The Double-Sided Information-Bottleneck Function

Michael Dikshtein
Technion
Email: michaeldic@campus.technion.ac.il

Or Ordentlich
Hebrew University of Jerusalem
Email: or.ordentlich@mail.huji.ac.il

Shlomo Shamai (Shitz)
Technion
Email: sshlomo@ee.technion.ac.il

***Abstract*—We consider a two-terminal variant (double-sided) of the information bottleneck problem, which is related to biclustering. In our setup, X and Y are dependent random variables and the problem is to find two independent channels $\mathsf{P}_{\mathsf{U}|\mathsf{X}}$ and $\mathsf{P}_{\mathsf{V}|\mathsf{Y}}$ (setting the Markovian structure $\mathsf{U} \to \mathsf{X} \to \mathsf{Y} \to \mathsf{V}$) that maximize $I(\mathsf{U};\mathsf{V})$ subject to constraints on the relevant mutual information expressions: $I(\mathsf{U};\mathsf{X})$ and $I(\mathsf{V};\mathsf{Y})$. For jointly Gaussian X and Y, we show that Gaussian channels are optimal in the low-SNR regime, but not for general SNR. Similarly, it is shown that for a doubly symmetric binary source, binary symmetric channels are optimal when the correlation is low, and are suboptimal for high correlation. We conjecture that Z and S channels are optimal when the correlation is 1 (i.e., $\mathsf{X} = \mathsf{Y}$), and provide supporting numerical evidence.**

I. Introduction

Let (X,Y) be a bivariate source characterized by a fixed joint probability law P_{XY} and consider all Markov chains $\mathsf{U} \to \mathsf{X} \to \mathsf{Y} \to \mathsf{V}$. The Double-Sided Information-Bottleneck (DSIB) function is defined as [1]

$$R_{\mathsf{P}_{\mathsf{XY}}}(C_u, C_v) \triangleq \max I(\mathsf{U};\mathsf{V}), \tag{1}$$

where the maximization is over all $\mathsf{P}_{\mathsf{U}|\mathsf{X}}$ and $\mathsf{P}_{\mathsf{V}|\mathsf{Y}}$ satisfying $I(\mathsf{U};\mathsf{X}) \le C_u$ and $I(\mathsf{V};\mathsf{Y}) \le C_v$. This problem is illustrated in Figure 1. In our study we aim to determine the maximum value and the achieving conditional distributions $(\mathsf{P}_{\mathsf{U}|\mathsf{X}}, \mathsf{P}_{\mathsf{V}|\mathsf{Y}})$ (test channels) of $I(\mathsf{U};\mathsf{V})$ for various fixed sources P_{XY}, and constraints C_u, C_v.

The problem we consider originates from the domain of clustering. Clustering is applied to organize similar entities in unsupervised learning [2]. It has numerous practical applications in data science, such as joint word-document clustering, gene expression [3], and pattern recognition. The data in those applications is arranged as a contingency table. Usually, clustering is performed on one dimension of the table, but sometimes it is useful to apply clustering on both dimensions of the contingency table [4]. For example, when there is a strong correlation between the rows and the columns of the table or when high-dimensional sparse structures are handled.

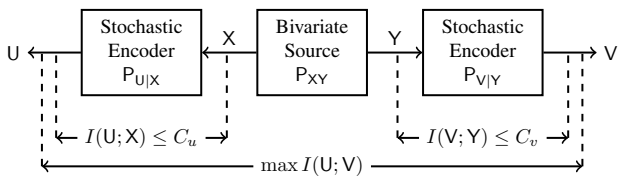

Fig. 1. Block diagram of the Double-Sided Information-Bottleneck function.

An information-theoretic approach to biclustering was initiated by Dhillon et al. [5]. They have regarded the normalized nonnegative contingency table as a joint probability distribution matrix of two random variables. Mutual information was proposed as a measure for optimal co-clustering. An optimization algorithm was presented that intertwines both row and column clustering at all stages. Distributed Clustering from a formal information-theoretic perspective was first explicitly considered by Pichler et al. [1], [6]. Consider a bivariate memoryless source with joint law P_{XY}. This source generates n i.i.d. copies $(\mathsf{X}^n, \mathsf{Y}^n)$, of (X,Y). Each sequence is observed at two different encoders, and each encoder generates a description of the observed sequence, $f(\mathsf{X}^n)$, and respectively, $g(\mathsf{Y}^n)$. The objective is to construct the mappings f and g such that the normalized mutual information between the descriptions would be maximal, while the description coding has bounded rate constraints. Single-letter inner and outer bounds for a general P_{XY} were derived. An example of a Doubly Symmetric Binary Source (DSBS) source was given, and several converse results were established. Furthermore, connections to the standard Information Bottleneck [7] and the Multiple Description CEO problems [8] were given. It was also shown that an information-theoretic biclustering problem is equivalent to hypothesis testing against independence with multiterminal data compression [9] and a pattern recognition problem [10].

The DSIB problem addressed in our paper is, in fact, a single letter version of the Distributed Clustering problem. The inner bound in [1] coincides with our definition of the problem. Moreover, if the Markov condition $\mathsf{U} \to \mathsf{X} \to \mathsf{Y} \to \mathsf{Z}$ is imposed on the multiletter variant, then those problems coincide.

A similar setting with a maximal correlation criterion between the reconstructed random variables has been considered in [11], [12]. Furthermore, it is sometimes the case that the optimal biclustering problem is easier to solve then the standard single-sided clustering problem. For example, Courtade-Kumar conjecture [13] for the standard single-sided clustering setting was ultimately proved for the biclustering setting [14]. A particular case, where (X,Y) are drawn from DSBS distribution, and the mappings f and g are restricted to be boolean function, was addressed in [14]. The bound $I(f(\mathsf{X}^n); g(\mathsf{Y}^n)) \le I(\mathsf{X};\mathsf{Y})$ was established, which is tight if and only if f and g are dictator functions. There are many

2021 IEEE International Symposium on Information Theory (ISIT) | 978-1-5386-8209-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISIT45174.2021.9517899

other variations of multi-user IB in the literature [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25].

A recent comprehensive tutorial on the Information Bottleneck method and related problems is given in [26]. Applications of Information Bottleneck problem in Machine Learning are detailed in [17], [27], [28], [29], [30], [31], [32].

II. Problem Formulation and Basic Properties

The DSIB function is closely related to the standard Single-Sided Information Bottleneck (SSIB) [7].

Definition 1 (SSIB): Let (X,V) be a pair of random variables with $|\mathcal{X}|=n$, $|\mathcal{V}|=m$ and fixed $\mathsf{P_{XV}}$. Denote by $\mathbf{q}_\mathsf{x}$ the marginal probability vector of X, and let T be the transition matrix from X to V, i.e.,

$$T_{ij} \triangleq \mathsf{P}\left(\mathsf{V}=i|\mathsf{X}=j\right), \qquad 1\le i\le m, 1\le j\le n.$$

Consider all random variables U satisfying the Markov chain $\mathsf{U}\to\mathsf{X}\to\mathsf{V}$. The SSIB function is defined as

$$\begin{aligned}\hat{R}_T(\mathbf{q}_\mathsf{x},C) \triangleq \ & \underset{\mathsf{P_{U|X}}}{\text{maximize}} \quad I(\mathsf{U};\mathsf{V}) \\ & \text{subject to} \quad I(\mathsf{X};\mathsf{U})\le C.\end{aligned} \tag{2}$$

The following tight cardinality bound in the single-sided counterpart of our problem was established in [33].

Lemma 1 ([33, Th. 9]): The optimization over U in (2) can be restricted to $|\mathcal{U}|\le n$.

This bound was actually already proved for the corresponding dual problem, namely, the IB Lagrangian, in [34]. But since $\hat{R}_T(\mathbf{q}_\mathsf{x},C)$ is generally not a strictly convex function of C, it cannot be directly applied for the primal problem (2).

Definition 2 (DSIB): Let (X,Y) be a pair of random variables with $|\mathcal{X}|=n$, $|\mathcal{Y}|=m$ and fixed $\mathsf{P_{XY}}$. Consider all the random variables U and V satisfying the Markov chain $\mathsf{U}\to\mathsf{X}\to\mathsf{Y}\to\mathsf{V}$. The DSIB function, $R:[0,H(\mathsf{X})]\times[0,H(\mathsf{Y})]\to\mathbb{R}_+$ is defined as

$$\begin{aligned}R_{\mathsf{P}_{\mathsf{xy}}}(C_u,C_v) \triangleq \ & \underset{\mathsf{P_{U|X}},\mathsf{P_{V|Y}}}{\text{maximize}} \quad I(\mathsf{U};\mathsf{V}) \\ & \text{subject to} \quad I(\mathsf{X};\mathsf{U})\le C_u, I(\mathsf{Y};\mathsf{V})\le C_v.\end{aligned} \tag{3}$$

Occasionally we will drop the subscript denoting the particular choice of the bivariate source $\mathsf{P_{XY}}$.

An equivalent form of (3) is

$$\begin{aligned}R(C_u,C_v) \triangleq \ & \underset{\mathsf{P_{V|Y}}}{\text{maximize}} \quad \underset{\mathsf{P_{U|X}}}{\text{maximize}} \quad I(\mathsf{U};\mathsf{V}). \\ & \text{subject to} \quad \text{subject to} \\ & I(\mathsf{Y};\mathsf{V})\le C_v \quad I(\mathsf{X};\mathsf{U})\le C_u\end{aligned} \tag{4}$$

Fix $\mathsf{P_{V|Y}}$ that satisfies $I(\mathsf{Y};\mathsf{V})\le C_v$. Denote by T_v the transition matrix from Y to V and respectively by W the transition matrix from X to Y. Since $\mathsf{P_{V|Y}}$ and $\mathsf{P_{XY}}$ are fixed, $\mathsf{P_{XV}}=\sum_y \mathsf{P_{V|Y}}(\cdot|y)\mathsf{P_{XY}}(\cdot,y)$ is also fixed. Denote by $T\triangleq T_vW$ the transition matrix from X to V. Therefore, the inner maximization term in (4) is just the SSIB function with parameters T and C_u, namely, $\hat{R}_T(\mathbf{q}_\mathsf{x},C_u)$. Hence, our problem can also be defined in the following two equivalent ways:

$$\begin{aligned}R(C_u,C_v) \triangleq \ & \underset{T_v}{\text{maximize}} \quad \hat{R}_{T_vW}(\mathbf{q}_\mathsf{x},C_u) \\ & \text{subject to} \quad I(\mathbf{q}_\mathsf{y},T_v)\le C_v\end{aligned}, \tag{5}$$

or

$$\begin{aligned}R(C_u,C_v) \triangleq \ & \underset{T_u}{\text{maximize}} \quad \hat{R}_{T_u\bar{W}}(\mathbf{q}_\mathsf{y},C_v) \\ & \text{subject to} \quad I(\mathbf{q}_\mathsf{x},T_u)\le C_u.\end{aligned}, \tag{6}$$

where T_u is the transition matrix from X to U, and $\bar{W}$ is the transition matrix from Y to X. This representation gives us a different perspective on our problem as an optimal channel for the SSIB setting.

The bound from Lemma 1 can be utilized to give cardinality bound for the double sided problem.

Proposition 2.1: For the DSIB optimization problem defined in (3) it suffices to consider random variables U and V with cardinalities $|\mathcal{U}|\le n$ and $|\mathcal{V}|\le m$.

Proof. Let T_u and T_v be two arbitrary transition matrices. By Lemma 1, there exists $T_{\tilde{u}}$ with $|\tilde{\mathcal{U}}|\le n$ such that $I(\tilde{\mathsf{U}};\mathsf{V})\ge I(\mathsf{U};\mathsf{V})$ and $I(\mathsf{X};\tilde{\mathsf{U}})\le C_u$. Similarly, T_v can be replaced with $T_{\tilde{v}}$, $|\tilde{\mathcal{V}}|\le m$ such that

$$I(\tilde{\mathsf{U}};\tilde{\mathsf{V}})\ge I(\tilde{\mathsf{U}},\mathsf{V})\ge I(\mathsf{U};\mathsf{V}),$$

and $I(\mathsf{Y};\tilde{\mathsf{V}})\le C_v$. Therefore, there exists an optimal solution with $|\mathcal{U}|=n$ and $|\mathcal{V}|=m$. ■

III. Main Results

In this section we will present the main analytical outcomes of our study. First we consider the scenario where our bivariate source is binary, and specifically DSBS. Then we treat the case where X and Y are jointly Gaussian.

A. Doubly Symmetric Binary Source (DSBS)

Let (X,Y) be a DSBS with parameter p, i.e., $\mathsf{P_{XY}}(x,y)=\frac{1}{2}(p\cdot\mathbb{1}(x\ne y)+(1-p)\mathbb{1}(x=y))$. Here we emphasize the dependence of the DSIB on the parameter p as $R(C_u,C_v,p)$.

A special case where we have a complete analytical solution is when p tends to $1/2$. Let $h_2(p):[0,1]\to[0,1]$ be the binary entropy function and $h_2^{-1}(\cdot)$ its inverse, restricted to $[0,1/2]$. Throughout this paper all logarithms are taken to base 2 unless stated otherwise.

Theorem 1: Suppose $p=\frac{1}{2}-\epsilon$, and consider $\epsilon\to 0$. Then

$$R(C_u,C_v,\epsilon)=2\epsilon^2\log e\cdot(1-2h_2^{-1}(1-C_u))^2(1-2h_2^{-1}(1-C_v))^2+o(\epsilon^2). \tag{7}$$

This theorem is proved in Appendix A. For the lower bound we take $\mathsf{P_{U|X}}$ and $\mathsf{P_{V|Y}}$ to be Binary Symmetric Channels (BSCs). In the following section we will give a numerical evidence that BSC test-channels are in fact optimal provided that p is sufficiently large. However, for small p this is no longer the case and we believe the following holds.

Conjecture 1: Let $\mathsf{X}=\mathsf{Y}$, i.e., $p=0$. The test channels $\mathsf{P_{U|X}}$ and $\mathsf{P_{V|X}}$ that achieve $R(C_u,C_v,0)$ are Z-channel and S-channel respectively.

Remark 1: Our results in the numerical section strongly support this conjecture. In fact they prove it within the resolution of the experiments, i.e., for optimizing over a dense set of test-channels rather then all test-channels. Nevertheless, we were not able to find an analytical proof for this result.

Remark 2: Suppose $\mathsf{X}=\mathsf{Y}$, $I(\mathsf{X};\mathsf{U})=C_u$, and $I(\mathsf{X};\mathsf{V})=C_v$. Since $I(\mathsf{U};\mathsf{V})=I(\mathsf{U};\mathsf{X})+I(\mathsf{V};\mathsf{X})-I(\mathsf{X};\mathsf{U},\mathsf{V})$ (as $\mathsf{U}\to\mathsf{X}\to\mathsf{Y}\to\mathsf{V}$ form a Markov chain in this order) then maximizing $I(\mathsf{U};\mathsf{V})$ is equivalent to minimizing $I(\mathsf{X};\mathsf{U},\mathsf{V})$, namely, minimizing information combining as in [35], [22]. Therefore, Conjecture 1 is equivalent to the conjecture that among all channels with $I(\mathsf{X};\mathsf{U})\geq C_u$ and $I(\mathsf{Y};\mathsf{V})\geq C_v$, Z and S are the worst channels for information combining.

This observation leads us the following additional conjecture.

Conjecture 2: The test channels $\mathsf{P}_{\mathsf{U}|\mathsf{X}}$ and $\mathsf{P}_{\mathsf{V}|\mathsf{X}}$ that maximize $I(\mathsf{X};\mathsf{U},\mathsf{V})$ are both Z channels.

Remark 3: Suppose now that p is arbitrary and assume that one of the channels $\mathsf{P}_{\mathsf{U}|\mathsf{X}}$ or $\mathsf{P}_{\mathsf{V}|\mathsf{Y}}$ is restricted to be a Binary Memoryless Symmetric (BMS) channel [36, Ch. 4], then the maximal $I(\mathsf{U};\mathsf{V})$ is attained by BSC channels, as those are the symmetric channels minimizing $I(\mathsf{X};\mathsf{U},\mathsf{V})$ [35]. It is not surprising that once the BMS constraint is removed, symmetric channels are no longer optimal (see the discussion in [35, Sec. VI.C]).

B. Gaussian Double-Sided Information Bottleneck (GDSIB)

In this subsection we consider a specific setting where (X,Y) is a Gaussian bivariate source, namely,

$$\begin{pmatrix}\mathsf{X}\\ \mathsf{Y}\end{pmatrix}\sim\mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1&\rho\\ \rho&1\end{pmatrix}\right). \tag{8}$$

We establish achievability schemes and show that Gaussian test-channels $\mathsf{P}_{\mathsf{U}|\mathsf{X}}$ and $\mathsf{P}_{\mathsf{V}|\mathsf{Y}}$ are optimal for vanishing SNR. Furthermore we show an elegant representation of the problem through *Hermite polynomials*. We denote the Gaussian DSIB function with explicit dependency on ρ as $R(C_u,C_v,\rho)$.

Proposition 3.1: Let $H_n(\cdot)$ be the nth order probabilistic Hermite polynomial, then,

$$I(\mathsf{U};\mathsf{V})=\mathbb{E}_{\mathsf{UV}}\left[\log\left(\sum_{n=0}^{\infty}\frac{\rho^n}{n!}\mathbb{E}\left[H_n(\mathsf{X})|\mathsf{U}\right]\mathbb{E}\left[H_n(\mathsf{Y})|\mathsf{V}\right]\right)\right]. \tag{9}$$

This representation follows by considering $I(\mathsf{U};\mathsf{V})=D(\mathsf{P}_{\mathsf{UV}}||\mathsf{P}_{\mathsf{U}}\cdot\mathsf{P}_{\mathsf{V}})$ and expressing $\frac{\mathsf{P}_{\mathsf{UV}}}{\mathsf{P}_{\mathsf{U}}\cdot\mathsf{P}_{\mathsf{V}}}$ using Mehler Kernel [37]. Mehler Kernel decomposition is a special case of a much richer family of Lancaster distributions [38].

Now we give two lower bounds on $R(C_u,C_v,\rho)$. Our first lower bound is established by choosing $\mathsf{P}_{\mathsf{U}|\mathsf{X}}$ and $\mathsf{P}_{\mathsf{V}|\mathsf{Y}}$ to be additive Gaussian channels, satisfying the mutual information (MI) constraints with equality.

Proposition 3.2: A lower bound on $R(C_u,C_v,\rho)$ is given by

$$R(C_u,C_v,\rho)\geq-\frac{1}{2}\log\left(1-\rho^2\left(1-2^{-2C_u}\right)\left(1-2^{-2C_v}\right)\right). \tag{10}$$

Although it was shown in [17] that choosing the test channel to be Gaussian is optimal for the single-sided variant, it is not the case for its double-sided extension. We will show this by examining a specific set of values for the rates constraints, $(C_u,C_v)=(1,1)$. Furthermore, we choose the test channels $\mathsf{P}_{\mathsf{U}|\mathsf{X}}$ and $\mathsf{P}_{\mathsf{V}|\mathsf{Y}}$ to be deterministic quantizers.

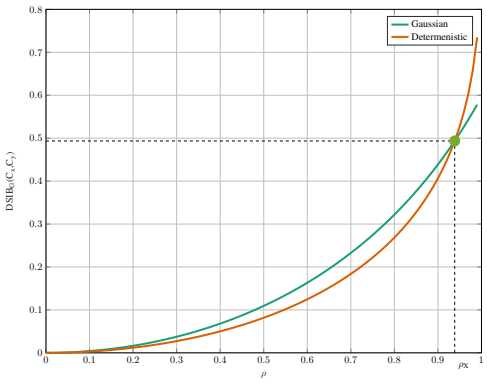

Fig. 2. Comparison of the lower bounds from Proposition 3.2 and Proposition 3.3.

Proposition 3.3: Let $(C_u,C_v)=(1,1)$, then

$$R(1,1,\rho)\geq 1-h_2\left(\frac{\arccos\rho}{\pi}\right). \tag{11}$$

The proof of this bound is developed in Appendix B.

We compare the bounds from Proposition 3.2 and Proposition 3.3 with $(C_u,C_v)=(1,1)$ in Figure 2. The most unexpected observation here is that the deterministic quantizers lower bound outperform the Gaussian test-channels for high values of ρ. The crossing point of those bounds is

$$\rho_{\text{cros}}=\frac{e}{\sqrt{1+e^2}}\to\sqrt{SNR_{\text{cros}}}=\frac{\rho_{\text{cros}}}{\sqrt{1-\rho_{\text{cros}}^2}}=e. \tag{12}$$

For $\rho\to 0$, the exact asymptotic behavior of the Gaussian and deterministic test-channels, respectively, for $C_u=C_v=1\,\text{bit}$ is

$$\lim_{\rho\to0}-\frac{1}{2}\log\left(1-\rho^2(1-2^{-2C_u})(1-2^{-2C_v})\right)=\frac{9\log e}{32}\rho^2+o(\rho^2),$$
$$\lim_{\rho\to0}1-h_2\left(\frac{\arccos\rho}{\pi}\right)=\frac{2\log e}{\pi^2}\rho^2+o(\rho^2).$$

Hence, the Gaussian choice outperforms the second lower bound for vanishing SNR.

Theorem 2: For small ρ, the GDSIB function is given by

$$R(C_u,C_v,\rho)=\frac{\rho^2\log e}{2}(1-2^{-2C_u})(1-2^{-2C_v})+o(\rho^2). \tag{13}$$

The lower bound follows from Proposition 3.2. The upper bound is established by considering the kernel representation from Proposition 3.1 in the limit of vanishing ρ. The detailed proof is given in Appendix C.

IV. Numerical Results

In this section we consider the case where P_{XY} is a DSBS with parameter p. Since V is binary, we can describe $\mathsf{P}_{\mathsf{V}|\mathsf{Y}}$ using the following transition matrix

$$T_v\triangleq\begin{pmatrix}\alpha&\beta\\ 1-\alpha&1-\beta\end{pmatrix}. \tag{14}$$

Now, let $\mathcal{F}_v$ be the set of all feasible $\mathsf{P}_{\mathsf{V}|\mathsf{Y}}$, i.e.,

$$\mathcal{F}_v(C_v) \triangleq \left\{(\alpha,\beta)\colon h\left(\frac{\alpha+\beta}{2}\right) - \frac{1}{2}\left(h(\alpha)+h(\beta)\right) = C_v\right\}$$

(equality constraint follows since we maximize a convex function over a convex set). This set defines a curve $\beta(\alpha) \triangleq \beta(\alpha, C_v)$. Examples of such curves for various values of C_v are shown in Figure 3.

Fix $(\alpha, \beta(\alpha)) \in \mathcal{F}_v$ and consider $\hat{R}_T(\mathbf{q}, C_u) = \hat{R}(C_u, \alpha, p)$ from (2), where $\bar{x} \triangleq 1 - x$ and

$$T = T_v W = \begin{pmatrix} \alpha\bar{p} + \beta(\alpha)p & \alpha p + \beta(\alpha)\bar{p} \\ \bar{\alpha}\bar{p} + \bar{\beta}(\alpha)p & \bar{\alpha}p + \bar{\beta}(\alpha)\bar{p} \end{pmatrix}. \tag{15}$$

Since SSIB is equivalent to the problem that was studied in [39] (as was shown in [40, Sec. 3]), we can utilize the recipe from [39, Sec. IV] to evaluate $\hat{R}(C_u, \alpha, p)$. With that tool in hand, we can compute $\hat{R}(C_u, \alpha, p)$ for every $\alpha \in \mathcal{F}_v$ and then find $\alpha^\star$ that maximizes $\hat{R}(C_u, \alpha, p)$. This is equivalent to

$$R(C_u, C_v, p) \triangleq \underset{\alpha \in \mathcal{F}_v}{\text{maximize}} \quad \hat{R}(C_u, \alpha, p). \tag{16}$$

The graph of $\hat{R}(C_u, \alpha, p)$ for $C_u = C_v = 0.3\,\text{bit}$ and various values of p is shown in Figure 4. For $p = 0$ ($\mathsf{X} = \mathsf{Y}$) we see that $\alpha^\star = 1$, in such case $\mathsf{P}_{\mathsf{V}|\mathsf{Y}}$ and $\mathsf{P}_{\mathsf{U}|\mathsf{X}}$ are Z and S channels, respectively. For small but increasing values of p we observe a smooth transition from Z-channel, $\alpha^\star(p=0)$, points on the curve $\beta(\alpha)$, towards the point $\alpha = 1-\beta$, which corresponds to a BSC. The graphs of $R(C_u, C_v, \cdot)$, $\alpha^\star(\cdot)$, $\beta^\star(\alpha^\star(\cdot))$ and $\alpha^\star(\cdot) + \beta(\alpha^\star(\cdot))$ are shown in Figure 5. As expected $R(C_u, C_v, \cdot)$ is a decreasing function of p. Furthermore, the channel transition probability has a threshold value, θ, which depends on C_u and C_v. If $p > \theta(C_u, C_v)$ then $\mathsf{P}_{\mathsf{V}|\mathsf{Y}}$ is a BSC (and thus also $\mathsf{P}_{\mathsf{U}|\mathsf{X}}$ [39, Sec. IV.A]).

V. Concluding Remarks

In this paper we have considered the Double-Sided Information-Bottleneck problem. Tight cardinality bounds on the auxiliary random variables were obtained for an arbitrary discrete bivariate source. For DSBS we have shown that BSC test-channel are optimal when $p \to 0.5$. Furthermore, we have shown numerical simulation for arbitrary p, indicating that Z and S channels are optimal for $p = 0$. As for the Gaussian bivariate source, representation of $I(\mathsf{U};\mathsf{V})$ utilizing Hermite polynomials was given. Also, the optimality of Gaussian test-channel was demonstrated for vanishing SNR. Moreover, we have constructed a lower bound attained by deterministic quantizers that outperforms the jointly Gaussian choice at high SNR. Note that the solution for the n-letter problem $\max \frac{1}{n} I(\mathsf{U};\mathsf{V})$ for $\mathsf{U} \to \mathsf{X}^n \to \mathsf{Y}^n \to \mathsf{V}$ under constraints $I(\mathsf{U};\mathsf{X}^n) \le nC_u$ and $I(\mathsf{V};\mathsf{Y}^n) \le nC_v$, does not tensorize in general. For $\mathsf{X}^n = \mathsf{Y}^n \sim \text{Ber}^{\otimes n}(0.5)$, we can easily achieve the cut-set bound $I(\mathsf{U};\mathsf{V})/n = \min\{C_u, C_v\}$. In addition, if time sharing is allowed, the results change drastically.

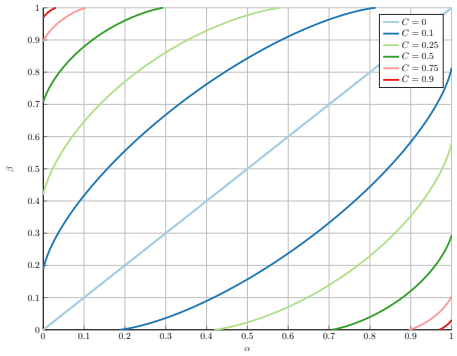

Fig. 3. Set of feasible curves $\beta(\alpha, C)$ for various values of C.

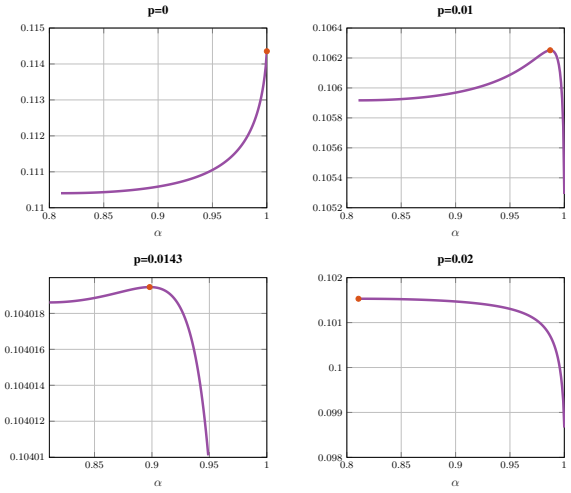

Fig. 4. Evaluation of $\hat{R}(C_u, \alpha, p)$ for $C_u = C_v = 0.3\,\text{bit}$ and various values of p. Recall that $1 - h_2^{-1}(1-0.3) = 0.8107$.

Acknowledgment

The work has been supported by the European Union's Horizon 2020 Research And Innovation Programme, grant

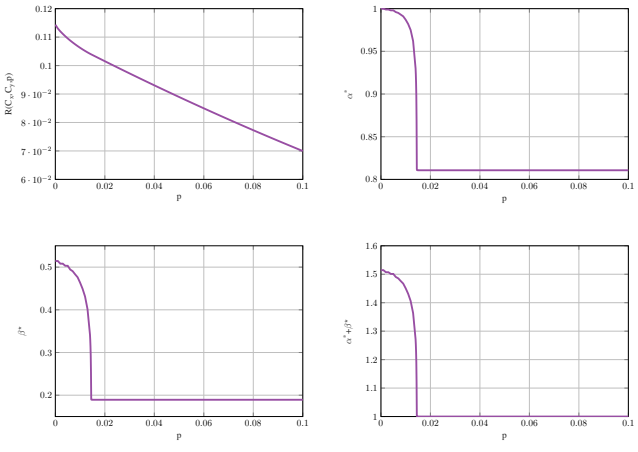

Fig. 5. Evaluation of $R(C_u, C_v, p)$, $\alpha^\star$, $\beta^\star \triangleq \beta(\alpha^\star)$, and $\alpha^\star + \beta^\star$ for $C_u = C_v = 0.3\,\text{bit}$.

agreement no. 694630, by the ISF under Grant 1791/17, and by the WIN consortium via the Israel minister of economy and science.

APPENDIX

A. Proof of Theorem 1

Fix $\mathsf{P}_{\mathsf{U}|\mathsf{X}}$ and $\mathsf{P}_{\mathsf{V}|\mathsf{Y}}$ that satisfy $I(\mathsf{X};\mathsf{U}) \le C_u$ and $I(\mathsf{Y};\mathsf{V}) \le C_v$. Denote $\alpha_u \triangleq \mathsf{P}(\mathsf{X}=1|\mathsf{U}=u)$, $\beta_v \triangleq \mathsf{P}(\mathsf{Y}=1|\mathsf{V}=v)$. Using this notation we have

$$I(\mathsf{U};\mathsf{V}) = \sum_{u,v} P_{\mathsf{U}}(u)P_{\mathsf{V}}(v)K(u,v,p)\log K(u,v,p), \quad (17)$$

where $K(u,v,p) \triangleq \frac{P_{\mathsf{UV}}(u,v)}{P_{\mathsf{U}}(u)P_{\mathsf{V}}(v)}$, and satisfies

$$K(u,v,p)=2(\bar{\alpha}_u\bar{p}\bar{\beta}_v+\alpha_u p\bar{\beta}_v+\bar{\alpha}_u p\beta_v+\alpha_u\bar{p}\beta_v)=2\alpha_u*\beta_v*\bar{p}. \quad (18)$$

Denoting $\epsilon \triangleq \frac{1}{2}-p$ ($p=\frac{1}{2}-\epsilon$), we obtain

$$K(u,v,\epsilon) = 1 + 2\epsilon(1-2\alpha_u)(1-2\beta_v). \quad (19)$$

Now we rewrite $I(\mathsf{U};\mathsf{V})$ with explicit dependency on ϵ as $I(\epsilon) = \sum_{u,v}P_{\mathsf{U}}(u)P_{\mathsf{V}}(v)K(u,v,\epsilon)\log K(u,v,\epsilon)$. We would like to expand $I(\epsilon)$ with Taylor series around $\epsilon = 0$. Note that $I(0) = 0 = I'(\epsilon)|_{\epsilon=0}$. Furthermore, the second derivative is given by

$$I''(\epsilon)|_{\epsilon=0}=4\log e\cdot\left(\sum_u P_{\mathsf{U}}(u)(1-2\alpha_u)^2\right)\left(\sum_v P_{\mathsf{V}}(v)(1-2\beta_v)^2\right).$$

Hence,

$$I(\epsilon)=2\epsilon^2\log e\cdot\left(\sum_u P_{\mathsf{U}}(u)(1-2\alpha_u)^2\right)\left(\sum_v P_{\mathsf{V}}(v)(1-2\beta_v)^2\right)+o(\epsilon^2).$$

Now, note that

$$\alpha_u = \begin{cases} h_2^{-1}(H(\mathsf{X}|\mathsf{U}=u)) & \alpha_u \le \frac{1}{2} \\ 1-h_2^{-1}(H(\mathsf{X}|\mathsf{U}=u)) & \alpha_u > \frac{1}{2} \end{cases} \quad (20)$$

with similar relation for β_v. Therefore,

$$\begin{aligned} I(\epsilon) &= \frac{2\epsilon^2}{\ln 2}\cdot\mathbb{E}_u\left[\left(1-2h_2^{-1}(H(\mathsf{X}|\mathsf{U}=u))\right)^2\right]\cdot\mathbb{E}_v\left[\left(1-2h_2^{-1}(H(\mathsf{Y}|\mathsf{V}=v))\right)^2\right]+o(\epsilon^2) \\ &\le 2\epsilon^2\log e\cdot(1-2h_2^{-1}(H(\mathsf{X}|\mathsf{U})))^2(1-2h_2^{-1}(H(\mathsf{Y}|\mathsf{V})))^2+o(\epsilon^2) \\ &\le 2\epsilon^2\log e\cdot(1-2h_2^{-1}(1-C_x))^2(1-2h_2^{-1}(1-C_y))^2+o(\epsilon^2), \end{aligned}$$

where the first inequality follows since the function $f: x \mapsto (1-2h_2^{-1}(x))^2$ is concave and applying Jensen's inequality (the proof is omitted due to space limitation), and the second inequality follows from rate constraints.

B. Proof of Proposition 3.3

We choose U and V to be deterministic functions of X and respectively Y, i.e., $\mathsf{U} = \mathrm{sign}(\mathsf{X})$ and $\mathsf{V} = \mathrm{sign}(\mathsf{Y})$. In such case the rate constraints are met with equality, namely, $I(\mathsf{U};\mathsf{X}) = 1 = I(\mathsf{Y};\mathsf{V})$. We proceed to evaluate the achievable rate,

$$\begin{aligned} I(\mathsf{U};\mathsf{V}) &= 1-\mathsf{P}(\mathsf{U}=0)h_2(\mathsf{P}(\mathsf{V}=1|\mathsf{U}=0))-\mathsf{P}(\mathsf{U}=1)h_2(\mathsf{P}(\mathsf{V}=0|\mathsf{U}=1)) \\ &\overset{(a)}{=} 1-h_2(\mathsf{P}(\mathsf{U}\neq\mathsf{V})), \end{aligned}$$

where equality in (a) follows since $\mathsf{P}(\mathsf{V}=1|\mathsf{U}=0) = \mathsf{P}(\mathsf{V}=0|\mathsf{U}=1)$ by symmetry. We therefore obtain the following formula for the "error probability":

$$\mathsf{P}(\mathsf{V}\neq\mathsf{U})=1-\mathsf{P}(\mathsf{X}<0,\mathsf{Y}<0)-\mathsf{P}(\mathsf{X}>0,\mathsf{Y}>0)\overset{(a)}{=}1-2\mathsf{P}(\mathsf{X}<0,\mathsf{Y}<0),$$

where (a) also follows from symmetry. Utilizing Sheppard's Formula [41, Ch. 5, p.107], we have $1-2\mathsf{P}(\mathsf{X}<0,\mathsf{Y}<0) = \frac{\arccos\rho}{\pi}$. This completes the proof of the proposition.

C. Proof of Theorem 2

We assume U and V are continuous RVs. The proof for the discrete case is identical. The joint density $f_{\mathsf{UV}}(u,v)$ can be expressed with explicit dependency on ρ as follows:

$$f(u,v;\rho) \triangleq f_{\mathsf{U}}(u)f_{\mathsf{V}}(v)\iint_{\mathbb{R}^2} f_{\mathsf{X}|\mathsf{U}}(x|u)M(x,y;\rho)f_{\mathsf{Y}|\mathsf{V}}(y|v)\mathrm{d}x\mathrm{d}y,$$

where $M(x,y;\rho) = \sum_{n=0}^{\infty}\frac{\rho^n}{n!}H_n(x)H_n(y)$ [37]. Similarly, $I(\mathsf{U};\mathsf{V})$ can also be written with explicit dependency on ρ

$$I(\rho) \triangleq I_\rho(\mathsf{U};\mathsf{V}) = \int\int f(u,v;\rho)\log\frac{f(u,v;\rho)}{f_{\mathsf{U}}(u)f_{\mathsf{V}}(v)}\mathrm{d}u\mathrm{d}v.$$

We would like to approximate $I(\rho)$ in the limit $\rho \to 0$ using a Taylor series up to a second order in ρ. As a first step we evaluate the first two derivatives of $f(u,v;\rho)$ at $\rho = 0$. Note that $M(x,y;0) = 1$ and

$$\frac{\mathrm{d}M}{\mathrm{d}\rho}\Big|_{\rho=0} = xy, \qquad \frac{\mathrm{d}^2M}{\mathrm{d}\rho^2}\Big|_{\rho=0} = (x^2-1)(y^2-1). \quad (21)$$

Thus, $f(u,v;0) = f_{\mathsf{U}}(u)f_{\mathsf{V}}(v)$,

$$\left.\frac{\mathrm{d}f}{\mathrm{d}\rho}\right|_{\rho=0} = f_{\mathsf{U}}(u)f_{\mathsf{V}}(v)\mathbb{E}[\mathsf{X}|\mathsf{U}=u]\,\mathbb{E}[\mathsf{Y}|\mathsf{V}=v],$$

and $\frac{\mathrm{d}^2f}{\mathrm{d}\rho^2}\big|_{\rho=0}=f_{\mathsf{U}}(u)f_{\mathsf{V}}(v)\big(\mathbb{E}[\mathsf{X}^2|\mathsf{U}=u]-1\big)\big(\mathbb{E}[\mathsf{Y}^2|\mathsf{V}=v]-1\big)$.

Expanding $I(\rho)$ in Taylor series around $\rho = 0$ gives us $I(0) = 0 = \frac{\mathrm{d}I(\rho)}{\mathrm{d}\rho}\big|_{\rho=0}$ and

$$\frac{\mathrm{d}^2I(\rho)}{\mathrm{d}\rho^2}\Big|_{\rho=0} = \log e\cdot\mathbb{E}\left[\left(\mathbb{E}[\mathsf{X}|\mathsf{U}]\right)^2\right]\mathbb{E}\left[\left(\mathbb{E}[\mathsf{Y}|\mathsf{V}]\right)^2\right].$$

Thus

$$I(\rho) = \frac{\rho^2\log e}{2}\mathbb{E}\left[\left(\mathbb{E}[\mathsf{X}|\mathsf{U}]\right)^2\right]\mathbb{E}\left[\left(\mathbb{E}[\mathsf{Y}|\mathsf{V}]\right)^2\right]+o(\rho^2). \quad (22)$$

Note that $\mathbb{E}[\mathsf{X}] = \mathbb{E}[\mathbb{E}[\mathsf{X}|\mathsf{U}]]$ and

$$1=\mathbb{E}[\mathsf{X}^2]=\mathbb{E}\big[\mathbb{E}[\mathsf{X}^2|\mathsf{U}]\big]=\mathbb{E}[\mathrm{var}[\mathsf{X}|\mathsf{U}]]+\mathbb{E}\left[\left(\mathbb{E}[\mathsf{X}|\mathsf{U}]\right)^2\right]. \quad (23)$$

Also, by [42, Corollary to Theorem 8.6.6], $\mathbb{E}[\mathrm{var}[\mathsf{X}|\mathsf{U}]] \ge \frac{1}{2\pi e}e^{2h(\mathsf{X}|\mathsf{U})}$. Moreover, from MI constraint, we have

$$I(\mathsf{X};\mathsf{U}) = h(\mathsf{X}) - h(\mathsf{X}|\mathsf{U}) = \frac{1}{2}\log(2\pi e) - h(\mathsf{X}|\mathsf{U}) \le C_u,$$

and therefore $h(\mathsf{X}|\mathsf{U}) \ge \log(2\pi e) - C_u$. Thus we get

$$-C_u \le \frac{1}{2}\log(\mathbb{E}[\mathrm{var}[\mathsf{X}|\mathsf{U}]]) \to \mathbb{E}[\mathrm{var}[\mathsf{X}|\mathsf{U}]] \ge 2^{-2C_u}. \quad (24)$$

Combining (23) and (24), we obtain $\mathbb{E}\left[\left(\mathbb{E}[\mathsf{X}|\mathsf{U}]\right)^2\right] \le 1-2^{-2C_u}$. In a very similar method one can show that $\mathbb{E}\left[\left(\mathbb{E}[\mathsf{Y}|\mathsf{V}]\right)^2\right] \le 1-2^{-2C_v}$. Thus, for $\rho \to 0$

$$I(\rho) \le \frac{\rho^2\log e}{2}(1-2^{-2C_u})(1-2^{-2C_v})+o(\rho^2). \quad (25)$$

REFERENCES

[1] G. Pichler, P. Piantanida, and G. Matz, "Distributed Information-Theoretic Biclustering," in *Proc. 2016 IEEE Int. Symp. Inf. Theory*, July 2016, pp. 1083–1087.

[2] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Prentice-Hall, 1988.

[3] N. Gupta and S. Aggarwal, "Modeling Biclustering as an optimization problem using Mutual Information," in *2009 Proceeding of International Conference on Methods and Models in Computer Science (ICM2CS)*, 2009, pp. 1–5.

[4] J. Hartigan, "Direct Clustering of a Data Matrix," *Journal of the American Statistical Association*, vol. 67, no. 337, pp. 123–129, 1972.

[5] I. S. Dhillon, S. Mallela, and D. S. Modha, "Information-Theoretic Co-clustering," in *Proc. The Ninth ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining*, ser. (KDD '03), 2003.

[6] G. Pichler, P. Piantanida, and G. Matz, "Distributed Information-Theoretic Clustering," *arXiv*, vol. abs/1602.04605, 2020.

[7] N. Tishby, F. C. N. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Annu. Allerton Conf. Commun. Control Comput.*, Sep. 1999, pp. 368—-377.

[8] T. A. Courtade and G. R. Kumar, "Which Boolean Functions Maximize Mutual Information on Noisy Inputs?" *IEEE Trans. Inf. Theory*, vol. 60, no. 8, pp. 4515–4525, 2014.

[9] T. S. Han, "Hypothesis Testing with Multiterminal Data Compression," *IEEE Trans. Inf. Theory*, vol. 33, no. 6, pp. 759–772, 1987.

[10] M. B. Westover and J. A. O'Sullivan, "Achievable Rates for Pattern Recognition," *IEEE Trans. Inf. Theory*, vol. 54, no. 1, pp. 299–320, 2008.

[11] A. Painsky, M. Feder, and N. Tishby, "An information-theoretic framework for non-linear canonical correlation analysis," *CoRR*, vol. abs/1810.13259, 2018.

[12] A. R. Williamson, "The Impacts of Additive Noise and 1-bit Quantization on the Correlation Coefficient in the Low-SNR Regime," in *Proc. 57th Annu. Allerton Conf. Commun. Control Comput.*, Sep. 2019, pp. 631–638.

[13] T. A. Courtade and T. Weissman, "Multiterminal Source Coding Under Logarithmic Loss," *IEEE Trans. Inf. Theory*, vol. 60, no. 1, pp. 740–761, 2014.

[14] G. Pichler, P. Piantanida, and G. Matz, "Dictator Functions Maximize Mutual Information," *Ann. Appl. Prob.*, vol. 28, no. 5, pp. 3094–3101, 2018.

[15] I. Estella and A. Zaidi, "Distributed Information Bottleneck Method for Discrete and Gaussian Sources," in *Proc. Int. Zurich Seminar Inf. Commun. (IZS)*. ETH Zurich, 2018, pp. 35–39.

[16] T. Berger, Zhen Zhang, and H. Viswanathan, "The CEO Problem," *IEEE Trans. Inf. Theory*, vol. 42, no. 3, pp. 887–902, 1996.

[17] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, "Information Bottleneck for Gaussian Variables," *J. Mach. Learn. Res.*, vol. 6, pp. 165–188, Dec. 2005.

[18] T. A. Courtade and J. Jiao, "An Extremal Inequality for Long Markov Chains," in *Proc. 52nd Annu. Allerton Conf. Commun., Control Comput.*, Oct. 2014, pp. 763–770.

[19] E. Erkip and T. M. Cover, "The Efficiency of Investment Information," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1026–1040, May 1998.

[20] I. E. Aguerri and A. Zaidi, "Distributed Variational Representation Learning," *IEEE Trans. Pattern Anal.*, vol. 43, no. 1, pp. 120–138, 2021.

[21] P. Gács and J. Körner, "Common information is far less than mutual information," *Probl. Contr. Inform. Theory*, vol. 2, no. 2, pp. 149–162, 1973.

[22] I. Land and J. Huber, "Information Combining," *Found. Trends Commun. Inf. Theory*, vol. 3, no. 3, pp. 227–330, 2006.

[23] Y. Ugur, I. E. Aguerri, and A. Zaidi, "Vector Gaussian CEO Problem Under Logarithmic Loss and Applications," *IEEE Trans. Inf. Theory*, vol. 66, no. 7, pp. 4183–4202, 2020.

[24] M. Vera, L. Rey Vega, and P. Piantanida, "Collaborative Information Bottleneck," *IEEE Trans. Inf. Theory*, vol. 65, no. 2, pp. 787–815, 2019.

[25] Q. Yang, P. Piantanida, and D. Gündüz, "The Multi-layer Information Bottleneck Problem," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Nov. 2017, pp. 404–408.

[26] A. Zaidi, I. E. Aguerri, and S. S. (Shitz), "On the Information Bottleneck Problems: Models, Connections, Applications and Information Theoretic Views," *Entropy*, vol. 22, no. 2, p. 151, 2020.

[27] N. Tishby and N. Zaslavsky, "Deep Learning and the Information Bottleneck Principle," in *Proc. Inf. Theory Workshop (ITW)*, Apr. 2015, pp. 1–5.

[28] A. Alemi, I. Fischer, J. Dillon, and K. Murphy, "Deep Variational Information Bottleneck," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017.

[29] P. Farajiparvar, A. Beirami, and M. Nokleby, "Information Bottleneck Methods for Distributed Learning," in *Proc. 56th Annu. Allerton Conf. Commun., Control Comput.*, 2018, pp. 24–31.

[30] R. A. Amjad and B. C. Geiger, "Learning Representations for Neural Network-Based Classification Using the Information Bottleneck Principle," *IEEE Trans. Pattern Anal.*, vol. 42, no. 9, pp. 2225–2239, 2020.

[31] A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox, "On the information bottleneck theory of deep learning," *J. Stat. Mech. Theory Exp.*, vol. 2019, no. 12, pp. 1–34, Dec. 2019.

[32] Z. Goldfeld and Y. Polyanskiy, "The Information Bottleneck Problem and its Applications in Machine Learning," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 19–38, 2020.

[33] S. Asoodeh and F. P. Calmon, "Bottleneck Problems: An Information and Estimation-Theoretic View," *Entropy*, vol. 22, no. 11, p. 1325, 2020.

[34] P. Harremoes and N. Tishby, "The Information Bottleneck Revisited or How to Choose a Good Distortion Measure," in *Proc. 2007 IEEE Int. Symp. Inf. Theory*, Jun. 2007, pp. 566–570.

[35] I. Sutskover, S. Shamai, and J. Ziv, "Extremes of information combining," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1313–1325, Apr. 2005.

[36] T. Richardson and R. Urbanke, *Modern Coding Theory*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[37] F. G. Mehler, "Ueber die Entwicklung einer Function von beliebig vielen Variablen nach Laplaceschen Functionen höherer Ordnung." *J. Reine Angew. Math.*, vol. 66, pp. 161–176, 1866.

[38] H. O. Lancaster, "The Structure of Bivariate Distributions," *Ann. Math. Statist.*, vol. 29, no. 3, pp. 719–736, 1958.

[39] H. S. Witsenhausen and A. D. Wyner, "A Conditional Entropy Bound for a Pair of Discrete Random Variables," *IEEE Trans. Inf. Theory*, vol. 21, no. 5, pp. 493–501, Sep. 1975.

[40] R. Gilad-Bachrach, A. Navot, and N. Tishby, "An information theoretic tradeoff between complexity and accuracy," in *Proc. 16th Conf. Comput. Theory (COLT)*, ser. Lecture Notes in Computer Science, B. Schölkopf and M. K. Warmuth, Eds., vol. 2777. Springer, 2003, pp. 595–609.

[41] R. O'Donnell, *Analysis of Boolean Functions*, 1st ed. New York, NY, USA: Cambridge Univ. Press, Jun. 2014.

[42] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2006.