The Multi-layer Information Bottleneck Problem

Qianqian Yang*, Pablo Piantanida†, Deniz Gündüz*

*Electrical and Electronic Engineering Department, Imperial College London, London SW7 2AZ, U.K.
† Laboratoire des Signaux et Systèmes, CentraleSupélec-CNRS-Université Paris-Sud, Gif-sur-Yvette, France

This work has been funded in part by the European Research Council (ERC) under Starting Grant BEACON (agreement 677854).

Abstract—The multi-layer information bottleneck (IB) problem, where information is propagated (or successively refined) from layer to layer, is considered. Based on information forwarded by the preceding layer, each stage of the network is required to preserve a certain level of relevance with regards to a specific hidden variable, quantified by the mutual information. The hidden variables and the source can be arbitrarily correlated. The optimal trade-off between rates of relevance and compression (or complexity) is obtained through a single-letter characterization, referred to as the rate-relevance region. Conditions of successive refinability are given. Binary source with BSC hidden variables and binary source with BSC/BEC mixed hidden variables are both proved to be successively refifiable. We further extend our result to Guassian models. A counterexample of successive refifiability is also provided.

I. INTRODUCTION

A fundamental problem in statistical learning is to extract the relevant essence of data from high-dimensional, noisy, salient sources. In supervised learning (e.g., speaker identification in speech recognition), a set of properties or statistical relationships is pre-specified as relevant information of interest (e.g., name, age or gender of the speaker) targeted to be learned from data; while in unsupervised learning, clusters (e.g., name, age or gender of the speaker) are automatically identified by a model. The IB framework characterizes the trade-off between the information rates (or complexity) of the reproduction signal \( X \), and the amount of mutual information it provides about \( Y \). The IB method has been found useful in a wide variety of learning applications, e.g., word clustering [2], image clustering [3], etc. In particular, interesting connections have been recently made between deep learning [4] and the successively refined IB method [5].

Despite the success of the IB method in the machine learning domain, less efforts have been invested in studying it from an information theoretical view. Gilad-Bachrach et al. [6] characterize the optimal trade-off between the rates of information and relevance, and provide a single-letter region. As a matter of fact, the conventional IB problem follows as a special instance of the novel noisy lossy source coding problem [7]. Extension of this information-theoretic framework address the collaborative IB problem by Vera et al. [8], and the distributed biclustering problem by Pichler et al. [9].

In this work, we introduce and investigate the multi-layer IB problem with non-identical hidden variables at each layer. This scenario is highly motivated by deep neural networks (DNN) and the recent work in [5]. Along the propagation of a DNN, each layer compresses its input, which is the output of the preceding layer, to a lower dimensional output, which is forwarded to the next layer. Another scenario may be the hierarchical, multi-layer network, in which information is propagated from higher layers to lower layers sequentially. Users in different layers may be interested in different properties of the original source. The main result of this paper is the full characterization of the rate-relevance region of the multi-layer IB problem. Conditions are provided for successive refifiability in the sense of the existence of codes that asymptotically achieve the rate-relevance function, simultaneously at all the layers. Binary source with BSC hidden variables and binary source with mixed BSC/BEC hidden variables\(^1\) are both proved to successively refifiable. The successive refifiability is also shown for Guassian sources. We further present a counterexample for which successive refifiability no longer holds. It is worth mentioning that the successive refifiability of the IB problem is also investigated in [14], with identical hidden variables.

The rest of the paper is organized as follows. Section II provides the definitions and presents the main result, the achievability and converse proofs of which are provided in the Appendices. The definition and conditions of successive refifiability are shown in Section III. Examples are presented in Section IV. Finally, we conclude the paper in Section V.

\(^1\)BSC hidden variables are obtained by passing the source through a binary symmetric channel, whereas BEC hidden variables are obtained through a binary erasure channel.
Theorem 2. Source $X$ is successively refinable for the L-layer IB problem with relevance constraints $\mu_1, \ldots, \mu_L$ with regards to hidden variables $Y_1, \ldots, Y_L$, if there exist random variables $U_1, \ldots, U_L$, satisfying $U_L \rightarrow \cdots \rightarrow U_1 \rightarrow X \leftrightarrow (Y_1, \ldots, Y_L)$, such that the following conditions hold simultaneously for $l = 1, \ldots, L$:

1) $R_{X \rightarrow Y_l}(\mu_l) = I(X; U_l)$, 
2) $\mu_l \leq I(Y_l; U_l)$.

Proof. Theorem 2 follows directly from Definition 2 and Theorem 1. \hfill \square

IV. EXAMPLES

A. Binary Source with Symmetric Hidden Variables

We consider $\mathcal{X} = \mathcal{Y} = \{0, 1\}$, $l = 1, \ldots, L$. The observable variable $X$ has a Bernoulli distribution $\frac{1}{2}$ (denoted as Bern $(\frac{1}{2})$), and the hidden variables are obtained by passing the source through independent BSCs, i.e., $Y_l = X \oplus N_l$, where $N_l \sim \text{Bern}(p_l)$. $0 \leq p_l \leq \frac{1}{2}$, is independent of $X$, and $\oplus$ denotes modulo-2 addition.

We first derive the rate-relevance function $R_{X \rightarrow Y_l}(\mu_l)$. Denote by $U_l$ any random variable for which $I(Y_l; U_l) \geq \mu_l$, $U_l \rightarrow X \leftrightarrow Y_l$. We have the following inequality:

$$\mu_l \leq H(Y_l) - H(X \oplus N_l | U_l) \leq 1 - H_b(p_l + H_b^{-1}(H(X | U_l))) = 1 - H_b(p_l + H_b^{-1}(1 - I(Y_l; U_l)))$$

where operation $*$ is defined as $a * b = a(1 - b) + b(1 - a)$, $H_b(\cdot)$ is the binary entropy function, defined as $H_b(p) = p \log 1/p + (1 - p) \log 1/(1 - p)$, and $H_b^{-1}(\cdot)$ is the inverse of the binary entropy function $H_b(p)$ with $p \in [0, 0.5]$.

(6) follows from Mrs. Gerber’s Lemma and the fact that $H(Y_l) = 1$. From (6), we obtain $I(X; U_l) \geq 1 - H_b\left(\frac{H_b^{-1}(1-(\mu_l)-p_l)}{1-2p_l}\right)$. Note by letting $U_l^\ast = X \oplus M_l$, where $M_l$ is independent of $X$ and $M_l \sim \text{Bern}\left(\frac{H_b^{-1}(1-(\mu_l)-p_l)}{1-2p_l}\right)$, we have $I(Y_l; U_l^\ast) = \mu_l$ and $I(X; U_l^\ast) = 1 - H_b\left(\frac{H_b^{-1}(1-(\mu_l)-p_l)}{1-2p_l}\right)$. We can conclude that $R_{X \rightarrow Y_l}(\mu_l) = 1 - H_b\left(\frac{H_b^{-1}(1-(\mu_l)-p_l)}{1-2p_l}\right)$ and $U_l^\ast$ given above is a rate-relevance function achieving auxiliary random variable.

Lemma 1. Binary sources as described above are always successively refinable for the L-layer IB problem if $R_{X \rightarrow Y_l}(\mu_l) \geq \cdots \geq R_{X \rightarrow Y_L}(\mu_L)$ and $\mu_l \leq 1 - H_b(p_l)$, for $l = 1, \ldots, L$.

Proof. Since $R_{X \rightarrow Y_l}(\mu_l) \geq \cdots \geq R_{X \rightarrow Y_L}(\mu_L)$, we can find binary variables $M_1, \ldots, M_L$, independent of each other and $X$, such that $M_l = \cdots = M_l \sim \text{Bern}(H_b^{-1}(1 - R_{X \rightarrow Y_l}(\mu_l)))$ for $l = 1, \ldots, L$. By choosing auxiliary random variables: $U_l = X \oplus M_1 \oplus \cdots \oplus M_l$, we have $I(X; U_l) = R_{X \rightarrow Y_l}(\mu_l)$ and $I(Y_l; U_l^\ast) = \mu_l$, for $l = 1, \ldots, L$, and $U_l \rightarrow \cdots \rightarrow U_1 \rightarrow X \leftrightarrow (Y_1, \ldots, Y_L)$. Together with Theorem 2, this conclude the proof of Lemma 1. \hfill \square
Since \( \mu \) for their setting. We first derive the rate-relevance function of Lemma 1. The proof follows the same arguments as in the proof of Section IV-A, except for the \( \mu \) variables as described above is never successively refinable. We consider a two-layer IB problem, i.e., \( L = 2 \). The joint distribution of \((X, Y_1, Y_2)\) is illustrated in Fig. 2, where \( X \) is a binary random variable of distribution Bernoulli \( \frac{1}{2} \) as in the previous example, but \( Y_1 \) is the output of a BEC with erasure probability \( \epsilon \) \((\epsilon \in [0,1/2])\) when \( X \) is the input, and \( Y_2 \) is the output of a (BSC) with crossover probability \( p \), \( p \in [0,1/2] \). A similar example can be found in [15] where the optimality of proposed coding scheme does not always hold for our setting. We first derive the rate-relevance function \( R_{X \rightarrow Y_1}(\mu_1) \). Denote by \( U_l \) any random variable such that \( I(Y_l; U_l) \geq \mu_l, U_l \rightarrow X \rightarrow Y_l \). We have the following sequence of inequalities:

\[
\mu_l \leq I(Y_l; U_l) = h(Y_l) - h(X + N_l|U_l) = h(Y_l) - h(X + N_l) + \frac{1}{2} \log \left( \frac{2\pi e \sigma_{N_l}^2}{\sigma_{N_l}^2} \right) = \frac{1}{2} \log \left( \frac{2\pi e \sigma_{N_l}^2}{\sigma_{N_l}^2} \right).
\]

where (8d) follows from the conditional Entropy Power Inequality (EPI) (Section 2.2 in [16]). We can also obtain an outer bound on \( R_{X \rightarrow Y_1}(\mu) \):

\[
R_{X \rightarrow Y_1}(\mu) \geq \frac{2\mu_2 \sigma_{2}^2}{2\mu_2 \sigma_{2}^2 + \sigma_{N_2}^2}
\]

by setting \( U_2^* = X + P_2 \), \( P_2 \sim N(0, \sigma_{P_2}^2) \), \( P_2 \leq L \), where \( \sigma_{P_2}^2 \) is given by:

\[
\sigma_{P_2}^2 = \frac{2\pi e \sigma_{x}^2 - 2^{2R_{X \rightarrow Y_1}(\mu)}}{2\pi e (2^{2R_{X \rightarrow Y_1}(\mu)} - 1)}.
\]

Lemma 3. Gaussian sources as described above are always successively refinable for the L-layer IB problem if \( R_{X \rightarrow Y_l}(\mu_l) \geq R_{X \rightarrow Y_{l+1}}(\mu_{l+1}) \) and \( \mu_L \leq I(X; Y_L) \).

D. Counterexample on successive refinability

In this section, we show that the multi-layer IB problem is not always successively refinable. We consider a two-layer IB problem, i.e., \( L = 2 \). Let \( X = (X_1, X_2) \), where \( X_1 \) and \( X_2 \) are two independent discrete random variables, and we have \( Y_1 = X_1 \) and \( Y_2 = X_2 \). We first derive the rate-relevance function despite BEC hidden variable.
function $R_{X \to Y_1}(\mu_1)$. Denote by $U_1$ any random variable such that $I(Y_1; U_1) \geq \mu_1$, $U_1 \sim X - Y_1$. We have:

$$I(X; U_1) = I(X_1, X_2; U_1)$$

(11a)

$$\geq I(X_1; U_1)$$

(11b)

$$\geq \mu_1.$$  

(11c)

By setting $U_1^*$ as

$$U_1^* = \begin{cases} X_1, & \text{with probability } \frac{\mu_1}{\mu(X_1)}, \\ 0, & \text{with probability } 1 - \frac{\mu_1}{\mu(X_1)}, \end{cases}$$

(12)

we have $I(Y_1; U_1^*) = I(X_1; U_1^*) = \mu_1$, and $I(X; U_1^*) = \mu_1$, which achieves the lower bound shown in (11). We can conclude that $R_{X \to Y_1}(\mu_1) = \mu_1$, and any rate-relevance function achieving random variable $U_1^*$ should satisfy $I(X_1; U_1^*) = 0$, since $I(X; U_1^*) = I(X_1^*; U_1^*) + I(X_2; U_1^* | X_1) = \mu_1$ and $I(X_1; U_1^*) = \mu_1$. Similarly, we can conclude that $R_{X \to Y_2}(\mu_2) = \mu_2$, and any rate-relevance function achieving random variable $U_2^*$ should satisfy $I(X_1; U_2^*) = 0$.

**Lemma 4.** Source $X$ with hidden variables $Y_1$ and $Y_2$ as described above is not successively refinable for the two-layer IB problem.

**Proof.** For any rate-relevance function achieving random variables $U_1^*$ and $U_2^*$, we have

$$I(U_2^*; X | U_1^*) = I(U_2^*; X_1, X_2 | U_1^*)$$

(13a)

$$= I(U_2^*; X_2 | U_1^*) + I(U_2^*; X_1 | U_1^*, X_2)$$

(13b)

$$\geq I(U_2^*; X_2 | U_1^*)$$

(13c)

$$= I(U_1^*; X_2) - I(U_1^*; X_2)$$

(13d)

$$= I(U_1^*; X_2) + I(U_1^*; X_2 | U_2^*)$$

(13e)

$$\geq I(U_2^*; X_2) \geq \mu_2,$$

(13f)

where (13e) is due to $I(U_1^*; X_2) = 0$, which follows from

$$I(X_1; U_2^*)$$

(14a)

$$= I(X_2; U_1^*) - I(U_1^*; X_2)$$

(14b)

$$= I(X_2; X_1) + I(X_2; U_1^* | X_1) - I(U_1^*; X_2)$$

(14c)

$$= -I(U_1^*; U_2^*).$$

(14d)

If $\mu_2 > 0$, $I(U_2^*; X | U_1^*) > 0$, which implies $U_2^*, U_1^*$ and $X$ cannot form a Markov chain for any rate-relevance function achieving random variables $U_1^*$ and $U_2^*$. With Theorem 2, we have proven Lemma 4.

**V. CONCLUSION**

The multi-layer IB problem with non-identical relevant variables was investigated. A single-letter expression of the rate-relevance region was given. The definition and conditions of successive refinability were presented, which was further investigated for the binary sources and Gaussian sources. A counterexample of successive refinability was also proposed.

**REFERENCES**

[1] N. Tishby, F. C. Pereira, and W. Bialek, “The information bottleneck method,” in Proc. of Annu. Allerton Conf. Commun., Contr. Comput., Monticello, IL, Sep. 1999, pp. 368–377.

[2] X. S. N. Slonim and N. Tishby, “Document clustering using word clusters via the information bottleneck method,” in Proc. of Annu. Conf. Research and Development in inform. retrieval, Athens, Greece, 2000, pp. 208–215.

[3] J. Goldberger, H. Greenspan, and S. Gordon, “Unsupervised image clustering using the information bottleneck method,” in Joint Pattern Recog. Symp., Zurich, Switzerland, Sep. 2002, pp. 158–165.

[4] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, May 2015.

[5] N. Tishby and N. Zaslavsky, “Deep learning and the information bottleneck principle,” in 2015 IEEE Information Theory Workshop (ITW), April 2015, pp. 1–5.

[6] R. Gilad-Bachrach, A. Navot, and N. Tishby, “An information theoretic tradeoff between complexity and accuracy,” in Proc. of Conf. Learning Theory (COLT), Washington, Aug. 2003, pp. 595–609.

[7] R. Dobrushin and B. Tsypak, “Information transmission with additional noise,” IRE Transactions on Information Theory, vol. 8, no. 5, pp. 993–1004, September 1962.

[8] M. Vera, L. R. Vega, and P. Piantanida, “Collaborative representation and lossless one-helper problem,” IEEE Trans. Inform. Theory, vol. 60, no. 2, 2014.

[9] E. Tuncel, “Capacity/storage tradeoff in high-dimensional identification systems,” IEEE Trans. Inform. Theory, no. 5, 2009.

[10] E. Tuncel and D. Gunduz, “Identification and lossy reconstruction in noisy databases,” IEEE Trans. Inform. Theory, no. 2, 2014.

[11] J. G. F. Willems, T. Kalker and J.-P. Linnartz, “Successive refinement for hypothesis testing and lossless one-helper problem,” IEEE Trans. Inform. Theory, vol. 54, no. 10, pp. 4666–4681, 2008.

[12] J. Villard and P. Piantanida, “Secure multiterminal source coding with side information at the eavesdropper,” IEEE Trans. Inform. Theory, vol. 59, no. 6, pp. 3668–3692, 2013.

[13] A. D. Gamal and Y. H. Kim, Network Information Theory. New York, NY, USA: Cambridge University Press, 2012.

**APPENDIX A**

**ACHIVABILITY OF THEOREM 1**

Consider first the direct part, i.e., every tuple $(R_1, \ldots, R_L, \mu_1, \ldots, \mu_L) \in \mathcal{R}$ is achievable.

**Code generation.** Fix a conditional probability mass function (pmf) $p(u_1, \ldots, u_L | x)$ such that $\mu_l \leq I(Y_l; U_1, \ldots, U_L)$, for $l = 1, \ldots, L$. First randomly generate $2^{nR_l}$ sequences $u_l^n(i_l)$, $i_l = [1 : 2^{nR_l}]$, independent and identically distributed (i.i.d) according to $p(u_l)$; then for each $u_l^n(i_l)$ randomly generate $2^{(R_{L+1} - R_l)}$ sequences $u_{L+l-1}^n(i_l, i_{L+l-1})$, $i_{L+l-1} = [1 : 2^{n(R_{L+1} - R_l)}]$, conditionally i.i.d. according to $p(u_{L+l-1} | u_l)$, and continue in the same manner, for each $u_{L+j}^n(i_L, \ldots, i_{L+j})$ randomly generate $2^{(R_{L+j} - R_{L+1})}$ sequences $u_{L+j}^n(i_L, \ldots, i_{L+j})$, $i_{L+j} = [1 : 2^{n(R_{L+j} - R_{L+1})}]$, conditionally i.i.d. according to $p(u_{L+j} | u_{L+j-1}, \ldots, u_j)$, for $j = [2 : L]$.

**Encoding and Decoding** After observing $x^n$, the first encoder finds an index tuple $(i_1, \ldots, i_L)$ such that $(x^n, u_l^n(i_L, \ldots, i_1), u_{L+1}^n(i_2, \ldots, i_2), u_{L+2}^n(i_2, \ldots, i_{L+1}), \ldots, u_L^n(i_L))$ is in the set $T_n(X, U_1, \ldots, U_L)$, which is the set of $\epsilon$ jointly typical $n$
vectors of random variables $X, U_1, \ldots, U_L$. If more than one such tuple exist, any one of them is selected. If no such tuple exists, we call it an error, and set $(i_1, \ldots, i_L) = (1, \ldots, 1)$. Then the $j$th encoder outputs $(i_j, \ldots, i_L)$, for $j = 1, \ldots, L$, and sends to the $j + 1$ encoder, if $j < L$, the index tuple $(i_j, \ldots, i_L)$ at a total rate of $R_j$. Given the index tuple $(i_j, \ldots, i_L)$, the $j$th decoder declares $u^n(i_L, \ldots, i_j)$ as its output, for $j = 1, \ldots, L$.

Relevance. First, we note that if there is no error in the encoding step, i.e., an index tuple $(i_1, \ldots, i_L)$ such that $(x^n, u^n_1(i_L), \ldots, i_1), u^n_2(i_L), \ldots, u^n_L(i_L)) \in T^n(X, U_1, \ldots, U_2)$ is found, then the relevance condition

$$\mu_l = I(Y; U_1, \ldots, U_L) \geq 0$$

for $l = 1, \ldots, L$, is satisfied by the definition of $T^n(X, U_1, \ldots, U_L)$ and the Markov lemma. Then we focus on the analysis of the probability of error, i.e., the probability that such an index tuple cannot be found in the encoding step.

An error occurs if one of the following events happens:

$$E_0 : x^n \notin T^n(X);$$

$$E_1 : x^n \in T^n(X), (x^n, u^n_L(i_L)) \notin T^n(X, U_L), \quad \text{for all } l_L = 1, \ldots, 2^{R_L};$$

$$E_l : (x^n, u^n_L-1+2(i_L-1+L), \ldots, u^n_L(i_L)) \in T^n(X, U_L-1+2, \ldots, U_L),$$

$$\notin T^n(X, U_L-1+1, \ldots, U_L), \quad \text{for all } l_L-1+1 = 1, \ldots, 2^{R_L-1+L-1}}$$

for $l = 2, \ldots, L$. It is clear that $P(E_0) \to 0$ as $n \to \infty$. Based on the properties of typical sequences:

$$\forall l \rightarrow \infty 0, \text{ if } R_L \geq I(X; U_L);$$

$$\forall l \rightarrow \infty 0, \text{ if } R_L-1+L-2 \geq I(X; U_L-1+1|U_L-1+2, \ldots, U_L),$$

for $l = 1, \ldots, L$.

APPENDIX B
CONVERSE OF THEOREM 1

Next, we prove that every achievable tuple $(R_1, \ldots, R_L, \mu_1, \ldots, \mu_L)$ must belong to $R$. The system achieving $(R_1, \ldots, R_L, \mu_1, \ldots, \mu_L)$ is specified by the encoding functions $\{f_1, \ldots, f_L\}$, i.e.,

$$f_1 : x^n \to Z_1;$$

$$f_l : Z_{l-1} \to Z_l, \quad l = 2, \ldots, L,$$

such that

$$R_L \geq \frac{1}{n} \log |Z_L|;$$

$$\mu_l \leq \frac{1}{n} I(Y^n_l; Z_l), \text{for } l = 1, \ldots, L.$$