On The Multi-View Information Bottleneck Representation

Teng-Hui Huang
Electrical and Computer Engineering
Purdue University
West Lafayette, IN, USA
huan1456@purdue.edu

Aly El Gamal
Emerging Technologies Lab
InterDigital
Los Altos, CA, USA
aly.elgamal@interdigital.com

Hesham El Gamal
Electrical and Information Engineering
The University of Sydney
NSW, Australia
hesham.elgamal@sydney.edu.au

Abstract—In this work, we generalize the information bottleneck (IB) approach to the multi-view learning context. The exponentially growing complexity of the optimal representation motivates the development of two novel formulations with more favorable performance-complexity tradeoffs. The first approach is based on forming a stochastic consensus and is suited for scenarios with significant representation overlap between the different views. The second method, relying on incremental updates, is tailored for the other extreme scenario with minimal representation overlap. In both cases, we extend our earlier work on the alternating directional methods of multiplier (ADMM) solver and establish its convergence and scalability. Empirically, we find that the proposed methods outperform state-of-the-art approaches in multi-view classification problems under a broad range of modeling parameters.

Index Terms—Information bottleneck, consensus ADMM, non-convex optimization, classification, multi-view learning.

I. INTRODUCTION

Recently, learning from multi-view data has drawn significant interests in the machine learning and data science communities (e.g., [1]–[5]). In this context, a view of data is a description or an observation about the source. For example, an object can be described in words or images. With multi-view data, one would naturally expect an improved performance from learning a more complete representation [6].

The challenges in multi-view learning are two-folded: First, one can combine all views of observations to form one giant view that loses no information contained within them but would suffer from exponential growth of the dimensionality of the merged observation. We call this the performance-complexity trade-off. Second, following this, if one instead opts to extract either view-shared or view-specific relevant features from each view, then heterogeneous forms of observations (e.g., audio and images) would make it difficult to learn low-complexity and meaningful representations with a unified method. This implies that the amount of representation overlap across all view observations is important for efficient multi-view learning.

In addressing the challenges, several recent works have attempted to apply the IB [7] principle to multi-view learning for it matches the objective well, that is, trading-off relevance and complexity in extracting both view-shared and view-specific features [8]–[11]. This generalized multi-view IB (MvIB) method is known to be a special case of the multi-terminal remote source coding problems with logarithmic loss. The logarithmic loss corresponds to soft reconstruction where the likelihood of all possible outcomes is received in contrast to a reconstructed symbol in conventional source coding problems [12]–[16].

In literature, the achievable region for the remote source coding problem is characterized in [17] for discrete cases, and recently, jointly Gaussian cases as well [13]. Along with the characterization, a variety of variational inference-based algorithms are proposed [14], [15]. This type of algorithm introduces extra variational variables to facilitate the optimization as by fixing one of the two sets of variables, the overall objective function is convex w.r.t. to the other set of variables. Optimizing in this alternating fashion, the convergence is assured.

Extending this line of research, our approach is rooted in a top-down information-theoretic formulation closely related to the optimal characterization of MvIB. Moreover, contrary to [11] which relies on black-box deep neural networks, we propose two constructive information theoretic formulations with performance comparable to that of the optimal joint view approach. The two approaches are motivated by two extreme multi-view learning scenarios: The first is characterized by a significant representation overlap between the different views which favors our consensus-complement two-stage formulation, whereas the second extreme scenario is characterized by a minimal representation overlap leading to our incremental update approach.

Different from existing variational inference-based algorithms that avoid dealing with the non-convexity of the overall objective function, in both of the proposed methods, we adopt the non-convex consensus ADMM as the main tool in deriving our solvers [18]–[20]. These new solvers can, therefore, be viewed as generalizations of our earlier work on the single-view ADMM algorithm [21]. More specifically, in the consensus-complement version, we separate the proposed Lagrangian into consensus and complement sub-objective functions and then proceed to solve the optimization problem in two steps. The new ADMM solver can hence efficiently form a consensus representation in large-scale multi-view learning problems with significantly lower dimensions.
compared to joint-view approaches. The same intuition is applied to the increment update approach as detailed in the sequel. Finally, we prove the convergence of our solvers under significantly milder constraints; as compared with earlier convergence results on this class of solvers [22]–[24]. More specifically, we relax the need for a strongly convex sub-objective function and, moreover, establish the linear rate of convergence around a locally stationary point in each case.

II. MULTIVIEW INFORMATION BOTTLENECK

Given $V$ views with observations $\{X^{(i)}\}_{i=1}^V$ generated from a target variable $Y$, we aim to find a set of individual representations $\{Z\}$ that is most compressed w.r.t. the observations of each view $X^{(i)}$, $\forall i \in [V]$ but at the same time maximizing their relevance toward the target variable $Y$ through $X^{(i)}$. Using a Lagrange multiplier formulation, the problem can be casted as:

$$\mathcal{L}_{\text{MvIB}} := \gamma I(\{X\}; Y) - I(Y; Z),$$

where $\{X\}_V$ denotes the set of all $V$-views of observations and $\{Z\}$ is the set of representations to be designed. Note that if the observations are combined in this manner and treated as one view, (1) reduces to the standard IB and can be solved with any existing single-view solver. However, combining all the observations in one giant view will result in an exponential increase in complexity (curse of dimensionality).

A basic assumption in the multiview learning literature is the conditional independence [9], [25], where the observations of all views $\{X^{(i)}\}_{i=1}^V$ are independent given the target variable $Y$. That is, $p(x_i | y) = \prod_{i=1}^V p(x^{(i)} | y)$. In the next two sections, we use this conditional independence assumption while constraining the set of allowable latent representations $\{Z\}$ to develop two novel information-theoretic formulations of the Multi-view IB (MvIB) problem.

A. Consensus-Complement Form

Inspired by the co-training methods in multi-view research [25], we design the set of latent representations $\{Z\}$ to consist of a consensus representation $Z_c$ and view-specific complement components $Z^{(i)}_c$, $\forall i \in [V]$. Then, by the chain rule of mutual information, the Lagrangian of (1) becomes:

$$\mathcal{L}_{\text{con}} = \gamma I(Z_c; \{X\}) - I(Z_c; Y) + \sum_{i=1}^V \gamma I(Z^{(i)}_c; \{X\}|Z_c, \{Z^{(i)}_{c,i-1}\}) - I(Y; Z^{(i)}_c|Z_c, \{Z^{(i)}_{c,i-1}\}),$$

(2)

where the sequence $\{Z^{(i)}_{c,i}\} := \{Z^{(1)}_c, \ldots, Z^{(i)}_c\}$ is defined to represent the accumulated complement views. To further simplify the above, the set of possible representations is subject to the following constraints (similar to [11], [25]):

- There always exist constants $\kappa_i, \forall i \in [V]$, independent of the observations $\{X\}$ such that $\kappa_i I(Z_c; X^{(i)}) = I(Z_c; X^{(i)}|X_{i-1})$, accounting the mutual dependency of view observations.
- $Y \rightarrow X^{(i)} \rightarrow Z^{(i)}_c \leftarrow Z_c$ forms a Markov chain. That is, $Z_c$ is side-information for $Z^{(i)}_c$.
- Given the consensus $Z_c$, $\{Z^{(i)}_c\}_{i=1}^V$ are independent.

Under these restrictions, (2) can be rewritten as the superposition of two parts, i.e., $\mathcal{L} := \mathcal{L} + \sum_{i=1}^V \mathcal{L}^{(i)}$, where $\mathcal{L}$ is defined as the consensus Lagrangian:

$$\mathcal{L} := \sum_{i=1}^V \gamma_i I(Z^{(i)}_c; X^{(i)}) - I(Z^{(i)}_c; Y),$$

while the second as the complement Lagrangian, consisting of $V$ terms with each view-specific sub-objective defined as:

$$\mathcal{L}^{(i)} := \gamma I(Z^{(i)}_c; X^{(i)}|Z_c) - I(Z^{(i)}_c; Y|Z_c), \forall i \in [V].$$

Then we recast $\mathcal{L}$ in (3) as:

$$\tilde{\mathcal{L}} := -\sum_{i=1}^V \gamma_i H(Z^{(i)}|Z_c) + (\gamma - 1) H(Z_c|Z_c) + H(Z^{(i)}|Z_c, Y).$$

(5)

Similarly, $\forall i \in [V]$, we rewrite (4) as:

$$\mathcal{L}^{(i)} := -\gamma H(Z^{(i)}_c|Z_c, X^{(i)}),$$

$$+ (\gamma - 1) H(Z^{(i)}_c|Z_c) + H(Z^{(i)}_c|Z_c, Y).$$

(6)

By representing the discrete (conditional) probabilities as vectors/tensors, we can solve (5) and (6) with augmented Lagrangian methods. Define the following vectors:

$$p_{z|x,i} := p(z|z_i x^{(i)}),$$

$$p_{z} := p(z),$$

$$p_{z|y} := p(z|y),$$

(7)

where $N_i := |X^{(i)}|, \forall i \in [V], L := |Z|, K := |Y|$. For clarity, we rewrite the primal variables for each view as $p_{z|x,i} := p_{z|x}^i$, and cascade the augmented variables which gives $q := [p_{z}^T, p_{z|y}^T]^T$. On the other hand, for the complement part, we define the following tensors:

$$\pi_{x,y,m,n,r} := P(Z^{(i)}_c = z^{(i)}_x|Z_c = z_{c,m}, X^{(i)} = x^{(i)}_r),$$

$$\pi_{x,y,m,n} := P(Z^{(i)}_c = z^{(i)}_x|Z_c = z_{c,m}, Y = y_r),$$

$$\pi_{z|x,y,m,n} := P(Z^{(i)}_c = z^{(i)}_x|Z_c = z_{c,m}).$$

Then we present the consensus-complement MvIB augmented Lagrangian as follows. For the consensus part:

$$\mathcal{L}_c(\{p_{i}\}_{i=1}^V, q, \{\nu_{i}\}_{i=1}^V) = \sum_{i=1}^V F_i(p_{i}) + (\nu_i A_i p_i - q) + \frac{c}{2} \|A_i p_i - q\|^2 + G(q),$$

(9)

where $\|\cdot\|$ is 2-norm, $c > 0$ is a penalty coefficient, $A_i p_i - q$ the linear penalty for each view $i \in [V]$, encouraging the
variables $q$ and each $p_i$ to satisfy the marginal probability and the Markov chain conditions with $A := [A^T_{x,i} \ A^T_{y,i}]^T$. Specifically, let $\otimes$ denote the Kronecker product, $A_{x,i} := I_L \otimes P^T_{x,i}$, $A_{y,i} := I_L \otimes P^T_{x,i|y}$, where $P_{x,i|y}$ is the matrix form of the conditional distribution $p(x(i)|y)$ with each entry $(m,n)$ equals to $p(x(i)|y_n)$. As for the complement part:

$$F_{e,i} = -\gamma H(Z_e^{(i)}|Z_e, X_e^{(i)}),$$

$$G_{e,i} = (\gamma - 1)H(Z_e^{(i)}) + H(Z_e^{(i)}|Z_e, Y).$$

Then to vectorize the tensors and express them as linear penalties, define $\pi_{y,i}[t]$ a realization of the consensus representation $z_e \in Z_e$, by Bayes’ rule, we have $\pi_{y,i}[t] = A_{z_e|y}[t]A_{x,i}[t]$, where $A_{z_e|y}[t]$ is a diagonal matrix formed from the vector $p_{z_e|y}[t]$, $\forall y \in Y$ and similarly for $\pi_{y,i}[t]$. To simplify notation, define $A_{x,i}[t] := A_{z_e|y}[t]A_{x,i}$ and $A_{x,i}[t] := A_{z_e|y}[t]A_{x,i}$, then the augmented Lagrangian of (6) becomes $L_{x,e} = \sum_{t \in Z_e} L_{x,e}^{(i)}[t]$ with each term defined as:

$$L_{x,e}^{(i)}[t] = F_{e,i}(\pi_{y,i}[t]) + G_{e,i}(\pi_{y,i}[t]) + \langle \mu_{y,i}, \bar{A}_{e,i} - \pi_{y,i}[t] \rangle,$$

where $\bar{A}_{e,i} := \left[A^T_{x,i} A^T_{y,i} \right]^T$, $\pi_{y,i}[t] := \left[\pi_{y,i}^{T} \pi_{y,i}^{T} \right]^T$.

Given the definitions, we propose a two-step algorithm to solve (2). The first step is to solve (9) through the following consensus ADMM, $\forall i \in \{V\}$:

$$p_i^{k+1} := \arg \min_{p_i \in \Omega_i} L_c(\{p_i^{k+1}, p_i, \{v_i^k\}, q_k\}),$$

$$\nu_i^{k+1} := \nu_i^k + c(A_{x,i} p_i^{k+1} - q_k),$$

$$q_{k+1} := \arg \min_{q_{k+1} \in \Omega_q} L_c(\{p_i^{k+1} \forall i \in \Omega_i\}, q_{k+1}).$$

Then in the second step (10) with two-block ADMM:

$$\pi_{e,i}^{k+1} := \arg \min_{\pi_{e,i}} L_{e,c}(\pi_{x,i}, \mu_i, \nu_i, \pi_{y,i}),$$

$$\mu_i^{k+1} := \mu_i^k + c(A_{x,i} \pi_{e,i}^{k+1} - \pi_{y,i}^k),$$

$$\nu_i^{k+1} := \arg \min_{\nu_i \in \Omega_q} L_c(\{p_i, \mu_i^{k+1}, \nu_i, p_i \forall i \in \Omega_q\}).$$

where in (11), we use the short-hand notation $\{p_i^{k+1}\} := \{p_i^{k+1} \forall i \in \Omega_i\}$ to denote the primal variables, up to $i-1$ views that are already updated to step $k+1$, and $\{p_i^{k}\} := \{p_i \forall i \in \Omega_i\}$ to denote the rest that are still at step $k$. We define $\{p_i^{k}\} := \{0\} \cup \{p_i \forall i \in \Omega_i\}$; in (11) and (12), the superscript $k$ denotes the step index; each of $\Omega_i, \Omega_q, \Pi_{1,i}, \Pi_{2,i}$ denotes a compound probability simplex. The algorithm starts with (11a), updating each view in succession. Then the augmented variables are updated with (11b) to complete step $k$. After convergence of (11), we run (12) in similar fashion for each view. And this completes the full algorithm.

### B. Incremental Update Form

Intuitively, the consensus-complement form works well in the case where the common information in the observations $\{X\}$ across all views is abundant. However, if the views are almost distinct, where each view is a complement to the others, then the previous form will be inefficient in the sense that learning the common may have negligible benefit. To address this, we design the representation set as $\{Z_e^{(i)}\}_{i=1}$ and propose the incremental update MvIB Lagrangian:

$$\mathcal{L}_{inc} := \sum_{i=1}^{V} \gamma I(X_i; Z_i^{(i)}|\{Z\}_{i=1}) - I(Y; Z_i^{(i)}|\{Z\}_{i=1}).$$

Again, to simplify the above, the incremental form is subject to the following constraints:

- For each view $i \in \{V\}$, the corresponding representation $Z^{(i)}$ only access $X^{(i)}$, so $Y \rightarrow X^{(i)} \rightarrow Z^{(i)} \leftarrow \{Z\}_{i=1}$ forms a Markov chain.

Under these conditions, in each step, we can replace observations of all views $\{X\}$ with the view-specific observation $X^{(i)}$ and rewrite (13) as:

$$\mathcal{L}_{inc} := \sum_{i=1}^{V} \gamma I(X_i; Z_i^{(i)}|\{Z\}_{i=1}) - I(Y; Z_i^{(i)}|\{Z\}_{i=1}).$$

In solving (14), we consider the following algorithm. At the $i$th step, we have:

$$P^{(i)}_{x|z_i,z_{<i}} := \arg \min_{P} \mathcal{L}_{inc}(P, \{P^{(j)}_{x|z_j,z_{<j}}\}_{j=1}^{i-1}),$$

$$p(z^{(i)}_i|z_{<i}) = \sum_{x_i} p(x^{(i)}_i|y) p(z^{(i)}_i|z_{<i}) p(x^{(i)}),$$

where $P^{(i)}_{x|z_i,z_{<i}}$ denotes the tensor form of a conditional probability $p(z^{(i)}_i|x^{(i)}_i,z^{(i-1)}_i$, $\ldots$, $z^{(1)}_i), \forall i \in \{V\}$. The tensor is the primal variable for step $i$ which belongs to a compound simplex $\Omega(i)$. In the algorithm, for each step (15a), we solve it with (11) by setting $V = 1$ and treating the estimators from the previous steps as priors. For example, $p(z^{(2)}_2|x^{(2)},z^{(1)}) = \frac{p(x^{(2)}_2|z^{(1)})p(z^{(2)}_2|z^{(1)})}{p(x^{(2)}|z^{(1)})}$, and $p(x^{(2)}|z^{(1)}) = \sum_y p(x^{(2)}|y)p(y|z^{(1)}).

### III. MAIN RESULTS

We propose two new information-theoretic formulations of MvIB and develop optimal-bound achieving algorithms that are in parallel to existing solvers [14], [15], [26]; our main results are the convergence proofs for the proposed two algorithms. The convergence analysis goes beyond the MvIB and the recent non-convex multi-block ADMM convergence results as we further show that strong convexity on $\{F_i\}_{i=1}^{V}$ is not necessary for proving convergence [24]. This new result connects our analysis to a more general class of functions that can be solved with multi-block non-convex consensus ADMM. For simplicity we denote the collective point at step $k$ as $w^k := (\{p_i^k\}, \{w_i^k\}, q^k), \mathcal{L}_k := \mathcal{L}_c(\{p_i^k\}, \{w_i^k\}, q^k)$ as the function value evaluated with $w^k$ and $\mu_B, \lambda_B$ denote the
smallest and largest singular value of a linear operator $B$ respectively.

**Theorem 1**: Suppose $F_i(p_i)$ is $L_i$-smooth and $M_i$-Lipschitz continuous $\forall i \in [V]$ and $G(q)$ is $\sigma_G$-weakly convex. Further, let $L_c$ be defined as in (9) and solved with the algorithm (11). If the penalty coefficient satisfies $c > \max_{i \in [V]} \left\{ (M_i \lambda_A L_i)/\mu_A A_i^T, \sigma_G \right\}$, then the sequence $\{u^k\}_{k \in \mathbb{N}}$ is finite and bounded. Moreover, $\{u^k\}_{k \in \mathbb{N}}$ converges linearly to a stationary point $w^*$ around a neighborhood such that $L^* \leq L < L^*_c + \delta$ for $|w - w^*| < \epsilon$ where $\delta, \epsilon > 0$.

**proof sketch**: The details of the proof are deferred to the full version [27]. Here we explain the key ideas.

The first step is to construct a sufficient decrease lemma (Lemma 3 in [27]) to assure that the function value $L_c$ decreases from step $k$ to $k+1$ by an amount lower bounded by the positive squared norm $\|u^k - w^{k+1}\|^2$. We decompose $L_c - L_c^{k+1}$ according to each step of the algorithm (11), $\forall i \in [V]$ as follows:

$$L^k_c - L_c^{k+1} = \sum_{i=1}^{V} \left[ L_c \left( \{p_i^{k+1}\}, \{p_i^k\}, \{p_{>i}^k\}, \{\nu^k\}, q^k \right) - L_c \left( \{p_i^{k+1}\}, \{p_i^k\}, \{p_{>i}^k\}, \{\nu^k\}, q^k \right) \right]$$

(16a)

$$\quad + \sum_{i=1}^{V} \left[ L_c \left( \{p_i^{k+1}\}, \{\nu_{>i}^{k+1}\}, \nu_i^k, \{p_{>i}^k\}, q^k \right) - L_c \left( \{p_i^{k+1}\}, \{\nu_{>i}^{k+1}\}, \nu_i^k, \{p_{>i}^k\}, q^k \right) \right]$$

(16b)

$$+ L_c \left( \{p_i^{k+1}\}, \{\nu^k\}, q^k \right) - L_c \left( \{p_i^{k+1}\}, \{\nu^k\}, q^{k+1} \right).$$

(16c)

For each view, each difference in (16a) can be lower bounded by using the convexity of $F_i$:

$$L_c \left( \{p_i^k\} \right) - L_c \left( \{p_i^{k+1}\} \right) \geq \sum_{i=1}^{V} \frac{c}{2} \|A_ip_i^k - A_ip_i^{k+1}\|^2. \quad (17)$$

On the other hand, in (16c), a similar lower bound for $G$ follows from its $\sigma_G$-weak convexity. This results in a negative squared norm $-\sigma_G/2\|q^k - q^{k+1}\|^2$. Nonetheless, by the first-order minimizer conditions (24) and the identity $2(u - v, w - u) = |u - v|^2 - |u - v|^2 - |u - w|^2$, the negative term is balanced by the penalty coefficient $c$ as the corresponding lower bound is (with other variables fixed):

$$L_c \left( q^k \right) - L_c \left( q^{k+1} \right) \geq \frac{c - \sigma_G}{2} \|q^k - q^{k+1}\|^2.$$ 

As for the dual update, (16b) gives a combination of negative norms $-1/c \sum_{i=1}^{V} \|\nu_i^k - \nu_i^{k+1}\|^2$. It turns out that by the first-order minimizer condition of $F_i$ and its smoothness:

$$\nabla F_i(p_i^{k+1}) = -A_i^T \nu_i^{k+1},$$

and that $A_i$ is full-row rank (holds for the complement step similarly):

$$\|\nu_i^k - \nu_i^{k+1}\|^2 \leq \frac{\lambda_i^2 A_i^2}{\mu_i A_i^T} \|p_i^k - p_i^{k+1}\|^2.$$

Then we need the following relation:

$$\|A_ip_i^k - A_ip_i^{k+1}\|^2 \geq \frac{M}{\|p_i^k - p_i^{k+1}\|}, \quad M > 0,$$

which is non-trivial because $A_i$ is full-row rank. To address this, we adopt the sub-minimization path method as in [20], which is applicable since $F_i$ is convex. Observe that (11a) is equivalent to a proximal operator:

$$\Psi_i(\eta) := \arg \min_{p_i \in E_i} \frac{c}{2} \|A_ip_i - \eta\|^2,$$

with $\eta := A_ip_i^k$ at step $k$. Using this technique, we can have the desired result using the Lipschitz continuity of $F_i$ (implied by the $L_i$-smoothness):

$$\|\Psi_i(A_ip_i^k) - \Psi_i(A_ip_i^{k+1})\| = \|p_i^k - p_i^{k+1}\| \leq M_i \|A_ip_i^k - A_ip_i^{k+1}\|.$$

This proves the sufficient decrease lemma and hence the convergence (see Appendix F of [27]). We further prove that the rate of convergence is linear by explicitly showing the Kurdyka-Łojasiewicz (KL) property [19], [28] is satisfied with an exponent $\theta = 1/2$ (Appendix D of [27]). It is known that using the KL inequality, the rate of convergence is characterized into three regions in terms of $\theta$ [19] ($\theta = 1/2$ corresponds to linear rate). The proof for $\theta = 1/2$ is again based on the convexity of $\{F_i\}$ and the weak convexity of $G$ and is referred to Appendix D of [27]. We note that the linear rate holds around a neighborhood of a stationary point $w^* := (p_i^*, \nu_i^*, q^*)$ which aligns with the results in [28].

As a remark, if the minimum element of a probability vector is bounded away from zero by a constant $\xi > 0$, a commonly adopted smoothness condition in density and entropy estimation research [29], the sub-objectives $\{F_i\}_{i=1}^{V}$ and $G$ can be shown to be Lipschitz continuous and smooth. Furthermore, under smoothness conditions, $G$ is a weakly convex function w.r.t. $q$ (Lemma 2 in [27]). From Theorem 1, the consensus-complement algorithm is convergent since the complement step is a special case of the algorithm (11) with $V = 1$ while treating $p(z_i|x^{(i)})$ as an additional prior probability. The incremental algorithm is convergent following the same reason.

IV. Numerical Results

We evaluate the proposed two approaches for two-view, synthetic distributions. In this part, the consensus-complement approach is denoted as Cons-Cmpl while the incremental update approach as Increment.

We simulate a classification task and compare the performance of the two proposed approaches to joint-view/single-view IB solvers [30], which are served as references for the best- and worst-case performance, and the state-of-the-art deep neural network-based method (DeepMvIB) [11], [14], with two layers of 4-neuron, fully connected weights plus ReLU activation for each view. Given (19), we randomly sample 10000 pairs of outcomes $(y, x^{(1)}, x^{(2)})$ as testing data.
Then we run the algorithms, sweeping through a range of $\gamma \in [0.1, 0.7]$ and record the best accuracy from 50 trials per $\gamma$. We use Bayes’ decoder to predict the testing data, where inverse transform sampling is adopted for the cumulative distribution of the decoders to obtain $\hat{y}$ for each pair of $(x^{(1)}, x^{(2)})$. The data-generating distribution is:

$$P(X^{(1)}|Y) = \begin{bmatrix} 0.75 & 0.05 \\ 0.20 & 0.20 \\ 0.05 & 0.75 \end{bmatrix}, \quad P(X^{(2)}|Y) = \begin{bmatrix} 0.85 & 0.15 \\ 0.15 & 0.85 \end{bmatrix},$$

with $P(Y) = [0.5 \ 0.5]^T$. The result is shown in Figure 1a. The dimension of each of $Z_c, Z_c^{(1)}, Z_c^{(2)}$ is 2, and 3 for each of $Z^{(1)}, Z^{(2)}$. Clearly, the two proposed approaches can achieve comparable performance to that of the joint-view IB solver and outperform the deepMvIB over the range of $\gamma$ we simulated. Interestingly, Cons-Cmpl outperforms Increment in the best accuracy. This might be due to the abundance of representation overlap. To better investigate this, we consider a different set of distributions with dimensions of all representations $|Z_c| = |Z_c^{(1)}| = |Z_c^{(2)}| = 3, \forall i \in \{1, 2\}$:

$$p(Y) = \begin{bmatrix} 1/3 & 1/3 & 1/3 \end{bmatrix}^T, \quad P(X^{(1)}|Y) := \begin{bmatrix} 0.90 & 0.20 & 0.20 \\ 0.05 & 0.45 & 0.35 \\ 0.05 & 0.35 & 0.45 \end{bmatrix},$$

$$P(X^{(2)}|Y) := \begin{bmatrix} 0.25 & 0.10 & 0.55 \\ 0.20 & 0.80 & 0.25 \\ 0.55 & 0.10 & 0.20 \end{bmatrix}. \quad (20)$$

Observe that for each view in (20), there is one class ($y_1$ in view 1 an $y_2$ in view 2), that is easy to infer through $X^{(i)}, i \in \{1, 2\}$ while the remaining two are ambiguous. This results in low representation overlap, a consensus is therefore difficult to form. In Figure 1b we examine the components of the relevance rate $I(\{Z\}; Y)$ where the

**V. Conclusion**

In this work, we propose two new information-theoretic formulations of MvIB and develop new optimal bound-achieving algorithms based on non-convex consensus ADMM, which are in parallel to existing solvers. We proposed two algorithms to solve the two forms respectively and prove their convergence and linear rates. Empirically, they achieve comparable performance to joint-view benchmarks and outperform state-of-the-art deep neural networks-based approaches in some synthetic datasets. For future works, we plan to evaluate the two methods on available multiview datasets [32], [33] and generalize the proposed MvIB framework to continuous distributions [34].
REFERENCES

[1] S. Sun, “A survey of multi-view machine learning,” Neural computing and applications, vol. 23, no. 7, pp. 2031–2038, 2013.

[2] Y. Yang and H. Wang, “Multi-view clustering: A survey,” Big Data Mining and Analytics, vol. 1, no. 2, pp. 83–107, 2018.

[3] M. Federici, A. Dutta, P. Forrë, N. Kushman, and Z. Akata, “Learning robust representations via multi-view information bottleneck,” in International Conference on Learning Representations, 2020.

[4] W. Wang, R. Arora, K. Livescu, and J. Bilmes, “On deep multi-view representation learning,” in International conference on machine learning. PMLR, 2015, pp. 1083–1092.

[5] Y. Li, M. Yang, and Z. Zhang, “A survey of multi-view representation learning,” IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 10, pp. 1863–1883, 2019.

[6] K. Zhan, F. Nie, J. Wang, and Y. Yang, “Multiview consensus graph clustering,” IEEE Transactions on Image Processing, vol. 28, no. 3, pp. 1261–1270, 2019.

[7] N. Tishby, F. C. Pereira, and W. Bialek, “The information bottleneck method,” arXiv preprint physics/0004057, 2000.

[8] C. Xu, D. Tao, and C. Xu, “Large-margin multi-view information bottleneck,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 8, pp. 1559–1572, 2014.

[9] Y. Gao, S. Gu, L. Xia, and Y. Fei, “Document clustering with multi-view information bottleneck,” in 2006 International Conference on Computational Intelligence for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and Internet Commerce (CIMCA’06), 2006, pp. 148–148.

[10] S. Hu, Z. Shi, and Y. Ye, “Dmb: Dual-correlated multivariate information bottleneck for multiview clustering,” IEEE Transactions on Cybernetics, pp. 1–15, 2020.

[11] Q. Wang, C. Boudreau, Q. Luo, P.-N. Tan, and J. Zhou, “Deep multi-view information bottleneck,” in Proceedings of the 2019 SIAM International Conference on Data Mining. SIAM, 2019, pp. 37–45.

[12] A. Blum and T. Mitchell, “Combining labeled and unlabeled data with co-training,” in Proceedings of the Eleventh Annual Conference on Computational Learning Theory, ser. COLT ’98. New York, NY, USA: Association for Computing Machinery, 1998, p. 92–100.

[13] Y. U˘gur, I. E. Aguerri, and A. Zaidi, “A generalization of blahut-arimoto algorithm to compute rate-distortion regions of multiterminal source coding under logarithmic loss,” arXiv preprint arXiv:1708.07309, 2017.

[14] T.-H. Huang, A. E. Gamal, and H. E. Gamal, “On the multi-view information bottleneck representation,” arXiv preprint arXiv:2202.02684, 2022.

[15] A. Blahut, “Computation of channel capacity and rate-distortion functions,” IEEE Transactions on Information Theory, vol. 18, no. 5, p. 740–761, 1972.

[16] O. Shamir, S. Sabato, and N. Tishby, “Learning and generalization with the information bottleneck,” Theoretical Computer Science, vol. 411, no. 29-30, pp. 2696–2711, 2010.

[17] E. Schubert and A. Zimek, “ELKI: A large open-source library for data analysis - ELKI release 0.7.5 heidelberg,” CoRR, vol. abs/1902.03616, 2019.

[18] M. Chao, D. Han, and X. Cai, “Convergence of the peaceman-rachford splitting method for a class of nonconvex programs,” Numerical Mathematics: Theory, Methods and Applications, vol. 14, no. 2, pp. 438–460, 2021.

[19] K. Guo, D. R. Han, and T. T. Wu, “Convergence of alternating direction method for minimizing sum of two nonconvex functions with linear constraints,” International Journal of Computer Mathematics, vol. 94, no. 8, pp. 1653–1669, 2017.