(y_j; z_j) \leq R_j} I(x; z). \tag{23}
\]

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 \( p(z_j | y_j) \), when fixing others. Multiplying (11) by \( -\lambda = -\frac{1}{\beta} \), the \( L_{\text{MB}} \) can be reformulated as \( L_{\text{MB}}^{(1)} = I(x; z) - \lambda \sum_j I(y_j; z_j) \), which gets simplified to \( L_{\text{MB}}^{(2)} = I(x; z) - \lambda I(y_j; z_j) \), when taking the required functional derivative w.r.t. \( p(z_j | y_j) \) and fixing \( p(z_j | y_j) \) for all \( j \neq \ell \). The corresponding functional for (23) is obtained as \( I(x; z) - \sum_j \lambda_j I(y_j; z_j) \), wherein each \( \lambda_j \) is, indeed, the pertinent Lagrange Multiplier for the \( j \)th constraint. Taking its functional derivative w.r.t. \( p(z_j | y_j) \) is then basically the same as taking the relevant functional derivative from the simplified version of \( L_{\text{MB}}^{(2)} \) up to the trivial substitution of \( \lambda \leftrightarrow \lambda_j \), 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 \( \beta \) input comprising all the individual \( \beta_j \) for \( 1 \leq j \leq J \) replaces the scalar-valued \( \beta \) input and, subsequently, the quantizer mapping for \( j \)th branch is updated w.r.t. its own \( \beta_j \). The corresponding procedure to obtain the adaptive vector \( \beta \) is then to alternate between all branches and 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 (\( \sigma^2 = 12 \)) 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, \( I(x; z) \), 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 \( I(x; z_j) \) 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 \( |Z_1| = 2 \) 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, \( I(x; z) \), 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) (\( \frac{P}{\sigma^2} \) per branch), the available mutual information, i.e., \( I(x; y) \),
reaches its upper-bound given by the source entropy \([15, H(x)]\), 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, \(\beta\). Thus, also in this sense, our approach expands the horizon of problem through enabling a complete sweep over the entire range of valid \(\beta\) 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