Received 6 May 2021; accepted 20 May 2021. Date of publication 27 May 2021; date of current version 9 June 2021.

Digital Object Identifier 10.1109/OJCOMS.2021.3083569

Reduced-Complexity Optimization of Distributed Quantization Using the Information Bottleneck Principle

STEFFEN STEINER [1] (Graduate Student Member, IEEE), VOLKER KUEHN [1] (Member, IEEE), MAXIMILIAN STARK [2], AND GERHARD BAUCH [2] (Fellow, IEEE)

[1] Institute of Communications Engineering, University of Rostock, 18119 Rostock, Germany
[2] Institute of Communications, Hamburg University of Technology, 21073 Hamburg, Germany
CORRESPONDING AUTHOR: S. STEINER (e-mail: steffen.steiner@uni-rostock.de)

This work was supported in part by the University of Rostock, and in part by the German Research Foundation (DFG) in the funding programme Open Access Publishing under Grant 325496636.

ABSTRACT This paper addresses the optimization of distributed compression in a sensor network. A direct communication among the sensors is not possible so that noisy measurements of a single relevant signal 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 Executive Officer (CEO) problem and represents a long-standing problem in information theory. In recent years significant progress has been achieved and the rate region has been completely characterized for specific distributions of involved processes and distortion measures. While algorithmic solutions of the CEO problem are principally known, their practical implementation quickly becomes challenging due to complexity reasons. In this contribution, an efficient greedy algorithm to determine feasible solutions of the CEO problem is derived using the information bottleneck (IB) approach. Following the Wyner-Ziv coding principle, the quantizers are successively designed using already optimized quantizer mappings as side-information. However, processing this side-information in the optimization algorithm becomes a major bottleneck because the memory complexity grows exponentially with number of sensors. Therefore, a sequential compression scheme leading to a compact representation of the side-information and ensuring moderate memory requirements even for larger networks is introduced. This internal compression is optimized again by means of the IB method. Numerical results demonstrate that the overall loss in terms of relevant mutual information can be made sufficiently small even with a significant compression of the side-information. The performance is compared to separately optimized quantizers and a centralized quantization. Moreover, the influence of the optimization order for asymmetric scenarios is discussed.

INDEX TERMS Chief executive officer, distributed compression, distributed source coding, information bottleneck.

I. INTRODUCTION

DISTRIBUTED sensing plays an important role in many areas. Smart environments, cities or homes shall improve safety and comfort. Environmental monitoring systems employ many sensors and fuse measurements to infer relevant information. Narrow-band signaling and low transmit powers often limit the available data rates and require the distributed sensors to compress their data before forwarding it to a common receiver. This leads to the well known distributed source coding scenario.

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 [1]. It discusses exemplary sensor networks and the separation of source and channel coding as well as digital versus analog sensing and transmission. The authors conclude 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 sensors.

This contribution considers a generic distributed sensing system whose optimization is known as the Chief Executive Officer (CEO) problem. More precisely, a single source is remotely sensed by multiple sensors. A direct communication among these sensors is not possible. Therefore, they compress their noisy observations locally before forwarding them to a common receiver over capacity limited links.

Many results exist for the quadratic Gaussian CEO problem considering jointly Gaussian signals and the mean squared error (MSE) distortion measure [2]–[5]. In [2], an asymptotic version of the sum-rate distortion function is analytically derived when the number of encoders goes to infinity. Moreover, the MSE distortion was shown to decrease asymptotically with the reciprocal sum-rate R for non-cooperating encoders while it decays exponentially (2^{-2R}) for cooperating sensors [4]. The non-asymptotic case was first investigated in [6] and the complete rate region of the quadratic Gaussian CEO problem is characterized independently by Prabhakaran *et al.* [3] and Oohama [7]. The CEO problem for a multivariate Gaussian relevant process under logarithmic loss distortion measure has been studied in [8], [9]. For more details about outer bounds and rate regions for this scenario the reader may refer to [10]–[12].

For arbitrary discrete source distributions and the logarithmic loss distortion measure, Courtade and Weissman completely characterized the CEO rate region in [13]. Moreover, asymptotic analyses for an infinite number of sensors have been performed in [14], [15]. Berger *et al.* investigated the error rate performance for a discrete source with the Hamming distance as a distortion measure and exhibited an inevitable loss due to non-cooperating sensors [14]. A scaling law on the sum-rate distortion function for arbitrary distortion measures has been derived in [15]. Despite of the large recent progress, rate regions have been characterized only for particular distributions of the relevant signal and specific distortion measures and are generally still unknown.

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 or quantization [16], [17]. In principle, a compromise between the preservation of relevant information and the compression is targeted. This trade-off can be controlled by a Lagrangian optimization approach leading to a non-convex optimization problem for which several IB algorithms exist [16]–[19]. 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 problem [20]. Meanwhile, a rich set of IB applications can be found in communications [21]–[26].

Algorithmic Solutions: Algorithmic approaches to solve the CEO problem have been introduced in [9], [27], [28]. Greedy algorithms are known to determine extreme points of the solution space and their convex hull include also intermediate rate tuples. Most solutions incorporate a modified Blahut-Arimoto algorithm which can be efficiently implemented and exhibits a good convergence behavior, but other solvers can be applied as well. As the algorithmic complexity grows exponentially with the number of sensors, solving the CEO problem for large networks becomes challenging. First, the greedy algorithm needs to consider all optimization orders to obtain the extreme points of the solution space.[1] Second, the dimensionality of the involved probability mass functions (pmfs) depends on the number of sensors and the required memory for storing them grows exponentially. Therefore, this paper proposes a novel approach to significantly reduce the memory requirements during the optimization. The improvement is achieved by compressing the side-information by means of the information bottleneck approach. Particularly, a direct and a sequential compression strategy are derived and evaluated. In this way the optimization can be performed even for larger networks.

Structure and Notation: Section II provides a brief overview on rate distortion theory and the IB method and defines the basic notation. Afterwards, Section III introduces the setup for distributed compression including the definitions of outer and inner bounds on the CEO rate region. The main contribution of this paper is presented in Section IV. First an algorithm to solve the CEO optimization problem allowing individual rate adjustments is introduced in Section IV-A. This approach is further optimized to reduce its memory requirements in Sections IV-B and IV-C. Section V presents numerical results and Section VI concludes this paper.

The following notation is used. Random variables are denoted by calligraphic letters $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$, their realizations x, y, z are elements of the sets $\mathbb{X}, \mathbb{Y}, \mathbb{Z}$ with cardinalities $|\mathbb{X}|$, $|\mathbb{Y}|$ and $|\mathbb{Z}|$, respectively. Vectors are denoted in bold letters $\mathbf{y} = [y_1 \ \dots \ y_M]^{\mathrm{T}}$ and multivariate random variables in boldface calligraphic letters $\boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{Z}}$, with $\boldsymbol{\mathcal{Z}}_{<m}$ covering only the processes $\mathcal{Z}_1$ to $\mathcal{Z}_{m-1}$. The terms $p(y|x)$, $p(x, y)$ and $I(\mathcal{X}; \mathcal{Y})$ represent conditional and joint probability density functions (pdfs) (or pmfs for discrete random variables) and the mutual information between $\mathcal{X}$ and $\mathcal{Y}$, respectively. Finally, $\mathbb{E}_{\mathcal{X}}\big[f(\mathcal{X})\big]$ represents the expectation of a function $f(x)$ with respect to the random variable $\mathcal{X}$.

II. RATE DISTORTION THEORY AND IB PRINCIPLE

A. RATE DISTORTION THEORY

Rate distortion theory goes back to the seminal work of Shannon [29], [30]. While the entropy $H(\mathcal{Y})$ of a discrete

1. For symmetric scenarios with identical signal-to-noise-ratios (SNRs), link capacities and cluster cardinalities, the optimization order is not important and complexity can be significantly reduced.

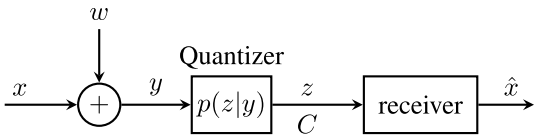

FIGURE 1. Illustration of noisy source coding (remote sensing) for a single sensor.

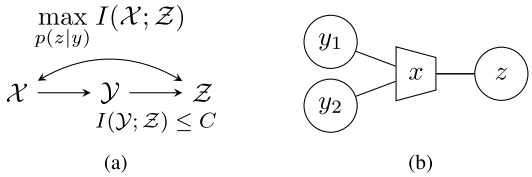

FIGURE 2. (a) Illustration of the IB setup, (b) Exemplary IB graph.

random process $\mathcal{Y}$ represents the ultimate lower bound for lossless compression, a lossy compression of $\mathcal{Y}$ to $\mathcal{Z}$ leads inevitably to a distortion of $\mathcal{Y}$. As opposed to clean sources, the measurements are often noisy which is widely known as remote sensing or noisy source coding [31]–[34]. The system model for this scenario is depicted in Fig. 1. Here, the quantizer input y is a noisy observation of a relevant signal x and the distortion is measured between x and z. The best trade-off between compression rate $I(\mathcal{Y};\mathcal{Z})$ and the average distortion $\mathbb{E}_{\mathcal{X},\mathcal{Z}}[\tilde{d}(x,z)]$ is defined by the distortion rate function

$$D(R) = \min_{p(z|y):\ I(\mathcal{Y};\mathcal{Z})\leq R} \mathbb{E}_{\mathcal{X},\mathcal{Z}}\Big[\tilde{d}(x,z)\Big]. \tag{1}$$

In (1), R is an upper bound on the compression rate. Generally, the mapping of y onto z denoted by the pmf $p(z|y)$ is stochastic, i.e., $p(z|y) \in [0,1]$ holds. Using the method of Lagrangian multipliers, the optimization problem can be rewritten to

$$p(z|y) = \operatorname*{argmin}_{\tilde{p}(z|y)} \mathbb{E}_{\mathcal{X},\mathcal{Z}}\Big[\tilde{d}(x,z)\Big] + \beta I(\mathcal{Y};\mathcal{Z}), \tag{2}$$

leading to the implicit solution[2] [33]

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

with

$$\begin{aligned} d_\beta(y,z) &= \frac{1}{\beta}\sum_x p(x|y)\cdot\tilde{d}(x,z) - \log p(z) \\ &= \frac{1}{\beta}\mathbb{E}_{\mathcal{X}|y}\Big[\tilde{d}(x,z)\Big] - \log p(z). \end{aligned} \tag{4}$$

The implicit equation in (3) can be efficiently solved by an iterative Blahut-Arimoto algorithm [35]–[37]. In (2), the Lagrange multiplier β serves as a trade-off parameter between distortion and compression.[3] The distortion $\tilde{d}(x,z)$ can be measured by different means. Its choice influences the mathematical nature of the optimization problem, i.e., whether it is convex or not. Hamming distance, squared Euclidean distance or the logarithmic loss function $\tilde{d}(x,z) = -\log p(x|z)$ are just some examples.

2. The distribution $p(z)$ also depends on the mapping $p(z|y)$.

3. In contrast to the most publications, the Lagrange multiplier β is placed in (2) in front of the compression rate $I(\mathcal{Y};\mathcal{Z})$ instead of the distortion measure $\mathbb{E}_{\mathcal{X},\mathcal{Z}}[\tilde{d}(x,z)]$ leading to the factor $\frac{1}{\beta}$ in the exponent defined in (4). This ensures a consistent notation when targeting distributed scenarios.

B. INFORMATION BOTTLENECK METHOD

The information bottleneck method is a clustering framework pairing concepts from machine learning and information theory [17]. For the general setup depicted in Fig. 1, the IB approach aims to optimize the mapping $p(z|y)$ such that a maximal relevant mutual information $I(\mathcal{X};\mathcal{Z})$ is preserved while the rate constraint $I(\mathcal{Y};\mathcal{Z}) \leq C$ is fulfilled, see also Fig. 2(a). According to [16], this problem can be formulated as a Lagrangian maximization problem

$$p(z|y) = \operatorname*{argmax}_{\tilde{p}(z|y)} I(\mathcal{X};\mathcal{Z}) - \beta I(\mathcal{Y};\mathcal{Z}) \tag{5}$$

$$= \operatorname*{argmin}_{\tilde{p}(z|y)} H(\mathcal{X}|\mathcal{Z}) + \beta I(\mathcal{Y};\mathcal{Z}). \tag{6}$$

The equality in (6) holds due to $I(\mathcal{X};\mathcal{Z}) = H(\mathcal{X}) - H(\mathcal{X}|\mathcal{Z})$ and the fact that $H(\mathcal{X})$ can be skipped because it does not depend on the mapping $p(z|y)$. The comparison of (5) and (6) with (2) reveals that the IB approach is a special formulation of the remote sensing problem using the logarithmic loss function $\tilde{d}(x,z) = -\log p(x|z)$ as a distortion measure whose expectation is $H(\mathcal{X}|\mathcal{Z}) = \mathbb{E}_{\mathcal{X},\mathcal{Z}}\big[-\log p(x|z)\big]$ [13], [20]. In this case, distortion minimization means maximization of the relevant mutual information $I(\mathcal{X};\mathcal{Z})$ for given $H(\mathcal{X})$. Since $I(\mathcal{X};\mathcal{Z})$ and $I(\mathcal{Y};\mathcal{Z})$ are convex functions of the mapping $p(z|y)$, (5) is a non-convex optimization problem having the implicit solution in (3), now with the exponent

$$\begin{aligned} d_\beta(y,z) &= \frac{1}{\beta} D_{\mathrm{KL}}\big[p(x|y)\|p(x|z)\big] - \log p(z) \\ &= \frac{1}{\beta}\mathbb{E}_{\mathcal{X}|y}\left[\log\frac{p(x|y)}{p(x|z)}\right] - \log p(z). \end{aligned} \tag{7}$$

In (7), $D_{\mathrm{KL}}\big[p(x|y)\|p(x|z)\big]$ denotes the Kullback-Leibler divergence. Again, the implicit solution can be efficiently solved by an iterative Blahut-Arimoto like algorithm. It has to be mentioned that the implementation of the algorithm generally requires a discretization of y with an appropriate fine resolution.

Equivalent to the remote sensing problem, the Lagrange multiplier β in (5) serves as a trade-off parameter between preservation of relevant mutual information and compression. For $\beta = 0$, the optimization solely focuses on the preservation of relevant mutual information and yields a deterministic clustering $p(z|y) \in \{0,1\}$. This is very attractive from a practical perspective because it allows an implementation by static lookup tables [17]. In this case, the compression is obtained by choosing $|\mathbb{Z}| < |\mathbb{Y}|$. For $\beta > 0$, the clustering $p(z|y) \in [0,1]$ is generally stochastic. For $\beta \to \infty$, no relevant mutual information is preserved and all values of y are

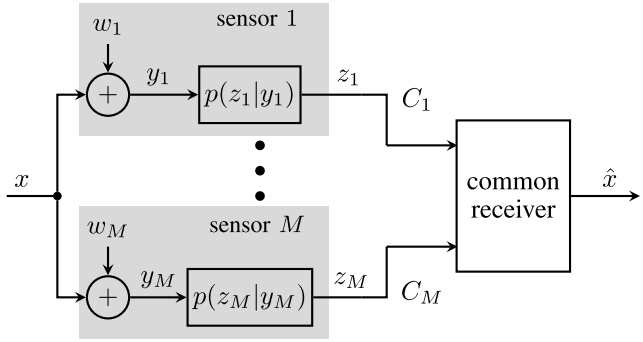

FIGURE 3. Distributed sensor system with M sensors, a common receiver and individual link capacities C_m.

mapped onto a single cluster index z. As the compression-rate curve is a monotonic increasing function, a simple bi-section search allows to adjust β such that a desired rate constraint $I(\mathcal{Y};\mathcal{Z}) \leq C$ is fulfilled. This technique will be applied in Section IV to obtain feasible solutions of the CEO problem. For a detailed review of IB algorithms the reader is referred to [17].

When discretizing signal processing algorithms, IB optimized lookup tables can be used to replace arithmetic operations [21]. IB graphs [38] are a graphical tool to visualize the decomposition of the algorithm and are exemplarily illustrated in Fig. 2(b). The relevant random variable is written inside the trapezoid which represents the IB compression. Closely related to factor graphs, IB graphs denote the involved random variables, i.e., the observations y_1, y_2 and cluster index z as variable nodes expressed by circles. This paper will make use of IB graphs in Section IV-B to derive a simplified architecture to efficiently compress the side-information used in the Greedy Distributed IB algorithm.

III. DISTRIBUTED SENSING SYSTEM

A. SYSTEM MODEL

Fig. 3 illustrates the considered system model. It consists of M sensors each observing a noisy version $y_m = x + w_m$ of the relevant signal x. The noise processes $\mathcal{W}_m$ at the sensors are assumed to be statistically independent yielding

$$p(\mathbf{y}|x) = \prod_{m=1}^{M} p(y_m|x). \tag{8}$$

As all processes have a zero mean, the measurement signal-to-noise-ratio (SNR) is given by $\gamma_m = \frac{\sigma_x^2}{\sigma_{w_m}^2}$, with σ_x^2, $\sigma_{w_m}^2$ denoting signal and noise variances, respectively. The noisy observation y_m of sensor m is quantized using the mapping $p(z_m|y_m)$. The sensors transmit compressed versions of the cluster indexes $z_1, \ldots, z_M$ to a common receiver over capacity limited links with capacities $C_1, \ldots, C_M$.

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

B. THE CHIEF EXECUTIVE OFFICER PROBLEM

Optimizing the distributed compression for a scenario depicted in Fig. 3 is termed the CEO problem [14]. In the literature, the two most popular distortion measures are the mean squared error and the logarithmic loss function. For the logarithmic loss function, the inner bound of the CEO rate region with M sensors is defined in [13] as

$$I\big(\boldsymbol{\mathcal{Y}}_{\mathbb{S}};\boldsymbol{\mathcal{Z}}_{\mathbb{S}}|\boldsymbol{\mathcal{Z}}_{\overline{\mathbb{S}}},\mathcal{Q}\big) \leq \sum_{m\in\mathbb{S}} C_m \quad \forall \quad \mathbb{S}\subseteq\{1,2,\ldots,M\} \tag{9}$$

$$H(\mathcal{X}|\boldsymbol{\mathcal{Z}},\mathcal{Q}) \leq D. \tag{10}$$

The corresponding outer bound is given by

$$\left[\sum_{m\in\mathbb{S}} I(\mathcal{Y}_m;\mathcal{Z}_m|\mathcal{X},\mathcal{Q}) + H\big(\mathcal{X}|\boldsymbol{\mathcal{Z}}_{\overline{\mathbb{S}}},\mathcal{Q}\big) - D\right]^+ \leq \sum_{m\in\mathbb{S}} C_m \;\forall\quad \mathbb{S}\subseteq\{1,2,\ldots,M\} \tag{11}$$

$$H(\mathcal{X}|\boldsymbol{\mathcal{Z}},\mathcal{Q}) \leq D \tag{12}$$

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

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

and it was shown in [13] that they match for this particular distortion measure. The rate constraints in (9) and (11) must hold for any subset $\mathbb{S}\subseteq\{1,2,\ldots,M\}$. The maximum permitted distortion is defined by the parameter D. Constructing the convex hull over all particular solutions by means of time-sharing via parameter $\mathcal{Q}$ reveals the complete CEO rate region for the log-loss distortion measure.

IV. ALGORITHMIC SOLUTIONS

A. GREEDY DISTRIBUTED INFORMATION BOTTLENECK APPROACH

This section presents the Greedy Distributed Information Bottleneck (GDIB) algorithm to determine feasible solutions of the CEO problem [39]. As it is based on the inner bound defined in (9) and (10), an obtained region does not represent the complete rate region but may be strictly smaller. Moreover, time sharing is generally required to determine the best Wyner-Ziv coding strategy [40], [41]. For notational simplicity, $\mathcal{Q}$ is omitted in subsequent equations.

As mentioned above, the expectation of the logarithmic loss function $H(\mathcal{X}|\boldsymbol{\mathcal{Z}})$ can be replaced by the relevant mutual information $I(\mathcal{X};\boldsymbol{\mathcal{Z}})$. Hence, the optimization goal is to maximize the relevant mutual information $I(\mathcal{X};\boldsymbol{\mathcal{Z}})$ subject to the rate constraints defined in (9). Having this in mind,

the optimization problem based on the inner bound can be reformulated to

$$\max_{\mathbf{P}} \; I(\mathcal{X};\boldsymbol{\mathcal{Z}}) \text{ s.t. } I(\boldsymbol{\mathcal{Y}}_{\mathbb{S}};\boldsymbol{\mathcal{Z}}_{\mathbb{S}}|\boldsymbol{\mathcal{Z}}_{\overline{\mathbb{S}}}) \leq \sum_{m\in\mathbb{S}} C_m \quad \forall \;\; \mathbb{S}\subseteq\{1,2,\ldots,M\} \tag{13}$$

with $\mathbf{P} = [p(z_1|y_1) \;\cdots\; p(z_M|y_M)]$ being the set of all mappings. According to [13], the compression rates $I(\boldsymbol{\mathcal{Y}}_{\mathbb{S}};\boldsymbol{\mathcal{Z}}_{\mathbb{S}}|\boldsymbol{\mathcal{Z}}_{\overline{\mathbb{S}}})$ are supermodular set functions w.r.t. the sets $\mathbb{S}$ [42], while the relevant information $I(\mathcal{X};\boldsymbol{\mathcal{Z}})$ does not depend on $\mathbb{S}$. Therefore, the extreme points of the solution space of (13) can be determined by a greedy approach and the optimization problem becomes

$$\max_{\mathbf{P},\pi} \; I(\mathcal{X};\boldsymbol{\mathcal{Z}}) \;\; \text{s.t. } I\big(\mathcal{Y}_{\pi(m)};\mathcal{Z}_{\pi(m)}|\boldsymbol{\mathcal{Z}}_{<\pi(m)}\big) \leq C_{\pi(m)} \quad \forall m\in\{1,2,\ldots,M\} \tag{14}$$

where $\pi(\cdot)$ denotes the optimization order. For each permutation, the obtained solution represents a specific vertex of the supermodular polyhedron and corresponds to a certain Wyner-Ziv coding strategy. As the best optimization order $\pi^*(\cdot)$ is not known in advance, (14) has to be solved for all permutations $\pi(\cdot)$ of the set $\{1,\ldots,M\}$. This increases the computational complexity exponentially with M. For the sake of notational simplicity, the permutations are skipped in subsequent equations. Using the method of Lagrangian multipliers and applying the chain rule to $I(\mathcal{X};\boldsymbol{\mathcal{Z}})$, the optimization problem in (14) can be expressed by the target function

$$\begin{aligned} L_{\text{GDIB}} &= I(\mathcal{X};\boldsymbol{\mathcal{Z}}) - \sum_{m=1}^{M} \beta_m \cdot I(\mathcal{Y}_m;\mathcal{Z}_m|\boldsymbol{\mathcal{Z}}_{<m}) \\ &= \sum_{m=1}^{M} I(\mathcal{X};\mathcal{Z}_m|\boldsymbol{\mathcal{Z}}_{<m}) - \beta_m \cdot I(\mathcal{Y}_m;\mathcal{Z}_m|\boldsymbol{\mathcal{Z}}_{<m}) \end{aligned} \tag{15}$$

which has to be maximized. Pursuing the greedy optimization approach, the objective in (15) can be decomposed into M target functions being successively maximized.

$$L_{\text{GDIB}}^{(1)} = I(\mathcal{X};\mathcal{Z}_1) - \beta_1 I(\mathcal{Y}_1;\mathcal{Z}_1) \tag{16a}$$

$$\vdots$$

$$L_{\text{GDIB}}^{(M)} = I(\mathcal{X};\mathcal{Z}_M|\boldsymbol{\mathcal{Z}}_{<M}) - \beta_M I(\mathcal{Y}_M;\mathcal{Z}_M|\boldsymbol{\mathcal{Z}}_{<M}) \tag{16b}$$

The first sensor is optimized by maximizing (16a) without side-information of other sensors. Afterwards, its mapping $p(z_1|y_1)$ serves as side-information for the optimization of the second sensor following the Wyner-Ziv coding strategy. The last sensor M exploits full side-information of all other sensors as indicated in (16b). The Lagrange multipliers β_m have to be chosen such that the individual rate constraints in (14) are fulfilled. The objectives in (16a)–(16b) can be solved by equating the derivative w.r.t. the mapping $p(z_m|y_m)$ to zero. The complete derivation for a particular sensor m is given in the Appendix and delivers the update rule

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

with the exponent

$$d_{\beta_m}(y_m,z_m) \coloneqq \mathbb{E}_{\boldsymbol{\mathcal{Z}}_{<m}|y_m}\Bigg[\frac{1}{\beta_m}\cdot D_{\text{KL}}\big[p(x|y_m,\mathbf{z}_{<m})\|p\big(x|\mathbf{z}_{\leq m}\big)\big] - \log p(z_m|\mathbf{z}_{<m})\Bigg]. \tag{18}$$

As described in Section II, this implicit[4] equation can be solved using an extension of the Blahut-Arimoto algorithm.

B. ONE-STEP COMPRESSION OF SIDE-INFORMATION

The derivations of the GDIB algorithm in Section IV-A hold for any network size. However, a closer look at the definition of $d_{\beta_m}(y_m,z_m)$ in (18) reveals that the dimensionality of the involved pmfs depends on the number of sensors in the network. Since a linear increase in the number of sensors, i.e., the dimensions of $\boldsymbol{\mathcal{Z}}_{<m}$, leads to an exponential increase in the number of elements in the representing data structure, large networks may cause memory problems especially when optimizing the last sensors. For a network with $M = 10$ sensors and cardinalities $|\mathbb{X}| = 4$, $|\mathbb{Y}_m| = 64$ and $|\mathbb{Z}_m| = 8\;\forall m$, the tuple $(y_M,\mathbf{z}_{<M})$ for the last sensor has $64\cdot 8^9 = 8.59\cdot 10^9$ elements. Thus, using 8 byte for double precision, it needs 256 GiB to store $p(x|y_M,\mathbf{z}_{<M})$.

To overcome this limitation for large networks, the strategy is to compress the high dimensional $\boldsymbol{\mathcal{Z}}_{<m}$ onto a single dimensional $\mathcal{Z}_{<m}^*$ of appropriate cardinality $|\mathbb{Z}_{<m}^*|$ using the IB method. Since $\mathcal{Z}_{<m}^*$ shall be nearly as informative about the relevant signal $\mathcal{X}$ as $\boldsymbol{\mathcal{Z}}_{<m}$, it can replace $\boldsymbol{\mathcal{Z}}_{<m}$ in the definition of $d_{\beta_m}(y_m,z_m)$. In turn, (18) can be rewritten as

$$d_{\beta_m}^{\text{c}}(y_m,z_m) \coloneqq \mathbb{E}_{\mathcal{Z}_{<m}^*|y_m}\Bigg[\frac{1}{\beta_m}\cdot D_{\text{KL}}\big[p\big(x|y_m,z_{<m}^*\big)\|p\big(x|z_m,z_{<m}^*\big)\big] - \log p\big(z_m|z_{<m}^*\big)\Bigg]. \tag{19}$$

The pmfs in (19) require much less memory than those in (18) as will be demonstrated in Section V-F.

The extended Blahut-Arimoto algorithm optimizing the quantizer of a particular sensor m with compressed side-information is given in Algorithm 1. The input pmf $p(z_{<m}^*,x)$ is determined in advance by one of the proposed compression schemes discussed in the next paragraphs. The KL divergence in (19) is determined in lines 6 to 10 by means of the joint pmf $p(z_{<m}^*,z_m,y_m,x)$ calculated in lines 3 to 5. The statistical distance $d_{\beta_m}^{\text{c}}(z_m,y_m)$ computed in lines 11 to 17 is used to update the quantizer mapping $p(z_m|y_m)$. The algorithm stops when a stopping criterion like the Jensen–Shannon divergence [43] between the quantizer mappings of successive iterations falls below a predefined threshold ϵ.

The compression of side-information $\mathbf{z}_{<m}$ to $z_{<m}^*$ is performed using the IB method with $\beta = 0$ to obtain

4. $p(z_m|\mathbf{z}_{<m})$ and $p(x|\mathbf{z}_{\leq m})$ depend on the mapping $p(z_m|y_m)$.

Algorithm 1: Blahut-Arimoto-Like Algorithm With Compressed Side-Information

input : $m, p(y_m, x), p^{\text{init}}(z_m|y_m), \beta_m, \epsilon, p(z^*_{<m}, x)$
output : $p(z_m|y_m)$

1 **begin**
initialization:
$p(z_m|y_m)^{(0)} \leftarrow p^{\text{init}}(z_m|y_m)$,
$l \leftarrow 1$
2 **do**
3 // calculate $p(z^*_{<m}, z_m, y_m, x)$
4 $p(z^*_{<m}, z_m, y_m, x) = p(z^*_{<m}|x)p(z_m|y_m)p(y_m, x)$
5 $p(z^*_{<m}, y_m, x) = \sum_{z_m} p(z^*_{<m}, z_m, y_m, x)$
6 // KL-Divergence D_{KL} of (19)
7 $p(x|z^*_{<m}, y_m) = p(z^*_{<m}, y_m, x)/p(z^*_{<m}, y_m)$
8 $p(z^*_{<m}, z_m, x) = \sum_{y_m} p(z^*_{<m}, z_m, y_m, x)$
9 $p(z^*_{<m}, z_m) = \sum_{x} p(z^*_{<m}, z_m, x)$
10 $p(x|z^*_{<m}, z_m) = p(z^*_{<m}, z_m, x)/p(z^*_{<m}, z_m)$
11 $D_{\text{KL}} = \sum_x p(x|z^*_{<m}, y_m) \cdot \log \frac{p(x|z^*_{<m}, y_m)}{p(x|z^*_{<m}, z_m)}$
12 // distance $d^{\text{c}}_{\beta_m}(z_m, y_m)$ (19)
13 $p(z^*_{<m}, y_m) = \sum_x p(z^*_{<m}, y_m, x)$
14 $p(z^*_{<m}|y_m) = p(z^*_{<m}, y_m)/p(y_m)$
15 $p(z_m|z^*_{<m}) = p(z^*_{<m}, z_m)/p(z^*_{<m})$
16 $d^{\text{c}}_{\beta_m}(z_m, y_m) =$
$\sum_{z^*_{<m}} p(z^*_{<m}|y_m) \cdot \left[\frac{1}{\beta_m} D_{\text{KL}} - \log p(z_m|z^*_{<m}) \right]$
17 // update quantizer $p(z_m|y_m)$
18 $p(z_m|y_m)^{(l)} = \frac{1}{\sum_z e^{-d^{\text{c}}_{\beta_m}(z_m, y_m)}} e^{-d^{\text{c}}_{\beta_m}(z_m, y_m)}$
19 $l++$
20 **while** $D_{JS}[\, p^{(l)}(z_m|y_m) \,||\, p^{(l-1)}(z_m|y_m)\,] < \epsilon$

deterministic mappings. We have investigated two different implementations, the *one-step compression scheme* and the *sequential compression scheme*.

The one-step compression scheme performs the compression in just a single step by mapping $\mathbf{z}_{<m}$ directly onto $z^*_{<m}$. It is depicted in Figure 4(a) for a network of $M = 6$ sensors. In this case, the side-information $\mathbf{z}_{<6}$ for sensor 6 consists of 5 variables which have to be compressed to $z^*_{<6}$. The corresponding algorithm is given in Algorithm 2. Note that the IB algorithm in line 3 requires the joint pmf $p(\mathbf{z}_{<m}, x)$, which is determined in line 2. However, this joint pmf still depends on the number of sensors in the network, and the one-step compression scheme quickly becomes itself infeasible for larger networks.

C. SEQUENTIAL COMPRESSION SCHEME

A solution to the curse of dimensionality provides a sequential compression strategy using a tree based structure. Here, the IB compression is always performed for two variables at a time starting with z_1 and z_2. The corresponding result $z^*_{<3}$ is obtained by Algorithm 3 and serves as an input for a subsequent joint IB compression step with z_3. Algorithm 4 represents the compression of side-information for all sensors $3 \leq m \leq M$. Repeating this procedure for the remaining variables leads to the tree based structure depicted in Fig. 4(b) for the example of $M = 6$ sensors.

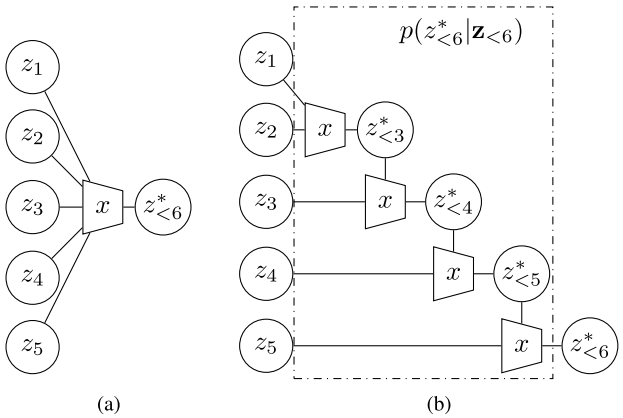

FIGURE 4. (a) one-step compression, (b) sequential compression of side-information $\mathbf{z}_{<6}$ for sensor 6.

Algorithm 2: One-Step Compression of Side-Information

input : $m, p(z_\nu|x)\ \forall \nu < m, p(x)$
output : $p(z^*_{<m}, x)$

1 **begin**
2 $p(\mathbf{z}_{<m}, x) = p(z_1|x) \cdot \ \ldots\ \cdot p(z_{m-1}|x)p(x)$
3 $p(z^*_{<m}, x) \leftarrow \text{IB}(p(\mathbf{z}_{<m}, x), \beta = 0)$

Algorithm 3: Sequential Compression of Side-Information $\mathbf{z}_{<3}$ for Sensor $m = 3$

input : $m, p(z_1|x), p(z_2|x), p(x)$
output : $p(z^*_{<3}, x)$

1 **begin**
2 $p(\mathbf{z}_{<3}, x) = p(z_1|x)p(z_2|x)p(x)$
3 $p(z^*_{<3}, x) \leftarrow \text{IB}(p(\mathbf{z}_{<3}, x), \beta = 0)$

Algorithm 4: Sequential Compression of Side-Information $\mathbf{z}_{<m}$ for Sensor $m > 3$

input : $m, p(z^*_{<m-1}|x), p(z_{m-1}|x), p(x)$
output : $p(z^*_{<m}, x)$

1 **begin**
2 $p(z^*_{<m-1}, z_{m-1}, x) = p(z^*_{<m-1}|x)p(z_{m-1}|x)p(x)$
3 $p(z^*_{<m}, x) \leftarrow \text{IB}(p(z^*_{<m-1}, z_{m-1}, x), \beta = 0)$

V. NUMERICAL RESULTS

A. REFERENCE SYSTEMS

In this section, numerical results for the proposed algorithms solving the CEO problem will be discussed. Their performance will be compared to a hypothetical centralized

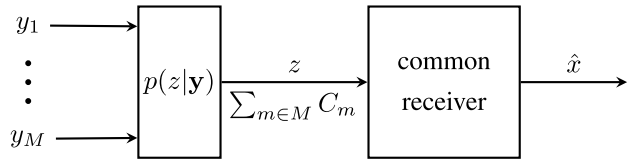

FIGURE 5. Setup of central IB optimized quantization.

quantization depicted in Figure 5. Here, a single quantizer having access to all measurements $\mathbf{y} = [y_1 \dots y_M]^{\mathrm{T}}$ performs the quantization of $\mathbf{y}$ to a scalar index z described by the mapping $p(z|\mathbf{y})$. Technically, the compression is done by assuming maximum ratio combining all samples y_m, which delivers a scalar sufficient statistics $\tilde{y}$ of $\mathbf{y}$ with an SNR $\gamma = \sum_m \gamma_m$. The joint pmf $p(x, \tilde{y})$ is then used to optimize a scalar quantizer by applying the IB principle described in Section II-B. The rate constraint to be met is the sum of all individual link capacities $\sum_{m\in M} C_m$. We denote this approach centralized IB (CIB). Since the communication among sensors is not possible in the original system model, the CIB approach provides a hypothetical upper bound.

Moreover, the independent scalar optimization of the quantizers applying the IB principle is abbreviated by IB and serves as a lower bound. The major difference to distributed compression according to Section IV is that no Wyner-Ziv coding is applied, i.e., no side-information $\mathbf{z}_{<m}$ is exploited in the optimization.

B. INFLUENCE OF NUMBER OF SENSORS

First, a varying number of sensors is considered in a symmetric scenario, i.e., the measurement SNRs γ_m, the link capacities C_m and the quantizer cardinalities $|\mathbb{Z}_m|$ are identical for all sensors $1 \le m \le M$. The measurement noise is Gaussian distributed and the relevant signal is taken from a uniformly distributed 4-ASK (Amplitude Shift Keying) alphabet. The sum-rate $C_{\text{sum}} = \sum_{m=1}^{M} C_m$ was fixed to a value independent of M. Consequently, the more sensors 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 sensors equally share a common medium in an orthogonal way and a round robin fashion.

Figs. 6 and 7 illustrate the influence of the number of sensors onto the relevant mutual information $I(\mathcal{X}; \mathcal{Z})$ for sum-rates $C_{\text{sum}} = 2.5$ bit/s/Hz and $C_{\text{sum}} = 4.0$ bit/s/Hz, respectively. The gray colored area represents the non-achievable region, since $I(\mathcal{X}; \mathcal{Z})$ cannot exceed $I(\mathcal{X}; \mathcal{Y})$ due to the data-processing inequality. For $\gamma_m = 8$ dB, it can be observed that independent scalar optimization of the quantizers (IB) looses relevant information with increasing number of sensors. 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 $\gamma_m = 3$ dB, $M = 3$

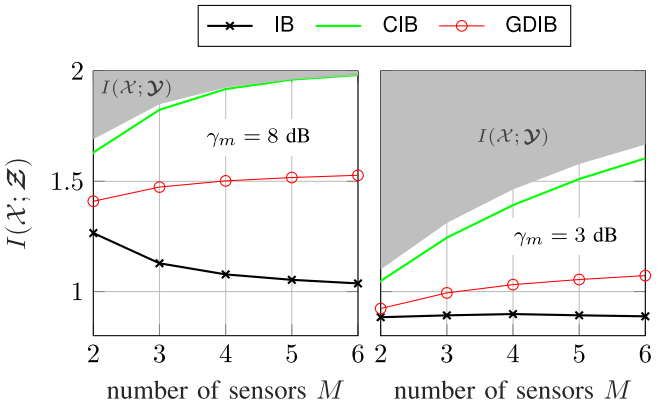

FIGURE 6. Relevant mutual information vs. number of sensors for fixed sum-rate C_{sum} =2.5 bit/s/Hz, $|\mathbb{X}| = 4$, $|\mathbb{Z}_m| = 4\ \forall m$.

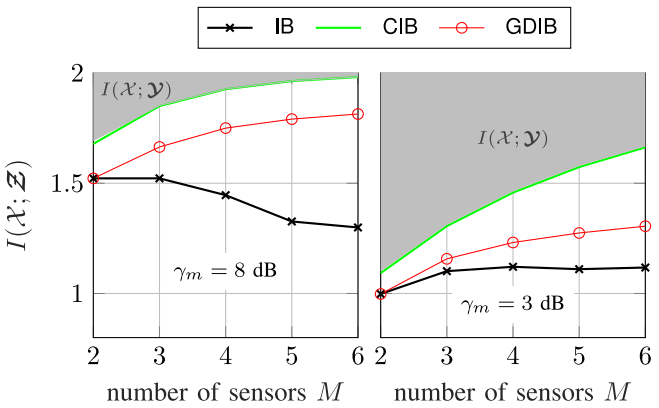

FIGURE 7. Relevant mutual information vs. number of sensors for fixed sum-rate C_{sum} =4.0 bit/s/Hz, $|\mathbb{X}| = 4$, $|\mathbb{Z}_m| = 4\ \forall m$.

and $M = 4$ sensors are slightly superior to only $M = 2$ sensors. However, a larger M leads again to a degradation.

Contrarily, the GDIB approach does not suffer from this effect and $I(\mathcal{X}; \mathcal{Z})$ increases with growing M. This result demonstrates that joint optimization of distributed quantizers leads to a significant gain compared to separately optimized quantization. The CIB approach clearly outperforms both other approaches as expected. It benefits most from an increasing number of sensors while the GDIB approach gains only moderately with growing M. This illustrates the limitation of non-cooperative distributed quantization not allowing an exchange of information between different sensors. A comparison of Figs. 6 and 7 reveals that the described effects are qualitatively similar for different sum-rates.

C. INFLUENCE OF SUM-RATE

Fig. 8 illustrates the relevant information versus the sum-rate for a symmetric scenario. Please note that due to $|\mathbb{Z}_m| = 4$ and $M = 5$, a sum-rate $C_{\text{sum}} \ge 10$ does not require lossy compressions with stochastic mappings $p(z_m|y_m)$ at the sensors. Only in this region the independent scalar IB optimization achieves the same relevant information as the GDIB approach. For smaller sum-rates, it performs worse.

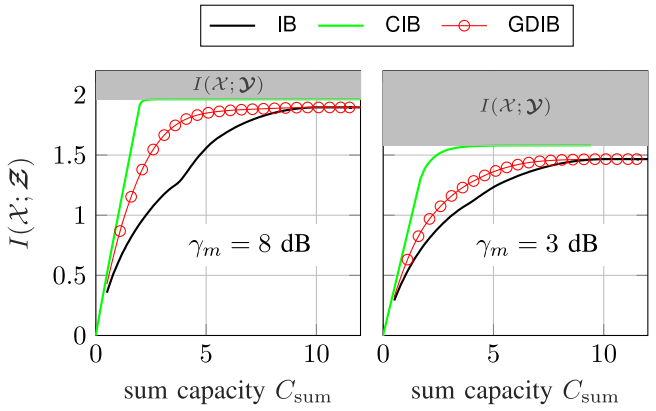

FIGURE 8. Relevant mutual information vs. sum capacity for a symmetric scenario with $M = 5$ sensors, $|\mathbb{X}| = 4$, SNRs $\gamma_m \in \{3, 8\}$ dB and $|\mathbb{Z}_m| = 4$.

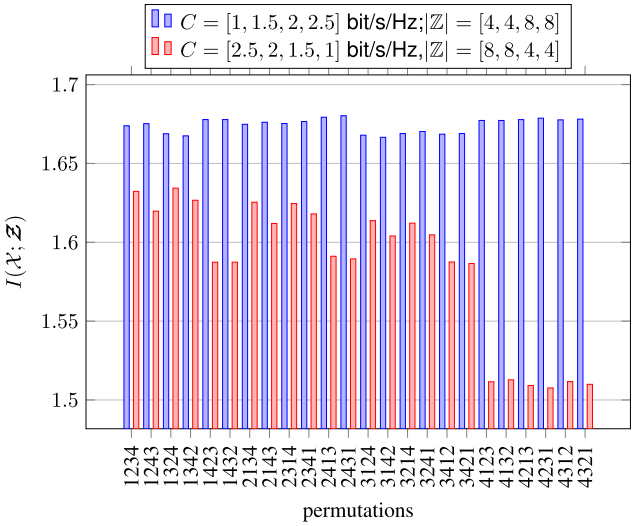

FIGURE 9. Relevant mutual information for non symmetric scenario with $M = 4$ sensors, SNRs $\gamma_m = [2, 4, 6, 8]$ dB and $|\mathbb{X}| = 4$ using GDIB optimization.

Furthermore, the comparison with the centralized quantization shows again the loss of distributed compression which is largest for moderate sum-rates between $2 \leq C_{\text{sum}} \leq 4$ bit/s/Hz. For large sum-rates, the gap between CIB and GDIB illustrates the asymptotic loss of non-cooperative versus fully-cooperative distributed compression.

D. OPTIMIZATION ORDER FOR ASYMMETRIC SCENARIOS

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

Fig. 9 illustrates the relevant information for $M = 4$ sensors and all permutations $\pi(\cdot)$ of the set $\{1, \ldots, 4\}$ using the 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 strategy does 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. 9, it seems that starting with the best measurement SNRs but strongest compression leads to worse performances. Contrarily, best results are achieved when sensors with high SNRs are optimized after those with small SNRs.

E. COMPRESSION OF SIDE-INFORMATION

In this subsection, the compression of side-information (SI) during the quantizer optimization is analyzed. Fig. 10 illustrates the uncompressed side-information $I(\mathcal{X}; \mathcal{Z}_{1:n})$ (NC, dotted lines) as well as one-step compressed (OC, solid lines) and sequentially compressed (SC, dashed lines) side-information $I(\mathcal{X}; \mathcal{Z}^*_{1:n})$ versus the sequential compression steps n when optimizing the last sensor $M = 10$. Note that $I(\mathcal{X}; \mathcal{Z}^*_{1:n})$ just represents the side-information available for the optimization of sensor M, not the overall relevant mutual information $I(\mathcal{X}; \mathcal{Z})$ after optimizing all sensors. Naturally, the side-information is independent of n for NC and OC. Moreover, the loss between compressed and uncompressed side-information becomes smaller for growing cardinality $|\mathbb{Z}^*|$. For the sequential compression scheme, the amount of side-information principally increases with each additional sensor. However, for low cardinalities $|\mathbb{Z}^*|$, the compressed side-information saturates early and a significant gap remains to the one-step compression performance, which serves as an upper bound. This can be explained by an accumulation of compression losses. Increasing the SI's cardinality to only $|\mathbb{Z}^*| = 12$ makes this loss very small for the considered scenario.

F. GDIB PERFORMANCE WITH COMPRESSED SIDE-INFORMATION

Fig. 11 illustrates the influence of compressed side-information on the overall performance. It depicts the relevant mutual information $I(\mathcal{X}; \mathcal{Z})$ versus the number of sensors in the network. The black dashed-dotted curve represents the result for an individual scalar IB optimization of each quantizer without considering the mappings of other sensors, whereas the black dotted curve represents the GDIB approach without compressing the side-information. They serve as lower and upper bounds, respectively. As already discussed in Section V-B, the GDIB approach preserves more relevant mutual information than the scalar IB case. In fact, the GDIB gains with increasing number of sensors and outperforms the scalar IB case for all M even with compressed side-information.

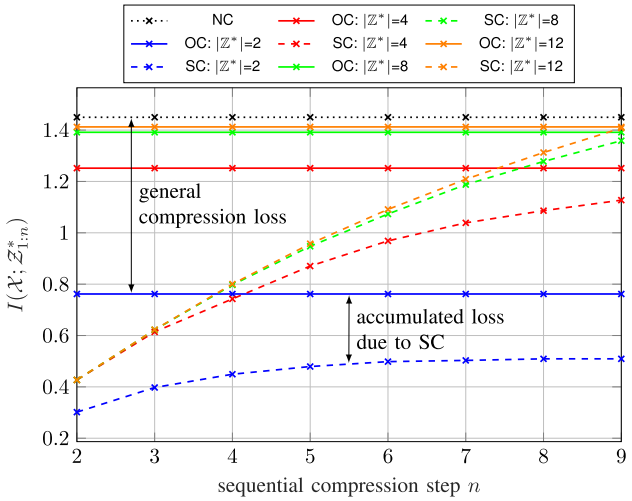

FIGURE 10. Compressed side-information versus sequential compression steps considering the last sensor; symmetric scenario: $M = 10$ sensors, sum-rate C_{sum} =2.5 bit/s/Hz, SNRs $\gamma_m = 8$ dB, $|\mathbb{X}| = 4$, $|\mathbb{Y}_m| = 64$ and $|\mathbb{Z}_m| = 4$; No Compression (NC), One-Step Compression (OC), Sequential Compression (SC).

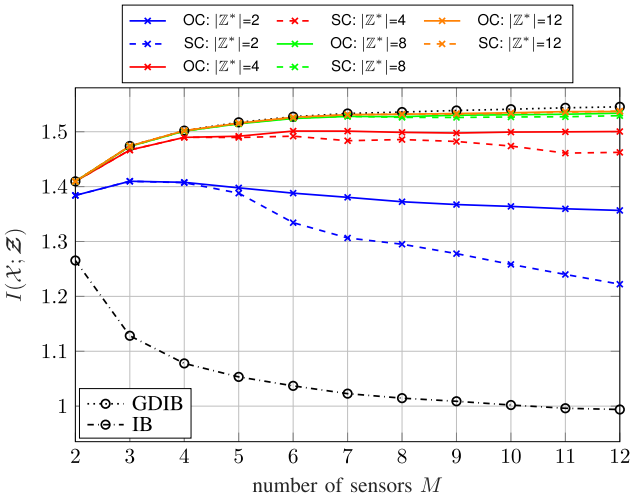

FIGURE 11. Relevant mutual information versus number of sensors; symmetric scenario, sum-rate C_{sum} =2.5 bit/s/Hz, SNRs $\gamma_m = 8$ dB, $|\mathbb{X}| = 4$, $|\mathbb{Y}_m| = 64$ and $|\mathbb{Z}_m| = 4$; One-Step Compression (OC), Sequential Compression (SC).

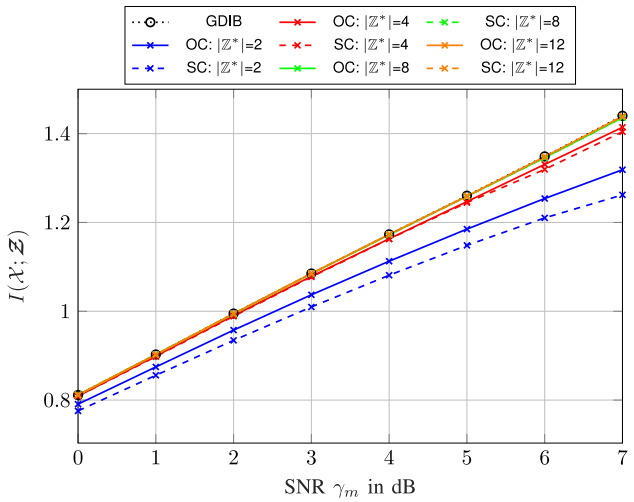

FIGURE 12. Relevant mutual information versus SNR γ_m in dB for symmetric scenario and sum-rate C_{sum} =2.5 bit/s/Hz, M=7, $|\mathbb{Y}_m| = 64$ and $|\mathbb{Z}_m| = 4$; One-Step Compression (OC), Sequential Compression (SC).

The solid curves represent the Greedy Distributed Information Bottleneck (GDIB) approach with one-step compressed side-information (OC). The comparison to the uncompressed GDIB curve illustrates the general loss due to compression of side-information. Due to stronger compression for growing M with fixed C_{sum}, more side-information for the quantizer optimization is required to preserve the relevant mutual information. If this side-information is not available due to low $|\mathbb{Z}^*|$, the overall performance degrades with growing M. With increasing $|\mathbb{Z}^*|$, the loss compared to the uncompressed case decreases. The required cardinality of $\mathbb{Z}^*$ depends on the number of sensors in the network. For the considered network with up to $M = 12$ sensors, a relatively small number $|\mathbb{Z}^*| = 12$ seems to be sufficient.

The dashed lines depict the Greedy Distributed Information Bottleneck approach with sequentially compressed side-information (SC). As already shown in Fig. 10, this approach generally performs worse than the GDIB with OC because the successive compression scheme leads to an accumulation of individual compression losses. Increasing the cardinality $|\mathbb{Z}^*|$ reduces these losses in relevant mutual information and allows to approach the performance of the uncompressed case. Compressing the side-information to only $|\mathbb{Z}^*| = 12$ clusters results in a negligible loss in relevant mutual information. In the scenario with $M = 12$ sensors, $|\mathbb{Z}_{1:11}| = 4^{11} = 4194304$ holds when optimizing the last sensor. Using 8 byte for double precision and $|\mathbb{Y}| = 64$, it needs 8 GiB to store $p(x|y_M, \mathbf{z}_{<M})$. Note that increasing the cardinality $|\mathbb{Z}_m|$ would lead to an exponential increase of the memory requirements. In contrast, it only needs 24 KiB for storing $p(x|y_M, z^*_{<M})$ with $|\mathbb{Z}^*| = 12$.

Finally, Fig. 12 depicts the relevant mutual information $I(\mathcal{X};\mathcal{Z})$ versus the measurement SNRs γ_m in dB for different cardinalities $|\mathbb{Z}^*|$ and a network size of $M = 7$. Again, the cardinality of the relevant signal and the number of output clusters are $|\mathbb{X}| = |\mathbb{Z}| = 4$. All sensors share a single channel with a sum-rate of $C_{\text{sum}} = 2.5$ bit/s/Hz. The black dotted curve depicts the upper bound, i.e., it represents the uncompressed GDIB approach. Naturally, increasing SNRs lead to growing relevant information. Moreover, the accumulated loss due to the sequential compression of side-information increases for larger SNRs. To overcome these losses, higher measurement SNRs require a larger cardinality $|\mathbb{Z}^*|$.

G. INFLUENCE OF RELEVANT SIGNAL

Figure 13 illustrates the loss in relevant mutual information $\Delta I(\mathcal{X};\mathcal{Z}) = I(\mathcal{X};\mathcal{Z}) - I^{\text{c}}(\mathcal{X};\mathcal{Z})$ due to one-step compression (OC) or sequential compression (SC) of side-information versus the cardinality $|\mathbb{Z}^*|$ for different alphabets

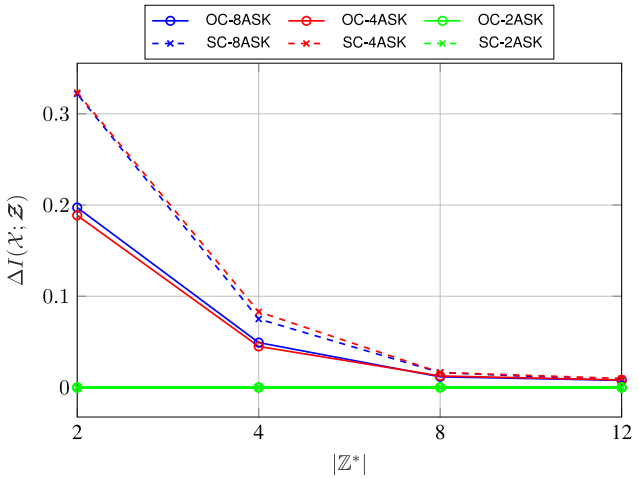

FIGURE 13. Compression loss $\Delta I(\mathcal{X};\mathcal{Z})$ vs $|\mathbb{Z}^*|$; Symmetric scenario with $M = 12$, sum-rate C_{sum} =2.5 bit/s/Hz, SNRs $\gamma_m = 8$ dB, $|\mathbb{X}|$ =2, 4, 8, $|\mathbb{Y}_m| = 64$ and $|\mathbb{Z}_m| = 4$; One-Step Compression (OC), Sequential Compression (SC).

of the relevant signal $\mathcal{X}$. It can be observed, that the compression does not introduce any error for a 2-ASK, independent of the cardinality $|\mathbb{Z}^*|$ and the compression scheme. This is different for the 4-ASK and the 8-ASK. The solid lines represent the loss introduced by one-step compression. As expected, it is smaller than for sequential compression depicted by dashed lines. With increasing cardinality $|\mathbb{Z}^*|$ the difference between the compression schemes becomes smaller and the general loss introduced by compression gets very small. In this simulation, no significant difference between the mappings 4-ASK and 8-ASK can be observed. However, for higher sum-rates C_{sum} the achievable relevant mutual information increases, which might also affect the compression loss in relevant mutual information using larger cardinalities $|\mathbb{X}|$.

VI. CONCLUSION

This paper derives an algorithmic solution for solving the CEO problem based on the well known IB principle. The GDIB algorithm optimizes the quantizer mappings successively and exploits already optimized mappings as side-information. The proposed algorithm substantially outperforms independently IB-optimized quantizers and benefits from a growing number of sensors which is not the case for the latter for fixed sum-rates. Moreover, the loss of distributed compression compared to centralized quantization is discussed. Since the complexity of the GDIB algorithm becomes infeasible for larger networks, side-information is sequentially compressed during the optimization process ensuring the feasibility even for larger network sizes. The loss in terms of relevant mutual information compared to the uncompressed case can be made arbitrary small for moderate cardinalities of the compressed side-information. Naturally, the optimal cardinality depends on the network size, on the measurement SNRs of the individual sensors and on the relevant random variable $\mathcal{X}$.

APPENDIX
DERIVATION OF GDIB SOLUTION

The solution for sensor m that maximizes the optimization problem in (16a)-(16b) can be found by equating the derivative of

$$L_{\text{GDIB}}^{(m)} = I(\mathcal{X};\mathcal{Z}_m|\boldsymbol{\mathcal{Z}}_{<m}) - \beta_m I(\mathcal{Y}_m;\mathcal{Z}_m|\boldsymbol{\mathcal{Z}}_{<m}) \tag{20}$$

w.r.t. $p(z_m|y_m)$ to zero. Note that the mapping $p(z_m|y_m)$ obviously influences both terms, $I(\mathcal{X};\mathcal{Z}_m|\boldsymbol{\mathcal{Z}}_{<m})$ and $I(\mathcal{Y}_m;\mathcal{Z}_m|\boldsymbol{\mathcal{Z}}_{<m})$. We will now derive the derivatives for all mutual information terms.

1) *Derivative of $I(\mathcal{X};\mathcal{Z}_m|\boldsymbol{\mathcal{Z}}_{<m})$:* The relevant mutual information in (20) can be rewritten such that the desired mapping occurs explicitly.

$$\begin{aligned} I(\mathcal{X};\mathcal{Z}_m|\boldsymbol{\mathcal{Z}}_{<m}) &= \mathbb{E}_{\mathcal{X},\boldsymbol{\mathcal{Z}}}\left[\log\frac{p(z_m|x,\mathbf{z}_{<m})}{p(z_m|\mathbf{z}_{<m})}\right] \\ &= \sum_{z_m}\sum_{y_m} p(z_m|y_m)\sum_{x} p(x,y_m) \\ &\quad\times \log\sum_{a\in\mathbb{Y}_m} p(z_m|a)p(a|x) \\ &\quad -\sum_{z_m}\sum_{y_m} p(z_m|y_m)\sum_{\mathbf{z}_{<m}} p(y_m,\mathbf{z}_{<m}) \\ &\quad\times \log\sum_{a\in\mathbb{Y}_m} p(z_m|a)p(a|\mathbf{z}_{<m}) \end{aligned} \tag{21}$$

The derivative of (21) delivers

$$\begin{aligned} &\frac{\partial I(\mathcal{X};\mathcal{Z}_m|\boldsymbol{\mathcal{Z}}_{<m})}{\partial p(z_m|y_m)} \\ &= \sum_{x} p(x,y_m)\log p(z_m|x) \\ &\quad + \sum_{x}\underbrace{\left[\sum_{y_m} p(z_m|y_m)p(x,y_m)\right]}_{=p(x,z_m)}\frac{p(y_m|x)}{p(z_m|x)} \\ &\quad - \sum_{\mathbf{z}_{<m}} p(y_m,\mathbf{z}_{<m})\log p(z_m|\mathbf{z}_{<m}) \\ &\quad - \sum_{\mathbf{z}_{<m}}\underbrace{\left[\sum_{y_m} p(z_m|y_m)p(y_m,\mathbf{z}_{<m})\right]}_{=p(\mathbf{z})}\frac{p(y_m|\mathbf{z}_{<m})}{p(z_m|\mathbf{z}_{<m})} \\ &= \sum_{x} p(x,y_m)\log p(z_m|x) \\ &\quad - \sum_{\mathbf{z}_{<m}} p(y_m,\mathbf{z}_{<m})\log p(z_m|\mathbf{z}_{<m}) \\ &= \sum_{\mathbf{z}_{<m}} p(y_m,\mathbf{z}_{<m})\sum_{x} p(x|y_m,\mathbf{z}_{<m})\log\frac{p(z_m|x)}{p(z_m|\mathbf{z}_{<m})}. \end{aligned} \tag{22}$$

Exploiting the Markov property $\mathcal{X}\to\mathcal{Y}_m\to\mathcal{Z}_m$ and the independence of $\mathcal{Z}_m$ given x, the argument of the logarithmic function can be extended to

$$\frac{p(z_m|x)}{p(z_m|\mathbf{z}_{<m})} = \frac{p(z_m|x,\mathbf{z}_{<m})}{p(z_m|\mathbf{z}_{<m})} = \frac{p(x|\mathbf{z}_{\le m})}{p(x|\mathbf{z}_{<m})}$$

$$= \frac{p\big(x|\mathbf{z}_{\leq m}\big)}{p(x|y_m, \mathbf{z}_{<m})} \frac{p(x|y_m, \mathbf{z}_{<m})}{p(x|\mathbf{z}_{<m})}. \tag{23}$$

The second ratio in (23) can be dropped because it does not depend on $p(z_m|y_m)$ and its contribution can be incorporated into the Lagrange multiplier β_m. The insertion of the first ratio into (22) yields the contribution of the derivative of the relevant mutual information

$$\begin{aligned}
&\frac{\partial I(\mathcal{X}; \mathcal{Z}_m|\boldsymbol{\mathcal{Z}}_{<m})}{\partial p(z_m|y_m)} \\
&\quad \to -\sum_{\mathbf{z}_{<m}} p(y_m, \mathbf{z}_{<m}) \sum_x p(x|y_m, \mathbf{z}_{<m}) \log \frac{p(x|y_m, \mathbf{z}_{<m})}{p\big(x|\mathbf{z}_{\leq m}\big)} \\
&\quad = -\sum_{\mathbf{z}_{<m}} p(y_m, \mathbf{z}_{<m}) D_{\mathrm{KL}}\big[p(x|y_m, \mathbf{z}_{<m}) \| p\big(x|\mathbf{z}_{\leq m}\big)\big]. \qquad (24)
\end{aligned}$$

2) *Derivative of $I(\mathcal{Y}_m; \mathcal{Z}_m|\boldsymbol{\mathcal{Z}}_{<m})$:* With the definition of the conditional compression rate

$$\begin{aligned}
I(\mathcal{Y}_m; \mathcal{Z}_m|\boldsymbol{\mathcal{Z}}_{<m}) &= \mathbb{E}_{\mathcal{Y}_m, \boldsymbol{\mathcal{Z}}}\left[\log \frac{p(z_m|y_m)}{p(z_m|\mathbf{z}_{<m})}\right] \\
&= \sum_{z_m} \sum_{y_m} p(z_m|y_m) p(y_m) \log p(z_m|y_m) \\
&\quad - \sum_{z_m} \sum_{y_m} p(z_m|y_m) \sum_{\mathbf{z}_{<m}} p(y_m, \mathbf{z}_{<m}) \\
&\quad \times \log \sum_{a \in \mathbb{Y}_m} p(z_m|a) p(a|\mathbf{z}_{<m}), \qquad (25)
\end{aligned}$$

its derivative becomes

$$\begin{aligned}
&\frac{\partial I(\mathcal{Y}_m; \mathcal{Z}_m|\boldsymbol{\mathcal{Z}}_{<m})}{\partial p(z_m|y_m)} \\
&= p(y_m) \log p(z_m|y_m) + p(y_m) \frac{p(z_m|y_m)}{p(z_m|y_m)} \\
&\quad - \sum_{\mathbf{z}_{<m}} p(y_m, \mathbf{z}_{<m}) \log p(z_m|\mathbf{z}_{<m}) \\
&\quad - \sum_{\mathbf{z}_{<m}} \underbrace{\left[\sum_{y_m} p(z_m|y_m) p(y_m, \mathbf{z}_{<m})\right]}_{= p(\mathbf{z}_{\leq m})} \frac{p(y_m|\mathbf{z}_{<m})}{p(z_m|\mathbf{z}_{<m})} \\
&= p(y_m) \log p(z_m|y_m) - \sum_{\mathbf{z}_{<m}} p(y_m, \mathbf{z}_{<m}) \log p(z_m|\mathbf{z}_{<m}). \qquad (26)
\end{aligned}$$

3) *Fusion of Derived Parts:* Combining the result in (24) and (26) delivers the complete derivative

$$\begin{aligned}
&-\sum_{\mathbf{z}_{<m}} p(y_m, \mathbf{z}_{<m}) \cdot D_{\mathrm{KL}}\big[p(x|y_m, \mathbf{z}_{<m}) \| p\big(x|\mathbf{z}_{\leq m}\big)\big] \\
&\quad - \beta_m p(y_m) \log p(z_m|y_m) \\
&\quad + \beta_m \sum_{\mathbf{z}_{<m}} p(y_m, \mathbf{z}_{<m}) \cdot \log p(z_m|\mathbf{z}_{<m}) = 0. \qquad (27)
\end{aligned}$$

Following the idea of Blahut and Arimoto [35]–[37], i.e., $p(x|\mathbf{z}_{\leq m})$ and $p(z_m|\mathbf{z}_{<m})$ are assumed to be independent of $p(z_m|y_m)$, (27) can be resolved w.r.t. the desired mapping of sensor m leading to the self-consistent solution

$$p(z_m|y_m) = \frac{e^{-d_{\beta_m}(y_m, z_m)}}{\sum_{z_m} e^{-d_{\beta_m}(y_m, z_m)}} \tag{28}$$

with

$$\begin{aligned}
d_{\beta_m}(y_m, z_m) &:= \sum_{\mathbf{z}_{<m}} p(\mathbf{z}_{<m}|y_m) \\
&\quad \times \left[\frac{1}{\beta_m} D_{\mathrm{KL}}\big[p(x|y_m, \mathbf{z}_{<m}) \| p\big(x|\mathbf{z}_{\leq m}\big)\big] - \log p(z_m|\mathbf{z}_{<m})\right] \qquad (29) \\
&= \mathbb{E}_{\boldsymbol{\mathcal{Z}}_{<m}|y_m}\left[\frac{1}{\beta_m} D_{\mathrm{KL}}\big[p(x|y_m, \mathbf{z}_{<m}) \| p\big(x|\mathbf{z}_{\leq m}\big)\big]\right. \\
&\quad \left. - \log p(z_m|\mathbf{z}_{<m})\right]. \qquad (30)
\end{aligned}$$

REFERENCES

[1] M. Gastpar, M. Vetterli, and P. Dragotti, "Sensing reality and communicating bits: A dangerous liaison," *IEEE Signal Process. Mag.*, vol. 23, no. 4, pp. 70–83, Jul. 2006.

[2] Y. Oohama, "The rate-distortion function for the quadratic Gaussian CEO problem," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1057–1070, May 1998.

[3] V. Prabhakaran, D. Tse, and K. Ramachandran, "Rate region of the quadratic Gaussian CEO problem," in *Proc. Int. Symp. Inf. Theory (ISIT)*, Jun. 2004, p. 119.

[4] H. Viswanathan and T. Berger, "The quadratic Gaussian CEO problem," *IEEE Trans. Inf. Theory*, vol. 43, no. 5, pp. 1549–1559, Sep. 1997. [Online]. Available: https://doi.org/10.1109/18.623151

[5] A. Wagner, S. Tavildar, and P. Viswanath, "Rate region of the quadratic Gaussian two-encoder source-coding problem," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 1938–1961, May 2008.

[6] J. Chen, X. 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 J. Sel. Areas Commun.*, vol. 22, no. 6, pp. 977–987, Aug. 2004.

[7] Y. Oohama, "Rate-distortion theory for Gaussian multiterminal source coding systems with several side informations at the decoder," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2577–2593, Jul. 2005.

[8] Y. Ugur, I. E. Aguerri, and A. Zaidi, "Vector Gaussian CEO problem under logarithmic loss," in *Proc. IEEE Inf. Theory Workshop (ITW)*, 2018, pp. 1–5.

[9] Y. Uğur, I. E. Aguerri, and A. Zaidi, "Vector Gaussian CEO problem under logarithmic loss and applications," *IEEE Trans. Inf. Theory*, vol. 66, no. 7, pp. 4183–4202, Jul. 2020.

[10] J. Wang and J. Chen, "On the vector Gaussian L-terminal CEO problem," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2012, pp. 571–575.

[11] J. Chen and J. Wang, "On the vector Gaussian CEO problem," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2011, pp. 2050–2054.

[12] Y. Xu and Q. Wang, "Rate region of the vector Gaussian CEO problem with the trace distortion constraint," *IEEE Trans. Inf. Theory*, vol. 62, no. 4, pp. 1823–1835, Apr. 2016.

[13] T. A. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," *IEEE Trans. Inf. Theory*, vol. 60, no. 1, pp. 740–761, Jan. 2014.

[14] T. Berger, Z. Zhang, and H. Viswanathan, "The CEO problem [multiterminal source coding]," *IEEE Trans. Inf. Theory*, vol. 42, no. 3, pp. 887–902, May 1996. [Online]. Available: https://doi.org/10.1109/18.490552

[15] K. Eswaran and M. Gastpar. (2018). *Remote Source Coding Under Gaussian Noise: Dueling Roles of Power and Entropy Power*. [Online]. Available: https://arxiv.org/abs/1805.06515v2

[16] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Annu. Allerton Conf. Commun. Control Comput.*, Sep. 1999, pp. 368–377.

[17] N. Slonim, "The information bottleneck theory and applications," Ph.D. dissertation, Dept. Comput. Sci., Hebrew Univ. Jerusalem, Jerusalem, Israel, Jan. 2002.

[18] S. Hassanpour, D. Wuebben, and A. Dekorsy, "Overview and investigation of algorithms for the information bottleneck method," in *Proc. SCC*, 2017, pp. 1–6.

[19] S. Hassanpour, D. Wübben, A. Dekorsy, and B. Kurkoski, "On the relation between the asymptotic performance of different algorithms for information bottleneck framework," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017, pp. 1–6.

[20] I. E. Aguerri and A. Zaidi, "Distributed variational representation learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 120–138, Jan. 2021.

[21] J. Lewandowsky and G. Bauch, "Information-optimum LDPC decoders based on the information bottleneck method," *IEEE Access*, vol. 6, pp. 4054–4071, 2018.

[22] G. Zeitler, "Low-precision analog-to-digital conversion and mutual information in channels with memory," in *Proc. 48th Annu. Allerton Conf. Commun. Control Comput.*, Sep. 2010, pp. 745–752.

[23] M. Meidlinger and G. Matz, "On irregular LDPC codes with quantized message passing decoding," in *Proc. IEEE 18th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jan. 2017, pp. 1–5.

[24] F. Romero and B. Kurkoski, "LDPC decoding mappings that maximize mutual information," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 9, pp. 2391–2401, Sep. 2016.

[25] G. Zeitler, *Low-Precision Quantizer Design for Communication Problems*, Technische Universitaet Muenchen, Munich, Germany, 2012.

[26] D. Chen and V. Kuehn, "Alternating information bottleneck optimization for the compression in the uplink of C-RAN," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–7.

[27] I. E. Aguerri and A. Zaidi, "Distributed information bottleneck method for discrete and Gaussian sources," in *Proc. Int. Zürich Seminar Inf. Commun. (IZS)*, 2018, pp. 35–39.

[28] 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 *Proc. IEEE Inf. Theory Workshop (ITW)*, 2017, pp. 349–353.

[29] C. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948.

[30] C. Shannon, "Coding theorems for a discrete source with a fidelity criterion," in *Proc. Inst. Radio Eng. Int. Convent. Rec.*, vol. 7, 1959, pp. 142–163.

[31] D. Sakrison, "Source encoding in the presence of random disturbance," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 1, pp. 165–167, Jan. 1968.

[32] J. Wolf and J. Ziv, "Transmission of noisy information to a noisy receiver with minimum distortion," *IEEE Trans. Inf. Theory*, vol. IT-16, no. 4, pp. 406–411, Jul. 1970.

[33] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Upper Saddle River, NJ, USA: Prentice-Hall, 1971.

[34] Y. Ephraim and R. Gray, "A unified approach for encoding clean and noisy sources by means of waveform and autoregressive model vector quantization," *IEEE Trans. Inf. Theory*, vol. IT-34, no. 4, pp. 826–834, Jul. 1988.

[35] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 1, pp. 14–20, Jan. 1972.

[36] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 4, pp. 460–473, Jul. 1972.

[37] T. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed. New York, NY, USA: Wiley, 2006.

[38] J. Lewandowsky, M. Stark, and G. Bauch, "Information bottleneck graphs for receiver design," in *Proc. IEEE ISIT*, 2016, pp. 2888–2892.

[39] S. Steiner and V. Kuehn, "Distributed compression using the information bottleneck principle," in *Proc. IEEE Int. Conf. Commun.*, 2021.

[40] I. E. Aguerri, A. Zaidi, G. Caire, and S. S. Shitz, "On the capacity of cloud radio access networks with oblivious relaying," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2017, pp. 2068–2072.

[41] I. E. Aguerri, A. Zaidi, G. Caire, and S. S. Shitz, "On the capacity of cloud radio access networks with oblivious relaying," *IEEE Trans. Inf. Theory*, vol. 65, no. 7, pp. 4575–4596, Jul. 2019.

[42] S. Fujishige, *Submodular Functions and Optimization*. Amsterdam, The Netherlands: Elsevier, 2005.

[43] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 145–151, Jan. 1991.

STEFFEN STEINER (Graduate Student Member, IEEE) received the Diploma degree in information technology/computer engineering from the University of Rostock in 2017, where he joined the Department of Communications Engineering in 2018. His focus lies on sensor networks and distributed compression using the information bottleneck principle.

VOLKER KUEHN (Member, IEEE) received the Diploma degree and the Ph.D. degree in electrical engineering from the University of Paderborn in 1993 and 1998, respectively. In 1998, he joined the Department of Communications Engineering, University of Bremen, where he wrote his post-doctoral thesis. From 2005 to 2006, he was with Qualcomm CDMA Technologies GmbH, Nuremberg, Germany. In 2006, he became a Full Professor of Communications Engineering with the University of Rostock, Germany. He has authored/coauthored three text books and numerous conference and journal contributions. He has worked on various fields of mobile radio communications like error control coding, multiple antenna systems as well as relay and sensor networks. In recent years, the focus has shifted to digital processing of sparse signals like compressed sensing and spectral estimation of finite rate of innovation signals and to distributed compression using the information bottleneck principle.

MAXIMILIAN STARK received the Bachelor of Science and Master of Science degrees in electrical engineering from the Hamburg University of Technology (TUHH) in 2014 and 2017, respectively, where he is currently pursuing the Ph.D. degree. In 2016, he studied with the Chalmers University of Technology during an Erasmus semester. Since 2017, he has been with the Institute of Communications, TUHH as a Research Assistant. In 2019, he was with Nokia Bell Labs, France, as a Visiting Researcher in the area of deep learning in communications. His research is in the area of machine learning methods for communications and signal processing, with particular interest in channel coding, the information bottleneck method and coarsely quantized signal processing. He was recipient of the Karl H. Dietze Award in 2017 for his master thesis.

GERHARD BAUCH (Fellow, IEEE) received the Dipl.-Ing. and Dr.-Ing. degrees in electrical engineering from the Munich University of Technology (TUM) in 1995 and 2001, respectively, and the Diplom-Volkswirt (master's in economics) degree from FernUniversitaet Hagen in 2001. In 1996, he was with the German Aerospace Center (DLR), Oberpfaffenhofen, Germany. From 1996 to 2001, he was a Member of Scientific Staff with TUM. In 1998 and 1999, he was also a Visiting Researcher with AT&T Labs Research, Florham Park, NJ, USA. In 2002, he joined DOCOMO Euro-Labs, Munich, Germany, where he has been managing the Advanced Radio Transmission Group. In 2007, he was additionally appointed as a Research Fellow of DOCOMO Euro-Labs. He was a Full Professor with the Universität der Bundeswehr Munich from 2009 to 2012. Since October 2012, he has been the Head of the Institute of Communications, Hamburg University of Technology.