Perspectives on Information Bottleneck Problems

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

***Abstract*—This paper focuses on variants of the bottleneck problem taking an information theoretic perspective. The intimate connections of this setting to: remote source-coding, information combining, common reconstruction, the Wyner-Ahlswede-Korner problem, the efficiency of investment information, CEO source coding under logarithmic-loss distortion measure and others will be highlighted. We discuss 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 tradeoffs between rates (i.e. complexity) and information (i.e. accuracy) 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 'accuracy' for the Gaussian information bottleneck, under 'complexity' constraints.**

I. Introduction

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 $\phi : \mathcal{X} \to \hat{\mathcal{Y}}$ and an additive 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 $\phi^\star \in \mathcal{F}$ that minimizes the expected risk

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

Because the joint distribution $P_{X,Y}$ is unknown, in practice the risk (1) (also called *population risk*) 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 $P_{X,Y}$. Also, it is important to restrict the set $\mathcal{F}$ of admissible predictors to a low-complexity class to prevent overfitting. This leads to the abstract inference problem shown in Figure 1.

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

Fig. 1. An abstract inference model for learning.

In this paper, we investigate the inference problem of Figure 1 in the case in which the loss function $\ell(\cdot)$ is the logarithmic-loss fidelity measure, given by

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

where $\hat{y}(\cdot)$ designates a probability distribution on $\mathcal{Y}$ and $\hat{y}(y)$ is the value of this distribution evaluated for the outcome $y \in \mathcal{Y}$. The choice of a 'good" loss function is often controversial in statistical learning theory, and although a complete and rigorous justification of the usage of logarithmic loss as a fidelity measure in learning theory is still awaited, partial explanations appeared in [1] and, especially in [2] where it is shown 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. Also, we constrain the complexity of the predictors by using mutual information as a regularizer term. This is inline with recent works [3], [4] that show that the generalization error can be upper-bounded using the mutual information between the input dataset and the output of the predictor – see also [5], [6] where the *stability* of an algorithm is controlled by constraining the mutual information between its input and output.

II. Information Bottleneck

The Information Bottleneck (IB) method [7] elegantly captures the above mentioned viewpoint of seeking the right balance between data fit and generalization (complexity) by using the mutual information both as a cost function and as a regularizer term. Specifically, IB formulates the problem of extracting the relevant information that some variable $X \in \mathcal{X}$ provides about another one $Y \in \mathcal{Y}$ which is of interest as hat of finding a representation U of X that is maximally informative about Y (i.e., large mutual information $I(U;Y)$) while being minimally informative about X (i.e., small $I(U;X)$). The solution of this problem can be found by solving a Lagrange formulation. For example, the representation U that maximizes $I(U;Y)$ while keeping $I(U;X)$ smaller than a prescribed threshold is the solution of the following Langrangian problem,

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

where β designates the Lagrange multiplier. In the IB framework, $I(U;Y)$ is referred to as the *accuracy* of U and $I(U;X)$ is referred to as the *complexity* of U, where complexity is measured here the minimum description length (or rate) at which the observation is compressed. If the joint distribution $P_{X,Y}$ is known, or can be approximated to high accuracy, a Blahut-Arimoto [8] type iterative algorithm enables to compute near[1] optimal accuracy-complexity mappings by iterating over a set of self-consistent equations [7]. In the case in which the joint distribution $P_{X,Y}$ is unknown, a variational inference type algorithm in which the mappings are parametrized by neural networks and the bound approximated by Markov sampling and optimized with stochastic gradient descent can be employed [9] – the approach makes usage of Kingma *et al.* reparametrization trick [10].

III. Connections

A. Remote Source Coding

As already observed in [11], the IB problem is essentially a remote source coding problem in which the distortion is measured under the logarithmic loss measure. More specifically, let Y designate a memoryless remote source; and X a noisy version of it that is observed at an encoder. The encoder uses R bits per sample to describe its observation to a decoder which is interested in reconstructing the remote source Y to within an average distortion level D, i.e.,

$$\mathbb{E}[\ell_{\log}^{(n)}(\mathbf{y},\hat{\mathbf{y}})] \leq D \tag{4}$$

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

$$\ell_{\log}^{(n)}(\mathbf{y},\hat{\mathbf{y}}) = \frac{1}{n}\sum_{i=1}^{n}\ell_{\log}(y_i,\hat{y}_i) \tag{5}$$

with the per-letter distortion defined as specified by (2). The minimum achievable rate for which the distortion constraint is met is given by [12], [13]

$$R = \min_{P_{U|X}} I(X,U) \tag{6}$$

where the minimization is over all auxiliary random variables U that satisfy that $U \multimap X \multimap Y$ forms a Markov Chain in this order and

$$D \geq H(Y|U). \tag{7}$$

Noting that the average log-loss distortion constraint (8) is equivalent to

$$I(U;Y) \leq H(Y) - D, \tag{8}$$

it is clear that the rate-distortion region of the aforementioned remote source-coding problem under logarithmic loss and the accuracy-complexity region of the IB method can be inferred from each other.

[1]While the BA algorithm is known to converge to the optimal solution for, e.g., the problem of rate-distortion computation of standard point-to-point lossy compression, it only converges to stationary points in the context of IB.

B. Common Reconstruction

Consider the problem of source coding with side information at the decoder, i.e., the well known Wyner-Ziv setting [14], 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

$$R + D \geq H(X|Y). \tag{9}$$

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

$$\psi(U,Y) = \Pr[X = x|U,Y] \tag{10}$$

for which it is easy to see that

$$\mathbb{E}[\ell_{\log}(X,\psi(U,Y))] = H(X|U,Y). \tag{11}$$

Now, assume that we constrain the coding in a manner that the encoder be able to produce an exact copy of the compressed source constructed by the decoder. This requirement, termed *common reconstruction* constraint (CR), was introduced and studied by Steinberg in [15] 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 (10) is no longer possible as it cannot be reproduced at the sender. 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

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

$$D \geq H(X|U) \tag{12b}$$

for some auxiliary random variable for which $U \multimap X \multimap Y$ holds. Observe that (12b) 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 (12b) with equality, i.e.,

$$R(D) = \min_{P_{U|X}\,:\, I(U;X)=H(X)-D} I(U;X|Y). \tag{13}$$

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

$$I(U;Y) = I(U;X) - I(U;X|Y). \tag{14}$$

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 [16], [17].

C. Information Combining

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

$$I(X;U,Y) = I(U;X) + I(Y;X) - I(U;Y) \tag{15}$$
$$= R + I(Y;X) - I(U;Y) \tag{16}$$

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* [18], [19], 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.

D. Wyner-Ahlswede-Körner 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, seperately by Wyner [20] and Ahlswede and Körner [21]. For given $R_X = R$, the minimum rate R_Y that is needed to recover Y losslessly is

$$R_Y^\star(R) = \min_{P_{U|X} \,:\, I(U;X) \le R} H(Y|U). \tag{17}$$

So, we get

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

E. Efficiency of Investment Information

Let Y model a stock market data; and X some correlated information. In [22], 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 [20] and Ahlswede-Körner [21]; and, so, also to the IB problem. The work [22] provides explicit analytic solutions for two horce race examples, jointly binary and jointly Gaussian horce races.

IV. 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 2. 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 $(X_1, \ldots, X_K)$ that are observed each at a distinct predictor. Throughout, we assume that the following Markov chain holds for all $k \in \mathcal{K} := \{1, \ldots, K\}$,

$$X_k \multimap Y \multimap X_{\mathcal{K}/k}. \tag{18}$$

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

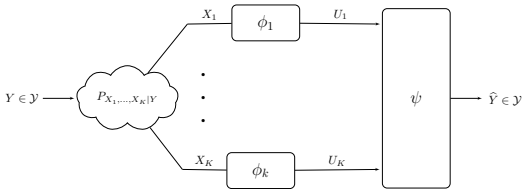

Fig. 2. A model for distributed, e.g., multi-view, learning.

A. An example: multi-view learning

In many data analytics problems, data is collected from various sources of information or feature extractors; and is 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* [23], has received growing interest; and various algorithms exist, sometimes under references such as *co-training* [24]–[27], *multiple kernel learning* [27] and *subspace learning* [28]. By using distinct encoder mappings to represent distinct groups of data, and jointly optimizing over all mappings to remove redundancy, multiview learning offers a degree of flexibility that is not only desirable in practice but is likely to result in better learning capability. Actually, as shown in [29], local learning algorithms produce less errors than global ones. Viewing the problem as that of function approximation, the intuition is that it is usually non-easy to find a unique function that holds good predictability properties in the entire data space.

Besides, the distributed learning of Figure 2 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 they access independent noisy versions of the entire dataset. Two such examples are Google Goggles and Siri in which the locally collected data is processed on clouds.

B. Optimal accuracy-complexity tradeoff region

The distributed IB problem of Figure 2 is studied in [30], [31] from information-theoretic grounds. For both discrete memoryless (DM) and memoryless vector Gaussian models, the authors establish fundamental limits of learning in terms of optimal tradeoffs between accuracy and complexity. The result for discrete sources is reproduced here for completeness.

Theorem 1. *([30], [31]) The accuracy-complexity region $\mathcal{IR}_{\mathrm{DIB}}$ of the distributed learning problem with $P_{X_\mathcal{K},Y}$ for which the Markov chain* (21) *holds, is given by the union of all tuples $(\Delta, R_1, \ldots, R_K) \in \mathbb{R}_+^{K+1}$ satisfying for all $\mathcal{S} \subseteq \mathcal{K}$,*

$$\Delta \le \sum_{k \in \mathcal{S}} [R_k - I(X_k; U_k|Y,T)] + I(Y; U_{\mathcal{S}^c}|T), \tag{19}$$

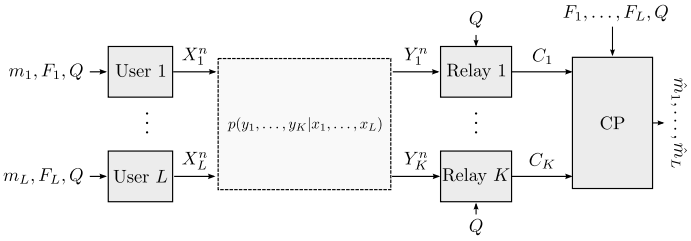

Fig. 3. CRAN model with oblivious relaying and time-sharing.

for some set of pmfs $\mathbf{P} := \{P_{U_k|X_k,T},\ldots,P_{U_k|X_k,T},P_T\}$ *with joint distribution of the form*

$$p_T(t)p_Y(y)\prod_{k=1}^{K}p_{X_k|Y}(x_k|y)\prod_{k=1}^{K}p_{U_k|X_k,T}(u_k|x_k,t). \quad (20)$$

C. Cloud RAN

Consider the discrete memoryless (DM) CRAN model shown in Figure 3. 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 [32], 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* [33], [34]. 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.

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 $i \in [1:n]$,

$$Y_{k,i} \multimap X_{\mathcal{L},i} \multimap Y_{\mathcal{K}/k,i} \quad (21)$$

forms a Markov chain in this order.

Theorem 2 ([35])**.** *For the class of DM CRANs with oblivious relay processing and enabled time-sharing for which* (21) *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*

$$p(q)\prod_{l=1}^{L}p(x_l|q)\prod_{k=1}^{K}p(y_k|x_{\mathcal{L}})\prod_{k=1}^{K}p(u_k|y_k,q). \quad (22)$$

The direct part of Theorem 2 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 [36, Theorem 3] with joint decompression and decoding (CF-JD). Alternatively, the rate region of Theorem 2 can also be obtained by a direct application of the noisy network coding (NNC) scheme of [37, Theorem 1].

D. Distributed Source Coding under Logarithmic Loss

Key element to the proof of the converse part of Theorem 2 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 [38]. A key step in [38] 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 on this result, in our converse proof of Theorem 2 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 $F_{\mathcal{L}}$. Also, the rate region of the vector Gaussian CEO problem under logarithmic loss distortion measure has been found recently in [39], [40].

V. Outlook

Among interesting problems that are left unaddressed in this paper 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 [33]. 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 [41], 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 rate could be conveyed over the CRAN uplink [42]. Alternatively, one may consider finding the worst-case noise under given input distributions, e.g., Gaussian, and rate-constrained compression at the relays. Finally, there are interesting aspects that address processing constraints of continuous waveforms, e.g., such as sampling at a given rate [43], [44] with focus on remote logarithmic distortion [38], which in turn boils down to the distributed bottleneck problem [30], [31].

References

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

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

[3] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Advances in Neural Information Processing Systems*, 2017, pp. 2521–2530.

[4] D. Russo and J. Zou, "How much does your data exploration overfit? controlling bias via information usage," *arXiv preprint arXiv:1511.05219*, 2015.

[5] O. Bousquet and A. Elisseeff, "Stability and generalization," *Journal of Machine Learning Research*, vol. 2, no. Mar, pp. 499–526, 2002.

[6] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, "Learnability, stability and uniform convergence," *Journal of Machine Learning Research*, vol. 11, no. Oct, pp. 2635–2670, 2010.

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

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

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

[10] D. P. Kingma and M. Welling, "Auto-encoding variational bayes." *CoRR*, vol. abs/1312.6114, 2013. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1312.html#KingmaW13

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

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

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

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

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

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

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

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

[19] I. Land and J. Huber, "Information combining," *Foundations and Trends in Commun. and Inform. Theory*, vol. 03, pp. 227–230, Nov. 2006.

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

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

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

[23] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013.

[24] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual conference on Computational learning theory*. ACM, 1998, pp. 92–100.

[25] P. Dhillon, D. P. Foster, and L. H. Ungar, "Multi-view learning of word embeddings via CCA," in *Advances in neural information processing systems*, 2011, pp. 199–207.

[26] A. Kumar and H. Daumé, "A co-training approach for multi-view spectral clustering," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 393–400.

[27] M. Gönen and E. Alpaydın, "Multiple kernel learning algorithms," *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2211–2268, 2011.

[28] Y. Jia, M. Salzmann, and T. Darrell, "Factorized latent spaces with structured sparsity," in *Advances in Neural Information Processing Systems*, 2010, pp. 982–990.

[29] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

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

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

[32] O. Simeone, E. Erkip, and S. Shamai, "On codebook information for interference relay channels with out-of-band relaying," *IEEE Trans. Inf. Theory,*, vol. 57, no. 5, pp. 2880–2888, May 2011.

[33] A. Sanderovich, S. Shamai, Y. Steinberg, and G. Kramer, "Communication via decentralized processing," *IEEE Tran. on Info. Theory,*, vol. 54, no. 7, pp. 3008–3023, Jul. 2008.

[34] A. Lapidoth and P. Narayan, "Reliable communication under channel uncertainty," *IEEE Trans. on Inf. Theory*, vol. 44, no. 6, pp. 2148–2177, Oct 1998.

[35] I. E. Aguerri, A. Zaidi, G. Caire, and S. S. (Shitz), "On the capacity of uplink cloud radio access networks with oblivious relaying," *IEEE Trans. on Info. Theory. Available at arxiv.org/abs/1710.09275*, 2017.

[36] T. M. Cover and A. El Gamal, "Capacity theorems for the relay channel," *IEEE Trans. Inf. Theory*, vol. IT-25, pp. 572–584, Sep. 1979.

[37] S. H. Lim, Y.-H. Kim, A. E. Gamal, and S.-Y. Chung, "Noisy network coding," *IEEE Trans. Inf. Theory*, vol. 57, pp. 3132–3152, May 2011.

[38] T.-A. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," *IEEE Trans. on Info. Theory*, vol. 60, pp. 740–761, Jan. 2014.

[39] Y. Ugur, I.-E. Aguerri, and A. Zaidi, "Rate region of the vector gaussian ceo problem under logarithmic loss," in *IEEE Int. Workshop on Info. Theory*, Nov. 2018.

[40] ——, "Vector gaussian ceo problem under logarithmic loss," *IEEE Trans. on Info. Theory. Submitted for publication*, 2018.

[41] A. Homri, M. Peleg, and S. S. (Shitz), "Oblivious fronthaul-constrained relay for a gaussian channel," *IEEE Trans. on Communications*, vol. 66, pp. 5112–5123, Nov. 2018.

[42] R. Karasik, O. Simeone, and S. Shamai, "Robust uplink communications over fading channels with variable backhaul connectivity," *IEEE Trans. on Communications*, vol. 12, pp. 5788–5799, Nov. 2013.

[43] Y. Chen, A.-J. Goldsmith, , and Y.-C. Eldar, "Channel capacity under sub-nyquist nonuniform sampling," *IEEE Trans. Inf. Theory*, vol. 60, pp. 4739–4756, Aug. 2014.

[44] A. Kipnis, Y.-C. Eldar, and A.-J. Goldsmith, "Analog-to-digital compression: A new paradigm for converting signals to bits," *IEEE Signal Processing Magazine*, pp. 16–39, May 2018.