The Information Bottleneck Method: Fundamental Idea and Algorithmic Implementations

Dirk Wübben
Department of Communications Engineering
University of Bremen, 28359 Bremen, Germany
Email: wuebben@ant.uni-bremen.de

***Abstract*—The quantized representation of signals is a general task of data processing. For lossy data compression the celebrated Rate-Distortion theory provides the compression rate in order to quantize a signal without exceeding a given distortion measure. Recently, with the Information Bottleneck method an alternative approach has been emerged in the field of machine learning and has been successfully applied for data processing. The fundamental idea is to include the original source into the problem setup when quantizing an observation variable and to use strictly information theoretic measures to design the quantizer. This paper introduces this framework and discusses algorithmic implementations for the quantizer design.**

I. Extended Abstract

A fundamental task in data processing is the quantized representation of noisy observations of an original source signal. Fig. 1 shows the considered system model consisting of a data source, a transmission channel and a quantizer.

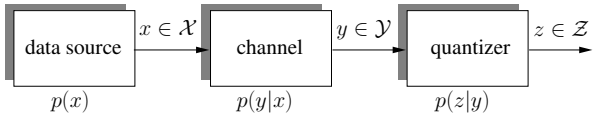

Fig. 1. General system model for the quantization of noisy observations

Without loss of generality, we assume the random variable x with realizations $x \in \mathcal{X}$ following the probability mass function (pmf) $p(x)$ as a discrete memoryless source (DMS). The observation variable y with realizations $y \in \mathcal{Y}$ is the output of a discrete memoryless channel (DMC) characterized by its transition probability distribution $p(y|x)$. Furthermore, the random variable z with realizations $z \in \mathcal{Z}$ is the output of the quantizer block being characterized by the conditional distribution $p(z|y)$. Thus, the quantizer output z is a compact representation of y with cardinality $|\mathcal{Z}| \leq |\mathcal{Y}|$. The design of the quantizer $p(z|y)$ realizes a trade-off between the *compression rate*, $I(\mathsf{y};\mathsf{z})$, and the *quality* of the compressed representation.

The Rate-Distortion (RD) theory provides the minimal number of bits per symbol in order to represent the received signal without exceeding an upper-bound on a given distortion measure, e.g., the mean square error (MSE) between the quantizer input signal and its representative at the output [1]. Specifically, the Blahut-Arimoto algorithm determines the lowest achievable compression rate for a certain maximum tolerable distortion. The main drawbacks of this formulation are the lack of a systematic way to choose a proper distortion measure for any case of pertinence and the fact, that the stochastic relation between the noisy observation and the original data source is not considered.

In [2], Tishby et al. have introduced the Information Bottleneck (IB) method for data compression. The central idea is to compress the observation y such that the quantizer output z preserves most of the information about the relevant variable, i.e., the original source x. Furthermore, IB avoids the a priori specification of a distortion measure by considering the mutual information $I(\mathsf{x};\mathsf{z})$ between the quantizer output and the original data source. In this fashion, the output of the quantizer becomes a compact representation of its input which is highly informative about the actual source of interest.

Given the joint probability distribution of the source and the channel output $p(x,y) = p(x)\,p(y|x)$ and assuming $\mathsf{x} \leftrightarrow \mathsf{y} \leftrightarrow \mathsf{z}$ to be a Markov chain, the quantizer should be designed such that the output z is a compact representation of the input y which is highly informative about x. Mathematically, the existent trade-off between the *compression rate*, $I(\mathsf{y};\mathsf{z})$, and the *relevant information*, $I(\mathsf{x};\mathsf{z})$, is established by the introduction of a non-negative Lagrange multiplier, β, in the design formulation. Hence, for an allowed number of quantizer output levels, n, the corresponding design problem follows as [2]

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

The optimal quantizer mapping for (1) can be derived 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

$$p(z|y) = \frac{p(z)}{\psi(y,\beta)} e^{-\beta D_{\mathrm{KL}}(p(x|y)\|p(x|z))} \quad (2)$$

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. The derived optimal mapping in (2) has an implicit form, as the cluster representative (in a conventional sense) $p(x|z)$ and the cluster probability $p(z)$ appearing on the right side of (2), depend on the quantizer mapping $p(z|y)$ by

$$p(z) = \sum_{y \in \mathcal{Y}} p(y)\,p(z|y) \quad (3)$$

and

$$p(x|z) = \frac{1}{p(z)} \sum_{y \in \mathcal{Y}} p(x,y)\, p(z|y) \ . \tag{4}$$

The iterative calculation of (2)-(4) leads to the Iterative Information Bottleneck (It-IB) algorithm [2]. Several alternative approaches to determine mapping functions in order to meet the trade-off between *compression rate* and *relevant information* have been discussed in the literature such as

- Agglomerative Information Bottleneck (Agg-IB) [3]
- Sequential Information Bottleneck (Seq-IB) [4]
- Deterministic Information Bottleneck (Det-IB) [5]
- KL-means Information Bottleneck (KL-means-IB) [6]
- Channel-Optimized Information Bottleneck (Ch-Opt-IB) [7].

For the special case of binary input alphabet computationally efficient adaptations exist as discussed in [8].

Subsequently, we compare the performance of the algorithms for 4-ASK ($x \in \{\pm 1, \pm 3\}$) input signals transmitted over AWGN channels with noise variance $\sigma_n^2 = 1$. Furthermore, to acquire the channel transition distribution $p(y|x)$, the continuous channel output is clipped at an amplitude of $3\,\sigma_n$ above the maximum input signal (i.e., 6 for 4-ASK) and 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 allowed number of clusters n.

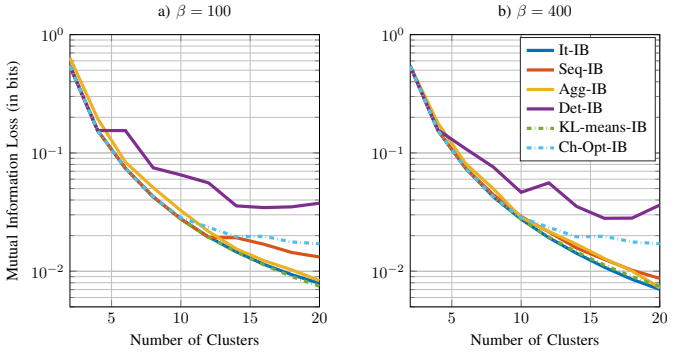

Fig. 2. Information loss ΔI for varying allowed number of bins n and 4-ASK input alphabet with a) $\beta = 100$ and b) $\beta = 400$

Fig. 2 shows the information loss ΔI of the investigated algorithms. One may note, as the resultant mapping of all algorithms (except for the Agg-IB) depends on the initialization, to achieve the corresponding curves, they have been run 10^5 times, with the best taken. Except for the KL-means-IB and the Ch-Opt-IB (both only consider $\beta \to \infty$) one can observe, that the accuracy of all algorithms is improved by increasing β from 100 to 400. For a fair comparison with the KL-means-IB and the Ch-Opt-IB we concentrate subsequently on Fig. 2 b) with a relatively high value of β.

First of all, the non-smooth behavior of the Det-IB is due to the fact that its provided mapping does not necessarily use the entire allowed number of clusters, i.e., $|Z| < n$. As an example, for $n = 12$ the used number of bins is smaller than the case of $n = 10$, leading to a coarser result. Furthermore, it can be seen that the It-IB and the KL-means-IB exhibit nearly the same performance over the entire range of allowed number of bins n. In addition, one notes that the Ch-Opt-IB also sweeps the corresponding curve of the It-IB for $n \leq 10$. The reason behind these observations is discussed in [9] where the asymptotic algorithmic equivalence of these algorithms is proven.

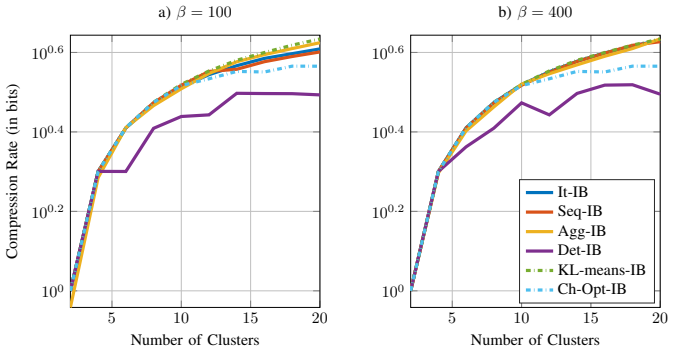

Fig. 3. Compression rate $I(\mathsf{y};\mathsf{z})$ for varying allowed number of bins n and 4-ASK input alphabet with a) $\beta = 100$ and b) $\beta = 400$

Fig. 3 displays the corresponding compression rates $I(\mathsf{y};\mathsf{z})$. It can be observed, that in general, the lower the information loss introduced by quantization, the higher the corresponding compression rate.

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] 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.

[3] N. Slonim and N. Tishby, "Agglomerative information bottleneck," in *Advances in Neural Information Processing Systems*, 1999, pp. 617–623.

[4] N. Slonim, N. Friedman, and N. Tishby, "Unsupervised document classification using sequential information maximization," in *25th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, Tampere, Finland, Aug. 2002, pp. 129–136.

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

[6] 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.

[7] 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.

[8] 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.

[9] S. Hassanpor, 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.