entropy

MDPI

Article

The Double-Sided Information Bottleneck Function †

Michael Dikshtein [1,*], Or Ordentlich [2] and Shlomo Shamai (Shitz) [1]

[1] Department of Electrical and Computer Engineering, Technion, Haifa 3200003, Israel
[2] School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem 9190401, Israel
* Correspondence: michaeldic@campus.technion.ac.il

† This paper is an extended version of our paper published in 2021 IEEE International Symposium on Information Theory.

Abstract: A double-sided variant of the information bottleneck method is considered. Let (X,Y) be a bivariate source characterized by a joint pmf P_{XY}. 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 correlations. 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. Furthermore, we present a Blahut–Arimoto type alternating maximization algorithm and demonstrate its performance for a representative setting. This problem is closely related to the domain of biclustering.

Keywords: information bottleneck; lossy compression; remote source coding; biclustering

check for updates

Citation: Dikshtein, M.; Ordentlich, O.; Shamai, S. The Double-Sided Information Bottleneck Function. *Entropy* **2022**, *24*, 1321. https://doi.org/10.3390/e24091321

Academic Editors: Gerhard Bauch and Jan Lewandowsky

Received: 25 July 2022
Accepted: 13 September 2022
Published: 19 September 2022

1. Introduction

The *information bottleneck* (IB) method [1] plays a central role in advanced lossy source compression. The analysis of classical source coding algorithms is mainly approached via the rate-distortion theory, where a fidelity measure must be defined. However, specifying an appropriate distortion measure in many real-world applications is challenging and sometimes infeasible. The IB framework introduces an essentially different concept, where another variable is provided, which carries the relevant information in the data to be compressed. The quality of the reconstructed sequence is measured via the mutual information metric between the reconstructed data and the relevance variables. Thus, the IB method provides a universal fidelity measure.

In this work, we extend and generalize the IB method by imposing an additional bottleneck constraint on the relevant variable and considering noisy observation of the source. In particular, 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 [2]:

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

where the maximization is over all $\mathsf{P}_{\mathsf{U}|\mathsf{X}}$ and $\mathsf{P}_{\mathsf{V}|\mathsf{Y}}$ satisfying $I(\mathsf{U};\mathsf{X}) \leq C_u$ and $I(\mathsf{V};\mathsf{Y}) \leq C_v$. 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 (1) for various fixed sources P_{XY} and constraints C_u and C_v.

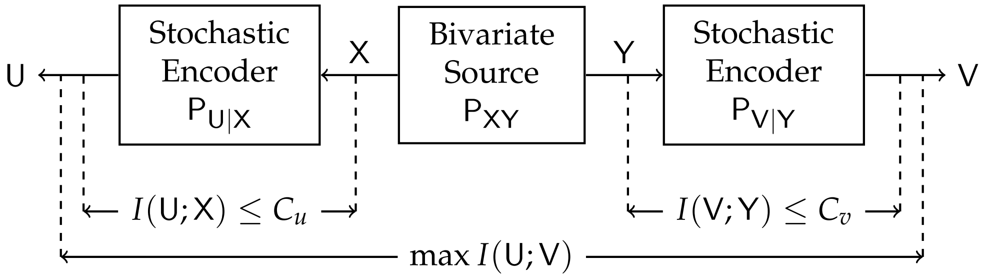

U
Stochastic Encoder $P_{U|X}$
X
Bivariate Source P_{XY}
Y
Stochastic Encoder $P_{V|Y}$
V
$I(U;X) \leq C_u$
$I(V;Y) \leq C_v$
$\max I(U;V)$

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

The problem we consider originates from the domain of clustering. Clustering is applied to organize similar entities in unsupervised learning [3]. It has numerous practical applications in data science, such as: joint word-document clustering, gene expression [4], and pattern recognition. The data in those applications are arranged as a contingency table. Usually, clustering is performed on one dimension of the table, but sometimes it is helpful to apply clustering on both dimensions of the contingency table [5], 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. The input and output of a typical biclustering algorithm are illustrated in Figure 2. Consider an $S \times T$ data matrix (a_{st}). Find partitions $\mathcal{B}_k \subseteq \{1, \ldots, S\}$ and $\mathcal{C}_l \subseteq \{1, \ldots, T\}$, $k = 1, \ldots, K$, $l = 1, \ldots, L$ such that all elements of the "biclusters" [6] $(a_{st})_{s \in \mathcal{B}_k, t \in \mathcal{C}_l}$ are homogeneous. The measure of homogeneity depends on the application.

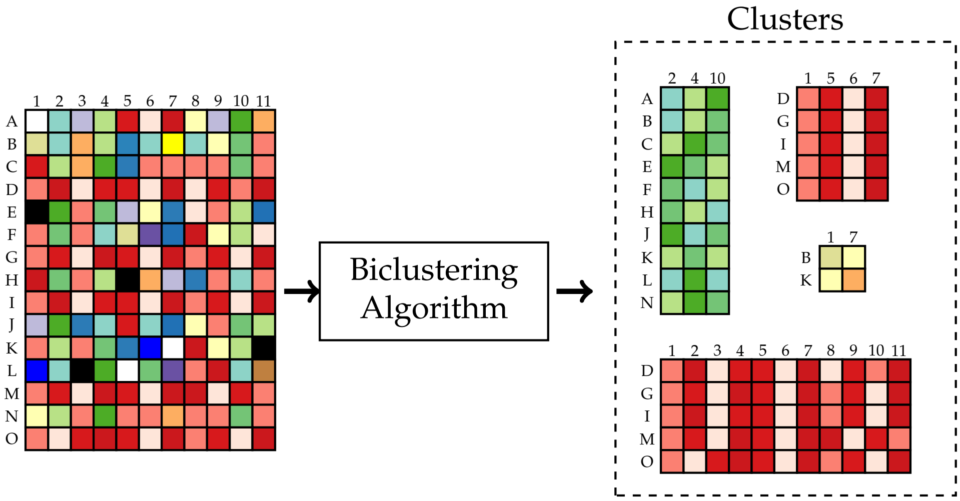
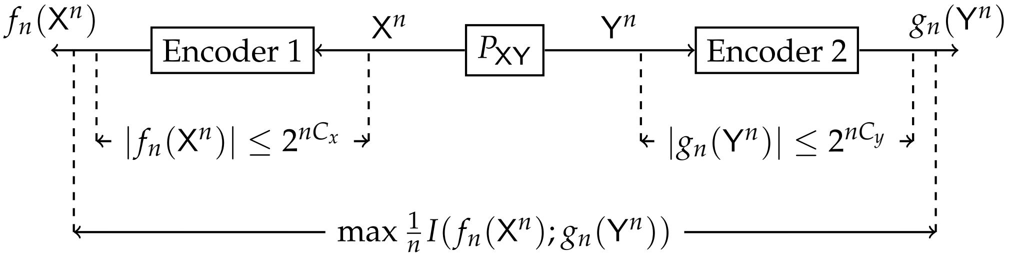


Figure 2. Illustration of a typical biclustering algorithm.

This problem can also be motivated by a remote source coding setting. Consider a latent random variable W, which satisfies U ← X ← W → Y → V and represents a source of information. We have two users that observe noisy versions of W, i.e., X and Y. Those users try to compress the observed noisy data so that their reconstructed versions, U and V, will be comparable under the maximum mutual information metric. The problem we consider also bears practical applications. Imagine a distributed sensor network where the different edges measure a noisy version of a particular signal but are not allowed to communicate with each other. Each of the nodes performs compression of the received signal. Under the DSIB framework, we can find the optimal compression schemes that preserve the reconstructed symbols' proximity subject to the mutual information measure.

Dhillon et al. [7] initiated an information-theoretic approach to biclustering. They have regarded the normalized non-negative 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 proper information-theoretic perspective was first explicitly considered by Pichler et al. [2]. Consider the model illustrated in Figure 3. A bivariate memory-less source with joint law P_{XY} 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, $f_n(\mathsf{X}^n)$ and $g_n(\mathsf{Y}^n)$. The objective is to construct the mappings f_n and g_n 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) was given, and several converse results were established. Furthermore, connections were made to the standard IB [1] and the *multiple description* CEO problems [8]. In addition, the equivalence of information-theoretic biclustering problem to hypothesis testing against independence with multiterminal data compression and a pattern recognition problem was established in [9,10], respectively.



$f_n(\mathsf{X}^n)$
Encoder 1
X^n
P_{XY}
Y^n
Encoder 2
$g_n(\mathsf{Y}^n)$
$|f_n(\mathsf{X}^n)| \leq 2^{nC_x}$
$|g_n(\mathsf{Y}^n)| \leq 2^{nC_y}$
$\max \frac{1}{n} I(f_n(\mathsf{X}^n); g_n(\mathsf{Y}^n))$

Figure 3. Block diagram of the information-theoretic biclustering problem.

The DSIB problem addressed in our paper is, in fact, a single-letter version of the *distributed clustering* setup [2]. The inner bound in [2] coincides with our problem definition. Moreover, if the Markov condition $\mathsf{U} \to \mathsf{X} \to \mathsf{Y} \to \mathsf{Z}$ is imposed on the multi-letter variant, then those problems are equivalent. A similar setting, but 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 more straightforward to solve than its standard, single-sided, clustering counterpart. For example, the Courtade–Kumar conjecture [13] for the standard single-sided clustering setting was ultimately proven for the biclustering setting [14]. A particular case, where (X, Y) are drawn from DSBS distribution and the mappings f_n and g_n are restricted to be Boolean functions, was addressed in [14]. The bound $I(f_n(\mathsf{X}^n); g_n(\mathsf{Y}^n)) \leq I(\mathsf{X}; \mathsf{Y})$ was established, which is tight if and only if f_n and g_n are dictator functions.

1.1. Related Work

Our work extends the celebrated standard (single-sided) IB (SSIB) method introduced by Tishby et al. [1]. Indeed, consider the problem illustrated in Figure 4. This single-sided counterpart of our work is essentially a remote source coding problem [15–17], choosing the distortion measure as the logarithmic loss. The random variable U represents the noisy version (X) of the source (Y) with a constrained number of bits ($I(\mathsf{U}; \mathsf{X}) \leq C$), and the goal is to maximize the relevant information in U regarding Y (measured by the mutual information between Y and U). In the standard IB setup, $I(\mathsf{U}; \mathsf{X})$ is referred to as the complexity of U, and $I(\mathsf{Y}; \mathsf{U})$ is referred to as the relevance of U.

For the particular case where $(\mathsf{U}, \mathsf{X}, \mathsf{Y})$ are discrete, an optimal $\mathsf{P}_{\mathsf{U}|\mathsf{X}}$ can be found by iteratively solving a set of self-consistent equations. A generalized Blahut–Arimoto algorithm [18–21] was proposed to solve those equations. The optimal test-channel $\mathsf{P}_{\mathsf{U}|\mathsf{X}}$ was characterized using a variation principle in [1]. A particular case of deterministic mappings from X to U was considered in [22], and algorithms that find those mappings were described.

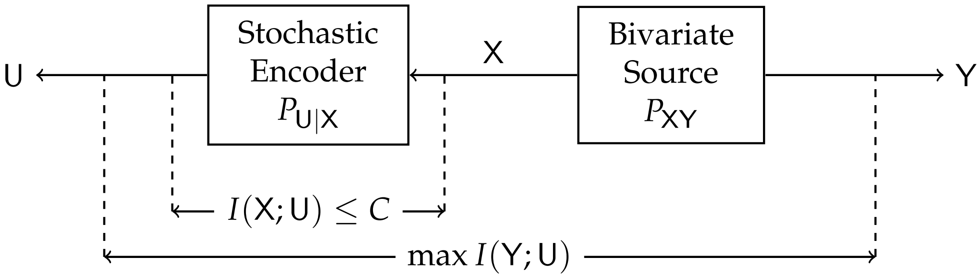

Stochastic Encoder
$P_{\mathsf{U}|\mathsf{X}}$
Bivariate Source
P_{XY}
U
X
Y
$I(\mathsf{X};\mathsf{U}) \leq C$
$\max I(\mathsf{Y};\mathsf{U})$

Figure 4. Block diagram of the Single-Sided Information Bottleneck function.

Several representing scenarios have been considered for the SSIB problem. The setting where the pair (X,Y) is a *doubly symmetric binary source* (DSBS) with transition probability p was addressed from various perspectives in [17,23,24]. Utilizing Mrs. Gerber's Lemma (MGL) [25], one can show that the optimal test-channel for the DSBS setting is a BSC. The case where $(\mathbf{X},\mathbf{Y})$ are jointly multivariate Gaussians in the SSIB framework was first considered in [26]. It was shown that the optimal distribution of $(\mathbf{U},\mathbf{X},\mathbf{Y})$ is also jointly Gaussian. The optimality of the Gaussian test channel can be proven using EPI [27], or exploiting I-MMSE and Single Crossing Property [28]. Moreover, the proof can be easily extended to jointly Gaussian random vectors $(\mathbf{X},\mathbf{Y})$ under the I-MMSE framework [29].

In a more general scenario where $\mathsf{X} = \mathsf{Y} + \mathsf{Z}$ and only Z is fixed to be Gaussian, it was shown that discrete signaling with deterministic quantizers as test-channel sometimes outperforms Gaussian P_{X} [30]. This exciting observation leads to a conjecture that discrete inputs are optimal for this general setting and may have a connection to the input amplitude constrained AWGN channels where it was already established that discrete input distributions are optimal [31–33]. One reason for the optimality of discrete distributions stems from the observation that constraining the compression rate limits the usable input amplitude. However, as far as we know, it remains an open problem.

There are various related problems considered in the literature that are equivalent to the SSIB; namely, they share a similar single-letter optimization problem. In the *conditional entropy bound* (CEB) function, studied in [17], given a fixed bivariate source (X,Y) and an equality constraint on the conditional entropy of X given U, the goal is to minimize the conditional entropy of Y given U over the set of U such that $\mathsf{U} \to \mathsf{X} \to \mathsf{Y}$ constitute a Markov chain. One can show that CEB is equivalent to SSIB. The *common reconstruction* CR setting [34] is a source coding with a side-information problem, also known as Wyner–Ziv coding, as depicted in Figure 5; with an additional constraint, the encoder can reconstruct the same sequence as the decoder. Additional assumption of log-loss fidelity results in a single-letter rate-distortion region equivalent to the SSIB. In the problem of *information combining* (IC) [23,35], motivated by message combining in LDPC decoders, a source of information, P_{Y}, is observed through two test-channels $\mathsf{P}_{\mathsf{X}|\mathsf{Y}}$ and $\mathsf{P}_{\mathsf{Z}|\mathsf{Y}}$. The IC framework aims to design those channels in two extreme approaches. For the first, IC asks what those channels should be to make the output pair (X,Z) maximally informative regarding Y. On the contrary, IC also considers how to design $\mathsf{P}_{\mathsf{X}|\mathsf{Y}}$ and $\mathsf{P}_{\mathsf{Z}|\mathsf{Y}}$ to minimize the information in (X,Z) regarding Y. The problem of minimizing IC can be shown to be equivalent to the SSIB. In fact, if (X,Y) is a DSBS, then by [23], $\mathsf{P}_{\mathsf{Z}|\mathsf{Y}}$ is a *binary symmetric channel* (BSC), recovering similar results from (Section IV.A of [17]).

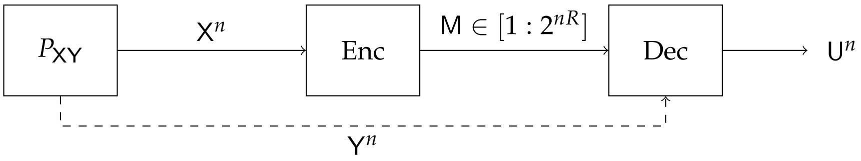


Figure 5. Block diagram of Source Coding with Side Information.

The IB method has been extended to various network topologies. A multilayer extension of the IB method is depicted in Figure 6. This model was first considered in [36]. A multivariate source $(\mathsf{X}, \mathsf{Y}_1, \ldots, \mathsf{Y}_L)$ generates a sequence of n i.i.d. copies $(\mathsf{X}^n, \mathsf{Y}_1^n, \ldots, \mathsf{Y}_L^n)$. The receiver has access only to the sequence X^n while $(\mathsf{Y}_1^n, \ldots, \mathsf{Y}_L^n)$ are hidden. The decoder performs a consecutive L-stage compression of the observed sequence. The representation at step k must be maximally informative about the respective hidden sequence Y_k, $k \in \{1, 2, \ldots, L\}$. This setup is highly motivated by the structure of deep neural networks. Specific results were established for the binary and Gaussian sources.

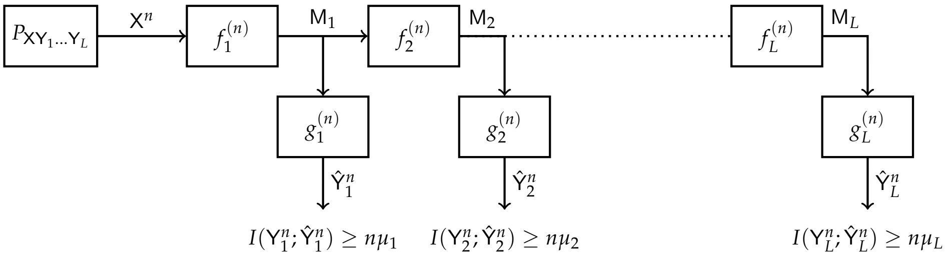

$P_{\mathsf{XY}_1\ldots\mathsf{Y}_L}$
X^n
$f_1^{(n)}$
M_1
$f_2^{(n)}$
M_2
$f_L^{(n)}$
M_L
$g_1^{(n)}$
$g_2^{(n)}$
$g_L^{(n)}$
$\hat{\mathsf{Y}}_1^n$
$\hat{\mathsf{Y}}_2^n$
$\hat{\mathsf{Y}}_L^n$
$I(\mathsf{Y}_1^n; \hat{\mathsf{Y}}_1^n) \geq n\mu_1$
$I(\mathsf{Y}_2^n; \hat{\mathsf{Y}}_2^n) \geq n\mu_2$
$I(\mathsf{Y}_L^n; \hat{\mathsf{Y}}_L^n) \geq n\mu_L$

Figure 6. Block diagram of the Multi-Layer IB.

The model depicted in Figure 7 represents a multiterminal extension of the standard IB. A set of receivers observe noisy versions $(\mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_K)$ of some source of information Y. The channel outputs $(\mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_K)$ are conditionally independent given Y. The receivers are connected to the central processing unit through noiseless but limited-capacity backhaul links. The central processor aims to attain a good prediction $\hat{\mathsf{Y}}$ of the source Y based on compressed representations of the noisy version of Y obtained from the receivers. The quality of prediction is measured via the mutual information merit between Y and $\hat{\mathsf{Y}}$. The Distributive IB setting is essentially a CEO source coding problem under logarithmic loss (log-loss) distortion measure [37]. The case where $(\mathsf{X}, \mathsf{Y}_1, \ldots, \mathsf{Y}_K)$ are jointly Gaussian random variables was addressed in [20], and a Blahut–Arimoto-type algorithm was proposed. An optimized algorithm to design quantizers was proposed in [38].

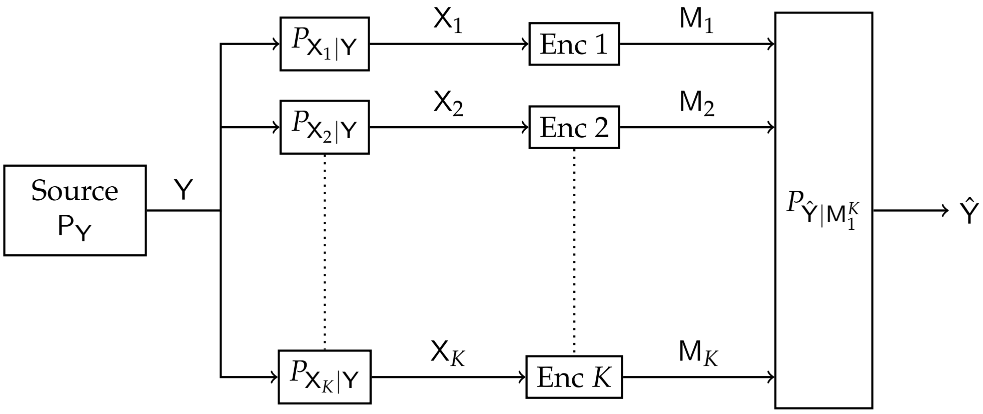


Figure 7. Block diagram of the Distributive IB.

A cooperative multiterminal extension of the IB method was proposed in [39]. Let $(\mathsf{X}_1^n, \mathsf{X}_2^n, \mathsf{Y}^n)$ be n i.i.d. copies of the multivariate source $(\mathsf{X}_1, \mathsf{X}_2, \mathsf{Y})$. The sequences X_1^n and X_2^n are observed at encoders 1 and 2, respectively. Each encoder sends a representation of the observed sequence through a noiseless yet rate-limited link to the other encoder and the mutual decoder. The decoder attempts to reconstruct the latent representation sequence Y^n based on the received descriptions. As shown in Figure 8, this setup differs from the CEO setup [40] since the encoders can cooperate during the transmission. The set of all feasible rates of complexity and relevance were characterized, and specific regions for the binary and Gaussian sources were established. There are many additional variations of multi-user IB in the literature [20,26,35–37,39–44].

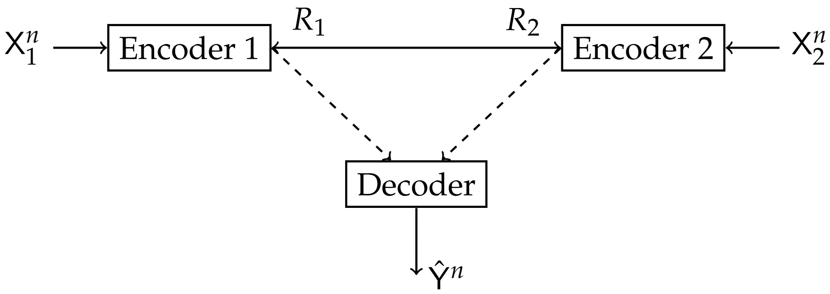


Figure 8. Block diagram of the Collaborative IB.

The IB problem connects to many timely aspects, such as *capital investment* [43], *distributed learning* [45], *deep learning* [46–52], and *convolutional neural networks* [53,54]. Moreover, it has been recently shown that the IB method can be used to reduce the data transfer rate and computational complexity in 5G LDPC decoders [55,56]. The IB method has also been connected with constructing good polar codes [57]. Due to the exponential output-alphabet growth of polarized channels, it becomes demanding to compute their capacities to identify the location of “frozen bits". Quantization is employed in order to reduce the computation complexity. The quality of the quantization scheme is assessed via mutual information preservation. It can be shown that the corresponding IB problem upper bounds the quantization technique. Quantization algorithms based upon the IB method were considered in [58–60]. Furthermore, a relationship between the KL means algorithm and the IB method has been discovered in [61].

A recent comprehensive tutorial on the IB method and related problems is given in [24]. Applications of IB problem in *machine learning* are detailed in [26,45–47,51,52,62].

1.2. Notations

Throughout the paper, random variables are denoted using a sans-serif font, e.g., X, their realizations are denoted by the respective lower-case letters, e.g., x, and their alphabets are denoted by the respective calligraphic letters, e.g., $\mathcal{X}$. Let $\mathcal{X}^n$ stand for the set of all n-tuples of elements from $\mathcal{X}$. An element from $\mathcal{X}^n$ is denoted by $x^n = (x_1, x_2, \ldots, x_n)$ and substrings are denoted by $x_i^j = (x_i, x_{i+1}, \ldots, x_j)$. The cardinality of a finite set, say $\mathcal{X}$, is denoted by $|\mathcal{X}|$. The probability mass function (pmf) of X, the joint pmf of X and Y, and the conditional pmf of X given Y are denoted by P_{X}, P_{XY}, and $\mathsf{P}_{\mathsf{X}|\mathsf{Y}}$, respectively. The expectation of X is denoted by $\mathbb{E}[\mathsf{X}]$. The probability of an event $\mathcal{E}$ is denoted as $\mathsf{P}(\mathcal{E})$.

Let X and Y be an n-ary and m-ary random variables, respectively. The marginal probability vector is denoted by a lowercase boldface letter, i.e.,

$$\mathbf{q} \triangleq (\mathsf{P}(\mathsf{X}=1), \mathsf{P}(\mathsf{X}=2), \ldots, \mathsf{P}(\mathsf{X}=n))^T. \tag{2}$$

The probability vector of an n-ary uniform random variable is denoted by $\mathbf{u}_n$. We denote by T the transition matrix from X to Y, i.e.,

$$T_{ij} \triangleq \mathsf{P}(\mathsf{Y}=i|\mathsf{X}=j), \qquad 1 \le i \le m, 1 \le j \le n. \tag{3}$$

The entropy of n-ary probability vector $\mathbf{q}$ is given by $h(\mathbf{q})$, where

$$h(\mathbf{q}) \triangleq -\sum_{i=1}^{n} q_i \log q_i. \tag{4}$$

Throughout this paper all logarithms are taken to base 2 unless stated otherwise. We denote the ones complemented with a bar, i.e., $\bar{x} = 1 - x$. The binary convolution of $x, y \in [0,1]$ is defined as $x * y \triangleq x\bar{y} + \bar{x}y$. The binary entropy function is defined by $h_b(p) : [0,1] \to [0,1]$, i.e., $h_b(p) \triangleq -p \log p - \bar{p} \log \bar{p}$, and $h_b^{-1}(\cdot)$ its inverse, restricted to $[0,1/2]$.

Let X and Y be a pair of random variables with joint pmf P_{XY} and marginal pmfs $\mathsf{P}_{\mathsf{X}} = \mathbf{q}_x$ and $\mathsf{P}_{\mathsf{Y}} = \mathbf{q}_y$. Furthermore, let T ($\bar{T}$) be the transition matrix from X (Y) to Y (X). The mutual information between X and Y is defined as:

$$I(\mathsf{X};\mathsf{Y}) = I(\mathbf{q}_{\mathsf{x}}, T) = I(\mathbf{q}_{\mathsf{y}}, \bar{T}) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \mathsf{P}_{\mathsf{XY}}(x,y) \log \frac{\mathsf{P}_{\mathsf{XY}}(x,y)}{\mathsf{P}_{\mathsf{X}}(x)\mathsf{P}_{\mathsf{Y}}(y)}. \tag{5}$$

1.3. Paper Outline

Section 2 gives a proper definition of the DSIB optimization problem, mentions various results directly related to this work, and provides some general preliminary results. The spotlight of Section 3 is on the binary (X,Y), where we derive bounds on the respective DSIB function and show a complete characterization for extreme scenarios. The jointly Gaussian (X,Y) is considered in Section 4, where an elegant representation of an objective function is presented, and complete characterization in the low-SNR regime is established. A Blahut–Arimoto-type alternating maximization algorithm will be presented in Section 5. Representative numerical evaluation of the bounds and the proposed algorithm will be provided in Section 6. Finally, a summary and possible future directions will be described in Section 7. The prolonged proofs are postponed to the Appendix A.

2. Problem Formulation and Basic Properties

The DSIB function is a multi-terminal extension of the standard IB [1]. First, we briefly remind the latter's definition and give related results that will be utilized for its double-sided counterpart. Then, we provide a proper definition of the DSIB optimization problem and present some general preliminaries.

2.1. The Single-Sided Information Bottleneck (SSIB) Function

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

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

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

$$\begin{aligned} \hat{R}_T(\mathbf{q}, C) \triangleq \quad & \underset{\mathsf{P}_{\mathsf{U}|\mathsf{X}}}{\mathit{maximize}} \quad I(\mathsf{U};\mathsf{V}) \\ & \mathit{subject\ to} \quad I(\mathsf{X};\mathsf{U}) \le C. \end{aligned} \tag{6}$$

Remark 1. *The SSIB problem defined in* (6) *is equivalent (has similar solution) to the CEB problem considered in [17].*

Although the optimization problem in (6) is well defined, the auxiliary random variable U may have an unbounded alphabet. The following lemma provides an upper bound on the cardinality of $\mathcal{U}$, thus relaxing the optimization domain.

Lemma 1 (Lemma 2.2 of [17])**.** *The optimization over U in* (6) *can be restricted to $|\mathcal{U}| \le n + 1$.*

Remark 2. *A tighter bound, namely $|\mathcal{U}| \le n$, was previously proved in [63] for the corresponding dual problem, namely, the IB Lagrangian. However, since $\hat{R}_T(\mathbf{q}, C)$ is generally not a strictly convex function of C, it cannot be directly applied for the primal problem* (6)*.*

Note that the SSIB optimization problem (6) is basically a convex function maximization over a convex set; thus, the maximum is attained on the boundary of the set.

Lemma 2 (Theorem 2.5 of [17])**.** *The inequality constraint in* (6) *can be replaced by equality constraint, i.e., $I(\mathsf{X};\mathsf{U}) = C$.*

2.2. The Double-Sided Information Bottleneck (DSIB) Function

Definition 2 (DSIB)**.** *Let* (X,Y) *be a pair of random variables with* $|\mathcal{X}|=n$*,* $|\mathcal{Y}|=m$ *and fixed* 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:*

$$R_{\mathsf{P}_{\mathsf{XY}}}(C_u,C_v)\triangleq\begin{array}{ll}\underset{\mathsf{P}_{\mathsf{U}|\mathsf{X}},\mathsf{P}_{\mathsf{V}|\mathsf{Y}}}{\textit{maximize}} & I(\mathsf{U};\mathsf{V})\\ \textit{subject to} & I(\mathsf{X};\mathsf{U})\leq C_u \textit{ and } I(\mathsf{Y};\mathsf{V})\leq C_v.\end{array}\tag{7}$$

The achieving conditional distributions $\mathsf{P}_{\mathsf{U}|\mathsf{X}}$ *and* $\mathsf{P}_{\mathsf{V}|\mathsf{Y}}$ *will be termed as the optimal test-channels. Occasionally, we will drop the subscript denoting the particular choice of the bivariate source* P_{XY}*.*

Note that (7) can be expressed in the following equivalent form:

$$R(C_u,C_v)\triangleq\begin{array}{lll}\underset{\mathsf{P}_{\mathsf{V}|\mathsf{Y}}}{\text{maximize}} & \underset{\mathsf{P}_{\mathsf{U}|\mathsf{X}}}{\text{maximize}} & I(\mathsf{U};\mathsf{V}).\\ \text{subject to} & \text{subject to} & \\ I(\mathsf{Y};\mathsf{V})\leq C_v & I(\mathsf{X};\mathsf{U})\leq C_u & \end{array}\tag{8}$$

Evidently, we can define (8) using (6). Indeed, fix $\mathsf{P}_{\mathsf{V}|\mathsf{Y}}$ so that it satisfies $I(\mathsf{Y};\mathsf{V})\leq C_v$. Denote by $T_{\mathsf{V}|\mathsf{Y}}$ the transition matrix from Y to V and by $T_{\mathsf{Y}|\mathsf{X}}$ the transition matrix from X to Y, respectively, i.e.,

$$\begin{aligned}(T_{\mathsf{V}|\mathsf{Y}})_{ik}&\triangleq\mathsf{P}(\mathsf{V}=i|\mathsf{Y}=k), && 1\leq i\leq|\mathcal{V}|, 1\leq k\leq m,\\ (T_{\mathsf{Y}|\mathsf{X}})_{kj}&\triangleq\mathsf{P}(\mathsf{Y}=k|\mathsf{X}=j), && 1\leq k\leq m, 1\leq j\leq n.\end{aligned}$$

Denote by $\mathbf{q}_{\mathsf{X}}$ and $\mathbf{q}_{\mathsf{Y}}$ the marginal probability vectors of X and Y, respectively, and consider the inner maximization term in (8). Since $\mathsf{P}_{\mathsf{V}|\mathsf{Y}}$ and P_{XY} are fixed, then $\mathsf{P}_{\mathsf{XV}}=\sum_y\mathsf{P}_{\mathsf{V}|\mathsf{Y}}(\cdot|y)\mathsf{P}_{\mathsf{XY}}(\cdot,y)$ is also fixed. Denote by $T_{\mathsf{V}|\mathsf{X}}\triangleq T_{\mathsf{V}|\mathsf{Y}}T_{\mathsf{Y}|\mathsf{X}}$ the transition matrix from X to V. Therefore, the inner maximization term in (8) is just the SSIB function with parameters $T_{\mathsf{V}|\mathsf{X}}$ and C_u, namely, $\hat{R}_{T_{\mathsf{V}|\mathsf{X}}}(\mathbf{q}_{\mathsf{X}},C_u)$. Hence, our problem can also be interpreted in the following two equivalent ways:

$$R(C_u,C_v)\triangleq\begin{array}{ll}\underset{T_{\mathsf{V}|\mathsf{Y}}}{\text{maximize}} & \hat{R}_{T_{\mathsf{V}|\mathsf{Y}}T_{\mathsf{Y}|\mathsf{X}}}(\mathbf{q}_{\mathsf{X}},C_u)\\ \text{subject to} & I(\mathbf{q}_{\mathsf{Y}},T_{\mathsf{V}|\mathsf{Y}})\leq C_v;\end{array}\tag{9}$$

or, similarly, by interchanging the order of maximization in (8), it can be expressed as follows:

$$R(C_u,C_v)\triangleq\begin{array}{ll}\underset{T_{\mathsf{U}|\mathsf{X}}}{\text{maximize}} & \hat{R}_{T_{\mathsf{U}|\mathsf{X}}T_{\mathsf{X}|\mathsf{Y}}}(\mathbf{q}_{\mathsf{Y}},C_v)\\ \text{subject to} & I(\mathbf{q}_{\mathsf{X}},T_{\mathsf{U}|\mathsf{X}})\leq C_u,\end{array}\tag{10}$$

where $T_{\mathsf{U}|\mathsf{X}}$ is the transition matrix from X to U, and $T_{\mathsf{X}|\mathsf{Y}}$ is the transition matrix from Y to X. This representation gives us a different perspective on our problem as an optimal compressed representation of the relevance random variable for the IB framework.

Remark 3. *Taking* $C_v=\infty$ *in* (9) *results in an deterministic channel from* Y *to* V*, i.e.,* $\mathsf{V}=\mathsf{Y}$*. Thus, the DSIB problem defined in* (7) *reduces to the SSIB problem* (6)*.*

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

Proposition 1. *For the DSIB optimization problem defined in* (7)*, it suffices to consider random variables* U *and* V *with cardinalities* $|\mathcal{U}| \leq n+1$ *and* $|\mathcal{V}| \leq m+1$*.*

Proof. Let $T_{\mathsf{U}|\mathsf{X}}$ and $T_{\mathsf{V}|\mathsf{Y}}$ be two arbitrary transition matrices. By Lemma 1, there exists $T_{\tilde{\mathsf{U}}|\mathsf{X}}$ with $|\tilde{\mathcal{U}}| \leq n+1$ such that $I(\tilde{\mathsf{U}};\mathsf{V}) \geq I(\mathsf{U};\mathsf{V})$ and $I(\mathsf{X};\tilde{\mathsf{U}}) \leq C_u$. Similarly, $T_{\mathsf{V}|\mathsf{Y}}$ can be replaced with $T_{\tilde{\mathsf{V}}|\mathsf{Y}}$, $|\tilde{\mathcal{V}}| \leq m+1$ such that $I(\tilde{\mathsf{U}};\tilde{\mathsf{V}}) \geq I(\tilde{\mathsf{U}},\mathsf{V}) \geq I(\mathsf{U};\mathsf{V})$, and $I(\mathsf{Y};\tilde{\mathsf{V}}) \leq C_v$. Therefore, there exists an optimal solution with $|\mathcal{U}| \leq n+1$ and $|\mathcal{V}| \leq m+1$. □

In the following two sections, we will present the primary analytical outcomes of our study. First, we consider the scenario where our bivariate source is binary, specifically DSBS. Then, we handle the case where X and Y are jointly Gaussian.

3. Binary (X,Y)

Let (X,Y) be a DSBS with parameter p, i.e.,

$$\mathsf{P}_{\mathsf{XY}}(x,y) = \frac{1}{2}(p \cdot \mathbb{1}(x \neq y) + (1-p)\mathbb{1}(x = y)). \tag{11}$$

We entitle the respective optimization problem (7) as the *binary double-sided information bottleneck* (BDSIB) and emphasize its dependence on the parameter p as $R(C_u, C_v, p)$.

The following proposition states that the cardinality bound from Lemma 1 can be tightened in the binary case.

Proposition 2. *Considering the optimization problem in* (6) *with* $\mathsf{X} = \mathrm{Ber}(q)$ *and* $|\mathcal{Y}| = 3$*, binary* U *is optimal.*

The proof of this proposition is postponed to Appendix A. Using similar justification for Proposition 1 combined with Proposition 2, we have the following strict cardinality formula for the BDSIB setting.

Proposition 3. *For the respective DSBS setting of* (7)*, it suffices to consider random variables* U *and* V *with cardinalities* $|\mathcal{U}| = |\mathcal{V}| = 2$*.*

Note that the above statement is not required for the results in the rest of this section to hold and will be mainly applied to justify our conjectures via numerical simulations.

We next show that the specific objective function for the binary setting of (7), i.e, the mutual information between U and V, has an elegant representation which will be useful in deriving lower and upper bounds.

Lemma 3. *The mutual information between* U *and* V *can be expressed as follows:*

$$I(\mathsf{U};\mathsf{V}) = \mathbb{E}_{\mathsf{P}_\mathsf{U} \times \mathsf{P}_\mathsf{V}}[K(\mathsf{U},\mathsf{V},p)\log K(\mathsf{U},\mathsf{V},p)], \tag{12}$$

where the expectation is taken over the product measure $\mathsf{P}_\mathsf{U} \times \mathsf{P}_\mathsf{V}$*,* U *and* V *are binary random variables satisfying:*

$$\mathsf{P}(\mathsf{U}=0) = \frac{\alpha_1 - \frac{1}{2}}{\alpha_1 - \alpha_0}, \qquad \mathsf{P}(\mathsf{V}=0) = \frac{\beta_1 - \frac{1}{2}}{\beta_1 - \beta_0}, \tag{13}$$

the kernel $K(u,v,p)$ *is given by:*

$$K(u,v,p) = 2\alpha_u * \beta_v * p = 1 - (1-2p)(1-2\alpha_u)(1-2\beta_v), \tag{14}$$

and the reverse test-channels are defined by: $\alpha_u \triangleq \mathsf{P}(\mathsf{X}=1|\mathsf{U}=u)$*,* $\beta_v \triangleq \mathsf{P}(\mathsf{Y}=0|\mathsf{V}=v)$*. Furthermore, since* $|(1-2p)(1-2\alpha_u)(1-2\beta_v)|<1$*, utilizing Taylor's expansion of* $\log(1-x)$*, we obtain:*

$$I(\mathsf{U};\mathsf{V}) = \sum_{n=2}^{\infty} \frac{(1-2p)^n \mathbb{E}[(1-2\alpha_{\mathsf{U}})^n]\mathbb{E}[(1-2\beta_{\mathsf{V}})^n]}{n(n-1)}. \tag{15}$$

The general cascade of test-channels and the DSBS, defined by $\{\alpha_u\}_{u=0}^1$, $\{\beta_v\}_{v=0}^1$ and p, is illustrated in Figure 9. The proof of Lemma 3 is postponed to Appendix B.

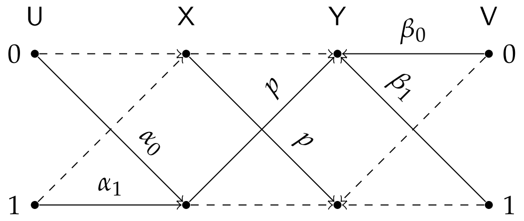


Figure 9. General test-channel construction of the BDSIB function.

We next examine some corner cases for which $R(C_u, C_v, p)$ is fully characterized.

3.1. Special Cases

A particular case where we have a complete analytical solution is when p tends to 1/2.

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

$$R(C_u, C_v, \epsilon) = 2\epsilon^2 \log e \cdot (1-2h_b^{-1}(1-C_u))^2(1-2h_b^{-1}(1-C_v))^2 + o(\epsilon^2), \tag{16}$$

and it is achieved by taking $\mathsf{P}_{\mathsf{U}|\mathsf{X}}$ *and* $\mathsf{P}_{\mathsf{V}|\mathsf{Y}}$ *as BSC test-channels satisfying the constraints with equality.*

This theorem follows by considering the low SNR regime in Lemma 3 and is proved in Appendix D. For the lower bound we take $\mathsf{P}_{\mathsf{U}|\mathsf{X}}$ and $\mathsf{P}_{\mathsf{V}|\mathsf{Y}}$ to be BSCs.

In Section 6 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 optimal test-channels* $\mathsf{P}_{\mathsf{U}|\mathsf{X}}$ *and* $\mathsf{P}_{\mathsf{V}|\mathsf{X}}$ *that achieve* $R(C_u, C_v, 0)$ *are Z-channel and S-channel respectively.*

Remark 4. *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 5. *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 [23,35]. 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 6. *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 (Chapter 4 of [64]), then the maximal $I(\mathsf{U};\mathsf{V})$ is attained by BSC channels, as those are the symmetric channels minimizing $I(\mathsf{X};\mathsf{U},\mathsf{V})$ [23]. It is not surprising that once the BMS constraint is removed, symmetric channels are no longer optimal (see the discussion in (Section VI.C of [23])).*

Consider now the case $\mathsf{X} = \mathsf{Y}$ ($p = 0$) with an additional symmetry assumption $C_u = C_v$. The most reasonable apriori guess is that the optimal test-channels $\mathsf{P}_{\mathsf{U}|\mathsf{X}}$ and $\mathsf{P}_{\mathsf{V}|\mathsf{X}}$ are the same up to some permutation of inputs and outputs. Surprisingly, this is not the case, unless they are BSC or Z channels, as the following negative result states.

Proposition 4. *Suppose $C_u = C_v$ and the transition matrix from X to V, given by*

$$T_{\mathsf{V}|\mathsf{X}} = \begin{pmatrix} a & b \\ 1-a & 1-b \end{pmatrix}, \tag{17}$$

satisfies $I(\mathbf{u}_2, T_{\mathsf{V}|\mathsf{X}}) = C_v$. Consider the respective SSIB optimization problem

$$\hat{R}_{T_{\mathsf{V}|\mathsf{X}}}(\mathbf{u}_2, C_u) = \max_{\mathsf{P}_{\mathsf{U}|\mathsf{X}}\colon I(\mathsf{U};\mathsf{X}) \le C_u} I(\mathsf{U};\mathsf{V}). \tag{18}$$

The optimal $\mathsf{P}_{\mathsf{U}|\mathsf{X}}$ that attains (6) with $\mathbf{q}_{\mathsf{X}} = \mathbf{u}_2$ and $C = C_u$ does not equal to $\mathsf{P}_{\mathsf{V}|\mathsf{X}}$ or any permutation of $\mathsf{P}_{\mathsf{V}|\mathsf{X}}$, unless $\mathsf{P}_{\mathsf{V}|\mathsf{X}}$ is a BSC or a Z channel.

The proof is based on [17] and is postponed to Appendix E.

As for the case of $\mathsf{X} \neq \mathsf{Y}$, i.e., $p \neq 0$, we have the following conjecture.

Conjecture 3. *For every $(C_u, C_v) \in [0,1] \times [0,1]$, there exists $\theta(C_u, C_v)$, such that for every $p > \theta(C_u, C_v)$ the achieving test-channels $P_{\mathsf{U}|\mathsf{X}}$ and $P_{\mathsf{V}|\mathsf{Y}}$ are BSC with parameters $\alpha = h_b^{-1}(1 - C_u)$ and $\beta = h_b^{-1}(1 - C_v)$ respectively.*

We will provide supporting arguments for this conjecture via numerical simulations in Section 6.

3.2. Bounds

In this section we present our lower and upper bounds on the BDSIB function, then we compare them for various channel parameters. The proofs are postponed to Appendix F. For the simplicity of the following presentation we define

$$g_b(x) \triangleq \frac{1}{2(1-x)} h_b(x), \quad x \in [0, 1/2], \tag{19}$$

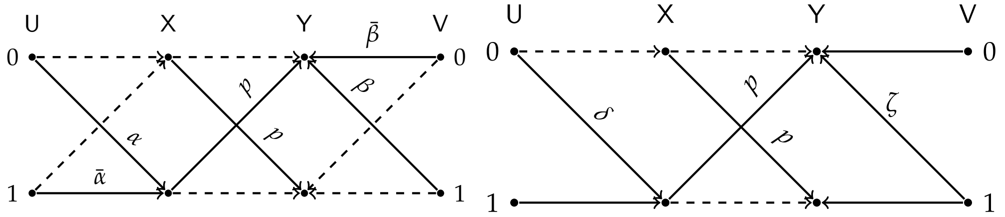
denote $g_b^{-1}(\cdot)$ as its inverse restricted to $[0,1]$, and $\hbar(x) \triangleq -x \log x$.

Proposition 5. *The BDSIB function is bounded from below by*

$$R(C_u, C_v, p) \ge \max\begin{cases} 1 - h_b(\alpha * \beta * p), \\ 1 - \frac{1}{2\bar{\delta}\bar{\zeta}}\left[\hbar(\delta * \zeta * p) + (1-2\zeta)\cdot\hbar(\bar{\delta} * p) + (1-2\delta)\cdot\hbar(\bar{\zeta} * p) + (1-2\delta)(1-2\zeta)\cdot\hbar(p)\right], \end{cases} \tag{20}$$

where $\alpha = h_b^{-1}(1 - C_u)$, $\beta = h_b^{-1}(1 - C_v)$, $\delta = g_b^{-1}(1 - C_u)$, and $\zeta = g_b^{-1}(1 - C_v)$.

All terms in the RSH of (20) are attained by taking test-channels that match the constraints with equality and plugging them in Lemma 3. In particular: the first term is achieved by BSC test channels with transition probabilities α and β; the second term is achieved by taking $\mathsf{P}_{\mathsf{U}|\mathsf{X}}$ be a $Z(\delta)$ channel and $\mathsf{P}_{\mathsf{V}|\mathsf{Y}}$ be an $S(\zeta)$ channel. The aforementioned test-channel configurations are illustrated in Figure 10.

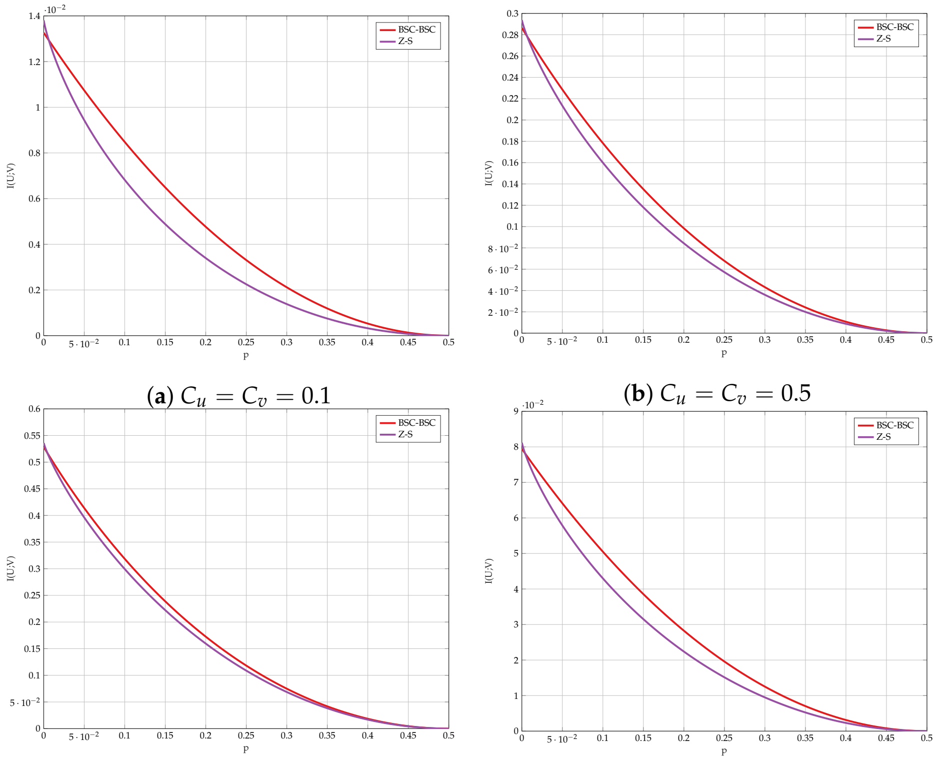


(**a**) Two BSCs.

(**b**) Z and S channels.

Figure 10. Test-channel that achieve the lower bound of Proposition 5.

We compare the different lower bounds derived in Proposition 5 for various values of constraints. The achievable rate vs channel transition probability p is shown in Figure 11. Our first observation is that BSC test-channels outperform all other choices for almost all values of p. However, Figure 12 gives a closer look on small values of p. It is evident that the combination of Z and S test-channels outperforms any other schemes for small values of p. We have used this observation as one supporting evidence to Conjecture 1.



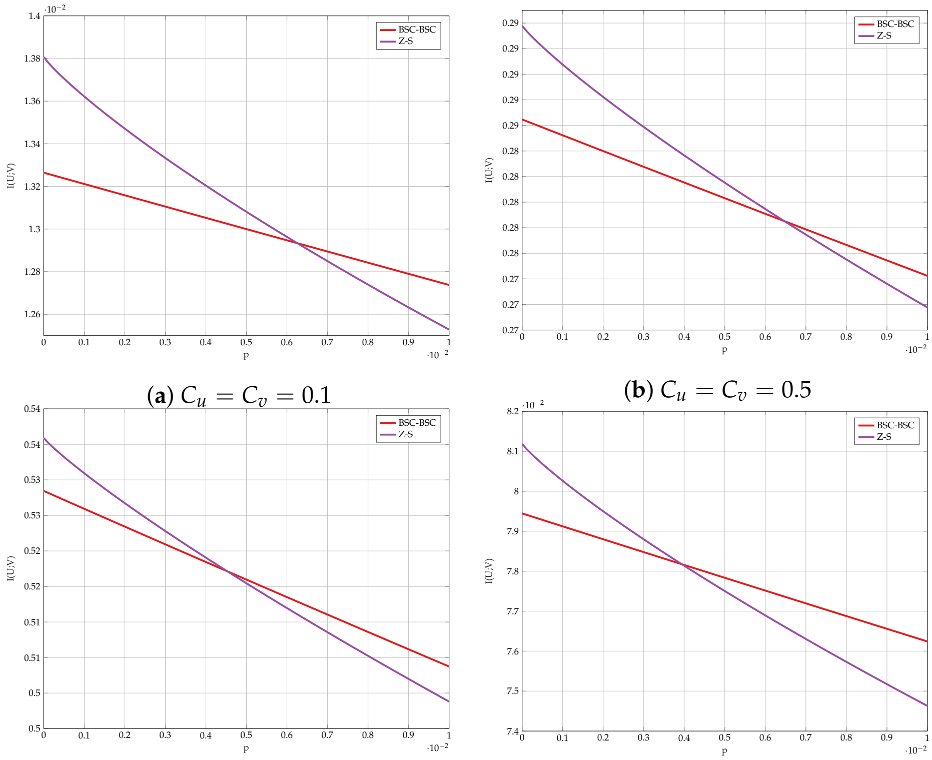
(**a**) $C_u = C_v = 0.1$

(**b**) $C_u = C_v = 0.5$

(**c**) $C_u = C_v = 0.7$

(**d**) $C_u = 0.1, C_v = 0.7$

Figure 11. Comparison of the lower bounds.



(**a**) $C_u = C_v = 0.1$

(**b**) $C_u = C_v = 0.5$

(**c**) $C_u = C_v = 0.7$

(**d**) $C_u = 0.1, C_v = 0.7$

Figure 12. Comparison of the lower bounds in high SNR regime.

We proceed to give an upper bound.

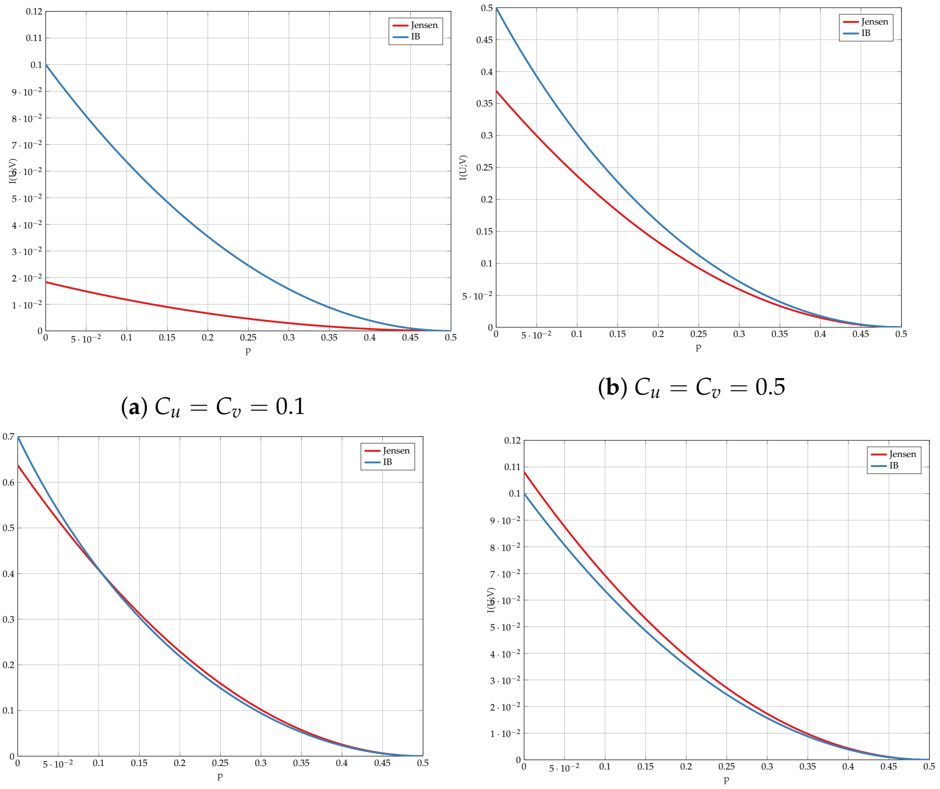
Proposition 6. *A general upper bound on BDSIB is given by*

$$R(C_u, C_v, p) \leq \min\begin{cases}(1-2p)^2(1-2h_b^{-1}(1-C_u))^2(1-2h_b^{-1}(1-C_v))^2, \\ \min\{1-h_b(h_b^{-1}(1-C_u)*p), 1-h_b(h_b^{-1}(1-C_v)*p)\}.\end{cases} \tag{21}$$

Note that the first term can be derived by applying Jensen's inequality on (12), and the second term is a combination of the standard IB and the cut-set bound. We postpone the proof of Proposition 6 to Appendix F.

Remark 7. *Since $p = \frac{1}{2} - \epsilon$, we have a factor 2 loss in the first term compared to the precise behavior we have found for $p \approx \frac{1}{2}$ in Theorem 1. This loss comes from the fact that the bound in (21) actually upper bounds the χ-squared mutual information between U and V. It is well-known that for very small $I(\mathsf{X};\mathsf{Y})$ we have that $I(\mathsf{X};\mathsf{Y}) \approx 1/2 I_{\chi^2}(\mathsf{X};\mathsf{Y})$, see [65].*

We compare the different upper bounds from Proposition 6 in Figure 13 for various bottleneck constraints, and in Figure 14 for various values of channel transition probabilities p. We observe that there are regions of C and p for which Jensen's based bound outperforms the standard IB bound.



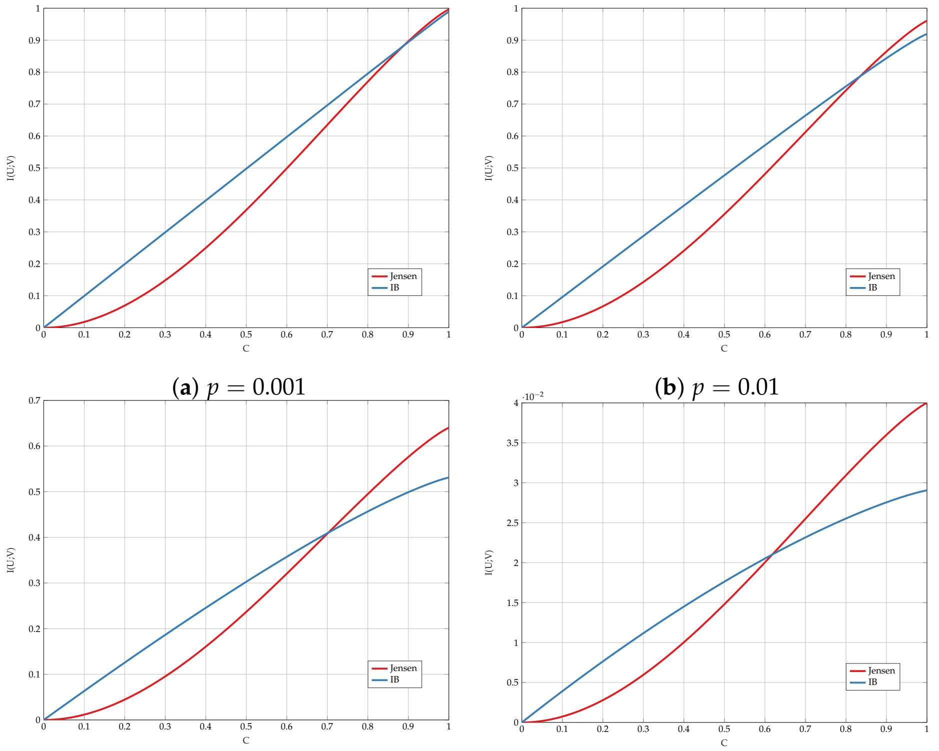
(**a**) $C_u = C_v = 0.1$

(**b**) $C_u = C_v = 0.5$

(**c**) $C_u = C_v = 0.7$

(**d**) $C_u = 0.1, C_v = 0.7$

Figure 13. Comparison of the upper bounds for various values of (C_u, C_v).



(**a**) $p = 0.001$

(**b**) $p = 0.01$

(**c**) $p = 0.1$

(**d**) $p = 0.4$

Figure 14. Comparison of the upper bounds for various values of p.

Finally, we compare the best lower and upper bounds from Propositions 5 and 6 for various values of channel parameters in Figure 15. We observe that the bounds are tighter for asymmetric constraints and high transition probabilities.

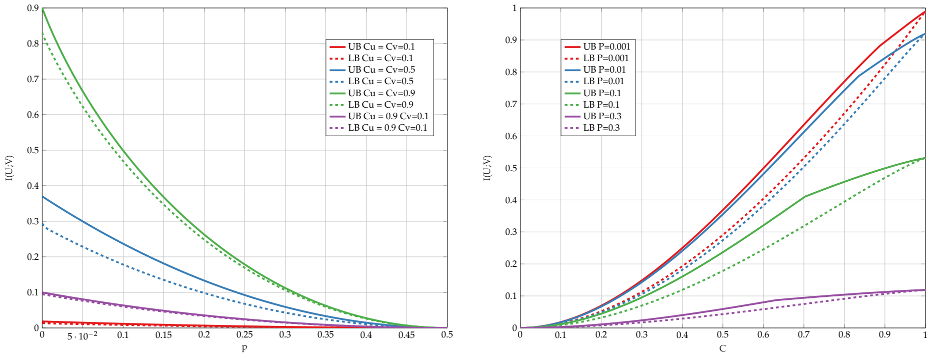


Figure 15. Capacity bounds for various values of p and $C = C_u = C_v$.

4. Gaussian (X, Y)

In this section we consider a specific setting where (X, Y) is the normalized zero mean Gaussian bivariate source, namely,

$$\begin{pmatrix} \mathsf{X} \\ \mathsf{Y} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right). \tag{22}$$

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 *probabilistic Hermite polynomials* which are defined by

$$H_n(x) \triangleq (-1)^n e^{\frac{x^2}{2}} \frac{\mathrm{d}^n}{\mathrm{d}x^n} e^{-\frac{x^2}{2}}, \quad n \in \mathbb{N}_0. \tag{23}$$

We denote the Gaussian DSIB function with explicit dependency on ρ as $R(C_u, C_v, \rho)$.

Proposition 7. *Let $H_n(\cdot)$ be the nth order probabilistic Hermite polynomial, then the objective function of (7) for the Gaussian setting is given by*

$$I(\mathsf{U;V}) = \mathbb{E}_{\mathsf{UV}}\left[\log\left(\sum_{n=0}^{\infty} \frac{\rho^n}{n!} \mathbb{E}[H_n(\mathsf{X})|\mathsf{U}]\mathbb{E}[H_n(\mathsf{Y})|\mathsf{V}] \right) \right]. \tag{24}$$

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 [66]. Mehler Kernel decomposition is a special case of a much richer family of Lancaster distributions [67]. The proof of Proposition 7 is relegated to Appendix G.

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 constraints with equality.

Proposition 8. *A lower bound on $R(C_u, C_v, \rho)$ is given by*

$$R(C_u, C_v, \rho) \geq -\frac{1}{2} \log\left(1 - \rho^2 \left(1 - 2^{-2C_u}\right)\left(1 - 2^{-2C_v}\right)\right). \tag{25}$$

The proof of this bound is developed in Appendix H.

Although it was shown in [26] 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 rate 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.

Proposition 9. *Let* $(C_u, C_v) = (1, 1)$*, then*

$$R(1, 1, \rho) \geq 1 - h_2\left(\frac{\arccos \rho}{\pi}\right). \tag{26}$$

The proof of this bound is developed in Appendix I.

We compare the bounds from Propositions 8 and 9 with $(C_u, C_v) = (1, 1)$ in Figure 16. 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 given by

$$\rho_{\text{cros}} = \frac{e}{\sqrt{1 + e^2}} \to \sqrt{SNR_{\text{cros}}} = \frac{\rho_{\text{cros}}}{\sqrt{1 - \rho^2_{\text{cros}}}} = e. \tag{27}$$

We proceed to present our upper bound on $R(C_u, C_v, \rho)$. This bound is a combination of the cutset bound and the single-sided Gaussian IB.

Proposition 10. *An upper bound on* (7) *with Gaussian* (X, Y) *setting* (22) *is given by*

$$R(C_u, C_v, \rho) \leq \min\left\{-\frac{1}{2}\log(1 - \rho^2(1 - 2^{-2C_u})), -\frac{1}{2}\log(1 - \rho^2(1 - 2^{-2C_v}))\right\}. \tag{28}$$

We compare the best lower and upper bounds from Propositions 8–10 in Figure 17. We observe that the bounds become tighter as the constraint increases and in the low-SNR regime.

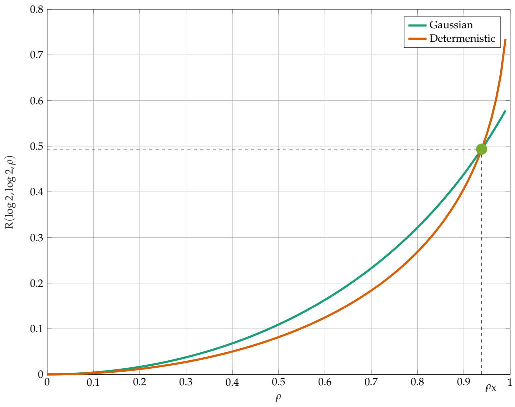


Figure 16. Comparison of the lower bounds from Propositions 8 and 9.

4.1. Low-SNR Regime

For $\rho \to 0$, the exact asymptotic behavior of the Gaussian (Proposition 8) and deterministic (Proposition 9) test-channels, respectively, for $C_u = C_v = 1$ bit is given by:

$$\lim_{\rho \to 0} -\frac{1}{2}\log\left(1 - \rho^2(1 - 2^{-2C_u})(1 - 2^{-2C_v})\right) = \frac{9 \log e}{32}\rho^2 + o(\rho^2),$$

$$\lim_{\rho \to 0} 1 - h_2\left(\frac{\arccos \rho}{\pi}\right) = \frac{2 \log e}{\pi^2}\rho^2 + o(\rho^2).$$

Hence, the Gaussian choice outperforms the second lower bound for vanishing SNR. The following theorem establishes that Gaussian test-channels are optimal for low-SNR.

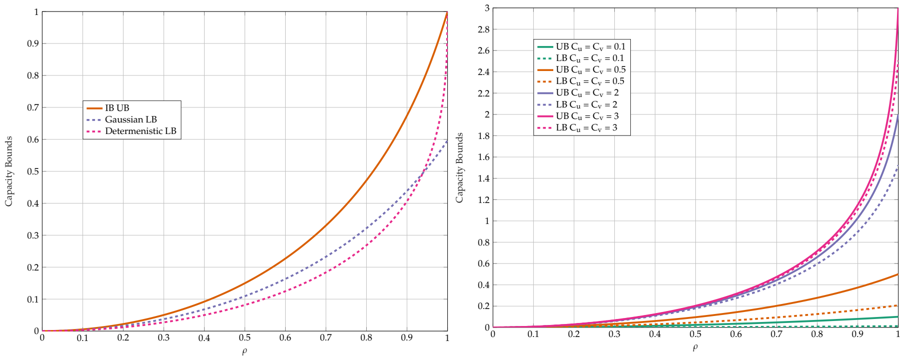


Figure 17. Capacity bounds for various values of p and $C = C_u = C_v = 1$.

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

$$R(C_u, C_v, \rho) = \frac{\rho^2 \log e}{2}(1 - 2^{-2C_u})(1 - 2^{-2C_v}) + o(\rho^2). \tag{29}$$

The lower bound follows from Proposition 8. The upper bound is established by considering the kernel representation from Proposition 7 in the limit of vanishing ρ. The detailed proof is given in Appendix J.

4.2. Optimality of Symbol-by-Symbol Quantization When $\mathsf{X} = \mathsf{Y}$

Consider an extreme scenario for which $\mathsf{X} = \mathsf{Y} \sim \mathcal{N}(0, 1)$. Taking the encoders $\mathsf{P}_{\mathsf{U}|\mathsf{X}}$ and $\mathsf{P}_{\mathsf{V}|\mathsf{X}}$ as a symbol-by-symbol deterministic quantizers satisfying:

$$H(\mathsf{U}) = H(\mathsf{V}) = \min\{C_u, C_v\},$$

we achieve the optimum

$$I(\mathsf{U}; \mathsf{V}) = \min\{C_u, C_v\}.$$

5. Alternating Maximization Algorithm

Consider the DSIB problem for DSBS with parameter p analyzed in Section 3. The respective optimization problem involves simultaneous search of the maximum over the sets $\{\mathsf{P}_{\mathsf{U}|\mathsf{X}}\}$ and $\{\mathsf{P}_{\mathsf{V}|\mathsf{Y}}\}$. An alternating maximization, namely, fixing $\mathsf{P}_{\mathsf{U}|\mathsf{X}}$, then finding the respective optimal $\mathsf{P}_{\mathsf{V}|\mathsf{Y}}$ and vice versa, is sub-optimal in general and may result in convergent to a saddle point. However, for the case $p = 0$ with symmetric bottleneck constraints, Proposition 4 implies that such point exists only for the BSC and Z/S channels. This motivates us to believe that performing an alternating maximization procedure on (9) will not result in sub-optimal saddle point, but rather converge to the optimal solution also for the general discrete (X, Y).

Thus, we propose an alternating maximization algorithm. The main idea is to fix $\mathsf{P}_{\mathsf{V}|\mathsf{Y}}$ and then compute $\mathsf{P}^*_{\mathsf{U}|\mathsf{X}}$ that attains the inner term in (9). Then, using $\mathsf{P}^*_{\mathsf{U}|\mathsf{X}}$, we find the optimal $\mathsf{P}^*_{\mathsf{V}|\mathsf{Y}}$ that attains the inner term in (10). Then, we repeat the procedure in alternating manner until convergence.

Note that inner terms of (9) and (10) are just the standard IB problem defined in (6). For completeness, we state here the main result from [1] and adjust it for our problem. Consider the respective Lagrangian of (6) given by:

$$L(\mathsf{P}_{\mathsf{U}|\mathsf{X}}, \lambda) = I(\mathsf{U}; \mathsf{V}) + \lambda(C - I(\mathsf{X}; \mathsf{U})). \tag{30}$$

Lemma 4 (Theorem 4 of [1])**.** *The optimal test-channel that maximizes* (30) *satisfies the equation:*

$$\mathsf{P}_{\mathsf{U}|\mathsf{X}}(u|x) = \frac{\mathsf{P}_{\mathsf{U}}(u)}{Z(x,\beta)} e^{-\beta D(\mathsf{P}_{\mathsf{V}|\mathsf{X}=x} \| \mathsf{P}_{\mathsf{V}|\mathsf{U}=u})}, \tag{31}$$

where $\beta \triangleq 1/\lambda$ *and* $\mathsf{P}_{\mathsf{V}|\mathsf{U}}$ *is given via Bayes' rule, as follows:*

$$\mathsf{P}_{\mathsf{V}|\mathsf{U}}(v|u) = \frac{1}{\mathsf{P}_{\mathsf{U}}(u)} \sum_x \mathsf{P}_{\mathsf{V}|\mathsf{X}}(v|x) \mathsf{P}_{\mathsf{U}|\mathsf{X}}(u|x) \mathsf{P}_{\mathsf{X}}(x). \tag{32}$$

In a very similar manner to the Blahut–Arimoto algorithm [18], the self-consistent equations can be adapted into converging, alternating iterations over the convex sets $\{\mathsf{P}_{\mathsf{U}|\mathsf{X}}\} = \Delta_n^{\otimes n}$, $\{\mathsf{P}_{\mathsf{U}}\} = \Delta_n$, and $\{\mathsf{P}_{\mathsf{V}|\mathsf{U}}\} = \Delta_n^{\otimes n}$, as stated in the following lemma.

Lemma 5 (Theorem 5 of [1])**.** *The self-consistent equations are satisfied simultaneously at the minima of the functional:*

$$F(\mathsf{P}_{\mathsf{U}|\mathsf{X}}, \mathsf{P}_{\mathsf{U}}, \mathsf{P}_{\mathsf{Y}|\mathsf{U}}) = I(\mathsf{U};\mathsf{X}) + \beta \mathbb{E}\Big[D(\mathsf{P}_{\mathsf{V}|\mathsf{X}} \| \mathsf{P}_{\mathsf{V}|\mathsf{U}})\Big], \tag{33}$$

where the minimization is performed independently over the convex sets of $\{\mathsf{P}_{\mathsf{U}|\mathsf{X}}\} = \Delta_n^{\otimes n}$*,* $\{\mathsf{P}_{\mathsf{U}}\} = \Delta_n$*, and* $\{\mathsf{P}_{\mathsf{V}|\mathsf{U}}\} = \Delta_n^{\otimes n}$*. The minimization is performed by the converging alternation iterations as described in Algorithm 1.*

Algorithm 1: IB iterative algorithm IBAM(args)

```
Input: $\mathsf{P}^{(0)}_{\mathsf{U}|\mathsf{X}}, \mathsf{P}_{\mathsf{XY}}, \beta, \epsilon$
$R^{(0)} = 0$
$t \leftarrow 0$
while $\Delta R \geq \epsilon$ do
    $\mathsf{P}_{\mathsf{U}}(u) \leftarrow \sum_x \mathsf{P}_{\mathsf{X}}(x) \mathsf{P}^{(t)}_{\mathsf{U}|\mathsf{X}}(u|x)$
    $\mathsf{P}_{\mathsf{X}|\mathsf{U}} \leftarrow \frac{\mathsf{P}^{(t)}_{\mathsf{U}|\mathsf{X}} \mathsf{P}_{\mathsf{X}}}{\mathsf{P}_{\mathsf{U}}}$
    $\mathsf{P}_{\mathsf{Y}|\mathsf{U}}(y|u) \leftarrow \sum_x \mathsf{P}_{\mathsf{Y}|\mathsf{X}}(y|x) \mathsf{P}_{\mathsf{X}|\mathsf{U}}(x|u)$
    $\mathsf{P}^{(t+1)}_{\mathsf{U}|\mathsf{X}}(u|x) \leftarrow \frac{\mathsf{P}_{\mathsf{U}}(u) \exp\left(-\beta D(\mathsf{P}_{\mathsf{Y}|\mathsf{X}=x} \| \mathsf{P}_{\mathsf{Y}|\mathsf{U}=u})\right)}{\sum_u \mathsf{P}_{\mathsf{U}}(u) \exp\left(-\beta D(\mathsf{P}_{\mathsf{Y}|\mathsf{X}=x} \| \mathsf{P}_{\mathsf{Y}|\mathsf{U}=u})\right)}$
    $R^{(t+1)} \leftarrow I(\mathsf{P}^{(t+1)}_{\mathsf{U}}, \mathsf{P}^{(t+1)}_{\mathsf{Y}|\mathsf{U}})$
    $\Delta R = |R^{(t+1)} - R^{(t)}|$
    $t \leftarrow t+1$
Output: $\mathsf{P}^{(t)}_{\mathsf{U}|\mathsf{X}}(u|x)$
```

Next, we propose a combined algorithm to solve the optimization problem from (7). The main idea is to fix one of the test-channels, i.e., $\mathsf{P}_{\mathsf{V}|\mathsf{Y}}$, and then find the corresponding optimal opposite test-channel, i.e., $\mathsf{P}_{\mathsf{U}|\mathsf{X}}$, using Algorithm 1. Then, we apply again Algorithm 1 by switching roles, i.e., fixing the opposite test-channel, i.e., $\mathsf{P}_{\mathsf{U}|\mathsf{X}}$, and then identifying the optimal $\mathsf{P}_{\mathsf{V}|\mathsf{Y}}$. We repeat this procedure until convergence of the objective function $I(\mathsf{U};\mathsf{V})$. We summarize the proposed composite method in Algorithm 2.

Remark 8. *Note that every alternating step of the algorithm involves finding an optimal* (β^*, η^*) *that corresponds to the respective problem constraints* (C_u, C_v)*. We have chosen to implement this exploration step using a bisection-type method. This may result that the actual pair* (C_u, C_v) *is* ϵ*-far away from the desired constraint.*

Algorithm 2: DSIB iterative algorithm DSIBAM(args)

```
Input: $P^{(0)}_{U|X}, P^{(0)}_{V|Y}, P_{XY}, C_u, C_v, \epsilon$
$R^{(0)} = 0$
$s \leftarrow 0$
while $\Delta R \geq \epsilon$ do
    $P_{XV}(x,v) \leftarrow \sum_y P^{(s)}_{V|Y}(v|y) P_{XY}(x,y)$
    $P_{U|X}(\beta) \leftarrow IBAM(P^{(s)}_{U|X}, P_{XV}, \beta, \epsilon)$
    $\beta^* \leftarrow \arg\min_\beta |I(P_X, P_{U|X}(\beta)) - C_u|$
    $P^{(s+1)}_{U|X} \leftarrow P_{U|X}(\beta^*)$
    $P_{YU}(y,u) \leftarrow \sum_x P^{(s+1)}_{U|X}(u|x) P_{XY}(x,y)$
    $P_{V|Y}(\eta) \leftarrow IBAM(P^{(s)}_{V|Y}, P_{YU}, \eta, \epsilon)$
    $\eta^* \leftarrow \arg\min_\eta |I(P_Y, P_{V|Y}(\eta)) - C_v|$
    $P^{(s+1)}_{V|Y} \leftarrow P_{V|Y}(\eta^*)$
    $P_{UV}(u,v) = \sum_y P^{(s+1)}_{V|Y}(v,y) P_{YU}(y,u)$
    $R^{(s+1)} \leftarrow I(P_{UV})$
    $\Delta R \leftarrow |R^{(s+1)} - R^{(s)}|$
    $s \leftarrow s + 1$
$C^{(s)}_u \leftarrow I(P_X, P^{(s)}_{U|X})$
$C^{(s)}_v \leftarrow I(P_Y, P^{(s)}_{V|Y})$
Output: $P^{(s)}_{U|X}, P^{(s)}_{V|Y}, R^{(s)}, C^{(s)}_u, C^{(s)}_v$
```

6. Numerical Results

In this section, we focus on the DSBS setting of Section 3. In the first part of this section, we will show using a brute-force method the existence of a sharp, phase-transition phenomena in the optimal test-channels $P_{U|X}$ and $P_{V|Y}$ vs. DSBS parameter p. In the second part of this section, we will evaluate the alternating maximization algorithm proposed in Section 5; then, we compare its performance to the brute-force method.

6.1. Exhaustive Search

In this set of simulations, we again fix the transition matrix from Y to V characterized by the parameters:

$$T = \begin{pmatrix} a & b \\ 1-a & 1-b \end{pmatrix}, \tag{34}$$

chosen such that $I(Y;V) = C_v$. This choice defines a path $b = f(a)$ in the (a,b) plain. Then, for every such T we optimize $I(U;V)$ for different values of the DSBS parameter p. The results for a specific choice of $(C_u, C_v) = (0.4, 0.6)$ vs. a for different values of p are plotted in Figure 18. Note that the region of a corresponds to the continuous conversion from a Z channel ($a = 0$) to a BSC ($a = a_{\max}$). We observe here a very sharp transition from the optimality of Z-S channels to BSC channels configuration for a small change in p. This kind of behavior continues to hold with a different choice of $(C_u = 0.1, C_v = 0.9)$, as can be seen in Figure 19.

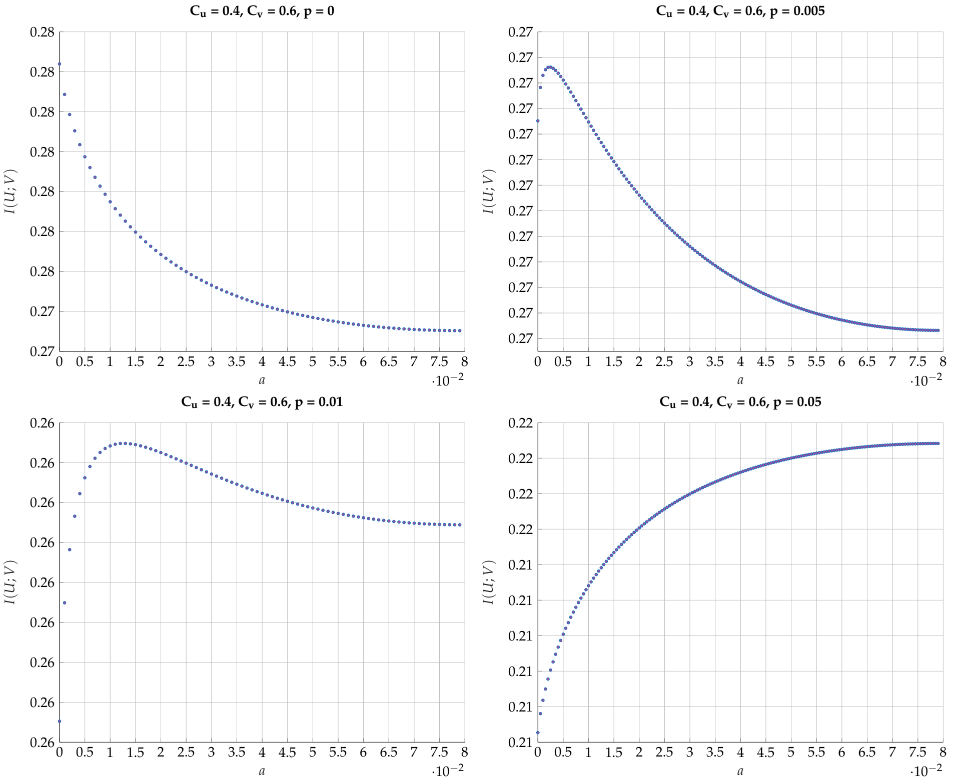


Figure 18. Maximal $I(\mathsf{U};\mathsf{V})$ for fixed values $(C_u, C_v) = (0.4, 0.6)$ and different values of p.

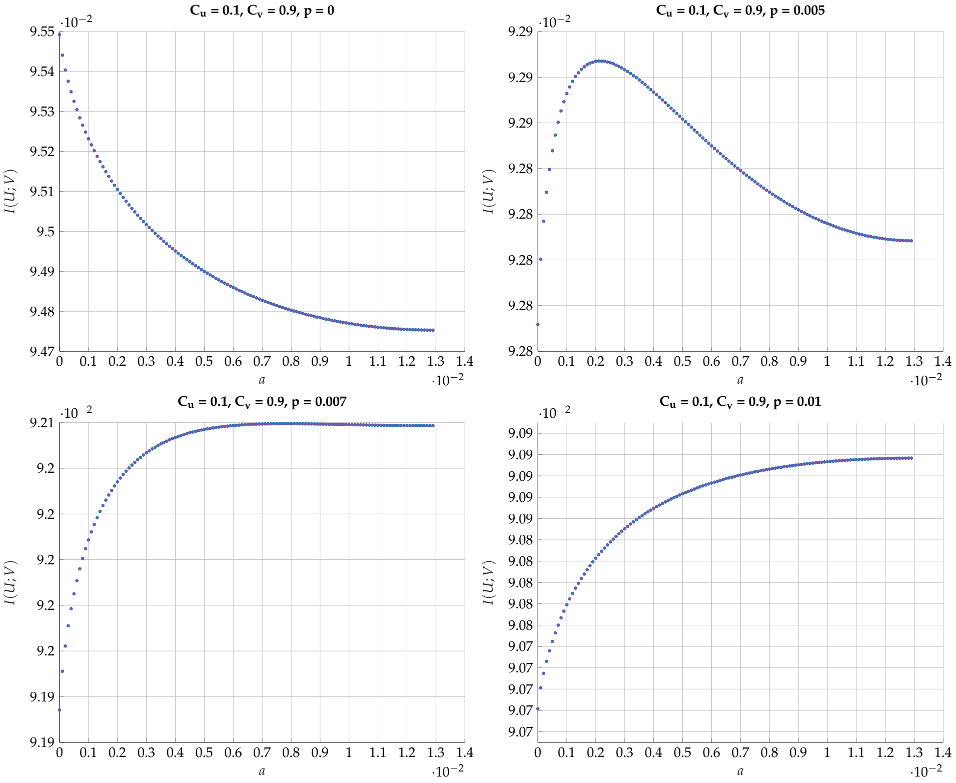


Figure 19. Maximal $I(\mathsf{U};\mathsf{V})$ for fixed values $(C_u, C_v) = (0.1, 0.9)$ and different values of p.

Next, we would like to emphasize this sharp phase transition phenomena by plotting the optimal a that achieves the maximal $I(\mathsf{U};\mathsf{V})$ vs the DSBS parameter p. The results for various combinations of C_u and C_v are presented in Figures 20 and 21. We observe that the curves are convex for $p \in [0, p_{th})$ and constant for $p > p_{th}$ with $a = a_{bsc}$. Furthermore, the derivative of $a(p)$ for $p \to p_{th}$ tends to ∞.

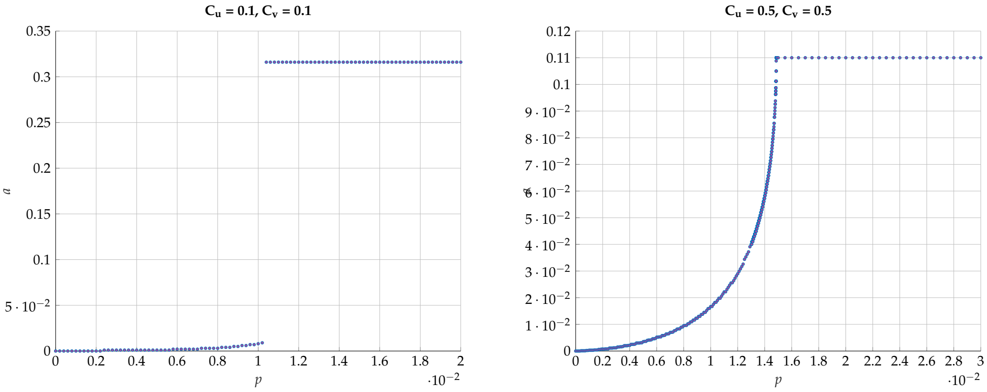


Figure 20. Optimal value of a for various values of C_u and C_v.

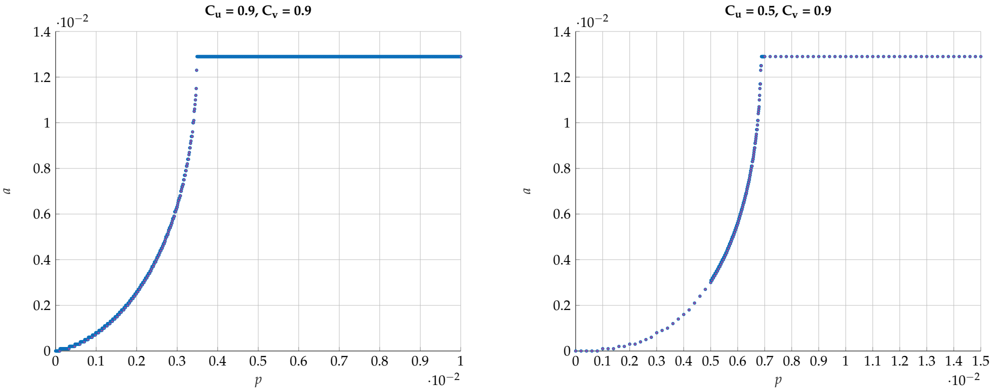


Figure 21. Optimal value of a for various values of C_u and $C_v = 0.9$.

One may further claim that there is no sharp transition to the BSC test-channels $\mathsf{P}_{\mathsf{U}|\mathsf{X}}$ and $\mathsf{P}_{\mathsf{V}|\mathsf{Y}}$ as p grows away from zero, but rather only approaches BSC. To convince the reader that the optimal test channels are exactly BSC, we performed an alternating maximization experiment. We fixed $p > 0$, C_u and C_v. Then we have chosen $\mathsf{P}_{\mathsf{V}|\mathsf{Y}}$ as an almost BSC channel satisfying $I(\mathsf{Y};\mathsf{V}) \le C_v$ and found the channel $P_{\mathsf{X}|\mathsf{U}}$ that maximizes $I(\mathsf{U};\mathsf{V})$ subject to $I(\mathsf{X};\mathsf{U}) \le C_u$. Then, we fixed the channel $\mathsf{P}_{\mathsf{X}|\mathsf{U}}$ and found the $\mathsf{P}_{\mathsf{Y}|\mathsf{V}}$ that maximizes $I(\mathsf{U};\mathsf{V})$ subject to $I(\mathsf{Y};\mathsf{V}) \le C_v$. We have repeated this alternating maximization procedure until it converges. The transition matrices were parameterized as follows:

$$T_{\mathsf{Y}|\mathsf{V}} = \begin{pmatrix} q_0 & q_1 \\ 1-q_0 & 1-q_1 \end{pmatrix}, \qquad T_{\mathsf{X}|\mathsf{U}} = \begin{pmatrix} p_0 & p_1 \\ 1-p_0 & 1-p_1 \end{pmatrix}. \tag{35}$$

The results for different values of p, C_u, and C_v are shown in Figures 22–24. We observe that p_0 and q_0 rapidly converge to their respective BSC values satisfying the mutual information constraints. Note that the last procedure is still an exhaustive search, but it is performed in alternating fashion between the sets $\{\mathsf{P}_{\mathsf{U}|\mathsf{X}}\}$ and $\{\mathsf{P}_{\mathsf{V}|\mathsf{Y}}\}$.

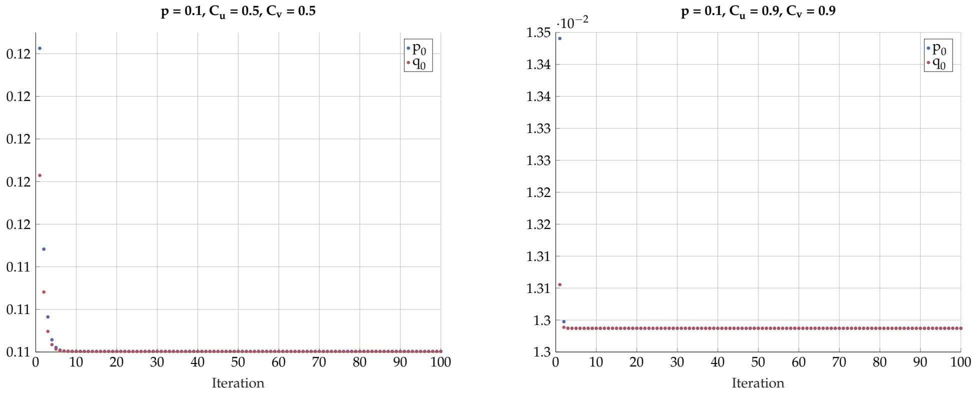


Figure 22. Alternating maximization with exhaustive search for various p, C_u, C_v.

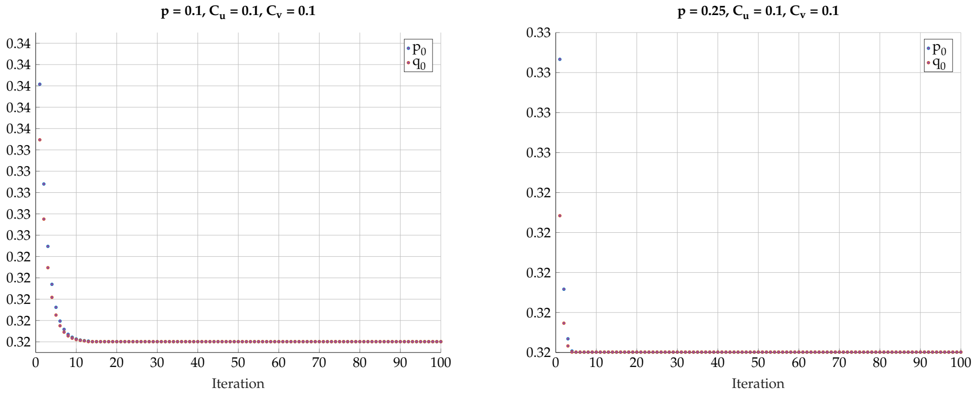


Figure 23. Alternating maximization with exhaustive search for various p, C_u, C_v.

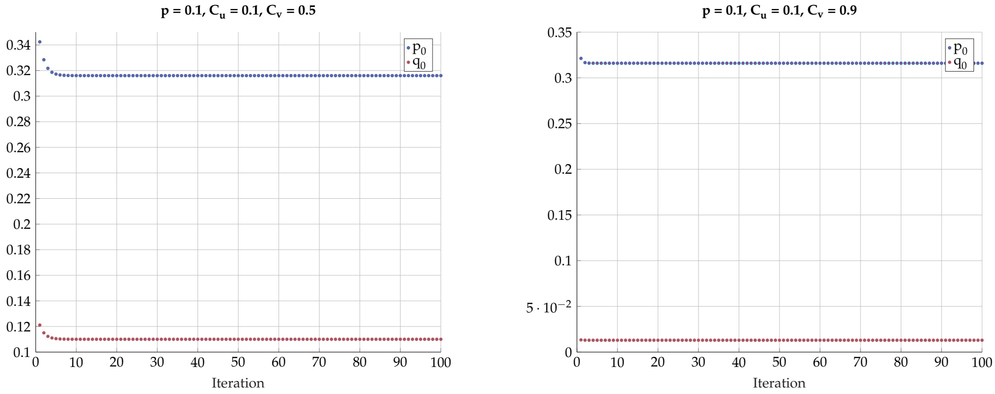
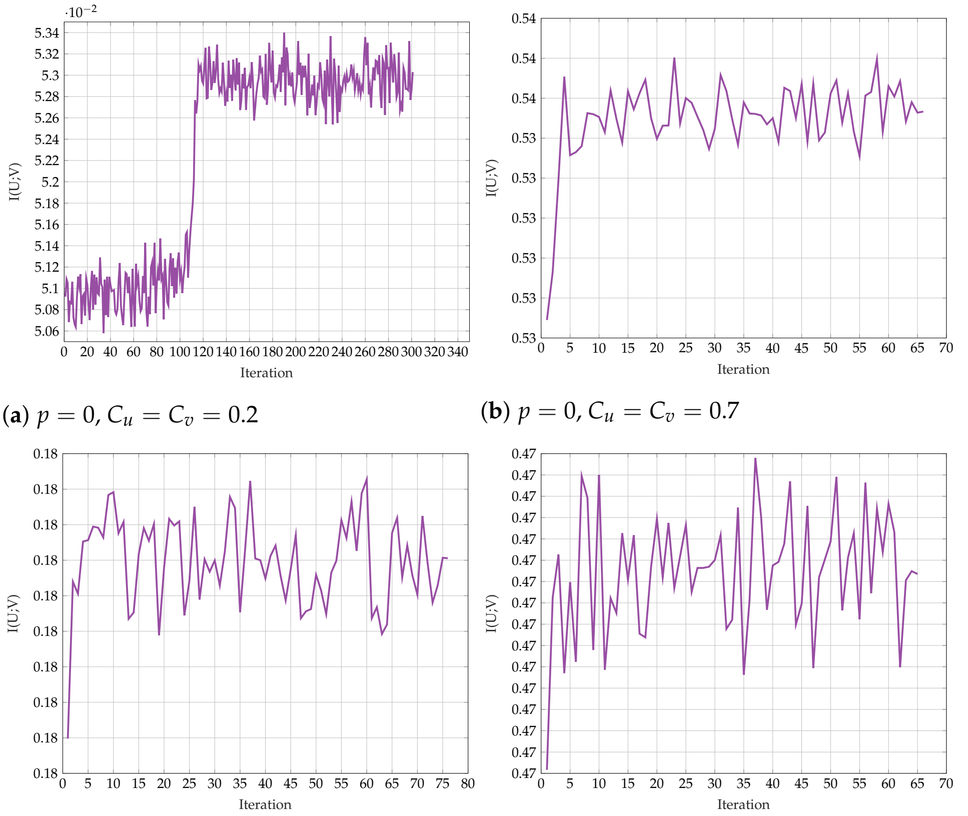


Figure 24. Alternating maximization with exhaustive search for various p, C_u, C_v.

6.2. *Alternating Maximization*

In this section, we will evaluate the algorithm proposed in Section 5. We focus on the DSBS setting of Section 3 with various values of problem parameters.

First, we explore the convergence behavior of the proposed algorithm. Figure 25 shows the objective function $I(\mathsf{U};\mathsf{V})$ on every iteration step for representative fixed-channel transition parameters p and the constraints C_u and C_v. We observe a slow convergence to a final value for $p = 0$ and $C_u = C_v = 0.2$, but once the constraints and the transition probability are increased, the algorithm converges much more rapidly. The non-monotonic behavior in some regimes is justified with the help of Remark 8. In Figure 26, we see the respective test-channel probabilities $\alpha_0 + \alpha_1$, $1 - \alpha_0$, $\beta_0 + \beta_1$, and $1 - \beta_1$. First, note that if $\alpha_0 + \alpha_1 = 1$, then $\mathsf{P}_{\mathsf{X}|\mathsf{U}}$ is a BSC. Similarly, if $\beta_0 + \beta_1 = 1$, then $\mathsf{P}_{\mathsf{Y}|\mathsf{V}}$ is a BSC. Second, if $1 - \alpha_0 = 1$, then $\mathsf{P}_{\mathsf{X}|\mathsf{U}}$ is a Z-channel. Similarly, if $1 - \beta_1 = 1$, then $\mathsf{P}_{\mathsf{Y}|\mathsf{V}}$ is an S-channel. We observe that for $p = 0$, the test-channels $\mathsf{P}_{\mathsf{X}|\mathsf{U}}$ and $\mathsf{P}_{\mathsf{Y}|\mathsf{V}}$ converge to Z- and S-channels, respectively. As for all other settings, the test-channels converge to BSC channels.



(**a**) $p = 0$, $C_u = C_v = 0.2$

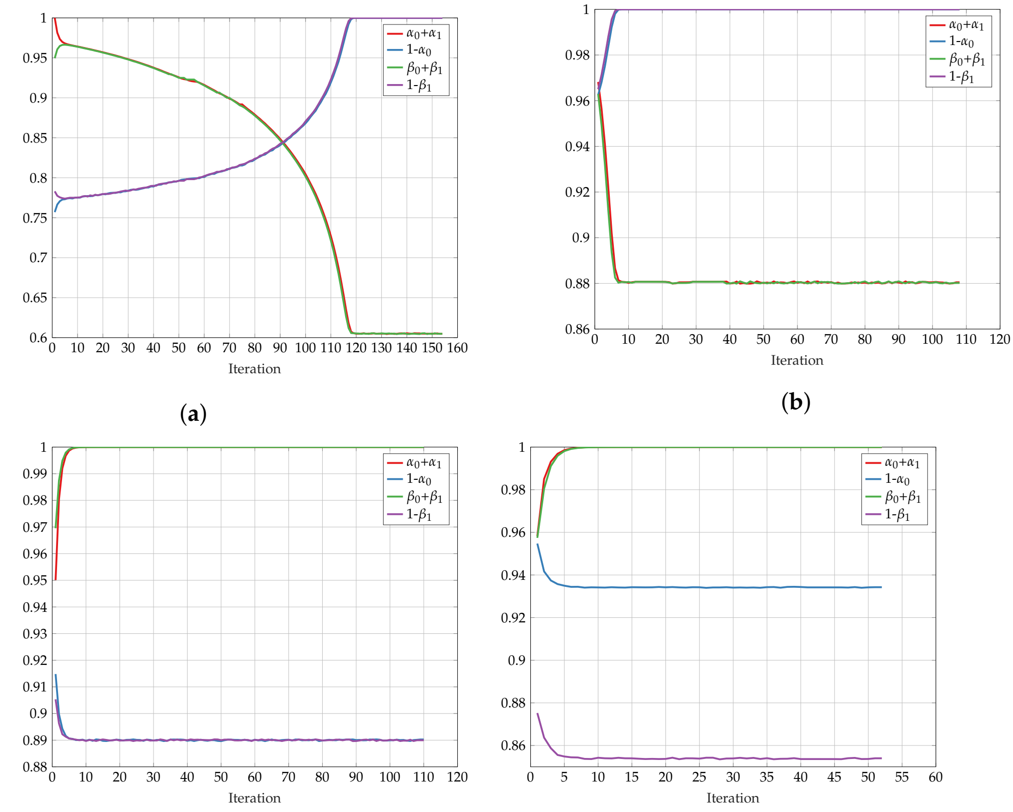
(**b**) $p = 0$, $C_u = C_v = 0.7$

(**c**) $p = 0.1$, $C_u = C_v = 0.5$

(**d**) $p = 0.001$, $C_u = 0.9$ and $C_v = 0.5$

Figure 25. Convergence of $I(\mathsf{U};\mathsf{V})$ for various values of p, C_u and C_v.

Finally, we compare the outcome of Algorithm 2 to the optimal solution achieved by the brute-force method, namely, evaluating (12) for every $\mathsf{P}_{\mathsf{U}|\mathsf{X}}$ and $\mathsf{P}_{\mathsf{V}|\mathsf{Y}}$ that satisfy the problem constraints. The results for various values of channel parameters are shown in Figure 27. We observe that the proposed algorithm achieves the optimum for any DSBS parameter p and some representative constraints C_u and C_v.



(**a**) (**b**)

(**c**) (**d**)

Figure 26. Convergence of $I(\mathsf{U};\mathsf{V})$ p with: (**a**) $C_u = C_v = 0.2$, $p = 0$; (**b**) $C_u = C_v = 0.7$, $p = 0$; (**c**) $C_u = C_v = 0.5$, $p = 0.1$; (**d**) $C_u = 0.65, C_v = 0.4$, $p = 0.1$.

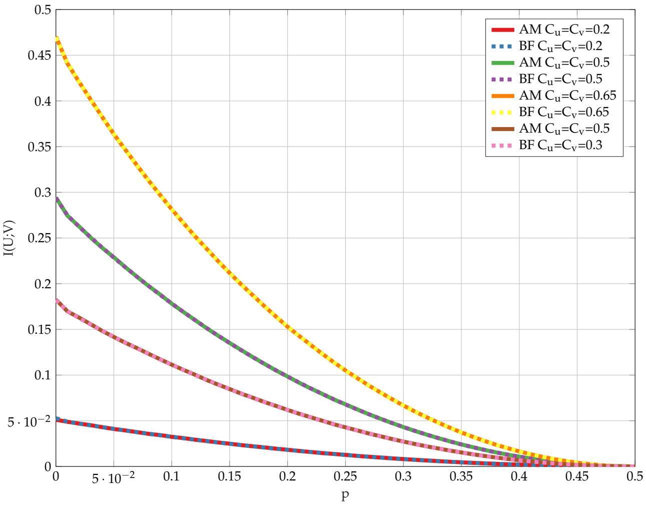


Figure 27. Comparison of the proposed alternating maximization algorithm and the brute-force search method for various problem parameters.

7. Concluding Remarks

In this paper, we have considered the Double-Sided Information Bottleneck problem. Cardinality bounds on the representation's alphabets were obtained for an arbitrary discrete bivariate source. When X and Y are binary, we have shown that taking binary auxiliary random variables is optimal. For DSBS, we have shown that BSC test-channels are optimal when $p \to 0.5$. Furthermore, numerical simulations for arbitrary p indicate 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. In addition, the optimality of the Gaussian test-channels 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 $I(\mathsf{U};\mathsf{X}^n) \leq nC_u$ and $I(\mathsf{V};\mathsf{Y}^n) \leq nC_v$ does not tensorize in general. For $\mathsf{X}^n = \mathsf{Y}^n \sim \mathrm{Ber}^{\otimes n}(0.5)$, 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.

Finally, we have proposed an alternating maximization algorithm based on the standard IB [1]. For the DSBS, it was shown that the algorithm converges to the global optimal solution.

Author Contributions: Conceptualization, O.O. and S.S.; methodology, M.D., O.O., and S.S.; software, M.D.; formal analysis, M.D.; writing—original draft preparation, M.D.; writing—review and editing, M.D.; supervision, S.S. All authors have read and agreed to the published version of the manuscript.

Funding: The work has been supported by the European Union's Horizon 2020 Research And Innovation Programme, grant agreement no. 694630, by the ISF under Grant 1641/21, and by the WIN consortium via the Israel minister of economy and science.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proof of Proposition 2

Before proceeding to proof Proposition 2, we need the following auxiliary results.

Lemma A1. *Let $\mathsf{P}_{\mathsf{Y}|\mathsf{X}}$ be an arbitrary binary-input, ternary-output channel, parameterized using the following transition matrix:*

$$T \triangleq \begin{pmatrix} a & b \\ c & d \\ 1-a-c & 1-b-d \end{pmatrix}. \tag{A1}$$

Consider the function $p \mapsto \phi(p,\lambda) = h(T\mathbf{p}) - \lambda h_b(p)$ defined on $[0,1]$. This function has the following properties:

1. *If $\phi(p,\lambda)$ is linear on a sub-interval of $[0,1]$, then it is linear for every $p \in [0,1]$.*
2. *Otherwise, it is strictly convex over $[0,1]$ or there are points p_l and p_u such that $0 < p_l < p_u < 1$ where*

$$\phi(p,\lambda) = \begin{cases} \text{strictly convex} & 0 < p < p_l = \mathcal{I}_1, \\ \text{strictly concave} & p_l < p < p_u = \mathcal{I}_2, \\ \text{strictly convex} & p_u < p < 1 = \mathcal{I}_3. \end{cases} \tag{A2}$$

We postpone the proof of this lemma to Appendix K.

Lemma A2. *The convex envelope of $\phi(\cdot)$ at any point $q \in [0,1]$ can be obtained as a convex combination of only points in $\mathcal{I}_1$ and $\mathcal{I}_3$.*

We postpone the proof of this lemma to Appendix L and proceed to proof Proposition 2. Note that if $F_T(x)$ is strictly convex in $[0, h_b(q)]$, then by the paragraph following (Theorem 2.3 of [17]) $|\mathcal{U}| = 2$, we are done.

From now on, we consider the case where $F_T(x)$ is not strictly convex. Then, there is an interval $\mathcal{L} \subset [0, h(q)]$ and $a \in \mathbb{R}_+$ such that

$$F_T(x) = a + \lambda_{\mathcal{L}} \cdot x \quad \forall x \in \mathcal{L}. \tag{A3}$$

Let $\mathbf{t}_0$ and $\mathbf{t}_1$ represent the columns of T corresponding to $\mathsf{X} = 0$ and $\mathsf{X} = 1$, respectively, Moreover let $q \triangleq \mathsf{P}(\mathsf{X} = 0)$ and $\mathbf{p} \triangleq (p, \bar{p})^T$ be the probability vector of an arbitrary binary random variable, where $\bar{p} \triangleq 1 - p$.

Assume $x_1, x_2 \in \mathcal{L}$ and $x_1 \neq x_2$. Then, there must be $\{\alpha_{1i}, p_{1i}\}_{i=1,2,3}$ and $\{\alpha_{2i}, p_{2i}\}_{i=1,2,3}$ such that

$$\sum_{i=1}^{3} \alpha_{1i} p_{1i} = q, \qquad \sum_{i=1}^{3} \alpha_{1i} h_b(p_{1i}) = x_1, \qquad \sum_{i=1}^{3} \alpha_{1i} h(T\mathbf{p}_{1i}) = a + \lambda_{\mathcal{L}} x_1, \tag{A4}$$

$$\sum_{i=1}^{3} \alpha_{2i} p_{2i} = q, \qquad \sum_{i=1}^{3} \alpha_{2i} h_b(p_{2i}) = x_2, \qquad \sum_{i=1}^{3} \alpha_{2i} h(T\mathbf{p}_{2i}) = a + \lambda_{\mathcal{L}} x_2. \tag{A5}$$

Lemma A3. *The set* $\{p_{11}, p_{12}, p_{13}, p_{21}, p_{22}, p_{23}\}$ *must contain at least three distinct points.*

We postpone the proof of this lemma to Appendix M.

Consider the function $p \mapsto \phi(p) = \phi(p, \lambda_{\mathcal{L}}) = h(T\mathbf{p}) - \lambda_{\mathcal{L}} h_b(p)$ defined on $[0, 1]$. We have that

$$\sum_{i=1}^{3} \alpha_{1i} \phi(p_{1i}) = \sum_{i=1}^{3} \alpha_{2i} \phi(p_{2i}) = a. \tag{A6}$$

In addition, if we define $\psi(\cdot)$ to be the lower convex envelope of $\phi(\cdot)$, then $\psi(q) = a$. Thus, the lower convex envelope of $\phi(\cdot)$ at q is attained by two linear combinations.

By Lemma Lemma A3, the set $\{p_{11}, p_{12}, p_{13}, p_{21}, p_{22}, p_{23}\}$ must contain at least three distinct points, say $\{p_{11}, p_{21}, p_{22}\}$. Due to Lemma A2, they are all in $\mathcal{I}_1 \cup \mathcal{I}_3$. Furthermore, by the pigeonhole principle, we must have that one of the intervals contains at least two points. Assume WLOG that $\{p_{11}, p_{21}\} \in \mathcal{I}_1$. For any $\gamma \in [0, 1]$, let $S = \bar{\gamma}\alpha_{11} + \gamma\alpha_{21}$ and consider the following set of weights/probabilities:

$$\left\{ \left(S, \frac{\bar{\gamma}\alpha_{11}}{S} \cdot p_{11} + \frac{\gamma\alpha_{21}}{S} \cdot p_{21} \right), (\bar{\gamma}\alpha_{12}, p_{12}), (\bar{\gamma}\alpha_{13}, p_{13}), (\gamma\alpha_{22}, p_{22}), (\gamma\alpha_{23}, p_{23}) \right\}. \tag{A7}$$

Note that

$$S + \bar{\gamma}\alpha_{12} + \bar{\gamma}\alpha_{13} + \gamma\alpha_{22} + \gamma\alpha_{23} = 1, \tag{A8}$$

and

$$\bar{\gamma}\alpha_{11} \cdot p_{11} + \gamma\alpha_{21} \cdot p_{21} \bar{\gamma}\alpha_{12} \cdot p_{12} + \bar{\gamma}\alpha_{13}, p_{13} + \gamma\alpha_{22} \cdot p_{22} + \gamma\alpha_{23} \cdot p_{23} = q, \tag{A9}$$

but since $\{p_{11}, p_{21}\} \in \mathcal{I}_1$

$$S \cdot \phi\left(\frac{\bar{\gamma}\alpha_{11}}{S} \cdot p_{11} + \frac{\gamma\alpha_{21}}{S} \cdot p_{21} \right) + \bar{\gamma}\alpha_{12}\phi(p_{12}) + \bar{\gamma}\alpha_{13}\phi(p_{13}) + \gamma\alpha_{22}\phi(p_{22}) + \gamma\alpha_{23}\phi(p_{23}) \tag{A10}$$

$$< S \cdot \left(\frac{\bar{\gamma}\alpha_{11}}{S} \cdot \phi(p_{11}) + \frac{\gamma\alpha_{21}}{S} \cdot \phi(p_{21}) \right) + \bar{\gamma}\alpha_{12}\phi(p_{12}) + \bar{\gamma}\alpha_{13}\phi(p_{13}) + \gamma\alpha_{22}\phi(p_{22}) + \gamma\alpha_{23}\phi(p_{23}) = a, \tag{A11}$$

thus, it attains a smaller value than a, provided that ϕ is strictly convex on $\mathcal{I}_1$. This contradicts our assumption that the convex envelope at q equals a, and thus $\phi(\cdot)$ must contain a linear segment in $\mathcal{I}_1$.

By Lemma A1, this can happen only if p is linear for every $p \in [0, 1]$. In particular:

$$h(T\mathbf{p}) - \lambda_{\mathcal{L}} h_b(p) = \phi(p) = (1 - p)\phi(0) + p\phi(1) = (1 - p)h(\mathbf{t}_0) + ph(\mathbf{t}_1). \tag{A12}$$

Note that for any choice of $\mathsf{P}_{\mathsf{X}|\mathsf{U}=u}$

$$\begin{aligned} H(\mathsf{Y}|\mathsf{U} = u) &= h(T\mathbf{p}_u) && \text{(A13)} \\ &= \phi(p_u) + \lambda_{\mathcal{L}} h_b(p_u) && \text{(A14)} \\ &= (1 - p_u)h(\mathbf{t}_0) + p_u h(\mathbf{t}_1) + \lambda_{\mathcal{L}} h_b(p_u). && \text{(A15)} \end{aligned}$$

Taking the expectation we obtain:

$$H(\mathsf{Y}|\mathsf{U}) = (1-q)h(\mathbf{t}_0) + qh(\mathbf{t}_1) + \lambda_{\mathcal{L}}x. \tag{A16}$$

This implies that

$$F_T(x) = (1-q)h(\mathbf{t}_0) + qh(\mathbf{t}_1) + \lambda_{\mathcal{L}}x, \tag{A17}$$

and this is attained by any choice of $\mathsf{P}_{\mathsf{U}|\mathsf{X}}$ satisfying $H(\mathsf{X}|\mathsf{U}) = x$. In particular the choice $\mathsf{U} = \mathsf{X} \oplus \mathsf{Z}$, where $\mathsf{Z} \sim \mathrm{Ber}(\delta)$ is statistically independent of X and is chosen such that $H(\mathsf{X}|\mathsf{U}) = x$, attains $F_T(x)$. Thus, $|\mathcal{U}| = 2$ suffices even if $F_T(x)$ is not strictly convex.

Appendix B. Proof of Lemma 3

Let $\mathsf{P}_{\mathsf{U}|\mathsf{X}}$ and $\mathsf{P}_{\mathsf{V}|\mathsf{Y}}$ be the test-channels from X to U and from Y to V, respectively. The joint probability function of U and V can be expressed via Bayes' rule and the Markov chain condition $\mathsf{U} \to \mathsf{X} \to \mathsf{Y} \to \mathsf{V}$ as:

$$\mathsf{P}_{\mathsf{UV}}(u,v) = 4 \cdot \mathsf{P}_{\mathsf{U}}(u)\mathsf{P}_{\mathsf{V}}(v) \sum_{x,y} \mathsf{P}_{\mathsf{X}|\mathsf{U}}(x|u)\mathsf{P}_{\mathsf{XY}}(x,y)\mathsf{P}_{\mathsf{Y}|\mathsf{V}}(y|v). \tag{A18}$$

Since $I(\mathsf{U};\mathsf{V}) = \mathbb{E}[\log(\mathsf{P}_{\mathsf{UV}}/\mathsf{P}_{\mathsf{U}} \times \mathsf{P}_{\mathsf{V}})]$, we define $K(u,v,p)$ as the ratio between the joint distribution of U and V relative to the respective product measure. Note that:

$$\begin{aligned}
K(u,v,p) &\triangleq \frac{\mathsf{P}_{\mathsf{UV}}(u,v)}{\mathsf{P}_{\mathsf{U}}(u)\mathsf{P}_{\mathsf{V}}(v)} && \text{(A19)}\\
&= 4\sum_{x,y} \mathsf{P}_{\mathsf{X}|\mathsf{U}}(x|u)\mathsf{P}_{\mathsf{XY}}(x,y)\mathsf{P}_{\mathsf{Y}|\mathsf{V}}(y|v) && \text{(A20)}\\
&= 2(\mathsf{P}_{\mathsf{X}|\mathsf{U}}(0|u) \cdot \bar{p} \cdot \mathsf{P}_{\mathsf{Y}|\mathsf{V}}(0|u) + \mathsf{P}_{\mathsf{X}|\mathsf{U}}(1|u) \cdot p \cdot \mathsf{P}_{\mathsf{Y}|\mathsf{V}}(0|u)) && \text{(A21)}\\
&\quad + 2(\mathsf{P}_{\mathsf{X}|\mathsf{U}}(0|u) \cdot p \cdot \mathsf{P}_{\mathsf{Y}|\mathsf{V}}(1|u) + \mathsf{P}_{\mathsf{X}|\mathsf{U}}(1|u) \cdot \bar{p} \cdot \mathsf{P}_{\mathsf{Y}|\mathsf{V}}(1|u)). && \text{(A22)}
\end{aligned}$$

Denoting $\alpha_u \triangleq \mathsf{P}_{\mathsf{X}|\mathsf{U}}(1|u)$ and $\beta_v \triangleq \mathsf{P}_{\mathsf{Y}|\mathsf{V}}(0|v)$, we obtain:

$$K(u,v,p) = 2(\bar{\alpha}_u\bar{p}\beta_v + \alpha_u p\beta_v + \bar{\alpha}_u p\bar{\beta}_v + \alpha_u\bar{p}\bar{\beta}_v) = 2\alpha_u * \beta_v * p. \tag{A23}$$

The last expression can also be represented as follows:

$$\begin{aligned}
2\alpha_u * \beta_v * p &= 2(1-p)(\alpha_u + \beta_v - 2\alpha_u\beta_v) + 2p(1 - \alpha_u - \beta_v + 2\alpha_u\beta_v) && \text{(A24)}\\
&= 2\alpha_u + 2\beta_v - 4\alpha_u\beta_v + 2p(1 - 2\alpha_u - 2\beta_v + 4\alpha_u\beta_v) && \text{(A25)}\\
&= 1 - (1-2p)(1 - 2\alpha_u - 2\beta_v + 4\alpha_u\beta_v) && \text{(A26)}\\
&= 1 - (1-2p)(1-2\alpha_u)(1-2\beta_v). && \text{(A27)}
\end{aligned}$$

Thus,

$$\begin{aligned}
I(\mathsf{U};\mathsf{V}) &= \sum_{u,v} \mathsf{P}_{\mathsf{UV}}(u,v) \log \frac{\mathsf{P}_{\mathsf{UV}}(u,v)}{\mathsf{P}_{\mathsf{U}}(u)\mathsf{P}_{\mathsf{V}}(v)} && \text{(A28)}\\
&= \sum_{u,v} \mathsf{P}_{\mathsf{U}}(u)\mathsf{P}_{\mathsf{V}}(v)K(u,v,p)\log K(u,v,p). && \text{(A29)}
\end{aligned}$$

Furthermore, note that since $|(1-2p)(1-2\alpha_u)(1-2\beta_v)| < 1$, we can utilize Taylor's expansion of $\log(1-x)$ to obtain:

$$\log K(u,v,p) = -\sum_{n=1}^{\infty} \frac{(1-2p)^n(1-2\alpha_u)^n(1-2\beta_v)^n}{n}, \tag{A30}$$

and

$$K(u,v,p)\log K(u,v,p) = -\sum_{n=1}^{\infty}\frac{(1-2p)^n(1-2\alpha_u)^n(1-2\beta_v)^n}{n} \\ +\sum_{n=1}^{\infty}\frac{(1-2p)^{n+1}(1-2\alpha_u)^{n+1}(1-2\beta_v)^{n+1}}{n}. \tag{A31}$$

Therefore:

$$I(\mathsf{U};\mathsf{V}) = -\sum_{n=1}^{\infty}\frac{(1-2p)^n\mathbb{E}[(1-2\alpha_{\mathsf{U}})^n]\mathbb{E}[(1-2\beta_{\mathsf{V}})^n]}{n} \tag{A32}$$

$$+\sum_{n=1}^{\infty}\frac{(1-2p)^{n+1}\mathbb{E}\big[(1-2\alpha_{\mathsf{V}})^{n+1}\big]\mathbb{E}\big[(1-2\beta_{\mathsf{V}})^{n+1}\big]}{n} \tag{A33}$$

$$\overset{(a)}{=} -\sum_{n=2}^{\infty}\frac{(1-2p)^n\mathbb{E}[(1-2\alpha_{\mathsf{U}})^n]\mathbb{E}[(1-2\beta_{\mathsf{V}})^n]}{n} \tag{A34}$$

$$+\sum_{n=1}^{\infty}\frac{(1-2p)^{n+1}\mathbb{E}\big[(1-2\alpha_{\mathsf{V}})^{n+1}\big]\mathbb{E}\big[(1-2\beta_{\mathsf{V}})^{n+1}\big]}{n} \tag{A35}$$

$$= -\sum_{n=1}^{\infty}\frac{(1-2p)^{n+1}\mathbb{E}\big[(1-2\alpha_{\mathsf{U}})^{n+1}\big]\mathbb{E}\big[(1-2\beta_{\mathsf{V}})^{n+1}\big]}{n+1} \tag{A36}$$

$$+\sum_{n=1}^{\infty}\frac{(1-2p)^{n+1}\mathbb{E}\big[(1-2\alpha_{\mathsf{V}})^{n+1}\big]\mathbb{E}\big[(1-2\beta_{\mathsf{V}})^{n+1}\big]}{n} \tag{A37}$$

$$= \sum_{n=1}^{\infty}(1-2p)^{n+1}\mathbb{E}\Big[(1-2\alpha_{\mathsf{U}})^{n+1}\Big]\mathbb{E}\Big[(1-2\beta_{\mathsf{V}})^{n+1}\Big]\cdot\frac{1}{n(n+1)}, \tag{A38}$$

where (a) follows since $\mathbb{E}[\alpha_{\mathsf{U}}] = \mathbb{E}[\beta_{\mathsf{V}}] = \frac{1}{2}$. This completes the proof.

Appendix C. Auxiliary Concavity Lemma

As a preliminary step to proving Theorem 1, we will need the following auxiliary lemma.

Lemma A4. *The function $f(x) = (1-2h_b^{-1}(x))^2$ is concave.*

Proof. Denoting $g(x) \triangleq h_b^{-1}(x)$, we have $f(x) = (1-2g(x))^2$. Since $f(x)$ is twice differentiable, it is sufficient to show that $f'(x)$ is decreasing. The first derivative is given by:

$$f'(x) = -4(1-2g(x))g'(x). \tag{A39}$$

Since

$$h_b(x) = -x\log x - (1-x)\log(1-x), \tag{A40}$$

$$h_b'(x) = \log\frac{1-x}{x}, \tag{A41}$$

$$h_b''(x) = -\frac{1}{(1-x)\ln 2} - \frac{1}{x\ln 2} = -\frac{1}{x(1-x)\ln 2}, \tag{A42}$$

utilizing the inverse function derivative property, we obtain:

$$g(x) = h_b^{-1}(x), \tag{A43}$$

$$g'(x) = \frac{1}{h'(g(x))} = \frac{1}{\log\frac{1-g(x)}{g(x)}}. \tag{A44}$$

In addition, the second order derivative is given by:

$$g''(x) = \frac{\log e}{\log^3 \frac{1-g(x)}{g(x)}} \frac{g(x)}{1-g(x)} \frac{1}{(g(x))^2} \tag{A45}$$

$$= \frac{\log e}{\log^3 \frac{1-g(x)}{g(x)}} \frac{1}{(1-g(x))g(x)} \tag{A46}$$

$$= \frac{(g'(x))^3}{\ln 2 g(x)(1-g(x))}. \tag{A47}$$

Define

$$r(t) \triangleq \frac{-4(1-2t)}{\log \frac{1-t}{t}}. \tag{A48}$$

Note that $f'(x) = r(g(x))$. Since $g(x)$ is increasing, in order to show that $f'(x)$ decreasing, it suffices to show that $r(t)$ decreasing. The first order derivative of $r(t)$ is given by:

$$r'(t) = \frac{8}{\log^2 \frac{1-t}{t}} \frac{1}{t(1-t)} \left(t(1-t) \log \frac{1-t}{t} - \frac{1-2t}{\ln 4} \right).$$

Define $\alpha \triangleq 1 - 2t$ such that $t = \frac{1}{2}(1-\alpha)$. Note that $\alpha \in [0,1]$. We obtain:

$$r'(t) = \frac{32}{\log^2 \frac{1+\alpha}{1-\alpha}} \frac{1}{1-\alpha^2} \left(\frac{1}{4}(1-\alpha^2) \log \frac{1+\alpha}{1-\alpha} - \frac{\alpha}{\ln 4} \right).$$

Now, making use of the expansion $\log(1+x) = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{x^k}{k}$, we have:

$$\log \frac{1+\alpha}{1-\alpha} = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{\alpha^k}{k} - \sum_{k=1}^{\infty} (-1)^{k+1} \frac{(-\alpha)^k}{k} = 2 \sum_{k \text{ odd}} \frac{\alpha^k}{k}.$$

Thus,

$$\begin{aligned}
&\frac{1}{4}(1-\alpha^2) \log \frac{1+\alpha}{1-\alpha} - \frac{\alpha}{\ln 4} \\
&= \frac{1}{2} \sum_{k \text{ odd}} \frac{\alpha^k}{k} - \frac{1}{2} \sum_{k \text{ odd}} \frac{\alpha^{k+2}}{k} - \frac{\alpha}{\ln 4} \\
&= \alpha \left(\frac{1}{2} - \frac{1}{\ln 4} \right) + \frac{1}{2} \sum_{\substack{k \text{ odd} \\ k \geq 3}} \alpha^k \left(\frac{1}{k} - \frac{1}{k-2} \right) \\
&\overset{(a)}{<} \alpha \left(\frac{1}{2} - \frac{1}{\ln e^2} \right) - \sum_{\substack{k \text{ odd} \\ k \geq 3}} \frac{\alpha^k}{k(k-2)} < 0,
\end{aligned}$$

where (a) follows since $\alpha > 0$. Thus, $r'(t) < 0$ and $f(x)$ is concave. □

Appendix D. Proof of Theorem 1

Plugging $p \leftarrow \frac{1}{2} - \epsilon$ ($\epsilon \triangleq \frac{1}{2} - p$) in (14), we obtain:

$$K(u,v,\epsilon) = 1 + 2\epsilon(1-2\alpha_u)(1-2\beta_v). \tag{A49}$$

Now, we rewrite $I(\mathsf{U};\mathsf{V})$ with explicit dependency on ϵ as:

$$I(\epsilon) = \sum_{u,v} P_{\mathsf{U}}(u) P_{\mathsf{V}}(v) K(u,v,\epsilon) \log K(u,v,\epsilon). \tag{A50}$$

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

$$\alpha_u = \begin{cases} h_2^{-1}(H(\mathsf{X}|\mathsf{U}=u)), & \alpha_u \leq \frac{1}{2} \\ 1 - h_2^{-1}(H(\mathsf{X}|\mathsf{U}=u)), & \alpha_u > \frac{1}{2} \end{cases} \tag{A51}$$

with similar relation for β_v. Therefore,

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

where the first inequality follows since the function $f : x \mapsto (1-2h_2^{-1}(x))^2$ is concave by Lemma A4 and applying Jensen's inequality, and the second inequality follows from rate constraints.

Appendix E. Proof of Proposition 4

Suppose that the optimal test-channel $\mathsf{P}_{\mathsf{V}|\mathsf{X}}$ is given by the following transition matrix:

$$T_{\mathsf{V}|\mathsf{X}} = \begin{pmatrix} a & b \\ 1-a & 1-b \end{pmatrix}. \tag{A52}$$

Assume in contradiction that the opposite optimal test-channel $\mathsf{P}_{\mathsf{U}|\mathsf{X}}$ is symmetric to $\mathsf{P}_{\mathsf{V}|\mathsf{X}}$ and is given by:

$$T_{\mathsf{U}|\mathsf{X}} = \begin{pmatrix} 1-b & 1-a \\ b & a \end{pmatrix}. \tag{A53}$$

Applying Bayes' rule on (A53), we obtain:

$$T_{\mathsf{X}|\mathsf{U}} = \begin{pmatrix} 1-\alpha_0 & 1-\alpha_1 \\ \alpha_0 & \alpha_1 \end{pmatrix} = \begin{pmatrix} \frac{\bar{b}}{\bar{a}+\bar{b}} & \frac{b}{a+b} \\ \frac{\bar{a}}{\bar{a}+\bar{b}} & \frac{a}{a+b} \end{pmatrix}. \tag{A54}$$

It was shown in (Section IV.D of [17]) that for fixed $\mathsf{P}_{\mathsf{V}|\mathsf{X}}$ given by (A52), the optimal $\mathsf{P}_{\mathsf{X}|\mathsf{U}}$ must satisfy the following equation:

$$\begin{aligned} (a-b)(h_b(\alpha_1)-h_b(\alpha_0))(h_b'(\hat{\alpha}_0)-h_b'(\hat{\alpha}_1)) + (h_b'(\alpha_1)-h_b'(\alpha_0))(h_b(\hat{\alpha}_1)-h_b(\hat{\alpha}_0)) \\ + (a-b)(\alpha_1-\alpha_0)(h_b'(\alpha_0)h_b'(\hat{\alpha}_1)-h_b'(\alpha_1)h_b'(\hat{\alpha}_0)) = 0, \end{aligned} \tag{A55}$$

where $\hat{\alpha}_0 \triangleq a\alpha_0 + b\bar{\alpha}_0$ and $\hat{\alpha}_1 \triangleq a\alpha_1 + b\bar{\alpha}_1$. Plugging α_0 and α_1 from (A54) in (A55) results in a contradiction, thus establishing the proof of Proposition 4.

Appendix F. Proof of Proposition 6

By Lemma 3, the objective function of (7) for a DSBS setting, denoted here by $I(p)$, is given by:

$$I(p) = \mathbb{E}_{\mathsf{P}_{\mathsf{U}}\times\mathsf{P}_{\mathsf{V}}}[K(\mathsf{U},\mathsf{V},p)\log K(\mathsf{U},\mathsf{V},p)], \tag{A56}$$

where $K(u,v,p)$ can be expressed as:

$$K(u,v,p) = 1 + (1-2p)(1-2\alpha_u * \beta_v) = 1 + (1-2p)(1-2\alpha_u)(1-2\beta_v). \tag{A57}$$

Since $\log(1+x) \leq x$, we have the following upper bound on $I(p)$:

$$\begin{aligned}
I(p) &= \sum_{u,v} P_{\mathsf{U}}(u)P_{\mathsf{V}}(v)K(u,v,p)\log K(u,v,p) && \text{(A58)}\\
&\leq \sum_{u,v} P_{\mathsf{U}}(u)P_{\mathsf{V}}(v)(1+(1-2p)(1-2\alpha_u)(1-2\beta_v))(1-2p)(1-2\alpha_u)(1-2\beta_v) && \text{(A59)}\\
&= (1-2p)(1-2\sum_u P_{\mathsf{U}}(u)\alpha_u)(1-2\sum_v P_{\mathsf{V}}(v)\beta_v) && \text{(A60)}\\
&\quad + (1-2p)^2\sum_u P_{\mathsf{U}}(u)(1-2\alpha_u)^2\sum_v P_{\mathsf{V}}(v)(1-2\beta_v)^2 && \text{(A61)}\\
&= (1-2p)(1-2\sum_u P_{\mathsf{U}}(u)\mathsf{P}(\mathsf{X}=1|\mathsf{U}=u))(1-2\sum_v P_{\mathsf{V}}(v)\mathsf{P}(\mathsf{Y}=1|\mathsf{V}=v)) && \text{(A62)}\\
&+ (1-2p)^2\sum_u P_{\mathsf{U}}(u)(1-2\mathsf{P}(\mathsf{X}=1|\mathsf{U}=u))^2\sum_v P_{\mathsf{V}}(v)(1-2\mathsf{P}(\mathsf{Y}=1|\mathsf{V}=v))^2 \\
&= (1-2p)(1-2\mathsf{P}(\mathsf{X}=1))(1-2\mathsf{P}(\mathsf{Y}=1)) && \text{(A63)}\\
&+ (1-2p)^2\sum_u P_{\mathsf{U}}(u)(1-2h_2^{-1}(H(\mathsf{X}|\mathsf{U}=u)))^2\sum_v P_{\mathsf{V}}(v)(1-2h_2^{-1}(H(\mathsf{Y}=|\mathsf{V}=v)))^2 \\
&\overset{(a)}{\leq} (1-2p)^2(1-2h_2^{-1}(H(\mathsf{X}|\mathsf{U})))^2(1-2h_2^{-1}(H(\mathsf{Y}=|\mathsf{V})))^2 && \text{(A64)}\\
&\overset{(b)}{\leq} (1-2p)^2(1-2h_2^{-1}(1-C_x)^2(1-2h_2^{-1}(1-C_y)^2, && \text{(A65)}
\end{aligned}$$

where the inequality in (a) follows from Lemma A4 and inequality in (b) follows from the problem constraints.

Appendix G. Proof of Proposition 7

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 $M(x,y;\rho) = \sum_{n=0}^{\infty}\frac{\rho^n}{n!}H_n(x)H_n(y)$ [66]. Similarly, $I(\mathsf{U};\mathsf{V})$ can also be written with explicit dependency on ρ

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

Appendix H. Proof of Proposition 8

Let $(\mathsf{U},\mathsf{X},\mathsf{Y},\mathsf{V})$ be jointly Gaussian Random variables, such that

$$\mathsf{X} = \sigma_{\mathsf{UX}}\mathsf{U} + \sqrt{1-\sigma_{\mathsf{UX}}^2}\mathsf{Z}_u, \qquad \mathsf{Y} = \sigma_{\mathsf{YV}}\mathsf{V} + \sqrt{1-\sigma_{\mathsf{YV}}^2}\mathsf{Z}_v,$$

where $\mathsf{Z}_u \sim \mathcal{N}(0,1)$, $\mathsf{Z}_v \sim \mathcal{N}(0,1)$, $\mathsf{Z}_u \perp \mathsf{U}$, $\mathsf{Z}_v \perp \mathsf{V}$. Due to Proposition 7, the mutual information for jointly Gaussian $(\mathsf{U},\mathsf{X},\mathsf{Y},\mathsf{V})$ is given by

$$\begin{aligned} I(\mathsf{U};\mathsf{V}) &= \mathbb{E}_{\mathsf{UV}}\left[\log\left(\sum_{n=0}^{\infty}\frac{\rho^n}{n!}\mathbb{E}[H_n(\mathsf{X})|\mathsf{U}]\mathbb{E}[H_n(\mathsf{Y})|\mathsf{V}]\right)\right] \\ &\overset{(a)}{=} \mathbb{E}_{\mathsf{UV}}\left[\log\left(\sum_{n=0}^{\infty}\frac{(\rho\sigma_{\mathsf{UX}}\sigma_{YV})^n}{n!}H_n(\mathsf{U})H_n(\mathsf{V})\right)\right] \\ &\overset{(b)}{=} \mathbb{E}_{\mathsf{UV}}\left[\log\left(\frac{1}{\sqrt{1-\rho^2\sigma^2_{\mathsf{UX}}\sigma^2_{\mathsf{YV}}}}\exp\left(\frac{2\rho\sigma_{\mathsf{UX}}\sigma_{\mathsf{YV}}\mathsf{UV}-\rho^2\sigma^2_{\mathsf{UX}}\sigma^2_{\mathsf{YV}}(\mathsf{U}^2+\mathsf{V}^2)}{2(1-\rho^2\sigma^2_{\mathsf{UX}}\sigma^2_{\mathsf{YV}})}\right)\right)\right] \\ &= -\frac{1}{2}\log(1-\rho^2\sigma^2_{\mathsf{UX}}\sigma^2_{\mathsf{YV}}) + \frac{\rho\sigma_{\mathsf{UX}}\sigma_{\mathsf{YV}}}{1-\rho^2\sigma^2_{\mathsf{UX}}\sigma^2_{\mathsf{YV}}}\mathbb{E}[\mathsf{UV}] - \frac{\rho^2\sigma^2_{\mathsf{UX}}\sigma^2_{\mathsf{YV}}}{2(1-\rho^2\sigma^2_{\mathsf{UX}}\sigma^2_{\mathsf{YV}})}\left(\mathbb{E}\left[\mathsf{U}^2\right]+\mathbb{E}\left[\mathsf{V}^2\right]\right) \\ &= -\frac{1}{2}\log(1-\rho^2\sigma^2_{\mathsf{UX}}\sigma^2_{\mathsf{YV}}), \end{aligned}$$

where (a) and (b) follow from the properties of Mehler Kernel [66].

By the Mutual Information constraints we have:

$$\sigma^2_{\mathsf{UX}} = 1 - e^{-2C_u} \qquad \sigma^2_{\mathsf{YV}} = 1 - e^{-2C_v}. \tag{A66}$$

Hence,

$$I(\mathsf{U};\mathsf{V}) = -\frac{1}{2}\log(1-\rho^2(1-e^{-2C_u})(1-e^{-2C_v})). \tag{A67}$$

Appendix I. Proof of Proposition 9

We choose U and V to be deterministic functions of X and Y, respectively, 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 (Chapter 5, p.107 of [68]), we have $1 - 2\mathsf{P}(\mathsf{X}<0,\mathsf{Y}<0) = \frac{\arccos\rho}{\pi}$. This completes the proof of the proposition.

Appendix J. Proof of Theorem 2

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

$$\frac{\mathrm{d}M}{\mathrm{d}\rho}\Big|_{\rho=0} = xy, \qquad \frac{\mathrm{d}^2M}{\mathrm{d}\rho^2}\Big|_{\rho=0} = (x^2-1)(y^2-1). \tag{A68}$$

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

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

and

$$\frac{\mathrm{d}^2 f}{\mathrm{d}\rho^2}\Big|_{\rho=0} = f_{\mathsf{U}}(u) f_{\mathsf{V}}(v) \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\mathsf{X}|\mathsf{U}}(x|u) \frac{\mathrm{d}^2 M(x,y;\rho)}{\mathrm{d}\rho^2}\Big|_{\rho=0} f_{\mathsf{Y}|\mathsf{V}}(y|v)\mathrm{d}x\mathrm{d}y \tag{A69}$$

$$= f_{\mathsf{U}}(u) f_{\mathsf{V}}(v)\left(\int_{-\infty}^{\infty}(x^2-1) f_{\mathsf{X}|\mathsf{U}}(x|u)\mathrm{d}x\right)\left(\int_{-\infty}^{\infty}(y^2-1) f_{\mathsf{Y}|\mathsf{V}}(y|v)\mathrm{d}y\right) \tag{A70}$$

$$= f_{\mathsf{U}}(u) f_{\mathsf{V}}(v)\left(\mathbb{E}[\mathsf{X}^2|\mathsf{U}=u]-1\right)\left(\mathbb{E}[\mathsf{Y}^2|\mathsf{V}=v]-1\right). \tag{A71}$$

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

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

Thus,

$$I(\rho) = \frac{\rho^2 \log e}{2}\mathbb{E}\left[(\mathbb{E}[\mathsf{X}|\mathsf{U}])^2\right]\mathbb{E}\left[(\mathbb{E}[\mathsf{Y}|\mathsf{V}])^2\right] + o(\rho^2). \tag{A72}$$

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

$$1 = \mathbb{E}[\mathsf{X}^2] = \mathbb{E}\Big[\mathbb{E}[\mathsf{X}^2|\mathsf{U}]\Big] = \mathbb{E}[\mathrm{var}[\mathsf{X}|\mathsf{U}]] + \mathbb{E}\left[(\mathbb{E}[\mathsf{X}|\mathsf{U}])^2\right]. \tag{A73}$$

In addition, by (Corollary to Theorem 8.6.6 of [69]), $\mathbb{E}[\mathrm{var}[\mathsf{X}|\mathsf{U}]] \geq \frac{1}{2\pi e}e^{2h(\mathsf{X}|\mathsf{U})}$. Moreover, from MI constraint, we have

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

and therefore $h(\mathsf{X}|\mathsf{U}) \geq \log(2\pi e) - C_u$. Thus, we obtain:

$$-C_u \leq \frac{1}{2}\log(\mathbb{E}[\mathrm{var}[\mathsf{X}|\mathsf{U}]]) \to \mathbb{E}[\mathrm{var}[\mathsf{X}|\mathsf{U}]] \geq 2^{-2C_u}. \tag{A74}$$

Combining (A73) and (A74), we obtain $\mathbb{E}\left[(\mathbb{E}[\mathsf{X}|\mathsf{U}])^2\right] \leq 1 - 2^{-2C_u}$.
In a very similar method, one can show that $\mathbb{E}\left[(\mathbb{E}[\mathsf{Y}|\mathsf{V}])^2\right] \leq 1 - 2^{-2C_v}$.
Thus, for $\rho \to 0$

$$I(\rho) \leq \frac{\rho^2 \log e}{2}(1-2^{-2C_u})(1-2^{-2C_v}) + o(\rho^2). \tag{A75}$$

Appendix K. Proof of Lemma A1

The function $\phi(p,\lambda)$ is a twice differentiable continuous function with respective second derivative given by

$$\frac{\partial^2\phi(p,\lambda)}{\partial p^2} = \phi_{pp}(p,\lambda) = -\frac{(a-b)^2}{ap+b\bar{p}} - \frac{(c-d)^2}{cp+d\bar{p}} - \frac{(a-b+c-d)^2}{1-(a+c)p-(b+d)\bar{p}} + \frac{\lambda}{p\bar{p}}. \tag{A76}$$

The former can also be written as a proper rational function [70], i.e., $\phi_{pp}(p,\lambda) = \frac{N(p)}{D(p)}$, where

$$\begin{aligned}N(p) = \lambda(ap+b\bar{p})(cp+d\bar{p})(1-(a+c)p-(b+d)\bar{p}) - (a-b)^2(cp+d\bar{p})(1-(a+c)p\\ -(b+d)\bar{p})p\bar{p} - (c-d)^2(ap+b\bar{p})(1-(a+c)p-(b+d)\bar{p})p\bar{p}\\ -(a-b+c-d)^2(ap+b\bar{p})(cp+d\bar{p})p\bar{p},\end{aligned} \tag{A77}$$

and

$$D(p) = p\bar{p}(ap+b\bar{p})(cp+d\bar{p})(1-(a+c)p-(b+d)\bar{p}). \tag{A78}$$

Note that $\phi_{pp}(p,\lambda)$ equals $+\infty$ for $p \in \{0,1\}$ and hence is positive for this set of points.

1. Suppose $\phi(p,\lambda)$ is linear over some interval $\mathcal{I} \subset [a,b]$. In such case, its second derivative must be zero over this interval, which implies that $N(p)$ is zero over this interval. Since $N(p)$ is a degree 3 polynomial, it can be zero over some interval if and only if it is zero everywhere. Thus, if $\phi(p,\lambda)$ is linear over some interval $\mathcal{I}$, then it is non-linear for every $p \in [0,1]$.
2. For $p \in (0,1)$, $D(p) > 0$ and $N(p)$ is a degree 3 polynomial in p. Since $N(0^+) > 0$ and $N(1^-) > 0$, this polynomial has no sign changes or has exactly two sign changes in $(0,1)$. Therefore, either $\phi(p,\lambda)$ is convex or there are two points p_1 and p_2, $0 < p_1 < p_2 < 1$, such that $\phi(p,\lambda)$ is convex in $p \in [0,p_1] \cup [p_2,1]$ and concave in $p \in [p_1,p_2]$.

Appendix L. Proof of Lemma A2

Let $\mathcal{I}_2 = [c,d] \subset [0,1]$ and assume in contradiction that $\{\alpha_i, p_i\}_{i=1,2,3}$ attains the lower convex envelope at point q, and that $p_2 \in \mathcal{I}_2$. By assumption, we have that

$$\alpha_1 p_1 + \alpha_2 p_2 + \alpha_3 p_3 = q. \tag{A79}$$

We can write $p_2 = \bar{\gamma} c + \gamma d$ for some $\gamma \in (0,1)$ and still

$$\alpha_1 p_1 + \alpha_2 \bar{\gamma} c + \alpha_2 \gamma d + \alpha_3 p_3 = q. \tag{A80}$$

However, due to concavity of $\phi(\cdot)$ in $\mathcal{I}_2$, we must have

$$\begin{aligned}\alpha_1\phi(p_1) + \alpha_2\bar{\gamma}\phi(c) + \alpha_2\gamma\phi(d) + \alpha_3\phi(p_3) &\le \alpha_1\phi(p_1) + \alpha_2\phi(\bar{\gamma}c + \gamma d) + \alpha_3\phi(p_3) \\ &= \alpha_1\phi(p_1) + \alpha_2\phi(p_2) + \alpha_3\phi(p_3).\end{aligned} \tag{A81}$$

This implies that there is a linear combination of point from $\mathcal{I}_1 \cup \mathcal{I}_3$ that attains a lower value than $\phi(q)$, contradicting the assumption that $\phi(q)$ is the lower convex envelope at point q. Since p_2 was arbitrary, the lemma holds.

Appendix M. Proof of Lemma A3

Assume in contradiction that there are no distinct points, i.e., it has $p_{11} = p_{12} = p_{13} = p_{21} = p_{22} = p_{23} = p$, then $p = q$ and $x_1 = x_2$,which contradicts the initial assumption that $x_1 \neq x_2$. Assume WOLG that $p_{11} = p_{12} = p_{13} = p_{21} = p_{22} = p$ but $p_{23} \neq p$. Since $p_{11} = p_{12} = p_{13} = p$ implies $p = q$, then p_{23} must be q as well in contradiction to the initial assumption.

Consider the following cases:

- $p_{11} = p_{12} = p_{13} = p_{21} = p_1, p_{22} = p_{23} = p_2$, $p_1 \neq p_2$: This implies $p_1 = q$. Furthermore,

$$\alpha_{21} q + \alpha_{22} p_2 + (1 - \alpha_{21} - \alpha_{22}) p_2 = q \rightarrow (1 - \alpha_{21}) p_2 = (1 - \alpha_{21}) q, \tag{A82}$$

which holds only if $p_2 = q$ in contradiction to our initial assumption.

- $p_{11} = p_{12} = p_{21} = p_{22} = p_1, p_{13} = p_{23} = p_2$, $p_1 \neq p_2$: This implies

$$(\alpha_{11} + \alpha_{12}) p_1 + (1 - \alpha_{21} - \alpha_{22}) p_2 = q = (\alpha_{21} + \alpha_{22}) p_1 + (1 - \alpha_{21} - \alpha_{22}) p_2, \tag{A83}$$

which holds only if $\alpha_{11} + \alpha_{12} = \alpha_{21} + \alpha_{22}$. In such case

$$\begin{aligned}x_1 &= (\alpha_{11} + \alpha_{12}) h(p_1) + (1 - \alpha_{11} - \alpha_{12}) h(p_2) \\ &= (\alpha_{21} + \alpha_{22}) h(p_1) + (1 - \alpha_{21} - \alpha_{22}) h(p_2) = x_2,\end{aligned} \tag{A84}$$

in contradiction to the assumption $x_1 \neq x_2$.

Thus, the lemma holds.

References

1. Tishby, N.; Pereira, F.C.N.; Bialek, W. The information bottleneck method. In Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing, Monticello, IL, USA, 22–24 September 1999; pp. 368–377.
2. Pichler, G.; Piantanida, P.; Matz, G. Distributed information-theoretic clustering. *Inf. Inference J. Ima* **2021**, *11*, 137–166. [CrossRef]
3. Jain, A.K.; Dubes, R.C. *Algorithms for Clustering Data*; Prentice-Hall: Hoboken, NJ, USA, 1988.
4. Gupta, N.; Aggarwal, S. Modeling Biclustering as an optimization problem using Mutual Information. In Proceedings of the International Conference on Methods and Models in Computer Science (ICM2CS), Delhi, India, 14–15 December 2009 ; pp. 1–5.
5. Hartigan, J. Direct Clustering of a Data Matrix. *J. Am. Stat. Assoc.* **1972**, *67*, 123–129. [CrossRef]
6. Madeira, S.; Oliveira, A. Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2004**, *1*, 24–45. [CrossRef] [PubMed]
7. Dhillon, I.S.; Mallela, S.; Modha, D.S. Information-Theoretic Co-Clustering. In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003, (KDD '03), Washington, DC, USA, 24–27 August 2003; pp. 89–98.
8. Courtade, T.A.; Kumar, G.R. Which Boolean Functions Maximize Mutual Information on Noisy Inputs? *IEEE Trans. Inf. Theory* **2014**, *60*, 4515–4525. [CrossRef]
9. Han, T.S. Hypothesis Testing with Multiterminal Data Compression. *IEEE Trans. Inf. Theory* **1987**, *33*, 759–772. [CrossRef]
10. Westover, M.B.; O'Sullivan, J.A. Achievable Rates for Pattern Recognition. *IEEE Trans. Inf. Theory* **2008**, *54*, 299–320. [CrossRef]
11. Painsky, A.; Feder, M.; Tishby, N. An Information-Theoretic Framework for Non-linear Canonical Correlation Analysis. *arXiv* **2018**, arXiv:1810.13259.
12. Williamson, A.R. The Impacts of Additive Noise and 1-bit Quantization on the Correlation Coefficient in the Low-SNR Regime. In Proceedings of the 57th Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, USA, 24–27 September 2019; pp. 631–638.
13. Courtade, T.A.; Weissman, T. Multiterminal Source Coding Under Logarithmic Loss. *IEEE Trans. Inf. Theory* **2014**, *60*, 740–761. [CrossRef]
14. Pichler, G.; Piantanida, P.; Matz, G. Dictator Functions Maximize Mutual Information. *Ann. Appl. Prob.* **2018**, *28*, 3094–3101. [CrossRef]
15. Dobrushin, R.; Tsybakov, B. Information transmission with additional noise. *IRE Trans. Inf. Theory* **1962**, *8*, 293–304. [CrossRef]
16. Wolf, J.; Ziv, J. Transmission of noisy information to a noisy receiver with minimum distortion. *IEEE Trans. Inf. Theory* **1970**, *16*, 406–411. [CrossRef]
17. Witsenhausen, H.S.; Wyner, A.D. A Conditional Entropy Bound for a Pair of Discrete Random Variables. *IEEE Trans. Inf. Theory* **1975**, *21*, 493–501. [CrossRef]
18. Arimoto, S. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Trans. Inf. Theory* **1972**, *18*, 14–20. [CrossRef]
19. Blahut, R. Computation of channel capacity and rate-distortion functions. *IEEE Trans. Inf. Theory* **1972**, *18*, 460–473. [CrossRef]
20. Aguerri, I.E.; Zaidi, A. Distributed Variational Representation Learning. *IEEE Trans. Pattern Anal.* **2021**, *43*, 120–138. [CrossRef]
21. Hassanpour, S.; Wuebben, D.; Dekorsy, A. Overview and Investigation of Algorithms for the Information Bottleneck Method. In Proceedings of the SCC 2017: 11th International ITG Conference on Systems, Communications and Coding, Hamburg, Germany, 6–9 February 2017; pp. 1–6.
22. Slonim, N. The Information Bottleneck: Theory and Applications. Ph.D. Thesis, Hebrew University of Jerusalem, Jerusalem, Israel, 2002.
23. Sutskover, I.; Shamai, S.; Ziv, J. Extremes of information combining. *IEEE Trans. Inf. Theory* **2005**, *51*, 1313–1325. [CrossRef]
24. Zaidi, A.; Aguerri, I.E.; Shamai, S. On the Information Bottleneck Problems: Models, Connections, Applications and Information Theoretic Views. *Entropy* **2020**, *22*, 151. [CrossRef]
25. Wyner, A.; Ziv, J. A theorem on the entropy of certain binary sequences and applications–I. *IEEE Trans. Inf. Theory* **1973**, *19*, 769–772. [CrossRef]
26. Chechik, G.; Globerson, A.; Tishby, N.; Weiss, Y. Information Bottleneck for Gaussian Variables. *J. Mach. Learn. Res.* **2005**, *6*, 165–188.
27. Blachman, N. The convolution inequality for entropy powers. *IEEE Trans. Inf. Theory* **1965**, *11*, 267–271. [CrossRef]
28. Guo, D.; Shamai, S.; Verdú, S. The interplay between information and estimation measures. *Found. Trends Signal Process.* **2013**, *6*, 243–429. [CrossRef]
29. Bustin, R.; Payaro, M.; Palomar, D.P.; Shamai, S. On MMSE Crossing Properties and Implications in Parallel Vector Gaussian Channels. *IEEE Trans. Inf. Theory* **2013**, *59*, 818–844. [CrossRef]
30. Sanderovich, A.; Shamai, S.; Steinberg, Y.; Kramer, G. Communication Via Decentralized Processing. *IEEE Trans. Inf. Theory* **2008**, *54*, 3008–3023. [CrossRef]
31. Smith, J.G. The information capacity of amplitude-and variance-constrained scalar Gaussian channels. *Inf. Control.* **1971**, *18*, 203–219. [CrossRef]
32. Sharma, N.; Shamai, S. Transition points in the capacity-achieving distribution for the peak-power limited AWGN and free-space optical intensity channels. *Probl. Inf. Transm.* **2010**, *46*, 283–299. [CrossRef]

33. Dytso, A.; Yagli, S.; Poor, H.V.; Shamai, S. The Capacity Achieving Distribution for the Amplitude Constrained Additive Gaussian Channel: An Upper Bound on the Number of Mass Points. *IEEE Trans. Inf. Theory* **2019**, *66*, 2006–2022. [CrossRef]
34. Steinberg, Y. Coding and Common Reconstruction. *IEEE Trans. Inf. Theory* **2009**, *55*, 4995–5010. [CrossRef]
35. Land, I.; Huber, J. Information Combining. *Found. Trends Commun. Inf. Theory* **2006**, *3*, 227–330. [CrossRef]
36. Yang, Q.; Piantanida, P.; Gündüz, D. The Multi-layer Information Bottleneck Problem. In Proceedings of the IEEE Information Theory Workshop (ITW), Kaohsiung, Taiwan, 6–10 November 2017; pp. 404–408.
37. Berger, T.; Zhang, Z.; Viswanathan, H. The CEO Problem. *IEEE Trans. Inf. Theory* **1996**, *42*, 887–902. [CrossRef]
38. Steiner, S.; Kuehn, V. Optimization Of Distributed Quantizers Using An Alternating Information Bottleneck Approach. In Proceedings of the WSA 2019: 23rd International ITG Workshop on Smart Antennas, Vienna, Austria, 24–26 April 2019; pp. 1–6.
39. Vera, M.; Rey Vega, L.; Piantanida, P. Collaborative Information Bottleneck. *IEEE Trans. Inf. Theory* **2019**, *65*, 787–815. [CrossRef]
40. Ugur, Y.; Aguerri, I.E.; Zaidi, A. Vector Gaussian CEO Problem Under Logarithmic Loss and Applications. *IEEE Trans. Inf. Theory* **2020**, *66*, 4183–4202. [CrossRef]
41. Estella, I.; Zaidi, A. Distributed Information Bottleneck Method for Discrete and Gaussian Sources. In Proceedings of the International Zurich Seminar on Information and Communication (IZS), Zurich, Switzerland, 21–23 February 2018; pp. 35–39.
42. Courtade, T.A.; Jiao, J. An Extremal Inequality for Long Markov Chains. In Proceedings of the 52nd Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, USA, 1–3 October 2014; pp. 763–770.
43. Erkip, E.; Cover, T.M. The Efficiency of Investment Information. *IEEE Trans. Inf. Theory* **1998**, *44*, 1026–1040. [CrossRef]
44. Gács, P.; Körner, J. Common information is far less than mutual information. *Probl. Contr. Inform. Theory* **1973**, *2*, 149–162.
45. Farajiparvar, P.; Beirami, A.; Nokleby, M. Information Bottleneck Methods for Distributed Learning. In Proceedings of the 56th Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, USA, 2–5 October 2018; pp. 24–31.
46. Tishby, N.; Zaslavsky, N. Deep Learning and the Information Bottleneck Principle. In Proceedings of the Information Theory Workshop (ITW), Jeju Island, Korea, 11–15 October 2015; pp. 1–5.
47. Alemi, A.; Fischer, I.; Dillon, J.; Murphy, K. Deep Variational Information Bottleneck. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
48. Shwartz-Ziv, R.; Tishby, N. Opening the black box of deep neural networks via information. *arXiv* **2017**, arXiv:1703.00810.
49. Gabrié, M.; Manoel, A.; Luneau, C.; Barbier, j.; Macris, N.; Krzakala, F.; Zdeborová, L. Entropy and mutual information in models of deep neural networks. In *Advances in NIPS*; Curran Associates, Inc.: Red Hook, NY, USA, 2018; Volume 31.
50. Goldfeld, Z.; van den Berg, E.; Greenewald, K.H.; Melnyk, I.; Nguyen, N.; Kingsbury, B.; Polyanskiy, Y. Estimating Information Flow in Neural Networks. *arXiv* **2018**, arXiv:1810.05728.
51. Amjad, R.A.; Geiger, B.C. Learning Representations for Neural Network-Based Classification Using the Information Bottleneck Principle. *IEEE Trans. Pattern Anal.* **2020**, *42*, 2225–2239. [CrossRef]
52. Saxe, A.M.; Bansal, Y.; Dapello, J.; Advani, M.; Kolchinsky, A.; Tracey, B.D.; Cox, D.D. On the information bottleneck theory of deep learning. *J. Stat. Mech. Theory Exp.* **2019**, *2019*, 1–34. [CrossRef]
53. Cheng, H.; Lian, D.; Gao, S.; Geng, Y. Evaluating Capability of Deep Neural Networks for Image Classification via Information Plane. In *Lecture Notes in Computer Science, Proceedings of the Computer Vision-ECCV 2018-15th European Conference, Munich, Germany, 8–14 September 2018, Proceedings, Part XI*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer: Berlin/Heidelberg, Germany, 2018; Volume 11215, pp. 181–195.
54. Yu, S.; Wickstrøm, K.; Jenssen, R.; Príncipe, J.C. Understanding Convolutional Neural Networks with Information Theory: An Initial Exploration. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 435–442. [CrossRef]
55. Lewandowsky, J.; Stark, M.; Bauch, G. Information bottleneck graphs for receiver design. In Proceedings of the IEEE International Symposium on Information Theory, Barcelona, Spain, 10–15 July 2016; pp. 2888–2892.
56. Stark, M.; Wang, L.; Bauch, G.; Wesel, R.D. Decoding rate-compatible 5G-LDPC codes with coarse quantization using the information bottleneck method. *IEEE Open J. Commun. Soc.* **2020**, *1*, 646–660. [CrossRef]
57. Bhatt, A.; Nazer, B.; Ordentlich, O.; Polyanskiy, Y. Information-distilling quantizers. *IEEE Trans. Inf. Theory* **2021**, *67*, 2472–2487. [CrossRef]
58. Stark, M.; Shah, A.; Bauch, G. Polar code construction using the information bottleneck method. In Proceedings of the 2018 IEEE Wireless Communications and Networking Conference Workshops (WCNCW), Barcelona, Spain, 15–18 April 2018; pp. 7–12.
59. Shah, S.A.A.; Stark, M.; Bauch, G. Design of Quantized Decoders for Polar Codes using the Information Bottleneck Method. In Proceedings of the SCC 2019: 12th International ITG Conference on Systems, Communications and Coding, Rostock, Germany, 11–14 February 2019; pp. 1–6.
60. Shah, S.A.A.; Stark, M.; Bauch, G. Coarsely Quantized Decoding and Construction of Polar Codes Using the Information Bottleneck Method. *Algorithms* **2019**, *12*, 192. [CrossRef]
61. Kurkoski, B.M. On the Relationship Between the KL Means Algorithm and the Information Bottleneck Method. In Proceedings of the 11th International ITG Conference on Systems, Communications and Coding (SCC), Hamburg, Germany, 6–9 February 2017; pp. 1–6.
62. Goldfeld, Z.; Polyanskiy, Y. The Information Bottleneck Problem and its Applications in Machine Learning. *IEEE J. Sel. Areas Inf. Theory* **2020**, *1*, 19–38. [CrossRef]
63. Harremoes, P.; Tishby, N. The Information Bottleneck Revisited or How to Choose a Good Distortion Measure. In Proceedings of the 2007 IEEE International Symposium on Information Theory, Nice, France, 24–29 June 2007; pp. 566–570.

64. Richardson, T.; Urbanke, R. *Modern Coding Theory*; Cambridge University Press: Cambridge, UK, 2008.
65. Sason, I. On f-divergences: Integral representations, local behavior, and inequalities. *Entropy* **2018**, *20*, 383. [CrossRef] [PubMed]
66. Mehler, F.G. Ueber die Entwicklung einer Function von beliebig vielen Variablen nach Laplaceschen Functionen höherer Ordnung. *J. Reine Angew. Math.* **1866**, *66*, 161–176.
67. Lancaster, H.O. The Structure of Bivariate Distributions. *Ann. Math. Statist.* **1958**, *29*, 719–736. [CrossRef]
68. O'Donnell, R. *Analysis of Boolean Functions*, 1st ed.; Cambridge University Press: New York, NY, USA, 2014.
69. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley: Hoboken, NJ, USA, 2006.
70. Corless, M.J. *Linear Systems and Control : An Operator Perspective*; Monographs and Textbooks in Pure and Applied Mathematics; Marcel Dekker: New York, NY, USA, 2003; Volume 254.