Algorithmic Solutions for Distributed Compression using the Information Bottleneck Principle

Steffen Steiner, Volker Kuehn, *Member, IEEE*

Abstract

This paper addresses the optimization of distributed compression in a sensor network. As a direct communication among the sensors is not possible, noisy measurements have to be locally compressed in order to meet the rate constraints of the communication links to a common receiver. This scenario is widely known as the Chief Execution (Estimation) Officer (CEO) problem. The corresponding literature mainly focuses on the characterization of the rate region for a specific target distortion, i.e. the set of feasible rates meeting the distortion constraint is determined. However, none of these solutions consider the converse problem allowing to minimize the distortion for given rate constraints. In this contribution, different algorithmic solutions for the converse problem are proposed minimizing the overall distortion while fulfilling the rate-constraints of individual links. In particular, two approaches for the joint quantizer design based on the information bottleneck (IB) method are derived. Following the Wyner-Ziv coding principle, the quantizers are successively designed using the statistics of other quantizers as side-information. The performance of the obtained quantizers is compared with other distributed approaches taken from literature and the vector quantizer. Moreover, the influences of different aspects like the optimization order and the asymmetry of scenarios are discussed.

Index Terms

CEO problem, information bottleneck, distributed compression, rate distortion theory.

S. Steiner and V. Kuehn are with the Faculty of Computer Science and Electrical Engineering, University of Rostock, Germany, e-mail: {steffen.steiner,volker.kuehn}@uni-rostock.de

I. INTRODUCTION

Distributed sensing plays an important role in many areas. Smart environments, cities or homes shall improve safety and comfort. Environmental monitoring systems, smart manufacturing (Industry 4.0) or systems for driving assistance in cars employ many sensors and fuse the measurements to infer relevant information. A current example in mobile radio communications are cloud-based radio-access networks where access points forward their received signals to a central processor in the cloud [1]. In all these scenarios, spatially separated sensors collect noisy measurements and forward them to a common control center over capacity-limited communications links. Therefore, sensing, communication, and signal processing have to be jointly optimized to efficiently infer the desired information.

This contribution considers a generic distributed sensing system. Multiple sensors measure the same process and communicate with a common receiver over capacity-limited links. A direct communication among the senors is not possible. The measurements are noisy which is widely known as remote sensing or noisy source coding problem as opposed to clean sources [2]–[5]. For reliable transmissions from the sensors to the receiver, the information rate of each sensor must be smaller than the corresponding link capacity. This is achieved by appropriate local compression at the sensors using Wyner-Ziv coding. The primary goal is to minimize an appropriately defined overall distortion while fulfilling the individual rate constraints. The following paragraphs discuss the state-of-the art of the theoretical background.

Distributed Compression and the CEO Problem

In information theory, distributed source coding has been of interest for decades and significant progress has been achieved in the past. The problem has been formulated in various facets and different assumptions on the distribution of variables and the distortion measure have been made. A survey on distributed source coding can be found in [6]. It discusses exemplary sensor networks and the separation of source and channel coding as well as digital versus analog sensing and transmission. The authors draw the conclusion that digital processing works fine for rich communication between sensors while it requires exponentially more sensors in order to achieve the same distortion as analog processing for limited communication between nodes.

Multi-terminal source coding generally considers M correlated sources and N decoders being interested in a subset of only $m \quad M$ source signals. In [7], an inner bound on the achievable rate region has been derived and [8] provides a solution for two encoders. Oohama developed

upper and lower bounds for the multi-terminal source coding problem with correlated source signals in [9], [10]. To the authors' knowledge, the rate region of the multi-terminal source coding problem is still unknown for more than two encoders.

The Chief Execution (Estimation) Officer (CEO) problem considers a single source being remotely sensed by distributed sensors. Each sensor compresses a noisy observation of the relevant process and forwards it to the CEO over capacity limited links. Many results exist for the quadratic Gaussian CEO problem considering Gaussian distributed signals and the mean squared error (MSE) distortion [11]–[14]. In [11], the sum-rate distortion function has been analytically derived for the quadratic Gaussian CEO problem and its complete rate region is characterized in [14]. Courtade and Weissman characterized the CEO rate region for arbitrary discrete source distributions and a logarithmic loss function [8]. Moreover, asymptotic analyses for an infinite number of sensors have been performed in [13], [15], [16]. Berger et al. investigated the error rate performance for a discrete source with the Hamming distance as a distortion measure and disclosed an inevitable loss due to non-cooperating sensors [15]. For the Gaussian case, the MSE distortion was shown to decrease asymptotically with the reciprocal sum-rate R for non-cooperating encoding while it decays exponentially (2^{-2R}) for cooperating sensors. A scaling law on the sum-rate distortion function for arbitrary distortion measures has been derived in [16]. The authors state that the distortion scales with $1/M$ for analog transmission and only with $1/\log(M)$ for digital transmission with M being the number of sensors. The CEO problem for a vector of relevant processes under logarithmic loss distortion measure has been studied in [17], [18]. For more details about outer bounds and rate regions the reader may refer to [19]–[21].

Despite of the large recent progress, rate regions have been characterized only for particular distributions of the relevant signal and distortion measures. Furthermore, the particular numerical computation of the rate region quickly becomes challenging due to the exponentially growing complexity. Most results are presented for two sensors or do focus on the sum-rate constraint only [22], [23] and not on the complete set of rate constraints defining the rate region.

Information Bottleneck

Independently of the distributed source coding problem, Tishby et al. introduced the information bottleneck (IB) framework as an information theoretic approach to optimize clustering (quantization) [24], [25]. In principle, a compromise between the compression rate and a relevant mutual information is targeted. This trade-off can be controlled by a Lagrangian optimization

approach leading to a non-convex optimization problem. Several IB algorithms exist [24]–[27] and the mathematical theory has been intensively studied in [28], [29]. Meanwhile, a rich set of IB applications can be found in communications. Analog-to-digital conversion for channels with memory has been investigated in [30]. The decoding of low density parity check (LDPC) codes using IB-optimized lookup tables was analyzed in [31]–[33]. The IB framework has been extended to distributed clustering of a common relevant information as well. In [34], cooperative Quantize-and-Forward relaying schemes have been optimized using the IB method. An alternating algorithm based on the IB framework was introduced in [35], [36] to deal with individual rate constraints in cloud-based radio access networks. Moreover, distributed sensor networks with imperfect forward links have been optimized in [37], [38]. Finally, [39] considered a multivariate IB variant to trade-off relevant information and compression sum-rate.

Although initiated in different areas, a tight connection between the CEO problem and the IB framework exists. For the logarithmic loss function as a distortion measure, the CEO problem can be formulated as a distributed IB approach. We exploit this connection to derive two new algorithmic solutions for the joint optimization of scalar quantizers in a distributed setup allowing to minimize the distortion while fulfilling individual rate constraints.

Structure and notation:

A brief overview on rate distortion theory and the IB method is provided in Section II. Afterwards, Section III introduces the setup for distributed compression including the definitions of outer and inner bounds on the rate region as well as the formulation of the vector quantization problem under log-loss distortion. Section IV presents different algorithmic solutions of the CEO optimization problem. First, two new approaches allowing individual rate adjustments are derived. Second, an approach based on the sum-rate constraint from [22] is reviewed. Section V presents numerical results and Section VI concludes this paper.

The following notation is used. Random variables are denoted by calligraphic letters $X; Y; Z$, their realizations $x; y; z$ are elements of the sets $\mathsf{X}; \mathsf{Y}; \mathsf{Z}$ with cardinalities $|\mathsf{X}|$, $|\mathsf{Y}|$ and $|\mathsf{Z}|$, respectively. Vectors are denoted in bold letters, $\mathbf{y} = \begin{bmatrix} y_1 & \dots & y_M \end{bmatrix}^{\mathsf{T}}$, and multivariate random variables in boldface calligraphic letters $\boldsymbol{Y}; \boldsymbol{Z}$, with $\boldsymbol{Z}_{<m}$ covering only the processes Z_1 to Z_{m-1}. The terms $p(y|x)$, $p(x; y)$ and $I(X; Y)$ represent conditional and joint probability density function (pdf) (or probability mass function (pmf) for discrete random variables) and the mutual

information between X and Y, respectively. Finally, $\mathsf{E}_X[f(X)]$ represents the expectation of a function $f(x)$ with respect to x.

II. Rate Distortion Theory and Information Bottleneck

A. Rate Distortion Theory

The rate distortion theory goes back to the seminal work of Shannon [40], [41]. While the entropy $H(Y)$ of a random process Y represents the ultimate lower bound for lossless compression, a lossy compression of Y to Z leads inevitably to a distortion of Y. The distortion $d(y;z)$ can be measured by different means. Hamming distance, squared Euclidean distance or the logarithmic loss function are just some examples. The expectation of the squared Euclidean distance $d(y;z) = |z - y|^2$ results into the well-known MSE, and requires z to be a physical representative of its cluster. It is often used in combination with Gaussian distributions due to convenient mathematical properties. Contrarily, the logarithmic loss function $d(y;z) = -\log p(y|z)$ is a pure statistical measure and does not require z to be a physical cluster representative (see Sec. II-B). Obviously, the choice of the distortion measure influences the mathematical nature of the optimization problem, e.g. whether it is convex or not. The distortion rate function $D(R)$ defines the best trade-off between compression rate $I(Y;Z)$ and the average distortion $\mathsf{E}_{Y,Z}[d(Y;Z)]$ and is obtained by solving

$$D(R) = \min_{p(z|y):\ I(Y;Z) \le R} \mathsf{E}_{Y,Z}[d(Y;Z)]. \tag{1}$$

From (1) it can be seen that the compression or quantization is expressed by a generally stochastic mapping of y onto z denoted by the pmf $p(z|y)$. Using the Lagrangian approach, (1) can be rewritten to

$$\min_{p(z|y)} \mathsf{E}_{Y,Z}[d(Y;Z)] + \beta I(Y;Z) \tag{2}$$

and leads to the self-consistent solution [42]

$$p(z|y) = \frac{p(z)\ e^{-\beta\, d(y;z)}}{\sum_z p(z)\ e^{-\beta\, d(y;z)}}. \tag{3}$$

In (2) and (3), the Lagrange multiplier β serves as trade-off parameter between distortion and compression.[1]

[1]In contrast to the most publications, the Lagrange multiplier β is placed in front of the compression rate $I(\mathcal{Y};\mathcal{Z})$ instead of the distortion measure $\mathbb{E}_{Y,Z}[d(\mathcal{Y};\mathcal{Z})]$ in (2) leading to the ratio $-\frac{1}{\beta}$ in the update equation (3). This ensures a consistent notation when targeting distributed scenarios with multiple compression rates but only a single distortion term.

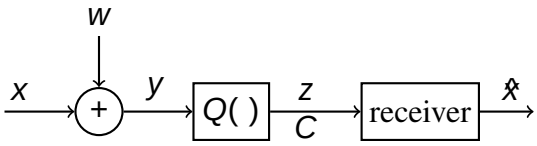

Fig. 1. Illustration of noisy source coding (remote sensing) for a single sensor

The system model for noisy source coding or remote sensing is depicted in Fig. 1. Here, the quantizer input y is a noisy observation of a relevant signal x, so that the distortion is measured between x and z. With

$$\min_{p(z|y)} \mathbb{E}_{X,Z}[d(X,Z)] + \quad I(Y;Z); \tag{4}$$

the solution becomes [4]

$$p(z|y) = \frac{p(z)\ e^{-\frac{1}{}\sum_x p(x|y)\ d(x,z)}}{\sum_z p(z)\ e^{-\frac{1}{}\sum_x p(x|y)\ d(x,z)}}: \tag{5}$$

A comparison of (5) with the classical rate distortion solution in (3) illuminates the similarity of both expressions. The main difference comes from the conditional expectation of the distance measure with respect to the relevant signal

$$d(y,z) = \sum_x p(x|y)\ \ d(x,z) = \mathbb{E}_{X|y}[d(X,z)]\ : \tag{6}$$

Hence, the algorithmic solution of (4) follows the same iterative approach.

B. Information Bottleneck Method

The information bottleneck method is a clustering framework pairing concepts from machine learning and information theory [25]. Following the general setup as depicted in Figure 1, the aim of the IB approach is to optimize the mapping $p(z|y)$ such that a maximal relevant information $I(X;Z)$ is preserved while the rate constraint $I(Y;Z) \quad C$ is fulfilled. According to [24], this problem can be formulated as a Lagrangian maximization problem

$$\max_{p(z|y)} I(X;Z) \qquad I(Y;Z) = \min_{p(z|y)} H(X|Z) + \quad I(Y;Z)\ : \tag{7}$$

The equality holds due to $I(X;Z) = H(X) \quad H(X|Z)$ and the fact that $H(X)$ can be skipped because it does not depend on the mapping $p(z|y)$. The comparison of (7) with (4) reveals that the IB approach is a special formulation of the rate distortion problem using the logarithmic loss function $d(x,z) = \quad \log p(x|z)$ as a distortion measure whose expectation is $H(X|Z) =$

$\mathsf{E}_{X,Z}[\ -\log p(x\,|\,z)]$. In this case, distortion minimization means maximization of the relevant mutual information $I(X;Z)$. Since this distortion is a function of the mapping $p(z\,|\,y)$, (7) is a non-convex optimization problem with the solution

$$p(z\,|\,y) = \frac{p(z)\ \ e^{\ -\frac{1}{\ }\ D_{\mathrm{KL}}[p(x|y)\|p(x|z)]}}{\sum_z p(z)\ \ e^{\ -\frac{1}{\ }\ D_{\mathrm{KL}}[p(x|y)\|p(x|z)]}} \tag{8}$$

where

$$D_{\mathrm{KL}}[p(x\,|\,y)\|p(x\,|\,z)] = \mathsf{E}_X\ \ \log\frac{p(x\,|\,y)}{p(x\,|\,z)}$$

denotes the Kullback-Leibler divergence. As this approach is solely based on probability distributions, $z \in Z$ is just a cluster index and needs not to be a physical representative of x or y. This allows the application of the IB method also in non-technical scenarios without signal representations.

Again, the Lagrange multiplier $\ $ in (7) serves as trade-off parameter between preservation of relevant information and compression. The choice $\ \to 0$ only focuses on the preservation of relevant information and yields a deterministic clustering $p(z|y) \in \{0,1\}$, which is very convenient from a practical perspective because it allows an implementation by a static lookup table [25]. In this case, a compression is obtained by choosing a compact representation $z \in Z$ with $|Z| < |Y|$. For $\ > 0$, the clustering $p(z|y) \in [0,1]$ is generally stochastic. For $\ \to 1$, no relevant information is preserved and all values of y are mapped onto a single cluster index z. The self-consistent solutions in (3), (5) and (8) can be efficiently solved with iterative Blahut-Arimoto like algorithms [43], [44]. As the compression-rate curve is a monotonic increasing function, a simple bi-section search allows to adjust $\ $ such that a desired rate constraint $I(Y;Z) \leq C$ is fulfilled. This technique will be applied in Sec. III to obtain solutions lying inside the CEO rate region. For a more detailed review of IB algorithms the reader may refer to [25].

III. Distributed Sensing System

A. System Model

Figure 2 illustrates the model of a distributed sensing system. It consists of M sensors forwarding their measurements to a common receiver over links with finite capacities C_m, $1 \leq m \leq M$. Each sensor m observes a noisy version $Y_m = X + W_m$ of the relevant process X. The additive measurement noise W_m is assumed to be white and Gaussian distributed. For zero-mean processes X and W_m, the measurement signal-to-noise-ratio (SNR) at sensor m is

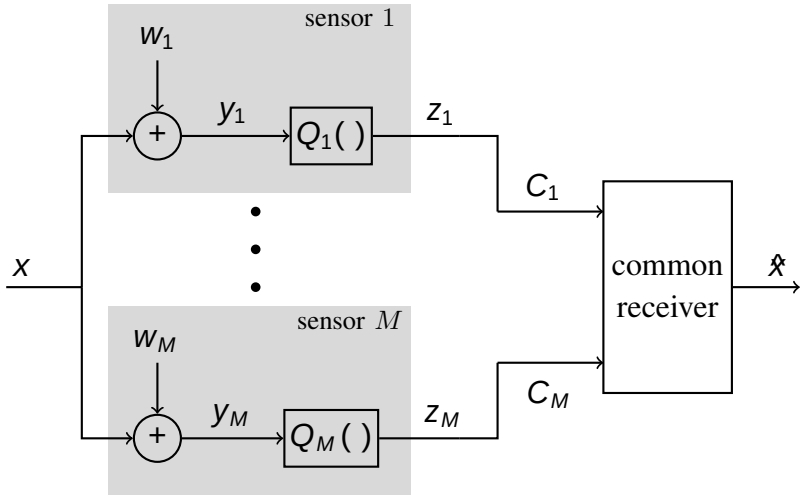

Fig. 2. Distributed sensor system with M sensors, a common receiver and individual link capacities C_m

given by $_m = \frac{^2_X}{^2_{W_m}}$, where ^2_X and $^2_{W_m}$ denote the powers of the relevant signal and the noise at sensor m, respectively. The noise processes W_m at the sensors are assumed to be statistically independent of each other yielding the likelihood function

$$p(\mathrm{y}jx) = \prod_{m=1}^{M} p(y_m jx): \tag{9}$$

In order to let the transmission rates meet the link capacities C_m of the forward links to the common receiver for all $1 \quad m \quad M$, each sensor m compresses its noisy observation y_m to z_m. This compression is described by a generally stochastic mapping $p(z_m jy_m)$.

In this distributed setup, it is assumed that the sensors are not allowed to exchange information. Instead, each sensor performs the compression locally without having observed the other sensor signals. However, the quantizers $Q_m(\,)$ can be jointly designed in order to optimize an overall fidelity criterion.

B. The Chief Execution Officer (CEO) Problem

Optimizing the distributed compression for a scenario depicted in Figure 2 is termed the Chief Execution (Estimation) Officer Problem [15]. In the literature, the two most popular distortion measures are the mean squared error and the logarithmic loss function. This paper focuses on the logarithmic loss function but uses the IB formulation with the relevant mutual information

$I(X;Z)$. According to (7), the problem formulations with $H(X|Z)$ and $I(X;Z)$ are equivalent. In this way, the inner bound of the CEO rate region with M sensors as defined in [45] becomes

$$I(Y_{\mathcal{S}}; Z_{\mathcal{S}} \,|\, Z_{\bar{\mathcal{S}}}) \quad C_{\mathcal{S}} = \sum_{m \in \mathcal{S}} C_m \quad \forall \quad \mathcal{S} \quad \{1,2,\ldots,M\} \tag{10}$$

$$I(X;Z) \quad D : \tag{11}$$

The corresponding outer bound is given by

$$\sum_{m \in \mathcal{S}} I(Y_m; Z_m \,|\, X) + H(X \,|\, Z_{\bar{\mathcal{S}}}) \quad D \Big)^{+} \quad C_{\mathcal{S}} = \sum_{m \in \mathcal{S}} C_m$$

$$\forall \quad \mathcal{S} \quad \{1,2,\ldots,M\} \tag{12}$$

$$I(X;Z) \quad D \tag{13}$$

with $[\,]^+ = \max(0, \,)$. Both bounds are defined for any distribution

$$p(x, \mathbf{y}, z) = p(x) \prod_{m=1}^{M} p(z_m|y_m) \; p(y_m|x)$$

and they match for the log-loss distortion [45]. Note that the rate constraints in (10) and (12) must hold for any subset of $\mathcal{S} \quad \{1,2,\ldots,M\}$ and its complementary subset $\bar{\mathcal{S}} = \{1,2,\ldots,M\} \setminus \mathcal{S}$. According to [46], the solution space forms a contra-polymatroid with $M!$ vertices making the computation of large systems very challenging.

C. Vector IB Quantization (VIB)

For benchmarking the distributed compression system, a vector quantization is considered as well. In contrast to the assumptions described in Subsection III-A, the vector quantizer is a single device having access to all observations $\mathbf{y} = [y_1 \ldots y_M]^{\mathsf{T}}$. Quantization is now performed on the entire vector $\mathbf{y}$ and delivers a single output z. In our particular setup, this hypothetical vector quantizer can be reduced to a simple scalar quantizer. Maximum ratio combining of all inputs y_m delivers a scalar sufficient statistics

$$y = \sum_{m=1}^{M} \quad _m \quad y_m$$

of the desired relevant signal x with an overall SNR $\quad = \sum_m \quad _m$. This scalar signal is then quantized to a cluster index z. In order to have a fair comparison with the distributed setup, the number of clusters is chosen to $|Z| = \prod_{m=1}^{M} |Z_m|$. Moreover, the forward link to the common

receiver has a channel capacity $C_{\text{sum}} = \sum_{m=1}^{M} C_m$ which equals the sum-rate in (10) for $\mathcal{S} = \{1,\dots,M\}$. Altogether, this results into the optimization problem

$$\max_{p(z|\mathbf{y})} I(X;Z) \quad I(Y;Z) = \max_{p(z|y)} I(X;Z) \quad I(Y;Z) \tag{14}$$

whose solution uses the update rule in (8) with the exponent

$$D_{\text{KL}}[p(x|y) \| p(x|z)] = \mathbb{E}_X \left[\log \frac{p(x|y)}{p(x|z)} \right] .$$

IV. Optimization Algorithms for Distributed Compression

This section presents different strategies to determine feasible solutions of the CEO problem. In contrast to many approaches which minimize the compression rates while keeping the average distortion below a threshold D, the converse problem is targeted. The optimization goal is to maximize the relevant mutual information $I(X;Z)$ (minimize the average distortion) given the rate constraints, i.e. the channel capacities of the links to the common receiver.

A. Distributed Information Bottleneck (DIB) Approach

From the rate constraints in (10) it becomes obvious that the set of sensors $\{1,\dots,M\}$ has to be split into all possible cut-sets $\mathcal{S}$ and its complements $\overline{\mathcal{S}}$. According to [23], [47], the CEO problem formulation based on the inner bound in (10) can be decomposed into a set of simpler problems. Inspired by this statement, we show in App. A for a particular but arbitrary subset $\mathcal{S}$ that the rate constraint $I(Y_{\mathcal{S}}; Z_{\mathcal{S}} | Z_{\overline{\mathcal{S}}}) \quad C_{\mathcal{S}}$ is fulfilled if the compression rates $I(Y_m; Z_m | Z_{<m})$ with partial side-information $Z_{<m}$ are upper bounded by the corresponding link capacities C_m for all $m \in \mathcal{S}$. Therefore, it is not necessary to check all $2^M - 1$ rate constraints for obtaining a feasible solution. Instead, it suffices to adjust the M Lagrange multipliers such that $I(Y_m; Z_m | Z_{<m})$ is upper bounded by the associated link capacity C_m for all $1 \quad m \quad M$.

While the authors in [23] proof for $M = 2$ sensors that the convex hull of the solutions for the simpler problems defines the CEO rate region, the extension to more than two sensors has not been proved. Due to $I(Y_m; Z_m | Z_{<m}) \quad I(Y_m; Z_m | Z_{\setminus m})$, using partial side-information $Z_{<m}$ to ensure $I(Y_m; Z_m | Z_{<m}) \quad C_m$ leads to stronger compression of Y_m to Z_m than the CEO rate constraints $I(Y_m; Z_m | Z_{\setminus m}) \quad C_m$ considering full side-information $Z_{\setminus m}$ require. Just the last sensor in the successive update procedure is restricted to the original rate constraint. Hence, less strict rate limitations might lead to a larger overall relevant mutual information $I(X;Z)$ while still fulfilling the CEO rate constraints. However, App. B shows that the consideration of full

side-information for all compression rates as used in the definition of the CEO rate region will lead to infeasible solutions violating at least some of the rate constraints. Consequently, not all rate constraints can be simultaneously fulfilled with equality. The main focus of this contribution is not the exact characterization of the rate region but the development of efficient algorithms determining feasible solutions for distributed compression problems.

As it is not known a priori which partial side-information shall be exploited by each sensor, the solutions have to be computed for all permutations $\pi(\cdot)$ of the set $\{1,\dots,M\}$. This increases the computational complexity exponentially with M. Each permutation represents a certain Wyner-Ziv coding strategy and a particular corner point of the rate region. Time-sharing between them delivers intermediate points in the rate region. With this argumentation, the optimization problem can be formulated as

$$\max_{\mathbf{P},\pi} I(X;Z) \text{ s.t. } I(Y_{\pi(m)};Z_{\pi(m)}|Z_{<\pi(m)}) \leq C_{\pi(m)} \quad \forall m \in \{1,2,\dots,M\} \tag{15}$$

where $\mathbf{P} = [p(z_1|y_1) \quad \dots \quad p(z_M|y_M)]$ denotes the set of all mappings and $\pi(m)$ represents an arbitrary order of the sensors. According to (15), the first sensor's compression rate $I(Y_{\pi(1)};Z_{\pi(1)})$ has no side-information and is upper bound by $C_{\pi(1)}$. The side-information increases for each further sensor until the last sensor exploits full side-information leading to $I(Y_{\pi(M)};Z_{\pi(M)}|Z_{<\pi(M)})$. For the sake of notational simplicity, the permutations are skipped in subsequent equations. Using Lagrange multipliers β_m to reformulate (15) as an unconstrained optimization problem yields the function

$$L_{\text{DIB}} = I(X;Z) - \sum_{m=1}^{M} \beta_m \, I(Y_m;Z_m|Z_{<m}) \tag{16}$$

which has to be maximized. The Lagrange multipliers β_m have to be chosen such that the individual rate constraints in (15) are fulfilled. As described in [48], the solution of (15) can be computed by successively updating one mapping $p(z_m|y_m)$ at a time keeping other mappings fixed. The solution of this optimization problem is derived in App. C and leads to the update equation

$$p(z_m|y_m) = \frac{e^{-d_{m:M}(y_m,z_m)}}{\sum_{z_m} e^{-d_{m:M}(y_m,z_m)}} \tag{17}$$

for sensor m with

$$d_{m:M}(y_m,z_m) := \frac{1}{\beta_m} \mathbb{E}_{Z_{\bar{m}}|y_m} \left[D_{\text{KL}}[p(x|y_m,z_{\bar{m}}) \| p(x|z)] - \sum_{m'=m}^{M} \beta_{m'} \log p(z_{m'}|z_{<m'}) \right]. \tag{18}$$

Please note that the relevant mutual information considers the mappings of all sensors as indicated by the Kullback-Leibler divergence. The compression rates represented by the log-terms in (18), however, exploit only partial side-information $\mathbf{z}_{<m'}$. The update rule in (17) has to be successively applied to all sensors. Neglecting the required permutations, Fig. 3 illustrates the flowchart of the DIB optimization. The block "BA algorithm" contains the most inner loop and updates quantizer m for a fixed $_m$ exploiting appropriate side-information as described above. This block basically contains a Blahut-Arimoto-like algorithm. Next, the "loop to adjust " finds the rate-fulfilling Lagrange multiplier $_m$ using a bisection search. The next outer loop runs over all sensors m and the most outer loop repeats the whole procedure until no significant changes in the mappings can be observed. An appropriate stopping criterion is the average Jensen-Shannon divergence between the mappings of the current and the previous iteration.

B. Greedy Distributed Information Bottleneck (GDIB) Approach

Applying the chain rule of information to $I(X;Z)$ in (16), the DIB approach can be reformulated to

$$L_{\text{DIB}} = \sum_{m=1}^{M} I(X;Z_m \mid Z_{<m}) \quad {}_m \, I(Y_m;Z_m|Z_{<m}). \tag{19}$$

The formulation in (19) suggests a greedy optimization strategy leading to the following set of M optimization problems.

$$L_{\text{GDIB}}^{(1)} = I(X;Z_1) \quad {}_1 I(Y_1;Z_1) \tag{20a}$$

$$\vdots$$

$$L_{\text{GDIB}}^{(M)} = I(X;Z_M \mid Z_{<M}) \quad {}_M I(Y_M;Z_M|Z_{<M}) \tag{20b}$$

Starting with the optimization of the first sensor without side-information in (20a), the process continues with the second sensor exploiting $p(z_1 \mid y_1)$ as side-information. This process continues until the mapping for sensor M has been obtained via (20b). The major difference to the DIB approach in (16) is that the relevant mutual information $I(X;Z)$ depending on all mappings $p(z_m \mid y_m)$, $1 \quad m \quad M$, is not considered at once but split into M terms $I(X;Z_m \mid Z_{<m})$. As these terms depend only on partial side-information $Z_{<m}$ like the compression rates, each summand in (19) is optimized separately. As a consequence, only a single rate constraint is

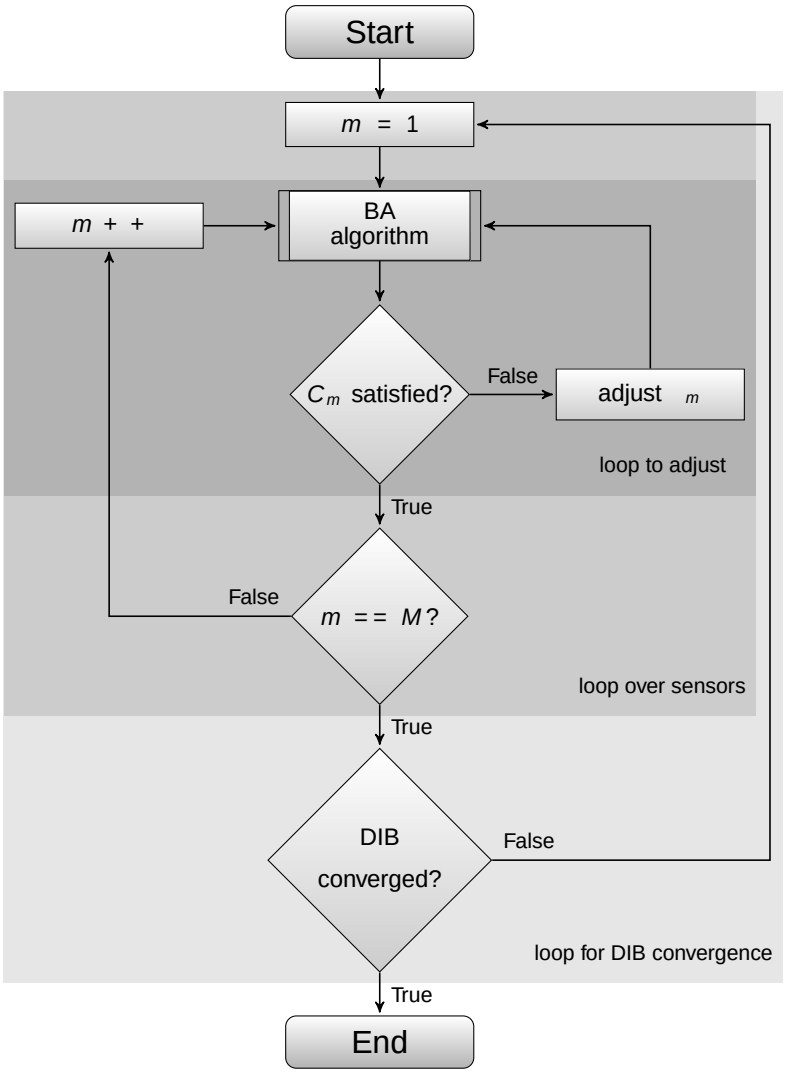

Fig. 3. Flowchart of the DIB optimization algorithm

considered in each step, which motivates the name successive IB approach. With this observation, the result in (18) can be easily modified for the GDIB approach yielding the update equation

$$p(z_m|y_m) = \frac{e^{d_m(y_m,z_m)}}{\sum_{z_m} e^{d_m(y_m,z_m)}} \tag{21}$$

with

$$d_m(y_m,z_m) := \mathrm{E}_{Z_{<m}|y_m}\left[\frac{1}{m} D_{\mathrm{KL}}[p(x|y_m,z_{<m}) \| p(x|z_m)] - \log p(z_m|z_{<m})\right]. \tag{22}$$

Again, the optimization can be efficiently solved by an iterative Blahut-Arimoto-like algorithm [42]. Equivalent to the DIB approach, (20a) - (20b) have to be solved for all possible permutations

because the optimal assignment of side-information is not known in advance. However, no outer loop as depicted in Fig. 3 is required because the GDIB optimization of sensor m does not depend on mappings of sensors $m' > m$. This reduces the computational complexity moderately.

Certainly, this greedy optimization strategy needs not to provide the same solution as the DIB approach. In (16), all mappings $p(z_m \mid y_m)$ are coupled via the relevant mutual information $I(X;Z)$. Optimizing instead each summand in (19) separately, partly neglects these dependencies. As each term in the sum depends in a non-convex way on all contributing mappings $p(z_m \mid y_m)$, there is no guarantee that DIB and GDIB solutions are identical. However, numerical results demonstrate identical performances for both approaches.

C. DIB Approach with Sum-Rate Constraint

In [22], Estella et al. proposed a generalization of the Blahut-Arimoto algorithm to minimize the average logarithmic loss under the sum-rate constraint $I(Y;Z) = \sum_{m=1}^{M} C_m$ based on the outer bound of the rate region in (12). Due to the restriction on the sum-rate, the outer bound (12) has only to be evaluated for the set $\mathcal{S} = \{1, 2, \ldots, M\}$. Their approach leads to the minimization problem

$$\min_{\mathbf{P}} H(X|Z) + \sum_{m=1}^{M} I(Y_m;Z_m) + H(X|Z_m) \tag{23}$$

for $\mathbf{P} := \{p(z_1|y_1), \ldots, p(z_M|y_M)\}$. Following similar steps as in [24] for deriving the Blahut-Arimoto algorithm results in a self-consistent solution

$$p(z_m|y_m) = \frac{p(z_m)\ e^{\,d\,(y_m,z_m)}}{\sum_{z_m} p(z_m)\ e^{\,d\,(y_m,z_m)}}\,. \tag{24}$$

As opposed to the scalar IB approach, the exponent

$$d\,(y_m,z_m) = D_{\mathrm{KL}}[p(x|y_m)\,\|\,p(x|z_m)] + \frac{1}{\beta}\,\mathsf{E}_{Z_{\bar{\mathcal{S}}m}|y_m}\left\{ D_{\mathrm{KL}}[p(x|y_m,z_{\bar{\mathcal{S}}m})\,\|\,p(x|z)] \right\} \tag{25}$$

contains now two Kullback-Leibler (KL) divergence terms $D_{\mathrm{KL}}[\,\|\,]$. The first term in (25) is also part of the scalar IB solution while the second term occurs only in the distributed setup. It represents the loss of relevant mutual information about x when quantization index z_m is available instead of the corresponding quantizer input y_m besides the other sensor outputs $z_{\bar{\mathcal{S}}m}$.

Note that this approach just considers the sum-rate constraint and uses only a single Lagrange multiplier. Consequently, it does not allow individual rate adjustments for each sensor. Instead, the solution provides mappings with compression rates $I(Y_{\mathcal{S}};Z_{\mathcal{S}} \mid Z_{\bar{\mathcal{S}}})$ which may not fulfill the

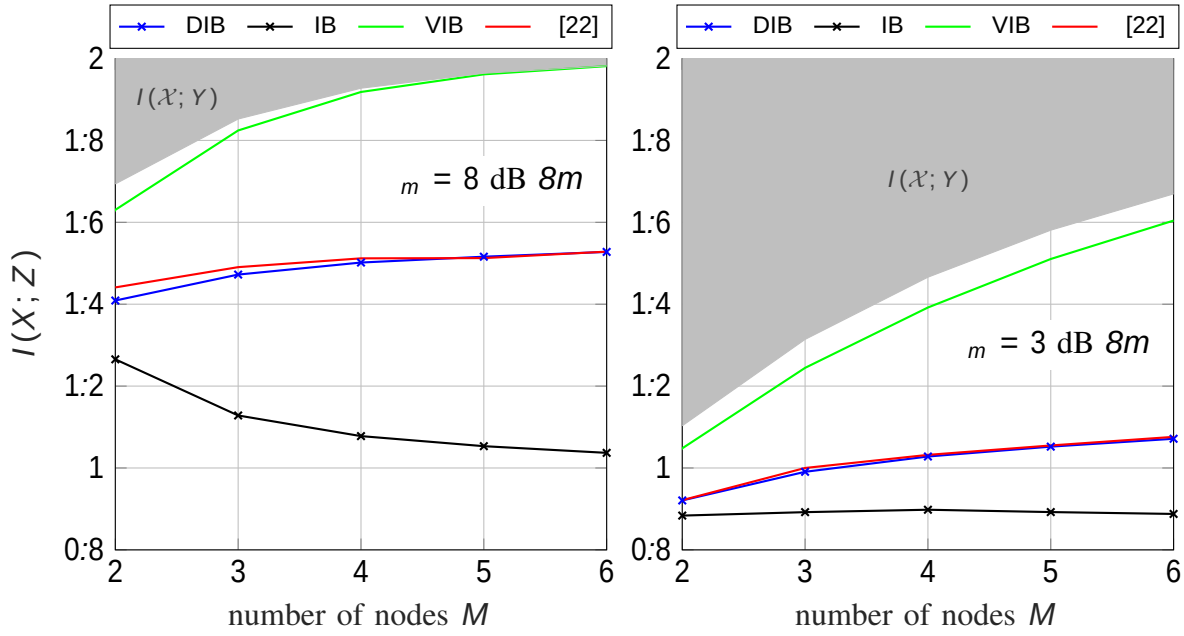

Fig. 4. Relevant mutual information vs. number of sensors for fixed sum-rate C_{sum} = 2:5 bit/s/Hz, $|X| = 4$, $|Z_m| = 4\ \forall m$

associated rate constraints C_S in a particular scenario. This effect will be analyzed in the next section.

V. Evaluation / Numerical Results

A. *Influence of Number of Nodes*

In this section, numerical results for the proposed algorithmic solutions of the CEO problem will be discussed. The first scenario considers of a varying number of sensors and is symmetric, i.e. the measurement SNRs $_m$, the link capacities C_m and the quantizer cardinalities jZ_mj are identical for all sensors $1 \quad m \quad M$. The measurement noise is Gaussian distributed and the relevant signal is a uniformly distributed 4-ASK (Amplitude Shift Keying). In order to allow a fair comparison with the approach from [22] (see Section IV-C), the sum-rate $C_{\text{sum}} = \sum_{m=1}^{M} C_m$ was fixed to a value independent of M. Consequently, the more nodes contribute to the distributed sensing, the smaller is the individually available capacity $C_m = \frac{C_{\text{sum}}}{M}$ of each forward link. This represents a scenario where all M nodes equally share a common medium in an orthogonal way and a round robin fashion.

Figs. 4 and 5 illustrate the influence of the number of nodes onto the relevant mutual information $I(X;Z)$. Two different measurement SNRs $_m$ 2 $f3;8g$ dB are considered, the quantizer

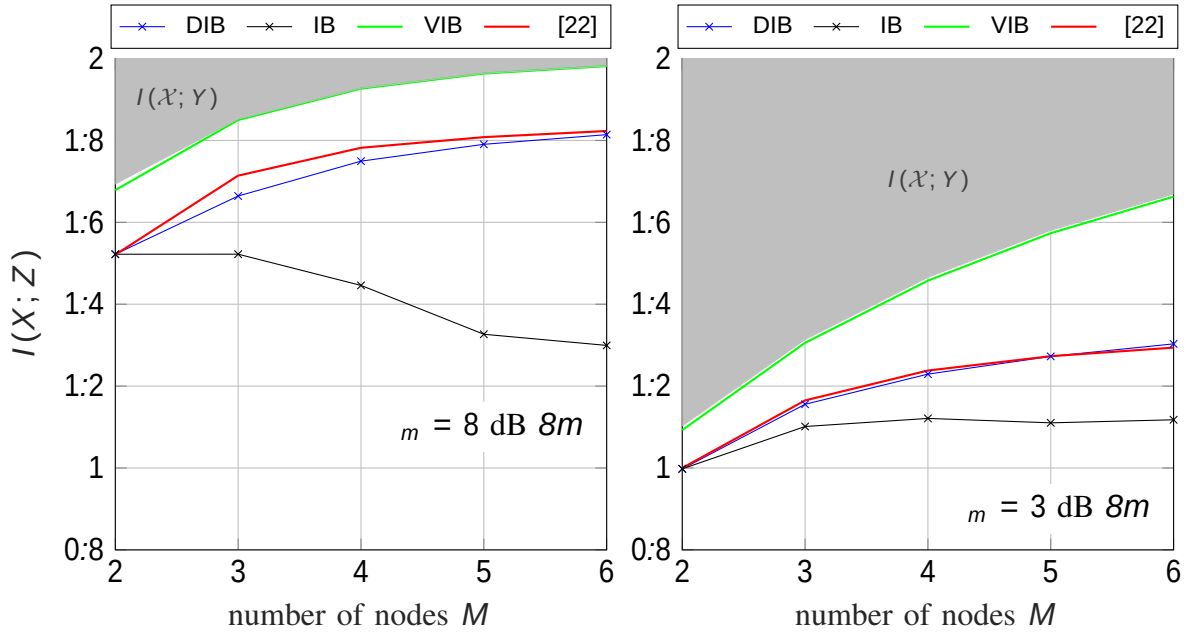

Fig. 5. Relevant mutual information vs. number of sensors for fixed sum-rate C_{sum} = 4:0 bit/s/Hz, $|\mathcal{X}|$ = 4, $|\mathcal{Z}_m|$ = 4 $\forall m$

cardinality is $|\mathcal{Z}_m| = 4$ and the sum-rate amounts to $C_{\text{sum}} \approx \{2.5, 4\}$ bit/s/Hz. In Fig. 4 and for = 8 dB, it can be observed that independent scalar optimization of the quantizers (IB) looses relevant information with increasing number of nodes. As the compression at each sensor has to become stronger with growing M, the overall relevant information decreases and it is beneficial to use less sensors with higher compression rates. For low SNRs like = 3 dB, $M = 3$ and $M = 4$ sensors are slightly superior to only $M = 2$ sensors. However, larger M leads again to a degradation. Second, the DIB approach does not suffer from this effect and can increase $I(X;Z)$ with growing M. This result demonstrates that joint optimization of distributed quantizers leads to a significant gain compared to separately optimized quantization. The greedy GDIB algorithm not shown in Figs. 4 and 5 achieves the same relevant information as the DIB approach. A comparison of Figs. 4 and 5 show that a higher sum-rate leads to qualitatively similar results, quantitatively a larger relevant mutual information $I(X;Z)$ can be achieved.

Compared to the sum-rate approach from [22], both DIB and GDIB seem to perform slightly worse. The difference is larger at high SNR and a few sensors and becomes negligible for $M > 5$ or at low SNRs. In the considered symmetric scenario, the sum-rate approach delivers identical mappings $p(z_m \mid y_m)$ for all sensors. This is not the case for the greedy DIB and GDIB algorithms whose quantizers are slightly different. It seems that these algorithms get stuck in sub-

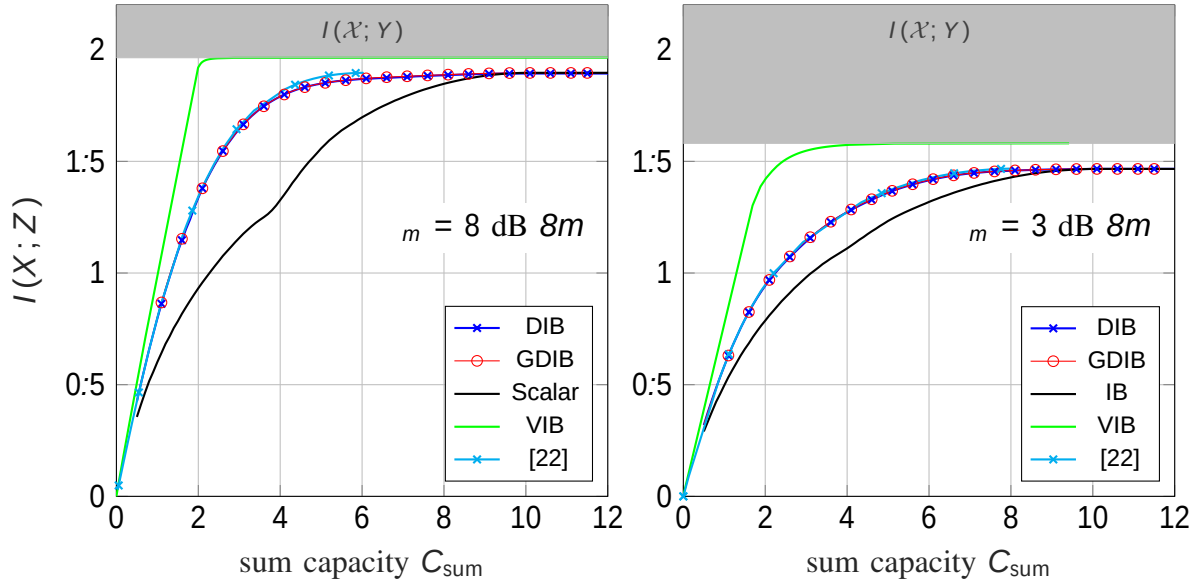

Fig. 6. Relevant mutual information vs. sum capacity for a symmetric scenario with $M = 5$ sensors, $|\mathsf{X}| = 4$, SNRs $_m \in \{3; 8\}$ dB and $|\mathsf{Z}_m| = 4$

optimal fixed points due to the successive optimization strategy. The achieved sum-rate $I(Y;Z)$ is slightly smaller than for the algorithm from [22]. Therefore, the algorithm proposed in [22] provides better solutions in strictly symmetric scenarios.

Finally, the hypothetical vector quantizer clearly outperforms all distributed approaches as expected. It benefits most from an increasing number of sensors while DIB and GDIB gain only moderately with growing M. This illustrates the limitation of distributed non-cooperative quantization not allowing an exchange of information between different sensors. In Fig. 4, the loss of the hypothetical vector quantizer compared to $I(X;Y)$ is significant due to the stricter sum-rate constraint. Clearly, this changes in Fig. 5 for $C_{\text{sum}} = 4$ bit/s/Hz where $I(X;Z)$ approaches $I(X;Y)$. For higher sum-rate constraints, the VIB gain over DIB and GDIB shrinks. All obtained solutions have been checked to be feasible, i.e. they fulfill the rate constraints.

B. Influence of Sum-Rate

Fig. 6 illustrates the relevant information versus the sum-rate for a symmetric scenario with $M = 5$ sensors, $jXj = 4$, SNRs $_m$ 2 $f3; 8g$ dB and $jZ_mj = 4$. Please note that due to $jZ_mj = 4$ and $M = 5$, a sum-rate C_{sum} 10 does not require any lossy compression at the sensors. Only in this region the independent scalar IB optimization reaches the same relevant information

as the DIB approach. For smaller sum-rates, it performs worse. Furthermore, the comparison with the vector quantization shows again the loss of distributed compression which is largest for moderate sum-rates between 2 C_{sum} 4 bit/s/Hz. Regarding the solutions derived in Section IV, GDIB and DIB deliver identical results. For high sum-rates of $C_{\text{sum}} > 4$ bis/s/Hz and $_m = 8$ dB, Zaidi's approach from [22] slightly outperforms GDIB and DIB. For smaller C_{sum} and $_m = 3$ dB, the results coincide.

C. Influence of Wyner-Ziv Coding Strategy for Asymmetric Scenarios

Finally, the influence of permutations, i.e. the Wyner-Ziv coding strategies is investigated. As no variations in the results for different permutations can be observed in symmetric scenarios, two asymmetric scenarios for $M = 4$ sensors are analyzed. In the first scenario, sensors with bad SNRs also have low link capacities, whereas in the second scenario sensors with bad SNRs have high link capacities and vice versa.

Fig. 7 illustrates the relevant information for all permutations ( ) of the set $\{1,\dots,4\}$ using the greedy GDIB algorithm. In the blue case, bad SNRs coincide with low link capacities and only minor differences regarding the achieved relevant mutual information can be observed. It seems that the Wyner-Ziv coding strategies do not have a big impact onto the result for this scenario. For the opposite red case, the overall performance is worse than for the blue case. This observation can be expected because accurate measurements have to be strongly compressed while unreliable measurements can contribute little to the overall result even if the corresponding link capacities are high. Moreover, significant differences occur between different permutations. Even though no clear conclusion about the optimal coding strategy can be drawn from Fig. 7, it seems that starting with the best measurements but strongest compression leads to the worst performance.

Fig. 8 depicts the different compression rates for the best permutation of the previously described asymmetric scenario, where a good measurement SNR corresponds to a high capacity link. Note that R represents the achieved compression rate $I(Y_{\mathcal{S}}; Z_{\mathcal{S}} \mid Z_{\bar{\mathcal{S}}})$, i.e. the left part of (10). The solid line illustrates the right part of (10) defining upper bounds on the compression rates given by the individual link capacities of this scenario. In order to provide a feasible solution the achieved compression rates must all lie below the solid rate constraint curve. The GDIB approach fulfills all rate constraints. In comparison, the sum-rate based approach of [22]

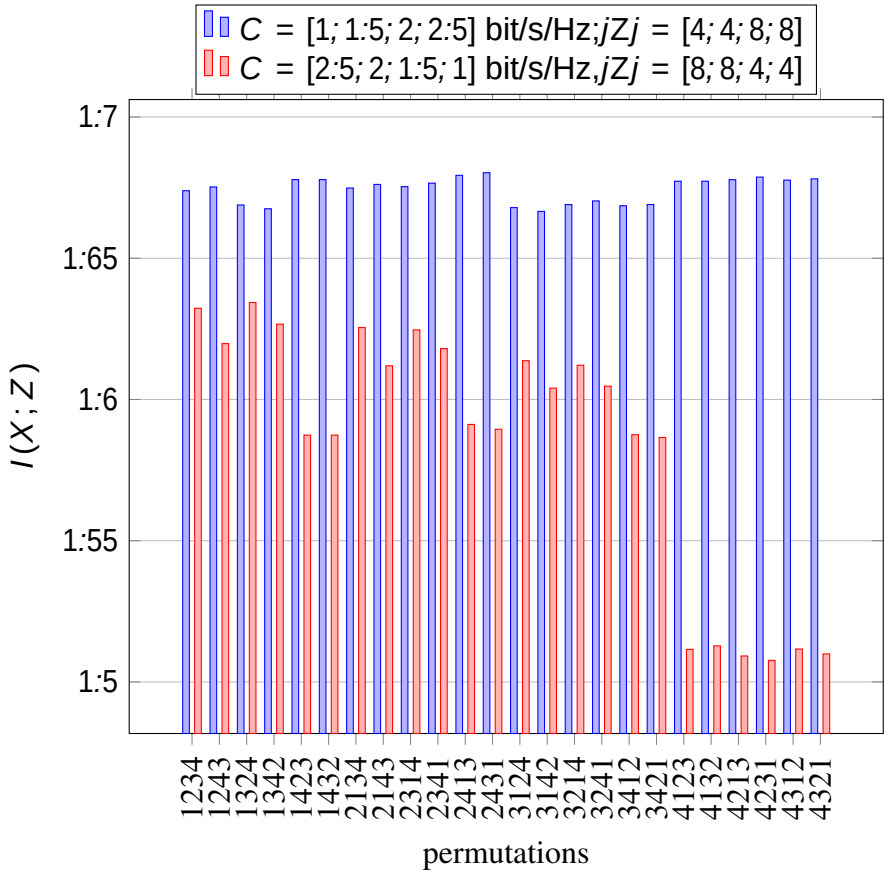

Fig. 7. Relevant mutual information for non symmetric scenario with $M = 4$ sensors, SNR = [2,4,6,8] dB and $|\mathsf{X}| = 4$ using GDIB optimization.

does not fulfill the first rate condition. Consequently, it does not provide a valid solution given the individual link capacities in this scenario.

Finally, Fig. 9 illustrates the compression rates for the best permutation of the previously described opposite scenario, where a good measurement SNR is associated to a low capacity link. Again, the GDIB approach fulfills all rate constraints while the sum-rate based approach of [22] violates 3 out of 15 rate constraints. This observation demonstrates that the proposed DIB and GDIB approaches deliver feasible solutions for all investigated scenarios while this is not possible if only the sum-rate is considered. If the cardinalities $j\mathsf{Z}_m j$ are not adapted to the link capacities C_m as done in Fig. 8 and Fig. 9 but are chosen equally, e.g. $j\mathsf{Z}_m j = 8\ \forall m$, the violation of the rate constraints when considering just the sum-rate becomes even more pronounced.

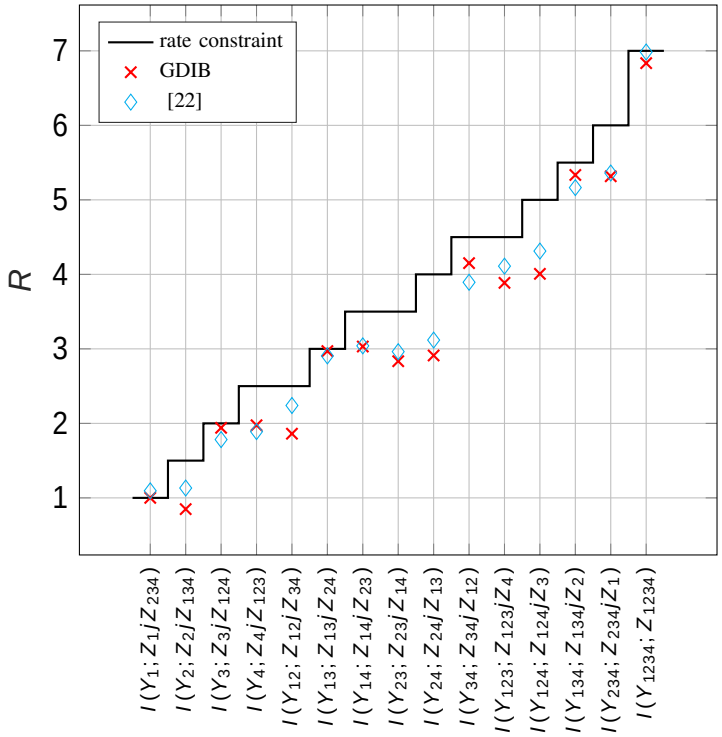

Fig. 8. Compression rates for non symmetric scenario (best permutation) with $M = 4$ sensors, SNR = [2,4,6,8] dB, $C = [1; 1{:}5; 2; 2{:}5]$ and $|\mathsf{X}| = 4$ using GDIB optimization. A good SNR corresponds to a link with high capacity.

VI. Conclusion

This paper proposed two algorithmic solutions for solving the CEO problem based on the well known IB method. The DIB algorithm considers all rate constraints simultaneously but optimizes the quantizers successively exploiting other mappings as side-information. The greedy GDIB approach also performs a successive optimization but considers just one rate constraint at a time. The proposed algorithms substantially outperform independently IB-optimized quantizers and benefit from a growing number of sensors which is not the case for the latter. Compared to a sum-rate optimization approach, the solutions are slightly worse in strictly symmetric scenarios. However, they are always feasible even in asymmetric scenarios where the sum-rate optimization often fails. Furthermore, numerical results illustrate the dependency of the solution from permutations of the Wyner-Ziv coding strategies. Finally, the loss of distributed compression compared to vector quantization is discussed.

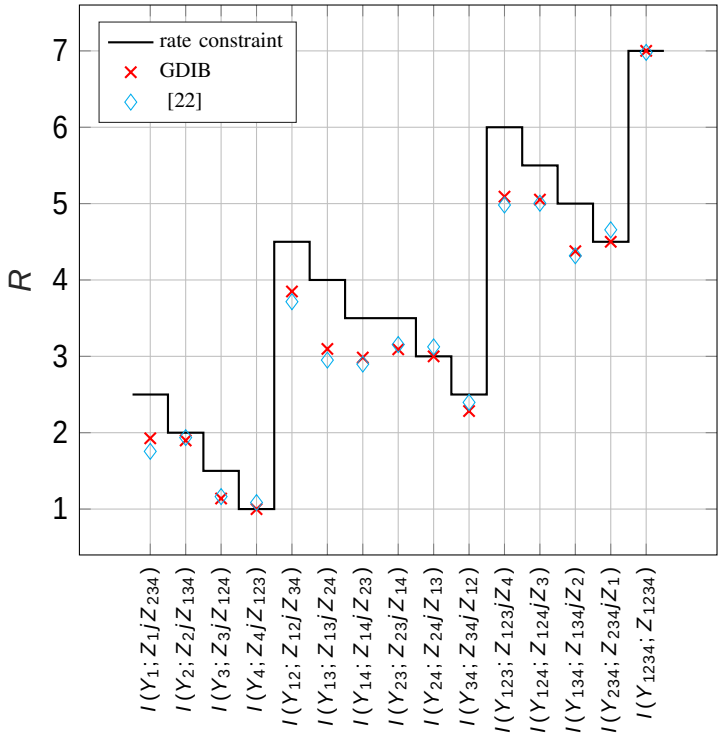

Fig. 9. Compression rates for non symmetric scenario (best permutation) with $M = 4$ sensors, SNR = [2,4,6,8] dB, C = [2:5; 2; 1:5; 1] and $|\mathcal{X}| = 4$ using GDIB optimization. A good SNR corresponds to a link with low capacity.

APPENDIX A
PROOF: PARTIAL SIDE-INFORMATION FULFILLS RATE-CONSTRAINTS

This appendix proofs that the DIB utility function with partial side-information yields a solution that fulfills all rate constraints defining the CEO rate-region. According to the formulation in (15), the DIB approach determines mappings $p(z_m|y_m)$ for which the single-node rate constraint ($|\mathcal{S}| = 1$)

$$I(Y_m; Z_m | Z_{<m}) \quad C_m \tag{26}$$

is fulfilled for all $1 \quad m \quad M$. For multi-node rate constraints with $|\mathcal{S}| > 1$, we obtain

$$\begin{aligned} I(Y_{\mathcal{S}}; Z_{\mathcal{S}} \mid Z_{\bar{\mathcal{S}}}) &= \mathrm{E}_{Y_{\mathcal{S}};Z}\left[\log \frac{p(\mathbf{z}_{\mathcal{S}} \mid \mathbf{y}_{\mathcal{S}})}{p(\mathbf{z}_{\mathcal{S}} \mid \mathbf{z}_{\bar{\mathcal{S}}})}\right] \\ &= \mathrm{E}_{Y_{\mathcal{S}};Z}\left[\sum_{m \in \mathcal{S}} \log \frac{p(z_m \mid y_m)}{p(z_m \mid \mathbf{z}_{<m})} - \log \frac{p(\mathbf{z}_{\mathcal{S}} \mid \mathbf{z}_{\bar{\mathcal{S}}})}{p(\mathbf{z}_{\mathcal{S}})}\right] \\ &= \sum_{m \in \mathcal{S}} I(Y_m; Z_m \mid Z_{<m}) - I(Z_{\mathcal{S}}; Z_{\bar{\mathcal{S}}}) \qquad (27) \\ &\leq \sum_{m \in \mathcal{S}} C_m. \qquad (28) \end{aligned}$$

The last inequality holds due to (26) and the fact that the mutual information $I(Z_{\mathcal{S}}; Z_{\bar{\mathcal{S}}})$ is non-negative. Hence, all rate constraints defining the CEO rate region are fulfilled. In (28), equality only holds if every compression rate equals the associated link capacity, $I(Y_m; Z_m \mid Z_{<m}) = C_m$, and $I(Z_{\mathcal{S}}; Z_{\bar{\mathcal{S}}}) = 0$, i.e. cluster indexes in $\mathcal{S}$ and $\bar{\mathcal{S}}$ do not share common information.

APPENDIX B
PROOF: FULL SIDE-INFORMATION DOES NOT FULFILL RATE-CONSTRAINTS

Similar to the proof for partial side-information in App. A, it is now shown that full side-information for all compression rates generally leads to infeasible solutions violating at least some of the CEO rate constraints. Full side-information means that the mappings of all sensors $\not\in m$ are considered for the compression rate of sensor m leading to the single-node rate constraint ($|\mathcal{S}| = 1$)

$$I(Y_m; Z_m \mid Z_{\not\in m}) \quad C_m \qquad (29)$$

for all $1 \quad m \quad M$. If the optimization fulfills the inequality in (29) with equality for all $1 \quad m \quad M$, the multi-node rate constraint with $|\mathcal{S}| > 1$ becomes

$$\begin{aligned} I(Y_{\mathcal{S}}; Z_{\mathcal{S}} \mid Z_{\bar{\mathcal{S}}}) &= \mathrm{E}_{Y_{\mathcal{S}};Z}\left[\log \frac{p(\mathbf{z}_{\mathcal{S}} \mid \mathbf{y}_{\mathcal{S}})}{p(\mathbf{z}_{\mathcal{S}} \mid \mathbf{z}_{\bar{\mathcal{S}}})}\right] = \mathrm{E}_{Y_{\mathcal{S}};Z}\left[\log \frac{\prod_{m \in \mathcal{S}} p(z_m \mid y_m)}{\prod_{m \in \mathcal{S}} p(z_m \mid \mathbf{z}_{\not\in m})} \cdot \frac{\prod_{m \in \mathcal{S}} p(z_m \mid \mathbf{z}_{\not\in m})}{p(\mathbf{z}_{\mathcal{S}} \mid \mathbf{z}_{\bar{\mathcal{S}}})}\right] \\ &= \mathrm{E}_{Y_{\mathcal{S}};Z}\left[\sum_{m \in \mathcal{S}} \log \frac{p(z_m \mid y_m)}{p(z_m \mid \mathbf{z}_{\not\in m})} - \log p(\mathbf{z}_{\mathcal{S}nm} \mid \mathbf{z}_{\bar{\mathcal{S}}})\right] \\ &= \sum_{m \in \mathcal{S}} I(Y_m; Z_m \mid Z_{\not\in m}) + H(Z_{\mathcal{S}nm} \mid Z_{\bar{\mathcal{S}}}) \\ &\sum_{m \in \mathcal{S}} C_m \qquad 8 \quad \mathcal{S} \quad \{1, 2, \ldots, M\}. \qquad (30) \end{aligned}$$

As indicated in (30), the last inequality cannot be ensured. If all single-node rate constraints in (29) can be fulfilled with equality, the non-negativity of the conditional entropy $H(Z_{\mathcal{S}nm} \mid Z_{\bar{\mathcal{S}}})$ leads to a violation of the corresponding rate-constraint. Hence, the multi-node rate constraints of the CEO rate region are not generally fulfilled. This is only the case if $H(Z_{\mathcal{S}nm} \mid Z_{\bar{\mathcal{S}}}) = 0$ holds, i.e. the random variables $Z_{\mathcal{S}nm}$ are completely determined by $Z_{\bar{\mathcal{S}}}$, or the compression rates in (29) are smaller than C_m. However, it is unknown how large the gap between $I(Y_m; Z_m \mid Z_{\mathcal{S}\setminus m})$ and C_m needs to be and whether this leads to a maximization of $I(X;Z)$.

APPENDIX C
DERIVATION OF DIB SOLUTION

The partial solution for sensor m that maximizes (16) can be found by equating the derivative of

$$L_{\text{DIB}} = I(X;Z) - \sum_{m'=1}^{M} \beta_{m'} \, I(Y_{m'}; Z_{m'} \mid Z_{<m'})$$

w.r.t. $p(z_m \mid y_m)$ to zero. Applying the chain rule, the relevant mutual information $I(X;Z)$ can be rewritten such that only the term $I(X; Z_m \mid Z_{\mathcal{S}\setminus m})$ depends on the desired mapping. Since only sensors $m' > m$ exploit the mapping $p(z_m \mid y_m)$ as side-information, all terms being independent of $p(z_m \mid y_m)$ can be skipped because they are constants for the derivative w.r.t. $p(z_m \mid y_m)$. This leads to the particular utility function for sensor m

$$L_{\text{DIB}}^{(m)} = I(X; Z_m \mid Z_{\mathcal{S}\setminus m}) - \beta_m \, I(Y_m; Z_m \mid Z_{<m}) - \sum_{m'>m} \beta_{m'} \, I(Y_{m'}; Z_{m'} \mid Z_{<m'}) \, . \tag{31}$$

Note that the mapping $p(z_m \mid y_m)$ obviously influences the terms $I(X; Z_m \mid Z_{\mathcal{S}\setminus m})$ and $I(Y_m; Z_m \mid Z_{<m})$ while it affects $I(Y_{m'}; Z_{m'} \mid Z_{<m'})$ as part of the side-information $Z_{<m'}$. We will now derive the derivatives for all three types of mutual information terms.

A. Derivative of $I(X; Z_m \mid Z_{\mathcal{S}\setminus m})$

The relevant mutual information in (31) can be rewritten such that the desired mapping occurs explicitly.

$$\begin{aligned} I(X; Z_m \mid Z_{\mathcal{S}\setminus m}) &= \mathsf{E}_{X;Z}\left[\log \frac{p(z_m \mid x; \mathbf{z}_{\mathcal{S}\setminus m})}{p(z_m \mid \mathbf{z}_{\mathcal{S}\setminus m})} \right] \\ &= \sum_{z_m} \sum_{y_m} p(z_m \mid y_m) \sum_{x} p(x; y_m) \log \sum_{a \in \mathcal{Y}_m} p(z_m \mid a) p(a \mid x) \\ &\quad - \sum_{z_m} \sum_{y_m} p(z_m \mid y_m) \sum_{\mathbf{z}_{\mathcal{S}\setminus m}} p(y_m; \mathbf{z}_{\mathcal{S}\setminus m}) \log \sum_{a \in \mathcal{Y}_m} p(z_m \mid a) p(a \mid \mathbf{z}_{\mathcal{S}\setminus m}) \end{aligned} \tag{32}$$

The derivative of (32) delivers

$$\begin{aligned}\frac{\partial I(X;Z_m \mid Z_{<m})}{\partial p(z_m \mid y_m)} &= \sum_x p(x,y_m)\,\log p(z_m \mid x) + \sum_x \Bigg[\underbrace{\sum_{y_m} p(z_m \mid y_m)\; p(x,y_m)}_{=p(x,z_m)}\Bigg] \frac{p(y_m \mid x)}{p(z_m \mid x)} \\ &\quad - \sum_{\mathbf{z}_{<m}} p(y_m,\mathbf{z}_{<m})\,\log p(z_m \mid \mathbf{z}_{<m}) \\ &\quad - \sum_{\mathbf{z}_{<m}} \Bigg[\underbrace{\sum_{y_m} p(z_m \mid y_m)\; p(y_m,\mathbf{z}_{<m})}_{=p(\mathbf{z})}\Bigg] \frac{p(y_m \mid \mathbf{z}_{<m})}{p(z_m \mid \mathbf{z}_{<m})} \qquad (33) \\ &= \sum_x p(x,y_m)\,\log p(z_m \mid x) - \sum_{\mathbf{z}_{<m}} p(y_m,\mathbf{z}_{<m})\,\log p(z_m \mid \mathbf{z}_{<m}) \\ &= \sum_{\mathbf{z}_{<m}} p(y_m,\mathbf{z}_{<m}) \sum_x p(x \mid y_m,\mathbf{z}_{<m})\,\log \frac{p(z_m \mid x)}{p(z_m \mid \mathbf{z}_{<m})}\,. \qquad (34)\end{aligned}$$

Exploiting the Markov property $X \to Y_m \to Z_m$ and the independence of Z_m given x, the argument of the logarithmic function can be extended to

$$\frac{p(z_m \mid x)}{p(z_m \mid \mathbf{z}_{<m})} = \frac{p(z_m \mid x,\mathbf{z}_{<m})}{p(z_m \mid \mathbf{z}_{<m})} = \frac{p(x \mid \mathbf{z})}{p(x \mid \mathbf{z}_{<m})} = \frac{p(x \mid \mathbf{z})}{p(x \mid y_m,\mathbf{z}_{<m})}\,\frac{p(x \mid y_m,\mathbf{z}_{<m})}{p(x \mid \mathbf{z}_{<m})}\,. \qquad (35)$$

The second ratio in (35) can be dropped because it does not depend on $p(z_m \mid y_m)$ and its contribution can be incorporated into the Lagrange multiplier λ_m. The insertion of the first ratio into (34) yields the contribution of the derivative of the relevant mutual information

$$\begin{aligned}\frac{\partial I(X;Z_m \mid Z_{<m})}{\partial p(z_m \mid y_m)} &\propto \sum_{\mathbf{z}_{<m}} p(y_m,\mathbf{z}_{<m}) \sum_x p(x \mid y_m,\mathbf{z}_{<m})\,\log \frac{p(x \mid y_m,\mathbf{z}_{<m})}{p(x \mid \mathbf{z})} \\ &= \sum_{\mathbf{z}_{<m}} p(y_m,\mathbf{z}_{<m})\; D_{\mathrm{KL}}\left[p(x \mid y_m,\mathbf{z}_{<m}) \,\|\, p(x \mid \mathbf{z})\right]\,. \qquad (36)\end{aligned}$$

B. Derivative of $I(Y_m;Z_m \mid Z_{<m})$

With the definition of the conditional compression rate

$$\begin{aligned}I(Y_m;Z_m \mid Z_{<m}) &= \mathrm{E}_{Y_m,Z}\left[\log \frac{p(z_m \mid y_m)}{p(z_m \mid \mathbf{z}_{<m})}\right] \\ &= \sum_{z_m}\sum_{y_m} p(z_m \mid y_m)\; p(y_m)\;\log p(z_m \mid y_m) \\ &\quad - \sum_{z_m}\sum_{y_m} p(z_m \mid y_m) \sum_{\mathbf{z}_{<m}} p(y_m,\mathbf{z}_{<m})\;\log \sum_{a \in \mathcal{Y}_m} p(z_m \mid a)\,p(a \mid \mathbf{z}_{<m}); \qquad (37)\end{aligned}$$

its derivative becomes

$$\begin{aligned}\frac{\partial I(Y_m; Z_m \mid Z_{<m})}{\partial p(z_m \mid y_m)} &= p(y_m) \log p(z_m \mid y_m) + p(y_m) \frac{p(z_m \mid y_m)}{p(z_m \mid y_m)} \\ &\quad - \sum_{\mathbf{z}_{<m}} p(y_m, \mathbf{z}_{<m}) \log p(z_m \mid \mathbf{z}_{<m}) \\ &\quad - \sum_{\mathbf{z}_{<m}} \Bigg[\underbrace{\sum_{y_m} p(z_m \mid y_m) p(y_m, \mathbf{z}_{<m})}_{=p(\mathbf{z}_{\le m})} \Bigg] \frac{p(y_m \mid \mathbf{z}_{<m})}{p(z_m \mid \mathbf{z}_{<m})} \\ &= p(y_m) \log p(z_m \mid y_m) - \sum_{\mathbf{z}_{<m}} p(y_m, \mathbf{z}_{<m}) \log p(z_m \mid \mathbf{z}_{<m}). \end{aligned} \tag{38}$$

C. Derivative of $I(Y_{m'}; Z_{m'} \mid Z_{<m'})$

The mutual information terms $I(Y_{m'}; Z_{m'} \mid Z_{<m'})$ in the sum of (31) can be expressed as

$$\begin{aligned} I(Y_{m'}; Z_{m'} \mid Z_{<m'}) &= \mathsf{E}\left[\log \frac{p(z_{m'} \mid y_{m'}, \mathbf{z}_{<m'})}{p(z_{m'} \mid \mathbf{z}_{<m'})} \right] \\ &= \sum_{y_{m'}} \sum_{\mathbf{z}_{\le m'}} p(y_{m'}, \mathbf{z}_{\le m'}) \left[\log p(z_{m'} \mid y_{m'}, \mathbf{z}_{<m'}) - \log p(z_{m'} \mid \mathbf{z}_{<m'}) \right] \\ &= \sum_{z_m} \sum_{y_m} \sum_{y_{m'}} \sum_{\mathbf{z}_{\le m' \setminus m}} p(z_m \mid y_m) p(y_m, y_{m'}, \mathbf{z}_{\le m' \setminus m}) \\ &\quad \cdot \left[\log p(z_{m'} \mid y_{m'}) - \log \frac{\sum_{a \in \mathcal{Y}_m} p(z_m \mid a) p(a, \mathbf{z}_{\le m' \setminus m})}{\sum_{a \in \mathcal{Y}_m} p(z_m \mid a) p(a, \mathbf{z}_{<m' \setminus m})} \right]. \end{aligned} \tag{39}$$

Its derivative becomes

$$\begin{aligned} &\frac{\partial I(Y_{m'}; Z_{m'} \mid Z_{<m'})}{\partial p(z_m \mid y_m)} \\ &= \sum_{y_{m'}} \sum_{\mathbf{z}_{\le m' \setminus m}} p(y_m, y_{m'}, \mathbf{z}_{\le m' \setminus m}) \log \frac{p(z_{m'} \mid y_{m'})}{p(z_{m'} \mid \mathbf{z}_{<m'})} \\ &\quad + \sum_{\mathbf{z}_{\le m' \setminus m}} \Bigg[\sum_{y_m} \sum_{y_{m'}} p(z_m \mid y_m) p(y_m, y_{m'}, \mathbf{z}_{\le m' \setminus m}) \Bigg] \left[\frac{p(y_m, \mathbf{z}_{<m' \setminus m})}{p(\mathbf{z}_{<m'})} - \frac{p(y_m, \mathbf{z}_{\le m' \setminus m})}{p(\mathbf{z}_{\le m'})} \right]. \end{aligned}$$

Be aware that the summations over y_m and $y_{m'}$ in the first square bracket do not affect the terms inside the second square bracket. With this, we obtain

$$\begin{aligned}\frac{\partial I(Y_{m'};Z_{m'}\,|\,Z_{<m'})}{\partial p(z_m\,|\,y_m)} &= \sum_{\mathbf{z}_{\setminus m}^{m'}}\Bigg[\sum_{y_{m'}} p(y_m,y_{m'};\mathbf{z}_{\setminus m}^{m'})\,\log\frac{p(z_{m'}\,|\,y_{m'})}{p(z_{m'}\,|\,\mathbf{z}_{<m'})}\Bigg] \\ &\quad + \sum_{\mathbf{z}_{\setminus m}^{m'}} p(\mathbf{z}_{\setminus m}^{m'})\left[\frac{p(y_m;\mathbf{z}_{<m'}^{\setminus m})}{p(\mathbf{z}_{<m'})} - \frac{p(y_m;\mathbf{z}_{\setminus m}^{m'})}{p(\mathbf{z}_{\setminus m}^{m'})}\right] \\ &= \sum_{\mathbf{z}_{\setminus m}^{m'}}\Bigg[\sum_{y_{m'}} p(y_m,y_{m'};\mathbf{z}_{\setminus m}^{m'})\,\log\frac{p(z_{m'}\,|\,y_{m'})}{p(z_{m'}\,|\,\mathbf{z}_{<m'})}\Bigg] \\ &\quad + p(y_m;\mathbf{z}_{<m'}^{\setminus m})\,p(z_{m'}\,|\,\mathbf{z}_{<m'}) - p(y_m;\mathbf{z}_{\setminus m}^{m'}) \\ &= \sum_{\mathbf{z}_{\setminus m}^{m'}}\sum_{y_{m'}} p(y_m,y_{m'};\mathbf{z}_{\setminus m}^{m'})\,\log\frac{p(z_{m'}\,|\,y_{m'})}{p(z_{m'}\,|\,\mathbf{z}_{<m'})}\,. \end{aligned} \tag{40}$$

Incorporating the numerator in the logarithm being independent of z_m into the Lagrange multiplier $\lambda_{m'}$ in (31), (40) can be formed to

$$\frac{\partial I(Y_{m'};Z_{m'}\,|\,Z_{<m'})}{\partial p(z_m\,|\,y_m)} \propto \sum_{\mathbf{z}_{\setminus m}^{m'}} p(y_m;\mathbf{z}_{\setminus m}^{m'})\,\log p(z_{m'}\,|\,\mathbf{z}_{<m'})\,. \tag{41}$$

D. Fusion of Derived Parts

Combining the result in (36), (38) and (41) for all $m' > m$ delivers the complete derivative

$$\begin{aligned} &\sum_{\mathbf{z}_{\setminus m}} p(y_m;\mathbf{z}_{\setminus m})\; D_{\mathrm{KL}}\left[p(x\,|\,y_m;\mathbf{z}_{\setminus m})\|p(x\,|\,\mathbf{z})\right] \\ &\quad \lambda_m p(y_m)\log p(z_m\,|\,y_m) + \lambda_m \sum_{\mathbf{z}_{<m}} p(y_m;\mathbf{z}_{<m})\;\log p(z_m\,|\,\mathbf{z}_{<m}) \\ &+ \sum_{m'>m}\lambda_{m'}\sum_{\mathbf{z}_{\setminus m}^{m'}} p(y_m;\mathbf{z}_{\setminus m}^{m'})\;\log p(z_{m'}\,|\,\mathbf{z}_{<m'}) \\ &= \lambda_m p(y_m)\;\log p(z_m\,|\,y_m) \\ &\quad \sum_{\mathbf{z}_{\setminus m}} p(y_m;\mathbf{z}_{\setminus m})\;\; D_{\mathrm{KL}}\left[p(x\,|\,y_m;\mathbf{z}_{\setminus m})\|p(x\,|\,\mathbf{z})\right] \sum_{m'\;\;m}\lambda_{m'}\log p(z_{m'}\,|\,\mathbf{z}_{<m'}) \;\;= 0\,: \end{aligned} \tag{42}$$

Following the idea of Blahut and Arimoto [42], $p(x \mid \mathbf{z})$ and $p(z_{m'} \mid \mathbf{z}_{<m'})$ are assumed to be independent of $p(z_m \mid y_m)$. With this trick, (42) can be resolved w.r.t. the desired mapping of sensor m leading to the self-consistent solution

$$p(z_m \mid y_m) = \frac{e^{\,d_{m:M}(y_m, z_m)}}{\sum_{z_m} e^{\,d_{m:M}(y_m, z_m)}} \tag{43}$$

with

$$\begin{aligned} d_{m:M}(y_m, z_m) &= \frac{1}{m} \sum_{\mathbf{z}_{\setminus m}} p(\mathbf{z}_{\setminus m} \mid y_m) \\ &\quad \Bigg[D_{\mathrm{KL}}\left[p(x \mid y_m, \mathbf{z}_{\setminus m}) \| p(x \mid \mathbf{z})\right] \sum_{m'} \sum_{m' \neq m} \log p(z_{m'} \mid \mathbf{z}_{<m'}) \Bigg] \\ &= \frac{1}{m} \mathbb{E}_{Z_{\setminus m} \mid y_m} \Bigg[D_{\mathrm{KL}}\left[p(x \mid y_m, \mathbf{z}_{\setminus m}) \| p(x \mid \mathbf{z})\right] \sum_{m'} \sum_{m' \neq m} \log p(z_{m'} \mid \mathbf{z}_{<m'}) \Bigg] : \end{aligned} \tag{44}$$

REFERENCES

[1] S.-H. Park, O. Simeone, O. Sahin, and S. Shitz, "Fronthaul compression for cloud radio access networks: Signal processing advances inspired by network information theory," *IEEE Signal Processing Magazine*, vol. 31, pp. 69–79, November 2014.

[2] D. Sakrison, "Source encoding in the presence of random disturbance," *IEEE Transactions on Information Theory*, vol. 14, pp. 165–167, January 1968.

[3] J. Wolf and J. Ziv, "Transmission of noisy information to a noisy receiver with minimum distortion," *IEEE Transactions on Information Theory*, vol. IT-16, pp. 406–411, July 1970.

[4] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall, 1971.

[5] Y. Ephraim and R. Gray, "A unified approach for encoding clean and noisy sources by means of waveform and autoregressive model vector quantization," *IEEE Transactions on Information Theory*, vol. 34, pp. 826–834, July 1988.

[6] M. Gastpar, M. Vetterli, and P. Dragotti, "Sensing reality and communicating bits: A dangerous liaison," *IEEE Signal Processing Magazine*, pp. 70–83, July 2006.

[7] T. Han and K. Kobayashi, "A unified achievable rate region for a general class of multiterminal source coding systems," *IEEE Transactions on Information Theory*, vol. IT-26, pp. 277–288, May 1980.

[8] T. A. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," *IEEE Transactions on Information Theory*, vol. 60, no. 1, pp. 740–761, 2014.

[9] Y. Oohama, "Distributed source coding of correlated gaussian observations," in *International Symposium on Information Theory and its Applications (ISITA 2008)*, (Auckland, New Zealand), pp. 119–, 7-10 December 2008.

[10] Y. Oohama, "Distributed source coding of correlated gaussian remote sources," *IEEE Transactions on Information Theory*, vol. 58, August 2012.

[11] Y. Oohama, "The rate-distortion function for the quadratic gaussian ceo problem," *IEEE Transactions on Information Theory*, vol. 44, pp. 1057–1070, May 1998.

[12] V. Prabhakaran, D. Tse, and K. Ramachandran, "Rate region of the quadratic gaussian ceo problem," in *International Symposium on Information Theory, 2004. ISIT 2004. Proceedings.*, pp. 119–, June 2004.

[13] H. Viswanathan and T. Berger, "The quadratic gaussian ceo problem," *IEEE Transactions on Information Theory*, vol. 43, p. 1549–1559, Sept. 1997.

[14] A. Wagner, S. Tavildar, and P. Viswanath, "Rate region of the quadratic gaussian two-encoder source-coding problem," *IEEE Transactions on Information Theory*, vol. 54, pp. 1938–1961, May 2008.

[15] T. Berger, Z. Zhang, and H. Viswanathan, "The ceo problem [multiterminal source coding]," *IEEE Transactions on Information Theory*, vol. 42, p. 887–902, May 1996.

[16] K. Eswaran and M. Gastpar, "Remote source coding under gaussian noise: Dueling roles of power and entropy power," *arxiv preprint: https://arxiv.org/abs/1805.06515v2*, 2018.

[17] Y. Ugur, I. E. Aguerri, and A. Zaidi, "Vector gaussian ceo problem under logarithmic loss," in *2018 IEEE Information Theory Workshop (ITW)*, pp. 1–5, IEEE, 2018.

[18] Y. Ugur, I. E. Aguerri, and A. Zaidi, "Vector gaussian ceo problem under logarithmic loss and applications," *arXiv preprint arXiv:1811.03933*, 2018.

[19] J. Wang and J. Chen, "On the vector gaussian l-terminal ceo problem," in *2012 IEEE International Symposium on Information Theory Proceedings*, pp. 571–575, July 2012.

[20] J. Chen and J. Wang, "On the vector gaussian ceo problem," in *2011 IEEE International Symposium on Information Theory Proceedings*, pp. 2050–2054, July 2011.

[21] Y. Xu and Q. Wang, "Rate region of the vector gaussian ceo problem with the trace distortion constraint," *IEEE Transactions on Information Theory*, vol. 62, pp. 1823–1835, April 2016.

[22] I. Estella Aguerri and A. Zaidi, "Distributed information bottleneck method for discrete and gaussian sources," in *The International Zurich Seminar on Information and Communication (IZS 2018) Proceedings*, pp. 35–39, ETH Zurich, 2018.

[23] Y. Uğur, I. E. Aguerri, and A. Zaidi, "A generalization of blahut-arimoto algorithm to compute rate-distortion regions of multiterminal source coding under logarithmic loss," in *2017 IEEE Information Theory Workshop (ITW)*, pp. 349–353, IEEE, 2017.

[24] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *37th Annual Allerton Conference on Communication, Control, and Computing*, pp. 368–377, Sept. 1999.

[25] N. Slonim, *The Information Bottleneck Theory and Applications*. PhD thesis, Hebrew University of Jerusalem, Jan. 2002.

[26] S. Hassanpour, D. Wuebben, and A. Dekorsy, "Overview and investigation of algorithms for the information bottleneck method," in *proceedings: SCC*, 2017.

[27] S. Hassanpour, D. Wübben, A. Dekorsy, and B. Kurkoski, "On the relation between the asymptotic performance of different algorithms for information bottleneck framework," in *IEEE International Conference on Communications (ICC)*, (Paris, France), May 2017.

[28] M. Meidlinger, A. Winkelbauer, and G. Matz, "On the relation between the gaussian information bottleneck and mse-optimal rate-distortion quantization," in *2014 IEEE Workshop on Statistical Signal Processing (SSP)*, pp. 89–92, IEEE, 2014.

[29] T. Gedeon, A. E. Parker, and A. G. Dimitrov, "The mathematical structure of information bottleneck methods," *Entropy*, vol. 14, no. 3, pp. 456–479, 2012.

[30] G. Zeitler, "Low-precision analog-to-digital conversion and mutual information in channels with memory," in *Proceedings 48th Annual Allerton Conference on Communication, Control and Computing*, pp. 745–752, Sept. 2010.

[31] J. Lewandowsky and G. Bauch, "Trellis based node operations for LDPC decoders from the Information Bottleneck method," in *2015 9th International Conference on Signal Processing and Communication Systems (ICSPCS)*, pp. 1–10, IEEE, Jan. 2015.

[32] M. Meidlinger and G. Matz, "On irregular LDPC codes with quantized message passing decoding," in *2017 IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC): 3-6 July 2017*, (Piscataway, NJ), pp. 1–5, IEEE, Jan. 2017.

[33] F. Romero and B. Kurkoski, "LDPC decoding mappings that maximize mutual information," *IEEE Journal on Selected Areas in Communications*, vol. 34, pp. 2391–2401, September 2016.

[34] G. Zeitler, *Low-Precision Quantizer Design for Communication Problems*. Dissertation, Technische Universitaet Muenchen, Muenchen, 2012.

[35] D. Chen and V. Kuehn, "Alternating information bottleneck optimization for the compression in the uplink of c-ran," in *2016 IEEE International Conference on Communications (ICC)*, pp. 1–7, May 2016.

[36] D. Chen and V. Kuehn, "Alternating information bottleneck optimization for weighted sum rate and resource allocation in the uplink of c-ran," in *WSA 2016; 20th International ITG Workshop on Smart Antennas*, pp. 1–7, March 2016.

[37] S. Movaghati and M. Ardakani, "Distributed channel-aware quantization based on maximum mutual information," *International Journal of Distributed Sensor Networks*, vol. 12, no. 5, p. 3595389, 2016.

[38] S. Hassanpour, D. Wübben, and A. Dekorsy, "On the equivalence of two information bottleneck-based routines devised for joint source-channel coding," in *25th Int. Conference on Telecommunication (ICT 2018)*, (Saint-Malo, France), Jun 2018.

[39] S. Hassanpour, D. Wübben, and A. Dekorsy, "A novel approach to distributed quantization via multivariate information bottleneck method," in *IEEE Global Communications Conference (GLOBECOM 2019)*, (Waikoloa, HI, USA), Dec 2019.

[40] C. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, p. 379–423, July 1948.

[41] C. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *Institute of Radio Engineers, International Convention Record*, vol. 7, p. 142–163, 1959.

[42] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley & Sons, second ed., 2006.

[43] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 14–20, 1972.

[44] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, 1972.

[45] T. A. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," in *2012 IEEE International Symposium on Information Theory Proceedings*, pp. 761–765, July 2012.

[46] Jun Chen, Xin Zhang, T. Berger, and S. B. Wicker, "An upper bound on the sum-rate distortion function and its corresponding rate allocation schemes for the ceo problem," *IEEE Journal on Selected Areas in Communications*, vol. 22, pp. 977–987, Aug 2004.

[47] J. Chen and T. Berger, "Successive wyner–ziv coding scheme and its application to the quadratic gaussian ceo problem," *IEEE Transactions on Information Theory*, vol. 54, pp. 1586–1603, April 2008.

[48] S. Steiner and V. Kuehn, "Optimization of distributed quantizers using an alternating information bottleneck approach," in *WSA 2019; 23rd International ITG Workshop on Smart Antennas*, pp. 1–6, April 2019.