On the Equivalence of Two Information Bottleneck-Based Routines Devised for Joint Source-Channel Coding

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

***Abstract*—Consider an extended version of noisy source coding in which the compressed data must be transmitted over an imperfect (forward) channel before being further processed. An interesting criterion for designing the quantizer in such cases is to maximize the end-to-end transmission rate. Since finding the globally optimal solution via the naive brute-force search is practically infeasible, a number of routines have been proposed aiming at providing complexity-wise tractable procedures at the expense of converging to local optima. In this paper, we study the relation between two particular devised heuristics appeared in totally different applications and prove their equivalence through an in-depth analytical investigation of their counterpart algorithmic steps. We further substantiate our presented line of argumentation by means of computer-generated simulations.**

I. Introduction

In this study, we focus on lossy joint source-channel coding. Explicitly, we consider the case for which the goal is to quantize an observed signal (e.g., at the output of an access channel) from a given source with the extra knowledge that the compressed signal has to be transmitted over a non-ideal (forward) channel to be fed into a distant signal processing chain. This, in fact, is the underlying scenario in a variety of practical applications, e.g., cooperative transmission through relaying with noisy links when quantize-and-forward strategy is chosen (see, e.g., [1]), distributed inference sensor networks with non-ideal links to the fusion center (see, e.g., [2]), reception schemes with unreliable memories (see, e.g., [3]), and Cloud-based Radio Access Networks (C-RANs) with noisy fronthaul links (see, e.g., [4], [5]). In such cases, the imperfect forward/fronthaul channel effects shall be incorporated into the quantizer design setup.

A rather straightforward approach to do so is to treat the observed variable as a *virtual* source and try to adapt the conventional methods from Rate-Distortion (RD) theory [6] by extending the distortion measure function such that the impacts of the imperfect forward channel are taken into account [7], [8]. The facts that in such procedures, the actual source is not explicitly brought into the design setup and furthermore there is no way to systematically achieve the appropriate distortion measure for any particular case of relevance, are incentives to think of a novel framework for quantization.

As an interesting alternative paradigm to work with, the Information Bottleneck (IB) method [9] can be deployed. It was firstly proposed in the context of machine learning wherein the goal was to extract the relevant information w.r.t. a desired variable from a typically huge dataset through clever clustering [10]. Performing this type of dimensionality reduction is an indispensable part in many practical fields exploiting statistical analysis to process data (see, e.g., [11]). To get an overall picture of the IB framework and a variety of pertinent algorithmic approaches and their relations, interested readers are referred to [12], [13].

Inspired by the primary IB methodology, instead of minimizing the average distortion w.r.t. an extended distortion measure, one can think of maximizing the Mutual Information (MI) between the source and the final variable to be processed. Unlike the conventional methods from RD theory, this leads to a symmetric design setup wherein both precision and complexity of the resultant outcome are characterized through MI terms. Consequently, the quantizer design becomes purely statistical and therefore irrelevant to the realizations of the variable to be compressed which makes it fundamentally different from conventional approaches. Moreover, in this fashion, for a given input statistics, the resultant quantizer maximizes the overall transmission rate which is absolutely desired for almost all communications systems. As will be discussed, for such quantizers obtaining the globally optimal solution within a tractable complexity is quite demanding. Thus, recently a number of iterative heuristics have been proposed in the literature aiming at yielding complexity-wise efficient routines at the expense of converging to local optima.

In this paper, we consider two specific devised routines, namely, the Channel-Optimized Information Bottleneck ("Algorithm 1" in [14]) and the Channel-Aware Double Maxima ("Algorithm 1" in [15]) approaches. The former has been developed to yield a novel rate-maximizing vector quantizer while the latter has been proposed in the context of distributed source coding when multiple observations of a given source shall be smartly quantized before being transmitted over noisy links to a fusion center for further processing. We carry out a comprehensive analysis to clearly prove their algorithmic equivalence for the overlapping scenario of the scalar quantization of a single noisy observation. To this end, we firstly introduce the presumed system model and provide the mathematical insights into the pertinent optimization task

978-1-5386-2321-3/18/$31.00 ©2018 IEEE

in Section II. Then, after an in-depth discussion about the aforementioned routines in Section III, we present our analysis in Section IV along with the simulation results of a typical digital transmission setup as corroborative evidence before summarizing the salient points at the end.

This theoretical inquiry elucidates the precise relation between the suggested procedures derived from totally different approaches taken to solve a given problem. One shall note although it may sound expectable that different routines attacking the same problem may be similar, proving their *identicality* is fundamentally different from having the coarse intuition of *similarity* that is our very contribution here.

II. Joint Source-Channel Coding Setup

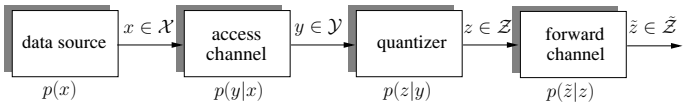

Fig. 1. System model for joint source-channel coding/compression

For the considered system model depicted in Fig. 1 we assume a discrete memoryless source x (with realizations $x \in \mathcal{X}$) having the a priori distribution $p(x)$ followed by a Discrete Memoryless Channel (DMC) being described via transition probability distribution $p(y|x)$. Presumably, the direct access to the source is not available. Thus, the observed variable y (with realizations $y \in \mathcal{Y}$) at the output of the access channel has to be compressed to the variable z (with realizations $z \in \mathcal{Z}$) before being transmitted over the forward DMC which is characterized through conditional distribution $p(\tilde{z}|z)$. Furthermore, we assume that $\mathsf{x} \leftrightarrow \mathsf{y} \leftrightarrow \mathsf{z} \leftrightarrow \tilde{\mathsf{z}}$ is a first-order Markov chain and the joint distribution $p(x,y) = p(x)\,p(y|x)$ and the forward channel transition probabilities $p(\tilde{z}|z)$ are given. In order to design a quantizer $p(z|y)$ that maximizes the overall transmission rate in this setup, the following problem must be addressed:

$$p^{\star}(z|y) = \underset{p(z|y)}{\operatorname{argmax}}\, I(\mathsf{x};\tilde{\mathsf{z}}) \;\text{ for } |\mathcal{Z}| \leq N, \tag{1}$$

wherein N is the allowed number of quantization levels and $|\cdot|$ denotes the cardinality (the number of elements) of a given set. To get an impression about the type of the optimization task at hand, we now investigate (1) in more details.

It is well known that for a given $p(x)$, the objective function[1] in (1) is convex w.r.t. the conditional distribution $p(\tilde{z}|x)$ [6]. Moreover, the relation between $p(\tilde{z}|x)$ and the quantizer mapping $p(z|y)$ is established through

$$p(\tilde{z}|x) = \sum_{y\in\mathcal{Y}} \sum_{z\in\mathcal{Z}} p(\tilde{z}|z)p(z|y)p(y|x), \tag{2}$$

that is of affine type preserving convexity. Hence, it is deduced that $I(\mathsf{x};\tilde{\mathsf{z}})$ is also convex w.r.t. the mapping $p(z|y)$. For a specific $y \in \mathcal{Y}$ it applies

$$\sum_{z\in\mathcal{Z}} p(\mathsf{z}=z|\mathsf{y}=y) = 1, \tag{3}$$

which defines a $(|\mathcal{Z}|-1)$-dimensional probability simplex. Therefore, the overall search space in (1) is achieved by the Cartesian product of $|\mathcal{Y}|$ of such simplices leading to a closed convex polytope in the space of dimensionality $|\mathcal{Y}|\times(|\mathcal{Z}|-1)$.

All in all, it is inferred that the optimization in (1) boils down to maximizing a convex function over a closed convex set which, in optimization theory, is referred to as *convex maximization* or *concave optimization*[2] being NP-hard in general [16]. Resorting to the well-known proposition that a convex function obtains its global maximum over a closed convex set at its extreme points, it is directly deducible that the optimal solution in (1) is achieved by deterministic mappings, i.e., $p(z|y) \in \{0,1\}$ for all pairs $(y,z) \in \mathcal{Y}\times\mathcal{Z}$. To clearly discern this, one may note that the extreme points of a polytope translate into its vertices and for the respective search space polytope in (1), each vertex is created by the Cartesian product of the vertices of its constituent simplices.

As the naive brute-force search over all vertices of the event space in (1) brings about the exponential complexity w.r.t. $|\mathcal{Y}|$ (the search space polytope has in total $|\mathcal{Z}|^{|\mathcal{Y}|}$ different vertices), evidently it cannot be considered as a promising strategy to obtain the desired mapping $p(z|y)$ in practice. This, in fact, is the motive behind the emergence of algorithms aiming at addressing (at least locally) the design problem (1) in an efficient manner. In the next section, we fully discuss two particular routines from this class of heuristics.

III. Considered Routines

A. Channel-Optimized Information Bottleneck (Ch-Opt-IB)

In [14], the authors have considered a similar setup as in Fig. 1 for the problem of channel-optimized vector quantization and developed an iterative algorithm which yields a vector quantizer that maximizes $I(\underline{\mathsf{x}};\tilde{\mathsf{z}})$, in which $\underline{\mathsf{x}}$ is a vector of length M comprising i.i.d. elements engendered by the source x. An interesting point about this algorithm is the implicit optimization of the quantizer output labels, obviating the NP-hard problem of label optimization which has to be addressed separately in conventional vector quantization approaches (see, e.g., [17]). Here, to present the Ch-Opt-IB we restrict ourselves to the scalar quantizer design, i.e., $M = 1$. It applies

$$I(\mathsf{x};\mathsf{y},\tilde{\mathsf{z}}) = I(\mathsf{x};\mathsf{y}) + I(\mathsf{x};\tilde{\mathsf{z}}|\mathsf{y}) = I(\mathsf{x};\tilde{\mathsf{z}}) + I(\mathsf{x};\mathsf{y}|\tilde{\mathsf{z}})\,. \tag{4}$$

Since $I(\mathsf{x};\tilde{\mathsf{z}}|\mathsf{y}) = 0$ by the Markovian assumption, the objective function in (1) can be rewritten as

$$I(\mathsf{x};\tilde{\mathsf{z}}) = I(\mathsf{x};\mathsf{y}) - I(\mathsf{x};\mathsf{y}|\tilde{\mathsf{z}})\,. \tag{5}$$

[1]The MI between discrete random variables a and b with the marginal and the joint distributions $p(a)$, $p(b)$ and $p(a,b)$, respectively is defined as $I(\mathsf{a};\mathsf{b}) \triangleq \sum_a \sum_b p(a,b)\log\frac{p(a,b)}{p(a)p(b)}$.

[2]Please note that this is a totally different task compared to the *convex optimization* wherein the aim is to find the minimum of a convex function.

As the *available mutual information* $I(\mathsf{x};\mathsf{y})$ is fixed (given by the joint distribution $p(x,y)$), maximizing $I(\mathsf{x};\tilde{\mathsf{z}})$ translates into minimizing $I(\mathsf{x};\mathsf{y}|\tilde{\mathsf{z}})$. Exploiting the following definition[3]

$$C(\mathsf{y}=y,\tilde{\mathsf{z}}=\tilde{z}) \triangleq D_{\mathrm{KL}}\big(p(x|y)\|p(x|\tilde{z})\big), \tag{6}$$

one can write

$$I(\mathsf{x};\mathsf{y}|\tilde{\mathsf{z}}) = \mathbb{E}_{\mathsf{y}}\big\{\mathbb{E}_{\tilde{\mathsf{z}}}\{C(\mathsf{y},\tilde{\mathsf{z}})|\mathsf{y}\}\big\}, \tag{7}$$

wherein the conditional expectation calculates as

$$\mathbb{E}_{\tilde{\mathsf{z}}}\{C(\mathsf{y},\tilde{\mathsf{z}})|\mathsf{y}=y\} = \sum_{z\in\mathcal{Z}} p(z|y) \sum_{\tilde{z}\in\tilde{\mathcal{Z}}} p(\tilde{z}|z) C(\mathsf{y}=y,\tilde{\mathsf{z}}=\tilde{z}) \ . \tag{8}$$

Since the inner sum term in (8) is constant for a specific $z\in\mathcal{Z}$, to minimize the conditional expectation (8) for each $y\in\mathcal{Y}$, the quantizer mapping must be chosen as[4] $p(z|y)=\delta_{z,z^\star(y)}$, where

$$z^\star(y) = \underset{z}{\operatorname{argmin}} \sum_{\tilde{z}\in\tilde{\mathcal{Z}}} p(\tilde{z}|z) C(\mathsf{y}=y,\tilde{\mathsf{z}}=\tilde{z}) \ . \tag{9}$$

Consequently, by doing so (7) is minimized for a given $C(\mathsf{y},\tilde{\mathsf{z}})$. Moreover, it is directly deducible that

$$p(\tilde{z}|y) = \sum_{z\in\mathcal{Z}} p(\tilde{z}|z)p(z|y) = p(\tilde{z}|z^\star(y)) \ . \tag{10}$$

To achieve an iterative heuristic that maximizes $I(\mathsf{x};\tilde{\mathsf{z}})$, one must be able to update $C(\mathsf{y},\tilde{\mathsf{z}})$ at the end of each iteration. This can be done by updating $p(x|\tilde{z})$ (taking into account the Markovian property) through

$$p(x|\tilde{z}) = \frac{\sum\limits_{y\in\mathcal{Y}_z} p(x,y)p(\tilde{z}|y)}{\sum\limits_{y\in\mathcal{Y}_z} p(\tilde{z}|y)p(y)}, \tag{11}$$

where $\mathcal{Y}_z$ denotes the subset of $\mathcal{Y}$ for which all members are allocated to the cluster z.

Explicitly, the Ch-Opt-IB is initialized by a random (valid) choice of $C(\mathsf{y},\tilde{\mathsf{z}})$ and iterates over the resultant mapping from (9) (*assignment* step) and the recalculated version of $C(\mathsf{y},\tilde{\mathsf{z}})$ obtained by (11) (*update* step) till convergence to a local optimum. Basically, this procedure can be regarded as an adapted version of the so-called Iterative Information Bottleneck algorithm proposed in [9] (hence the name).

B. Channel-Aware Double Maxima (Ch-Aware-Double-Max)

Lately, the authors in [15] have devised an iterative routine (we refer to as Channel-Aware Double Maxima algorithm) to address the problem of distributed joint source-channel coding. Specifically, they have considered the scenario depicted in Fig. 2 wherein a number of K noisy observations (measured values) of the source x have to be quantized (locally but not independently, i.e., in a jointly manner) first and then transmitted to the fusion center over imperfect forward channels. As the

[3] $D_{\mathrm{KL}}(\cdot\|\cdot)$ is the Kullback-Leibler (KL) divergence which is defined for probability distributions $p(a)$ and $q(a)$ over the same event space $\mathcal{A}$ of the random variable a as $D_{\mathrm{KL}}\big(p(a)\|q(a)\big) \triangleq \sum_{a\in\mathcal{A}} p(a)\log\frac{p(a)}{q(a)}$ [6]. The relation between MI and KL divergence is established through $I(\mathsf{a};\mathsf{b}) = D_{\mathrm{KL}}\big(p(a,b)\|p(a)p(b)\big)$.

[4] Here, δ refers to the Kronecker delta function.

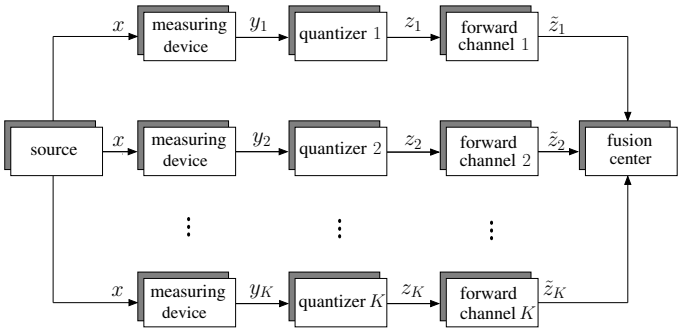

Fig. 2. System model for distributed joint source-channel coding

design criterion, they aimed at maximizing the end-to-end rate, which is quantified through the MI between the source and the resultant vector comprising all variables entering the fusion center, i.e., $I(\mathsf{x};\tilde{\mathsf{z}}_1,\tilde{\mathsf{z}}_2,\cdots,\tilde{\mathsf{z}}_K)$. What makes their derived routine attractive is, as shown in [15], it results in a *high-quality* set of general-purpose quantizers that can be deployed successfully for a wide variety of different applications, e.g., the Chief Executive Officer (CEO) problem [18]. Explicitly, it has been shown that performance-wise its acquired result is quite comparable with (and in some cases even better than) the resultant outcomes of the schemes particularly designed for estimation [19] or detection [20] purposes.

Here again, to adhere to the presumed system model in Fig. 1 for discussion of the Ch-Aware-Double-Max approach we restrict ourselves to the single measurement transmission case ($K=1$). Utilizing the chain rule of MI, the objective function in (1) can be expanded as[5]

$$I(\mathsf{x};\tilde{\mathsf{z}}) = I(\mathsf{x},\mathsf{y};\tilde{\mathsf{z}}) - I(\mathsf{y};\tilde{\mathsf{z}}|\mathsf{x}) \tag{12a}$$

$$= H(\mathsf{x},\mathsf{y}) - H(\mathsf{x},\mathsf{y}|\tilde{\mathsf{z}}) - H(\mathsf{y}|\mathsf{x}) + H(\mathsf{y}|\mathsf{x},\tilde{\mathsf{z}}) \ . \tag{12b}$$

Since the entropies $H(\mathsf{x},\mathsf{y})$ and $H(\mathsf{y}|\mathsf{x})$ in (12b) are fixed (given by the joint distribution $p(x,y)$), it is deduced that (1) boils down to

$$p^\star(z|y) = \underset{p(z|y)}{\operatorname{argmax}} \big[H(\mathsf{y}|\mathsf{x},\tilde{\mathsf{z}}) - H(\mathsf{x},\mathsf{y}|\tilde{\mathsf{z}})\big] \ . \tag{13}$$

The objective function in (13) can be rewritten as

$$\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\sum_{\tilde{z}\in\tilde{\mathcal{Z}}} p(x,y)p(\tilde{z}|y)\log p(x|\tilde{z}) \tag{14a}$$

$$= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}}\sum_{\tilde{z}\in\tilde{\mathcal{Z}}} p(x,y)p(z|y)p(\tilde{z}|z)\log p(x|\tilde{z}), \tag{14b}$$

where to attain (14b) the presumed Markovian property is exploited. Defining $q(y,z)=p(z|y)$ and $f(x,\tilde{z})$ as an arbitrary function such that for each specific value $\tilde{z}\in\tilde{\mathcal{Z}}$ it applies $\sum_{x\in\mathcal{X}} f(\mathsf{x}=x,\tilde{z})=1$, the authors in [15] have introduced a generalized objective function, L, as

$$L(q,f) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}}\sum_{\tilde{z}\in\tilde{\mathcal{Z}}} p(x,y)q(y,z)p(\tilde{z}|z)\log f(x,\tilde{z}). \tag{15}$$

[5] $I(\mathsf{a};\mathsf{b}) = H(\mathsf{a}) - H(\mathsf{a}|\mathsf{b}) = H(\mathsf{b}) - H(\mathsf{b}|\mathsf{a})$ where the entropy function is defined as $H(\mathsf{a}) \triangleq -\sum\limits_{a} p(a)\log p(a)$.

Then, utilizing the method of Lagrange multipliers, i.e., deriving the *augmented* objective function achieved by adding some extra terms regarding the side constraints, it has been shown in [15] that for a given $q(y,z)$, the optimal function $f^\star(x,\tilde{z})$ that maximizes L is achieved by

$$f^\star(x,\tilde{z}) = \frac{\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}} p(x,y)p(z|y)p(\tilde{z}|z)}{\sum_{x'\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}} p(x',y)p(z|y)p(\tilde{z}|z)} \ . \tag{16}$$

Having the assumed Markovian property in mind and noting the respective marginalization of the joint distribution $p(x,y,z,\tilde{z})$ at both the numerator and the denominator of (16) reveals that $f^\star(x,\tilde{z}) = \frac{p(x,\tilde{z})}{p(\tilde{z})}$ which is $p(x|\tilde{z})$ by definition.

Correspondingly, to acquire the optimal mapping $q^\star(y,z)$ that maximizes L for a given $f(x,\tilde{z})$, it should be satisfied that for each $y\in\mathcal{Y}$, $p(z|y)=\delta_{z,z^\star(y)}$, where

$$z^\star(y) = \underset{z}{\operatorname{argmax}} \sum_{x\in\mathcal{X}}\sum_{\tilde{z}\in\tilde{\mathcal{Z}}} p(x,y)p(\tilde{z}|z)\log f(x,\tilde{z}) \ . \tag{17}$$

Since the objective function in (14b) is nothing else than the maximum of the generalized objective function L over $f(x,\tilde{z})$ for a given $q(y,z)$, the respective optimization task can be secured by solving an enlarged maximization problem, i.e., performing double (alternating) maximization of L over $f(x,\tilde{z})$ and $q(y,z)$ in an iterative manner (hence the name), analogous to the proposed methodology in [21].

Specifically, the Ch-Aware-Double-Max routine is initialized by a (valid) random deterministic mapping $p(z|y)$ and iterates over (16) (*update* step) and the resultant mapping by (17) (*assignment* step) till convergence to a local optimum.

In the subsequent section, we conduct a comprehensive analysis to prove the equivalence of the aforementioned algorithms. This conclusion, actually, is quite interesting and insightful since the followed mathematical methodology for derivation of the considered algorithms are totally different. This is directly observable, noting that the *variational calculus* is applied within the derivation of the Ch-Aware-Double-Max routine, while that is not the case for the Ch-Opt-IB approach.

IV. Stepwise Comparison of Ch-Opt-IB and Ch-Aware-Double-Max Algorithms

Our main contribution lies in this section, where through a detailed analysis over the parallel algorithmic steps of the Ch-Opt-IB and the Ch-Aware-Double-Max approaches, we lucidly evince their equivalence. To this end, basically, we have to demonstrate that the corresponding assignment and update steps are identical for both algorithms.

A. Analysis

We begin our analysis by considering the assignment step in Ch-Aware-Double-Max routine. Replacing the resultant $f(x,\tilde{z})$ from (16) to (17), for each $y\in\mathcal{Y}$ the allocated cluster is determined by

$$z^\star(y) = \underset{z}{\operatorname{argmax}} \sum_{\tilde{z}\in\tilde{\mathcal{Z}}}\sum_{x\in\mathcal{X}} p(x,y)p(\tilde{z}|z)\log p(x|\tilde{z}) \tag{18a}$$

$$= \underset{z}{\operatorname{argmin}} \sum_{\tilde{z}\in\tilde{\mathcal{Z}}} p(\tilde{z}|z)\sum_{x\in\mathcal{X}} p(x,y)\big(-\log p(x|\tilde{z})\big), \tag{18b}$$

wherein the maximization is substituted by the minimization through introduction of the minus sign. The inner sum term in (18b) can be rewritten as

$$\sum_{x\in\mathcal{X}} p(x,y)\big(\log\frac{p(x|y)}{p(x|\tilde{z})} - \log p(x|y)\big) \tag{19a}$$

$$= p(y)\Big(\sum_{x\in\mathcal{X}} p(x|y)\big(\log\frac{p(x|y)}{p(x|\tilde{z})} - \log p(x|y)\big)\Big) \tag{19b}$$

$$= p(y)\Big(D_{\mathrm{KL}}\big(p(x|y)\|p(x|\tilde{z})\big) + H(\mathsf{x}|\mathsf{y}=y)\Big) \ . \tag{19c}$$

Since the respective minimization in (18b) is independent of $p(y)$, substituting the inner term in (18b) by (19c) yields

$$z^\star(y) = \underset{z}{\operatorname{argmin}} \sum_{\tilde{z}\in\tilde{\mathcal{Z}}} p(\tilde{z}|z)\Big(D_{\mathrm{KL}}\big(p(x|y)\|p(x|\tilde{z})\big) + H(x|\mathsf{y}=y)\Big). \tag{20}$$

Expanding the objective function in (20), it applies

$$\sum_{\tilde{z}\in\tilde{\mathcal{Z}}} p(\tilde{z}|z)\Big(D_{\mathrm{KL}}\big(p(x|y)\|p(x|\tilde{z})\big) + H(x|\mathsf{y}=y)\Big) \tag{21a}$$

$$= \sum_{\tilde{z}\in\tilde{\mathcal{Z}}} p(\tilde{z}|z)D_{\mathrm{KL}}\big(p(x|y)\|p(x|\tilde{z})\big) + \sum_{\tilde{z}\in\tilde{\mathcal{Z}}} p(\tilde{z}|z)H(x|\mathsf{y}=y) \tag{21b}$$

$$= \sum_{\tilde{z}\in\tilde{\mathcal{Z}}} p(\tilde{z}|z)D_{\mathrm{KL}}\big(p(x|y)\|p(x|\tilde{z})\big) + H(x|\mathsf{y}=y), \tag{21c}$$

where the second term in (21c) is derived noting the fact that the conditional entropy $H(x|\mathsf{y}=y)$ is fixed (given by the joint distribution $p(x,y)$) and $\sum_{\tilde{z}\in\tilde{\mathcal{Z}}} p(\tilde{z}|z)=1$. Substituting (21c) in (20), it can be realized that the ultimate cluster allocation's rule for Ch-Aware-Double-Max algorithm is determined through

$$z^\star(y) = \underset{z}{\operatorname{argmin}} \sum_{\tilde{z}\in\tilde{\mathcal{Z}}} p(\tilde{z}|z)D_{\mathrm{KL}}\big(p(x|y)\|p(x|\tilde{z})\big), \tag{22}$$

since the required minimization is independent of the fixed entropy term in (21c). Considering (22) and (9) together, the equality of the assignment steps for Ch-Opt-IB and Ch-Aware-Double-Max routines is immediately inferable.

Now, we consider the corresponding update steps. In particular, regarding (16) and (11) it can be readily observed that both routines update the same distribution $p(x|\tilde{z})$. Nevertheless, to plainly discern that (16) is indeed identical to (11), it shall be noted that the mapping $p(z|y)$ in (16) is deterministic, i.e., it is equal to 1 iff $y\in\mathcal{Y}_z$. Moreover, due to the assumed Markovian property, it applies $p(\tilde{z}|y) = \sum_{z\in\mathcal{Z}} p(\tilde{z}|z)p(z|y)$ and therefore

$$\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}} p(x,y)p(z|y)p(\tilde{z}|z) = \sum_{y\in\mathcal{Y}_z} p(x,y)\sum_{z\in\mathcal{Z}} p(z|y)p(\tilde{z}|z) \tag{23a}$$

$$= \sum_{y\in\mathcal{Y}_z} p(x,y)p(\tilde{z}|y) \ . \tag{23b}$$

Hence, it becomes clear that the respective numerators in (16) and (11) are the same. In addition, the present summation

over all $x' \in \mathcal{X}$ at the denominator of (16) results in the marginal probability $p(y)$ from the joint distribution $p(x', y)$ and therefore

$$\sum_{x' \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}} p(x', y) p(z|y) p(\tilde{z}|z) \tag{24a}$$

$$= \sum_{y \in \mathcal{Y}_z} \sum_{x' \in \mathcal{X}} p(x', y) \sum_{z \in \mathcal{Z}} p(\tilde{z}|z) \tag{24b}$$

$$= \sum_{y \in \mathcal{Y}_z} \sum_{z \in \mathcal{Z}} p(y) p(\tilde{z}|z) \, . \tag{24c}$$

Thus, one deduces that both denominators are identical as well.

Altogether, via the developed analysis, we clearly proved the algorithmic equivalence of the considered approaches. This yields the profound insight that although these heuristics aim at solving the design problem in (1) following totally different strategies, surprisingly, they eventually provide *identical* solution procedures.

Please note that since both heuristics converge to a local optimum, their respective outcome heavily depends on the choice of initialization. Hence, it can be asserted that, assuming a sufficiently large number of runs (to achieve independence from initialization), both routines generate the same result $p(z|y)$.

B. Simulation Results

In this part, we set about investigating the performance of discussed approaches over a typical digital transmission scenario. Explicitly, we consider the equiprobable bipolar 4-ASK signaling ($\mathcal{X} = \{\pm 1, \pm 3\}$) at the input with the corresponding variance of $\sigma_x^2 = 5$. To attain the transition probability distribution $p(y|x)$, we firstly clip the corresponding conditional probability density functions (pdf) of an Additive White Gaussian Noise (AWGN) channel with three different noise variances ($\sigma_n^2 = 1, 2, 3$) to the part with the absolute value not higher than 6, 7.2 and 8.1, respectively (to set the border guard intervals of $3\sigma_n$ to assure 99.7% coverage) and then uniformly discretize them into $|\mathcal{Y}| = 128$ parts. For the forward channel we consider an N-ary symmetric model being purely characterized by the reliability parameter e in a sense that for each symbol, the correct reception occurs with probability $1 - e$ and the erroneous reception to every other symbol occurs with probability $\frac{e}{N-1}$.

To compare the performances of discussed algorithms, we calculated the resultant overall MI, $I(\mathsf{x}; \tilde{\mathsf{z}})$, over the varying maximum number of output levels for two different scenarios to cover the effects of both DMCs in our presumed system model. As the first case, we kept the forward channel constant (by choosing a specific value for the reliability parameter $e = 0.01$) and varied the noise variance of the first DMC in Fig. 1 (which henceforth we refer to as the access channel). As the second case, we fixed the access channel (by choosing a specific noise variance $\sigma_n^2 = 1$) and varied the reliability parameter e of the forward channel in Fig. 1. The corresponding plots are illustrated in Figs. 3 and 4, respectively. Please note that to obtain (quasi-)independence from choice of initialization each algorithm was run 10^5 times with the best outcome taken.

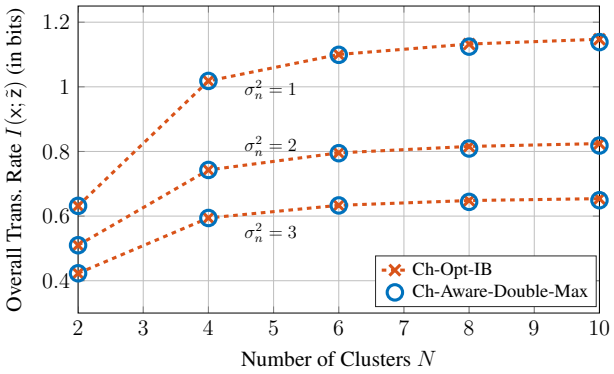

Fig. 3. End-to-end mutual information $I(\mathsf{x}; \tilde{\mathsf{z}})$ vs. maximum number of output levels N for three noise variances σ_n^2 and fixed reliability parameter $e = 0.01$

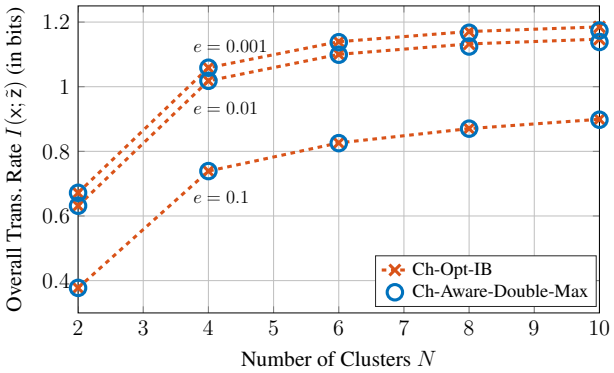

Fig. 4. End-to-end mutual information $I(\mathsf{x}; \tilde{\mathsf{z}})$ vs. maximum number of output levels N for three values of e and fixed noise variance $\sigma_n^2 = 1$

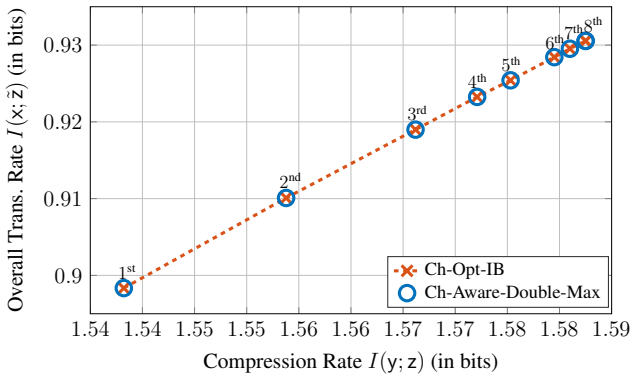

Fig. 5. Overall MI, $I(\mathsf{x}; \tilde{\mathsf{z}})$ vs. compression rate $I(\mathsf{y}; \mathsf{z})$ through iterations with arbitrary choices $N = 8$, $\sigma_n^2 = 1$ and $e = 0.001$

Regarding both plots, the main observation is that irrespective of the specific choices of the parameters of the assumed model, i.e., N, σ_n^2 and e, the Ch-Opt-IB and the Ch-Aware-Double-Max engender (almost) identical results.

Focusing on Fig. 3, it can be seen that by increasing the noise variance σ_n^2 the end-to-end attainable transmission rate $I(\mathsf{x}; \tilde{\mathsf{z}})$ decreases. The reason behind is due to the fact that the overall MI is upper-bounded by the capacity of the access channel. This is directly deduced by applying data processing inequality for the presumed Markov chain. It is well known that the capacity of the discrete input AWGN

channel is reversely related to its noise variance (assuming fixed input variance) [22]. Hence, the lower the noise variance, the higher the capacity of the access channel and consequently the chance of achieving higher values of the end-to-end MI. Concerning Fig. 4, the effect of the forward channel on the overall obtainable transmission rate shows itself in the choice of the reliability parameter e. Explicitly, it can be seen that by decreasing e the end-to-end MI increases. This can be justified analogously, noting the fact that the overall MI is upper-bounded by the capacity of the forward channel C_{FC} as well. To clearly discern this, one may note that $\mathsf{x} \leftrightarrow \mathsf{y} \leftrightarrow \mathsf{z} \leftrightarrow \tilde{\mathsf{z}}$ implies $\tilde{\mathsf{z}} \leftrightarrow \mathsf{z} \leftrightarrow \mathsf{y} \leftrightarrow \mathsf{x}$. It is rather straightforward to show that for a given N, the forward channel capacity C_{FC} which is calculated as [23]

$$C_{\text{FC}}(N, e) = \log N + (1-e)\log(1-e) + e \log \frac{e}{N-1} \quad (25)$$

increases by decreasing e, giving chance to the overall transmission rate to reach higher values.

From the discussion above, it is deducible that the end-to-end MI is upper-bounded by the minimum capacity among the present DMCs in Fig. 1. To vividly see that, as an example one may consider Fig. 4 for the case of $N = 10$. There, although the capacity of the forward channel C_{FC} is calculated as 2.54, 3.21 and 3.31 bits (per channel use) for different values of the reliability parameter e (in descending order), the overall transmission rate is limited by the capacity (more accurately the input-output MI under equiprobable input signaling) of the access channel which amounts to 1.22 bits.

A closer look at Figs. 3 and 4 for relatively large values of N reveals a minor performance mismatch between Ch-Opt-IB and Ch-Aware-Double-Max. This can be attributed to the fact that 10^5 runs does not bring about perfect independence from the choice of initialization. Nevertheless, to rigorously depict the equivalence of both routines, we generated Fig. 5 in which instead of initializing the Ch-Aware-Double-Max randomly, we fed it by the resultant mapping at the end of the first iteration of the Ch-Opt-IB that was initialized randomly with the specific choices of $N = 8$ and $e = 0.001$. To obtain this plot, we calculated the overall transmission rate $I(\mathsf{x}; \tilde{\mathsf{z}})$ and the corresponding compression rate $I(\mathsf{y}; \mathsf{z})$ achieved by the resultant mapping at the end of each iteration for both algorithms. The demonstrated evolution of the outcomes through iterations plainly confirms the conclusion of our conducted analysis.

V. Summary

In this article, we considered the quantizer design problem for the joint source-channel coding with mutual information as the fidelity criterion. Specifically, after providing the mathematical insights into the respective optimization task, we discussed two candidate solution procedures, namely, the Ch-Opt-IB and the Ch-Aware-Double-Max, both quite recently appeared in the literature. Subsequently, by conducting a thorough analysis over the parallel algorithmic steps, we plainly proved their algorithmic equivalence. Finally, performing Monte Carlo simulations for a practical transmission scenario, we also corroborated our presented argument.

Acknowledgment

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

References

[1] G. C. Zeitler, "Low-Precision Quantizer Design for Communication Problems," Ph.D. dissertation, TU Munich, Germany, 2012.

[2] B. Chen, L. Tong, and P. K. Varshney, "Channel-Aware Distributed Detection in Wireless Sensor Networks," *IEEE Signal Processing Magazine*, vol. 23, no. 4, pp. 16–26, 2006.

[3] C. Novak, C. Studer, A. Burg, and G. Matz, "The Effect of Unreliable LLR Storage on the Performance of MIMO-BICM," in *Proc. Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, Nov. 2010, pp. 736–740.

[4] D. Wübben, P. Rost, J. Bartelt, M. Lalam, V. Savin, M. Gorgoglione, A. Dekorsy, and G. Fettweis, "Benefits and Impact of Cloud Computing on 5G Signal Processing," *Special Issue "The 5G Revolution" of the IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 35–44, Nov. 2014.

[5] J. Bartelt, L. Landau, and G. Fettweis, "Improved Uplink I/Q-Signal Forwarding for Cloud-Based Radio Access Networks with Millimeter Wave Fronthaul," in *IEEE Int. Symposium on Wireless Communication Systems (ISWCS)*, Brussels, Belgium, Aug. 2015.

[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, 2006.

[7] A. Kurtenbach and P. Wintz, "Quantizing for Noisy Channels," *IEEE Trans. on Communication Technology*, vol. 17, no. 2, pp. 291–302, 1969.

[8] N. Farvardin and V. Vaishampayan, "Optimal Quantizer Design for Noisy Channels: An Approach to Combined Source-Channel Coding," *IEEE Trans. on Information Theory*, vol. 33, no. 6, pp. 827–838, 1987.

[9] N. Tishby, F. C. Pereira, and W. Bialek, "The Information Bottleneck Method," in *37th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, Sep. 1999, pp. 368–377.

[10] N. Slonim, "The Information Bottleneck: Theory and Applications," Ph.D. dissertation, Hebrew University of Jerusalem, Israel, 2002.

[11] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[12] S. Hassanpour, D. Wübben, and A. Dekorsy, "Overview and Investigation of Algorithms for the Information Bottleneck Method," in *11th Int. Conference on Systems, Communications and Coding (SCC)*, Hamburg, Germany, Feb. 2017.

[13] S. Hassanpour, D. Wübben, A. Dekorsy, and B. M. Kurkoski, "On the Relation Between the Asymptotic Performance of Different Algorithms for Information Bottleneck Framework," in *IEEE Int. Conference on Communications (ICC)*, Paris, France, May 2017.

[14] A. Winkelbauer, G. Matz, and A. Burg, "Channel-Optimized Vector Quantization with Mutual Information as Fidelity Criterion," in *Proc. Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, Nov. 2013, pp. 851–855.

[15] S. Movaghati and M. Ardakani, "Distributed Channel-Aware Quantization Based on Maximum Mutual Information," *International Journal of Distributed Sensor Networks*, vol. 12, no. 5, May 2016.

[16] R. Horst and P. M. Pardalos, *Handbook of Global Optimization*. Springer Science & Business Media, 2013, vol. 2.

[17] Y. Linde, A. Buzo, and R. Gray, "An Algorithm for Vector Quantizer Design," *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 84–95, 1980.

[18] T. Berger, Z. Zhang, and H. Viswanathan, "The CEO Problem," *IEEE Transactions on Information Theory*, vol. 42, no. 3, pp. 887–902, 1996.

[19] J. A. Gubner, "Distributed Estimation and Quantization," *IEEE Transactions on Information Theory*, vol. 39, no. 4, pp. 1456–1459, 1993.

[20] M. Longo, T. D. Lookabaugh, and R. M. Gray, "Quantization for Decentralized Hypothesis Testing under Communication Constraints," *IEEE Trans. on Information Theory*, vol. 36, no. 2, pp. 241–255, 1990.

[21] R. E. Blahut, "Computation of Channel Capacity and Rate-Distortion Functions," *IEEE Trans. on Information Theory*, vol. 18, no. 4, pp. 460–473, Jul. 1972.

[22] G. Caire, G. Taricco, and E. Biglieri, "Bit-Interleaved Coded Modulation," *IEEE Transactions on Information Theory*, vol. 44, no. 3, pp. 927–946, 1998.

[23] D. Hankerson, G. A. Harris, and P. D. Johnson Jr., *Introduction to Information Theory and Data Compression*, 2nd ed., CHAPMAN & HALL/CRC, 2003.