The Compound Information Bottleneck Program

Michael Dikshtein, Nir Weinberger, and Shlomo Shamai (Shitz)
Department of Electrical and Computer Engineering, Technion, Haifa 3200003, Israel
Email: {michaeldic@campus., nirwein@, sshlomo@ee.}technion.ac.il

2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834812

***Abstract*—Motivated by the emerging technology of oblivious processing in remote radio heads with universal decoders, we formulate and analyze in this paper a compound version of the *information bottleneck* problem. In this problem, a Markov chain $\mathsf{X} \to \mathsf{Y} \to Z$ is assumed, and the marginals P_{X} and P_{Y} are set. The mutual information between X and Z is sought to be maximized over the choice of the conditional probability of Z given Y from a given class, under the worst choice of the joint probability of the pair (X,Y) from a different class. We provide values, bounds, and various characterizations for specific instances of this problem: the binary symmetric case, the scalar Gaussian case, the vector Gaussian case, the symmetric modulo-additive case, and the total variation constraints case. Finally, for the general case, we propose a Blahut-Arimoto type of alternating iterations algorithm to find a consistent solution to this problem.**

I. Introduction and Problem Formulation

The *information bottleneck* (IB) methodology [1] provides a universal distortion measure for data compression when the desired distortion measure is either unavailable or cannot be defined. Nonetheless, in most practical cases, the distribution of the source involved in the IB problem is also not known with perfect accuracy (e.g., when it is estimated from a finite sample). In this paper, this aspect motivates us to introduce a *compound* version of the IB problem, in which the source distribution is only known to belong to a given class, and the representation chosen by the IB method is chosen to be the best possible under the worst-case choice within the class.

We consider the compound remote source coding system [2]–[4]. Let P_{X} be a source of information generating the sequence X^n. The encoder observes Y^n which is a noisy version of X^n. Then, the encoder produces a compressed representation M, which is later on mapped by the decoder to the reconstructed sequence Z^n. The distortion is evaluated between X^n and Z^n, while the rate is the relative number of bits required to represent M. The encoder's goal is to find a compression strategy that extracts from Y^n the relevant information regarding X^n, when the distribution of the channel $\mathsf{P}_{\mathsf{Y}|\mathsf{X}}$ is not known in advance and cannot be accurately learned. This compound setting generalizes the classical remote source coding model studied by Dobrushin and Wolf [5], [6].

This model motivates one to formulate a *compound* version of the *information bottleneck* (IB) optimization problem [1]. Specifically, let (X,Y) be a pair of random variables and fix their marginals to P_{X} and P_{Y}, respectively. Consider all random variables Z satisfying the Markov chain $\mathsf{X} \to \mathsf{Y} \to \mathsf{Z}$. Unlike the standard IB problem, in which the joint distribution of P_{XY} is fixed, here we consider an uncertainty set for this joint distribution, and aim to solve the following problem:

$$R^{\mathrm{com}}_{\mathsf{P}_{\mathsf{X}}\mathsf{P}_{\mathsf{Y}}}(\mathcal{P}_{\mathcal{XY}},\mathcal{D}_{\mathcal{Z}|\mathcal{Y}}) = \sup_{\mathsf{P}_{\mathsf{Z}|\mathsf{Y}}\in\mathcal{D}_{\mathcal{Z}|\mathcal{Y}}} \inf_{\mathsf{P}_{\mathsf{XY}}\in\mathcal{P}_{\mathcal{XY}}} I(\mathsf{X};\mathsf{Z}). \quad (1)$$

In this paper, we take the set $\mathcal{D}_{\mathcal{Z}|\mathcal{Y}}$ as the set of possible representations, and the set $\mathcal{P}_{\mathcal{XY}}$ is the uncertainty set of the joint distribution. The class $\mathcal{D}_{\mathcal{Z}|\mathcal{Y}}$ will be the usual IB class, i.e., $\mathcal{D}_{\mathcal{Z}|\mathcal{Y}} = \left\{\mathsf{P}_{\mathsf{Z}|\mathsf{Y}} \colon I(\mathsf{Y};\mathsf{Z}) \le C_2\right\}$, or a restricted subset of this class, with an additional structure. The class $\mathcal{P}_{\mathcal{XY}}$ will take one of the following variants:

- *Privacy Funnel* (PF): $\mathcal{P}_{\mathcal{XY}} = \{\mathsf{P}_{\mathsf{XY}} \colon I(\mathsf{X};\mathsf{Y}) \ge C_1\}$ as motivated by capacity-guaranteed links;
- *Minimal Correlation*: $\mathcal{P}_{\mathcal{XY}} = \{\mathsf{P}_{\mathsf{XY}} \colon \mathbb{E}\left[\mathsf{XY}\right] \ge \rho_1\}$ as motivated by the Gaussian setting;
- *Total Variation* (TV): $\mathcal{P}_{\mathcal{XY}} = \{\mathsf{P}_{\mathsf{XY}} \colon d_{\mathrm{TV}}(\mathsf{P}_{\mathsf{XY}},\mathsf{P}_1) \le D_1\}$ as motivated by finite data samples analysis, and P_1 is some nominal distribution;

where all are constrained to the given marginals, i.e., $\sum_x \mathsf{P}_{\mathsf{XY}}(x,y) = \mathsf{P}_{\mathsf{Y}}(y)$ and $\sum_y \mathsf{P}_{\mathsf{XY}}(x,y) = \mathsf{P}_{\mathsf{X}}(x)$. For the above sets of optimization, we have $\max\min$ in (1).

As said, choosing the class $\mathcal{P}_{\mathcal{XY}}$ to a singleton recovers the standard IB problem [1], which for discrete alphabets was initially studied in [7] as a method to characterize common information [8]. The IB method is essentially a remote source coding problem [5], [6], choosing the distortion measure as the logarithmic-loss, and thus recovers remote source coding by taking $\mathcal{D}_{\mathcal{Z}|\mathcal{Y}}$ as a maximal distortion set.

In addition, PF, a dual problem to the IB framework [9], [10], can also be recovered from (1) by setting $\mathcal{P}_{\mathcal{XY}}$ as the PF family (removing the marginalization constraint on P_{X}) and $\mathcal{D}_{\mathcal{Z}|\mathcal{Y}}$ to contain a singleton. Therefore, the problem introduced in (1) is a composition of the IB and PF problems. This observation makes the problem in (1) rather delicate – e.g., if (X,Y) are jointly Gaussian, even the standard PF rate is zero since one can use the channel from Y to Z to describe the less significant bits of Y [11]. We also mention that the PF is directly connected to *information combining* (IC) [12], [13]. For example, if the channel from Y to X is a *binary memoryless symmetric* (BMS) [14, Ch. 4], then by [12], $\mathsf{P}_{\mathsf{Z}|\mathsf{Y}}$ is a *binary erasure channel* (BEC). Furthermore, the additive noise Helper problem, studied in [15], is directly linked to the PF. By reformulating the former as an IC problem, the solution follows directly, as was demonstrated in [11].

The IB problem can be approached via several strategies. When (X,Y) is a *doubly symmetric binary source* (DSBS) with transition probability p [16], it can be shown that binary symmetric channels are optimal via Mrs. Gerber's lemma [17] (see also the examples in [7] and [12]). When (X,Y) are jointly multivariate Gaussians, it was shown in [18] that the optimal distribution of $(\mathsf{X},\mathsf{Y},\mathsf{Z})$ is also jointly Gaussian. The optimality of the Gaussian test channel can also be proved using EPI or utilizing I-MMSE and Single Crossing Property [19], [20]. In a different and more general case, when $(\mathsf{X},\mathsf{Y},\mathsf{Z})$ are discrete random variables, a locally optimal $\mathsf{P}_{\mathsf{Z}|\mathsf{Y}}$ can be found by iteratively solving a set of self–consistent equations. A generalized Blahuto-Arimoto algorithm was proposed to

solve those equations [1], [21]–[23]. Finally, a particular case of deterministic mappings from X to Y was considered in [24].

In this work, we address the compound setting for the IB problem with the goal of providing similar results. First, we address the DSBS and Gaussian (scalar and vector) settings. Second, we consider general modulo additive channels, with modulo additive representations, and provide various bounds on the compound IB function with PF-based compound set and then with TV-based compound set, and again derive various bounds on the compound IB function. Finally, we return to the general discrete alphabet case with a PF-based compound set and propose an alternating algorithm, which essentially iterates between the maximization over $\mathsf{P}_{\mathsf{Z}|\mathsf{Y}}$ (an IB problem) and minimization over P_{XY} (a PF problem). Omitted proofs and other details are in the full version of this paper [25].

Related work: The IB framework is closely related to a variety of problems in information theory, such as *remote source coding* [6], *conditional entropy bound* (CEB) [7], *common reconstruction* [26], and *information combining* (IC) [12], [13], see an overview in [16]. Applications of information bottleneck method in machine learning are detailed in [18], [27]–[32]. Furthermore, the IB problem connects to many timely aspects, such as *capital investment* [33], *distributed learning* [29], *deep learning* [27], [28], [30], [31], [34]–[36] and *convolutional neural networks* [37], [38].

II. Binary Symmetric and Gaussian channels

A simple way to obtain precise analytical solutions to (1) is by establishing a saddle point property [39, Sec. 5.4.2].

Lemma 1 (Optimality of Saddle Point): suppose there exists a saddle point $(\tilde{w},\tilde{z})$, satisfying $f(\tilde{w},\tilde{z}) = \inf_{w\in\mathcal{W}} f(w,\tilde{z})$ and $f(\tilde{w},\tilde{z}) = \sup_{z\in\mathcal{Z}} f(\tilde{w},z)$, then $f(\tilde{w},\tilde{z}) = \sup_{z\in\mathcal{Z}}\inf_{w\in\mathcal{W}} f(w,z)$.

In the rest of this section we provide basic examples for which full characterization of the problem in (1) is known.

A. Binary Y

Consider X and Y being both $\mathrm{Ber}(0.5)$ random variables with PF type of $\mathcal{P}_{\mathcal{X}\mathcal{Y}}$, and $C_1, C_2 \in [0, \log 2]$. Let $R^{\text{bin}}(C_1, C_2)$ denote the compound IB with a PF constraint for this setting. In such case, (X,Y) are restricted to be distributed as a DSBS with parameter α, i.e.,

$$\mathsf{P}_{\mathsf{XY}}(x,y) = \frac{1}{2}(\alpha\cdot\mathbb{1}(x\neq y) + (1-\alpha)\mathbb{1}(x=y)), \quad (2)$$

where $\alpha = h_b^{-1}(1-C_1)$, with $h_b(\cdot)$ being the binary entropy function and $h_b^{-1}(\cdot)$ its inverse. Furthermore, the optimal $\mathsf{P}_{\mathsf{Z}|\mathsf{Y}}$ in this case is a BSC with parameter $\beta = h_b^{-1}(1-C_2)$ [7]. The compound rate is thus $R^{\text{bin}}(C_1,C_2) = 1 - h_b(\alpha * \beta)$, where $*$ is the binary convolution operator.

Next, assume that Y is $\mathrm{Ber}(0.5)$, but there are no constraints on X nor Z. In such case the optimal $\mathsf{P}_{\mathsf{Z}|\mathsf{Y}}$ is a BSC with parameter $\delta = h_b^{-1}(1-C_2)$, while the optimal $\mathsf{P}_{\mathsf{X}|\mathsf{Y}}$ is a BEC with parameter $\epsilon = 1 - C_1$. The optimal rate in such case is $R^{\text{bin}}(C_1,C_2) = C_1\cdot C_2$. This result can be established by combining [7, IV.C] with [12, Thm. 1] and Lemma 1.

B. Scalar Gaussian Y

We proceed to consider another fundamental scenario where the marginal distributions of X and Y are both Gaussian. Note that in contrast to the symmetric $\mathrm{Ber}(0.5)$ setting, which restricts the channel from X to Y being a BSC, here, Gaussianity of the marginals does not imply the joint distribution of (X,Y) being Gaussian [40, Ch. 4.7]. Thus, the result of the following theorem is not trivial. Let $R^{\text{sc-G}}(\rho,C)$ denote the value of (1) with $\mathcal{P}_{\mathcal{X}\mathcal{Y}}$ being the minimum correlation class with parameter $\rho>0$ and $\mathcal{Q}_{\mathcal{Z}|\mathcal{Y}}$ being the IB bottleneck class with parameter $C\in\mathbb{R}$.

Theorem 1: $R^{\text{sc-G}}(\rho,C) = -\frac{1}{2}\log(1-\rho^2\rho_C^2)$, with $\rho_C^2 = 1-2^{-2C}$, and jointly Gaussian $(\mathsf{X},\mathsf{Y},\mathsf{Z})$ is the unique optimizer of (1).

C. Vector Gaussian $(\mathbf{X},\mathbf{Y})$

Now, suppose that $\mathbf{X}$ and $\mathbf{Y}$ are jointly Gaussian random vectors of dimension n. Let $R^{\text{vec-G}}(C_1,C_2)$ denote the value of (1) with $\mathcal{P}_{\mathcal{X}\mathcal{Y}}$ being the PF constraint with capacity $C_1\in\mathbb{R}$ and $\mathcal{Q}_{\mathcal{Z}|\mathcal{Y}}$ is the IB bottleneck class with capacity $C_2\in\mathbb{R}$.

Theorem 2:

$$R^{\text{vec-G}}(C_1,C_2) = -\frac{n}{2}\log(1-\rho_1^2\rho_2^2), \quad (3)$$

where $\rho_k^2 = 1-2^{-2C_k/n}$ for $k\in\{1,2\}$. The optimal triplet $(\mathbf{X},\mathbf{Y},\mathbf{Z})$ is jointly Gaussian with independent components. In particular, this result establishes that the worst case channel $\mathsf{P}_{\mathsf{Y}|\mathsf{X}}$ is an Additive White Gaussian Noise, and its optimal representation $\mathsf{P}_{\mathsf{Z}|\mathsf{Y}}$ is also white.

III. Modulo Additive Channels with PF Constraint

In this section, we return to the general discrete alphabet case, yet we restrict our attention to a symmetric setting with the following assumptions:

$$\mathcal{P}_{\mathcal{X}\mathcal{Y}} \triangleq \{\mathsf{P}_{\mathsf{XY}}\colon \mathsf{P}_{\mathsf{X}} = \mathbf{u}_n, \mathsf{Y} = \mathsf{X}\oplus\mathsf{W}, H(\mathsf{W})\le\eta_1, \mathsf{X}\perp\mathsf{W}\}, \quad (4)$$

$$\mathcal{Q}_{\mathcal{Z}|\mathcal{Y}} \triangleq \{\mathsf{P}_{\mathsf{Z}|\mathsf{Y}}\colon \mathsf{Z} = \mathsf{Y}\oplus\mathsf{V}, H(\mathsf{V})\ge\eta_2, \mathsf{Y}\perp\mathsf{V}\}, \quad (5)$$

where $\mathbf{u}_n$ is the probability vector of uniform distribution on n, and $\perp$ stands for statistical independence. This setting implies $|\mathcal{Y}| = |\mathcal{Z}| = n$. Moreover, it also holds that $\mathsf{Z} = \mathsf{X}\oplus\mathsf{W}\oplus\mathsf{V}$. Using $H(\mathsf{W}) \equiv H(\mathsf{P}_{\mathsf{W}})$ and $H(\mathsf{V}) \equiv H(\mathsf{P}_{\mathsf{V}})$, we observe that $I(\mathsf{X};\mathsf{Z}) = \log n - H(\mathsf{P}_{\mathsf{W}} * \mathsf{P}_{\mathsf{V}})$, where $*$ is the n-ary convolution operator, and $\eta_1,\eta_2\in[0,\log n]$. Thus, the solution to (1) is equivalent to the solution of

$$R^{\text{mod}}(\eta_1,\eta_2) \triangleq \min_{\mathsf{P}_{\mathsf{V}}\colon H(\mathsf{P}_{\mathsf{V}})\ge\eta_2}\ \max_{\mathsf{P}_{\mathsf{W}}\colon H(\mathsf{P}_{\mathsf{W}})\le\eta_1} H(\mathsf{P}_{\mathsf{W}} * \mathsf{P}_{\mathsf{V}}). \quad (6)$$

In (4) we have confined the channel $\mathsf{P}_{\mathsf{Z}|\mathsf{Y}}$ to be modulo additive, which may be too restrictive in general. Nonetheless, when the IB function is *strictly* convex, the modulo additive channel assumption for $\mathcal{Q}_{\mathcal{Z}|\mathcal{Y}}$ can be relaxed. Indeed:

Proposition 1: Fix a joint pmf $\mathsf{P}_{\mathsf{XY}}\in\mathcal{P}_{\mathcal{X}\mathcal{Y}}$, where $\mathcal{P}_{\mathcal{X}\mathcal{Y}}$ defined in (4). Denote by T the transition probability matrix from Y to X. Assume that function $R_T^{\text{CEB}}(\eta)$ defined by

$$R_T^{\text{CEB}}(\eta) \triangleq \min_{\mathsf{P}_{\mathsf{Z}|\mathsf{Y}}\colon\ H(\mathsf{Y}|\mathsf{Z})\ge\eta} H(\mathsf{X}|\mathsf{Z}), \quad (7)$$

is a strictly convex function of η, then it is equivalent to the following problem:

$$g_T(\eta) \triangleq \min_{\mathbf{p}\in\Delta_n:\, h_n(\mathbf{p})\geq\eta} h_n(T\mathbf{p}), \tag{8}$$

where Δ_n is the n-dimensional simplex, and the optimal channel from Y to Z is also a modulo additive channel.
Thus, if the strict convexity holds then modulo additive channels will form a saddle point in (6) and thus optimal via Lemma 1 (assuming that P_{XY} is modulo additive).

Remark 1: Proposition 1 establishes equivalence between the problems addressed in [41] and [7]. But, as was shown in [41], the function $g_T(\eta)$ is not convex in general, therefore we cannot universally utilize Proposition 1, but only for regions of η where it is convex.

We will next show that in the low-SNR regime, specifically, when $\eta_1 \geq \log(n-1)$, the optimal distribution achieving (6) has a unique structure, characterized by Hamming channels. We first give a proper definition of the generalized Hamming channel. A pmf $\mathbf{p} \in \Delta_n$ is called (α, n)-Hamming [41], if for some $\alpha \in [0,1]$, $\bar{x} \triangleq 1\text{–}x$, it is of the form

$$\mathbf{p} = \alpha \cdot \mathbf{e}_n + \bar{\alpha} \cdot \mathbf{u}_n = \left(\alpha + \frac{\bar{\alpha}}{n}, \frac{\bar{\alpha}}{n}, \dots \frac{\bar{\alpha}}{n}\right), \tag{9}$$

where $\mathbf{e}_n = (1, 0, \dots)$ is an extreme point of Δ_n. For negative values of α, the vector on the RHS of (9) is a pmf only if $\alpha \in [-\frac{1}{n-1}, 0)$. In that case it has a full support, and the first probability is the smallest and all the other $n-1$ probabilities are the largest and equal to each other. Note also that $\mathbf{p} = \mathbf{u}_n$ for $\alpha = 0$ and then $H(\mathbf{p}) = \log n$, while $\mathbf{p} = (0, \mathbf{u}_{n-1}^T)^T$ for $\alpha = -\frac{1}{n-1}$ and then $H(\mathbf{p}) = \log(n-1)$. We thus generalize the Hamming pmf for all $\alpha \in [-1, 0]$ as follows. A pmf $\mathbf{p}$ is (α, n, k) negative-Hamming if

$$\mathbf{p} = [\alpha \cdot \mathbf{e}_k + \bar{\alpha}\mathbf{u}_k, \mathbf{0}_{n-k}], \tag{10}$$

where $k \in [n-1]$ is such that $\alpha \in (-\frac{1}{k-1}, 0]$ and $k = n$ otherwise.

Theorem 3: Consider the optimization problem defined in (6), and assume that $\eta_1 \geq \log(n-1)$, then, the optimal P_{V} and P_{W} are a regular Hamming channel with parameter α and a negative Hamming channel with parameters (β, n, n), respectively, where $\alpha \in [0,1]$ is the positive root of

$$\eta_2 + \left(\alpha + \frac{\bar{\alpha}}{n}\right)\log\left(\alpha + \frac{\bar{\alpha}}{n}\right) + \frac{(n-1)\bar{\alpha}}{n}\log\frac{\bar{\alpha}}{n} = 0, \tag{11}$$

and $\beta \in [-1/(n-1), 0]$ is the negative root of

$$\eta_1 + \left(\beta + \frac{\bar{\beta}}{n}\right)\log\left(\beta + \frac{\bar{\beta}}{n}\right) + \frac{(n-1)\bar{\beta}}{n}\log\frac{\bar{\beta}}{n} = 0. \tag{12}$$

Furthermore,

$$R^{\text{mod}}(\eta_1, \eta_2) = -\left(\alpha\beta + \frac{\overline{\alpha\beta}}{n}\right)\log\left(\alpha\beta + \frac{\overline{\alpha\beta}}{n}\right) - \frac{(n-1)\overline{\alpha\beta}}{n}\log\frac{\overline{\alpha\beta}}{n}.$$

Remark 2: This elegant result does not extends to the regime $\eta_1 \in (0, \log(n-1))$, as the following counterexample demonstrates. Suppose $\mathsf{P}_{\mathsf{W}} = \mathbf{p}$ is a negative Hamming channel with parameters $(0.46, 3, 2)$, and take $\eta_2 = 0.7$. In this case the positive Hamming point is given by $\mathbf{q}^+ = (0.866, 0.067, 0.067)^T$ which achieves an output entropy of $h(\mathbf{p} * \mathbf{q}^+) = 1.179$ (bits). However, taking $\mathbf{q}^* = (0.857, 0.031, 0.112)^T$ gives us $h(\mathbf{p} * \mathbf{q}^*) = 1.165 < h(\mathbf{p} * \mathbf{q}^+)$ (bits).

We next provide bounds on (6) which complement the result of Theorem 3.

Theorem 4: Let α be the positive root of (11), β be the parameter of the negative Hamming pmf (10) with entropy η_1, and ζ be the positive root of (12). If $\eta_1 \in (0, \log(n-1))$, then

$$R^{\text{mod}}(\eta_1, \eta_2) \leq -\left(\alpha\beta + \frac{\alpha\bar{\beta}}{k} + \frac{\bar{\alpha}}{n}\right)\log\left(\alpha\beta + \frac{\alpha\bar{\beta}}{k} + \frac{\bar{\alpha}}{n}\right)$$
$$-(k\text{–}1)\left(\frac{\alpha\bar{\beta}}{k} + \frac{\bar{\alpha}}{n}\right)\log\left(\frac{\alpha\bar{\beta}}{k} + \frac{\bar{\alpha}}{n}\right) - (n\text{–}k)\left(\frac{\bar{\alpha}}{n}\right)\log\left(\frac{\bar{\alpha}}{n}\right).$$

If $n = 3$, then

$$R^{\text{mod}}(\eta_1, \eta_2) \geq (1+\beta)h_b\left(\frac{1\text{–}\alpha}{3}\right) + (1+\beta)\left(1\text{–}\frac{1\text{–}\alpha}{3}\right) - \beta\eta_2.$$

If $n > 3$, then

$$R^{\text{mod}}(\eta_1, \eta_2) \geq -\left(\alpha\zeta + \frac{\overline{\alpha\zeta}}{n}\right)\log\left[\alpha\zeta + \frac{\overline{\alpha\zeta}}{n}\right] - \frac{(n\text{–}1)\overline{\alpha\zeta}}{n}\log\frac{\overline{\alpha\zeta}}{n}.$$

Finally, we consider the high-SNR regime, namely the scenario where η_1 is small. In such case we have the following characterization of the optimal distributions and rate.

Theorem 5: Suppose $\eta_2 > \log(n-1)$, then

$$R(\eta_1, \eta_2) - \eta_2 = \alpha\beta\log\left(1 + \frac{\alpha n}{1-\alpha}\right)\cdot(1 + o(1)), \tag{13}$$

with α and β being the positive roots of (11), and (12), and $o(1)$ vanishes when $\eta_1 \downarrow 0$. Optimal P_{W} and P_{V} are both positive Hamming distributions satisfying the constraint with equality.

IV. Modulo Additive Channels with TV Constraint

Let $\delta \in (0, 2)$ be given, and a nominal channel modulo additive channel represented by $P_{\mathsf{W}}^{(0)}$. In this section, the constraint $H(\mathsf{W}) \leq \eta_1$ in P_{XY} from the previous section is replaced with the constraint $d_{\text{TV}}(\mathsf{P}_{\mathsf{W}}, \mathsf{P}_{\mathsf{W}}^{(0)}) \leq \delta$ (the set $\mathcal{Q}_{\mathcal{Z}|\mathcal{Y}}$ remains the same). We denote the resulting compound IB value as $R^{\text{TV}}(\delta, \eta_2)$.

A natural approach is to relate $R^{\text{TV}}(\delta, \eta_2)$ to the standard bottleneck problem $R(0, \eta_2) \equiv R_T^{\text{CEB}}(\eta_2)$ via the continuity of entropy in the total variation metric. This idea was used, e.g., in [42], to establish generalization bounds for the bottleneck problem, that is, in the regime of vanishing δ. Here, we present a slightly tighter result, valid for any $\delta \in (0, 1)$. To this end, recall that the entropy difference of two pmfs in Δ_n of total variation δ is bounded by [43], [44] $\omega(\delta, n) \triangleq \frac{1}{2}\delta\log(n-1) + h_b\left(\frac{\delta}{2}\right)$.

Proposition 2: For any $\delta \in (0, 1)$

$$\left|R^{\text{TV}}(\delta, \eta_2) - R_T^{\text{CEB}}(\eta_2))\right| \leq \omega(\delta, n) \tag{14}$$

where $R_T^{\text{CEB}}(\eta_2)$ is computed at $\mathsf{P}_{\mathsf{W}}^{(0)}$.

Proposition 2 relates the compound IB to the standard IB problem, however, the latter is, in general, difficult to compute (and requires, for example, alternating minimization algorithm as in Section V). In what follows, we will state computable

upper and lower bounds to $R^{\mathrm{TV}}(\delta,\eta_2)$. To this end, let T be a channel transition matrix, and let $\theta(T)\in[0,1]$ be the Dobrushin contraction coefficient of T [45]

$$\theta(T) \triangleq \max_{\mathbf{p},\mathbf{q}\in\Delta_n:\,\mathbf{p}\neq\mathbf{q}} \frac{d_{\mathrm{TV}}(T\mathbf{p},T\mathbf{q})}{d_{\mathrm{TV}}(\mathbf{p},\mathbf{q})} = \frac{1}{2}\max_{i,i'\in[n]:\,i\neq i'} d_{\mathrm{TV}}(T_i,T_{i'}), \tag{15}$$

where T_i is the ith row of T (the second inequality is a "two-point characterization"). Thus, at worst case, $\theta(T)$ is computable by merely n^2-n total variation distance calculations. Furthermore, if $T\in[0,1]^{n\times n}$ is obtained by n permutations of a pmf, then only $n-1$ total variation distance calculations are required. Second, let $\Gamma(\delta)\triangleq\min_{\mathbf{q}\in\Delta_n:\,d_{\mathrm{TV}}(\mathbf{q},\mathbf{u}_n)\le\delta} H(\mathbf{q})$ be the minimal entropy over a total variation ball centered at $\mathbf{u}_n$. This problem has a closed-form solution [46, Thm. 3] as follows: If $1-1/n\le\delta/2$ then the optimal solution is $\mathbf{q}=(1,0,\ldots,0)$ and $\Gamma(\delta)=0$. Otherwise, let $n_0(\delta)\triangleq\lfloor n+1-n\delta/2\rfloor$. Then the optimal solution is $\mathbf{q}^*=(1/n+\delta/2,1/n\ldots,1/n,(n-n_0(\delta)+1)/n-\delta/2,0,\ldots,0)$ (there are n_0-2 terms of $1/n$ so the support size of this solution is n_0). Therefore, for $\delta\in[0,2-2/n]$ the function $\Gamma(\delta)$ is strictly positive and strictly decreasing with extreme values of $\Gamma(0)=\log n$ and $\Gamma(2-2/n)=0$. So, there exists an inverse function to $\Gamma(\delta)$, which we denote by $D(\eta):[0,\log n]\to[0,2-2/n]$. Third, for a given $\mathbf{p}^{(0)}\in\Delta_n$, let $\Phi(\delta;\mathbf{p}^{(0)})\triangleq\max_{\mathbf{q}\in\Delta_n:\,d_{\mathrm{TV}}(\mathbf{q},\mathbf{p}^{(0)})\le\delta} H(\mathbf{q})$ be the maximal entropy over a total variation ball centered at $\mathbf{u}_n$. This problem also has a closed-form solution [46, Thm. 2] as follows: Let μ and ν be such that $\sum_{i=1}^n(p_i^{(0)}-\mu)_+=\sum_{i=1}^n(\nu-p_i^{(0)})_+=\delta/2$. If $\nu\ge\mu$ then $\Phi(\delta;\mathbf{p}^{(0)})=\log n$ and the maximizing distribution $\mathbf{q}^*=\mathbf{u}_n$ is uniform. Otherwise, $\mathbf{q}^*$ is such that $\mathbf{q}_i^*=\min\{\max(p_i^{(0)},\mu),\nu\}$, and its entropy is the maximum.

Theorem 6: Let $T(\mathsf{P_W})$ be the channel transition matrix which corresponds to n cyclic permutations of $\mathsf{P_W}$. Then,

$$R^{\mathrm{TV}}(\delta,\eta_2)\ge\max_{\mathsf{P_W}:\,d_{\mathrm{TV}}(\mathsf{P_W},\mathsf{P_W^{(0)}})\le\delta}\Gamma\left(\theta(T(\mathsf{P_W}))\cdot D(\eta)\right), \tag{16}$$

and that

$$R^{\mathrm{TV}}(\delta,\eta_2)\le\min_{\mathsf{P_V}:\,H(\mathsf{P_V})=\eta_2}\Phi\left(\theta(T(\mathsf{P_V}))\cdot\delta;T(\mathsf{P_V})\mathbf{p}^{(0)}\right). \tag{17}$$

Since $\Gamma(\delta)$, its inverse $D(\eta)$, as well as $\Phi(\delta;\mathbf{p}^{(0)})$ are all computable, the expressions in the lower bound can be computed for any given $T(\mathsf{P_W})$. In general, the optimization over $\mathsf{P_W}$ in the lower bound is computationally difficult. However, *any* arbitrary choice of $\mathsf{P_W}$ which satisfies the constraint leads to a valid lower bound. Analogous statements hold for $\mathsf{P_V}$ in the upper bound. It should be noted that the optimization of the lower bound requires finding the minimal $\theta(T(\mathsf{P_W}))$, whereas $\mathsf{P_V}$ in the upper bound affects both the contraction coefficient $\theta(T(\mathsf{P_V}))$ and the transformed nominal pmf $T(\mathsf{P_V})\mathbf{p}^{(0)}$.

Note that as $g_T(\eta)\ge\eta$ always holds [41, Lemma 5 (c)], and so the lower bound of Thm. 6 requires optimizing over $\mathsf{P_W}$ for which $\theta(T(\mathsf{P_W}))<1$. In general $\theta(T)<1$ only if no two rows of T are orthogonal. Here, since the rows of $T(\mathsf{P_W})$ are circular permutations of $\mathsf{P_W}$, it holds that $\theta(T)<1$ if and only if the support of $\mathsf{P_W}$ is strictly larger than $n/2$.

Algorithm 1: pf_iterator(args)

Input: $\mathsf{P_X}$, $\mathsf{P_Y}$, $\mathsf{P_{Z|Y}}$ and β_1

Initialize: Arbitrary $\mathsf{P}_{\mathsf{XY}}^{(0)}$ with valid marginals, $t=1$.

while *Variation in* $I(\mathsf{X};\mathsf{Z})$ *is greater then* ϵ **do**

Compute $\mathsf{P}_{\mathsf{Z|X}}^{(t)}(z|x)=\frac{\sum_{y\in\mathcal{Y}}\mathsf{P_{Z|Y}}(z|y)\mathsf{P}_{\mathsf{XY}}^{(t-1)}(x,y)}{\mathsf{P_X}(x)}$;

Set $\mathsf{P}_{\mathsf{XY}}^{(t)}(x,y)=\frac{\mathsf{P_X}(x)\mathsf{P_Y}(y)e^{-\beta_1 D\left(\mathsf{P_{Z|Y}}(\cdot|y)||\mathsf{P}_{\mathsf{Z|X}}^{(t)}(\cdot|x)\right)}}{Z_1(x,y,\beta_1)}$;

Find $Z_1(x,y,\beta_1)$ s.t. $\mathsf{P}_{\mathsf{XY}}^{(t)}$ has valid marginals;

$t=t+1$;

end

Output: $\mathsf{P}_{\mathsf{XY}}^*$

Remark 3: The proof of Thm. 6 provides a lower bound on $g_T(\eta)$ Witsenhausen's function from [41], which may be of independent interest.

V. Alternating Optimization Algorithm

We return in this section to the general (C_1,C_2) PF compound set. Applying a two-phase Lagrangian methodology, we obtain a set of self consistent equations for $\mathsf{P_{XY}}$ and $\mathsf{P_{Z|Y}}$. We then propose a Blahuto-Arimoto type iterative algorithm that solves those equations.

A. The Inner Lagrangian

Fix $\mathsf{P_{Z|Y}}$ that satisfies $I(\mathsf{Y};\mathsf{Z})\le C_2$ and consider the inner minimization problem from (1), given by

$$f(\mathsf{P_{Z|Y}},C_1)=\min_{\mathsf{P_{XY}}:I(\mathsf{X};\mathsf{Y})\ge C_1} I(\mathsf{X};\mathsf{Z}). \tag{18}$$

For $\lambda_1\ge0$, the respective Lagrangian of (18) is given by,

$$\mathcal{L}_{\min}(\mathsf{P_{XY}},\lambda_1,\boldsymbol{\mu},\boldsymbol{\nu})=I(\mathsf{X};\mathsf{Z})+\lambda_1\left(C_1-I(\mathsf{X};\mathsf{Y})\right)+\sum_{x\in\mathcal{X}}\mu_x\sum_{y\in\mathcal{Y}}\mathsf{P_{XY}}(x,y)+\sum_{y\in\mathcal{Y}}\nu_y\sum_{x\in\mathcal{X}}\mathsf{P_{XY}}(x,y). \tag{19}$$

Proposition 3: Any stationary point $\mathsf{P}_{\mathsf{XY}}^*$ of (19) satisfies

$$\mathsf{P}_{\mathsf{XY}}^*(x,y)=\frac{\mathsf{P_X}(x)\mathsf{P_Y}(y)e^{-\beta_1 D\left(\mathsf{P_{Z|Y}}(\cdot|y)||\mathsf{P_{Z|X}}(\cdot|x)\right)}}{Z(x,y,\beta_1)}, \tag{20}$$

where $\beta_1\triangleq1/\lambda_1$ and $Z(x,y,\beta_1)$ is the normalization constant. Furthermore, the optimal $\mathsf{P_{Z|X}}(z|x)$ is given by

$$\mathsf{P_{Z|X}}(z|x)=\frac{1}{\mathsf{P_X}(x)}\sum_{y\in\mathcal{Y}}\mathsf{P_{Z|Y}}(z|y)\mathsf{P}_{\mathsf{XY}}^*(x,y). \tag{21}$$

The system of equations characterizing the stationary points in (20) and (21) must hold simultaneously for consistency. An alternating iteration algorithm is a common approach to solve these equations.

Proposition 4: Equations (20) and (21) are satisfied simultaneously at the minimum of the Lagrangian (19) where the minimization is performed independently over the convex sets of $\{\mathsf{P_{XY}}(x,y)\}$ and $\{\mathsf{P_{Z|X}}(z|x)\}$,

$$\min_{\mathsf{P_{Z|X}}(z|x)}\min_{\mathsf{P_{XY}}(x,y)}\mathcal{L}_{\min}(\mathsf{P_{XY}},\lambda_1,\mu,\nu). \tag{22}$$

These independent conditions correspond precisely to alternating interactions of (20) and (21). Denoting by t the iteration step, we obtain Algorithm 1.

B. The Outer Lagrangian

Note that maximization of $I(\mathsf{X};\mathsf{Z})$ for a fixed P_{XY} that satisfies $I(\mathsf{X};\mathsf{Y}) \geq C_1$ is just the standard *information bottleneck*, the proposed here technique is identical to the one suggested in [1]. The respective algorithm from [1, Theorem 5] is summarized in Algorithm 2.

Algorithm 2: ib_iterator(args)

Input: P_{XY}, and β_2
Initialize: Arbitrary $\mathsf{P}^{(0)}_{\mathsf{Z}|\mathsf{Y}}$, $s=1$.
while *Variation in* $I(\mathsf{X};\mathsf{Z})$ *is greater then* ϵ **do**
  $\mathsf{P}^{(s)}_{\mathsf{Z}|\mathsf{Y}}(z|y) = \frac{\mathsf{P}^{(s-1)}_{\mathsf{Z}}(z)}{Z(y,\beta_2)} \cdot e^{-\beta_2 D\left(\mathsf{P}_{\mathsf{X}|\mathsf{Y}}(\cdot|y)||\mathsf{P}^{(s-1)}_{\mathsf{X}|\mathsf{Z}}(\cdot|z)\right)}$;
  $\mathsf{P}^{(s)}_{\mathsf{Z}}(z) = \sum_{y\in\mathcal{Y}} \mathsf{P}_{\mathsf{Y}}(y)\mathsf{P}^{(s-1)}_{\mathsf{Z}|\mathsf{Y}}(z|y)$;
  $\mathsf{P}^{(s)}_{\mathsf{X}|\mathsf{Z}}(x|z) = \sum_{y\in\mathcal{Y}} \mathsf{P}_{\mathsf{X}|\mathsf{Y}}(x|y)\mathsf{P}^{(s)}_{\mathsf{Y}|\mathsf{Z}}(y|z)$;
  $s = s+1$;
end
Output: $\mathsf{P}^{*}_{\mathsf{Z}|\mathsf{Y}}$

C. The Compound Algorithm

To this end, two algorithms were proposed that aim to solve (1) in a isolated manner. In this section we propose a method that intervenes them together with an objective to find the solution simultaneously. There are two natural approaches to handle this problem. The first one is to alternate between the steps of each algorithm until convergence. The second one is to run the first algorithm until convergence and then the other one, and so on. We have found the second type of algorithms to be more effective, and this is summarized in Algorithm 3. We have no global convergence guarantees here as the standard IB does not have such [1].

Algorithm 3: COMIB Programming

Input: P_{X}, P_{Y}, C_1 and C_2
Initialize: $\mathsf{P}^{(0)}_{\mathsf{Z}|\mathsf{Y}}$.
while *Variation in* $I(\mathsf{X};\mathsf{Z})$ *is greater then* ϵ **do**
  for $\beta_1 \in \mathbb{R}_+$ **do**
    $\mathsf{P}^{*}_{\mathsf{XY}}(\beta_1^*) = pf\_iterator(\mathsf{P}_{\mathsf{X}}, \mathsf{P}_{\mathsf{Y}}, \mathsf{P}^{(0)}_{\mathsf{Z}|\mathsf{Y}}, \beta_1)$;
  end
  Find $\mathsf{P}^{*}_{\mathsf{XY}}(\beta_1^*)$ s.t. $I(\mathsf{X};\mathsf{Y}) = C_1$;
  for $\beta_2 \in \mathbb{R}_+$ **do**
    $\mathsf{P}^{*}_{\mathsf{Z}|\mathsf{Y}}(\beta_2) = ib\_iterator(\mathsf{P}^{*}_{\mathsf{XY}}(\beta_1^*), \beta_2)$;
  end
  Find β_2^* s.t. $I(\mathsf{P}^{*}_{\mathsf{Z}|\mathsf{Y}}(\beta_2^*)) = C_2$. $\mathsf{P}^{*}_{\mathsf{Z}|\mathsf{Y}}(\beta_2^*) \mapsto \mathsf{P}^{(0)}_{\mathsf{Z}|\mathsf{Y}}$.
end
Output: P^{*}_{XY},$P^{*}_{\mathsf{Z}|\mathsf{Y}}$

VI. Numerical Simulations

We evaluate both the analytical bounds derived in Thm. 4 and the algorithm developed in Section V by comparing their results on a common example. A representative example of $n=5$ and various rate constraints is shown in Figure 1. As expected, the algorithm's output lies in the medium between the upper and lower bounds. It is also somewhat closer to the lower bound, which hints that the upper bounds might be improved.

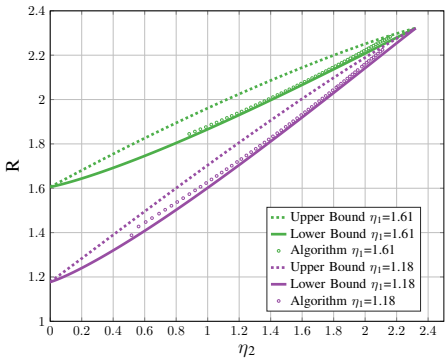

Fig. 1. Bounds on $R^{\text{mod}}(\eta_1,\eta_2)$ function for $n=5$ and $\eta_1 = \{1.61, 1.18\}$.

We also evaluate the bounds derived for the TV class setting in Section IV. An example for $n=15$, and $\delta=0.3$, and $\mathsf{P}^{(0)}_{\mathsf{W}} \propto \exp(2i)$ for $i \in [15]$ (and 0 otherwise) is illustrated in Figure 2. The bounds are fairly close and tighten for large values of η_2.

VII. Concluding Remarks

We have defined the COMIB programming problem. We obtained various characterizations for the binary setting, the Gaussian settings, and derived upper and lower bounds for modulo additive channels with PF constraints, and with TV constraints. Under some qualifying conditions, Gaussian distributions and Hamming channels were shown to be extermal. Finally, we have proposed an alternating iteration algorithm that finds a locally optimal solution. Future research calls for further tightening these bounds, and establishing additional settings in which the optimal channels and representations can be analytically characterized.

Acknowledgment

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 1791/17, and by the WIN consortium via the Israel minister of economy and science.

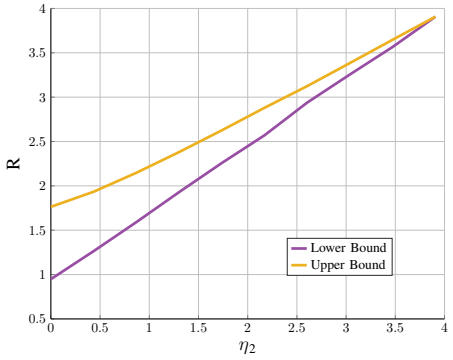

Fig. 2. Bounds on $R^{\text{TV}}(\delta,\eta_2)$ for $n=15$ and $\delta=0.3$.

REFERENCES

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

[2] R. Fontana, "On universal coding for classes of composite and remote sources with memory (corresp.)," *IEEE Trans. Inf. Theory*, vol. 27, no. 6, pp. 784–786, Nov. 1981.

[3] A. Dembo and T. Weissman, "The minimax distortion redundancy in noisy source coding," *IEEE Trans. Inf. Theory*, vol. 49, no. 11, pp. 3020–3030, Nov. 2003.

[4] T. Weissman, "Universally attainable error exponents for rate-distortion coding of noisy sources," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1229–1246, Jun. 2004.

[5] R. Dobrushin and B. Tsybakov, "Information transmission with additional noise," *IRE Trans. Inf. Theory*, vol. 8, no. 5, pp. 293–304, Sep. 1962.

[6] J. Wolf and J. Ziv, "Transmission of noisy information to a noisy receiver with minimum distortion," *IEEE Trans. Inf. Theory*, vol. 16, pp. 406–411, Jul. 1970.

[7] H. S. Witsenhausen and A. D. Wyner, "A conditional entropy bound for a pair of discrete random variables," *IEEE Trans. Inf. Theory*, vol. 21, no. 5, pp. 493–501, Sep. 1975.

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

[9] A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Médard, "From the information bottleneck to the privacy funnel," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Nov. 2014, pp. 501–505.

[10] F. du Pin Calmon, A. Makhdoumi, M. Médard, M. Varia, M. Christiansen, and K. R. Duffy, "Principal inertia components and applications," *IEEE Trans. Inf. Theory*, vol. 63, no. 8, pp. 5011–5038, Aug. 2017.

[11] S. Shamai, "The information bottleneck: A unified information theoretic view," National Conference on Communications (NCC2021), Jul. 2021, plenary Address.

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

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

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

[15] S. I. Bross and A. Lapidoth, "The additive noise channel with a helper," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Visby, Sweden, Aug. 2019, pp. 1–5.

[16] A. Zaidi, I. E. Aguerri, and S. S. (Shitz), "On the information bottleneck problems: Models, connections, applications and information theoretic views," *Entropy*, vol. 22, no. 2, 2020.

[17] A. Wyner and J. Ziv, "A theorem on the entropy of certain binary sequences and applications I," *IEEE Trans. Inf. Theory*, vol. 19, pp. 769–772, Nov. 1973.

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

[19] D. Guo, S. Shamai (Shitz), and S. Verdú, "The interplay between information and estimation measures," *Found. Trends Signal Process.*, vol. 6, no. 4, pp. 243–429, 2012.

[20] R. Bustin, M. Payaro, D. P. Palomar, and S. Shamai (Shitz), "On MMSE crossing properties and implications in parallel vector Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 59, no. 2, pp. 818–844, Feb. 2013.

[21] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. 18, pp. 14–20, Jan. 1972.

[22] S. Hassanpour, D. Wuebben, and A. Dekorsy, "Overview and investigation of algorithms for the information bottleneck method," in *Proc. 11th Int. ITG Conf. Syst., Commun. Coding (SCC)*, Feb. 2017, pp. 1–6.

[23] I. Estella-Aguerri and A. Zaidi, "Distributed variational representation learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 120–138, Jan. 2021.

[24] N. Slonim, "The information bottleneck: Theory and applications," Ph.D. dissertation, Hebrew University of Jerusalem, Jerusalem, Israel, 2002.

[25] M. Dikshtein, N. Weinberger, and S. Shamai (Shitz), "The compound information bottleneck outlook," *CoRR*, vol. abs/2205.04567v1, 2022. [Online]. Available: https://arxiv.org/abs/2205.04567v1

[26] Y. Steinberg, "Coding and common reconstruction," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 4995–5010, Nov. 2009.

[27] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Jerusalem, Israel, Apr. 2015, pp. 1–5.

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

[29] P. Farajiparvar, A. Beirami, and M. Nokleby, "Information bottleneck methods for distributed learning," in *Proc. 56th Annu. Allerton Conf. Commun., Control Comput.*, 2018, pp. 24–31.

[30] R. A. Amjad and B. C. Geiger, "Learning representations for neural network-based classification using the information bottleneck principle," *IEEE Trans. Pattern Anal.*, vol. 42, no. 9, pp. 2225–2239, Sep. 2020.

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

[32] Z. Goldfeld and Y. Polyanskiy, "The information bottleneck problem and its applications in machine learning," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 19–38, May 2020.

[33] E. Erkip and T. M. Cover, "The efficiency of investment information," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1026–1040, May 1998.

[34] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *CoRR*, vol. abs/1703.00810, 2017. [Online]. Available: http://arxiv.org/abs/1703.00810

[35] M. Gabrié, A. Manoel, C. Luneau, J. Barbier, N. Macris, F. Krzakala, and L. Zdeborová, "Entropy and mutual information in models of deep neural networks," *J. Stat. Mech. Theory Exp.*, vol. 2019, no. 12, Dec. 2019.

[36] Z. Goldfeld, E. van den Berg, K. H. Greenewald, I. Melnyk, N. Nguyen, B. Kingsbury, and Y. Polyanskiy, "Estimating information flow in neural networks," *CoRR*, vol. abs/1810.05728, 2018.

[37] H. Cheng, D. Lian, S. Gao, and Y. Geng, "Evaluating capability of deep neural networks for image classification via information plane," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 168–182.

[38] S. Yu, K. Wickstrøm, R. Jenssen, and J. C. Príncipe, "Understanding convolutional neural networks with information theory: An initial exploration," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 435–442, Jan. 2021.

[39] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA.: Cambridge University Press, 2014.

[40] D. P. Bertsekas and J. N. Tsitsiklis, *Introduction to probability*. Belmont, MA: Athena Scientific, 2002.

[41] H. S. Witsenhausen, "Entropy inequalities for discrete channels," *IEEE Trans. Inf. Theory*, vol. 20, no. 5, pp. 610–616, Sep. 1974.

[42] O. Shamir, S. Sabato, and N. Tishby, "Learning and generalization with the information bottleneck," *Theoretical Computer Science*, vol. 411, no. 29, pp. 2696–2711, 2010, Algorithmic Learning Theory (ALT 2008).

[43] K. M. Audenaert, "A sharp Fannes-type inequality for the von Neumann entropy," *J. Phys. A*, vol. 40, pp. 8127–8136, 2007.

[44] Z. Zhang, "Estimating mutual information via Kolmogorov distance," *IEEE Trans. Inform. Theory*, vol. 53, no. 9, pp. 3280–3282, Sep. 2007.

[45] R. L. Dobrushin, "Central limit theorem for nonstationary Markov chains. i," *Theory Prob. Applications*, vol. 1, no. 1, pp. 65–80, 1956.

[46] S.-W. Ho and R. W. Yeung, "The interplay between entropy and variational distance," *IEEE Trans. Inform. Theory*, vol. 56, no. 12, pp. 5906–5929, Dec. 2010.