A Graph-Based Message Passing Approach for Noisy Source Coding via Information Bottleneck Principle

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*—The main focus of this paper is on the problem of noisy source coding wherein observed signals from an inaccessible source shall be compressed. To that end, rather than resorting to the conventional methods from *Rate-Distortion* theory, the so-called *Information Bottleneck* paradigm is deployed in order to obtain a highly informative representing signal w.r.t. the given source. An efficient, generic and highly flexible graph-based message passing routine for clustering, known as the *Affinity Propagation* is successfully applied here as a novel treatment for that purpose. The fundamental differences and the performance-wise comparison w.r.t. the state-of-the-art *KL-Means-IB* algorithm is provided as well.**

I. Introduction

Under the presumption of direct access to the source, the lossy data compression/source coding problem is dealt with by the celebrated *Rate-Distortion* (RD) theory [1]. Within the RD framework, a *distortion measure function* has to be defined a-priori, i.e., before the quantizer block design, to characterize the precision of the outcome. Principally, it quantifies the amount of distortion between the original signal and its representative after compression. Alas, the RD theory fails in answering the fundamental question of how to *systematically* obtain the *proper* distortion function in any case of pertinence. Hence, for a wide variety of practical cases, regardless of the structure of the signals and solely for the sake of simplicity, the *squared Euclidean distance* between the quantizer's input/output values is chosen.

For cases in which merely a noisy version of the source is available for the compression, one may either resort to the well-established conventional methods, by treating the observed signal as a *virtual source* [2] or, instead, think about applying a novel framework that directly incorporates the actual source of interest into the design formulation. Deciding in favor of the latter with the intention of bypassing the present faults in the conventional theory, the *Information Bottleneck* (IB) paradigm [3] can be deployed successfully.

The emergence of the IB framework is traced back to the context of *machine learning* applications. Specifically, it was proposed as a novel approach for the crucial task of dimensionality reduction through clustering [4]. However, this methodology can be employed in a broad range of applications concerning data transmission systems as well, among others, construction of polar codes [6], designing analog-to-digital converters (ADCs) [5] and implementation of modern discrete decoding schemes [7], [8] with reduced complexity and yet quite satisfactory performance. Within this work, we focus on the *noisy source coding* scenario in which through the IB formulation, the complexity of the outcome is determined by the so-called *compression rate* which is the mutual information (MI) between the input and the output of the quantizer block. Contrary to the RD theory, the precision of the outcome is quantified via the so-called *relevant information*, i.e., the MI between the actual source and the quantized representative signal. Consequently, a symmetric (in the sense of employing two MI terms to mathematically found the underlying precision-complexity trade-off) design setup [3] results which obviates the demand for the a-priori distortion measure specification. Furthermore, unlike (most of) the other approaches, IB-based quantization wherein pure entropy calculations are involved, is absolutely statistical and totally independent of the specific realizations of the variable(s).

The pertinent optimization task is the focal challenge in the IB-based quantization setup. As it will be discussed, obtaining the *globally* optimal solution via practically feasible algorithms is far from trivial and up to now it is solely achieved for the special case of *binary* input alphabets [9]. Consequently, following a pragmatic approach, one may resort to efficient heuristics which aim at solving the design problem at least *locally*.

Here, we tackle this problem from a new perspective taking advantage of a generic and highly flexible message passing based clustering procedure known as *Affinity Propagation* (AP) [10]. AP is an efficient tool designed for exemplar-based clustering, which takes as input a matrix of pairwise similarities among the set of *articles* to be clustered and provides a high-quality grouping result.

To be able to make use of this powerful tool, the application-specific grouping task shall be translated to an equivalent exemplar-based clustering problem with clear determination of the proper measure of pairwise similarities. In this paper, we address this translation in case of the IB-based quantization and show then what exactly are the corresponding *articles* to be clustered and what is the *proper measure* of pairwise similarities among them. To that end, we introduce the presumed system model and the general IB framework

978-1-5386-4727-1/18/$31.00 ©2018 IEEE

for noisy source coding in the first part of Section II along with a short discussion about the state-of-the-art *KL-Means-IB* routine on its second part. In Section III first we provide the basic understanding of the underlying mechanism for AP that is the well-known *Max-Product* algorithm tailored for a specific *factor graph* and then we show how to achieve an equivalent exemplar-based clustering problem by solving which we obtain the required quantizer for IB-based source coding setup. Finally, we provide some performance assessment results at Section IV before summarizing the most salient points in Section V.

II. Information Bottleneck Source Coding Setup

A. System Model and General IB Framework

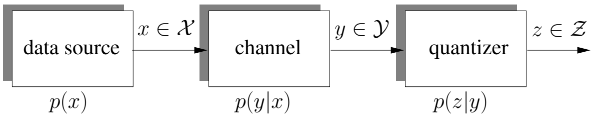

Fig. 1: The assumed system model for the noisy source coding/quantization

We consider the system model illustrated in Fig. 1 for the noisy source coding scenario. The underlying assumption is that there is no immediate 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 aim is then to quantize the observed signal y ($y \in \mathcal{Y} = \{y_1, \cdots, y_{|\mathcal{Y}|}\}$) at the output of the discrete memoryless channel specified by the transition probabilities $p(y|x)$, to the random variable z ($z \in \mathcal{Z} = \{z_1, \cdots, z_{|\mathcal{Z}|}\}$). It may be noted that, in general, $\mathcal{Z}$ should not necessarily be a subset of $\mathcal{Y}$. Moreover, we presume that the joint probability distribution $p(x, y) = p(x)\, p(y|x)$ is given and $\mathsf{x} \leftrightarrow \mathsf{y} \leftrightarrow \mathsf{z}$ constitutes a first-order Markov chain, i.e., $p(z|x, y) = p(z|y)$. Within the IB framework, the *compression rate*, given as the quantizer's input/output MI[2], $I(\mathsf{y}; \mathsf{z})$, quantifies the complexity of the outcome and the *relevant information*, $I(\mathsf{x}; \mathsf{z})$, quantifies its resultant precision. A non-negative Lagrange multiplier $0 \leqslant \beta < \infty$ is utilized to establish the existent trade-off and, thus, the design setup for the quantizer $p(z|y)$ is formulated as [3]

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

in which M denotes the allowed number of output bins and the factor $\frac{1}{\beta+1}$ is solely considered for the sake of mathematical clarity when investigating the extreme cases of β. It shall be noted that the trade-off parameter β can be twiddled in order to weaken (or strengthen) the information preservation capability of the quantizer block. Additionally, it is noteworthy that the resultant quantizer $p(z|y)$ has a *stochastic* or *soft* 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 $y \in \mathcal{Y}$.

Evidently, the case of $\beta \to 0$ is not of interest since then the relevant information term $I(\mathsf{x}; \mathsf{z})$ in (1) is dropped and the minimum compression rate $I(\mathsf{y}; \mathsf{z}) = 0$ can be obtained by making the quantizer's output z being statistically independent of y. In case of finite values of β, it can be shown that the objective function in (1) is neither concave nor convex w.r.t. the quantizer $p(z|y)$ [11]. Hence, the optimization itself is of neither type and consequently finding the *globally optimal* solution becomes quite challenging. Regarding the extreme case of β being asymptotically large, taking the limit of (1) by letting $\beta \to \infty$, the design formulation boils down to

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

wherein, by omitting the minus sign, the minimization in (1) is substituted with the maximization term. It is provable that the present optimization task in (2) is of *convex maximization*[3] type (being *NP-hard* in general [12]) and thus the optimal solution is achieved through *deterministic* mappings [9]. One may note that employing the naive *brute-force* search over all deterministic quantizers results in an exponential complexity w.r.t. $|\mathcal{Y}|$ which clearly makes it intractable in practice.

All things considered, it can be deduced that for non-zero values of β, the pertinent optimization task is far from trivial and thus heuristics shall be proposed to treat the corresponding design problem efficiently. Henceforth, we focus on the salient case of β being asymptotically large (2) wherein the aim is to maximize the end-to-end transmission rate via retaining as much relevant information as possible under the side-constraint on the cardinality of the output representative signal.

In the following part, we present the so-called *KL-Means-IB* algorithm as an instance of the state-of-the-art routines that so far have been proposed in the literature to address the design problem in (2). The provided discussion there will pave the way to perceive the IB-based quantization as an exemplar-based clustering task which opens up the chance of exploiting AP as a novel and quite efficient treatment (see Section III).

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

As discussed before, for asymptotically large values of β, (2) shall be considered as the corresponding design formulation. Based on the definition of MI as difference of entropies[4], in [13] authors have shown that the following holds

$$I(\mathsf{x};\mathsf{z}) = I(\mathsf{x};\mathsf{y}) - \big(H(\mathsf{x}|\mathsf{z}) - H(\mathsf{x}|\mathsf{y})\big) \quad (3a)$$
$$= I(\mathsf{x};\mathsf{y}) - \mathbb{E}_{\mathsf{y},\mathsf{z}}\left\{D_{\mathsf{KL}}\big(p(x|\mathsf{y}) \| p(x|\mathsf{z})\big)\right\} \ . \quad (3b)$$

Therefore, the maximization of the relevant information $I(\mathsf{x}; \mathsf{z})$ corresponds to the minimization of the expectation term in (3b), since the *available information* $I(\mathsf{x}; \mathsf{y})$ is already fixed (it is a function of the joint distribution $p(x, y)$, assumed

[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 $I(\mathsf{a}; \mathsf{b}) \triangleq \sum\limits_a \sum\limits_b p(a, b) \log \frac{p(a,b)}{p(a)p(b)}$.

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

to be given). At this point, the present connection between the design problem at hand and the renowned K-*Means* clustering routine can be perceived. The traditional K-*Means* algorithm [14] is designed to cluster a set of points into K bins such that the average squared Euclidean distance between the points and the pertinent empirical means of the engendered groups is minimized. This task is accomplished in an iterative manner via two distinct steps, namely the *assignment* and the *update* steps. Within the assignment phase, each particular point is allocated to the very cluster with the closest mean among all candidates. Subsequently, in the update phase, the clusters' representatives are recalculated as the respective means (hence the name). This methodology can be generalized to encompass different types of distortion measures. For example, in [16] a certain family of divergences including the *Kullback-Leibler* (KL) divergence[5] has been considered. As an interesting interpretation (regarding its *model-based* derivation [15] with the isotropic spherical Gaussian noise assumption) from the communications point of view, one may treat the points to be clustered as the noisy versions of K *originally* transmitted points (where in communications terminology it is referred to as the signal constellation) and target at learning the most fitting underlying constellation.

Starting to describe it more tangibly, it shall be noted that irrespective of the certain choice for $y \in \mathcal{Y}$, it must hold $\sum_{x\in\mathcal{X}} p(\mathsf{x}=x|\mathsf{y}=y)=1$, which introduces a $(|\mathcal{X}|-1)$-dimensional *probability simplex*, referred to as the *backward channel simplex*. Thus, as suggested in [13], through the transformation of the primary quantization space (i.e., the space wherein the y values are defined) into the backward channel simplex by considering $p(x|\mathsf{y}=y)$[6] and $p(x|\mathsf{z}=z)$ (the corresponding points in the transformed space) instead of y and z, respectively, and treating the KL divergence as the appropriate distortion measure, the design problem in (2) can be perceived as a special K-*Means* clustering task. Explicitly, the *KL-Means-IB* [13] is initiated by a random pick of M distinct points $p(x|y)$ as the primary means. Subsequently, through the assignment phase, each point $p(x|y)$ is grouped into the particular bin z for which the corresponding representative $p(x|z)$ shows the least KL divergence. Mathematically, $p(z|y)=\delta_{z,z^\star(y)}$[7] where the host bin z for each particular y value is chosen as

$$z^\star(y) = \operatorname*{argmin}_{z} D_{\mathrm{KL}}\big(p(x|y)\|p(x|z)\big) \ . \tag{4}$$

Next, the representants are updated as the clusters' centers of mass [16]

$$p(x|z) = \frac{\sum\limits_{y\in\mathcal{Y}_z} p(y)p(x|y)}{\sum\limits_{y\in\mathcal{Y}_z} p(y)}, \tag{5}$$

where $\mathcal{Y}_z$ denotes the subset of $\mathcal{Y}$ for which all members are allocated to the bin z. The mentioned procedure is perpetuated till either a convergence criterion or a maximum number of iterations is met.

[5] Also known as relative entropy, among two probability distributions $p(a)$ and $q(a)$ over the same event space $\mathcal{A}$ of the random variable a, is defined as $D_{\mathrm{KL}}\big(p(a)\|q(a)\big) \triangleq \sum_{a\in\mathcal{A}} p(a)\log\frac{p(a)}{q(a)}$ [1].

[6] It is defined as $p(x|\mathsf{y}=y) \triangleq [p(x_1|\mathsf{y}=y),\cdots,p(x_{|\mathcal{X}|}|\mathsf{y}=y)]$.

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

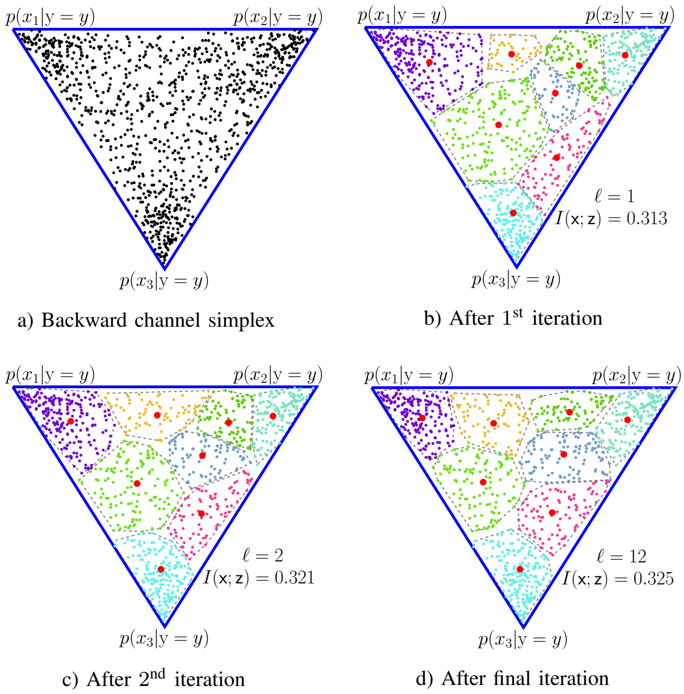

Fig. 2: Evolution of the *KL-Means-IB* outcome ($M=8$) through iterations ℓ, 3-PSK signaling over an AWGNC ($\sigma_n^2=1$), the red dots demonstrate the clusters' representatives $p(x|z)$, the acquired $I(\mathsf{x};\mathsf{z})$ is provided as well

To vividly visualize what exactly happens in the backward channel simplex during the iterations of the *KL-Means-IB*, we illustrate the corresponding 2-dimensional simplex for an example of 3-PSK signaling in Fig. 2. Specifically, each point inside the depicted simplex corresponds to a certain $p(x|\mathsf{y}=y)=[p(x_1|\mathsf{y}=y),p(x_2|\mathsf{y}=y),p(x_3|\mathsf{y}=y)]$ for a particular 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 grouping procedure for $M=8$ clusters is then carried out in this space and as can be observed, through iterations till convergence, the clustering results are getting refined.

The complexity of the *KL-Means-IB* routine is dominated by the assignment phase. Hence, as suggested by (4), per iteration, the *KL-Means-IB* has the complexity of $O(|\mathcal{X}|\cdot|\mathcal{Y}|\cdot|\mathcal{Z}|)$, as for every specific value y, the KL divergence to all $|\mathcal{Z}|$ possible candidates has to be calculated where each of these calculations sums up $|\mathcal{X}|$ terms.

III. Affinity Propagation

A. Description and Discussion

We commence this part by providing a *general explanation* of the *Affinity Propagation* (AP) [10]. AP is a recursive message passing routine for exemplar-based clustering. An *exemplar* is the representative of a certain group (engendered after clustering) that is chosen from the primary set of *articles*, the *mother-set*, to be clustered. Therefore, the *exemplar-set* can be regarded as a reduced (concise) version of the *mother-set* that represents it (w.r.t. a certain criterion) in a best fashion. Describing it roughly, AP considers all primary articles as nodes of a network being delineated by a fully-connected graph in which there is an edge between any two articles. These articles communicate with each other through the edges in a recursive manner (bilateral messages per edge) in order to gradually decide about the most suitable *exemplar-set* and, consequently, the corresponding *non-exemplar* to *exemplar* allocations. The aforementioned bilateral recursive inter-node communication between the articles can be described as follows: based on the direction of the transmitted messages, they can be interpreted either as *responsibilities* or *availabilities*. In AP terminology, the *responsibility* is the pertinent designation for the message being transmitted from an *article* to a *potential exemplar*. Conversely, the *availability* is utilized to signify that the message is transmitted from a *potential exemplar* to an *article*. At the initial stage, all N different articles are treated as potential exemplars. Hence, each specific article i shall generate a message $r(i,k)$ for the article k with $i,k \in \{1,2,...,N\}$. The *responsibility* $r(i,k)$ indicates to which extent the *article* i conceives the *potential exemplar* k to be responsible to serve it as an exemplar. The first round of responsibilities are solely created by the pairwise *similarities* $s(i,j)$ for $i,j \in \{1,2,...,N\}$, given as the input of the AP. A quite interesting feature of AP that secures a great deal of flexibility is the fact that the *mother-set* does not have to lay in a metric or continuous or even ordinal space and, consequently, the pairwise similarities do not have to be calculated based on a metric, i.e., they do not have to be *symmetric* and, moreover, they do not have to satisfy the *triangle inequality*.

In the next phase of the AP dynamism the availability messages are generated from the received responsibilities. The *availability* $a(i,k)$ indicates to which extent the *potential exemplar* k conceives itself to be available as an exemplar for the *article* i. This bilateral message passing procedure is perpetuated till a highly satisfactory set of exemplars and the corresponding clusters emerge.

Another compelling feature of AP is about its astonishingly simple and intuitive update equations. Explicitly, the update rule for the *responsibility* calculations is given as

$$r(i,k) = s(i,k) - \max_{j \neq k} \{a(i,j) + s(i,j)\}, \tag{6}$$

and the corresponding update rules for *availabilities* are

$$a(i,k) = \min \left\{ 0, r(k,k) + \sum_{j \notin \{i,k\}} \max\{0, r(j,k)\} \right\} \text{ for } i \neq k \tag{7}$$

and in case of self-availability

$$a(k,k) = \sum_{j \neq k} \max\{0, r(j,k)\} \ . \tag{8}$$

The stated update rules are derived from the utilization of the *Max-Product* routine [17] to approximate the marginals of the global function pertaining to the specific *factor graph* depicted in Fig. 3. The *variable nodes* c_i and the *factor nodes* f_i with $i \in \{1,...,N\}$ represent the *chosen exemplar* and the *coherency check* for the article i, respectively. By coherency check f_i, it is meant that if any other article $j \neq i$ decides in favor of the article i as its exemplar, then the article i should be the exemplar of itself as well. Considering the singleton factor nodes in Fig. 3, it is rather straightforward to observe that the aim of the computation over such a specific factor graph is to find out the certain *coherent configuration* of the variables which maximizes the sum of the overall similarities between all the articles and their corresponding exemplars. The immediate application of the *Max-Product* routine requires the messages to be vector-valued. Nonetheless, performing some intelligent mathematical tricks as stated in [18], the update rules are shrunk to the scalar-valued versions (6)-(8) which brings about the complexity of $O(N^2)$ per iteration.

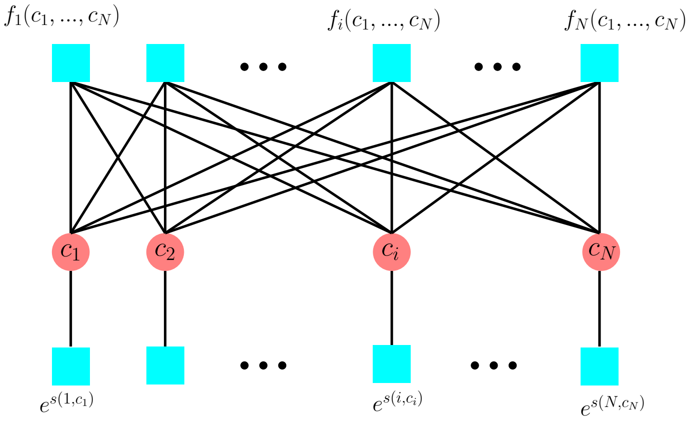

Fig. 3: The factor graph on which AP is developed, the *variable nodes* c_i and the *factor nodes* f_i with represent the *chosen exemplar* and the *coherency check* for the article i, respectively

At the starting point of AP, to treat all the articles equally, i.e., providing the same chance to be chosen as an exemplar, all the availabilities $a(i,j)$ in (6) are set to zero. Nevertheless, after a while if an article gets assigned to another one, its availability becomes negative, which directly influences the pertinent *effective similarity* in (6), bringing it out of the ongoing exemplarship competition. Focusing on the availability update (7) for the article k, it is basically calculated as the sum of its self-responsibility $r(k,k)$ and all the positive feedbacks from the other articles. The negative feedbacks (responsibilities) are ignored while those are related to the articles for which the article k is not an appropriate exemplar and for a decent exemplar, it suffices to represent some and not all of the articles well.

Another intriguing aspect of AP is about its so-called *automatic model selection* capability [18]. One may note that the cardinality of the *exemplar-set* is not given a-priori but it rather comes out naturally (at the end of the recursive process) for each specific choice of the common *self-similarity*, which in AP terminology is referred to as the common *preference*. In general, the preference values do not have to be the same for all the articles and the larger the value of the preference $s(i,i)$, the higher the chance of the article i to be chosen as an exemplar at the end. Consequently, in case of having a common preference, it is principally the very parameter which can be twiddled in order to have an influence on the cardinality of the resultant *exemplar-set*. The rationale behind is the fact that the sum of the preferences can be regarded as the penalty term (w.r.t. the complexity) being present at the objective function of the AP factor graph to avoid, e.g., the case in which each point is treated as an exemplar (for and only for itself). As a general observation, performing the *belief propagation* over loopy graphs may lead to an instable behavior. Hence, as an important implementation detail, the messages which shall be transmitted over the fully-connected AP graph of Fig. 3 have to be dampened. Thus, irrespective of being either availability or responsibility, the messages are calculated as the weighted combination of their previous and current values. Finally, it has to be mentioned that the cluster allocation for the article i at the end of each round of message exchange can be estimated as

$$\hat{c}_i = \underset{k}{\operatorname{argmax}} \left\{ a(i,k) + r(i,k) \right\} . \tag{9}$$

B. IB-Based Quantization Utilizing AP

In this part, we establish the connection between the IB-based noisy source coding problem (2) and the generic exemplar-based clustering task behind AP. Consequently, as an upshot, we propose the AP usage as a novel and efficient treatment of the quantization task at hand.

Commencing with (3b), it is directly deducible that to maximize the end-to-end transmission rate, $I(\mathsf{x};\mathsf{z})$, one shall maximize the average term

$$\begin{aligned} &\mathbb{E}_{\mathsf{y},\mathsf{z}} \left\{ -D_{\mathrm{KL}}\big(p(x|\mathsf{y}) \| p(x|\mathsf{z})\big) \right\} \\ &= \sum_{y\in\mathcal{Y}} p(y) \sum_{z\in\mathcal{Z}} -p(z|y) \cdot D_{\mathrm{KL}}\big(p(x|y) \| p(x|z)\big) . \end{aligned} \tag{10}$$

As already discussed, the deterministic mapping is an optimal choice of the quantizer for the present *convex maximization* problem. Therefore, assuming *hard* quantization and denoting the chosen z for each specific y as $z^\star(y)$, i.e., $p(z|y) = 1$ for $z = z^\star(y)$ and zero otherwise, (10) can be rewritten as

$$\sum_{y\in\mathcal{Y}} -p(y) \cdot D_{\mathrm{KL}}\big(p(x|y) \| p(x|z^\star(y))\big) . \tag{11}$$

The structure of (11) is reminiscent of the equivalent global AP objective function $\sum_i s(i, c_i)$, which is the overall sum of the similarities between the individual articles i and their corresponding exemplars c_i. Principally, by considering the realizations $y^{(i)}$ and $y^{(j)}$ of the observed (channel output) random variable y in Fig. 1 as the *articles* i and j, respectively, and by defining the *pairwise similarities* as

$$s(i,j) = -p(i) \cdot D_{\mathrm{KL}}\big(p(x|i) \| p(x|j)\big), \tag{12}$$

the global AP objective function coincides with (11). As the aim of AP is to find the coherent configuration that maximizes its objective function, it is exactly in line with the IB-based quantization design criterion which boils down to maximizing the average term in (10). Noting the *bijective correspondence* between each y and the pertinent point $p(x|y)$ in the *backward channel simplex*, one can better comprehend the AP clustering methodology. Focusing on the provided example in the previous section, unlike the *KL-Means-IB*, AP treats all the present points in the simplex of Fig. 2 a) as potential exemplars at first. By exchanging responsibilities and availabilities between these points in a recursive manner, the final *exemplar-set* (set of all red dots) gradually appears. It is also noteworthy that by utilizing AP, contrary to the *KL-Means-IB* approach, the cluster's representatives, i.e., the red dots are chosen from the primary set of points in the depicted backward channel simplex rather than being calculated as the corresponding center of mass per cluster.

IV. Simulation Results

In this part, we set about investigating the performance behavior of the proposed AP-based quantization approach. To that end, we consider a typical digital transmission scenario in which 1000 equiprobable symbols from a 16-QAM constellation ($\sigma_x^2 = 10$) are transmitted over an AWGN channel with three different noise variances ($\sigma_n^2 = 1, 2, 3$). The received points are then clustered into a varying number of groups (2 to 40). As the performance indicator, we calculate the resultant overall transmission rate, $I(\mathsf{x};\mathsf{z})$. Furthermore, we perform the similar investigation for the KL-Means-IB with $U = 100$ runs (retaining the best outcome) and depict the obtained curve as well to be able to compare both approaches.

Fig. 4 illustrates the pertinent results. It shall be mentioned that, as the cardinality of the output clusters depends on the specific choice of the common preference, to generate the respective curve of the AP-based approach, we exploited the well-known *bisection method* [18]. It is clearly observable that irrespective of the specific choices of the model parameters, i.e., the noise variance σ_n^2 and the number of output clusters M, both routines engender almost identical results. However, there are fundamental differences between the two approaches which must be taken into account. First of all, it should be noted that the outcome of *one* run of the KL-Means-IB heavily depends on its choice of initialization and, therefore, to ameliorate the final result, it must be repeated a number of times, e.g., $U = 100$. Secondly, unlike the KL-Means-IB, the AP-based method does not follow any random initialization and, consequently, does not need to be repeated to deliver a better outcome. Nonetheless, to produce any specific number M of output clusters, the proposed algorithm must be repeated a number of times to determine the relevant value of the common preference (utilizing bisection method) yielding that certain output cardinality.

To obviate the utilization of the bisection method and therefore substantially reducing the required computational effort of the proposed AP-based treatment, we calculated the *approximate common preference* (ACP) pertaining to any number of output cardinality. Fig. 5 depicts the corresponding values. To obtain these curves, for each specific noise variance, we repeated the aforementioned transmission scenario for 100 times and saved the corresponding CPs. The resultant ACP for each value of M is then calculated as the arithmetic mean of

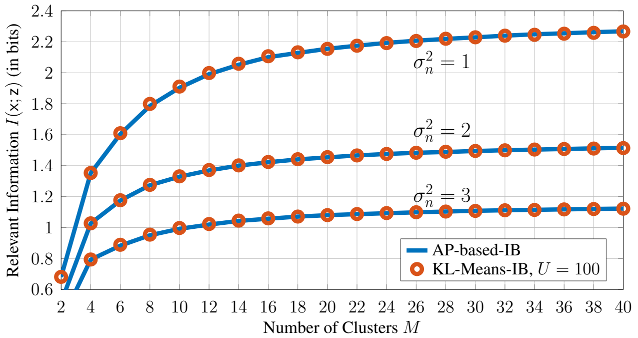

Fig. 4: Relevant information $I(\mathsf{x};\mathsf{z})$ vs. number of clusters M, 16-QAM signaling ($\sigma_x^2 = 10$), AWGNC with different noise variances

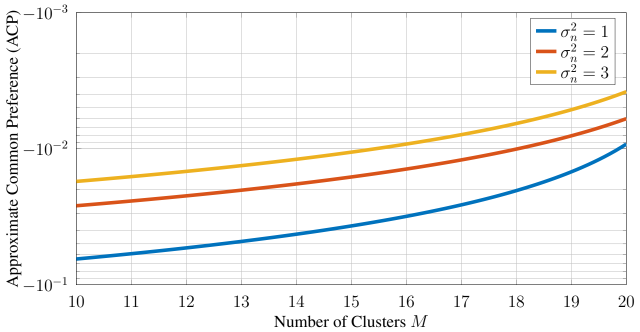

Fig. 5: Approximate common preference vs. number of clusters M, 16-QAM signaling ($\sigma_x^2 = 10$), AWGNC with different noise variances

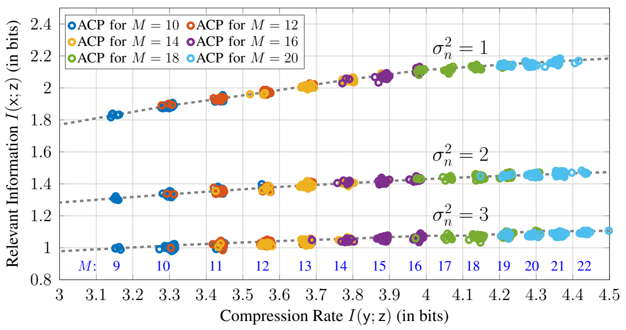

Fig. 6: Relevant information $I(\mathsf{x};\mathsf{z})$ vs. compression rate $I(\mathsf{y};\mathsf{z})$, 16-QAM ($\sigma_x^2 = 10$), AWGNC with different noise variances, bisection result (····)

the lowest and the highest values among 100 trials. To check the usability of such a preference guideline (which can be done once offline), we repeated the previous transmission setup 50 times more and for each of the new trials, we inserted the calculated values for six choices of M (taken from Fig. 5) as the chosen CP and executed the AP-based routine only once. The corresponding results are illustrated in Fig. 6. The immediate observation verifies the fact that the provided preference guideline works satisfactorily well. As a result, in cases for which the strict upholding of a certain output cardinality is not a must and a relatively small variation range can be tolerated, only *one run* of the proposed algorithm provides a quite promising result. This, generally, brings about a noticeable gain in computational effort w.r.t. the KL-Means-IB usage. To perceive this, one shall consider the overall complexity of both approaches to yield comparable results. While in case of KL-Means-IB, it amounts to $O(|\mathcal{X}| \cdot |\mathcal{Y}| \cdot |\mathcal{Z}| \cdot \ell_1 \cdot U)$ with ℓ_1 denoting the average number of iterations per run, for the proposed method, it will be $O(|\mathcal{Y}|^2 \cdot \ell_2)$ with ℓ_2 being the counterpart of ℓ_1, and, usually, it applies that $|\mathcal{X}| \cdot |\mathcal{Z}| \cdot \ell_1 \cdot U \gg |\mathcal{Y}| \cdot \ell_2$.

V. Summary

We considered noisy source coding and rather than methods of *Rate-Distortion* theory, we deployed *Information Bottleneck* framework and provided the respective mathematical insights. As the major contribution, we then proposed a novel treatment utilizing *Affinity Propagation* which is an efficient graph-based message passing approach for clustering. Finally, we investigated the performance of our proposed treatment w.r.t. the state-of-the-art KL-Means-IB routine and showed that the novel ACP-approach (with only *one* run) can be taken as a quite efficient and competitive alternative.

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 Trans. on Information Theory*, vol. 36, no. 6, pp. 1450–1452, Nov. 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, pp. 368–377, Sep. 1999.
[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, Sep. 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] B. J. Frey and D. Dueck, “Clustering by Passing Messages between Data Points,” *Science*, vol. 315, no. 5814, pp. 972–976, Feb. 2007.
[11] 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.
[12] R. Horst and P. M. Pardalos, *Handbook of Global Optimization*, Springer Science & Business Media, vol. 2, 2013.
[13] 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.
[14] A. K. Jain, “Data Clustering: 50 Years Beyond K-means,” *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, June 2010.
[15] A. D. Gordon, *Classification: Methods for the Exploratory Analysis of Multivariate Data*, Chapman and Hall, 1981.
[16] 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.
[17] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, “Factor Graphs and the Sum-Product Algorithm,” *IEEE Transactions on information theory*, vol. 47, no. 2, pp. 498–519, Feb. 2001.
[18] D. Dueck, “Affinity Propagation: Clustering Data by Passing Messages,” Ph.D. dissertation, University of Toronto, Canada, 2009.