A DATA-DRIVEN QUANTIZATION DESIGN FOR DISTRIBUTED TESTING AGAINST INDEPENDENCE WITH COMMUNICATION CONSTRAINTS

Sebastian Espinosa, Jorge F. Silva

Department of Electrical Engineering
University of Chile
Santiago de Chile, Chile

Pablo Piantanida

Laboratoire des signaux et systèmes (L2S)
CentraleSupélec CNRS Université Paris-Saclay
Gif-sur-Yvette, France

ABSTRACT

This paper studies the problem of designing a quantizer (encoder) for the task of distributed detection of independence subject to one-side communication (limited bits) constraints. By exploiting the asymptotic performance limits as an objective to train a quantization scheme, we propose an algorithm that addresses an *info-max problem* for this lossy compression task. Tools from machine learning are incorporated to facilitate our data-driven optimization. Experiments on synthetic data support our design principle and approximations, expressing that the devised solutions are effective in compressing data while preserving the relevant information for the underlying task of testing against independence.

***Index Terms*—** Distributed decision, quantization design, lossy data compression, information bottleneck.

1. INTRODUCTION

The study of distributed decision and inference, where measurements are collected remotely with communication constraints, is an important problem in signal processing over networks. In particular, the problem of detecting independence between measurements taken remotely is fundamental from a theoretical standpoint and relevant in applications. For example, in cognitive radio networks arises the problem of spectrum sensing where there is a primary agent and a group of secondary agents; the presence of the primary user's signal introduces dependence on the decentralized sensors and then, the secondary agents collaborate to detect whether the primary agent is present or not [1, 2].

In this work, we focus on the distributed setting first introduced by Ahlswede and Csiszar in [3]. This problem consists of testing against independence where the observations (the evidence) are available at two remote nodes, as shown in Figure 1. In particular, one of the observations needs to be transmitted to the remote agent (the detector) subject to a rate-constraint in terms of bits per-sample. In this framework, the information-theoretic analysis of performance have been addressed in [3] where the asymptotic limit for the Type 2 error subject to a fixed Type 1 constraint were derived. More recently, finite-length performance bounds (non-asymptotic) have been derived [4, 5]. These theoretical results are important because they offer bounds to the performance limits for the Type 1 and Type 2 errors. Unfortunately, these results do not provide implementable coding schemes. The design of practical algorithms (quantizers) achieving good performance in this distributed setting has been an elusive topic.

The main contribution of this paper is a design criterion for the two agents: the encoder $f_n(\cdot)$ and the detector $\phi_n(\cdot)$ with the objective of minimizing the probability of misdetection subject to a fixed-rate constraint on $f_n(\cdot)$. Our criterion exploits the asymptotic information limits in [3, Th.2] to devise a practical info-max design principle. In particular, a data-driven optimization algorithm is proposed to learn the quantizer (encoder) for this problem. By borrowing ideas from unsupervised learning [6–8], a collection of soft quantizer, i.e. based on conditional probabilities of bins given the features, is used to formalize the info-max problem. In particular, we consider the rich collection of *Boltzmann distributions* to represent the space of soft-quantizers. Interestingly, our methodology connects with the so-called Variational *Information Bottleneck* (IB) problem in [9–11] as we are optimizing the empirical mutual information between a class variable and a representation variable (the output of the quantizer) subject to a compression constraint [12]. In terms of experimental evaluation, the proposed solution in this paper offers performance advantages (Neyman–Pearson trade-off) when compared with a strategy that is based on an unsupervised design principle. We observe that this gain is proportional to the mutual (relevant) information of the underlying model. Finally, we also evaluate how the number of samples (block-length) and the communication constraint (number of bits) affect the performance of the proposed test.

2. PRELIMINARIES

Let us consider a finite product space $\mathbb{Z} = \mathbb{X} \times \mathbb{Y}$, where $\mathcal{P}(\mathbb{Z})$ denotes the space of probability measures on $\mathbb{Z}$. We have a pair (X, Y) of random variables with values in $\mathbb{Z}$ equipped with a probability $P_{X,Y} \in \mathcal{P}(\mathbb{Z})$ (the model), where $P_X \in$

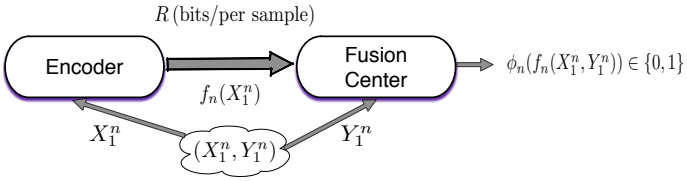

Fig. 1: The one-side distributed test: $f_n(\cdot)$ is the encoder and $\phi_n(\cdot)$ is the detector acting on $(f_n(X_1^n), Y_1^n)$.

$\mathcal{P}(\mathbb{X})$ and $P_Y \in \mathcal{P}(\mathbb{Y})$ denote the marginals distributions of X and Y, respectively. $X_1^n = (X_1, ..., X_n)$ and $Y_1^n = (Y_1, ..., Y_n)$ denote the (i.i.d.) finite block vectors with product distribution $P_{X_1^n, Y_1^n} \equiv P_{X,Y}^n \in \mathcal{P}(\mathbb{X}^n \times \mathbb{Y}^n)$ (the $n-$fold distribution). The two hypotheses for the n-length test against independence are:

$$\begin{array}{ll} H_0: & (X_1^n, Y_1^n) \sim P_{XY}^n, \\ H_1: & (X_1^n, Y_1^n) \sim Q_{XY}^n, \end{array} \tag{1}$$

where $P_{X,Y} \in \mathcal{P}(\mathbb{Z})$ and $Q_{X,Y} \equiv P_X \cdot P_Y$ denote the product probability induced by the marginals of $P_{X,Y}$. For the decision with communication constraint (illustrated in Figure 1), we introduce the pair (f_n, ϕ_n) of encoding and decision rule of length n and rate R (in bits per sample) by:

$$\begin{aligned} & f_n : \mathbb{X}^n \to \{1, \ldots, 2^{nR}\}, \text{ (encoder)} \\ & \phi_n : \{1, \ldots, 2^{nR}\} \times \mathbb{Y}^n \to \Theta = \{0, 1\}, \text{ (detector).} \end{aligned} \tag{2}$$

$f_n(\cdot)$ is a fixed-rate (lossy) encoder of X_1^n operating at rate R (bits per sample) and $\phi_n(\cdot)$ represents the decision rule (or detector) acting on the one-sided compressed observations $(f_n(X_1^n), Y_1^n) \in \{1, ..., 2^{nR}\} \times \mathbb{Y}^n$. For any pair (f_n, ϕ_n) of length n and rate R, we can introduce its Type 1 and Type 2 errors [13], [14]:

$$P_0(f_n, \phi_n) \equiv P_{XY}^n(\mathcal{A}^c(f_n, \phi_n)) \text{ and} \tag{3}$$

$$P_1(f_n, \phi_n) \equiv Q_{XY}^n(\mathcal{A}(f_n, \phi_n)) \text{ with ,} \tag{4}$$

$A(f_n, \phi_n) \equiv \{(x_1^n, y_1^n) \in \mathbb{X}^n \times \mathbb{Y}^n : \phi_n(f_n(x_1^n), y_1^n) = 0\}$. Given $\epsilon > 0$, the performance analysis of this problem is:

$$\beta_n(\epsilon, R) \equiv \min_{(f_n, \phi_n)} \{P_1(f_n, \phi_n) : P_0(f_n, \phi_n) \leq \epsilon\}. \tag{5}$$

The minimum in (5) is over the collection of encoder-decision pairs of the form presented in (2). Then, $\beta_n(\epsilon, R)$ is the optimum Type 2 error that can be achieved for a given Type 1 error restriction. Importantly, $\forall \epsilon > 0$, we have the following asymptotic result [3, Th.3]:

$$\xi(R) = \lim_{n \to \infty} -\frac{1}{n} \log \beta(\epsilon, R) = \max_{\substack{U : U \multimap X \multimap Y \\ I(U;X) \leq R \\ |\mathbb{U}| \leq |\mathbb{X}| + 1}} I(U; Y), \tag{6}$$

where $U \multimap X \multimap Y$ indicates a Markov chain. This result states that $\beta_n(\epsilon, R)$ admits an exponential decreasing pattern as n tends to infinity with an error exponent given by $\xi(R)$. Furthermore, the recent work [5, Th.2] shows upper and lower bounds for $\beta_n(\epsilon, R)$ indicating that for n sufficiently (reasonable) large $\beta_n(\epsilon, R)$ can be accurately approximated by $e^{-n\xi(R)}$. This result supports the adoption of $e^{-n\xi(R)}$ as a useful proxy for Type 2 error probability. As a matter of fact, the design of (f_n, ϕ_n) as being the solution to the operational problem in (5) is not tractable. As an alternative, we consider the results in [3, 5] to guide the encoder from the expression $\xi(R)$, i.e., using the info-max problem in (6).

3. MAIN CONTRIBUTION

We propose a concrete data-driven info-max criterion for the encoder f_n (and implicitly ϕ_n) for testing independence. The fundamental limit of the Type 2 error in (6) (independent of $\epsilon > 0$) is given by the mutual information maximization between a soft (lossy) representation of X (represented by U) and the class level Y. Using this optimization, we propose to maximize the mutual information between the representation $U = f_n(X^n)$ and Y_1^n, i.e., $I(U; Y_1^n)$, given a size constraint on the range of f_n. This yields the info-max problem:

$$\max_{f_n : \mathbb{X}^n \to \mathbb{U} = \{1, \ldots, K\}} I(U; Y_1^n). \tag{7}$$

The problem in (7) can be seen as a multi-letter version of (6), where we use deterministic mappings (quantizers) instead of the soft mappings (conditional probabilities) expressed in (6).

3.1. Approximations and Design Considerations

The problem in (7) is practically challenging when considering a large n. Some approximations are needed to make it tractable from an optimization viewpoint. In this process, we relax some assumptions that make the resulting problem more realistic (data-driven). The approximations adopted to make (7) tractable for large n is the focus of this section.

3.1.1. Empirical Version of (7)

First, we note that the sequence $U = f(X_1^n) \multimap X_1^n \multimap Y_1^n$ (and the model $P_{U,X,Y}$) forms a *Markov chain*. From this, the mutual information $I(U; Y_1^n)$ in (7) can be conveniently expressed as:

$$\sum_{\substack{x_1^n \in \mathbb{X}^n \\ u \in \mathbb{U}}} P_{U|X_1^n} P_{X_1^n} \log \left(\frac{1}{\sum_{x_1^n \in \mathbb{X}^n} P_{U|X_1^n} P_{X_1^n}} \right) \tag{8}$$

$$- \sum_{\substack{x_1^n \in \mathbb{X}^n \\ u \in \mathbb{U} \\ y_1^n \in \mathbb{Y}^n}} P_{Y_1^n|X_1^n} P_{U|X_1^n} P_{X_1^n} \log \left(\frac{\sum_{x_1^n \in \mathbb{X}^n} P_{Y_1^n|X_1^n} P_{X_1^n}}{\sum_{x_1^n \in \mathbb{X}^n} P_{Y_1^n|X_1^n} P_{U|X_1^n} P_{X_1^n}} \right), \tag{9}$$

where $P_{X_1^n}, P_{Y_1^n|X_1^n}$ and $P_{U|X_1^n}$ are short-hand notations for $P_{X_1^n}(x_1^n)$, $P_{Y_1^n|X_1^n}(y_1^n|x_1^n)$ and $P_{U|X_1^n}(u|x_1^n)$, respectively. In practical communication problems, the distribution of the sources at the nodes is often unknown. Then, instead of assuming P_X in (8), we assume a training (i.i.d) set $\{\overline{x}_1, \ldots, \overline{x}_m\}$, with $\overline{x}_i \in \mathbb{X}^n$ that will be used to approximate the expectations in (8) (w.r.t. P_{X^n}) by their respective empirical means. In addition, the expectation in (9) is proportional to the size of $\mathbb{X}^n \times \mathbb{Y}^n$. For large n, this expectation is impractical to compute. Alternatively, we assume that i.i.d. samples $S_{x_1^n} = \{\overline{y}_1, \ldots, \overline{y}_{m'}\}$ of Y_1^n given $X_1^n = x_1^n$ are available. These samples are used to approximate the expectation in (9) (w.r.t. $P_{X_1^n, Y_1^n}$) by their respective empirical average. Then, our empirical (and computationally tractable) version of $I(U; Y_1^n)$ writes as: $\hat{I}_\alpha(U; Y_1^n) \triangleq$

$$
\begin{aligned}
&= \frac{1}{m} \sum_{i=1}^{m} \sum_{u \in \mathbb{U}} P_{U|X_1^n}(u|\overline{x}_i) \log \left(\frac{1}{\frac{1}{m} \sum_{l=1}^{m} P_{U|X_1^n}(u|\overline{x}_l)} \right) \\
&- \frac{\alpha}{mm'} \sum_{i=1}^{m} \sum_{\overline{y}_j \in S_{\overline{x}_i}} \sum_{u \in \mathbb{U}} P_{U|X_1^n}(u|\overline{x}_i) \\
&\cdot \log \left(\frac{\sum_{l=1}^{m} P_{Y_1^n|X_1^n}(\overline{y}_j|\overline{x}_l)}{\sum_{l=1}^{m} P_{Y_1^n|X_1^n}(\overline{y}_j|\overline{x}_l) P_{U|X_1^n}(u|\overline{x}_l)} \right). \qquad (10)
\end{aligned}
$$

3.1.2. Soft quantizers based on Boltzmann distributions

We need to determine the collection of models $P_{U|X_1^n}$ in (10). Every $P_{U|X_1^n}$ is directly linked to the encoder f_n. Instead of using a deterministic mapping, we relax this assumption and we consider general soft-quantizers (or conditional distributions from X_1^n to U). We propose to relax this deterministic assumption to enrich the space of hypotheses on solving (7). In the process, our problem connect naturally with the type of info-max optimization addressed in representation learning. Following [6, 15], we consider the family of *Boltzmann distribution* as it is a rich and expressive enough collection of parametric distributions for $P_{U|X_1^n}$. More precisely, let

$$
p^W(u|x_1^n) \triangleq P_{U|X_1^n}(u|x_1^n, W, \tau) = \frac{e^{-\frac{\tau \|w_u - x_1^n\|^2}{2}}}{\sum_{l \in \mathbb{U}} e^{-\frac{\tau \|w_l - x_1^n\|^2}{2}}}, \qquad (11)
$$

where W is a weight matrix $W \in \mathbb{R}^{n \times |\mathbb{U}|}$, such that, $W = [w_1; \ldots; w_{|\mathbb{U}|}]$, $w_u \in \mathbb{X}^n$, $u \in \mathbb{U}$. This family of posterior distributions has been widely used in machine learning, because of its expressiveness and learning properties [16, 17]. At the end, we can address our main design problem as follows:

$$
W^* = \arg\max_W \hat{I}_\alpha(U; Y_1^n). \qquad (12)
$$

Importantly, the expression in (10) is smooth and differentiable with respect W, i.e., our collection $(p^W(u|x_1^n))$. Consequently, the solution of (12) can be approximated based on the Stochastic Gradient Descent (SGD) algorithm [18]. A pseudo-code for this mutual information maximization algorithm is provided here:

Algorithm 1 Mutual Information Maximization

1: Initialize: τ, α, λ, m (number of iterations), W
2: **for** $i \leftarrow 0$ **to** m **do**
3: $\quad p^W(u|\overline{x}_i) \leftarrow e^{-\frac{\tau \|w_u - \overline{x}_i\|^2}{2}}$ $u \in \mathbb{U}$
4: $\quad p^W(u|\overline{x}_i) \leftarrow \frac{p^W(u|\overline{x}_i)}{\sum_{l \in \mathbb{U}} p^W(u|\overline{x}_l)}$
5: $\quad w_u \leftarrow w_u - \lambda \frac{\partial \hat{I}_\alpha(Y_1^n; U)}{\partial w_u}$
6: **end for**
7: **Result:** Prediction $f(\overline{x}_i) = \arg\max_{u \in \mathbb{U}} p^W(u|\overline{x}_i)$

Finally, the solution in (12) is a weight matrix that produce a soft quantizer. The hard-quantizer or encoder (denoted by $f_n^W(\cdot)$) is obtained with the MAP (soft-max) rule:

$$
f_n^W(x_1^n) = \arg\max_{u \in \mathbb{U}} \frac{e^{-\frac{\tau \|w_u - x_1^n\|^2}{2}}}{\sum_{l \in \mathbb{U}} e^{-\frac{\tau \|w_l - x_1^n\|^2}{2}}}. \qquad (13)
$$

4. EXPERIMENTAL RESULTS

We evaluate the performance of our data-driven design $f_n^{W^*}(\cdot)$ in (13). Given $U = f_n^{W^*}(X_1^n)$, the decision rule ϕ_n is given by the (optimal) *Neyman-Pearson (NP) test* acting on U. Therefore, the decision (decoder) is given by the family: $\phi_n^\tau(u, y_1^n) = 0$ if $\frac{P_{U,Y_1^n}(u, y_1^n)}{P_U(u) P_{Y_1^n}(y_1^n)} > \tau$ and $\phi_n^\tau(u, y_1^n) = 1$, otherwise. Exploring a rich range of values for $\tau > 0$, we can determine the complete range of Type 1 and Type 2 errors associated to our design $f_n^{W^*}(\cdot)$. For the experimental setting, we consider a joint space of size $|\mathbb{X}| \times |\mathbb{Y}| = 17 \times 17$. We derive a discrete probability by partitioning $\mathbb{R}^2$ with a Gaussian density in $\mathbb{R}^2$ of parameters $(\mu_X, \mu_Y, \sigma_X, \sigma_Y . \rho)$. To this end, we assume that $\mu_X = \mu_Y$ and $\sigma_X^2 = \sigma_Y^2$ where ρ is the cross-correlation. Using this construction, we control the statistical dependency of the induced discrete vector (X, Y) with the parameter $\rho > 0$ of the continuous model.

	Quantization level $\lvert\mathbb{U}\rvert$				
ρ	4	20	50	160	200
0.2	85.59 %	73.41 %	65.42%	56.14%	50.34%
0.5	88.15 %	71.23 %	60.21%	51.84%	48.21%
0.8	58.82 %	30.57 %	26.54%	24.80%	22.39%

Table 1: Relative discrepancy of the power of the test with respect to the restriction-free case for different quantization levels. We fixed the Type I error to 0.02 and $n = 8$.

Figures 2 presents the ROC curve for different level statistical dependency or discrimination (indexed by ρ), for different quantization levels $|\mathbb{U}|$, and for different sample sizes n. On these curves, we contrast our strategy (continuous line)

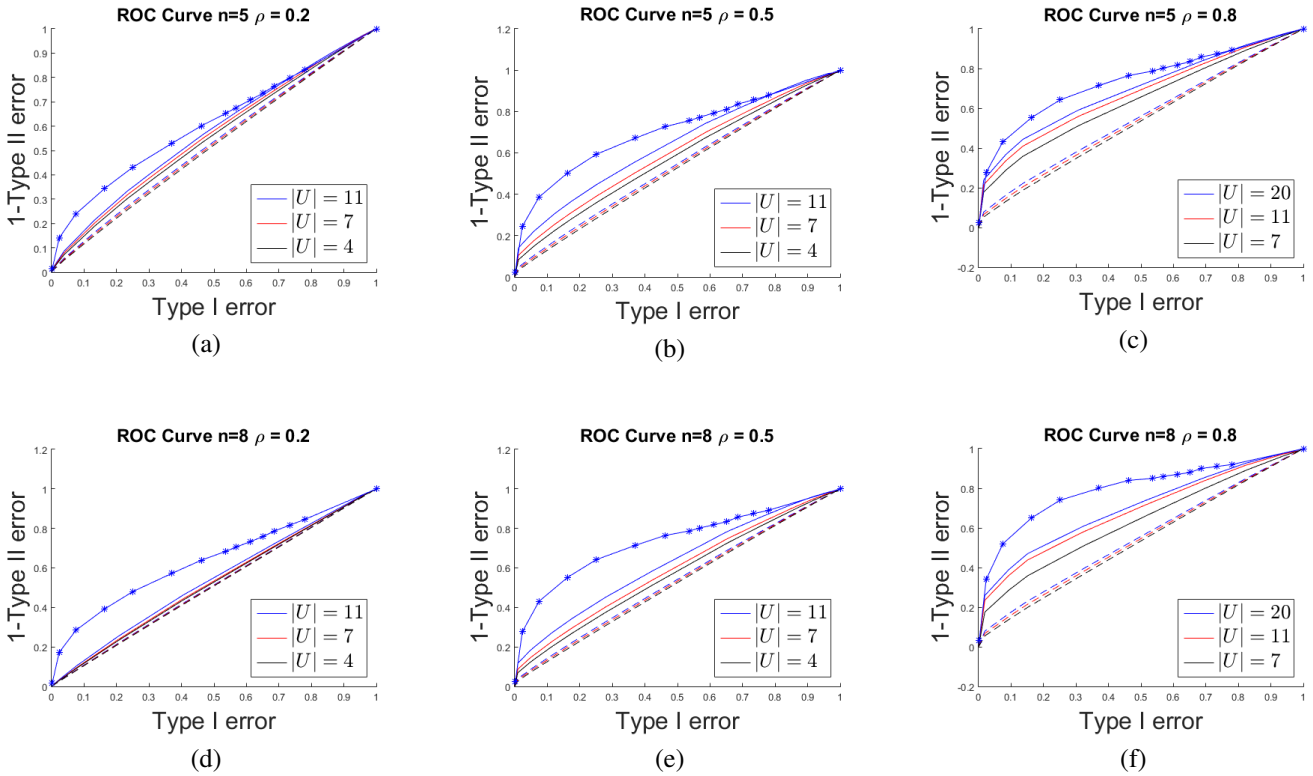

Fig. 2: Errors of the proposed design in (13) for different level ($|\mathbb{U}|$) and correlation (ρ) (solid lines). Dashed lines corresponds to the performances of a quantizer that minimizes $I(X; f_n(X^n))$. Blue line with asterisk is the oracle Neyman-Pearson test.

with an unsupervised method that is agnostic to the task and only tries to preserve the information of X_1^n (dashed line). As expected, in all the settings (indexed by n, ρ), as $|\mathbb{U}|$ increases the performance of both strategies improves: a large $|\mathbb{U}|$ implies that more bits $\log_2(|\mathbb{U}|)$ are transferred to the decision stage. However, our data-driven design shows a clear advantage in all the operational points of the ROC curve. Remarkable, this advantage is more prominent when the level of dependency between X and Y increases (by increasing ρ): a greater ρ implies a higher mutual information between (X, Y) and then a better trade-off between the errors [19].

From these results, our quantizer is effective in representing with few bits the information that the lossless sample offers to discriminate H_0 from H_1. This is observed when comparing the performances of our lossy strategy with the oracle (lossless) NP test acting on (X^n, Y^n), for the regime of higher mutual information (sub-figs. (b) and (c)). Here, $|\mathbb{U}|$ is a small (almost zero) fraction of the size of $\mathbb{X}^n$, however $f_n^{W^*}(X_1^n)$ produces modest degradation in performance with respect to the use of X_1^n. Table 1 shows the relative reduction of the power of the test with respect to the error-free case using a fixed Type I error of 0.02 and $n = 8$. Interestingly, for $\rho = 0.8$ we observe a relative degradation of 22.39% using $|\mathbb{U}|= 200$, which is a compression of almost 99.99% with respect to $|\mathbb{X}|^n= 17^8$. This observation implies that our infomax compression scheme can achieve close to oracle results by using a negligible fraction of the size of $\mathbb{X}^n$.

5. SUMMARY AND CONCLUDING REMARKS

This work investigated a data-driven quantization scheme for the problem of testing against independence with one-sided communication constraints. Building on the performance limit of this problem, an algorithm is proposed to tackle a multi-letter info-max learning task reminiscent of the type of representation for learning algorithms used in modern ML algorithms. Our algorithm is data-driven as it builds upon information obtained from i.i.d. samples (empirical averages). Although the proposed scheme uses partial information of the underlying model $P_{Y|X}$ (the channel), it can be directly extended to a scenario where $P_{Y|X}$ is also estimated from data. In the effectiveness of our design, concrete gains in detection errors are observed when compared with an unsupervised quantization design without exploiting the relevant information for the underlying test. For the design of distributed communications system oriented to binary detection, our work offers an effective information-driven strategy that improves decisions in distributed scenarios.

6. REFERENCES

[1] Minna Chen, Wei Liu, Biao Chen, and John Matyjas, "Quantization for distributed testing of independence," in *2010 13th International Conference on Information Fusion*. IEEE, 2010, pp. 1–5.

[2] Maggie Mhanna, Pablo Piantanida, and Pierre Duhamel, "Privacy-preserving quantization learning with applications to smart meters," in *IEEE International Conference on Communications, ICC 2017, Paris, France, May 21-25, 2017*. 2017, pp. 1–6, IEEE.

[3] Rudolf Ahlswede and Imre Csiszár, "Hypothesis testing with communication constraints," *IEEE Transactions on Information Theory*, vol. 32, no. 4, pp. 533–542, 1986.

[4] Sebastian Espinosa, Jorge F Silva, and Pablo Piantanida, "Finite-length bounds on hypothesis testing subject to vanishing type i error restrictions," *IEEE Signal Processing Letters*, vol. 28, pp. 229–233, 2021.

[5] Sebastian Espinosa, Jorge F Silva, and Pablo Piantanida, "New results on testing against independence with rate-limited constraints," in *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2019, pp. 1–5.

[6] Svetlana Lazebnik and Maxim Raginsky, "Supervised learning of quantizer codebooks by information loss minimization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 7, pp. 1294–1309, 2008.

[7] Zhaoqing Peng, Libo Zhang, and Tiejian Luo, "Learning to communicate via supervised attentional message processing," in *Proceedings of the 31st International Conference on Computer Animation and Social Agents*, 2018, pp. 11–16.

[8] Aaron van den Oorod, Oriol Vinyals, and Koray Kavukcuoglu, "Neural discrete representation learning," *arXiv preprint arXiv:1711.00937*, 2017.

[9] Naftali Tishby, Fernando C Pereira, and William Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[10] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy, "Deep variational information bottleneck," *arXiv preprint arXiv:1612.00410*, 2016.

[11] Bin Dai, Chen Zhu, Baining Guo, and David Wipf, "Compressing neural networks using the variational information bottleneck," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1135–1144.

[12] Matias Vera, Leonardo Rey Vega, and Pablo Piantanida, "Compression-based regularization with an application to multitask learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 5, pp. 1063–1076, 2018.

[13] Maurice Kendall, Alan Stuart, Keith J Ord, and Steven Arnold, *Kendall's Advanced Theory of Statistics: Volume 2A–Classical Inference and and the Linear Model*, 1999.

[14] Solomon Kullback and Richard A Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[15] Malik Boudiaf, Imtiaz Ziko, Jérôme Rony, Jose Dolz, Pablo Piantanida, and Ismail Ben Ayed, "Information maximization for few-shot learning," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 2445–2457, Curran Associates, Inc.

[16] Hannes Schulz, Andreas Müller, Sven Behnke, et al., "Investigating convergence of restricted boltzmann machine learning," in *NIPS 2010 Workshop on Deep Learning and Unsupervised Feature Learning*, 2010, vol. 1, pp. 6–1.

[17] F Xabier Albizuri, Alicia d'Anjou, Manuel Graña, and J Antonio Lozano, "Convergence properties of high-order boltzmann machines," *Neural networks*, vol. 9, no. 9, pp. 1561–1567, 1996.

[18] Kevin P Murphy, *Probabilistic Machine Learning: An introduction*, MIT Press, 2022.

[19] Thomas M Cover, *Elements of information theory*, John Wiley & Sons, 1999.