On the Equivalence of Double Maxima and KL-Means for Information Bottleneck-Based Source 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*—In the context of noisy source coding, contrary to the conventional Rate-Distortion theory, the so-called Information Bottleneck method formulates the existent fundamental complexity-precision trade-off in a symmetric and purely information-theoretic fashion. Since the pertinent optimization task to design the quantizer is quite demanding, a number of heuristics have been developed to provide practically feasible procedures at the expense of yielding suboptimal solutions. In this paper, we consider two pertinent routines originally appeared in totally different applications and set out to precisely prove their algorithmic equivalence by conducting a thorough analysis over the corresponding algorithmic steps. We further corroborate our theoretical investigation employing computer-based simulations.**

I. Introduction

The problem of lossy data compression/source coding is dealt with by the celebrated Rate-Distortion (RD) theory under the presumption of direct access to the source [1]. There, to characterize the precision of the outcome a distortion measure function has to be defined a priori (before the quantizer design) which quantifies the amount of distortion among the original signal and its representative after compression. Alas, the RD theory has no answer regarding the basic question of how to systematically achieve the *proper* distortion function in any case of relevance. Therefore, in many practical situations, irrespective of the structure of the signals and solely for the sake of simplicity, the square Euclidean distance (between the quantizer's input/output values) is chosen.

In cases for which only a noisy version of the source is available for compression, one can either resort to the conventional established methods by treating this observed variable as a *virtual* source [2] or, instead, think of having a new framework which directly incorporates the actual source of interest into the design setup. Moving in the direction of the latter with the aim of bypassing the present faults in the conventional theory, the Information Bottleneck (IB) paradigm [3] can be exploited successfully.

The IB framework primarily emerged in the context of machine learning as a novel method performing dimensionality reduction through clustering [4]. This approach can be exploited in a variety of applications concerning data transmission systems, among others, designing analog-to-digital converters (ADCs) [5], construction of polar codes [6] and implementation of modern discrete decoding schemes [7], [8] with reduced complexity and still quite promising performance.

In this study, we focus on the noisy source coding scenario. There, contrary to the RD theory, through the IB formulation the precision of the outcome is quantified by the so-called *relevant information*, i.e., the mutual information (MI) between the actual source and the quantized representation. Hence, a novel symmetric (in the sense of employing two MI terms to mathematically found the underlying trade-off) design setup [3] results that obviates the requirement of an a priori distortion measure specification. Moreover, unlike other approaches, performing the IB-based quantization wherein merely entropy calculations are involved, is purely statistics-based and completely independent of specific realizations of the variable to be compressed.

The focal challenge in the IB-based quantizer design setup lies in the pertinent optimization task. As it will be discussed, finding the globally optimal solution through a practically feasible algorithm is far from trivial and up to now it is only achieved for the special case of binary input alphabets [9]. Consequently, in recent years a number of heuristics have been developed aiming at yielding complexity-wise tractable routines at the expense of converging to *local* optima.

In this article, we consider two specific routines, explicitly, the KL-Means algorithm [10] and the Double Maxima approach ([11], Algorithm 1). The former can be regarded as an adapted version of the well-known K-Means algorithm [12] and the latter has been proposed to address the problem of distributed noisy source coding. To analyze the relation between the mentioned approaches, we first introduce the general IB setup and after providing the mathematical insights into the respective optimization tasks, we discuss both methods in detail. Then, through scrutinizing the corresponding steps that each algorithm performs within the iterations to produce its final result, we evince their equivalence.

This theoretical investigation yields the valuable comprehension of the exact relation between different procedures trying to find *locally optimal* solutions for a given problem by following quite various strategies. Please note that although it may seem expectable that different heuristics attacking the same problem may be similar, proving their *identicality* is fundamentally different from having the coarse intuition of *similarity* and that is our very contribution in this work.

978-1-5386-1734-2/18/$31.00 ©2018 IEEE

II. Information Bottleneck Source Coding Setup

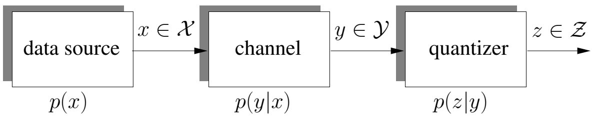

Fig. 1: System model for noisy source coding/compression

For the noisy source coding scenario we consider the system model depicted in Fig. 1. It is assumed that there is no direct access to the discrete memoryless source x (with realizations $x \in \mathcal{X} = \{x_1, \cdots, x_{|\mathcal{X}|}\}$)[1] characterized by the a priori distribution $p(x)$. The goal is then to compress the observed variable y (with realizations $y \in \mathcal{Y} = \{y_1, \cdots, y_{|\mathcal{Y}|}\}$) at the output of the discrete memoryless channel specified by transition probabilities $p(y|x)$, to the random variable z (with realizations $z \in \mathcal{Z} = \{z_1, \cdots, z_{|\mathcal{Z}|}\}$). It shall be noted that, in general, $\mathcal{Z}$ is not necessarily a subset of $\mathcal{Y}$. Furthermore, it is presumed that the joint probability distribution $p(x, y) = p(x)\, p(y|x)$ is known and $\mathsf{x} \leftrightarrow \mathsf{y} \leftrightarrow \mathsf{z}$ is a first-order Markov chain, i.e., $p(z|x, y) = p(z|y)$. Within the IB framework, the complexity of the outcome is quantified through the *compression rate* given by the quantizer's input/output MI[2], $I(\mathsf{y}; \mathsf{z})$, and its resultant precision is quantified through the *relevant information*, $I(\mathsf{x}; \mathsf{z})$. To establish the existent trade-off, a non-negative Lagrange multiplier $0 \leqslant \beta \leqslant \infty$ is exploited and the design setup for the quantizer $p(z|y)$ is formulated via [3]:

$$p^{\star}(z|y) = \underset{p(z|y)}{\operatorname{argmin}} \; \frac{1}{\beta+1}\Big(I(\mathsf{y}; \mathsf{z}) - \beta I(\mathsf{x}; \mathsf{z})\Big) \; \text{for } |\mathcal{Z}| \leq N, \quad (1)$$

where N is the allowed number of output clusters and the factor $\frac{1}{\beta+1}$ is considered solely for the sake of mathematical clarity when investigating the extreme cases for the trade-off parameter. One may note that the trade-off parameter β can be twiddled in order to strengthen (or weaken) the information preservation capability. In addition, it is noteworthy that the resultant mapping $p(z|y)$ has *stochastic* nature in general, i.e., $0 \leq p(z|y) \leq 1$ fulfilling $\sum_{z \in \mathcal{Z}} p(\mathsf{z} = z | \mathsf{y} = y) = 1$ for each $y \in \mathcal{Y}$.

Obviously, the case of $\beta \to 0$ is not of interest as the relevant information term $I(\mathsf{x}; \mathsf{z})$ in (1) is dropped and the minimum compression rate $I(\mathsf{y}; \mathsf{z}) = 0$ could be achieved by making the quantizer's output z being statistically independent of y. For finite β, it can be shown that the objective function in (1) is neither convex nor concave w.r.t. the mapping $p(z|y)$ [13]. Thus, the optimization itself is of neither type and therefore finding the globally optimal solution becomes quite challenging. Regarding asymptotically large values of β, taking the limit of (1) by letting $\beta \to \infty$, the design formulation reduces to:

$$p^{\star}(z|y) = \underset{p(z|y)}{\operatorname{argmax}} \; I(\mathsf{x}; \mathsf{z}) \;\; \text{for } |\mathcal{Z}| \leq N, \quad (2)$$

where the minimization in (1) is substituted by the maximization in (2) by omitting the minus sign. It can be shown that the optimization task in (2) is of *convex maximization*[3] type and the optimal solution is achieved via *deterministic* mappings [9]. It is worth bearing in mind that executing the naive brute-force search over all deterministic mappings results in an exponential complexity w.r.t. $|\mathcal{Y}|$ that makes it evidently intractable in practice.

All in all, it is inferred that for non-zero values of β the relevant optimization task is far from trivial and therefore heuristics have been proposed to address the corresponding design problems efficiently. In the next section after presenting the primarily suggested algorithm for the general IB-based quantizer design (1), we focus on the salient case of β being asymptotically large (2) in which the goal is to maximize the end-to-end transmission rate via keeping as much relevant information as possible under the side-constraint on the cardinality of the output representative signal. One shall note that the present constraint on the cardinality of the representatives $|\mathcal{Z}| \leq N$ in the problem formulation restricts the compression rate $I(\mathsf{y}; \mathsf{z})$ in any case. Thus, even for the extreme case of $\beta \to \infty$, the compression rate is upper-bounded by $I(\mathsf{y}; \mathsf{z}) \leq \log_2(N)$ bits.

III. Considered Routines

A. Iterative Information Bottleneck (It-IB)

After introducing the IB framework in [3], Tishby et al. exploited the variational calculus to deduce the optimal mapping

$$p(z|y) = \frac{p(z)}{\psi(y, \beta)} e^{-\beta D_{\mathrm{KL}}\left(p(x|y) \| p(x|z)\right)} \quad (3)$$

as the stationary point of the objective function in (1). The normalization function $\psi(y, \beta)$ secures a valid mapping $p(z|y)$ for each $y \in \mathcal{Y}$ and $D_{\mathrm{KL}}(\cdot \| \cdot)$ is the Kullback-Leibler (KL) divergence[4]. It should be noted that the provided solution in (3) has an implicit form, since distributions $p(z)$ and $p(x|z)$ appearing on the right-side of (3) are related to the quantizer $p(z|y)$ through

$$p(z) = \sum_{y \in \mathcal{Y}} p(y) p(z|y) \quad (4)$$

and

$$p(x|z) = \frac{1}{p(z)} \sum_{y \in \mathcal{Y}} p(x, y) p(z|y) \; . \quad (5)$$

The principal idea behind the Iterative IB (It-IB) algorithm is to utilize the derived optimal mapping (3) in an iterative manner (commencing with a random valid distribution $p(z|y)$) through exploiting the consistency conditions (4) and (5).

For β being finite, the outcome will be usually a *soft*, i.e., *stochastic* mapping. Contrarily, letting $\beta \to \infty$, the resultant mapping becomes *hard*, i.e., *deterministic* ([14]).

[1] $|\cdot|$ denotes the cardinality (the number of elements) of a given set.

[2] 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 $I(\mathsf{a}; \mathsf{b}) \triangleq \sum_a \sum_b p(a, b) \log \frac{p(a,b)}{p(a)p(b)}$.

[3] Also known as *concave optimization*, is about finding the maxima of a convex function over a closed convex set. This is totally different compared to the *convex optimization* wherein the aim is to find the minimum of a convex function.

[4] The KL divergence, also known as relative entropy, between two probability distributions $p(a)$ and $q(a)$ over the same event space $\mathcal{A}$ of the random variable a, is defined as $D_{\mathrm{KL}}(p(a) \| q(a)) \triangleq \sum_{a \in \mathcal{A}} p(a) \log \frac{p(a)}{q(a)}$ [1]. The relation between MI and KL divergence is established through $I(\mathsf{a}; \mathsf{b}) = D_{\mathrm{KL}}(p(a, b) \| p(a)p(b))$.

Explicitly, in every iteration, each channel's output element y is mapped with probability 1 to the specific cluster z which shows the least KL divergence among all candidate bins.

B. KL-Means Information Bottleneck (KL-Means-IB)

As already discussed, in case of β being asymptotically large, (2) shall be regarded as the corresponding design setup. Considering the definition of MI as difference of entropies[5] it has been shown in [10] that the following holds

$$I(\mathsf{x};\mathsf{z}) = I(\mathsf{x};\mathsf{y}) - \big(H(\mathsf{x}|\mathsf{z}) - H(\mathsf{x}|\mathsf{y})\big) \quad (6a)$$
$$= I(\mathsf{x};\mathsf{y}) - \mathbb{E}_{\mathsf{y},\mathsf{z}}\left\{D_{\text{KL}}\big(p(x|\mathsf{y})\|p(x|\mathsf{z})\big)\right\} . \quad (6b)$$

Hence, maximization of the relevant information $I(\mathsf{x};\mathsf{z})$ translates into minimization of the expectation term in (6b), as the *available information* $I(\mathsf{x};\mathsf{y})$ is given and fixed (it is a function of the joint distribution $p(x, y)$, assumed to be known).

At this point, the connection of the present design problem to the renowned K-Means clustering method can be realized. The conventional K-Means algorithm [15] intends to cluster a number of points into K bins such that the average square Euclidean distance between the points and the empirical means of the clusters is minimized. This task is done in an iterative fashion through two distinct steps, namely the *assignment* and the *update* steps. In the assignment step, each particular point is allocated to the very bin with the closest respective mean among all candidates. Subsequently, in the update step clusters' representatives are recalculated as the corresponding means (hence the name).

An interesting interpretation of this approach of clustering (regarding the *model-based* derivation of K-Means [16] with isotropic spherical Gaussian noise assumption) from the communications perspective, is to treat the points to be clustered as noisy versions of K *originally* transmitted points[6] and aiming for learning the most fitting underlying constellation. This methodology can be generalized to apply different types of distortion measures. For instance, in [12] a specific family of divergences including the KL divergence has been considered.

One may note that irrespective of the specific choice of $y \in \mathcal{Y}$ it applies $\sum_{x\in\mathcal{X}} p(\mathsf{x}=x|\mathsf{y}=y) = 1$, introducing a $(|\mathcal{X}|-1)$-dimensional probability simplex, referred to as the *backward channel simplex*. Therefore, as suggested in [10], transforming the primary quantization space (i.e., the space in which y values are defined) to the backward channel simplex by considering $p(x|\mathsf{y}=y)$[7] and $p(x|\mathsf{z}=z)$(the pertinent points in the transformed space) instead of y and z, respectively and treating the KL divergence as the given distortion measure, the design problem (2) can be perceived as a special K-Means clustering task. Specifically, the KL-Means-IB [10] is initiated by choosing N different points $p(x|y)$ as the primary means. Subsequently, through the assignment step,

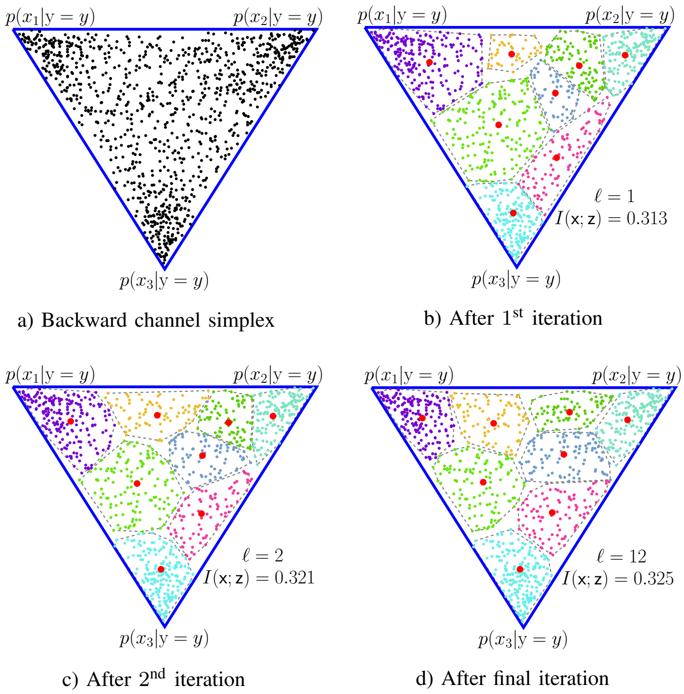

Fig. 2: Evolution of the KL-Means-IB result ($N=8$) through iterations ℓ for 3-PSK signaling over an AWGNC ($\sigma_n^2=1$) with the red dots demonstrating the clusters' representatives $p(x|z)$ and the acquired $I(\mathsf{x};\mathsf{z})$

each point $p(x|y)$ is clustered to the particular bin z for which the corresponding representative $p(x|z)$ shows the least KL divergence. Mathematically, $p(z|y) = \delta_{z,z^\star(y)}$[8] where the host cluster z for each particular y value is chosen as

$$z^\star(y) = \operatorname*{argmin}_z D_{\text{KL}}\big(p(x|y)\|p(x|z)\big) . \quad (7)$$

Next, clusters' representatives are updated as the bins' centers of mass [12]

$$p(x|z) = \frac{\sum\limits_{y\in\mathcal{Y}_z} p(y)p(x|y)}{\sum\limits_{y\in\mathcal{Y}_z} p(y)}, \quad (8)$$

wherein $\mathcal{Y}_z$ denotes the subset of $\mathcal{Y}$ for which all members are allocated to the bin z. The mentioned procedure is repeated till either a convergence criterion or a maximum number of iterations is met. To clearly visualize what happens in the backward channel simplex during the iterations of the KL-Means-IB, we demonstrate the corresponding 2-dimensional simplex for an example of 3-PSK signaling in Fig. 2. Explicitly, each point inside the drawn simplex corresponds to a particular $p(x|\mathsf{y}=y) = \{p(x_1|\mathsf{y}=y), p(x_2|\mathsf{y}=y), p(x_3|\mathsf{y}=y)\}$ for a certain y value received at the output of an additive white Gaussian noise channel (AWGNC) with the noise variance of unity ($\sigma_n^2=1$). The aforementioned clustering procedure for $N=8$ is then performed in this space and as can be seen, through iterations the clustering results are getting refined.

The complexity of the KL-Means-IB is dominated by its assignment step. Thus, as suggested by (7), per iteration, the KL-Means-IB has the complexity of $O(|\mathcal{X}|\cdot|\mathcal{Y}|\cdot|\mathcal{Z}|)$, since

[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 $H(\mathsf{a}) \triangleq -\sum_a p(a)\log p(a)$.

[6] In communications terminology it is referred to as the signal constellation.

[7] It is defined as $p(x|\mathsf{y}=y) \triangleq [p(x_1|\mathsf{y}=y), \cdots, p(x_{|\mathcal{X}|}|\mathsf{y}=y)]$. Each entry of this vector can be considered as a coordinate of the pertinent point.

[8] Here, δ represents the Kronecker delta function.

for each specific value y, the KL divergence to all $|\mathcal{Z}|$ possible candidates has to be calculated where each of these divergence calculations sums up $|\mathcal{X}|$ terms. It is worth noting that the KL-Means-IB is algorithmically equivalent to the It-IB under the assumption of β being asymptotically large [14].

C. Double Maxima Information Bottleneck (Double-Max-IB)

Quite recently, inspired by [17], a new heuristic has been proposed in [11] (called Algorithm 1) which henceforth we refer to as Double-Max-IB. This routine addresses the problem of distributed noisy source coding. To this end, the general system setup in [11] considers M different noisy observations (measurements) $\mathsf{y}_1, \mathsf{y}_2, \cdots, \mathsf{y}_M$ of the source x and the suggested solution encodes/quantizes them locally (but not independently, i.e., in a jointly fashion), such that the MI between the source and the random vector comprising all M individual representatives $\mathsf{z}_1, \mathsf{z}_2, \cdots, \mathsf{z}_M$ is maximized.

What makes this algorithm interesting is, as investigated in [11], in this fashion a set of *high-quality* general purpose quantizers is achieved that can be successfully employed for a variety of applications, e.g., the 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 specifically designed for estimation [19] or detection [20] purposes. In addition, the proposed routine can be readily adapted to the more realistic scenario of having non-ideal transmission channels from distributed sensors to the fusion center, thereby providing a novel distributed joint source-channel coding scheme.

Having the assumed system model of Fig. 1 in mind, to discuss the Double-Max-IB, we restrict ourselves to the case in which only one noisy observation is available, i.e., $M = 1$. Exploiting the chain rule of MI, the objective function in (2) can be expanded as

$$I(\mathsf{x};\mathsf{z}) = I(\mathsf{x},\mathsf{y};\mathsf{z}) - I(\mathsf{y};\mathsf{z}|\mathsf{x}) \tag{9a}$$

$$= H(\mathsf{x},\mathsf{y}) - H(\mathsf{x},\mathsf{y}|\mathsf{z}) - H(\mathsf{y}|\mathsf{x}) + H(\mathsf{y}|\mathsf{x},\mathsf{z}) \ . \tag{9b}$$

Since the entropies $H(\mathsf{x},\mathsf{y})$ and $H(\mathsf{y}|\mathsf{x})$ in (9b) are fixed (functions of the given joint distribution $p(x,y)$), it is directly concluded that (2) translates into

$$p^\star(z|y) = \underset{p(z|y)}{\operatorname{argmax}} \ \big[H(\mathsf{y}|\mathsf{x},\mathsf{z}) - H(\mathsf{x},\mathsf{y}|\mathsf{z})\big] \tag{10a}$$

$$= \underset{p(z|y)}{\operatorname{argmax}} \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}} p(x,y,z) \log p(x|z) \tag{10b}$$

$$= \underset{p(z|y)}{\operatorname{argmax}} \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}} p(x,y) p(z|y) \log p(x|z), \tag{10c}$$

where from (10b) to (10c) the assumed Markovian property is exploited. Assuming $q(y,z) = p(z|y)$ and $f(x,z)$ as an arbitrary function such that for each specific value $z \in \mathcal{Z}$ it applies $\sum_{x\in\mathcal{X}} f(\mathsf{x}=x, z) = 1$, the authors in [11] have defined a generalized objective function, L, as

$$L(q,f) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \sum_{z\in\mathcal{Z}} p(x,y)\, q(y,z)\, \log f(x,z) \ . \tag{11}$$

Then, utilizing the method of Lagrange multipliers, it has been shown that for a given $q(y,z)$, the optimal function $f^\star(x,z)$ that maximizes L is achieved by

$$f^\star(x,z) = \frac{\sum\limits_{y\in\mathcal{Y}} p(x,y) p(z|y)}{\sum\limits_{x'\in\mathcal{X}} \sum\limits_{y\in\mathcal{Y}} p(x',y) p(z|y)} \ . \tag{12}$$

Having the assumed Markovian property in mind and noting the respective marginalization of the joint distribution $p(x,y,z)$ at both the numerator and the denominator of (12) reveals that $f^\star(x,z) = \frac{p(x,z)}{p(z)}$ which is $p(x|z)$ by definition.

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

$$z^\star(y) = \underset{z}{\operatorname{argmax}} \sum_{x\in\mathcal{X}} p(x,y) \log f(x,z) \ . \tag{13}$$

As the objective function in (10c) is nothing else than the maximum of the generalized objective function L over $f(x,z)$ for a given $q(y,z)$, the pertinent optimization task can be secured by performing double (alternating) maximization of L over $f(x,z)$ and $q(y.z)$ in an iterative manner (hence the name). Explicitly, the Double-Max-IB algorithm is initialized to a valid random deterministic mapping $p(z|y)$ and then iterates over (12) (*update* step) and the resultant mapping by (13) (*assignment* step), till convergence to a local optimum.

Analogous to the KL-Means-IB, the complexity of the Double-Max-IB approach is dominated by its assignment step. Thus, as suggested by (13), per iteration, the Double-Max-IB routine has the complexity of $O(|\mathcal{X}|\cdot|\mathcal{Y}|\cdot|\mathcal{Z}|)$. Specifically, for each y value, the objective function in (13) that comprises the summation of $|\mathcal{X}|$ terms, must be calculated for all $|\mathcal{Z}|$ possible candidates.

The main contribution of this work lies in the next section, in which through an in-depth examination of the KL-Means-IB and the Double-Max-IB steps, we plainly demonstrate their algorithmic equivalence. While the KL-Means-IB reformulates the optimization (2) through a double (alternating) minimization [10], the Double-Max-IB offers a totally different approach of alternating maximization. Surprisingly, it turns out that they provide *identical* solution procedures.

IV. Stepwise Comparison of KL-Means-IB and Double-Max-IB Algorithms

A. Analysis

Basically, to prove the equivalence of the considered approaches, we have to illustrate that the respective assignment and update steps are identical for both routines.

Here, we embark on our analysis by considering the assignment step in Double-Max-IB. Substituting the resultant $f(x,z)$ from (12) to (13), for each $y \in \mathcal{Y}$ the allocated cluster z is determined by

$$z^\star(y) = \underset{z}{\operatorname{argmax}} \sum_{x\in\mathcal{X}} p(x,y) \log p(x|z) \tag{14a}$$

$$= \underset{z}{\operatorname{argmin}} \sum_{x\in\mathcal{X}} p(x,y) \big(-\log p(x|z)\big), \tag{14b}$$

wherein the maximization is substituted by minimization through introduction of the minus sign. (14b) can be further rewritten as

$$z^{\star}(y) = \underset{z}{\operatorname{argmin}} \sum_{x\in\mathcal{X}} p(x,y)\big(\log\frac{p(x|y)}{p(x|z)} - \log p(x|y)\big) \; . \quad (15)$$

Recalling $p(x,y) = p(y)p(x|y)$ and noting that the respective minimization in (15) is independent of $p(y)$, it applies

$$z^{\star}(y) = \underset{z}{\operatorname{argmin}} \sum_{x\in\mathcal{X}} p(x|y)\big(\log\frac{p(x|y)}{p(x|z)} - \log p(x|y)\big) \; . \quad (16)$$

Expanding the objective function in (16) yields

$$z^{\star}(y) = \underset{z}{\operatorname{argmin}}\Big(\sum_{x\in\mathcal{X}} p(x|y)\log\frac{p(x|y)}{p(x|z)} - \sum_{x\in\mathcal{X}} p(x|y)\log p(x|y)\Big)$$

or equivalently

$$z^{\star}(y) = \underset{z}{\operatorname{argmin}}\Big(D_{\mathrm{KL}}\big(p(x|y)\|p(x|z)\big) + H(x|\mathsf{y}=y)\Big) \; . \quad (17)$$

Since the conditional entropy $H(x|\mathsf{y}{=}y)$ in (17) is fixed by the given distribution $p(x,y)$, the ultimate cluster allocation's rule for Double-Max-IB is determined as

$$z^{\star}(y) = \underset{z}{\operatorname{argmin}}\, D_{\mathrm{KL}}\big(p(x|y)\|p(x|z)\big) \; . \quad (18)$$

Considering (18) and (7) together, it is immediately realizable that the assignment steps of the KL-Means-IB and the Double-Max-IB routines are identical.

Next, we consider the corresponding update steps. Specifically, regarding (12) and (8) it can be readily seen that both algorithms update the same distribution $p(x|z)$. Nonetheless, to clearly discern that (12) is indeed identical to (8), one may note that the mapping $p(z|y)$ in (12) is deterministic, i.e., it is equal to 1 iff $y \in \mathcal{Y}_z$. Moreover, since $p(x,y) = p(y)p(x|y)$, it applies

$$\sum_{y\in\mathcal{Y}} p(x,y)p(z|y) = \sum_{y\in\mathcal{Y}_z} p(x,y) = \sum_{y\in\mathcal{Y}_z} p(y)p(x|y) \; . \quad (19)$$

Thus, it becomes clear that the numerators in (12) and (8) are the same. In addition, the present summation over all $x'{\in}\mathcal{X}$ at denominator of (12) marginalizes the joint distribution $p(x',y)$ into $p(y)$ and consequently it applies

$$\sum_{x'\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x',y)p(z|y) = \sum_{x'\in\mathcal{X}}\sum_{y\in\mathcal{Y}_z} p(x',y) = \sum_{y\in\mathcal{Y}_z} p(y) \; . \quad (20)$$

Therefore, the identicality of both denominators becomes evident as well. All in all, via the performed analysis, we plainly proved the algorithmic equivalence of KL-Means-IB and Double-Max-IB. This brings about the profound insight that although these heuristics aim to solve the design problem in (2) through different approaches, surprisingly, they eventually provide exactly *identical* solution procedures.

Please note that since both heuristics converge to a local optimum, their respective outcome heavily depends on the choice of initialization. So, it can be asserted that assuming a sufficiently large number of runs (to achieve independence from initialization), both routines engender the same mapping.

B. Simulation Results

In this part, we set out to investigate the performance of discussed approaches over a typical digital transmission scenario. Specifically, we consider the equiprobable bipolar 4-ASK signaling ($\mathcal{X}{=}\{\pm1,\pm3\}$) at the input with the variance $\sigma_x^2 = 5$. To obtain the transition probability distribution $p(y|x)$, we firstly clip the corresponding conditional probability density functions (pdf) of an 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 interval of $3\sigma_n$ to assure 99.7% coverage) and then uniformly discretize them into $|\mathcal{Y}| = 128$ parts.

First, we aim for comparing conventional and IB-based quantization approaches. To do so, as a prevalent RD-based quantizer, we deploy the well-known Lloyd-Max algorithm [21], [22] wherein the square Euclidean distance is chosen as distortion measure function. As an IB-based approach, we consider the It-IB with $\beta = 400$ which results in (an almost) hard clustering. Fig. 3 depicts the *mutual information loss* defined as $\Delta I{=}I(\mathsf{x};\mathsf{y}){-}I(\mathsf{x};\mathsf{z})$ for different allowed number of clusters N. It shall be mentioned that to obtain the corresponding curves, each algorithm was run for $U{=}10^5$ different initializations with the best result taken. It is observed that irrespective of the channel quality (noise variance), the It-IB algorithm always outperforms the Lloyd-Max routine in preservation of the information (about the source) contained in the quantizer output. Although the performances are comparable for relatively high signal-to-noise ratio (SNR) values, the results suggest that in general, the superiority of the It-IB performance becomes more tangible in critical cases of degraded channel quality (i.e., the region of low SNR values).

Next, we focus on the performances of the three discussed IB-based routines. For that, Fig. 4 is generated wherein the noise variance is assumed to be $\sigma_n^2 = 1$. It can be seen from Fig. 4 a) that for $\beta = 400$, the performances of the It-IB and the KL-Means-IB are (almost) the same over the entire range of N. The reason behind is thoroughly discussed in [14] in which the algorithmic equivalence of these two routines is shown for $\beta \to \infty$. Regarding the KL-Means-IB and the Double-Max-IB curves, at first glance it may seem that their performances are not the same over the entire range of N. Nonetheless, it can be observed that by increasing U the respective curve of the KL-Means-IB is swept by the Double-Max-IB for higher values of N. This clearly indicates that by increasing U, the present gap between the performances can be bridged for higher and higher values of N. In other words, as will be shown subsequently, for sufficiently large values of U (to assure independence from choice of initialization), irrespective of N, both routines produce the same result. To get an impression about the complexity-precision trade-off, the corresponding compression rates are drawn in Fig. 4 b). The main message inferable is the higher the required precision, the higher the pertinent complexity (in information theoretic sense, i.e., the resultant compression rate $I(\mathsf{y};\mathsf{z})$ by provided mapping $p(z|y)$) as well.

As a final inquiry, instead of initializing the Double-Max-IB

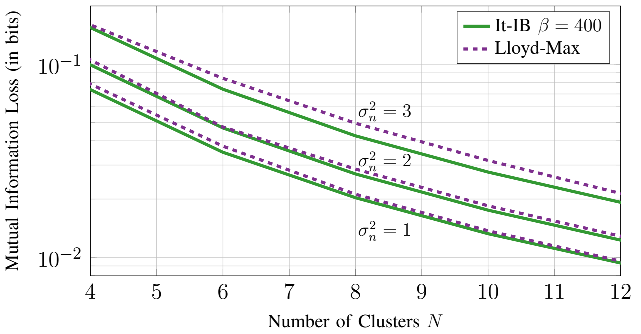

Fig. 3: Mutual information loss ΔI vs. the allowed number of clusters N for RD and IB-based quantizers, AWGN channel with different noise variances

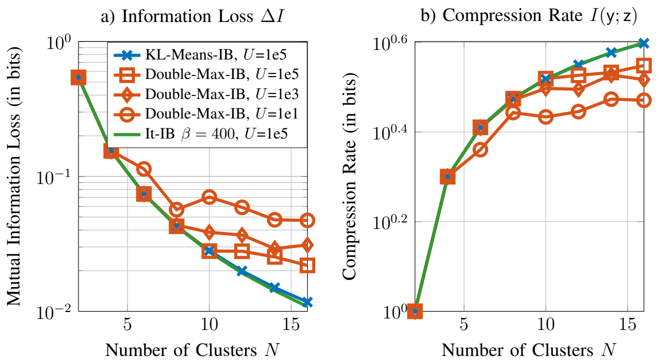

Fig. 4: a) Mutual information loss ΔI and b) compression rate $I(\mathsf{y};\mathsf{z})$ vs. the allowed number of clusters N, AWGN channel with $\sigma_n^2 = 1$

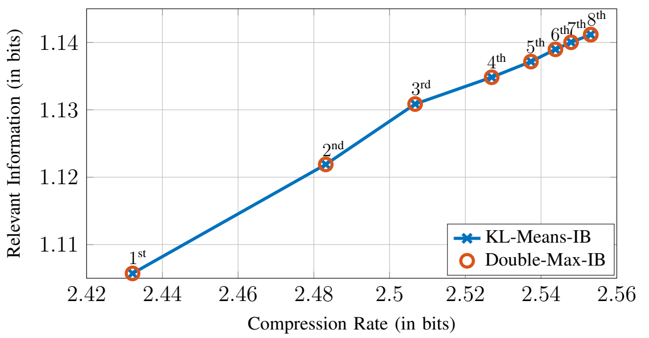

Fig. 5: Relevant information $I(\mathsf{x};\mathsf{y})$ vs. compression rate $I(\mathsf{y};\mathsf{z})$ through iterations, AWGN channel with $\sigma_n^2 = 1$, $N = 8$

algorithm by any random deterministic mapping, we fed it by the resultant mapping at the end of the first iteration of the KL-Means-IB routine and registered the evolution of the acquired solution (with the arbitrary choices of $N = 8$ and σ_n^2=1) for both algorithms over the iterations. Fig. 5 illustrates the corresponding results. At this point, it becomes clear that initializing both routines identically, their behavior through convergence would be exactly the same.

V. Summary

In this paper, we firstly presented the Information Bottleneck framework as an alternative to the conventional methods on the subject of noisy source coding and then discussed three individual heuristics appeared in the literature to address the corresponding design problem. Afterwards, we analyzed the relation between KL-Means-IB and Double-Max-IB routines and plainly proved their equivalence via the scrutiny of the respective algorithmic steps they make to produce the favorable result. Finally, we substantiated our conducted analysis by performing simulations over a typical transmission scenario.

Acknowledgment

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

References

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

[2] E. Ayanoglu, "On Optimal Quantization of Noisy Sources," *IEEE Transactions on Information Theory*, vol. 36, no. 6, pp. 1450–1452, 1990.

[3] 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.

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

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

[6] I. Tal and A. Vardy, "How to Construct Polar Codes," *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6562–6582, Oct. 2013.

[7] J. Lewandowsky and G. Bauch, "Trellis based Node Operations for LDPC Decoders from the Information Bottleneck Method," in *9th Int. Conference on Signal Processing and Communication Systems (ICSPCS 2015)*, Cairns, Australia, Dec. 2015.

[8] F. J. C. Romero and B. M. Kurkoski, "LDPC Decoding Mappings That Maximize Mutual Information," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 9, pp. 2391–2401, 2016.

[9] B. M. Kurkoski and H. Yagi, "Quantization of Binary-Input Discrete Memoryless Channels," *IEEE Transactions on Information Theory*, vol. 60, no. 8, pp. 4544–4552, Aug. 2014.

[10] A. Zhang and B. M. Kurkoski, "Low-Complexity Quantization of Discrete Memoryless Channels," in *Int. Symposium on Information Theory and Its Applications (ISITA)*, Monterey, CA, USA, Oct. 2016.

[11] 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.

[12] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman Divergences," *The Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, Oct. 2005.

[13] 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.

[14] 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.

[15] A. K. Jain, "Data Clustering: 50 Years Beyond K-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, June 2010.

[16] A. D. Gordon, *Classification: Methods for the Exploratory Analysis of Multivariate Data*, Chapman and Hall, 1981.

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

[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 Transactions on Information Theory*, vol. 36, no. 2, pp. 241–255, 1990.

[21] S. P. Lloyd, "Least Squares Quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.

[22] J. Max, "Quantizing for Minimum Distortion," *IRE Transactions on Information Theory*, vol. 6, no. 1, pp. 7–12, Mar. 1960.