MDPI

Tutorial

On the Information Bottleneck Problems: Models, Connections, Applications and Information Theoretic Views

Abdellatif Zaidi [1,2,*] **and Iñaki Estella-Aguerri** [2] **and Shlomo Shamai (Shitz)** [3]

[1]   Institut d'Électronique et d'Informatique Gaspard-Monge, Université Paris-Est,
      77454 Champs-sur-Marne, France
[2]   Mathematics and Algorithmic Sciences Lab, Paris Research Center, Huawei Technologies France, 92100
      Boulogne-Billancourt , France; inaki.estella@gmail.com
[3]   Technion Institute of Technology, Technion City, Haifa 32000, Israel; sshlomo@ee.technion.ac.il
[*]   Correspondence: abdellatif.zaidi@u-pem.fr

Abstract: This tutorial paper focuses on the variants of the bottleneck problem taking an information theoretic perspective and discusses practical methods to solve it, as well as its connection to coding and learning aspects. The intimate connections of this setting to remote source-coding under logarithmic loss distortion measure, information combining, common reconstruction, the Wyner–Ahlswede–Korner problem, the efficiency of investment information, as well as, generalization, variational inference, representation learning, autoencoders, and others are highlighted. We discuss its extension to the distributed information bottleneck problem with emphasis on the Gaussian model and highlight the basic connections to the uplink Cloud Radio Access Networks (CRAN) with oblivious processing. For this model, the optimal trade-offs between relevance (i.e., information) and complexity (i.e., rates) in the discrete and vector Gaussian frameworks is determined. In the concluding outlook, some interesting problems are mentioned such as the characterization of the optimal inputs ("features") distributions under power limitations maximizing the "relevance" for the Gaussian information bottleneck, under "complexity" constraints.

Keywords: information bottleneck; rate distortion theory; logarithmic loss; representation learning

1. Introduction

A growing body of works focuses on developing learning rules and algorithms using information theoretic approaches (e.g., see [1–6] and references therein). Most relevant to this paper is the Information Bottleneck (IB) method of Tishby *et al.* [1], which seeks the right balance between data fit and generalization by using the mutual information as both a cost function and a regularizer. Specifically, IB formulates the problem of extracting the relevant information that some signal $X \in \mathcal{X}$ provides about another one $Y \in \mathcal{Y}$ that is of interest as that of finding 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 $I(U;X)$). In the IB framework, $I(U;Y)$ is referred to as the *relevance* of U and $I(U;X)$ is referred to as the *complexity* of U, where complexity here is measured by the minimum description length (or rate) at which the observation is compressed. Accordingly, the performance of learning with the IB method and the optimal mapping of the data are found by solving the Lagrangian formulation

$$\mathcal{L}_{\beta}^{\text{IB},*} := \max_{P_{U|X}} I(U;Y) - \beta I(U;X), \tag{1}$$

where $P_{U|X}$ is a stochastic map that assigns the observation X to a representation U from which Y is inferred and β is the Lagrange multiplier. Several methods, which we detail below, have been proposed to obtain solutions $P_{U|X}$ to the IB problem in Equation (4) in several scenarios, e.g., when the distribution of the sources (X, Y) is perfectly known or only samples from it are available.

The IB approach, as a method to both characterize performance limits as well as to design mapping, has found remarkable applications in supervised and unsupervised learning problems such as classification, clustering, and prediction. Perhaps key to the analysis and theoretical development of the IB method is its elegant connection with information-theoretic rate-distortion problems, as it is now well known that the IB problem is essentially a remote source coding problem [7–9] in which the distortion is measured under logarithmic loss. Recent works show that this connection turns out to be useful for a better understanding the fundamental limits of learning problems, including the performance of deep neural networks (DNN) [10], the emergence of invariance and disentanglement in DNN [11], the minimization of PAC-Bayesian bounds on the test error [11,12], prediction [13,14], or as a generalization of the evidence lower bound (ELBO) used to train variational auto-encoders [15,16], geometric clustering [17], or extracting the Gaussian "part" of a signal [18], among others. Other connections that are more intriguing exist also with seemingly unrelated problems such as privacy and hypothesis testing [19–21] or multiterminal networks with oblivious relays [22,23] and non-binary LDPC code design [24]. More connections with other coding problems such as the problems of information combining and common reconstruction, the Wyner–Ahlswede–Korner problem, and the efficiency of investment information are unveiled and discussed in this tutorial paper, together with extensions to the distributed setting.

The abstract viewpoint of IB also seems instrumental to a better understanding of the so-called *representation learning* [25], which is an active research area in machine learning that focuses on identifying and disentangling the underlying explanatory factors that are hidden in the observed data in an attempt to render learning algorithms less dependent on feature engineering. More specifically, one important question, which is often controversial in statistical learning theory, is the choice of a "good" loss function that measures discrepancies between the true values and their estimated fits. There is however numerical evidence that models that are trained to maximize mutual information, or equivalently minimize the error's entropy, often outperform ones that are trained using other criteria such as mean-square error (MSE) and higher-order statistics [26,27]. On this aspect, we also mention Fisher's dissertation [28], which contains investigation of the application of information theoretic metrics to blind source separation and subspace projection using Renyi's entropy as well as what appears to be the first usage of the now popular Parzen windowing estimator of information densities in the context of learning. Although a complete and rigorous justification of the usage of mutual information as cost function in learning is still awaited, recently, a partial explanation appeared in [29], where the authors showed that under some natural data processing property Shannon's mutual information uniquely quantifies the reduction of prediction risk due to side information. Along the same line of work, Painsky and Wornell [30] showed that, for binary classification problems, by minimizing the logarithmic-loss (log-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. Perhaps, this justifies partially why mutual information (or, equivalently, the corresponding loss function, which is the log-loss fidelity measure) is widely used in learning theory and has already been adopted in many algorithms in practice such as the *infomax* criterion [31], the tree-based algorithm of Quinlan [32], or the well known Chow–Liu algorithm [33] for learning tree graphical models, with various applications in genetics [34], image processing [35], computer vision [36], etc. The logarithmic loss measure also plays a central role in the theory of prediction [37, Ch. 09] where it is often referred to as the *self-information* loss function, as well as in Bayesian modeling [38] where priors are usually designed to maximize the mutual information between the parameter to be estimated and the observations. The goal of learning, however, is not merely to learn model parameters accurately for previously seen data. Rather, in essence, it is the ability to successfully apply rules that are extracted from previously seen data to

characterize new unseen data. This is often captured through the notion of "generalization error". The generalization capability of a learning algorithm hinges on how sensitive the output of the algorithm is to modifications of the input dataset, i.e., its *stability* [39,40]. In the context of deep learning, it can be seen as a measure of how much the algorithm overfits the model parameters to the seen data. In fact, efficient algorithms should strike a good balance between their ability to fit training dataset and that to generalize well to unseen data. In statistical learning theory [37], such a dilemma is reflected through that the minimization of the "population risk" (or "test error" in the deep learning literature) amounts to the minimization of the sum of the two terms that are generally difficult to minimize simultaneously, the "empirical risk" on the training data and the generalization error. To prevent over-fitting, regularization methods can be employed, which include parameter penalization, noise injection, and averaging over multiple models trained with distinct sample sets. Although it is not yet very well understood how to optimally control model complexity, recent works [41,42] show that the generalization error can be upper-bounded using the mutual information between the input dataset and the output of the algorithm. This result actually formalizes the intuition that the less information a learning algorithm extracts from the input dataset the less it is likely to overfit, and justifies, partly, the use of mutual information also as a regularizer term. The interested reader may refer to [43] where it is shown that regularizing with mutual information alone does not always capture all desirable properties of a latent representation. We also point out that there exists an extensive literature on building optimal estimators of information quantities (e.g. entropy, mutual information), as well as their Matlab/Python implementations, including in the high-dimensional regime (see, e.g., [44–49] and references therein).

This paper provides a review of the information bottleneck method, its classical solutions, and recent advances. In addition, in the paper, we unveil some useful connections with coding problems such as 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 learning problems such as inference, generalization, and representation learning. Leveraging these connections, we discuss its extension to the distributed information bottleneck problem with emphasis on its solutions and the Gaussian model and highlight the basic connections to the uplink Cloud Radio Access Networks (CRAN) with oblivious processing. For this model, the optimal trade-offs between relevance and complexity in the discrete and vector Gaussian frameworks is determined. In the concluding outlook, some interesting problems are mentioned such as the characterization of the optimal inputs distributions under power limitations maximizing the "relevance" for the Gaussian information bottleneck under "complexity" constraints.

Notation

Throughout, uppercase letters denote random variables, e.g., X; lowercase letters denote realizations of random variables, e.g., x; and calligraphic letters denote sets, e.g., $\mathcal{X}$. The cardinality of a set is denoted by $|\mathcal{X}|$. For a random variable X with probability mass function (pmf) P_X, we use $P_X(x) = p(x)$, $x \in \mathcal{X}$ for short. Boldface uppercase letters denote vectors or matrices, e.g., $\mathbf{X}$, where context should make the distinction clear. For random variables $(X_1, X_2, \cdots)$ and a set of integers $\mathcal{K} \subseteq \mathbb{N}$, $X_{\mathcal{K}}$ denotes the set of random variables with indices in the set $\mathcal{K}$, i.e., $X_{\mathcal{K}} = \{X_k : k \in \mathcal{K}\}$. If $\mathcal{K} = \varnothing$, $X_{\mathcal{K}} = \varnothing$. For $k \in \mathcal{K}$, we let $X_{\mathcal{K}/k} = (X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_K)$, and assume that $X_0 = X_{K+1} = \varnothing$. In addition, for zero-mean random vectors $\mathbf{X}$ and $\mathbf{Y}$, the quantities $\mathbf{\Sigma_x}$, $\mathbf{\Sigma_{x,y}}$ and $\mathbf{\Sigma_{x|y}}$ denote, respectively, the covariance matrix of the vector $\mathbf{X}$, the covariance matrix of vector $(\mathbf{X}, \mathbf{Y})$, and the conditional covariance matrix of $\mathbf{X}$, conditionally on $\mathbf{Y}$, i.e., $\mathbf{\Sigma_x} = \mathrm{E}[\mathbf{X}\mathbf{X}^H]$ $\mathbf{\Sigma_{x,y}} := \mathrm{E}[\mathbf{X}\mathbf{Y}^H]$, and $\mathbf{\Sigma_{x|y}} = \mathbf{\Sigma_x} - \mathbf{\Sigma_{x,y}}\mathbf{\Sigma_y^{-1}}\mathbf{\Sigma_{y,x}}$. Finally, for two probability measures P_X and Q_X on the random variable $X \in \mathcal{X}$, the relative entropy or Kullback–Leibler divergence is denoted as $D_{\mathrm{KL}}(P_X \| Q_X)$. That is, if P_X is absolutely continuous with respect to Q_X, $P_X \ll Q_X$ (i.e., for every $x \in \mathcal{X}$, if $P_X(x) > 0$, then $Q_X(x) > 0$), $D_{\mathrm{KL}}(P_X \| Q_X) = \mathbb{E}_{P_X}[\log(P_X(X)/Q_X(X))]$, otherwise $D_{\mathrm{KL}}(P_X \| Q_X) = \infty$.

2. the Information Bottleneck Problem

The Information Bottleneck (IB) method was introduced by Tishby et al. [1] 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, as shown in Figure 1.
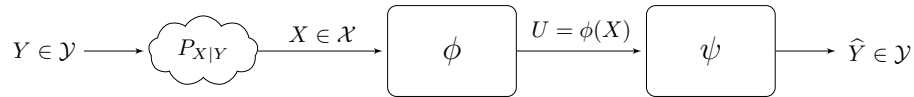

Figure 1. Information bottleneck problem.

Specifically, the IB method consists of finding the stochastic mapping $P_{U|X} : \mathcal{X} \to \mathcal{U}$ that from an observation X outputs a representation $U \in \mathcal{U}$ 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 $I(U;X)$ (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 [50].) The auxiliary random variable U satisfies that $U \multimap X \multimap Y$ is a Markov Chain in this order; that is, that the joint distribution of (X, U, Y) satisfies

$$p(x, u, y) = p(x)p(y|x)p(u|x), \tag{2}$$

and the mapping $P_{U|X}$ is chosen such that U strikes 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 $I(U;X)$. In particular, such U, or effectively the mapping $P_{U|X}$, can be determined to maximize the IB-Lagrangian defined as

$$\mathcal{L}_\beta^{\mathrm{IB}}(P_{U|X}) := I(U;Y) - \beta I(U;X) \tag{3}$$

over all mappings $P_{U|X}$ that satisfy $U \multimap X \multimap Y$ and the trade-off parameter β is a positive Lagrange multiplier associated with the constraint on $I(U;Y)$.

Accordingly, for a given β and source distribution $P_{X,Y}$, the optimal mapping of the data, denoted by $P_{U|X}^{*,\beta}$, is found by solving the IB problem, defined as

$$\mathcal{L}_\beta^{\mathrm{IB},*} := \max_{P_{U|X}} I(U;Y) - \beta I(U;X)., \tag{4}$$

over all mappings $P_{U|Y}$ that satisfy $U \multimap X \multimap Y$. It follows from the classical application of Carathéodory's theorem [51] that without loss of optimality, U can be restricted to satisfy $|\mathcal{U}| \le |\mathcal{X}| + 1$.

In Section 3 we discuss several methods to obtain solutions $P_{U|X}^{*,\beta}$ to the IB problem in Equation (4) in several scenarios, e.g., when the distribution of (X, Y) is perfectly known or only samples from it are available.
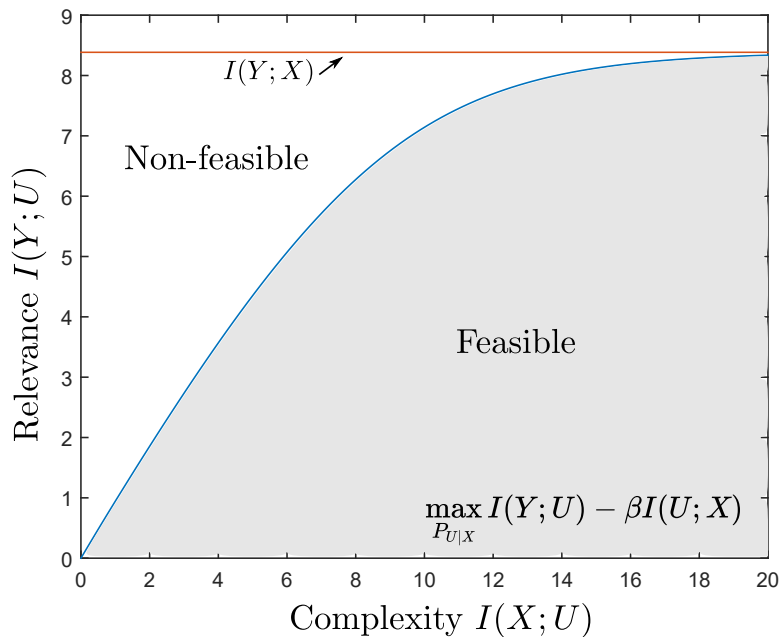
2.1. the Ib Relevance–Complexity Region

The minimization of the IB-Lagrangian $\mathcal{L}_\beta$ in Equation (4) for a given $\beta \ge 0$ and $P_{X,Y}$ results in an optimal mapping $P_{U|X}^{*,\beta}$ and a relevance–complexity pair (Δ_β, R_β) where $\Delta_\beta = I(U_\beta, X)$ and $R_\beta = I(U_\beta, Y)$ are, respectively, the relevance and the complexity resulting from generating U_β with the solution $P_{U|X}^{*,\beta}$. By optimizing over all $\beta \ge 0$, the resulting relevance–complexity pairs (Δ_β, R_β) characterize the boundary of the region of simultaneously achievable relevance–complexity pairs for a

distribution $P_{X,Y}$ (see Figure 2). In particular, for a fixed $P_{X,Y}$, we define this region as the union of relevance–complexity pairs (Δ, R) that satisfy

$$\Delta \le I(U, Y), \quad R \ge I(X, U) \tag{5}$$

where the union is over all $P_{U|X}$ such that U satisfies $U \multimap X \multimap Y$ form a Markov Chain in this order. Any pair (Δ, R) outside of this region is not simultaneously achievable by any mapping $P_{U|X}$.

Figure 2. Information bottleneck relevance–complexity region. For a given β, the solution $P_{U|X}^{*,\beta}$ to the minimization of the IB-Lagrangian in Equation (3) results in a pair (Δ_β, R_β) on the boundary of the IB relevance–complexity region (colored in grey).

3. Solutions to the Information Bottleneck Problem

As shown in the previous region, the IB problem provides a methodology to design mappings $P_{U|X}$ performing at different relevance–complexity points within the region of feasible (Δ, R) pairs, characterized by the IB relevance–complexity region, by minimizing the IB-Lagrangian in Equation (3) for different values of β. However, in general, this optimization is challenging as it requires computation of mutual information terms.

In this section, we describe how, for a fixed parameter β, the optimal solution $P_{U|X}^{\beta,*}$, or an efficient approximation of it, can be obtained under: (i) particular distributions, e.g., Gaussian and binary symmetric sources; (ii) known general discrete memoryless distributions; and (iii) unknown memory distributions and only samples are available.

3.1. Solution for Particular Distributions: Gaussian and Binary Symmetric Sources

In certain cases, when the joint distribution $P_{X,Y}$ is know, e.g., it is binary symmetric or Gaussian, information theoretic inequalities can be used to minimize the IB-Lagrangian in (4) in closed form.

3.1.1. Binary IB

Let X and Y be a doubly symmetric binary sources (DSBS), i.e., $(X, Y) \sim \text{DSBS}(p)$ for some $0 \le p \le 1/2$. (A DSBS is a pair (X, Y) of binary random variables $X \sim \text{Bern}(1/2)$ and $Y \sim \text{Bern}(1/2)$ and $X \oplus Y \sim \text{Bern}(p)$, where $\oplus$ is the sum modulo 2. That is, Y is the output of a binary symmetric channel with crossover probability p corresponding to the input X, and X is the output of the same channel

with input Y.) Then, it can be shown that the optimal U in (4) is such that $(X, U) \sim \text{DSBS}(q)$ fo $0 \le q \le 1$. Such a U can be obtained with the mapping $P_{U|X}$ such that

$$U = X \oplus Q, \quad \text{with } Q \sim \text{DSBS}(q). \tag{6}$$

In this case, straightforward algebra leads to that the complexity level is given by

$$I(U; X) = 1 - h_2(q), \tag{7}$$

where, for $0 \le x \le 1$, $h_2(x)$ is the entropy of a Bernoulli-(x) source, i.e., $h_2(x) = -x \log_2(x) - (1 - x) \log_2(1 - x)$, and the relevance level is given by

$$I(U; Y) = 1 - h_2(p \star q) \tag{8}$$

where $p \star q = p(1 - q) + q(1 - p)$. The result extends easily to discrete symmetric mappings $Y \longrightarrow X$ with binary X (one bit output quantization) and discrete non-binary Y.

3.1.2. Vector Gaussian IB

Let $(\mathbf{X}, \mathbf{Y}) \in \mathbb{C}^{N_x} \times \mathbb{C}^{N_y}$ be a pair of jointly Gaussian, zero-mean, complex-valued random vectors, of dimension $N_x > 0$ and $N_y > 0$, respectively. In this case, the optimal solution of the IB-Lagrangian in Equation (3) (i.e., test channel $P_{U|X}$) is a noisy linear projection to a subspace whose dimensionality is determined by the tradeoff parameter β. The subspaces are spanned by basis vectors in a manner similar to the well known canonical correlation analysis [52]. For small β, only the vector associated to the dimension with more energy, i.e., corresponding to the largest eigenvalue of a particular hermitian matrix, will be considered in U. As β increases, additional dimensions are added to U through a series of critical points that are similar to structural phase transitions. This process continues until U becomes rich enough to capture all the relevant information about Y that is contained in X. In particular, the boundary of the optimal relevance–complexity region was shown in [53] to be achievable using a test channel $P_{\mathbf{U}|\mathbf{X}}$, which is such that $(\mathbf{U}, \mathbf{X})$ is Gaussian. Without loss of generality, let

$$\mathbf{U} = \mathbf{A}\mathbf{X} + \boldsymbol{\zeta} \tag{9}$$

where $\mathbf{A} \in \mathcal{M}_{N_u, N_x}(\mathbb{C})$ is an $N_u \times N_x$ complex valued matrix and $\boldsymbol{\zeta} \in \mathbb{C}^{N_u}$ is a Gaussian noise that is independent of $(\mathbf{X}, \mathbf{Y})$ with zero-mean and covariance matrix $\mathbf{I}_{N_u}$. For a given non-negative trade-off parameter β, the matrix $\mathbf{A}$ has a number of rows that depends on β and is given by [54, Theorem 3.1]

$$\mathbf{A} = \begin{cases} \left[\mathbf{0}^T; \ldots; \mathbf{0}^T\right], & 0 \le \beta < \beta_1^c \\ \left[\alpha_1 \mathbf{v}_1^T; \mathbf{0}^T; \ldots; \mathbf{0}^T\right], & \beta_1^c \le \beta < \beta_2^c \\ \left[\alpha_1 \mathbf{v}_1^T; \alpha_2 \mathbf{v}_2^T; \mathbf{0}^T; \ldots; \mathbf{0}^T\right], & \beta_2^c \le \beta < \beta_3^c \\ \vdots \end{cases} \tag{10}$$

where $\{\mathbf{v}_1^T, \mathbf{v}_2^T, \cdots, \mathbf{v}_{N_x}^T\}$ are the left eigenvectors of $\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}} \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}$ sorted by their corresponding ascending eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_{N_x}$. Furthermore, for $i = 1, \cdots, N_x$, $\beta_i^c = \frac{1}{1-\lambda_i}$ are critical β-values, $\alpha_i = \sqrt{\frac{\beta(1-\lambda_i)-1}{\lambda_i r_i}}$ with $r_i = \mathbf{v}_i^T \boldsymbol{\Sigma}_{\mathbf{X}} \mathbf{v}_i$, $\mathbf{0}^T$ denotes the N_x-dimensional zero vector and semicolons separate the rows of the matrix. It is interesting to observe that the optimal projection consists of eigenvectors of $\boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{Y}} \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}$, combined in a judicious manner: for values of β that are smaller than β_1^c, reducing complexity is of prime importance, yielding extreme compression $\mathbf{U} = \boldsymbol{\zeta}$, i.e., independent noise and no information preservation at all about $\mathbf{Y}$. As β increases, it undergoes a series of critical points $\{\beta_i^c\}$, at each of which a new eignevector is added to the matrix $\mathbf{A}$, yielding a more complex but richer representation—the rank of $\mathbf{A}$ increases accordingly.

For the specific case of scalar Gaussian sources, that is $N_x = N_y = 1$, e.g., $X = \sqrt{\text{snr}}Y + N$ where N is standard Gaussian with zero-mean and unit variance, the above result simplifies considerably. In this case, let without loss of generality the mapping $P_{U|X}$ be given by

$$X = \sqrt{a}X + Q \tag{11}$$

where Q is standard Gaussian with zero-mean and variance σ_q^2. In this case, for $I(U;X) = R$, we get

$$I(U;Y) = \frac{1}{2}\log(1 + \text{snr}) - \frac{1}{2}\log\left(1 + \text{snr}\exp(-2R)\right). \tag{12}$$

3.2. Approximations for Generic Distributions

Next, we present an approach to obtain solutions to the the information bottleneck problem for generic distributions, both when this solution is known and when it is unknown. The method consists in defining a variational (lower) bound on the IB-Lagrangian, which can be optimized more easily than optimizing the IB-Lagrangian directly.

3.2.1. A Variational Bound

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 optimizing the IB-Lagrangian

$$\mathcal{L}_\beta^{\text{IB}}(P_{U|X}) := I(U;Y) - \beta I(U;X) \tag{13}$$

where the maximization is over all stochastic mappings $P_{U|X}$ such that $U \multimap X \multimap Y$ and $|\mathcal{U}| \leq |\mathcal{X}| + 1$. In this section, we show that minimizing Equation (13) is equivalent to optimizing the variational cost

$$\mathcal{L}_\beta^{\text{VIB}}(P_{U|X}, Q_{Y|U}, S_U) := \mathbb{E}_{P_{U|X}}\left[\log Q_{Y|U}(Y|U)\right] - \beta D_{\text{KL}}(P_{U|X}|S_U), \tag{14}$$

where $Q_{Y|U}(y|u)$ is an given stochastic map $Q_{Y|U} : \mathcal{U} \to [0,1]$ (also referred to as the variational approximation of $P_{Y|U}$ or decoder) and $S_U(u) : \mathcal{U} \to [0,1]$ is a given stochastic map (also referred to as the variational approximation of P_U), and $D_{\text{KL}}(P_{U|X}|S_U)$ is the relative entropy between $P_{U|X}$ and S_U.

Then, we have the following bound for a any valid $P_{U|X}$, i.e., satisfying the Markov Chain in Equation (2),

$$\mathcal{L}_\beta^{\text{IB}}(P_{U|X}) \geq \mathcal{L}_\beta^{\text{VIB}}(P_{U|X}, Q_{Y|U}, S_U), \tag{15}$$

where the equality holds when $Q_{Y|U} = P_{Y|U}$ and $S_U = P_U$, i.e., the variational approximations correspond to the true value.

In the following, we derive the variational bound. Fix $P_{U|X}$ (an encoder) and the variational decoder approximation $Q_{Y|U}$. The relevance $I(U;Y)$ can be lower-bounded as

$$I(U;Y) = \oint_{u \in \mathcal{U}, y \in \mathcal{Y}} P_{U,Y}(u,y) \log \frac{P_{Y|U}(y|u)}{P_Y(y)} d_y d_u \tag{16}$$

$$\overset{(a)}{=} \oint_{u \in \mathcal{U}, y \in \mathcal{Y}} P_{U,Y}(u,y) \log \frac{Q_{Y|U}(y|u)}{P_Y(y)} d_y d_u + D\left(P_Y \| Q_Y | U\right) \tag{17}$$

$$\overset{(b)}{\geq} \oint_{u \in \mathcal{U}, y \in \mathcal{Y}} P_{U,Y}(u,y) \log \frac{Q_{Y|U}(y|u)}{P_Y(y)} d_y d_u \tag{18}$$

$$= H(Y) + \oint_{u \in \mathcal{U}, y \in \mathcal{Y}} P_{U,Y}(u,y) \log Q_{Y|U}(y|u) d_y d_u \tag{19}$$

$$\overset{(c)}{\geq} \oint_{u \in \mathcal{U}, y \in \mathcal{Y}} P_{U,Y}(u,y) \log Q_{Y|U}(y|u) d_y d_u \tag{20}$$

$$\stackrel{(d)}{=} \oint_{u \in \mathcal{U}, \, x \in \mathcal{X}, \, y \in \mathcal{Y}} P_X(x) P_{Y|X}(y|x) P_{U|X}(u|x) \log Q_{Y|U}(y|u) d_x d_y d_u, \tag{21}$$

where in (a) the term $D\left(P_Y \| Q_Y | U\right)$ is the conditional relative entropy between P_Y and Q_Y, given P_U; (b) holds by the non-negativity of relative entropy; (c) holds by the non-negativity of entropy; and (d) follows using the Markov Chain $U \, \leftrightarrow \, X \, \leftrightarrow \, Y$.

Similarly, let S_U be a given the variational approximation of P_U. Then, we get

$$I(U; X) = \oint_{u \in \mathcal{U}, \, x \in \mathcal{X}} P_{U,X}(u, x) \log \frac{P_{U|X}(u|x)}{P_U(u)} d_x d_u \tag{22}$$

$$= \oint_{u \in \mathcal{U}, \, x \in \mathcal{X}} P_{U,X}(u, x) \log \frac{P_{U|X}(u|x)}{S_U(u)} d_x d_u - D\left(P_U \| S_U\right) \tag{23}$$

$$\leq \oint_{u \in \mathcal{U}, \, x \in \mathcal{X}} P_{U,X}(u, x) \log \frac{P_{U|X}(u|x)}{S_U(u)} d_x d_u \tag{24}$$

where the inequality follows since the relative entropy is non-negative.

Combining Equations (21) and (24), we get

$$I(U; Y) - \beta I(U; X) \geq \oint_{u \in \mathcal{U}, \, x \in \mathcal{X}, \, y \in \mathcal{Y}} P_X(x) P_{Y|X}(y|x) P_{U|X}(u|x) \log Q_{Y|U}(y|u) d_x d_y d_u$$

$$- \beta \oint_{u \in \mathcal{U}, \, x \in \mathcal{X}} P_{U,X}(u, x) \log \frac{P_{U|X}(u|x)}{S_U(u)} d_x d_u. \tag{25}$$

The use of the variational bound in Equation (14) over the IB-Lagrangian in Equation (13) shows some advantages. First, it allows the derivation of alternating algorithms that allow to obtain a solution by optimizing over the encoders and decoders. Then, it is easier to obtain an empirical estimate of Equation (14) by sampling from: (i) the joint distribution $P_{X,Y}$; (ii) the encoder $P_{U|X}$; and (iii) the prior S_U. Additionally, as noted in Equation (15), when evaluated for the optimal decoder $Q_{Y|U}$ and prior S_U, the variational bound becomes tight. All this allows obtaining algorithms to obtain good approximate solutions to the IB problem, as shown next. Further theoretical implications of this variational bound are discussed in [55].

3.2.2. Known Distributions

Using the variational formulation in Equation (14), when the data model is discrete and the joint distribution $P_{X,Y}$ is known, the IB problem can be solved by using an iterative method that optimizes the variational IB cost function in Equation (14) alternating over the distributions $P_{U|X}, Q_{Y|U}$, and S_U. In this case, the maximizing distributions $P_{U|X}, Q_{Y|U}$, and S_U can be efficiently found by an alternating optimization procedure similar to the expectation-maximization (EM) algorithm [56] and the standard Blahut–Arimoto (BA) method [57]. In particular, a solution $P_{U|X}$ to the constrained optimization problem is determined by the following self-consistent equations, for all $(u, x, y) \in \mathcal{U} \times \mathcal{X} \times \mathcal{Y}$, [1]

$$P_{U|X}(u|x) = \frac{P_U(u)}{Z(\beta, x)} \exp\left(-\beta D_{\mathrm{KL}}\left(P_{Y|X}(\cdot|x) \| P_{Y|U}(\cdot|u)\right)\right) \tag{26a}$$

$$P_U(u) = \sum_{x \in \mathcal{X}} P_X(x) P_{U|X}(u|x) \tag{26b}$$

$$P_{Y|U}(y|u) = \sum_{x \in \mathcal{X}} P_{Y|X}(y|x) P_{X|U}(x|u) \tag{26c}$$

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 [1] 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 [57,58], 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 $I(X;Y)$.

For discrete sources with (small) alphabets, the updating equations described by Equation (26) are relatively easy computationally. However, if the variables X and Y lie in a continuum, solving the equations described by Equation (26) is very challenging. In the case in which X and Y are joint multivariate Gaussian, the problem of finding the optimal representation U is analytically tractable in [53] (see also the related [54,59]), as discussed in Section 3.1.2. Leveraging the optimality of Gaussian mappings $P_{U|X}$ to restrict the optimization of $P_{U|X}$ to Gaussian distributions as in Equation (9), allows reducing the search of update rules to those of the associated parameters, namely covariance matrices. When Y is a deterministic function of X, the IB curve cannot be explored, and other Lagrangians have been proposed to tackle this problem [60].

3.3. Unknown Distributions

The main drawback of the solutions presented thus far for the IB principle is that, in the exception of small-sized discrete (X,Y) for which iterating Equation (26) converges to an (at least local) solution and jointly Gaussian (X,Y) for which an explicit analytic solution was found, solving Equation (3) is generally computationally costly, especially for high dimensionality. Another important barrier in solving Equation (3) directly is that IB necessitates knowledge of the joint distribution $P_{X,Y}$. In this section, we describe a method to provide an approximate solution to the IB problem in the case in which the joint distribution is unknown and only a give training set of N samples $\{(x_i,y_i)\}_{i=1}^{N}$ is available.

A major step ahead, which widened the range of applications of IB inference for various learning problems, appeared in [48], where the authors used neural networks to parameterize the variational inference lower bound in Equation (14) and show that its optimization can be done through the classic and widely used stochastic gradient descendent (SGD). This method, denoted by Variational IB in [48] and detailed below, has allowed handling handle high-dimensional, possibly continuous, data, even in the case in which the distributions are unknown.

3.3.1. Variational IB

The goal of the variational IB when only samples $\{(x_i,y_i)\}_{i=1}^{N}$ are available is to solve the IB problem optimizing an approximation of the cost function. For instance, for a given training set $\{(x_i,y_i)\}_{i=1}^{N}$, the right hand side of Equation (14) can be approximated as

$$\mathcal{L}_{\text{low}} \approx \frac{1}{N}\sum_{i=1}^{N}\left[\oint_{u\in\mathcal{U}}\left(P_{U|X}(u|x_i)\log Q_{Y|U}(y_i|u) - \beta P_{U|X}(u|x_i)\log\frac{P_{U|X}(u|x_i)}{S_U(u)}\right)d_u\right]. \tag{27}$$

However, in general, the direct optimization of this cost is challenging. In the variational IB method, this optimization is done by parameterizing the encoding and decoding distributions $P_{U|X}$, $Q_{Y|U}$, and S_U that are to optimize using a family of distributions whose parameters are determined by DNNs. This allows us to formulate Equation (14) in terms of the DNN parameters, i.e., its weights, and optimize it by using the reparameterization trick [15], Monte Carlo sampling, and stochastic gradient descent (SGD)-type algorithms.

Let $P_\theta(u|x)$ denote the family of encoding probability distributions $P_{U|X}$ over $\mathcal{U}$ for each element on $\mathcal{X}$, parameterized by the output of a DNN f_θ with parameters θ. A common example is the family of multivariate Gaussian distributions [15], which are parameterized by the mean $\boldsymbol{\mu}^\theta$ and covariance matrix $\boldsymbol{\Sigma}^\theta$, i.e., $\gamma := (\boldsymbol{\mu}^\theta, \boldsymbol{\Sigma}^\theta)$. Given an observation X, the values of $(\boldsymbol{\mu}^\theta(x), \boldsymbol{\Sigma}^\theta(x))$ are determined by the output of the DNN f_θ, whose input is X, and the corresponding family member is given by

$P_\theta(u|x) = \mathcal{N}(u; \boldsymbol{\mu}^\theta(x), \boldsymbol{\Sigma}^\theta(x))$. For discrete distributions, a common example are concrete variables [61] (or Gumbel-Softmax [62]). Some details are given below.

Similarly, for decoder $Q_{Y|U}$ over $\mathcal{Y}$ for each element on $\mathcal{U}$, let $Q_\psi(y|u)$ denote the family of distributions parameterized by the output of the DNNs f_{ψ_k}. Finally, for the prior distributions $S_U(u)$ over $\mathcal{U}$ we define the family of distributions $S_\varphi(u)$, which do not depend on a DNN.

By restricting the optimization of the variational IB cost in Equation (14) to the encoder, decoder, and prior within the families of distributions $P_\theta(u|x)$, $Q_\psi(y|u)$, and $S_\varphi(u)$, we get

$$\max_{P_{U|X}} \max_{Q_{Y|U}, S_U} \mathcal{L}_\beta^{\text{VIB}}(P_{U|X}, Q_{Y|U}, S_U) \geq \max_{\theta, \phi, \varphi} \mathcal{L}_\beta^{\text{NN}}(\theta, \phi, \varphi), \tag{28}$$

where θ, ϕ, and φ denote the DNN parameters, e.g., its weights, and the cost in Equation (29) is given by

$$\mathcal{L}_\beta^{\text{NN}}(\theta, \phi, \varphi) := \mathbb{E}_{P_{Y,X}} \mathbb{E}_{\{P_\theta(U|X)\}} \left[\log Q_{\phi(Y|U)}(Y|U) \right] - \beta D_{\text{KL}}(P_\theta(U|X) \| S_\varphi(U)). \tag{29}$$

Next, using the training samples $\{(x_i, y_i)\}_{i=1}^N$, the DNNs are trained to maximize a Monte Carlo approximation of Equation (29) over $\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\varphi}$ using optimization methods such as SGD or ADAM [63] with backpropagation. However, in general, the direct computation of the gradients of Equation (29) is challenging due to the dependency of the averaging with respect to the encoding P_θ, which makes it hard to approximate the cost by sampling. To circumvent this problem, the reparameterization trick [15] is used to sample from $P_\theta(U|X)$. In particular, consider $P_\theta(U|X)$ to belong to a parametric family of distributions that can be sampled by first sampling a random variable Z with distribution $P_Z(z)$, $z \in \mathcal{Z}$ and then transforming the samples using some function $g_\theta : \mathcal{X} \times \mathcal{Z} \to \mathcal{U}$ parameterized by θ, such that $U = g_\theta(x, Z) \sim P_\theta(U|x)$. Various parametric families of distributions fall within this class for both discrete and continuous latent spaces, e.g., the Gumbel-Softmax distributions and the Gaussian distributions. Next, we detail how to sample from both examples:

1. Sampling from Gaussian Latent Spaces: When the latent space is a continuous vector space of dimension D, e.g., $\mathcal{U} = \mathbb{R}^D$, we can consider multivariate Gaussian parametric encoders of mean ($\boldsymbol{\mu}^\theta$, and covariance $\boldsymbol{\Sigma}^\theta$), i.e., $P_\theta(u|x) = \mathcal{N}(u; \boldsymbol{\mu}^\theta, \boldsymbol{\Sigma}^\theta)$. To sample $\mathbf{U} \sim \mathcal{N}(u; \boldsymbol{\mu}^\theta, \boldsymbol{\Sigma}^\theta)$, where $\boldsymbol{\mu}^\theta(x) = f_{e,\theta}^\mu(x)$ and $\boldsymbol{\Sigma}^\theta(x) = f_{e,\theta}^\Sigma(x)$ are determined as the output of a NN, sample a random variable $Z \sim \mathcal{N}(z; \mathbf{0}, \mathbf{I})$ i.i.d. and, given data sample $x \in \mathcal{X}$, and generate the jth sample as

$$u_j = f_{e,\theta}^\mu(x) + f_{e,\theta}^\Sigma(x) z_j \tag{30}$$

where z_j is a sample of $Z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, which is an independent Gaussian noise, and $f_e^\mu(x)$ and $f_e^\Sigma(x)$ are the output values of the NN with weights θ for the given input sample x.

An example of the resulting DIB architecture to optimize with an encoder, a latent space, and a decoder parameterized by Gaussian distributions is shown in Figure 3.

2. Sampling from a discrete latent space with the Gumbel-Softmax:

If U is categorical random variable on the finite set $\mathcal{U}$ of size D with probabilities $\pi := (\pi_1, \ldots, \pi_D))$, we can encode it as D-dimensional one-hot vectors lying on the corners of the $(D-1)$-dimensional simplex, Δ_{D-1}. In general, costs functions involving sampling from categorical distributions are non-differentiable. Instead, we consider Concrete variables [62] (or Gumbel-Softmax [61]), which are continuous differentiable relaxations of categorical variables on the interior of the simplex, and are easy to sample. To sample from a Concrete random variable $U \in \Delta_{D-1}$ at temperature $\lambda \in (0, \infty)$, with probabilities $\pi \in (0, 1)^D$, sample $G_d \sim \text{Gumbel}(0, 1)$ i.i.d. (The Gumbel$(0, 1)$ distribution can be sampled by drawing $u \sim \text{Uniform}(0, 1)$ and calculating $g = -\log(-\log(u))$.), and set for each of the components of $U = (U_1, \ldots, U_D)$

$$U_d = \frac{\exp((\log(\pi_d + G_d)/\lambda))}{\sum_{j=1}^D \exp((\log(\pi_j + G_j)/\lambda))}, \quad d = 1, \ldots, D. \tag{31}$$

We denote by $Q_{\pi,\lambda(u,x)}$ the Concrete distribution with parameters $(\pi(x),\lambda)$. When the temperature λ approaches 0, the samples from the concrete distribution become one-hot and $\Pr\{\lim_{\lambda\to 0} U_d\} = \pi_d$ [61]. Note that, for discrete data models, standard application of Caratheodory's theorem [64] shows that the latent variables U that appear in Equation (3) can be restricted to be with bounded alphabets size.
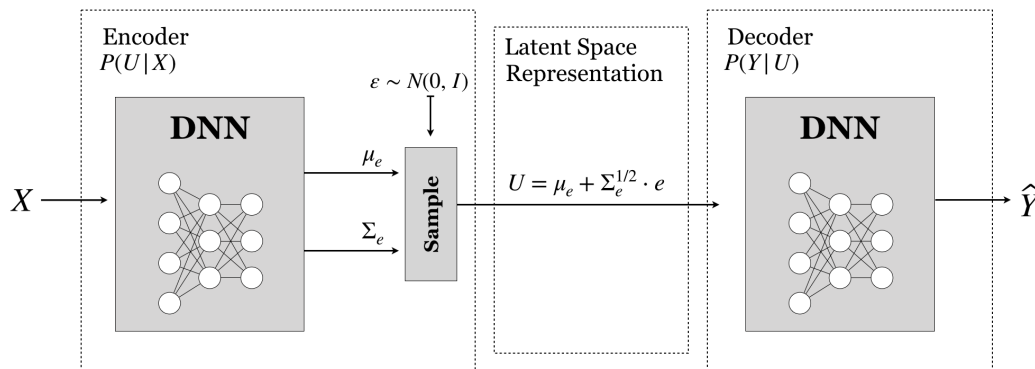

Figure 3. Example parametrization of Variational Information Bottleneck using neural networks.

The reparametrization trick transforms the cost function in Equation (29) into one which can be to approximated by sampling M independent samples $\{u_m\}_{m=1}^M \sim P_\theta(u|x_i)$ for each training sample (x_i, y_i), $i = 1, \ldots, N$ and allows computing estimates of the gradient using backpropagation [15]. Sampling is performed by using $u_{i,m} = g_\phi(x_i, z_m)$ with $\{z_m\}_{m=1}^M$ i.i.d. sampled from P_Z. Altogether, we have the empirical-DIB cost for the ith sample in the training dataset:

$$\mathcal{L}_{\beta,i}^{\text{emp}}(\theta,\phi,\varphi) := \frac{1}{M} \sum_{m=1}^M \left[\log Q_\phi(y_i|u_{i,m}) - \beta D_{\text{KL}}(P_\theta(U_i|x_i)\|Q_\varphi(U_i)) \right]. \tag{32}$$

Note that, for many distributions, e.g., multivariate Gaussian, the divergence $D_{\text{KL}}(P_\theta(U_i|x_i)\|Q_\varphi(U_i))$ can be evaluated in closed form. Alternatively, an empirical approximation can be considered.

Finally, we maximize the empirical-IB cost over the DNN parameters θ, ϕ, φ as,

$$\max_{\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\varphi}} \frac{1}{N} \sum_{i=1}^N \mathcal{L}_{\beta,i}^{\text{emp}}(\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\varphi}). \tag{33}$$

By the law of large numbers, for large N, M, we have $1/N \sum_{i=1}^M \mathcal{L}_{\beta,i}^{\text{emp}}(\theta,\phi,\varphi) \to \mathcal{L}_\beta^{\text{NN}}(\theta,\phi,\varphi)$ almost surely. After convergence of the DNN parameters to $\theta^*, \phi^*, \varphi^*$, for a new observation X, the representation U can be obtained by sampling from the encoders $P_{\theta_k^*}(U_k|X_k)$. In addition, note that a soft estimate of the remote source Y can be inferred by sampling from the decoder $Q_{\phi^*}(Y|U)$. The notion of encoder and decoder in the IB-problem will come clear from its relationship with lossy source coding in Section 4.1.

4. Connections to Coding Problems

The IB problem is a one-shot coding problem, in the sense that the operations are performed letter-wise. In this section, we consider now the relationship between the IB problem and (asymptotic) coding problem in which the coding operations are performed over blocks of size n, with n assumed to be large and the joint distribution of the data $P_{X,Y}$ is in general assumed to be known a priori. The connections between these problems allow extending results from one setup to another, and to consider generalizations of the classical IB problem to other setups, e.g., as shown in Section 6.

4.1. Indirect Source Coding under Logarithmic Loss

Let us consider the (asymptotic) indirect source coding problem shown in Figure 4, in which Y designates a memoryless remote source and X a noisy version of it that is observed at the encoder.
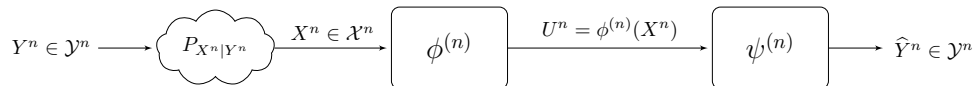

Figure 4. A remote source coding problem.

A sequence of n samples $X^n = (X_1, \ldots, X_n)$ is mapped by an encoder $\phi^{(n)} : \mathcal{X}^n \rightarrow \{1, \ldots, 2^{nR}\}$ which outputs a message from a set $\{1, \ldots, 2^{nR}\}$, that is, the encoder uses at most R bits per sample to describe its observation and the range of the encoder map is allowed to grow with the size of the input sequence as

$$\|\phi^{(n)}\| \leq nR. \tag{34}$$

This message is mapped with a decoder $\phi^{(n)} : \{1, \ldots, 2^{nR}\} \rightarrow \hat{\mathcal{Y}}$ to generate a reconstruction of the source sequence Y^n as $\mathcal{Y}^n \in \hat{\mathcal{Y}}^n$. As already observed in [50], the IB problem in Equation (3) is essentially equivalent to a remote point-to-point source coding problem in which distortion between Y^n as $\mathcal{Y}^n \in \hat{\mathcal{Y}}^n$ is measured under the logarithm loss (log-loss) fidelity criterion [65]. That is, rather than just assigning a deterministic value to each sample of the source, the decoder gives an assessment of the degree of confidence or reliability on each estimate. Specifically, given the output description $m = \phi^{(n)}(x^n)$ of the encoder, the decoder generates a soft-estimate $\hat{y}^n$ of y^n in the form of a probability distribution over $\mathcal{Y}^n$, i.e., $\hat{y}^n = \hat{P}_{Y^n|M}(\cdot)$. The incurred discrepancy between y^n and the estimation $\hat{y}^n$ under log-loss for the observation x^n is then given by the per-letter logarithmic loss distortion, which is defined as

$$\ell_{\log}(y, \hat{y}) := \log \frac{1}{\hat{y}(y)}. \tag{35}$$

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 $y \in \mathcal{Y}$.

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, using a per-letter distortion metric, i.e.,

$$\mathbb{E}[\ell_{\log}^{(n)}(Y^n, \hat{Y}^n)] \leq D \tag{36}$$

where the incurred distortion between two sequences Y^n and $\hat{Y}^n$ is measured as

$$\ell_{\log}^{(n)}(Y^n, \hat{Y}^n) = \frac{1}{n} \sum_{i=1}^{n} \ell_{\log}(y_i, \hat{y}_i) \tag{37}$$

and the per-letter distortion is measured in terms of that given by the logarithmic loss in Equation (53). The rate distortion region of this model is given by the union of all pairs (R, D) that satisfy [7,9]

$$R \geq I(U; X) \tag{38a}$$

$$D \geq H(Y|U) \tag{38b}$$

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 [66, p. 310], it is easy to see that this region is not altered if one restricts U to satisfy $|\mathcal{U}| \leq |\mathcal{X}| + 1$. In addition, 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

$$R \geq I(U; X) \tag{39a}$$

$$\Delta \leq I(U; Y) \tag{39b}$$

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

The boundary of this region is equivalent to the one described by the IB principle in Equation (3) if solved for all β, and therefore the IB problem is essentially a remote source coding problem in which the distortion is measured under the logarithmic loss measure. Note that, operationally, the IB problem is equivalent to that of finding an encoder $P_{U|X}$ which maps the observation X to a representation U that satisfies the bit rate constraint R and such that U captures enough relevance of Y so that the posterior probability of Y given U satisfies an average distortion constraint.

4.2. Common Reconstruction

Consider the problem of source coding with side information at the decoder, i.e., the well known Wyner–Ziv setting [67], 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

$$R + D \geq H(X|Y). \tag{40}$$

The optimal coding scheme utilizes standard Wyner–Ziv compression [67] at the encoder and the decoder map $\psi : \mathcal{U} \times \mathcal{Y} \to \hat{\mathcal{X}}$ is given by

$$\psi(U, Y) = \Pr[X = x|U, Y] \tag{41}$$

for which it is easy to see that

$$\mathbb{E}[\ell_{\log}(X, \psi(U, Y))] = H(X|U, Y). \tag{42}$$

Now, assume that we constrain the coding in a manner that the encoder is 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 [68] 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 in Equation (41) 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

$$R \leq I(U; X|Y) \tag{43a}$$

$$D \geq H(X|U) \tag{43b}$$

for some auxiliary random variable for which $U \multimap X \multimap Y$ holds. Observe that Equation (43b) 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 in Equation (43b) with equality, i.e.,

$$R(D) = \min_{P_{U|X} : I(U;X) = H(X) - D} I(U; X|Y). \tag{44}$$

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

$$I(U; Y) = I(U; X) - I(U; X|Y). \tag{45}$$

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 [69,70].

4.3. Information Combining

Consider again the IB problem. Assume one wishes to 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, we have

$$I(X;U,Y) = I(U;X) + I(Y;X) - I(U;Y) \tag{46}$$
$$= R + I(Y;X) - I(U;Y) \tag{47}$$

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* [71,72], 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.

4.4. 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 [73] and Ahlswede and Körner [74]. For given $R_X = R$, the minimum rate R_Y that is needed to recover Y losslessly is

$$R_Y^\star(R) = \min_{P_{U|X}\,:\,I(U;X)\,\le\,R} H(Y|U). \tag{48}$$

Thus, we get

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

and therefore, solving the IB problem is equivalent to solving the Wyner–Ahlswede–Korner Problem.

4.5. the Privacy Funnel

Consider again the setting of Figure 4, and let us assume that the pair (Y,X) models data that a user possesses and which have the following properties: the data Y are some sensitive (private) data that are not meant to be revealed at all, or else not beyond some level Δ; and the data X are non-private and are 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 trade off between the amount of information that the user shares about X and the information that he keeps private about Y. The data X are 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

$$C(\ell,P) = \inf_{\hat{y}\in\hat{\mathcal{Y}}} \mathbb{E}_P[\ell(Y,\hat{y})] - \inf_{\hat{Y}(\phi(X))} \mathbb{E}_P[\ell(Y,\hat{Y})] \tag{49}$$

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 [75] (see also [76]), it is shown that that under some mild conditions the inference cost gain $C(\ell, P)$ as defined by Equation (49) is upper-bounded as

$$C(\ell, P) \le 2\sqrt{2}L\sqrt{I(U; Y)} \tag{50}$$

where L is a constant. The inequality in Equation (50) holds irrespective to the choice of the loss function ℓ, and this justifies the usage of the logarithmic loss function as given by Equation (53) in the context of finding a suitable trade off between utility and privacy, since

$$I(U; Y) = H(Y) - \inf_{\hat{Y}(U)} \mathbb{E}_P[\ell_{\log}(Y, \hat{Y})]. \tag{51}$$

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 $I(U; Y)$.

4.6. Efficiency of Investment Information

Let Y model a stock market data and X some correlated information. In [77], 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 [73] and Ahlswede-Körner [74], and, thus, also to the IB problem. The work in [77] provides explicit analytic solutions for two horse race examples, jointly binary and jointly Gaussian horse races.

5. Connections to Inference and Representation Learning

In this section, we consider the connections of the IB problem with learning, inference and generalization, for which, typically, the joint distribution $P_{X,Y}$ of the data is not known and only a set of samples is available.

5.1. Inference Model

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

$$\mathcal{C}_{P_{X,Y}}(\psi, \ell) = \mathbb{E}_{P_{X,Y}}[\ell(Y, \psi(X))]. \tag{52}$$

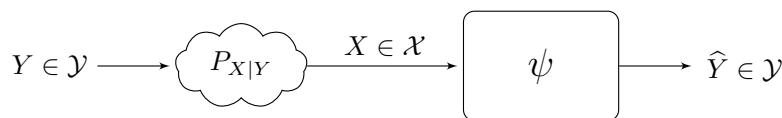An abstract inference model is shown in Figure 5.

Figure 5. 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 [26,27]. This corresponds to choosing the loss function given by the logarithmic loss, which is defined as

$$\ell_{\log}(y, \hat{y}) := \log \frac{1}{\hat{y}(y)} \tag{53}$$

for $y \in \mathcal{Y}$, where $\hat{y} \in \mathcal{P}(\mathcal{Y})$ 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 [30] where Painsky and Wornell showed 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 [29] showed 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 [31], the tree-based algorithm of Quinlan [32], or the well known Chow–Liu algorithm [33] for learning tree graphical models, with various applications in genetics [34], image processing [35], computer vision [36], and others. The logarithmic loss measure also plays a central role in the theory of prediction [37, Ch. 09], where it is often referred to as the *self-information* loss function, as well as in Bayesian modeling [38] where priors are usually designed to maximize the mutual information between the parameter to be estimated and the observations.

When the join distribution $P_{X,Y}$ is known, the optimal predictor and the minimum expected (population) risk can be characterized. Let, for every $x \in \mathcal{X}$, $\psi(x) = Q(\cdot|x) \in \mathcal{P}(\mathcal{Y})$. It is easy to see that

$$\mathbb{E}_{P_{X,Y}}[\ell_{\log}(Y, Q)] = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{X,Y}(x, y) \log\left(\frac{1}{Q(y|x)}\right) \tag{54a}$$

$$= \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{X,Y}(x, y) \log\left(\frac{1}{P_{Y|X}(y|x)}\right) + \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{X,Y}(x, y) \log\left(\frac{P_{Y|X}(y|x)}{Q(y|x)}\right) \tag{54b}$$

$$= H(Y|X) + D(P_{Y|X}\|Q) \tag{54c}$$

$$\geq H(Y|X) \tag{54d}$$

with equality iff the predictor is given by the conditional posterior $\psi(x) = P_Y(Y|X = x)$. That is, the minimum expected (population) risk is given by

$$\min_{\psi} \mathcal{C}_{P_{X,Y}}(\psi, \ell_{\log}) = H(Y|X). \tag{55}$$

If the joint distribution $P_{X,Y}$ is unknown, which is most often the case in practice, the population risk as given by Equation (56) 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, one is interested in optimizing the empirical population risk, which for a set of n i.i.d. samples from $P_{X,Y}$, $\mathcal{D}_n := \{(x_i, y_i)\}_{i=1}^n$, is defined as

$$\hat{\mathcal{C}}_{P_{X,Y}}(\psi, \ell, \mathcal{D}_n) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \psi(x_i)). \tag{56}$$

The difference between the empirical and population risks is normally measured in terms of the generalization gap, defined as

$$\text{gen}_{P_{X,Y}}(\psi, \ell, \mathcal{D}_n) := \mathcal{C}_{P_{X,Y}}(\psi, \ell_{\log}) - \hat{\mathcal{C}}_{P_{X,Y}}(\psi, \ell, \mathcal{D}_n). \tag{57}$$

5.2. Minimum Description Length

One popular approach to reducing the generalization gap is by restricting the set $\mathcal{F}$ of admissible predictors to a low-complexity class (or constrained complexity) to prevent over-fitting. One way to limit the model's complexity is by restricting the range of the prediction function, as shown in Figure 6. This is the so-called minimum description length complexity measure, often used in the learning literature to limit the description length of the weights of neural networks [78]. A connection between the use of the minimum description complexity for limiting the description length of the input encoding and accuracy studied in [79] and with respect to the weight complexity and accuracy is given in [11]. Here, the stochastic mapping $\phi : \mathcal{X} \longrightarrow \mathcal{U}$ is a compressor with

$$\|\phi\| \le R \tag{58}$$

for some prescribed "input-complexity" value R, or equivalently prescribed average description-length.
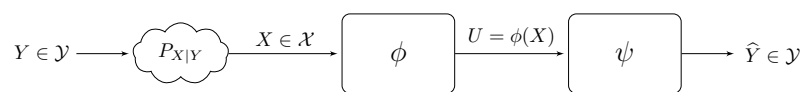

Figure 6. Inference problem with constrained model's complexity.

Minimizing the constrained description length population risk is now equivalent to solving

$$\mathcal{C}_{P_{X,Y},\mathrm{DLC}}(R) = \min_{\phi} \mathbb{E}_{P_{X,Y}}[\ell_{\log}(Y^n, \psi(U^n))] \tag{59}$$

$$\text{s.t. } \|\phi(X^n)\| \le nR. \tag{60}$$

It can be shown that this problem takes its minimum value with the choice of $\psi(U) = P_{Y|U}$ and

$$\mathcal{C}_{P_{X,Y},\mathrm{DLC}}(R) = \min_{P_{U|X}} H(Y|U) \quad \text{s.t. } R \ge I(U;X), \tag{61}$$

The solution to Equation (61) for different values of R is effectively equivalent to the IB-problem in Equation (4). Observe that the right-hand side of Equation (61) 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.

5.3. Generalization and Performance Bounds

The IB-problem appears as a relevant problem in fundamental performance limits of learning. In particular, when $P_{X,Y}$ is unknown, and instead n samples i.i.d from $P_{X,Y}$ are available, the optimization of the empirical risk in Equation (56) leads to a mismatch between the true loss given by the population risk and the empirical risk. This gap is measured by the generalization gap in Equation (57). Interestingly, the relationship between the true loss and the empirical loss can be bounded (in high probability) in terms of the IB-problem as [80]

$$\mathcal{C}_{P_{X,Y}}(\psi, \ell_{\log}) \le \hat{\mathcal{C}}_{P_{X,Y}}(\psi, \ell, \mathcal{D}_n) + \mathrm{gen}_{P_{X,Y}}(\psi, \ell, \mathcal{D}_n)$$

$$= \underbrace{H_{\hat{P}_{X,Y}^{(n)}}(Y|U)}_{\hat{\mathcal{C}}_{P_{X,Y}}(\psi, \ell, \mathcal{D}_n)} + \underbrace{A\sqrt{I(\hat{P}_X^{(n)}; P_{U|X})} \cdot \frac{\log n}{n} + \frac{B\sqrt{\Lambda(P_{U|X}, \hat{P}_{Y|U}, P_{\hat{Y}|U})}}{\sqrt{n}} + \mathcal{O}\left(\frac{\log n}{n}\right)}_{\text{Bound on } \mathrm{gen}_{P_{X,Y}}(\psi, \ell, \mathcal{D}_n)}$$

where $\hat{P}_{U|X}$ and $\hat{P}_{Y|U}$ are the empirical encoder and decoder and $P_{\hat{Y}|U}$ is the optimal decoder. $H_{\hat{P}_{X,Y}^{(n)}}(Y|U)$ and $I(\hat{P}_X^{(n)}; P_{U|X})$ are the empirical loss and the mutual information resulting from the

dataset $\mathcal{D}_n$ and $\Lambda(P_{U|X}, \hat{P}_{Y|U}, P_{\hat{Y}|U})$ is a function that measures the mismatch between the optimal decoder and the empirical one.

This bound shows explicitly the trade-off between the empirical relevance and the empirical complexity. The pairs of relevance and complexity simultaneously achievable is precisely characterized by the IB-problem. Therefore, by designing estimators based on the IB problem, as described in Section 3, one can perform at different regimes of performance, complexity and generalization.

Another interesting connection between learning and the IB-method is the connection of the logarithmic-loss as metric to common performance metrics in learning:

- The logarithmic-loss gives an upper bound on the probability of miss-classification (accuracy):

$$\epsilon_{Y|X}(Q_{\hat{Y}|X}) := 1 - \mathrm{E}_{P_{XY}}[Q_{\hat{Y}|X}] \leq 1 - \exp\left(-\mathrm{E}_{P_{X,Y}}[\ell_{\log}(Y, Q_{\hat{Y}|X})]\right)$$

- The logarithmic-loss is equivalent to maximum likelihood for large n:

$$-\frac{1}{n}\log P_{Y^n|X^n}(y^n|x^n) = -\frac{1}{n}\sum_{i=1}^{n}\log P_{Y|X}(y_i|x_i) \overset{n\to\infty}{\longrightarrow} \mathrm{E}_{X,Y}[-\log P_{Y|X}(Y|X)]$$

- The true distribution P minimizes the expected logarithmic-loss:

$$P_{Y|X} = \arg\min_{Q_{\hat{Y}|X}} \mathrm{E}_P \log \frac{1}{Q_{\hat{Y}|X}} \quad \text{and} \quad \min_{Q_{\hat{Y}|X}} \mathrm{E}[\ell_{\log}(Y, Q_{\hat{Y}|X})] = H(Y|X)$$

Since for $n \to \infty$ the joint distribution P_{XY} can be perfectly learned, the link between these common criteria allows the use of the IB-problem to derive asymptotic performance bounds, as well as design criteria, in most of the learning scenarios of classification, regression, and inference.

5.4. Representation Learning, Elbo and Autoencoders

The performance of machine learning algorithms depends strongly on the choice of data representation (or features) on which they are applied. For that reason, feature engineering, i.e., the set of all pre-processing operations and transformations applied to data in the aim of making them support effective machine learning, is important. However, because it is both data- and task-dependent, such feature-engineering is labor intensive and highlights one of the major weaknesses of current learning algorithms: their inability to extract discriminative information from the data themselves instead of hand-crafted transformations of them. In fact, although it may sometimes appear useful to deploy feature engineering in order to take advantage of human know-how and prior domain knowledge, it is highly desirable to make learning algorithms less dependent on feature engineering to make progress towards true artificial intelligence.

Representation learning is a sub-field of learning theory that aims at learning representations of the data that make it easier to extract useful information, possibly without recourse to any feature engineering. That is, the goal is to identify and disentangle the underlying explanatory factors that are hidden in the observed data. In the case of probabilistic models, a good representation is one that captures the posterior distribution of the underlying explanatory factors for the observed input. For related works, the reader may refer, e.g., to the proceedings of the International Conference on Learning Representations (ICLR), see https://iclr.cc/.

The use of the Shannon's mutual information as a measure of similarity is particularly suitable for the purpose of learning a good representation of data [81]. In particular, a popular approach to representation learning are autoencoders, in which neural networks are designed for the task of representation learning. Specifically, we design a neural network architecture such that we impose a

bottleneck in the network that forces a compressed knowledge representation of the original input, by optimizing the Evidence Lower Bound (ELBO), given as

$$\mathcal{L}^{\text{ELBO}}(\theta, \phi, \varphi) := \frac{1}{N} \sum_{i=1}^{N} \Big[\log Q_\phi(x_i|u_i) - D_{\text{KL}}(P_\theta(U_i|x_i) \| Q_\varphi(U_i)) \Big]. \tag{62}$$

over the neural network parameters θ, ϕ, φ. Note that this is precisely the variational-IB cost in Equation (32) for $\beta = 1$ and $Y = X$, i.e., the IB variational bound when particularized to distributions whose parameters are determined by neural networks. In addition, note that the architecture shown in Figure 3 is the classical neural network architecture for autoencoders, and that is coincides with the variational IB solution resulting from the optimization of the IB-problem in Section 3.3.1. In addition, note that Equation (32) provides an operational meaning to the β-VAE cost [82], as a criterion to design estimators on the relevance–complexity plane for different β values, since the β-VAE cost is given as

$$\mathcal{L}^{\beta-\text{VAE}}(\theta, \phi, \varphi) := \frac{1}{N} \sum_{i=1}^{N} \Big[\log Q_\phi(x_i|u_i) - \beta D_{\text{KL}}(P_\theta(U_i|x_i) \| Q_\varphi(U_i)) \Big], \tag{63}$$

which coincides with the empirical version of the variational bound found in Equation (32).

5.5. Robustness to Adversarial Attacks

Recent advances in deep learning has allowed the design of high accuracy neural networks. However, it has been observed that the high accuracy of trained neural networks may be compromised under nearly imperceptible changes in the inputs [83–85]. The information bottleneck has also found applications in providing methods to improve robustness to adversarial attacks when training models. In particular, the use of the variational IB method of Alemi et al. [48] showed the advantages of the resulting neural network for classification in terms of robustness to adversarial attacks. Recently, alternatives strategies for extracting features in supervised learning are proposed in [86] to construct classifiers robust to small perturbations in the input space. Robustness is measured in terms of the (statistical)-Fisher information, given for two random variables (Y, Z) as

$$\Phi(Z|Y) = \mathbb{E}_{Y,Z} \left| \frac{\partial}{\partial y} \log p(Z|Y) \right|^2. \tag{64}$$

The method in [86] builds upon the idea of the information bottleneck by introducing an additional penalty term that encourages the Fisher information in Equation (64) of the extracted features to be small, when parametrized by the inputs. For this problem, under jointly Gaussian vector sources (X, Y), the optimal representation is also shown to be Gaussian, in line with the results in Section 6.2.1 for the IB without robustness penalty. For general source distributions, a variational method is proposed similar to the variational IB method in Section 3.3.1. The problem shows connections with the I-MMSE [87], de Brujin identity[88,89], Cramér–Rao inequality [90], and Fano's inequality [90].

6. Extensions: Distributed Information Bottleneck

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 7. 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 $k \in \mathcal{K} := \{1, \cdots, K\}$,

$$X_k \multimap Y \multimap X_{\mathcal{K}/k}. \tag{65}$$

The variable Y is a target variable and we seek to characterize how accurately 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.
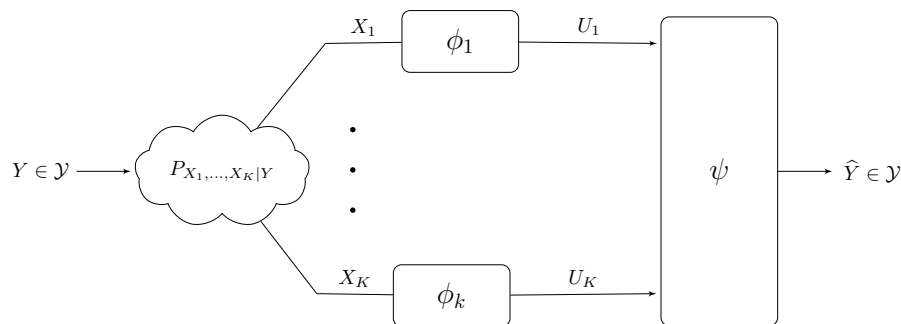

Figure 7. A model for distributed, e.g., multi-view, learning.

6.1. the Relevance–Complexity Region

The distributed IB problem of Figure 7 is studied in [91,92] from information-theoretic grounds. For both discrete memoryless (DM) and memoryless vector Gaussian models, the authors established fundamental limits of learning in terms of optimal trade-offs between relevance and complexity, leveraging on the connection between the IB-problem and source coding. The following theorem states the result for the case of discrete memoryless sources.

Theorem 1. *([91,92]) 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*

$$\Delta \le \sum_{k \in \mathcal{S}} \left[R_k - I(X_k; U_k | Y, T) \right] + I(Y; U_{\mathcal{S}^c} | T), \quad \forall \mathcal{S} \subseteq \mathcal{K} \tag{66}$$

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

Proof. The proof of Theorem 1 can be found in [92, Section 7.1] and is reproduced in Section 8.1 for completeness. □

For a given joint data distribution $P_{X_{\mathcal{K}}, Y}$, Theorem 1 extends the single encoder IB principle of Tishby in Equation (3) to the distributed learning model with K encoders, which we denote by Distributed Information Bottleneck (DIB) problem. The result characterizes the optimal relevance–complexity trade-off as a region of achievable tuples $(\Delta, R_1, \ldots, R_K)$ in terms of a distributed representation learning problem involving the optimization over K conditional pmfs $P_{U_k|X_k,T}$ and a pmf P_T. The pmfs $P_{U_k|X_k,T}$ correspond to stochastic encodings of the observation X_k to a latent variable, or representation, U_k which captures the relevant information of Y in observation X_k. Variable T corresponds to a time-sharing among different encoding mappings (see, e.g., [51]). For such encoders, the optimal decoder is implicitly given by the conditional pmf of Y from $U_1, \ldots, U_K$, i.e., $P_{Y|U_{\mathcal{K}},T}$.

The characterization of the relevance–complexity region can be used to derive a cost function for the D-IB similarly to the IB-Lagrangian in Equation (3). For simplicity, 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

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

We define the DIB-Lagrangian (under sum-rate) as

$$\mathcal{L}_s(\mathbf{P}) := -H(Y|U_{\mathcal{K}}) - s \sum_{k=1}^{K} [H(Y|U_k) + I(X_k; U_k)]. \tag{67}$$

The optimization of Equation (67) over the encoders $P_{U_k|X_k, T}$ allows obtaining mappings that perform on the boundary of the relevance–sum complexity region $\mathcal{RI}_{\mathrm{DIB}}^{\mathrm{sum}}$. To see that, note that it is easy to see that the relevance–sum complexity 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

$$\Delta(R_{\mathrm{sum}}, P_{X_{\mathcal{K}}, Y}) = \max_{\mathbf{P}} \min \left\{ I(Y; U_{\mathcal{K}}), R_{\mathrm{sum}} - \sum_{k=1}^{K} I(X_k; U_k|Y) \right\}, \tag{68}$$

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.

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*

$$\Delta_s = \frac{1}{(1+s)} \left[(1 + sK)H(Y) + sR_s + \max_{\mathbf{P}} \mathcal{L}_s(\mathbf{P}) \right], \tag{69}$$

$$R_s = I(Y; U_{\mathcal{K}}^*) + \sum_{k=1}^{K} [I(X_k; U_k^*) - I(Y; U_k^*)], \tag{70}$$

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 in Equation (67)

Proof. The proof of Proposition 1 can be found in [92, Section 7.3] and is reproduced here in Section 8.2 for completeness. □

The optimization of the distributed IB cost function in Equation (67) generalizes the centralized Tishby's information bottleneck formulation in Equation (3) to the distributed learning setting. Note that for $K = 1$ the optimization in Equation (69) reduces to the single encoder cost in Equation (3) with a multiplier $s/(1 + s)$.

6.2. Solutions to the Distributed Information Bottleneck

The methods described in Section 3 can be extended to the distributed information bottleneck case in order to find the mappings $P_{U_1|X_1, T}, \cdots, P_{U_K|X_K, T}$ in different scenarios.

6.2.1. 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. In addition, 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 in Equation (83). 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

$$\mathbf{X}_k = \mathbf{H}_k \mathbf{Y} + \mathbf{N}_k, \tag{71}$$

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, covariance matrix $\mathbf{\Sigma}_k$, and independent from all other noises and $\mathbf{Y}$.

For the vector Gaussian model Equation (71), the result of Theorem 1, which can be extended to continuous sources using standard techniques, characterizes the relevance–complexity region of this model. The following theorem characterizes the relevance–complexity region, 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

$$\mathbf{U}_k = \mathbf{A}_k \mathbf{X}_k + \mathbf{Z}_k \sim \mathcal{CN}(\mathbf{u}_k; \mathbf{A}_k \mathbf{X}_k, \mathbf{\Sigma}_{z,k}), \tag{72}$$

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

Theorem 2. *For the vector Gaussian data model, the relevance–complexity region $\mathcal{RI}_{\mathrm{DIB}}^{\mathrm{G}}$ is given by the union of all tuples $(\Delta, R_1, \ldots, R_L)$ that satisfy*

$$\Delta \le \sum_{k \in \mathcal{S}} \left(R_k + \log \left| \mathbf{I} - \mathbf{\Sigma}_k^{1/2} \mathbf{\Omega}_k \mathbf{\Sigma}_k^{1/2} \right| \right) + \log \left| \mathbf{I} + \sum_{k \in \mathcal{S}^c} \mathbf{\Sigma}_{\mathbf{y}}^{1/2} \mathbf{H}_k^{\dagger} \mathbf{\Omega}_k \mathbf{H}_k \mathbf{\Sigma}_{\mathbf{y}}^{1/2} \right|, \quad \forall \mathcal{S} \subseteq \mathcal{K},$$

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

Proof. The proof of Theorem 2 can be found in [92, Section 7.5] and is reproduced here in Section 8.4 for completeness. □

Theorem 2 extends the result of [54,93] on the relevance–complexity trade-off characterization of the single-encoder IB problem for jointly Gaussian sources to K encoders. The theorem also shows that the optimal test channels $P_{U_k|X_k}$ are multivariate Gaussian, as given by Equation (72).

Consider the following symmetric distributed scalar Gaussian setting, in which $Y \sim \mathcal{N}(0,1)$ and

$$X_1 = \sqrt{\mathrm{snr}}\, Y + N_1 \tag{73a}$$
$$X_2 = \sqrt{\mathrm{snr}}\, Y + N_2 \tag{73b}$$

where N_1 and N_2 are standard Gaussian with zero-mean and unit variance, both independent of Y. In this case, for $I(U_1; X_1) = R$ and $I(U; X_2) = R$, the optimal relevance is

$$\Delta^\star(R, \mathrm{snr}) = \frac{1}{2} \log \left(1 + 2\mathrm{snr}\exp(-4R) \left(\exp(4R) + \mathrm{snr} - \sqrt{\mathrm{snr}^2 + (1 + 2\mathrm{snr})\exp(4R)} \right) \right). \tag{74}$$

An easy upper bound on the relevance can be obtained by assuming that X_1 and X_2 are encoded jointly at rate $2R$, to get

$$\Delta_{\mathrm{ub}}(R, \mathrm{snr}) = \frac{1}{2} \log(1 + 2\mathrm{snr}) - \frac{1}{2} \log \left(1 + 2\mathrm{snr}\exp(-4R) \right). \tag{75}$$

The reader may notice that, if X_1 and X_2 are encoded independently, an achievable relevance level is given by

$$\Delta_{\mathrm{lb}}(R, \mathrm{snr}) = \frac{1}{2} \log(1 + 2\mathrm{snr} - \mathrm{snr}\exp(-2R)) - \frac{1}{2} \log \left(1 + \mathrm{snr}\exp(-2R) \right). \tag{76}$$

6.3. Solutions for Generic Distributions

Next, we present how the distributed information bottleneck can be solved for generic distributions. Similar to the case of single encoder IB-problem, the solutions are based on a variational bound on the DIB-Lagrangian. For simplicity, we look at the D-IB under sum-rate constraint [92].

6.4. a Variational Bound

The optimization of Equation (67) generally requires computing 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

$$\mathbf{Q} := \{Q_{Y|U_1,\ldots,U_K}, Q_{Y|U_1}, \ldots, Q_{Y|U_K}, Q_{U_1}, \ldots, Q_{U_K}\}. \tag{77}$$

The variational D-IB cost for the DIB-problem is given by

$$\mathcal{L}_s^{\mathrm{VB}}(\mathbf{P},\mathbf{Q}) := \underbrace{\mathbb{E}[\log Q_{Y|U_K}(Y|U_K)]}_{\text{av. logarithmic-loss}} + s \underbrace{\sum_{k=1}^K \Big(\mathbb{E}[\log Q_{Y|U_k}(Y|U_k)] - D_{\mathrm{KL}}(P_{U_k|X_k}\|Q_{U_k}) \Big)}_{\text{regularizer}}. \tag{78}$$

Lemma 1. *For fixed* $\mathbf{P}$, *we have*

$$\mathcal{L}_s(\mathbf{P}) \geq \mathcal{L}_s^{\mathrm{VB}}(\mathbf{P},\mathbf{Q}), \qquad \text{for all pmfs } \mathbf{Q}. \tag{79}$$

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,* $\forall k \in \mathcal{K}$,

$$Q_{U_k}^* = P_{U_k} \tag{80a}$$

$$Q_{Y|U_k}^* = P_{Y|U_k} \tag{80b}$$

$$Q_{Y|U_1,\ldots,U_k}^* = P_{Y|U_1,\ldots,U_K}, \tag{80c}$$

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* $\mathbf{P}$.

Proof. The proof of Lemma 1 can be found in [92, Section 7.4] and is reproduced here in Section 8.3 for completeness. □

Then, the optimization in Equation (69) can be written in terms of the variational DIB cost function as follows,

$$\max_{\mathbf{P}} \mathcal{L}_s(\mathbf{P}) = \max_{\mathbf{P}} \max_{\mathbf{Q}} \mathcal{L}_s^{\mathrm{VB}}(\mathbf{P},\mathbf{Q}). \tag{81}$$

The variational DIB cost in Equation (78) is a generalization to distributed learning with *K*-encoders of the evidence lower bound (ELBO) of the target variable *Y* given the representations $U_1, \cdots, U_K$ [15]. If $Y = (X_1,\ldots,X_K)$, the bound generalizes the ELBO used for VAEs to the setting of $K \geq 2$ encoders. In addition, note that Equation (78) also generalizes and provides an operational meaning to the β-VAE cost [82] with $\beta = s/(1+s)$, as a criteria to design estimators on the relevance–complexity plane for different β values.

6.5. Known Memoryless Distributions

When the data model is discrete and the joint distribution $P_{X,Y}$ is known, the DIB problem can be solved by using an iterative method that optimizes the variational IB cost function in Equation (81) alternating over the distributions $\mathbf{P}, \mathbf{Q}$. The optimal encoders and decoders of the D-IB under sum-rate constraint satisfy the following self consistent equations,

$$p(u_k|y_k) = \frac{p(u_k)}{Z(\beta, u_k)} \exp\left(-\psi_s(u_k, y_k)\right),$$

$$p(x|u_k) = \sum_{y_k \in \mathcal{Y}_k} p(y_k|u_k) p(x|y_k)$$

$$p(x|u_1, \ldots, u_K) = \sum_{y_\mathcal{K} \in \mathcal{Y}_\mathcal{K}} p(y_\mathcal{K}) p(u_\mathcal{K}|y_\mathcal{K}) p(x|y_\mathcal{K})/p(u_\mathcal{K})$$

where $\psi_s(u_k, y_k) := D_{\mathrm{KL}}(P_{X|y_k} \| Q_{X|u_k}) + \frac{1}{s} \mathbb{E}_{U_{\mathcal{K} \setminus k}|y_k}[D_{\mathrm{KL}}(P_{X|U_{\mathcal{K} \setminus k}, y_k} \| Q_{X|U_{\mathcal{K} \setminus k}, u_k}))]$.

Alternating iterations of these equations converge to a solution for any initial $p(u_k|x_k)$, similarly to a Blahut–Arimoto algorithm and the EM.
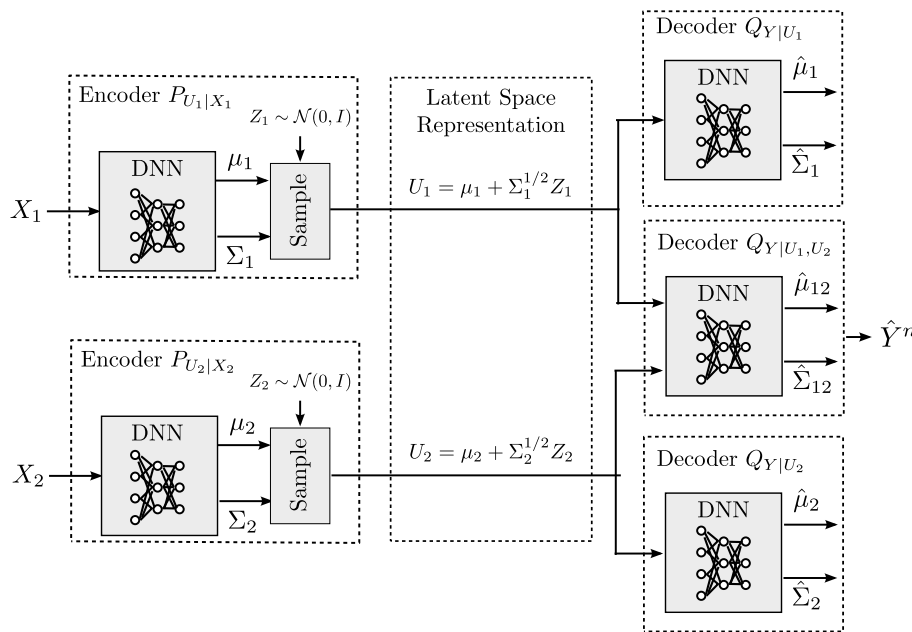
6.5.1. Distributed Variational IB

When the data distribution is unknown and only data samples are available, the variational DIB cost in Equation (81) can be optimized following similar steps as for the variational IB in Section 3.3.1 by parameterizing the encoding and decoding distributions $\mathbf{P}, \mathbf{Q}$ using a family of distributions whose parameters are determined by DNNs. This allows us to formulate Equation (81) in terms of the DNN parameters, i.e., its weights, and optimize it by using the reparameterization trick [15], Monte Carlo sampling, and stochastic gradient descent (SGD)-type algorithms.

Considering encoders and decoders $\mathbf{P}, \mathbf{Q}$ parameterized by DNN parameters $\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\varphi}$, the DIB cost in Equation (81) can be optimized by considering the following empirical Monte Carlo approximation:

$$\max_{\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\varphi}} \frac{1}{n} \sum_{i=1}^n \left[\log Q_{\phi_\mathcal{K}}(y_i|u_{1,i,j}, \ldots, u_{K,i,j}) + s \sum_{k=1}^K \left(\log Q_{\phi_k}(y_i|u_{k,i,j}) - D_{\mathrm{KL}}(P_{\theta_k}(U_{k,i}|x_{k,i}) \| Q_{\varphi_k}(U_{k,i})) \right) \right], \quad (82)$$

where $u_{k,i,j} = g_{\phi_k}(x_{k,i}, z_{k,j})$ are samples obtained from the reparametrization trick by sampling from K random variables P_{Z_k}. The details of the method can be found in [92]. The resulting architecture is shown in Figure 8. This architecture generalizes that from autoencoders to the distributed setup with K encoders.

Figure 8. Example parameterization of the Distributed Variational Information Bottleneck method using neural networks.

6.6. Connections to Coding Problems and Learning

Similar to the point-to-point IB-problem, the distributed IB problem also has abundant connections with (asymptotic) coding and learning problems.

6.6.1. Distributed Source Coding under Logarithmic Loss

Key element to the proof of the converse part of Theorem 3 is the connection with the Chief Executive Officer (CEO) source coding problem. For the case of $K \geq 2$ encoders, while the characterization of the optimal rate-distortion region of this problem for general distortion measures has eluded the information theory for now more than four decades, a characterization of the optimal region in the case of logarithmic loss distortion measure has been provided recently in [65]. A key step in [65] is that the log-loss distortion measure admits a lower bound in the form of the entropy of the source conditioned on the decoders' input. Leveraging this result, in our converse proof of Theorem 3, we derive a single letter upper bound on the entropy of the channel inputs conditioned on the indices $J_\mathcal{K}$ that are sent by the relays, in the absence of knowledge of the codebooks indices $F_\mathcal{L}$. In addition, the rate region of the vector Gaussian CEO problem under logarithmic loss distortion measure has been found recently in [94,95].

6.6.2. Cloud RAN

Consider the discrete memoryless (DM) CRAN model shown in Figure 9. In this model, L users communicate with a common destination or central processor (CP) through K relay nodes, where $L \geq 1$ and $K \geq 1$. Relay node k, $1 \leq k \leq K$, is connected to the CP via an error-free finite-rate fronthaul link of capacity C_k. In what follows, we let $\mathcal{L} := [1:L]$ and $\mathcal{K} := [1:K]$ indicate the set of users and relays, respectively. Similar to Simeone *et al.* [96], the relay nodes are constrained to operate without knowledge of the users' codebooks and only know a time-sharing sequence Q^n, i.e., a set of time instants at which users switch among different codebooks. The obliviousness of the relay nodes to the actual codebooks of the users is modeled via the notion of *randomized encoding* [97,98]. That is, users or transmitters select their codebooks at random and the relay nodes are *not* informed about the currently selected codebooks, while the CP is given such information.
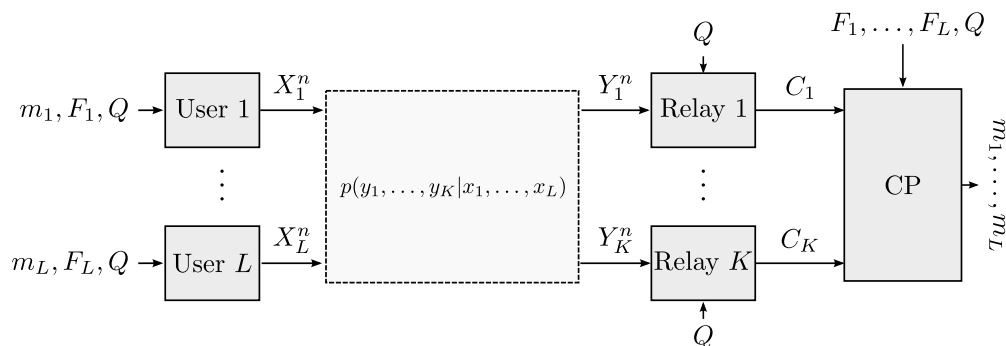

Figure 9. CRAN model with oblivious relaying and time-sharing.

Consider the following class of DM CRANs in which the channel outputs at the relay nodes are independent conditionally on the users' inputs. That is, for all $k \in \mathcal{K}$ and all $i \in [1:n]$,

$$Y_{k,i} \multimap X_{\mathcal{L},i} \multimap Y_{\mathcal{K}/k,i} \tag{83}$$

forms a Markov Chain in this order.

The following theorem provides a characterization of the capacity region of this class of DM CRAN problem under oblivious relaying.

Theorem 3 ([22,23]). *For the class of DM CRANs with oblivious relay processing and enabled time-sharing for which Equation (83) holds, the capacity region $\mathcal{C}(C_\mathcal{K})$ is given by the union of all rate tuples $(R_1, \ldots, R_L)$ which satisfy*

$$\sum_{t \in \mathcal{T}} R_t \leq \sum_{s \in \mathcal{S}} [C_s - I(Y_s; U_s | X_\mathcal{L}, Q)] + I(X_\mathcal{T}; U_{\mathcal{S}^c} | X_{\mathcal{T}^c}, Q),$$

for all non-empty subsets $\mathcal{T} \subseteq \mathcal{L}$ *and all* $\mathcal{S} \subseteq \mathcal{K}$, *for some joint measure of the form*

$$p(q) \prod_{l=1}^{L} p(x_l|q) \prod_{k=1}^{K} p(y_k|x_{\mathcal{L}}) \prod_{k=1}^{K} p(u_k|y_k, q). \tag{84}$$

The direct part of Theorem 3 can be obtained by a coding scheme in which each relay node compresses its channel output by using Wyner–Ziv binning to exploit the correlation with the channel outputs at the other relays, and forwards the bin index to the CP over its rate-limited link. The CP jointly decodes the compression indices (within the corresponding bins) and the transmitted messages, i.e., Cover-El Gamal compress-and-forward [99, Theorem 3] with joint decompression and decoding (CF-JD). Alternatively, the rate region of Theorem 3 can also be obtained by a direct application of the noisy network coding (NNC) scheme of [64, Theorem 1].

The connection between this problem, source coding and the distributed information bottleneck is discussed in [22,23], particularly in the derivation of the converse part of the theorem. Note also the similarity between the resulting capacity region in Theorem 3 and the relevance complexity region of the distributed information bottleneck in Theorem 1, despite the significant differences of the setups.

6.6.3. Distributed Inference, ELBO and Multi-View Learning

In many data analytics problems, data are collected from various sources of information or feature extractors and are intrinsically *heterogeneous*. For example, an image can be identified by its color or texture features and a document may contain text and images. Conventional machine learning approaches concatenate all available data into one big row vector (or matrix) on which a suitable algorithm is then applied. Treating different observations as a single source might cause overfitting and is not physically meaningful because each group of data may have different statistical properties. Alternatively, one may partition the data into groups according to samples homogeneity, and each group of data be regarded as a separate *view*. This paradigm, termed *multi-view learning* [100], has received growing interest, and various algorithms exist, sometimes under references such as *co-training* [101–104], *multiple kernel learning* [104], and *subspace learning* [105]. By using distinct encoder mappings to represent distinct groups of data, and jointly optimizing over all mappings to remove redundancy, multi-view learning offers a degree of flexibility that is not only desirable in practice but is also likely to result in better learning capability. Actually, as shown in [106], local learning algorithms produce fewer errors than global ones. Viewing the problem as that of function approximation, the intuition is that it is usually not easy to find a unique function that holds good predictability properties in the entire data space.

Besides, the distributed learning of Figure 7 clearly finds application in all those scenarios in which learning is performed collaboratively but distinct learners either only access subsets of the entire dataset (e.g., due to physical constraints) or access independent noisy versions of the entire dataset.

In addition, similar to the single encoder case, the distributed IB also finds applications in fundamental performance limits and formulation of cost functions from an operational point of view. One of such examples is the generalization of the commonly used ELBO and given in Equation (62) to the setup with K views or observations, as formulated in Equation (78). Similarly, from the formulation of the DIB problem, a natural generalization of the classical autoencoders emerge, as given in Figure 8.

7. Outlook

A variant of the bottleneck problem in which the encoder's output is constrained in terms of its entropy, rather than its mutual information with the encoder's input as done originally in [1], was considered in [107]. The solution of this problem turns out to be a deterministic encoder map as opposed to the stochastic encoder map that is optimal under the IB framework of Tishby *et al.* [1], which results in a reduction of the algorithm's complexity. This idea was then used and extended to the case of available resource (or time) sharing in [108].

In the context of privacy against inference attacks [109], the authors of [75,76] considered a dual of the information bottleneck problem in which $X \in \mathcal{X}$ represents some private data that are correlated with the non-private data $Y \in \mathcal{Y}$. A legitimate receiver (analyst) wishes to infer as much information as possible about the non-private data Y but does not need to infer any information about the private data X. 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 trade-off between the amount of information that the user shares about X as measured by the mutual information $I(U; X)$ and the information that he keeps private about Y as measured by the mutual information $I(U; Y)$, where $U = \phi(X)$.

Among interesting problems that are left unaddressed in this paper is that of characterizing optimal input distributions under rate-constrained compression at the relays where, e.g., discrete signaling is already known to sometimes outperform Gaussian signaling for single-user Gaussian CRAN [97]. It is conjectured that the optimal input distribution is discrete. Other issues might relate to extensions to continuous time filtered Gaussian channels, in parallel to the regular bottleneck problem [108], or extensions to settings in which fronthauls may be not available at some radio-units, and that is unknown to the systems. That is, the more radio units are connected to the central unit, the higher is the rate that could be conveyed over the CRAN uplink [110]. Alternatively, one may consider finding the worst-case noise under given input distributions, e.g., Gaussian, and rate-constrained compression at the relays. Furthermore, there are interesting aspects that address processing constraints of continuous waveforms, e.g., sampling at a given rate [111,112] with focus on remote logarithmic distortion [65], which in turn boils down to the distributed bottleneck problem [91,92]. We also mention finite-sample size analysis (i.e., finite block length n, which relates to the literature on finite block length coding in information theory). Finally, it is interesting to observe that the bottleneck problem relates to interesting problem when R is not necessarily scaled with the block length n.

8. Proofs

8.1. Proof of Theorem 1

The proof relies on the equivalence of the studied distributed learning problem with the Chief-Executive Officer (CEO) problem under logarithmic-loss distortion measure, which was studied in [65, Theorem 10]. For the K-encoder CEO problem, let us consider K encoding functions $\phi_k : \mathcal{X}_k \to \mathcal{M}_k^{(n)}$ satisfying $nR_k \geq \log |\phi_k(X_k^n)|$ and a decoding function $\tilde{\psi} : \mathcal{M}_1^{(n)} \times \cdots \times \mathcal{M}_K^{(n)} \to \hat{\mathcal{Y}}^n$, which produces a probabilistic estimate of Y from the outputs of the encoders, i.e., $\hat{\mathcal{Y}}^n$ is the set of distributions on $\mathcal{Y}$. The quality of the estimation is measured in terms of the average log-loss.

Definition 1. *A tuple $(D, R_1, \ldots, R_K)$ is said to be achievable in the K-encoder CEO problem for $P_{X_\mathcal{K}, Y}$ for which the Markov Chain in Equation (83) holds, if there exists a length n, encoders ϕ_k for $k \in \mathcal{K}$, and a decoder $\tilde{\psi}$, such that*

$$D \geq \mathrm{E}\left[\frac{1}{n} \log \frac{1}{\hat{P}_{Y^n|J_\mathcal{K}}(Y^n | \phi_1(X_1^n), \cdots, \phi_K(X_K^n))}\right], \tag{85}$$

$$R_k \geq \frac{1}{n} \log |\phi_k(X_k^n)| \quad \text{for all } k \in \mathcal{K}. \tag{86}$$

The rate-distortion region $\mathcal{RD}_{\mathrm{CEO}}$ is given by the closure of all achievable tuples $(D, R_1, \ldots, R_K)$. $\square$

The following lemma shows that the minimum average logarithmic loss is the conditional entropy of Y given the descriptions. The result is essentially equivalent to [65, Lemma 1] and it is provided for completeness.

Lemma 2. *Let us consider $P_{X_{\mathcal{K}},Y}$ and the encoders $J_k = \phi_k(X_k^n)$, $k \in \mathcal{K}$ and the decoder $\hat{Y}^n = \tilde{\psi}(J_{\mathcal{K}})$. Then,*

$$\mathrm{E}[\ell_{\log}(Y^n, \hat{Y}^n)] \geq H(Y^n|J_{\mathcal{K}}), \tag{87}$$

with equality if and only if $\tilde{\psi}(J_{\mathcal{K}}) = \{P_{Y^n|J_{\mathcal{K}}}(y^n|J_{\mathcal{K}})\}_{y^n \in \mathcal{Y}^n}$.

Proof. Let $Z := (J_1, \ldots, J_K)$ be the argument of $\tilde{\psi}$ and $\hat{P}(y^n|z)$ be a distribution on $\mathcal{Y}^n$. We have for $Z = z$:

$$\mathrm{E}[\ell_{\log}(Y^n, \hat{Y}^n)|Z = z] = \sum_{y^n \in \mathcal{Y}^n} P(y^n|z) \log\left(\frac{1}{\hat{P}(y^n|z)}\right) \tag{88}$$

$$= \sum_{y^n \in \mathcal{Y}^n} P(y^n|z) \log\left(\frac{P(y^n|z)}{\hat{P}(y^n|z)}\right) + H(Y^n|Z = z) \tag{89}$$

$$= D_{\mathrm{KL}}(P(y^n|z)\|\hat{P}(y^n|z)) + H(Y^n|Z = z) \tag{90}$$

$$\geq H(Y^n|Z = z), \tag{91}$$

where Equation (91) is due to the non-negativity of the KL divergence and the equality holds if and only if for $\hat{P}(y^n|z) = P(y^n|z)$ where $P(y^n|z) = \mathrm{Pr}\{Y^n = y^n|Z = z\}$ for all z and $y^n \in \mathcal{Y}^n$. Averaging over Z completes the proof.　□

Essentially, Lemma 2 states that minimizing the average log-loss is equivalent to maximizing relevance as given by the mutual information $I\left(Y^n; \psi\left(\phi_1(X_1^n), \cdots, \phi_K(X_K^n)\right)\right)$. Formally, the connection between the distributed learning problem under study and the K-encoder CEO problem studied in [65] can be formulated as stated next.

Proposition 2. *A tuple $(\Delta, R_1, \ldots, R_K) \in \mathcal{RI}_{\mathrm{DIB}}$ if and only if $(H(Y) - \Delta, R_1, \ldots, R_K) \in \mathcal{RD}_{\mathrm{CEO}}$.*

Proof. Let the tuple $(\Delta, R_1, \ldots, R_K) \in \mathcal{RI}_{\mathrm{DIB}}$ be achievable for some encoders ϕ_k. It follows by Lemma 2 that, by letting the decoding function $\tilde{\psi}(J_{\mathcal{K}}) = \{P_{Y^n|J_{\mathcal{K}}}(y^n|J_{\mathcal{K}})\}$, we have $\mathrm{E}[\ell_{\log}(Y^n, \hat{Y}^n)|J_{\mathcal{K}}] = H(Y^n|J_{\mathcal{K}})$, and hence $(H(Y) - \Delta, R_1, \ldots, R_K) \in \mathcal{RD}_{\mathrm{CEO}}$.
Conversely, assume the tuple $(D, R_1, \ldots, R_K) \in \mathcal{RD}_{\mathrm{CEO}}$ is achievable. It follows by Lemma 2 that $H(Y) - D \leq H(Y^n) - H(Y^n|J_{\mathcal{K}}) = I(Y^n; J_{\mathcal{K}})$, which implies $(\Delta, R_1, \ldots, R_K) \in \mathcal{RI}_{\mathrm{DIB}}$ with $\Delta = H(Y) - D$.　□

The characterization of rate-distortion region $\mathcal{R}_{\mathrm{CEO}}$ has been established recently in [65, Theorem 10]. The proof of the theorem is completed by noting that Proposition 2 implies that the result in [65, Theorem 10] can be applied to characterize the region $\mathcal{RI}_{\mathrm{DIB}}$, as given in Theorem 1.

8.2. Proof of Proposition 1

Let $\mathbf{P}^*$ be the maximizing in Equation (69). Then,

$$(1+s)\Delta_s = (1+sK)H(Y) + sR_s + \mathcal{L}_s(\mathbf{P}^*) \tag{92}$$

$$= (1+sK)H(Y) + sR_s + \left(-H(Y|U_{\mathcal{K}}^*) - s\sum_{k=1}^{K}[H(Y|U_k^*) + I(X_k; U_k^*)]\right) \tag{93}$$

$$= (1+sK)H(Y) + sR_s + (-H(Y|U_{\mathcal{K}}^*) - s(R_s - I(Y; U_{\mathcal{K}}^*) + KH(Y))) \tag{94}$$

$$= (1+s)I(Y; U_{\mathcal{K}}^*) \tag{95}$$

$$\leq (1+s)\Delta(R_s, P_{X_{\mathcal{K}},Y}), \tag{96}$$

where Equation (94) is due to the definition of $\mathcal{L}_s(\mathbf{P})$ in Equation (67); Equation (95) holds since $\sum_{k=1}^{K}[I(X_k; U_k^*) + H(Y|U_k^*)] = R_s - I(Y; U_\mathcal{K}^*) + KH(Y)$ using Equation (70); and Equation (96) follows by using Equation (68).

Conversely, if $\mathbf{P}^*$ is the solution to the maximization in the function $\Delta(R_{\mathrm{sum}}, P_{X_\mathcal{K}, Y})$ in Equation (68) such that $\Delta(R_{\mathrm{sum}}, P_{X_\mathcal{K}, Y}) = \Delta_s$, then $\Delta_s \leq I(Y; U_\mathcal{K}^*)$ and $\Delta_s \leq R - \sum_{k=1}^{K} I(X_k; U_k^*|Y)$ and we have, for any $s \geq 0$, that

$$\Delta(R_{\mathrm{sum}}, P_{X_\mathcal{K}, Y}) = \Delta_s \tag{97}$$

$$\leq \Delta_s - (\Delta_s - I(Y; U_\mathcal{K}^*)) - s\left(\Delta_s - R_{\mathrm{sum}} + \sum_{k=1}^{K} I(X_k; U_k^*|Y)\right) \tag{98}$$

$$= I(Y; U_\mathcal{K}^*) - s\Delta_s + sR_{\mathrm{sum}} - s\sum_{k=1}^{K} I(X_k; U_k^*|Y) \tag{99}$$

$$= H(Y) - s\Delta_s + sR_{\mathrm{sum}} - H(Y|U_\mathcal{K}^*) - s\sum_{k=1}^{K}[I(X_k; U_k^*) + H(Y|U_k^*)] + sKH(Y) \tag{100}$$

$$\leq H(Y) - s\Delta_s + sR_{\mathrm{sum}} + \mathcal{L}_s^* + sKH(Y) \tag{101}$$

$$= H(Y) - s\Delta_s + sR_{\mathrm{sum}} + sKH(Y) - ((1 + sK)H(Y) + sR_s - (1+s)\Delta_s) \tag{102}$$

$$= \Delta_s + s(R_{\mathrm{sum}} - R_s), \tag{103}$$

where in Equation (100) we use that $\sum_{k=1}^{K} I(X_k; U_k|Y) = -KH(Y) + \sum_{k=1}^{K} I(X_k; U_k) + H(Y|U_k)$. which follows by using the Markov Chain $U_k \multimap X_k \multimap Y \multimap (X_{\mathcal{K}\setminus k}, U_{\mathcal{K}\setminus k})$; Equation (101) follows since $\mathcal{L}_s^*$ is the maximum over all possible distributions $\mathbf{P}$ (possibly distinct from the $\mathbf{P}^*$ that maximizes $\Delta(R_{\mathrm{sum}}, P_{X_\mathcal{K}, Y})$); and Equation (102) is due to Equation (69). Finally, Equation (103) is valid for any $R_{\mathrm{sum}} \geq 0$ and $s \geq 0$. Given s, and hence (Δ_s, R_s), letting $R = R_s$ yields $\Delta(R_s, P_{X_\mathcal{K}, Y}) \leq \Delta_s$. Together with Equation (96), this completes the proof of Proposition 1.

8.3. Proof of Lemma 1

Let, for a given random variable Z and $z \in \mathcal{Z}$, a stochastic mapping $Q_{Y|Z}(\cdot|z)$ be given. It is easy to see that

$$H(Y|Z) = \mathbb{E}[-\log Q_{Y|Z}(Y|Z)] - D_{\mathrm{KL}}(P_{Y|Z} \| Q_{Y|Z}). \tag{104}$$

In addition, we have

$$I(X_k; U_k) = H(U_k) - H(U_k|X_k) \tag{105}$$

$$= D_{\mathrm{KL}}(P_{U_k|X_k} \| Q_{U_k}) - D_{\mathrm{KL}}(P_{U_k} \| Q_{U_k}). \tag{106}$$

Substituting it into Equation (67), we get

$$\mathcal{L}_s(\mathbf{P}) = \mathcal{L}_s^{\mathrm{VB}}(\mathbf{P}, \mathbf{Q}) + D_{\mathrm{KL}}(P_{Y|U_\mathcal{K}} \| Q_{Y|U_\mathcal{K}}) + s\sum_{k=1}^{K}(D_{\mathrm{KL}}(P_{Y|U_k} \| Q_{Y|U_k}) + D_{\mathrm{KL}}(P_{U_k} \| Q_{U_k})) \tag{107}$$

$$\geq \mathcal{L}_s^{\mathrm{VB}}(\mathbf{P}, \mathbf{Q}), \tag{108}$$

where Equation (108) follows by the non-negativity of relative entropy. In addition, note that the inequality in Equation (108) holds with equality iff $\mathbf{Q}^*$ is given by Equation (80).

8.4. Proof of Theorem 2

The proof of Theorem 2 relies on deriving an outer bound on the relevance–complexity region, as given by Equation (66), and showing that it is achievable with Gaussian pmfs and without time-sharing.

In doing so, we use the technique of [89, Theorem 8], which relies on the de Bruijn identity and the properties of Fisher information and MMSE.

Lemma 3. *[88,89] Let* $(\mathbf{X}, \mathbf{Y})$ *be a pair of random vectors with pmf* $p(\mathbf{x}, \mathbf{y})$. *We have*

$$\log|(\pi e)\mathbf{J}^{-1}(\mathbf{X}|\mathbf{Y})| \le h(\mathbf{X}|\mathbf{Y}) \le \log|(\pi e)\mathrm{mmse}(\mathbf{X}|\mathbf{Y})|, \tag{109}$$

where the conditional Fischer information matrix is defined as

$$\mathbf{J}(\mathbf{X}|\mathbf{Y}) := \mathrm{E}[\nabla \log p(\mathbf{X}|\mathbf{Y}) \nabla \log p(\mathbf{X}|\mathbf{Y})^\dagger] \tag{110}$$

and the minimum mean square error (MMSE) matrix is

$$\mathrm{mmse}(\mathbf{X}|\mathbf{Y}) := \mathrm{E}[(\mathbf{X} - \mathrm{E}[\mathbf{X}|\mathbf{Y}])(\mathbf{X} - \mathrm{E}[\mathbf{X}|\mathbf{Y}])^\dagger]. \tag{111}$$

For $t \in \mathcal{T}$ and fixed $\prod_{k=1}^K p(\mathbf{u}_k|\mathbf{x}_k, t)$, choose $\mathbf{\Omega}_{k,t}$, $k = 1, \dots, K$ satisfying $\mathbf{0} \preceq \mathbf{\Omega}_{k,t} \preceq \mathbf{\Sigma}_k^{-1}$ such that

$$\mathrm{mmse}(\mathbf{Y}_k|\mathbf{X}, \mathbf{U}_{k,t}, t) = \mathbf{\Sigma}_k - \mathbf{\Sigma}_k \mathbf{\Omega}_{k,t} \mathbf{\Sigma}_k. \tag{112}$$

Note that such $\mathbf{\Omega}_{k,t}$ exists since $\mathbf{0} \preceq \mathrm{mmse}(\mathbf{X}_k|\mathbf{Y}, \mathbf{U}_{k,t}, t) \preceq \mathbf{\Sigma}_k^{-1}$, for all $t \in \mathcal{T}$, and $k \in \mathcal{K}$.
Using Equation (66), we get

$$\begin{aligned} I(\mathbf{X}_k; \mathbf{U}_k|\mathbf{Y}, t) &\ge \log|\mathbf{\Sigma}_k| - \log|\mathrm{mmse}(\mathbf{X}_k|\mathbf{Y}, \mathbf{U}_{k,t}, t)| \\ &= -\log|\mathbf{I} - \mathbf{\Sigma}_k^{1/2} \mathbf{\Omega}_{k,t} \mathbf{\Sigma}_k^{1/2}|, \end{aligned} \tag{113}$$

where the inequality is due to Lemma 3, and Equation (113) is due to Equation (112).
In addition, we have

$$I(\mathbf{Y}; \mathbf{U}_{\mathcal{S}^c, t}|t) \le \log|\mathbf{\Sigma}_\mathbf{y}| - \log|\mathbf{J}^{-1}(\mathbf{Y}|\mathbf{U}_{\mathcal{S}^c, t}, t)| \tag{114}$$

$$= \log\left| \sum_{k \in \mathcal{S}^c} \mathbf{\Sigma}_\mathbf{y}^{1/2} \mathbf{H}_k^\dagger \mathbf{\Omega}_{k,t} \mathbf{H}_k \mathbf{\Sigma}_\mathbf{y}^{1/2} + \mathbf{I} \right|, \tag{115}$$

where Equation (114) is due to Lemma 3 and Equation (115) is due to to the following equality, which relates the MMSE matrix in Equation (112) and the Fisher information, the proof of which follows,

$$\mathbf{J}(\mathbf{Y}|\mathbf{U}_{\mathcal{S}^c, t}, t) = \sum_{k \in \mathcal{S}^c} \mathbf{H}_k^\dagger \mathbf{\Omega}_{k,t} \mathbf{H}_k + \mathbf{\Sigma}_\mathbf{y}^{-1}. \tag{116}$$

To show Equation (116), we use de Brujin identity to relate the Fisher information with the MMSE as given in the following lemma, the proof of which can be found in [89].

Lemma 4. *Let* $(\mathbf{V}_1, \mathbf{V}_2)$ *be a random vector with finite second moments and* $\mathbf{N} \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma}_N)$ *independent of* $(\mathbf{V}_1, \mathbf{V}_2)$. *Then,*

$$\mathrm{mmse}(\mathbf{V}_2|\mathbf{V}_1, \mathbf{V}_2 + \mathbf{N}) = \mathbf{\Sigma}_N - \mathbf{\Sigma}_N \mathbf{J}(\mathbf{V}_2 + \mathbf{N}|\mathbf{V}_1) \mathbf{\Sigma}_N. \tag{117}$$

From the MMSE of Gaussian random vectors [51],

$$\mathbf{Y} = \mathrm{E}[\mathbf{Y}|\mathbf{X}_{\mathcal{S}^c}] + \mathbf{Z}_{\mathcal{S}^c} = \sum_{k \in \mathcal{S}^c} \mathbf{G}_k \mathbf{X}_k + \mathbf{Z}_{\mathcal{S}^c}, \tag{118}$$

where $\mathbf{G}_k = \mathbf{\Sigma}_{\mathbf{y}|\mathbf{x}_{\mathcal{S}^c}} \mathbf{H}_k^\dagger \mathbf{\Sigma}_k^{-1}$ and $\mathbf{Z}_{\mathcal{S}^c} \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{y}|\mathbf{x}_{\mathcal{S}^c}})$, and

$$\mathbf{\Sigma}_{\mathbf{y}|\mathbf{x}_{\mathcal{S}^c}}^{-1} = \mathbf{\Sigma}_\mathbf{y}^{-1} + \sum_{k \in \mathcal{S}^c} \mathbf{H}_k^\dagger \mathbf{\Sigma}_k^{-1} \mathbf{H}_k. \tag{119}$$

Note that $\mathbf{Z}_{\mathcal{S}^c}$ is independent of $\mathbf{Y}_{\mathcal{S}^c}$ due to the orthogonality principle of the MMSE and its Gaussian distribution. Hence, it is also independent of $\mathbf{U}_{\mathcal{S}^c,q}$.

Thus, we have

$$\text{mmse}\left(\sum_{k\in\mathcal{S}^c}\mathbf{G}_k\mathbf{X}_k\Big|\mathbf{Y},\mathbf{U}_{\mathcal{S}^c,t},t\right)=\sum_{k\in\mathcal{S}^c}\mathbf{G}_k\text{mmse}\left(\mathbf{X}_k|\mathbf{Y},\mathbf{U}_{\mathcal{S}^c,t},t\right)\mathbf{G}_k^\dagger \tag{120}$$

$$=\boldsymbol{\Sigma}_{\mathbf{y}|\mathbf{x}_{\mathcal{S}^c}}\sum_{k\in\mathcal{S}^c}\mathbf{H}_k^\dagger\left(\boldsymbol{\Sigma}_k^{-1}-\boldsymbol{\Omega}_k\right)\mathbf{H}_k\boldsymbol{\Sigma}_{\mathbf{y}|\mathbf{x}_{\mathcal{S}^c}}, \tag{121}$$

where Equation (120) follows since the cross terms are zero due to the Markov Chain $(\mathbf{U}_{k,t},\mathbf{X}_k)\,\circ\!\!-\,\mathbf{Y}\,\circ\!\!-\,(\mathbf{U}_{\mathcal{K}/k,t},\mathbf{X}_{\mathcal{K}/k})$ (see [89, Appendix V]); and Equation (121) follows due to Equation (112) and $\mathbf{G}_k$. Finally, we have

$$\mathbf{J}(\mathbf{Y}|\mathbf{U}_{\mathcal{S}^c,t},t)=\boldsymbol{\Sigma}_{\mathbf{y}|\mathbf{x}_{\mathcal{S}^c}}^{-1}-\boldsymbol{\Sigma}_{\mathbf{y}|\mathbf{x}_{\mathcal{S}^c}}^{-1}\text{mmse}\left(\sum_{k\in\mathcal{S}^c}\mathbf{G}_k\mathbf{X}_k\Big|\mathbf{Y},\mathbf{U}_{\mathcal{S}^c,t},t\right)\boldsymbol{\Sigma}_{\mathbf{y}|\mathbf{x}_{\mathcal{S}^c}}^{-1} \tag{122}$$

$$=\boldsymbol{\Sigma}_{\mathbf{y}|\mathbf{x}_{\mathcal{S}^c}}^{-1}-\sum_{k\in\mathcal{S}^c}\mathbf{H}_k^\dagger\left(\boldsymbol{\Sigma}_k^{-1}-\boldsymbol{\Omega}_{k,t}\right)\mathbf{H}_k \tag{123}$$

$$=\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}+\sum_{k\in\mathcal{S}^c}\mathbf{H}_k^\dagger\boldsymbol{\Omega}_{k,t}\mathbf{H}_k, \tag{124}$$

where Equation (122) is due to Lemma 4; Equation (123) is due to Equation (121); and Equation (124) follows due to Equation (119).

Then, averaging over the time sharing random variable T and letting $\bar{\boldsymbol{\Omega}}_k:=\sum_{t\in\mathcal{T}}p(t)\boldsymbol{\Omega}_{k,t}$, we get, using Equation (113),

$$I(\mathbf{X}_k;\mathbf{U}_k|\mathbf{Y},T)\geq-\sum_{t\in\mathcal{T}}p(t)\log|\mathbf{I}-\boldsymbol{\Sigma}_k^{1/2}\boldsymbol{\Omega}_{k,t}\boldsymbol{\Sigma}_k^{1/2}|$$

$$\geq-\log|\mathbf{I}-\boldsymbol{\Sigma}_k^{1/2}\bar{\boldsymbol{\Omega}}_k\boldsymbol{\Sigma}_k^{1/2}|, \tag{125}$$

where Equation (125) follows from the concavity of the log-det function and Jensen's inequality.

Similarly, using Equation (115) and Jensen's Inequality, we have

$$I(\mathbf{Y};\mathbf{U}_{\mathcal{S}^c}|T)\leq\log\left|\sum_{k\in\mathcal{S}^c}\boldsymbol{\Sigma}_{\mathbf{y}}^{1/2}\mathbf{H}_k^\dagger\bar{\boldsymbol{\Omega}}_k\mathbf{H}_k\boldsymbol{\Sigma}_{\mathbf{y}}^{1/2}+\mathbf{I}\right|. \tag{126}$$

The outer bound on $\mathcal{RI}_{\text{DIB}}$ is obtained by substituting into Equation (66), using Equations (125) and (126), noting that $\boldsymbol{\Omega}_k=\sum_{t\in\mathcal{T}}p(t)\boldsymbol{\Omega}_{k,t}\preceq\boldsymbol{\Sigma}_k^{-1}$ since $\mathbf{0}\preceq\boldsymbol{\Omega}_{k,t}\preceq\boldsymbol{\Sigma}_k^{-1}$, and taking the union over $\boldsymbol{\Omega}_k$ satisfying $\mathbf{0}\preceq\boldsymbol{\Omega}_k\preceq\boldsymbol{\Sigma}_k^{-1}$.

Finally, the proof is completed by noting that the outer bound is achieved with $T=\varnothing$ and multivariate Gaussian distributions $p^\star(\mathbf{u}_k|\mathbf{x}_k,t)=\mathcal{CN}(\mathbf{x}_k,\boldsymbol{\Sigma}_k^{1/2}(\boldsymbol{\Omega}_k-\mathbf{I})\boldsymbol{\Sigma}_k^{1/2})$.

References

1. Tishby, N.; Pereira, F.; Bialek, W. The information bottleneck method. In Proceedings of the Thirty-Seventh Annual Allerton Conference on Communication, Control, and Computing, Allerton House, Monticello, IL, USA, 22–24 September, 1999; pp. 368–377.
2. Pratt, W.K. *Digital Image Processing*; John Willey & Sons INC.: New York, NY, USA, 1991.
3. Yu, S.; Principe, J.C. Understanding Autoencoders with Information Theoretic Concepts. *arXiv* **2018**, arXiv:1804.00057.
4. Yu, S.; Jenssen, R.; Principe, J.C. Understanding Convolutional Neural Network Training with Information Theory. *arXiv* **2018**, arXiv:1804.06537.
5. Kong, Y.; Schoenebeck, G. Water from Two Rocks: Maximizing the Mutual Information. *arXiv* **2018**, arXiv:1802.08887.
6. Ugur, Y.; Aguerri, I.E.; Zaidi, A. A generalization of Blahut-Arimoto algorithm to computing rate-distortion regions of multiterminal source coding under logarithmic loss. In Proceedings of IEEE Information Theory Workshop, ITW, Kaohsiung, Taiwan, 6–10 November 2017.
7. Dobrushin, R.L.; Tsybakov, B.S. Information transmission with additional noise. *IRE Trans. Inf. Theory* **1962**, *85*, 293–304.
8. Witsenhausen, H.S.; Wyner, A.D. A conditional Entropy Bound for a Pair of Discrete Random Variables. *IEEE Trans. on Inf. Theory* **1975**, *IT-21*, 493–501.
9. Witsenhausen, H.S. Indirect Rate Distortion Problems. *IEEE Trans. on Info. Theory* **1980**, *IT-26*, 518–521.
10. Shwartz-Ziv, R.; Tishby, N. Opening the Black Box of Deep Neural Networks via Information. *CoRR* **2017**, abs/1703.00810.
11. Achille, A.; Soatto, S. Emergence of Invariance and Disentangling in Deep Representations. *ArXiv* **2017**, arXiv:1706.01350.
12. McAllester, D.A. A PAC-Bayesian Tutorial with a Dropout Bound. *CoRR* **2013**, abs/1307.2118,
13. Alemi, A.A. Variational Predictive Information Bottleneck. *arXiv* **2019**, arXiv:1910.10831.
14. Mukherjee, S. Machine Learning using the Variational Predictive Information Bottleneck with a Validation Set. *arXiv* **2019**, arXiv:1911.02210.
15. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *CoRR* **2013**, abs/1312.6114.
16. Mukherjee, S. General Information Bottleneck Objectives and their Applications to Machine Learning. *arXiv* **2019**, arXiv:1912.06248.
17. Strouse, D.; Schwab, D.J. The information bottleneck and geometric clustering. *Neural Comput.* **2019**, *31*, 596–612.
18. Painsky, A.; Tishby, N. Gaussian Lower Bound for the Information Bottleneck Limit. *J. Mach. Learn. Res. (JMLR)* **2018** *18*, 7908–7936.
19. Kittichokechai, K.; Caire, G. Privacy-constrained remote source coding. In Proceedings of the 2016 IEEE International Symposium on Information Theory (ISIT), Barcelona, Spain, 10–15 July 2016; pp. 1078–1082.
20. Tian, C.; Chen, J. Successive Refinement for Hypothesis Testing and Lossless One-Helper Problem. *IEEE Trans. Info. Theory* **2008**, *54*, 4666–4681.
21. Sreekumar, S.; Gündüz, D.; Cohen, A. Distributed Hypothesis Testing Under Privacy Constraints. *arXiv* **2018**, arXiv:1807.02764.
22. Aguerri, I.E.; Zaidi, A.; Caire, G.; Shamai (Shitz), S. On the Capacity of Cloud Radio Access Networks with Oblivious Relaying. In Proceedings of 2017 IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, 25–30 June 2017; pp. 2068–2072.
23. Aguerri, I.E.; Zaidi, A.; Caire, G.; Shamai (Shitz), S. On the capacity of uplink cloud radio access networks with oblivious relaying. *IEEE Trans. Inf. Theory* **2019**, *65*, 4575–4596.
24. Stark, M.; Bauch, G.; Lewandowsky, J.; Saha, S. Decoding of Non-Binary LDPC Codes Using the Information Bottleneck Method. In Proceedings of ICC 2019-2019 IEEE International Conference on Communications (ICC), Shanghai, China, 20–24 May 2019; pp. 1–6.
25. Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828.
26. Erdogmus, D. Information theoretic learning: Renyi's entropy and its applications to adaptive system training. Ph.D. Thesis, University of Florida Gainesville, Florida, FL, USA, 2002.

27. Principe, J.C.; Euliano, N.R.; Lefebvre, W.C. *Neural and Adaptive Systems: Fundamentals Through Simulations*; Wiley New York: New York, NY, USA, 2000; Volume 672.

28. Fisher, J.W. *Nonlinear Extensions to the Minumum Average Correlation Energy Filter*; University of Florida: Gainesville, FL, USA, 1997.

29. Jiao, J.; Courtade, T.A.; Venkat, K.; Weissman, T. Justification of logarithmic loss via the benefit of side information. *IEEE Trans. Inf. Theory* **2015**, *61*, 5357–5365.

30. Painsky, A.; Wornell, G.W. On the Universality of the Logistic Loss Function. *arXiv* **2018**, arXiv:1805.03804.

31. Linsker, R. Self-organization in a perceptual network. *Computer* **1988**, *21*, 105–117.

32. Quinlan, J.R. *C4. 5: Programs for Machine Learning*; Elsevier: Amsterdam, The Netherlands, 2014.

33. Chow, C.; Liu, C. Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inf. Theory* **1968**, *14*, 462–467.

34. Olsen, C.; Meyer, P.E.; Bontempi, G. On the impact of entropy estimation on transcriptional regulatory network inference based on mutual information. *EURASIP J. Bioinf. Syst. Biol.* **2008**, *2009*, 308959.

35. Pluim, J.P.; Maintz, J.A.; Viergever, M.A. Mutual-information-based registration of medical images: A survey. *IEEE Trans. Med. Imaging* **2003**, *22*, 986–1004.

36. Viola, P.; Wells III, W.M. Alignment by maximization of mutual information. *Intern. J. Comput. Vision* **1997**, *24*, 137–154.

37. Cesa-Bianchi, N.; Lugosi, G. *Prediction, Learning and Games*; Cambridge University Press: New York, NY, USA, 2006.

38. Lehmann, E.L.; Casella, G. *Theory of Point Estimation*; Springer Science & Business Media: Berlin, Germany, 2006.

39. Bousquet, O.; Elisseeff, A. Stability and generalization. *J. Mach. Learn. Res.* **2002**, *2*, 499–526.

40. Shalev-Shwartz, S.; Shamir, O.; Srebro, N.; Sridharan, K. Learnability, stability and uniform convergence. *J. Mach. Learn. Res.* **2010**, *11*, 2635–2670.

41. Xu, A.; Raginsky, M. Information-theoretic analysis of generalization capability of learning algorithms. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 2521–2530.

42. Russo, D.; Zou, J. How much does your data exploration overfit? Controlling bias via information usage. *arXiv* **2015**, arXiv:1511.05219.

43. Amjad, R.A.; Geiger, B.C. Learning Representations for Neural Network-Based Classification Using the Information Bottleneck Principle. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, doi:10.1109/TPAMI.2019.2909031.

44. Paninski, L. Estimation of entropy and mutual information. *Neural Comput.* **2003**, *15*, 1191–1253.

45. Jiao, J.; Venkat, K.; Han, Y.; Weissman, T. Minimax estimation of functionals of discrete distributions. *IEEE Trans. Inf. Theory* **2015**, *61*, 2835–2885.

46. Valiant, P.; Valiant, G. Estimating the unseen: improved estimators for entropy and other properties. In Proceedings of the Advances in Neural Information Processing Systems 26 (NIPS 2013), Lake Tahoe, USA, 5–10 December 2013; pp. 2157–2165.

47. Chalk, M.; Marre, O.; Tkacik, G. Relevant sparse codes with variational information bottleneck. *arXiv* **2016**, arXiv:1605.07332.

48. Alemi, A.; Fischer, I.; Dillon, J.; Murphy, K. Deep Variational Information Bottleneck. In Proceedings of the International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 April 2017.

49. Achille, A.; Soatto, S. Information Dropout: Learning Optimal Representations Through Noisy Computation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *PP*, 1.

50. Harremoes, P.; Tishby, N. The information bottleneck revisited or how to choose a good distortion measure. In Proceedings of 2007 IEEE International Symposium on Information Theory, Nice, France, 24–29 June 2007; pp. 566–570.

51. Gamal, A.E.; Kim, Y.H. *Network Information Theory*; Cambridge University Press: Cambridge, UK, 2011.

52. Hotelling, H. The most predictable criterion. *J. Educ. Psycology* **1935**, *26*, 139–142.

53. Globerson, A.; Tishby, N. *On the Optimality of the Gaussian Information Bottleneck Curve*; Technical Report, Hebrew University: Jerusalem, Israel, 2004.

54. Chechik, G.; Globerson, A.; Tishby, N.; Weiss, Y. Information Bottleneck for Gaussian Variables. *J. Mach. Learn. Res.* **2005**, *6*, 165–188.

55. Wieczorek, A.; Roth, V. On the Difference Between the Information Bottleneck and the Deep Information Bottleneck. *arXiv* **2019**, arXiv:1912.13480.

56. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* **1977**, *39*, 1–38.

57. Blahut, R. Computation of channel capacity and rate-distortion functions. *IEEE Trans. Inf. Theory* **1972**, *18*, 460–473.

58. Arimoto, S. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Trans. Inf. Theory* **1972**, *IT-18*, 12–20.

59. Winkelbauer, A.; Matz, G. Rate-information-optimal gaussian channel output compression. In Proceedings of 48th Annual Conference on Information Sciences and Systems (CISS), Princeton, NJ, USA, 19–21 March 2014; pp. 1–5.

60. Gálvez, B.R.; Thobaben, R.; Skoglund, M. The Convex Information Bottleneck Lagrangian. *Entropy* **2020**, *20(1)*.

61. Jang, E.; Gu, S.; Poole, B. Categorical Reparameterization with Gumbel-Softmax. *arXiv* **2017**, arXiv:1611.01144.

62. Maddison, C.J.; Mnih, A.; Teh, Y.W. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. *CoRR* **2016**, abs/1611.00712, [1611.00712].

63. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *CoRR* **2014**, abs/1412.6980.

64. Lim, S.H.; Kim, Y.H.; Gamal, A.E.; Chung, S.Y. Noisy Network Coding. *IEEE Trans. Inf. Theory* **2011**, *57*, 3132–3152.

65. Courtade, T.A.; Weissman, T. Multiterminal source coding under logarithmic loss. *IEEE Trans. Inf. Theory* **2014**, *60*, 740–761.

66. I. Csiszár.; J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*; Academic Press: London, UK, 1981.

67. Wyner, A.D.; Ziv, J. The rate-distortion function for source coding with side information at the decoder. *IEEE Trans. Inf. Theory* **1976**, *22*, 1–10.

68. Steinberg, Y. Coding and Common Reconstruction. *IEEE Trans. Inf. Theory* **2009**, *IT-11*, 4995–5010.

69. Benammar, M.; Zaidi, A. Rate-Distortion of a Heegard-Berger Problem with Common Reconstruction Constraint. In Proceedings of International Zurich Seminar on Information and Communication, Cambridge, MA, USA, 1–6 July 2016.

70. Benammar, M.; Zaidi, A. Rate-distortion function for a heegard-berger problem with two sources and degraded reconstruction sets. *IEEE Trans. Inf. Theory* **2016**, *62*, 5080–5092.

71. Sutskover, I.; Shamai, S.; Ziv, J. Extremes of Information Combining. *IEEE Trans. Inf. Theory* **2005**, *51*, 1313–1325.

72. Land, I.; Huber, J. Information Combining. *Found. Trends Commun. Inf. Theory* **2006**, *03*, 227–230.

73. Wyner, A.D. On source coding with side information at the decoder. *IEEE Trans. Inf. Theory* **1975**. *21*, 294–300.

74. Ahlswede, R.; Korner, J. Source coding with side information and a converse for degraded broadcast channels. *IEEE Trans. Inf. Theory* **1975**, *21*, 629–637.

75. Makhdoumi, A.; Salamatian, S.; Fawaz, N.; Médard, M. From the information bottleneck to the privacy funnel. In Proceedings of IEEE Information Theory Workshop, ITW, Hobart, Tasamania, Australia, 2–5 November 2014; pp. 501–505.

76. Asoodeh, S.; Diaz, M.; Alajaji, F.; Linder, T. Information Extraction Under Privacy Constraints. *IEEE Trans. Inf. Theory* **2019**, *65*, 1512–1534.

77. Erkip, E.; Cover, T.M. The efficiency of investment information. *IEEE Trans. Inf. Theory* **1998**, *44*, 1026–1040.

78. Hinton, G.E.; van Camp, D. Keeping the Neural Networks Simple by Minimizing the Description Length of the Weights. *Proceedings of the Sixth Annual Conference on Computational Learning Theory*; ACM: New York, NY, USA, 1993; COLT '93, pp. 5–13.

79. Gilad-Bachrach, R.; Navot, A.; Tishby, N. An Information Theoretic Tradeoff between Complexity and Accuracy. In *Learning Theory and Kernel Machines*; Lecture Notes in Computer Science, Springer: Berlin/Heidelberg, Germany, 2003; pp. 595–609.

80. Vera, M.; Piantanida, P.; Vega, L.R. The Role of Information Complexity and Randomization in Representation Learning, 2018, [arXiv:stat.ML/1802.05355].

81. Huang, S.L.; Makur, A.; Wornell, G.W.; Zheng, L. On Universal Features for High-Dimensional Learning and Inference. *arXiv* **2019**, arXiv:1911.09105.

82. Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; Lerchner, A. β-VAE: Learning basic visual concepts with a constrained variational framework. In Proceedings of International Conference on Learning Representations, Toulon, France, 24–26 April 2017.

83. Kurakin, A.; Goodfellow, I.; Bengio, S. Adversarial Machine Learning at Scale. *arXiv* **2016**, arXiv:1611.01236.

84. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv* **2017**, arXiv:1706.06083.

85. Engstrom, L.; Tran, B.; Tsipras, D.; Schmidt, L.; Madry, A. A rotation and a translation suffice: Fooling cnns with simple transformations. *arXiv* **2017**, arXiv:1712.02779.

86. Pensia, A.; Jog, V.; Loh, P.L. Extracting robust and accurate features via a robust information bottleneck. *arXiv* **2019**, arXiv:1910.06893.

87. Guo, D.; Shamai, S.; Verdu, S. Mutual Information and Minimum Mean-Square Error in Gaussian Channels. *IEEE Trans. Inf. Theory* **2005**, *51*, 1261–1282.

88. Dembo, A.; Cover, T.M.; Thomas, J.A. Information theoretic inequalities. *IEEE Trans. Inf. Theory* **1991**, *37*, 1501–1518.

89. Ekrem, E.; Ulukus, S. An Outer Bound for the Vector Gaussian CEO Problem. *IEEE Trans. Inf. Theory* **2014**, *60*, 6870–6887.

90. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Willey & Sons INC.: New York, NY, USA, 1991.

91. Aguerri, I.E.; Zaidi, A. Distributed information bottleneck method for discrete and Gaussian sources. In Proceedings of International Zurich Seminar on Information and Communication, IZS, Zurich, Switzerland, 21–23 February 2018.

92. Aguerri, I.E.; Zaidi, A. Distributed Variational Representation Learning. *arXiv* **2018**, arxiv:1807.04193.

93. Winkelbauer, A.; Farthofer, S.; Matz, G. The rate-information trade-off for Gaussian vector channels. In Proceedings of the The 2014 IEEE International Symposium on Information Theory, Honolulu, HI, USA, 29 June–4 July 2014; pp. 2849–2853.

94. Ugur, Y.; Aguerri, I.E.; Zaidi, A. Rate region of the vector Gaussian CEO problem under logarithmic loss. In Proceedings of 2018 IEEE Information Theory Workshop (ITW), Guangzhou, China , 25–29 November 2018.

95. Ugur, Y.; Aguerri, I.E.; Zaidi, A. Vector Gaussian CEO Problem under Logarithmic Loss. *arXiv* **2018**, arXiv:1811.03933.

96. Simeone, O.; Erkip, E.; Shamai, S. On Codebook Information for Interference Relay Channels With Out-of-Band Relaying. *IEEE Trans. Inf. Theory* **2011**, *57*, 2880–2888.

97. Sanderovich, A.; Shamai, S.; Steinberg, Y.; Kramer, G. Communication Via Decentralized Processing. *IEEE Tran. Inf. Theory* **2008**, *54*, 3008–3023.

98. Lapidoth, A.; Narayan, P. Reliable communication under channel uncertainty. *IEEE Trans. Inf. Theory* **1998**, *44*, 2148–2177.

99. Cover, T.M.; El Gamal, A. Capacity Theorems for the Relay Channel. *IEEE Trans. Inf. Theory* **1979**, *25*, 572–584.

100. Xu, C.; Tao, D.; Xu, C. A survey on multi-view learning. *arXiv* **2013**, arXiv:1304.5634.

101. Blum, A.; Mitchell, T. Combining labeled and unlabeled data with co-training. In Proceedings of the eleventh annual conference on Computational learning theory, Madison, WI, USA, 24–26 July 1998; pp. 92–100.

102. Dhillon, P.; Foster, D.P.; Ungar, L.H. Multi-view learning of word embeddings via CCA. In Proceedings of the 2011 Advances in neural information processing systems, Granada, Spain, 12–17 December 2011; pp. 199–207.

103. Kumar, A.; Daumé, H. A co-training approach for multi-view spectral clustering. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), Bellevue, WA, USA, 28 June–2 July 2011; pp. 393–400.

104. Gönen, M.; Alpaydın, E. Multiple kernel learning algorithms. *J. Mach. Learn. Res.* **2011**, *12*, 2211–2268.

105. Jia, Y.; Salzmann, M.; Darrell, T. Factorized latent spaces with structured sparsity. In Proceedings of the Advances in Neural Information Processing Systems, Brno, Czech, 24–25 June 2010; pp. 982–990.

106. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer science & business media: Berlin, Germany, 2013.

107. Strouse, D.J.; Schwab, D.J. The deterministic information bottleneck. *Mass. Inst. Tech. Neural Comput.* **2017**, *26*, 1611–1630.

108. Homri, A.; Peleg, M.; Shitz, S.S. Oblivious Fronthaul-Constrained Relay for a Gaussian Channel. *IEEE Trans. Commun.* **2018**, *66*, 5112–5123.

109. du Pin Calmon, F.; Fawaz, N. Privacy against statistical inference. In Proceedings of 2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 1–5 October 2012.

110. Karasik, R.; Simeone, O.; Shamai, S. Robust Uplink Communications over Fading Channels with Variable Backhaul Connectivity. *IEEE Trans. Commun.* **2013**, *12*, 5788–5799.

111. Chen, Y.; Goldsmith, A.J.; Eldar, Y.C. Channel capacity under sub-Nyquist nonuniform sampling. *IEEE Trans. Inf. Theory* **2014**, *60*, 4739–4756.

112. Kipnis, A.; Eldar, Y.C.; Goldsmith, A.J. Analog-to-Digital Compression: A New Paradigm for Converting Signals to Bits. *IEEE Signal Process. Mag.* **2018**, *35*, 16–39.