On the Relation Between the Asymptotic Performance of Different Algorithms for Information Bottleneck Framework

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

Brian M. Kurkoski
School of Information Science
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan 923-1292
Email: kurkoski@jaist.ac.jp

***Abstract*—The general problem of quantizing observation signals appears in different aspects of data processing, from special code designs to realization of low-complexity receivers. To this end, a new framework, known as the Information Bottleneck method, has recently attracted a great deal of attention. In this paper, after introducing this framework and providing the Iterative Information Bottleneck algorithm as the primary pertinent solution, we also discuss three other heuristics aiming to solve the similar problem efficiently. Since the resultant solution of considered approaches is locally optimum, it strongly depends on the choice of initialization. The main contribution of this work is to prove the equivalence of these algorithms asymptotically, i.e., assuming an infinite run of algorithms for the extreme case of infinitely large trade-off parameter. We also substantiate this claim by means of computer-based simulations.**

I. Introduction

Regarding communication systems, the quantization of noisy observations of the original source signal is a fundamental task. The Rate-Distortion (RD) theory deals with the underlying complexity-precision trade-off [1]. Explicitly, for an upper-bound on the tolerable value of a given distortion measure, the Blahut-Arimoto algorithm provides the minimal number of bits per symbol in order to represent the received signal [2]. This conventional formulation has some drawbacks. First of all, it does not suggest a systematic way to choose a proper distortion measure for any case of relevance. Irrespective of the characteristics of the signals and for the sake of simplicity, a common choice in many applications is the mean square error (MSE) between the noisy observation as quantizer input signal and its representant at the output. Obviously, this may not be always the best choice. Furthermore, the stochastic relation between the original data source and its noisy observation is not considered by traditional design setup.

With the so-called Information Bottleneck (IB), Tishby et al. have introduced an alternative approach for data compression [3]. There, the focal idea is to compress the observation in a way that the quantizer output preserves most of the information about the relevant variable, e.g., the original source. Contrary to the RD theory, the IB formulation obviates the a priori specification of a distortion measure by considering the mutual information between the original data source and the quantizer output. Moreover, purely dealing with entropy calculations for which only distributions are required results in an arbitrary choice of the representation set, e.g., a finite set of integers at the quantizer output. As a result, a compact representation of the quantizer input signal is achieved that is highly informative about the actual source of interest.

A major subject in learning theory which comprises a significant part of techniques dealing with the problem of unsupervised learning is dimensionality reduction through clustering [4]. In addition to the learning theory motivations that led to the advent of the IB concept, similar problems of quantization/compression arise in different aspects of data transmission like construction of polar codes [5], analog-to-digital converter (ADC) at receiver front-ends [6] and many other potential cases.

This paper is structured as follows: the general IB framework is introduced in Section II. Next, four different algorithmic approaches are presented in Section III. The main contribution of this work lies in Section IV where we show the equivalence of all considered algorithms when letting the IB trade-off parameter grow asymptotically large, i.e., $\beta \to \infty$. In the same section, performance results are presented to corroborate the claimed argumentation. Finally, we conclude by providing a summary of this work in Section V.

II. Information Bottleneck Method

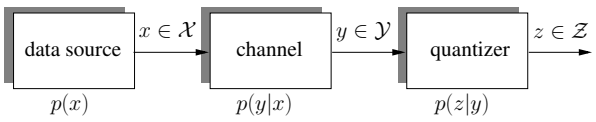

Fig. 1. General system model for the quantization of noisy observations

We consider the system model shown in Fig. 1 comprising a data source and a transmission channel, both assumed to be discrete and memoryless, followed by a quantizer. The source is modeled by the random variable x taking realizations $x \in \mathcal{X}$ according to the probability mass function (pmf) $p(x)$. The channel is characterized by its transition probability distribution $p(y|x)$ and its output is modeled by the random variable y with realizations $y \in \mathcal{Y}$. Finally, we take the random variable z with realizations $z \in \mathcal{Z}$ as the output of the quantization block that is characterized by the conditional distribution $p(z|y)$. In what follows, the mutual information between x and y is denoted by $I(\mathsf{x};\mathsf{y}) = H(\mathsf{x}) - H(\mathsf{x}|\mathsf{y})$ with the source entropy $H(\mathsf{x})$ and the conditional entropy $H(\mathsf{x}|\mathsf{y})$.

978-1-4673-8999-0/17/$31.00 ©2017 IEEE

The quantizer design problem would then be as follows: assume the joint probability distribution $p(x,y)=p(x)\,p(y|x)$ is given and $\mathsf{x} \leftrightarrow \mathsf{y} \leftrightarrow \mathsf{z}$ is a Markov chain. The quantizer output z shall then be a compact representant of its input y which keeps most of the information it contains about the source x. Mathematically, in the design formulation the present trade-off among the *compression rate*, $I(\mathsf{y};\mathsf{z})$, and the *relevant information*, $I(\mathsf{x};\mathsf{z})$, is established through a non-negative Lagrange multiplier, β. Therefore, denoting the allowed number of quantizer output levels by n, the relative design problem follows as[1]:

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

Two important questions must be answered at this point to be able to find out the entity of the optimization task at hand:

1) What is the event space of the mapping made by the quantizer $p(z|y)$?
2) Is the objective function in (1) concave or convex over the pertinent event space?

To address the first question, one may note that the resultant $p(z|\mathsf{y}=y)$ for each specific value y of the random variable y is a $(|\mathcal{Z}|-1)$-dimensional probability simplex, simply due to the fact that $\sum_{z\in\mathcal{Z}} p(\mathsf{z}=z|\mathsf{y}=y)=1$ holds. Consequently, the overall event space of the mapping $p(z|y)$ would be the product set of $|\mathcal{Y}|$ of such simplices that is a closed convex polytope in the $|\mathcal{Y}|\times(|\mathcal{Z}|-1)$ Euclidean space [7].

To address the second question, we subsequently cover the entire interval of allowed values of β by considering three different cases, specifically two extreme cases of $\beta\to 0$ and $\beta\to\infty$ and the third case of finite values.

Letting $\beta\to 0$, the objective function in (1) will be reduced to the compression rate $I(\mathsf{y};\mathsf{z})$. It is well known that $I(\mathsf{y};\mathsf{z})$ is a convex function of $p(z|y)$ for fixed $p(y)$ [1]. Thus, the optimization problem is *convex* and any valid stochastic allocation of y to $|\mathcal{Z}|$ bins which is repeated for all $y\in\mathcal{Y}$ would be a solution. In this fashion, the compression rate takes its global minimum value of $I(\mathsf{y};\mathsf{z})=0$ since y and z become statistically independent. Evidently, $\beta\to 0$ is not of any interest, as no relevant information is kept.

Letting $\beta\to\infty$, i.e., aiming to keep as much relevant information as possible, the design problem (1) reduces to

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

Recalling the present Markov chain $\mathsf{x} \leftrightarrow \mathsf{y} \leftrightarrow \mathsf{z}$, the conditional probabilities $p(z|x)$ and $p(z|y)$ are connected through $p(z|x)=\sum_{y\in\mathcal{Y}} p(z|y)p(y|x)$. This relation is of affine type which preserves convexity. Since $I(\mathsf{x};\mathsf{z})$ is convex w.r.t. $p(z|x)$ for fixed $p(x)$, it will also be a convex function of $p(z|y)$. As a result, the maximization in (2) becomes a *concave* optimization problem[2] [8]. A well-known proposition in concave optimization theory [9] asserts, that a convex function $f:\mathcal{D}\to\mathbb{R}$ attains its global maximum over $\mathcal{D}$ at an extreme point of $\mathcal{D}$. Thus, it can be shown that there exists an optimal solution of deterministic type, i.e., $p(z|y)\in\{0,1\}$ for all pairs of $(y,z)\in\mathcal{Y}\times\mathcal{Z}$. To this end, one shall note that extreme points of a convex polytope translate into its vertices. For the event space of the mapping $p(z|y)$, each vertex corresponds to the product set of vertices of its constituent probability simplices, leading to a deterministic mapping for each $y\in\mathcal{Y}$. As the naive exhaustive search over all $|\mathcal{Z}|^{|\mathcal{Y}|}$ vertices of the event space of quantizer mappings $p(z|y)$ is evidently intractable for the relatively large cardinality of elements to be clustered, most heuristics exploit the existence of a deterministic solution for (2) to solve the problem at least locally. It is noteworthy, that maximizing the relevant information $I(\mathsf{x};\mathsf{z})$ in (2) results in the highest achievable rate between x and z.

In case for which β takes finite values, the objective function in (1) is the sum of a concave $(\frac{-\beta}{\beta+1}I(\mathsf{x};\mathsf{z}))$ and a convex $(\frac{1}{\beta+1}I(\mathsf{y};\mathsf{z}))$ function of the mapping $p(z|y)$ which in general, is neither convex nor concave. Therefore, also the present optimization would be of neither concave nor convex type. This, in fact, makes the task of finding the optimal mapping to be very difficult. As a result, a number of heuristics have been developed aiming to converge to locally optimal solutions for each value of β. Some of these algorithmic approaches are presented in Section III. Interested readers are referred to [10] for more details.

It must be noticed that the present constraint on the cardinality of the representation set $|\mathcal{Z}|\leq n$ in the problem setup always restricts the compression rate $I(\mathsf{y};\mathsf{z})$. Hence, 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. Algorithmic Approaches

In this section, we discuss four different algorithms developed to deal with the IB-based quantizer design problem. Please note that in what follows, bin and cluster refer to the same concept and hence are used interchangeably.

A. Iterative Information Bottleneck (It-IB)

In [3], Tishby et al. derived the optimal quantizer mapping for (1) by means of variational calculus. Explicitly, for a specific value of β the mapping $p(z|y)$ is a stationary point of the objective function in (1), if and only if

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

is met for all pairs $(y,z)\in\mathcal{Y}\times\mathcal{Z}$. The function $\psi(y,\beta)$ normalizes the mapping $p(z|y)$ to ensure a valid distribution for each $y\in\mathcal{Y}$ and $D_{\mathrm{KL}}(\cdot\,\|\,\cdot)$ is the Kullback-Leibler (KL) divergence[3]. The derived optimal mapping in (3) has an implicit form, as the cluster representatives (in a conventional

[1]Note that the mapping $p^{\star}(z|y)$ is independent of the present multiplier $\frac{1}{\beta+1}$. It is only considered for the sake of mathematical clarity when investigating the extreme cases of asymptotically small/large values of β.

[2]Please note that *concave optimization* is about searching for the *maxima* of a *convex function* $(\cup)$ over a feasible region and, thus, is essentially different from *convex optimization* that aims for the *minima* of a *convex function*.

[3]The KL divergence, also known as relative entropy, between two probability distributions $p(x)$ and $q(x)$ over the same event space $\mathcal{X}$ of the random variable x, is defined as $D_{\mathrm{KL}}\big(p(x)\|q(x)\big)=\sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}$ [1].

sense) $p(x|z)$ and the cluster probability $p(z)$ appearing on the right side of (3), depend on the quantizer mapping $p(z|y)$ by

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

and

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

The Iterative IB (It-IB) algorithm is initialized by a valid random mapping $p(z|y)$ and then iterates over (4), (5) and (3), until a specific convergence criterion is met. It is important to mention that for finite values of β, the resultant mapping $p(z|y)$ is of *stochastic* nature, i.e., each y is mapped to all bins z with a certain probability. Usually, the iterative procedure needs to be repeated for different initializations, as the algorithm converges only to a locally optimal solution.

B. Deterministic Information Bottleneck (Det-IB)

In [11], Strouse and Schwab introduced the generalized objective function as

$$\mathcal{L}_\alpha = H(\mathsf{z}) - \alpha\, H(\mathsf{z}|\mathsf{y}) - \beta\, I(\mathsf{x};\mathsf{z}) \tag{6}$$

where the parameter $\alpha \in [0,1]$. The stochastic nature of the solution provided by the IB algorithm stems from the presence of the term $H(\mathsf{z}|\mathsf{y})$ in the respective functional (1) that coincides with (6) for $\alpha = 1$. Letting $\alpha \to 0$, i.e., trying to suppress the origin of stochasticity, leads to a solution $p(z|y)$ which, contrary to the resultant mapping from IB algorithm, is of *deterministic* type even for finite values of β.

Exploiting variational calculus once again, for a specific value of α the optimal mapping is found as

$$p(z|y) = \frac{1}{\psi(y,\alpha,\beta)} e^{\frac{1}{\alpha}\Big(\log p(z) - \beta D_{\mathrm{KL}}\left(p(x|y)\|p(x|z)\right)\Big)} \tag{7}$$

in which the normalization function $\psi(y,\alpha,\beta)$ ensures a valid mapping $p(z|y)$ for each $y \in \mathcal{Y}$. Clearly, by decreasing the value of α, the power of the exponential function in (7) grows asymptotically large. Therefore, letting $\alpha \to 0$ the mapping $p(z|y)$ for each $y \in \mathcal{Y}$ degenerates to a delta function which results in a deterministic quantizer. Mathematically, the resultant mapping for each $y \in \mathcal{Y}$ is given by $p(z|y) = \delta_{z,z^\star(y)}$ with the Kronecker function δ and the optimum cluster $z^\star(y)$ for quantizer input signal y is obtained as

$$z^\star(y) = \underset{z}{\operatorname{argmax}} \Big(\log p(z) - \beta D_{\mathrm{KL}}\left(p(x|y)\|p(x|z)\right)\Big) \ . \tag{8}$$

Similar to the It-IB (3), the provided solution (7) is of implicit form. Consequently, to achieve the required quantizer mapping $p(z|y)$, the Deterministic IB (Det-IB) algorithm is initialized with a valid random deterministic mapping $p(z|y)$ and iterates over equations (4), (5), and (8), until a specific convergence criterion is met. It is important to mention, that in general the resultant mapping does not use the entire allowed number of bins, i.e., $|\mathcal{Z}| < n$, as the term $\log p(z)$ in (8) encourages the assignment of elements into already used clusters.

C. KL-means Information Bottleneck (KL-means-IB)

In the extreme case of $\beta \to \infty$ as already discussed, the general IB problem (1) is reduced to finding the mapping $p(z|y)$ that maximizes the relevant information $I(\mathsf{x};\mathsf{z})$. Considering the definition of mutual information and the Lemma for the difference of conditioned uncertainties provided in [12], for the given Markov chain $\mathsf{x} \leftrightarrow \mathsf{y} \leftrightarrow \mathsf{z}$ one can write

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

$$= I(\mathsf{x};\mathsf{y}) - \mathbb{E}_{\mathsf{y},\mathsf{z}}\left\{D_{\mathrm{KL}}\big(p(x|\mathsf{y})\|p(x|\mathsf{z})\big)\right\} \ . \tag{9b}$$

It is readily seen that the maximization of the relevant information $I(\mathsf{x};\mathsf{z})$ corresponds to the minimization of the average KL divergence $\mathbb{E}_{\mathsf{y},\mathsf{z}}\left\{D_{\mathrm{KL}}\big(p(x|\mathsf{y})\|p(x|\mathsf{z})\big)\right\}$, since the mutual information $I(\mathsf{x};\mathsf{y})$ is fixed. With an analogous approach to the Lloyd-Max algorithm [13], [14], the KL-means-IB algorithm finds a locally optimal solution by alternating minimization in the mapping $p(z|y)$ (assignment step) and in the conditional probability distribution $p(x|z)$ (update step). As a main difference, the squared Euclidean norm used within the Lloyd-Max algorithm is substituted by the KL divergence that is the proper distance measure for the IB setup.

In the initialization, the KL-means-IB algorithm picks randomly n conditional probability distributions $p(x|\mathsf{y} = y)$ corresponding to n different values of y as means of clusters. Then, through the assignment step the points[4] with smallest KL divergence to each mean, are clustered in the same bin. Mathematically, let $\mathcal{Y}_z$ denote the subset of $\mathcal{Y}$ mapped to the cluster z with corresponding mean value $p(x|z)$. For $z' \neq z$, the following holds

$$\mathcal{Y}_z = \Big\{ y \,|\, D_{\mathrm{KL}}\big(p(x|y)\|p(x|z)\big) \leq D_{\mathrm{KL}}\big(p(x|y)\|p(x|z')\big) \Big\} \ .$$

Subsequently, in the update step, the mean of each cluster is calculated as its corresponding center of mass [15]. Explicitly, the corresponding mean for the cluster z is calculated as

$$p(x|z) = \frac{\sum\limits_{y \in \mathcal{Y}_z} p(y)p(x|y)}{\sum\limits_{y \in \mathcal{Y}_z} p(y)} \ . \tag{10}$$

The iteration between the assignment and the update steps is continued until either a specific convergence criterion is met or a maximum number of iterations is reached.

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

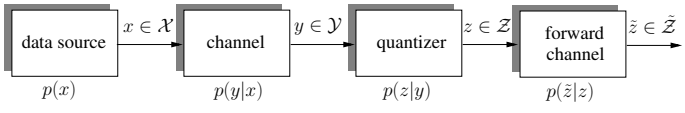

Fig. 2. The extended system model featuring forward channel

With the underlying assumption of $\beta \to \infty$, Winkelbauer et al. extended the quantizer design problem through the transmission of the quantizer output signals z over a forward channel [16]. The new model is depicted in Fig. 2. Denoting

[4]One shall note that each conditional probability distribution $p(x|\mathsf{y} = y)$ can be interpreted as a point in the space of dimension $|\mathcal{X}|$.

the forward channel output by $\tilde{\mathsf{z}}$ with realizations $\tilde{z} \in \tilde{\mathcal{Z}}$ and characterizing the extra discrete memoryless channel (DMC) by the transition probability distribution $p(\tilde{z}|z)$, the IB problem in (2) can then be reformulated as

$$p^\star(z|y) = \underset{p(z|y)}{\operatorname{argmax}} \; I(\mathsf{x};\tilde{\mathsf{z}}) \;\; \text{for } |\mathcal{Z}| \leq n \; . \tag{11}$$

For the Markov chain $\mathsf{x} \leftrightarrow \mathsf{y} \leftrightarrow \mathsf{z} \leftrightarrow \tilde{\mathsf{z}}$, the optimization in (11) is again of *concave* type, as $I(\mathsf{x};\tilde{\mathsf{z}})$ is a convex function of $p(\tilde{z}|x)$ for fixed $p(x)$ and for a given forward channel $p(\tilde{z}|z)$, $p(z|y)$ and $p(\tilde{z}|x)$ are connected by an affine relation which preserves convexity. Evidently, the problem (11) becomes equal to (2) for special case of error-free forward channel, since in this case no loss of information occurs from z to $\tilde{z}$. With $I(\mathsf{x};\tilde{\mathsf{z}}) = I(\mathsf{x};\mathsf{y}) - I(\mathsf{x};\mathsf{y}|\tilde{\mathsf{z}})$ and for a given constant $I(\mathsf{x};\mathsf{y})$, the maximization of $I(\mathsf{x};\tilde{\mathsf{z}})$ in (11) translates to the minimization of $I(\mathsf{x};\mathsf{y}|\tilde{\mathsf{z}})$. Defining $C(\mathsf{y}=y,\tilde{\mathsf{z}}=\tilde{z}) = D_{\mathrm{KL}}\big(p(x|y)\|p(x|\tilde{z})\big)$ the relation

$$I(\mathsf{x};\mathsf{y}|\tilde{\mathsf{z}}) = \mathbb{E}_{\mathsf{y}}\big\{\mathbb{E}_{\tilde{\mathsf{z}}}\{C(\mathsf{y},\tilde{\mathsf{z}})|\mathsf{y}\}\big\} \tag{12}$$

has been derived in [16], in which the conditional expectation is calculated as

$$\mathbb{E}_{\tilde{\mathsf{z}}}\{C(\mathsf{y},\tilde{\mathsf{z}})|\mathsf{y}\} = \sum_{z\in\mathcal{Z}} p(z|y) \sum_{\tilde{z}\in\tilde{\mathcal{Z}}} p(\tilde{z}|z)\, C(\mathsf{y}=y,\tilde{\mathsf{z}}=\tilde{z}) \; . \tag{13}$$

Aiming to minimize (13) for each $y \in \mathcal{Y}$, the corresponding mapping is chosen as $p(z|y) = \delta_{z,z^\star(y)}$ in which the optimum cluster $z^\star(y)$ is obtained by

$$z^\star(y) = \underset{z}{\operatorname{argmin}} \sum_{\tilde{z}\in\tilde{\mathcal{Z}}} p(\tilde{z}|z) C(\mathsf{y}=y,\tilde{\mathsf{z}}=\tilde{z}) \; . \tag{14}$$

Therefore, the conditional probability distribution $p(\tilde{z}|y)$ of the combination of the forward channel and the quantizer calculates as

$$p(\tilde{z}|y) = \sum_{z\in\mathcal{Z}} p(\tilde{z}|z)p(z|y) = p(\tilde{z}|z^\star(y)) \; . \tag{15}$$

It is clear that in this fashion, the conditional mutual information (12) is minimized for a given $C(\mathsf{y},\tilde{\mathsf{z}})$. The main idea behind the Ch-Opt-IB algorithm is to adapt the iterative IB discussed in Section III-A to the new model of Fig. 2. Precisely, it is initialized by a random $C(\mathsf{y}=y,\tilde{\mathsf{z}}=\tilde{z})$ for all $(y,\tilde{z}) \in \mathcal{Y}\times\tilde{\mathcal{Z}}$ and iterates over the modified versions of (4), (5) and (15) (substituting z by $\tilde{z}$), until a specific convergence criterion is met. Obviously, $C(\mathsf{y},\tilde{\mathsf{z}})$ is updated accordingly after each iteration. It is noteworthy that in [17], the special case of this algorithm, assuming an ideal forward channel has already been proposed.

IV. Algorithmic Step Comparison

As already mentioned, since the resultant mapping $p(z|y)$ of the considered heuristics is locally optimal, it strongly depends on the choice of initialization. In this part we show that asymptotically, i.e., for an infinite run of algorithms which are initialized randomly (to assure independence from initialization) and for $\beta \to \infty$, all discussed algorithms in Section III will perform equivalently. We also provide corresponding simulation results to confirm our reasoning.

A. Argumentation

We start this part by expressing the line of reasoning provided in [18] which proves the equivalence of the KL-means-IB and the It-IB algorithms. We further investigate the other two algorithms, namely the Ch-Opt-IB and the Det-IB algorithms, and clearly show that through convergence to the resultant mapping $p(z|y)$, they perform equivalent algorithmic steps. Moreover, we point out that the algorithmic step equivalence of the Ch-Opt-IB and the Det-IB algorithms is exactly in the same fashion of the KL-means-IB and the It-IB algorithms. Hence, in such a way, the asymptotic equivalence of all four presented approaches is proven.

Starting with the It-IB algorithm and assuming $\beta \to \infty$, for each observation $y \in \mathcal{Y}$, (3) degenerates to a hard clustering in which the bin with the minimum KL divergence among all is chosen with probability of 1. To discern this, one may note that letting $\beta \to \infty$, the exponential term in (3) becomes asymptotically small for all different values of z but the one with minutest KL divergence becomes small least. As a result, after the normalization done by the function $\psi(y,\beta)$, the respective values of $p(z|y)$ for all other bins tend to zero. Mathematically, $p(z|y) = \delta_{z,z^\star(y)}$ where the host cluster $z^\star(y)$ is chosen as

$$z^\star(y) = \underset{z}{\operatorname{argmin}}\, D_{\mathrm{KL}}\big(p(x|y)\|p(x|z)\big) \; . \tag{16}$$

This is the same as the assignment step in the KL-means-IB algorithm, as performing (16) for all observation values $y \in \mathcal{Y}$ will be the resultant mapping by Section III-C. Next, we focus on the corresponding update steps. In the It-IB algorithm, (4) and (5) together can be regarded as the update step. With this in mind, comparing the respective steps in the It-IB algorithm and the KL-means-IB algorithm, it can be deduced that they are basically the same. This can be seen by noting the present deterministic mapping $p(z|y)$ for $\beta \to \infty$ in the It-IB algorithm. Since for a specific cluster z, $p(z|y) = 1$ only for $y \in \mathcal{Y}_z$ and having in mind that $p(x,y) = p(y)p(x|y)$, it becomes clear at this point that (5) together with (4) reduce to (10). Consequently, one can conclude that the It-IB and the KL-means-IB are algorithmically equivalent, as the assignment (mapping) and the update steps are identical for both.

Now we follow a similar procedure to investigate the Ch-Opt-IB and the Det-IB algorithms. Before commencing the analysis, one may note that in order to make these two algorithms comparable, the forward channel in the Ch-Opt-IB must be error-free, as only in this case the underlying extended framework in Fig. 2 reduces to the conventional IB setup in Fig. 1. In the following, w.l.o.g. we assume an ideal forward channel, i.e., $p(\tilde{z}|z) = \delta_{z,\tilde{z}}$. Starting with the Ch-Opt-IB and following the steps in Section III-D, it can be seen that for the ideal forward channel, determination of the optimum cluster $z^\star(y)$ in (14) degenerates to

$$z^\star(y) = \underset{z}{\operatorname{argmin}}\, C(\mathsf{y}=y,\mathsf{z}=z) \; . \tag{17}$$

Then, we consider the Det-IB for $\beta \to \infty$. In this case, as the first term in (8) becomes negligible compared to its second

term, determination of the optimum cluster $z^\star(y)$ reduces to

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

where the minus sign is dropped by changing the maximization to minimization. As the minimization task at hand is irrelevant to β, it has been dropped, too. At this point, one may observe that the cluster assignment step for each $y \in \mathcal{Y}$ is identical for both algorithms, since $C(\mathsf{y} = y, \mathsf{z} = z)$ in (17) is $D_{\mathrm{KL}}\big(p(x|y)\|p(x|z)\big)$ by definition. Hence, to prove the overall equivalence of both algorithms, it suffices to show that the respective update step is also the same for both approaches. One may note that both algorithms iterate over the same equations. Specifically, they use (4) and (5) to update their distributions with the minor difference that for the Ch-Opt-IB, z is substituted by $\tilde{z}$. But bearing in mind that for an ideal forward channel, $\tilde{z}$ and z are basically the same, one notices that this step is also completely equal for both approaches. As a result, it becomes apparent that although these two algorithmic approaches appear to be different at first glance, they are algorithmically equivalent when $\beta \to \infty$.

All in all, comparing the assignment and the update steps of the four discussed approaches, one can deduce that they all are algorithmically equivalent in an asymptotic sense of $\beta \to \infty$. Explicitly, the equivalence of the respective assignment step is readily seen by comparing (16) with (18). Moreover, concerning the update step, it must be noticed that (4) together with (5) is used for all heuristics. As a result, one infers that running all algorithms with the same initialization, e.g., taking the deterministic mapping at the end of the first iteration of the Ch-Opt-IB algorithm as the starting point of the other three algorithms, the resultant mapping would be the same for all.

B. Performance Assessment

For the numerical evaluations we apply equiprobable bipolar 4-ASK with $\sigma_x^2 = 5$ as input signal and assume AWGN channel with noise variance $\sigma_n^2 = 1$. Furthermore, to acquire the channel transition distribution $p(y|x)$, the continuous channel output values with the absolute value of less or equal than 6 (to set the border guard interval of 3 times the noise standard deviation) is uniformly discretized to $|\mathcal{Y}| = 128$ values. In particular, we investigate the accuracy by the *mutual information loss* $\Delta I = I(\mathsf{x};\mathsf{y}) - I(\mathsf{x};\mathsf{z})$ and the complexity-precision trade-off by the corresponding *compression rate* $I(\mathsf{y};\mathsf{z})$ for different values of β over varying the allowed number of clusters n. Moreover, to get an impression about the complexity of the considered algorithms their required runtime in MATLAB (C source MEX files) has been provided. Finally, based on statistics of the pertinent performance of the presented algorithms over a relatively large number of runs with random initialization, their average required time to acquire precision-specific results are also calculated.

Fig. 3 a) visualizes the mutual information loss ΔI over varying the allowed number of bins n in case of $\beta \to \infty$ with the exception of the It-IB for which β is set to 400 to preserve numerical stability. Different algorithms are run for different numbers of initializations, U, to investigate their performance

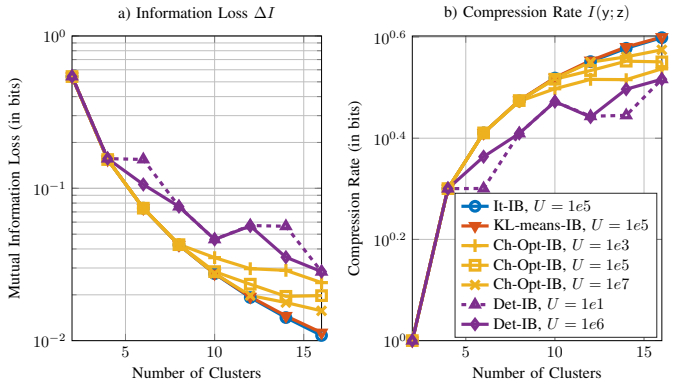

Fig. 3. a) Information loss ΔI and b) compression rate $I(\mathsf{y};\mathsf{z})$ over varying the allowed number of bins n

sensitivity to the size of the event space of their initialization[5]. One may observe that for $U = 10^5$, the performance of the It-IB and the KL-means-IB are almost the same. Furthermore, regarding the corresponding curves of the Ch-Opt-IB, it can be seen that by increasing the number of runs, the respective performance curve of the It-IB and the KL-means-IB is swept for higher number of bins. This clearly indicates, that irrespective of the chosen number of clusters, these three algorithms end up to the same performance and the only difference between them is about their required number of runs to achieve a precision-specific result. Considering the corresponding curves for the Det-IB, it is seen that by increasing the number of runs from $U = 10$ to $U = 10^6$, the respective performance does not change significantly. The observed non-smooth behavior is due to the fact, that usually the provided mapping by the Det-IB does not use the entire allowed number of bins, i.e., $|\mathcal{Z}| < n$. As an example, considering the corresponding curve for $U = 10$, increasing the allowed number of bins from $n = 4$ to $n = 6$ does not provide a significant gain, since the used number of clusters remains the same. Nevertheless, to rigorously check the algorithmic equivalence of the Det-IB and the Ch-Opt-IB, instead of initializing the Det-IB by any random deterministic mapping, we took the resultant mapping at the end of the first iteration of the Ch-Opt-IB, as the starting point of the Det-IB. As a result, identical curves to the Ch-Opt-IB were achieved[6]. This signifies that compared to the Ch-Opt-IB, performance sensitivity of the Det-IB to the size of the event space of its initialization is much less, i.e., significantly higher numbers of trials (in a sense of random initialization of the algorithm) are required for the Det-IB to sweep the performance curve of the Ch-Opt-IB for $n > 4$. The respective compression rate $I(\mathsf{y};\mathsf{z})$ of the investigated approaches over varying the allowed number of clusters n is shown in Fig. 3 b). Considering both subfigures, the main message inferred is the higher the accuracy, the higher would also be the corresponding compression rate.

Next, to investigate the convergence behavior of the algorithms, their average runtime per execution over varying the

[5]Since all heuristics converge to local optima, by increasing the number of runs more local optima are searched and consequently better result is expected.

[6]To avoid overloading the legend in Fig. 3, the explicit mention is dropped.

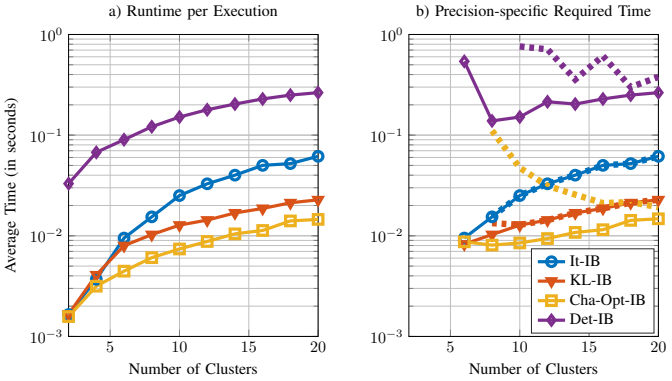

Fig. 4. a) Average runtime per execution and b) precision-specific average required time over varying the allowed number of bins n, 0.9 precision (—) and 0.95 precision (····)

allowed number of clusters is provided in Fig. 4 a). To achieve the required runtime per execution for each algorithm, the corresponding arithmetic mean is calculated for $U = 10^3$ runs. It can be noticed that by increasing the number of clusters, the required time to converge to a local solution also increases for all approaches. One may note, that in Fig. 4 a) the precision of the resultant solution is not conditioned. In the next step, we investigate the average required time for each approach to produce precision-specific results. To do so, assuming a sufficiently large number of runs, one shall also take the statistics of the performance of different approaches into account. The corresponding results demonstrating the average required time for each algorithmic approach when keeping 90 and 95 percent of the *available mutual information*, $I(\mathsf{x};\mathsf{y}) = 1.2187$ bits, is provided in Fig. 4 b).

To obtain these curves, for each specific allowed number of clusters all algorithms were run $U = 10^5$ times and their resultant mutual information $I(\mathsf{x};\mathsf{z})$ was stored. Then, counting the number of trials for which the $I(\mathsf{x};\mathsf{z})$ is at least $0.9I(\mathsf{x};\mathsf{y})$ and $0.95I(\mathsf{x};\mathsf{y})$ respectively, one can calculate the expected number of trials needed for each algorithm to produce a favorable result. Finally, the average required time to have a precision-specific performance is derived by multiplying the expected number of trials with corresponding average required time per trial. Please note that for the chosen number of runs, none of the algorithms converge to a solution with at least 95 percent precision for 6 clusters or less.

As suggested in Fig. 4 b), going from 90 to 95 percent precision, the average required time is almost the same for It-IB and KL-means-IB, while it increases for the other two approaches. The observed time increase stems from the higher expected number of required trials to reach higher precision results. Furthermore, it is seen, that in general, for the case of requiring results of not less than 90 percent precision, the Ch-Opt-IB can be exploited as the fastest approach. For higher values of the required precision, the KL-means-IB would be the favorable choice. Last, comparing the It-IB and the KL-means-IB, it is noticed that the similar behavior as Fig. 4 a) is still present after requiring both algorithms to provide a precision-specific result.

V. Summary

In this work, we presented the general IB framework to quantize an observation variable and provided the mathematical insight into the optimization task of relevance. Following that, we presented different approaches as possible solution candidates. Explicitly, after discussing the It-IB as the primary offered solution and its variants, i.e., Det-IB and Ch-Opt-IB, we also considered the KL-means-IB. For the extreme case of $\beta \to \infty$, we proved, that these approaches are algorithmically equivalent. To corroborate our argumentation, we also provided the corresponding simulation results. Finally, to get an impression about the complexity of different approaches, exploiting the statistics of its performance over a sufficiently large number of runs, we further drew pertinent curves of the average required time for each algorithm to reach a precision-specific outcome.

References

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

[2] 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.

[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, p. 368–377.

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

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

[6] G. Zeitler, A. C. Singer, and G. Kramer, "Low-Precision A/D Conversion for Maximum Information Rate in Channels with Memory," *IEEE Trans. on Communications*, vol. 60, no. 9, pp. 2511–2521, Sep. 2012.

[7] T. Gedeon, A. E. Parker, and A. G. Dimitrov, "The Mathematical Structure of Information Bottleneck Methods," *Entropy*, vol. 14, no. 3, pp. 456–479, Mar. 2012.

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

[9] R. Horst, P. M. Pardalos, and N. Van Thoai, *Introduction to Global Optimization*, 2nd ed. Springer Science & Business Media, 2000.

[10] 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.

[11] D. Strouse and D. Schwab, "The Deterministic Information Bottleneck," in *Conference on Uncertainty in Artificial Intelligence*, New York City, NY, USA, Jun. 2016.

[12] 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.

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

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

[15] 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.

[16] 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.

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

[18] B. M. Kurkoski, "On the Relation Between the KL Means Algorithm and the Information Bottleneck Method," in *11th Int. Conference on Systems, Communications and Coding (SCC)*, Hamburg, Germany, Feb. 2017.