Conference Paper

On the Information Bottleneck Problems: An Information Theoretic Perspective

Author(s):
Zaidi, Abdellatif; Shamai (Shitz), Shlomo

Publication Date:
2020-02-26

Permanent Link:

https://doi.org/10.3929/ethz-b-000402952 →

International Zurich Seminar on Information and Communication (IZS), February 26 – 28, 2020

On the Information Bottleneck Problems: An Information Theoretic Perspective

Abdellatif Zaidi[†] [‡] Shlomo Shamai[†]

[‡] Université Paris-Est, Champs-sur-Marne, 77454, France
[†] Paris Research Center, Huawei Technologies, Boulogne-Billancourt, 92100, France
[†] Technion Institute of Technology, Technion City, Haifa 32000, Israel
{abdellatif.zaidi@u-pem.fr, sshlomo@ee.technion.ac.il}

***Abstract*—This paper focuses on variants of the bottleneck problem taking an information theoretic perspective. The intimate connections of this setting to: remote source-coding, information combining, common reconstruction, the Wyner-Ahlswede-Korner problem, the efficiency of investment information, CEO source coding under logarithmic-loss distortion measure and others are highlighted. We discuss the distributed information bottleneck problem with emphasis on the Gaussian model. For this model, the optimal tradeoffs between relevance (i.e., information) and complexity (i.e., rates) in the discrete and vector Gaussian frameworks is determined.**

I. Statistical Inference

Let a measurable variable $X \in \mathcal{X}$ and a target variable $Y \in \mathcal{Y}$ with unknown joint distribution $P_{X,Y}$ be given. In the classic problem of statistical learning, one wishes to infer an accurate predictor of the target variable $Y \in \mathcal{Y}$ based on observed realizations of $X \in \mathcal{X}$. That is, for a given class $\mathcal{F}$ of admissible predictors $\psi : \mathcal{X} \to \hat{\mathcal{Y}}$ and a loss function $\ell : \mathcal{Y} \to \hat{\mathcal{Y}}$ that measures discrepancies between true values and their estimated fits, one aims at finding the mapping $\psi \in \mathcal{F}$ that minimizes the expected (population) risk

$$\mathcal{C}_{P_{X,Y}}(\psi, \ell) = \mathbb{E}_{P_{X,Y}}[\ell(Y, \psi(X))]. \tag{1}$$

An abstract inference model is shown in Figure 1.

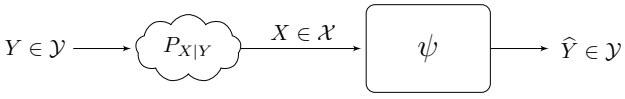

Fig. 1. An abstract inference model for learning.

The choice of a "good" loss function $\ell(\cdot)$ is often controversial in statistical learning theory. There is however numerical evidence that models that are trained to minimize the error's entropy often outperform ones that are trained using other criteria such as mean-square error (MSE) and higher-order statistics [1], [2]. This corresponds to choosing the loss function given by the logarithmic loss, which is defined as

$$\ell_{\log}(y, \hat{y}) := \log \frac{1}{\hat{y}(y)} \tag{2}$$

for $y \in \mathcal{Y}$ and $\hat{y} \in \mathcal{P}(\mathcal{Y})$ designates here a probability distribution on $\mathcal{Y}$ and $\hat{y}(y)$ is the value of that distribution evaluated at the outcome $y \in \mathcal{Y}$. Although a complete and rigorous justification of the usage of the logarithmic loss as distortion measure in learning is still awaited, recently a partial explanation appeared in [3] where Painsky and Wornell show that, for binary classification problems, by minimizing the logarithmic-loss one actually minimizes an upper bound to any choice of loss function that is smooth, proper (i.e., unbiased and Fisher consistent) and convex. Along the same line of work, the authors of [4] show that under some natural data processing property Shannon's mutual information uniquely quantifies the reduction of prediction risk due to side information. Perhaps, this justifies partially why the logarithmic-loss fidelity measure is widely used in learning theory and has already been adopted in many algorithms in practice such as the *infomax* criterion [5]. The logarithmic loss measure also plays a central role in the theory of prediction [6, Ch. 09], where it is often referred to as the *self-information* loss function, as well as in Bayesian modeling [7] where priors are usually designed so as to maximize the mutual information between the parameter to be estimated and the observations.

Let for every $x \in \mathcal{X}$, $\psi(x) = Q(\cdot|x) \in \mathcal{P}(\mathcal{Y})$. It is easy to see that

$$\mathbb{E}_{P_{X,Y}}[\ell_{\log}(Y, Q)] = \sum_{x \in \mathcal{X},\, y \in \mathcal{Y}} P_{X,Y}(x, y) \log\Big(\frac{1}{Q(y|x)}\Big) \tag{3a}$$

$$= H(Y|X) + D\big(P_{Y|X} \| Q\big) \tag{3b}$$

$$\geq H(Y|X) \tag{3c}$$

with equality iff $\psi(X) = P_{Y|X}$. That is,

$$\min_{\psi} \mathcal{C}_{P_{X,Y}}(\psi, \ell_{\log}) = H(Y|X). \tag{4}$$

If the joint distribution $P_{X,Y}$ is unknown, which is most often the case in practice, the population risk as given by (1) cannot be computed directly; and, in the standard approach, one usually resorts to choosing the predictor with minimal risk on a training dataset consisting of n labeled samples $\{(x_i, y_i)\}_{i=1}^n$ that are drawn independently from the unknown joint distribution $P_{X,Y}$. In this case, it is important to restrict the set $\mathcal{F}$ of admissible predictors to a low-complexity class to prevent overfitting. One way to reduce the model's complexity is by restricting the range of the prediction function as shown in Figure 2. Here, the stochastic mapping $\phi : \mathcal{X} \longrightarrow \mathcal{U}$ is a compressor with

$$\|\phi\| \leq R \tag{5}$$

for some prescribed 'input-complexity' value R.

Let $U = \phi(X)$. The expected logarithmic loss is now given by

$$\mathcal{C}_{P_{X,Y}}(\phi, \psi; \ell_{\log}) = \mathbb{E}_{P_{X,Y}}[\ell_{\log}(Y, \psi(U))] \tag{6}$$

and takes its minimum value with the choice $\psi(U) = P_{Y|U}$,

$$\min_{\psi} \mathcal{C}_{P_{X,Y}}(\phi, \psi; \ell_{\log}) = H(Y|U) \tag{7}$$

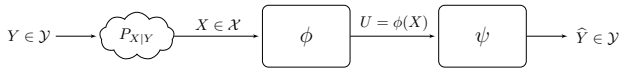

Fig. 2. Inference problem with constrained model's complexity.

where the choice of U is subjected to the input constraint (5). Noting that the right-hand-side (RHS) of (7) is larger for small values of R, it is clear that a good predictor ϕ should strike a right balance between reducing the model's complexity and reducing the error's entropy, or, equivalently, maximizing the mutual information $I(U;Y)$ about the target variable Y.

A. Remote Source Coding under Logarithmic Loss

The aforementioned inference problem is a one-shot coding problem, in the sense that the prediction and estimation operations are performed letter-wise. Consider now the (asymptotic) remote source coding problem shown in Figure 3 in which the coding operations are performed over blocks of size n, with n assumed to be large. Here, Y designates a memoryless remote source; and X a noisy version of it that is observed at the encoder. The range of the encoder map is allowed to grow with the size of the input sequence as

$$\|\phi^{(n)}\| \leq nR. \tag{8}$$

That is, the encoder uses at most R bits per sample to describe its observation to a decoder which is interested in reconstructing the remote source Y^n to within an average distortion level D, i.e.,

$$\mathbb{E}[\ell_{\log}^{(n)}(\mathbf{y},\hat{\mathbf{y}})] \leq D \tag{9}$$

where the incurred distortion between two vectors $\mathbf{y}$ and $\hat{\mathbf{y}}$ is given by

$$\ell_{\log}^{(n)}(\mathbf{y},\hat{\mathbf{y}}) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\log}(y_i,\hat{y}_i) \tag{10}$$

with the per-letter distortion defined as specified by (2).

Fig. 3. A remote source coding problem.

The rate distortion region of this model is given by the union of all pairs (R,D) that satisfy [8], [9]

$$R \geq I(U;X) \tag{11a}$$
$$D \geq H(Y|U) \tag{11b}$$

where the union is over all auxiliary random variables U that satisfy that $U \multimap X \multimap Y$ forms a Markov Chain in this order. Invoking the support lemma [10, p. 310], it is easy to see that this region is not altered if one restricts U to satisfy $|\mathcal{U}| \leq |\mathcal{X}| + 1$. Also, using the substitution $\Delta := H(Y) - D$, the region can be written equivalently as the union of all pairs $(R, H(Y) - \Delta)$ that satisfy

$$R \geq I(U;X) \tag{12a}$$
$$\Delta \leq I(U;Y) \tag{12b}$$

where the union is over all U's that satisfy $U \multimap X \multimap Y$, with $|\mathcal{U}| \leq |\mathcal{X}| + 1$.

B. Information Bottleneck

The Information Bottleneck (IB) method has been introduced by Tishby *et al.* in [11] as a method for extracting the information that some variable $X \in \mathcal{X}$ provides about another one $Y \in \mathcal{Y}$ that is of interest. Specifically, IB finds a representation U that is maximally informative about Y, i.e., large mutual information $I(U;Y)$, while being minimally informative about X, i.e., small mutual information $I(U;X)$ [1]. The auxiliary random variable U satisfies that $U \multimap X \multimap Y$ is a Markov chain in this order; and is chosen so a to strike a suitable balance between the degree of *relevance* of the representation as measured by the mutual information $I(U;Y)$ and its degree of *complexity* as measured by the mutual information $I(U;X)$. For example, U can be determined so as to minimize the IB-Lagrangian

$$\mathcal{L} : I(U;X) - \beta I(U;Y) \tag{13}$$

over all mappings that satisfy $U \multimap X \multimap Y$. The tradeoff parameter β is a positive Lagrange multiplier associated with the constraint on $I(U;Y)$. The solution of this constrained optimization problem is determined by the following self-consistent equations, for all $(u,x,y) \in \mathcal{U} \times \mathcal{X} \times \mathcal{Y}$,

$$P_{U|X}(u|x) = \frac{P_U(u)}{Z(\beta,x)} \exp\Big(-\beta D\Big(P_{Y|X}(\cdot|x)\|P_{Y|U}(\cdot|u)\Big)\Big) \tag{14a}$$
$$P_U(u) = \sum_{x\in\mathcal{X}} P_X(x) P_{U|X}(u|x) \tag{14b}$$
$$P_{Y|U}(y|u) = \sum_{x\in\mathcal{X}} P_{Y|X}(y|x) P_{X|U}(x|u) \tag{14c}$$

where $P_{X|U}(x|u) = P_{U|X}(u|x)P_X(x)/P_U(u)$ and $Z(\beta,x)$ is a normalization term. It is shown in [11] that alternating iterations of these equations converges to a solution of the problem for any initial $P_{U|X}$. However, by opposition to the standard Blahut-Arimoto algorithm [13], [14] which is classically used in the computation of rate-distortion functions of discrete memoryless sources for which convergence to the optimal solution is guaranteed, convergence here may be to a local optimum only. If $\beta = 0$ the optimization is non-constrained and one can set $U = \varnothing$, which yields minimal relevance and complexity levels. Increasing the value of β steers towards more accurate and more complex representations, until $U = X$ in the limit of very large (infinite) values of β for which the relevance reaches its maximal value $I(X;Y)$.

C. Variational Inference

Recall the IB goal of finding a representation U of X that is maximally informative about Y while being concise enough (i.e., bounded $I(U;X)$). This corresponds to the Lagrangian formulation

$$\mathcal{L} : \max I(U;Y) - \beta I(U;X) \tag{15}$$

where the maximization is over all stochastic mappings $P_{U|X}$ such that $U \multimap X \multimap Y$ and $|\mathcal{U}| \leq |\mathcal{X}| + 1$. The main drawback of the IB principle is that in the exception of small-sized discrete (X,Y) for which iterating (14) converges to an (at least local) solution and jointly Gaussian (X,Y) for which an explicit analytic solution was found, solving (15) is generally computationally costly

[1] As such, the usage of Shannon's mutual information seems to be motivated by the intuition that such a measure provides a natural quantitative approach to the questions of meaning, relevance and common-information, rather than the solution of a well-posed information-theoretic problem – a connection with source coding under logarithmic loss measure appeared later on in [12].

especially for high-dimensional data since it requires computation of mutual information terms. Another important barrier in solving (15) directly is that IB necessitates knowledge of the joint distribution $P_{X,Y}$. A major step ahead, which widened up the range of applications of IB inference for various learning problems, appeared in [15] where the authors use variational inference to derive a lower bound on (15) and show that its optimization can be done through the classic and widely used stochastic gradient descendent (SGD). This has allowed to use deep neural networks to parametrize the involved distributions (including the test channel $P_{U|X}$); and, thus, to handle high-dimensional, possibly continuous, data.

II. Connections

A. Common Reconstruction

Consider the problem of source coding with side information at the decoder, i.e., the well known Wyner-Ziv setting [16], with the distortion measured under logarithmic-loss. Specifically, a memoryless source X is to be conveyed lossily to a decoder that observes a statistically correlated side information Y. The encoder uses R bits per sample to describe its observation to the decoder which wants to reconstruct an estimate of X to within an average distortion level D, where the distortion is evaluated under the log-loss distortion measure. The rate distortion region of this problem is given by the set of all pairs (R, D) that satisfy

$$R + D \geq H(X|Y). \tag{16}$$

The optimal coding scheme utilizes standard Wyner-Ziv compression at the encoder and the decoder map $\psi : \mathcal{U} \times \mathcal{Y} \to \hat{\mathcal{X}}$ is given by

$$\psi(U, Y) = \Pr[X = x|U, Y] \tag{17}$$

for which it is easy to see that

$$\mathbb{E}[\ell_{\log}(X, \psi(U, Y))] = H(X|U, Y). \tag{18}$$

Now, assume that we constrain the coding in a manner that the encoder be able to produce an exact copy of the compressed source constructed by the decoder. This requirement, termed *common reconstruction* constraint (CR), was introduced and studied by Steinberg in [17] for various source coding models, including the Wyner-Ziv setup, in the context of a "general distortion measure. For the Wyner-Ziv problem under log-loss measure that is considered in this section, such a CR constraint causes some rate loss because the reproduction rule (17) is no longer possible. In fact, it is not difficult to see that under the CR constraint the above region reduces to the set of pairs (R, D) that satisfy

$$R \leq I(U; X|Y) \tag{19a}$$
$$D \geq H(X|U) \tag{19b}$$

for some auxiliary random variable for which $U \multimap X \multimap Y$ holds. Observe that (19b) is equivalent to $I(U; X) \geq H(X) - D$ and that, for a given prescribed fidelity level D, the minimum rate is obtained for a description U that achieves the inequality (19b) with equality, i.e.,

$$R(D) = \min_{P_{U|X} \,:\, I(U;X)=H(X)-D} I(U; X|Y). \tag{20}$$

Because $U \multimap X \multimap Y$, we have

$$I(U; Y) = I(U; X) - I(U; X|Y). \tag{21}$$

Under the constraint $I(U; X) = H(X) - D$ it is easy to see that minimizing $I(U; X|Y)$ amounts to maximizing $I(U; Y)$, an aspect which bridges the problem at hand with the IB problem.
In the above, the side information Y is used for binning but not for the estimation at the decoder. If the encoder ignores whether Y is present or not at the decoder side, the benefit of binning is reduced – see the Heegard-Berger model with common reconstruction studied in [18], [19].

B. Information Combining

Consider again the IB problem. Say one wishes the find the representation U that maximizes the relevance $I(U; Y)$ for a given prescribed complexity level, e.g., $I(U; X) = R$. For this setup,

$$I(X; U, Y) = I(U; X) + I(Y; X) - I(U; Y) \tag{22}$$
$$= R + I(Y; X) - I(U; Y) \tag{23}$$

where the first equality holds since $U \multimap X \multimap Y$ is a Markov chain. Maximizing $I(U; Y)$ is then equivalent to minimizing $I(X; U, Y)$. This is reminiscent of the problem of *information combining* [20], where X can be interpreted as a source information that is conveyed through two channels: the channel $P_{Y|X}$ and the channel $P_{U|X}$. The outputs of these two channels are conditionally independent given X; and they should be processed in a manner such that, when combined, they preserve as much information as possible about X.

C. Wyner-Ahlswede-Korner Problem

Here, the two memoryless sources X and Y are encoded separately at rates R_X and R_Y respectively. A decoder gets the two compressed streams and aims at recovering Y losslessly. This problem was studied and solved separately by Wyner [21] and Ahlswede and Körner [22]. For given $R_X = R$, the minimum rate R_Y that is needed to recover Y losslessly is

$$R_Y^\star(R) = \min_{P_{U|X} \,:\, I(U;X) \leq R} H(Y|U). \tag{24}$$

So, we get

$$\max_{P_{U|X} \,:\, I(U;X) \leq R} I(U; Y) = H(Y) - R_Y^\star(R).$$

D. The Privacy Funnel

Consider again the setting of Figure 3; and let us assume that the pair (Y, X) models data that a user possesses and which has the following properties: the data Y is some sensitive (private) data that is not meant to be revealed at all, or else not beyond some level Δ; and the data X is non-private and is meant to be shared with another user (analyst). Because X and Y are correlated, sharing the non-private data X with the analyst possibly reveals information about Y. For this reason, there is a tradeoff between the amount of information that the user shares about X and the information that he keeps private about Y. The data X is passed through a randomized mapping ϕ whose purpose is to make $U = \phi(X)$ maximally informative about X while being minimally informative about Y.
The analyst performs an inference attack on the private data Y based on the disclosed information U. Let $\ell : \mathcal{Y} \times \hat{\mathcal{Y}} \longrightarrow \bar{\mathbb{R}}$ be an arbitrary loss function with reconstruction alphabet $\hat{\mathcal{Y}}$ that measures the cost of inferring Y after observing U. Given $(X, Y) \sim P_{X,Y}$ and under the given loss function ℓ, it is natural to quantify the difference between the prediction losses in predicting $Y \in \mathcal{Y}$ prior and after observing $U = \phi(X)$. Let

$$C(\ell, P) = \inf_{\hat{y} \in \hat{\mathcal{Y}}} \mathbb{E}_P[\ell(Y, \hat{y})] - \inf_{\hat{Y}(\phi(X))} \mathbb{E}_P[\ell(Y, \hat{Y})] \tag{25}$$

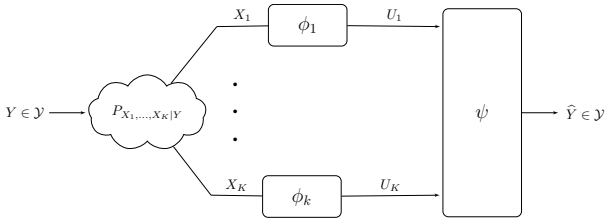

Fig. 4. A model for distributed, e.g., multi-view, learning.

where $\hat{y} \in \hat{\mathcal{Y}}$ is deterministic and $\hat{Y}(\phi(X))$ is any measurable function of $U = \phi(X)$. The quantity $C(\ell, P)$ quantifies the reduction in the prediction loss under the loss function ℓ that is due to observing $U = \phi(X)$, i.e., the inference cost gain. In [23] (see also [24]) it is shown that that under some mild conditions the inference cost gain $C(\ell, P)$ as defined by (25) is upper-bounded as

$$C(\ell, P) \leq 2\sqrt{2}L\sqrt{I(U;Y)} \tag{26}$$

where L is a constant. The inequality (26) holds irrespective to the choice of the loss function ℓ; and this justifies the usage of the logarithmic loss function as given by (2) in the context of finding a suitable tradeoff between utility and privacy, since

$$I(U;Y) = H(Y) - \inf_{\hat{Y}(U)} \mathbb{E}_P[\ell_{\log}(Y, \hat{Y})]. \tag{27}$$

Under the logarithmic loss function, the design of the mapping $U = \phi(X)$ should strike a right balance between the utility for inferring the non-private data X as measured by the mutual information $I(U;X)$ and the privacy metric about the private date Y as measured by the mutual information $I(U;Y)$.

E. Efficiency of Investment Information

Let Y model a stock market data and X some correlated information. In [25], Erkip and Cover investigated how the description of the correlated information X improves the investment in the stock market Y. Specifically, let $\Delta(C)$ denote the maximum increase in growth rate when X is described to the investor at rate C. Erkip and Cover found a single-letter characterization of the incremental growth rate $\Delta(C)$. When specialized to the horse race market, this problem is related to the aforementioned source coding with side information of Wyner [21] and Ahlswede-Körner [22]; and, so, also to the IB problem. The work [25] provides explicit analytic solutions for two horce race examples, jointly binary and jointly Gaussian horce races.

III. Distributed Learning

Consider now a generalization of the IB problem in which the prediction is to be performed in a distributed manner. The model is shown in Figure 4. Here, the prediction of the target variable $Y \in \mathcal{Y}$ is to be performed on the basis of samples of statistically correlated random variables $(X_1, \cdots, X_K)$ that are observed each at a distinct predictor. Throughout, we assume that the following Markov chain holds for all $k \in \mathcal{K} := \{1, \cdots, K\}$,

$$X_k \multimap Y \multimap X_{\mathcal{K}/k}. \tag{28}$$

The variable Y is a target variable and we seek to characterize how accurate it can be predicted from a measurable random vector $(X_1, \cdots, X_K)$ when the components of this vector are processed separately, each by a distinct encoder.

A. Optimal relevance-complexity tradeoff region

The distributed IB problem of Figure 4 is studied in [26], [27] from information-theoretic grounds. For both discrete memoryless (DM) and memoryless vector Gaussian models, the authors establish fundamental limits of learning in terms of optimal tradeoffs between relevance and complexity. The following theorem [26], [27] states the result for the case of discrete memoryless sources.

Theorem 1. *The relevance-complexity region $\mathcal{IR}_{\mathrm{DIB}}$ of the distributed learning problem is given by the union of all non-negative tuples $(\Delta, R_1, \ldots, R_K) \in \mathbb{R}_+^{K+1}$ that satisfy*

$$\Delta \leq \sum_{k \in \mathcal{S}} [R_k - I(X_k; U_k|Y,T)] + I(Y; U_{\mathcal{S}^c}|T), \quad \forall \mathcal{S} \subseteq \mathcal{K} \tag{29}$$

for some joint distribution of the form $P_T P_Y \prod_{k=1}^K P_{X_k|Y} \prod_{k=1}^K P_{U_k|X_k,T}$.

B. A Variational Bound

Let us consider the problem of maximizing the relevance under a sum-complexity constraint. Let $R_{\mathrm{sum}} = \sum_{k=1}^K R_k$ and

$$\mathcal{RI}_{\mathrm{DIB}}^{\mathrm{sum}} := \Big\{(\Delta, R_{\mathrm{sum}}) \in \mathbb{R}_+^2 : \exists (R_1, \ldots, R_K) \in \mathbb{R}_+^K \text{ s.t. } \sum_{k=1}^K R_k = R_{\mathrm{sum}} \text{ and } (\Delta, R_1, \ldots, R_K) \in \mathcal{RI}_{\mathrm{DIB}}\Big\}. \tag{30}$$

It is easy to see that the region $\mathcal{RI}_{\mathrm{DIB}}^{\mathrm{sum}}$ is composed of all the pairs $(\Delta, R_{\mathrm{sum}}) \in \mathbb{R}_+^2$ for which $\Delta \leq \Delta(R_{\mathrm{sum}}, P_{X_{\mathcal{K}},Y})$, with

$$\Delta(R_{\mathrm{sum}}, P_{X_{\mathcal{K}},Y}) = \max_{\mathbf{P}} \min \left\{ I(Y; U_{\mathcal{K}}), R_{\mathrm{sum}} - \sum_{k=1}^K I(X_k; U_k|Y) \right\}, \tag{31}$$

where the maximization is over joint distributions that factorize as $P_Y \prod_{k=1}^K P_{X_k|Y} \prod_{k=1}^K P_{U_k|X_k}$. The pairs $(\Delta, R_{\mathrm{sum}})$ that lie on the boundary of $\mathcal{RI}_{\mathrm{DIB}}^{\mathrm{sum}}$ can be characterized as given in the following proposition [27, Section 7.3].

Proposition 1. *For every pair $(\Delta, R_{\mathrm{sum}}) \in \mathbb{R}_+^2$ that lies on the boundary of the region $\mathcal{RI}_{\mathrm{DIB}}^{\mathrm{sum}}$ there exists a parameter $s \geq 0$ such that $(\Delta, R_{\mathrm{sum}}) = (\Delta_s, R_s)$, with*

$$\Delta_s = \frac{1}{(1+s)} \Big[(1+sK)H(Y) + sR_s + \max_{\mathbf{P}} \mathcal{L}_s(\mathbf{P}) \Big], \tag{32}$$

$$R_s = I(Y; U_{\mathcal{K}}^*) + \sum_{k=1}^K [I(X_k; U_k^*) - I(Y; U_k^*)], \tag{33}$$

where $\mathbf{P}^*$ is the set of conditional pmfs $\mathbf{P} = \{P_{U_1|X_1}, \cdots, P_{U_K|X_K}\}$ that maximize the cost function

$$\mathcal{L}_s(\mathbf{P}) := -H(Y|U_{\mathcal{K}}) - s \sum_{k=1}^K [H(Y|U_k) + I(X_k; U_k)]. \tag{34}$$

The optimization of (34) generally requires to compute marginal distributions that involve the descriptions $U_1, \cdots, U_K$, which might not be possible in practice. In what follows, we derive a variational lower bound on $\mathcal{L}_s(\mathbf{P})$ on the DIB cost function in terms of families of stochastic mappings $Q_{Y|U_1,\ldots,U_K}$ (a decoder), $\{Q_{Y|U_k}\}_{k=1}^K$ and priors $\{Q_{U_k}\}_{k=1}^K$. For the simplicity of the notation, we let

$$\mathbf{Q} := \{Q_{Y|U_1,\ldots,U_K}, Q_{Y|U_1}, \ldots, Q_{Y|U_K}, Q_{U_1}, \ldots, Q_{U_K}\}. \tag{35}$$

Let

$$\mathcal{L}_s^{\mathrm{VB}}(\mathbf{P}, \mathbf{Q}) := \underbrace{\mathbb{E}[\log Q_{Y|U_{\mathcal{K}}}(Y|U_{\mathcal{K}})]}_{\text{av. logarithmic-loss}}$$

$$+ s \underbrace{\sum_{k=1}^{K} \Big(\mathbb{E}[\log Q_{Y|U_k}(Y|U_k)] - D_{\mathrm{KL}}(P_{U_k|X_k} \| Q_{U_k}) \Big)}_{\text{regularizer}}. \quad (36)$$

Lemma 1. *([27, Section 7.4]) For fixed* $\mathbf{P}$*, we have*

$$\mathcal{L}_s(\mathbf{P}) \geq \mathcal{L}_s^{\mathrm{VB}}(\mathbf{P}, \mathbf{Q}), \qquad \text{for all pmfs } \mathbf{Q}. \quad (37)$$

In addition, there exists a unique $\mathbf{Q}$ *that achieves the maximum* $\max_{\mathbf{Q}} \mathcal{L}_s^{\mathrm{VB}}(\mathbf{P}, \mathbf{Q}) = \mathcal{L}_s(\mathbf{P})$*, and is given by,* $\forall k \in \mathcal{K}$*,*

$$Q^*_{U_k} = P_{U_k} \quad (38a)$$
$$Q^*_{Y|U_k} = P_{Y|U_k} \quad (38b)$$
$$Q^*_{Y|U_1,\ldots,U_k} = P_{Y|U_1,\ldots,U_K}, \quad (38c)$$

where the marginals P_{U_k} *and the conditional marginals* $P_{Y|U_k}$ *and* $P_{Y|U_1,\ldots,U_K}$ *are computed from* $\mathbf{P}$*.*

C. Vector Gaussian Model

In this section, we show that for the jointly vector Gaussian data model it is enough to restrict to Gaussian auxiliaries $(\mathbf{U}_1, \cdots, \mathbf{U}_K)$ in order to exhaust the entire relevance-complexity region. Also, we provide an explicit analytical expression of this region. Let $(\mathbf{X}_1, \ldots, \mathbf{X}_K, \mathbf{Y})$ be a jointly vector Gaussian vector that satisfies the Markov chain (28). Without loss of generality, let the target variable be a complex-valued, zero-mean multivariate Gaussian $\mathbf{Y} \in \mathbb{C}^{n_y}$ with covariance matrix $\mathbf{\Sigma_y}$, i.e., $\mathbf{Y} \sim \mathcal{CN}(\mathbf{y}; \mathbf{0}, \mathbf{\Sigma_y})$, and $\mathbf{X}_k \in \mathbb{C}^{n_k}$ given by

$$\mathbf{X}_k = \mathbf{H}_k \mathbf{Y} + \mathbf{N}_k, \quad (39)$$

where $\mathbf{H}_k \in \mathbb{C}^{n_k \times n_y}$ models the linear model connecting $\mathbf{Y}$ to the observation at encoder k, and $\mathbf{N}_k \in \mathbb{C}^{n_k}$ is the noise vector at encoder k, assumed to be Gaussian with zero-mean and covariance matrix $\mathbf{\Sigma}_k$, and independent from all other noises and $\mathbf{Y}$.

The following theorem [27, Section 7.5] characterizes the relevance-complexity region of the model (39), which we denote hereafter as $\mathcal{RI}_{\mathrm{DIB}}^{\mathrm{G}}$. The theorem also shows that in order to exhaust this region it is enough to restrict to no time sharing, i.e., $T = \varnothing$ and multivariate Gaussian test channels

$$\mathbf{U}_k = \mathbf{A}_k \mathbf{X}_k + \mathbf{Z}_k \sim \mathcal{CN}(\mathbf{u}_k; \mathbf{A}_k \mathbf{X}_k, \mathbf{\Sigma}_{z,k}), \quad (40)$$

where $\mathbf{A}_k \in \mathbb{C}^{n_k \times n_k}$ projects the observation $\mathbf{X}_k$ and $\mathbf{Z}_k$ is a zero-mean Gaussian noise with covariance $\mathbf{\Sigma}_{z,k}$.

Theorem 2. *For the model* (39) *the region* $\mathcal{RI}_{\mathrm{DIB}}^{\mathrm{G}}$ *is given by the union of all tuples* $(\Delta, R_1, \ldots, R_L)$ *that satisfy* $\forall \mathcal{S} \subseteq \mathcal{K}$

$$\Delta \leq \sum_{k \in \mathcal{S}} \Big(R_k + \log \Big| \mathbf{I} - \mathbf{\Sigma}_k^{1/2} \mathbf{\Omega}_k \mathbf{\Sigma}_k^{1/2} \Big| \Big) + \log \Big| \mathbf{I} + \sum_{k \in \mathcal{S}^c} \mathbf{\Sigma}_{\mathbf{y}}^{1/2} \mathbf{H}_k^\dagger \mathbf{\Omega}_k \mathbf{H}_k \mathbf{\Sigma}_{\mathbf{y}}^{1/2} \Big|$$

for some matrices $\mathbf{0} \preceq \mathbf{\Omega}_k \preceq \mathbf{\Sigma}_k^{-1}$*.*

Acknowledgment: The work of S. Shamai has been supported by the European Union's Horizon 2020 Research And Innovation Programme, grant agreement no. 694630.

References

[1] D. Erdogmus, "Information theoretic learning: Renyi's entropy and its applications to adaptive system training," Ph.D. dissertation, University of Florida Gainesville, Florida, 2002.

[2] J. C. Principe, N. R. Euliano, and W. C. Lefebvre, *Neural and adaptive systems: fundamentals through simulations*. Wiley New York, 2000, vol. 672.

[3] A. Painsky and G. W. Wornell, "On the universality of the logistic loss function," *arXiv preprint arXiv:1805.03804*, 2018.

[4] J. Jiao, T. A. Courtade, K. Venkat, and T. Weissman, "Justification of logarithmic loss via the benefit of side information," *IEEE Transactions on Information Theory*, vol. 61, no. 10, pp. 5357–5365, 2015.

[5] R. Linsker, "Self-organization in a perceptual network," *Computer*, vol. 21, no. 3, pp. 105–117, 1988.

[6] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning and games*. New York, USA: Cambridge, Univ. Press, 2006.

[7] E. L. Lehmann and G. Casella, *Theory of point estimation*. Springer Science & Business Media, 2006.

[8] R.-L. Dobrushin and B. S. Tsybakov, "Information transmission with additional noise," *IRE Trans. on Info. Theory*, vol. 85, pp. 293–304, 1962.

[9] H.-S. Witsenhausen, "Indirect rate distortion problems," *IEEE Trans. on Info. Theory*, vol. IT-26, pp. 518–521, Sep. 1980.

[10] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. London, U. K.: Academic Press, 1981.

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

[12] P. Harremoes and N. Tishby, "The information bottleneck revisited or how to choose a good distortion measure," in *Proc. IEEE Int. Symp. Information Theory*, Jun. 2007, pp. 566–570.

[13] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, Jul 1972.

[14] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. IT-18, no. 1, pp. 12 – 20, Jan. 1972.

[15] A. Alemi, I. Fischer, J. Dillon, and K. Murphy, "Deep variational information bottleneck," in *ICLR*, 2017. [Online]. Available: https://arxiv.org/abs/1612.00410

[16] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. 22, pp. 1–10, Jan. 1976.

[17] Y. Steinberg, "Coding and common reconstruction," *IEEE Trans. Inf. Theory*, vol. IT-11, pp. 4995–5010, Nov. 2009.

[18] M. Benammar and A. Zaidi, "Rate-distortion of a heegard-berger problem with common reconstruction constraint," in *Proc. of International Zurich Seminar on Information and Communication*. IEEE, Mar. 2016.

[19] ——, "Rate-distortion function for a heegard-berger problem with two sources and degraded reconstruction sets," *IEEE Transactions on Information Theory*, vol. 62, no. 9, pp. 5080–5092, 2016.

[20] I. Sutskover, S. Shamai, and J. Ziv, "Extremes of information combining," *IEEE Trans. Inform. Theory*, vol. 51, no. 04, pp. 1313–1325, 2005.

[21] A. D. Wyner, "On source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. IT-21, pp. 294–300, May 1975.

[22] R. Ahlswede and J. Korner, "Source coding with side information and a converse for degraded broadcast channels," *IEEE Trans. Inf. Theory*, vol. 21, no. 6, pp. 629–637, November 1975.

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

[24] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder, "Information extraction under privacy constraints," *IEEE Trans. Info. Theory*, vol. 65, no. 03, pp. 1512–1534, Mar. 2019.

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

[26] I. E. Aguerri and A. Zaidi, "Distributed information bottleneck method for discrete and Gaussian sources," in *Proc. of Int. Zurich Seminar on Information and Communication, IZS*, Zurich, Switzerland, 2018.

[27] ——, "Distributed variational representation learning," *IEEE Trans. on Pattern Analysis and Machine Intelligence. To appear. Available at https://arxiv.org/abs/1807.04193*, 2018.