An Information Bottleneck Problem with Rényi's Entropy

Jian-Jia Weng, Fady Alajaji, and Tamás Linder
Department of Mathematics and Statistics
Queen's University
Kingston, ON K7L 3N6, Canada
Emails: jianjia.weng@queensu.ca, {fady, linder}@mast.queensu.ca

***Abstract*—This paper considers an information bottleneck problem with the objective of obtaining a most informative representation of a hidden feature subject to a Rényi entropy complexity constraint. The optimal bottleneck trade-off between relevance (measured via Shannon's mutual information) and Rényi entropy cost is defined and an iterative algorithm for finding approximate solutions is provided. We also derive an operational characterization for the optimal trade-off by demonstrating that the optimal Rényi entropy-relevance trade-off is achievable by a simple time-sharing scalar coding scheme and that no coding scheme can provide better performance. Two examples where the optimal Shannon entropy-relevance trade-off can be exactly determined are further given.**

***Index Terms*—Information bottleneck, entropy-constrained optimization, Rényi entropy, coding theorem, time-sharing.**

I. Introduction

In the past decade, the optimization of information measures such as entropy, cross-entropy, and mutual information has been widely and successfully adopted in machine learning algorithms [1]–[4] and transmission systems [5]–[10]. In particular, numerous results are related to the so-called information bottleneck (IB) method [11] whose objective is to extract from observed data the maximal relevant information about a hidden variable subject to a mutual information complexity constraint. Significant efforts are still devoted to studying the IB method and its variants, including its variational approximation [12], its application to analyze the effectiveness of deep neural networks [13], and its generalizations [14], [15]. This paper studies a constrained information optimization problem that is close to a deterministic variant of the IB method, the so-called deterministic IB (DIB) method [16].

The DIB method was proposed to capture the notion of compression in the IB method. In view of Shannon's lossless source coding theorem [17], the authors of [16] suggested replacing the mutual information complexity constraint in the IB method with a Shannon entropy complexity constraint to take into account the cost of storing the extracted information. Here, the storage cost is assumed to vary linearly with the length of the codeword that represents the extracted information. Motivated by Campbell's source coding theorem [18], an extension of Shannon's result, we consider a Rényi entropy [19] constraint to associate the codeword length with an exponential storage cost (which is more appropriate for applications where the processing cost of decoding and buffer overflow problems caused by long codewords are significant). As Shannon entropy is a limiting case of Rényi entropy [19] (as the order goes to 1), this consideration extends the DIB method in a certain sense.

Our Rényi extension may provide a way to improve the performance of the DIB method in machine learning tasks [2] or other applications such as channel quantization [7] and relay transmission [9]. In addition, the use of Rényi entropy generated interest in its own right in information theory and it has played an important role in a variety of studies, including generalized source-coding cut-off rates [20], [21], quantization [22], encoding tasks [23], guessing [24], information combining [25], generative deep networks [26], etc. It is thus of interest to examine the role of Rényi entropy in bottleneck problems.

We now formulate our bottleneck problem. Consider a pair of discrete random variables (Y, X) with joint probability distribution $P_{Y,X}$ over a finite alphabet $\mathcal{Y} \times \mathcal{X}$ and another (representation) random variable $W \in \mathcal{W}$, which form a Markov chain: $Y \multimap X \multimap W$, i.e., $P_{Y,X,W} = P_{Y,X}P_{W|X}$. The objective is to determine the maximal amount of relevant information about Y that can be extracted from X and conveyed in W subject to a Rényi entropy constraint $H_\alpha(W) \le \gamma$ for $\alpha \in (0,1)$,[1] where $H_\alpha(\cdot)$ denotes the Rényi entropy of order α [19]. We call this problem an information-Rényi entropy bottleneck problem and study the function $F_{\alpha,M} : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ in connection with the problem:

$$F_{\alpha,M}(\gamma) := \max_{\substack{P_{W|X}: Y \multimap X \multimap W \\ |\mathcal{W}| \le M, H_\alpha(W) \le \gamma}} I(Y;W) \tag{1}$$

where $\gamma \ge 0$, M is a finite positive integer, and $|\mathcal{W}|$ denotes the cardinality of $\mathcal{W}$. To simplify our presentation, we let $F_{1,M}$ denote (1) with the constraint $H_\alpha(W) \le \gamma$ replaced by a Shannon entropy constraint $H(W) \le \gamma$, which is a constrained optimization formulation of the DIB problem in [16].

Define $\bar{I}_{\alpha,M}$ as the upper concave envelope of $F_{\alpha,M}$ for $\alpha \in (0,1]$. By [27, Corollary 17.1.5], we have that

This work was supported in part by NSERC of Canada.

[1] Although $H_\alpha(W)$ is defined for $\alpha \in (0,1) \cup (1,\infty)$, we only consider $\alpha \in (0,1)$ since $H_\alpha(W)$ with such α has an operational meaning in the application of Campbell's source coding theorem [18]. Moreover, $H_\alpha(W)$ is a concave and continuous function in P_W in this range of α; these two properties allow us to replace the supremum with maximum in (1) and (2) and develop efficient numerical methods to compute (2).

$$\bar{I}_{\alpha,M}(\gamma) = \max \sum_{i=1}^{2} \lambda_i F_{\alpha,M}(\gamma_i) \tag{2}$$

where the maximum is taken over all convex combinations of two pairs $(\gamma_i, F_{\alpha,M}(\gamma_i))$, $i = 1, 2$, such that $\sum_{i=1}^{2} \lambda_i \gamma_i = \gamma$. In this paper, we show that the function $\bar{I}_{\alpha,M}$ describes the optimal Rényi entropy-relevance trade-off for our bottleneck problem. Specifically, we establish an operational characterization of $\bar{I}_{\alpha,M}$ for any $\alpha \in (0,1)$ and finite M. This finding is analogous to the IB coding theorem [28] and clarifies the operational meaning of the information quantities in (1). We note that a closed-form expression of $\bar{I}_{\alpha,M}$ is only available for very special cases. We also derive bounds for $\bar{I}_{\alpha,M}$ and provide numerical methods to compute $\bar{I}_{\alpha,M}$.

The rest of this paper is organized as follows. In Section II, the system model is given and the IB and DIB results are reviewed. In Section III, bounds and properties for $\bar{I}_{\alpha,M}$ are established, followed by the derivation of an operational characterization. In Section IV, methods to compute $\bar{I}_{\alpha,M}$ and two examples are given. Conclusions are drawn in Section V.

II. Preliminaries

Given discrete random variables A_i on a common alphabet $\mathcal{A}$, $i = 1, 2, \ldots, n$, we let $A^n = (A_1, A_2, \ldots, A_n)$. Throughout this paper, we assume the following system model when developing information-theoretic results. The system input is a sequence of independent and identically distributed (i.i.d.) random variables $X_i \in \mathcal{X}$, $i = 1, 2, \ldots, n$, which is correlated with another sequence of i.i.d. hidden variables $Y_i \in \mathcal{Y}$. The joint probability distribution is given by $P_{Y^n,X^n}(y^n, x^n) = \prod_{i=1}^{n} P_{Y,X}(y_i, x_i)$ for some joint distribution $P_{Y,X}$. The goal is to transform X^n into the most informative W^n subject to a complexity constraint. In [28], a complete coding theorem is derived for the IB method, but there seems to have no corresponding result for the DIB method. One of our objectives is to fill this gap under a more general framework.

We begin with the definition of Rényi entropy [19] of order $\alpha \in (0,1) \cup (1,\infty)$ of a random variable W with alphabet $\mathcal{W}$ and distribution P_W:

$$H_\alpha(W) := \frac{1}{1-\alpha} \log_2 \left(\sum_{w \in \mathcal{W}} P_W^\alpha(w) \right).$$

Some properties of $H_\alpha(W)$ for $\alpha \in (0,1)$ are summarized [29]:

- $H(W) = H_1(W) := \lim_{\alpha \to 1} H_\alpha(W)$;
- $0 \le H_\alpha(W) \le \log_2 |\mathcal{W}|$;
- $H_\alpha(W)$ is non-increasing in α;
- $H_\alpha(W)$ is concave in P_W for $\alpha \in [0,1]$;
- $H_\alpha(W)$ is continuous in α at any $\alpha \in (0,1)$ and finite $\mathcal{W}$.

In this paper, all information quantities are measured in bits. We next review some IB and DIB results related to our work.

A. The IB Method [11] and A Complete Coding Theorem [28]

In short, the IB method aims to extract from an observation X the most relevant representation W about a hidden variable Y under a complexity constraint. For any fixed $P_{Y,X}$, such an objective is associated with the function $\bar{I}_{\text{IB}} : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ defined as

$$\bar{I}_{\text{IB}}(r) := \max_{\substack{P_{W|X}: Y \multimap X \multimap W \\ I(X;W) \le r}} I(Y;W) \tag{3}$$

which is non-decreasing, continuous, and concave in r [30]. One can also set $|\mathcal{W}| = |\mathcal{X}| + 1$ without changing $\bar{I}_{\text{IB}}(\gamma)$. An operational interpretation of the $\bar{I}_{\text{IB}}$ function is described next.

Definition 1. *A $(2^{nr_n}, n)$ IB code consists of an encoding function $\mathcal{E}_n : \mathcal{X}^n \to \{1, 2, \ldots, 2^{nr_n}\}$ and a decoding function $\mathcal{D}_n : \{1, 2, \ldots, 2^{nr_n}\} \to \mathcal{W}^n$.*

The average symbol-wise mutual information between Y^n and W^n associated with the above code is computed as $\eta_n := \frac{1}{n} \sum_{i=1}^{n} I(Y_i; W_i)$, where $W^n = \mathcal{D}_n(\mathcal{E}_n(X^n))$.

Definition 2. *A complexity-relevance pair (r, η) is said to be achievable if there exists a sequence of IB codes $\{\mathcal{E}_n, \mathcal{D}_n\}$ such that $\limsup_{n\to\infty} r_n \le r$ and $\liminf_{n\to\infty} \eta_n \ge \eta$. The achievable IB region $\mathcal{R}_{\text{IB}} \subset \mathbb{R}^2_{\ge 0}$ is defined as the closure of all achievable pairs.*

Letting $I_{\text{IB}}(r) = \max(\eta \colon (r, \eta) \in \mathcal{R}_{\text{IB}})$, Gilad *et al.* proved the following proposition in [28]. Here, we rephrase their original statement in [28, Theorem 2] in terms of the functions I_{IB} and $\bar{I}_{\text{IB}}$, which is more convenient for our use.

Proposition 1 ([28]). *$I_{\text{IB}}(r) = \bar{I}_{\text{IB}}(r)$ for $r \ge 0$.*

B. The DIB Method [16]

As mentioned before, the DIB method borrows the source coding idea from information theory and intends to minimize the representation cost for W. Specifically, the DIB method is associated with the following optimization problem

$$\max_{P_{W|X}: Y \multimap X \multimap W} \big[\beta I(Y;W) - H(W)\big] \tag{4}$$

where $\beta \in [0, \infty)$ controls the trade-off between $I(Y;W)$ and $H(W)$. When $\mathcal{W}$ is finite, (4) is a convex optimization problem since, as one can verify, $H(W)$ is concave in $P_{W|X}$ and $I(Y;W)$ is convex in $P_{W|X}$ for any fixed $P_{Y,X}$ [31]. Moreover, the feasible set of all valid conditional probability distributions $P_{W|X}$ is compact and convex. Therefore, we know that the maximum value of the objective function is attainable by some extreme point of the feasible set [27, Corollary 32.3.1]. In our case, the extreme points are conditional distributions $P_{W|X}$ where $P_{W|X}(w|x)$ is either 0 or 1 for any $x \in \mathcal{X}$ and $w \in \mathcal{W}$, i.e., the optimizer of (4) is deterministic.[2]

We remark that the objective function to be maximized in (4) can be viewed as the Lagrangian corresponding to the constrained optimization problem in (1) with $H_\alpha(W)$ replaced by $H(W)$, where β denotes the Lagrange multiplier. From this viewpoint, the DIB problem is a special case of our information-Rényi entropy bottleneck problem. In [16], the DIB method was empirically shown to attain a relevance level similar to that of the IB method with smaller entropy $H(W)$. However, unlike the IB method, the DIB method does not have an operational meaning in terms of a coding theorem.

III. The Rényi Entropy-Relevance Trade-off and Operational Characterization

In Section I, we defined $\bar{I}_{\alpha,M}$ as the upper concave function of $F_{\alpha,M}$. In this section, we investigate the properties of $\bar{I}_{\alpha,M}$,

[2] [16] adopted another more complex approach to solve (4) and concluded the same deterministic structure for the maximizing distribution.

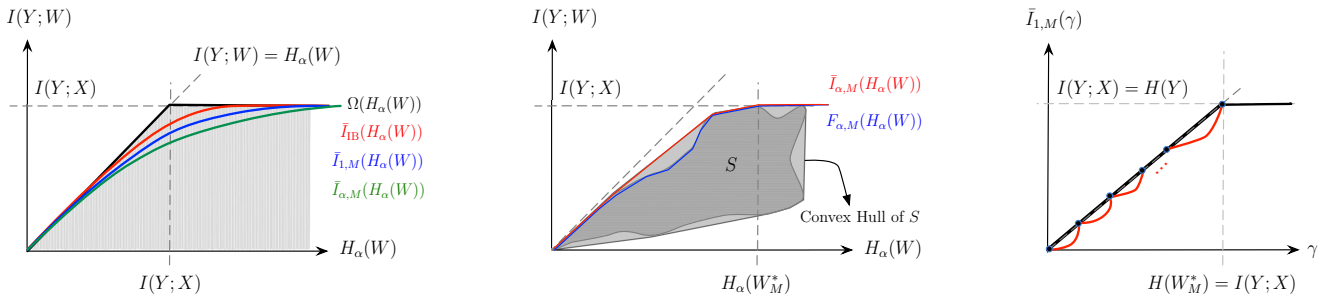

Fig. 1: (a) All feasible Rényi entropy-relevance pairs must lie in the shaded region, (b) Typical shapes of the set $S = \{(H_\alpha(W), I(Y;W)) : P_{W|X} \in \mathcal{P}(\mathcal{W}|\mathcal{X})\}$ and the convex hull of S, and (c) The plot of $\bar{I}_{1,M}(\gamma)$ for the case $Y = f(X)$, where the black dots are given by deterministic $P_{W|X}$'s and the curve is obtained by the DIB method [32].

which will be used in deriving an operational characterization and to develop efficient numerical methods to estimate $\bar{I}_{\alpha,M}$. Below, we assume that $P_{Y,X} = P_X P_{Y|X}$, $M < \infty$, and $\alpha \in (0,1)$ are fixed, unless otherwise stated.

A. Bounds on $\bar{I}_{\alpha,M}(\gamma)$ and Properties of $\bar{I}_{\alpha,M}(\gamma)$

We first recall under the Markov chain constraint $Y \multimap X \multimap W$, we have

$$I(Y;W) \le I(Y;X) \tag{5a}$$

and

$$I(Y;W) \le H(W) \le H_\alpha(W) \le \log_2 |\mathcal{W}| \tag{5b}$$

where (5a) is due to the data-processing inequality [33] and (5b) holds since $H_\alpha(W)$ is non-increasing in α. Based on (5), we construct a concave function Ω given by $\Omega(\gamma) = \gamma$ for $\gamma \in [0, I(Y;X)]$ and $\Omega(\gamma) = I(Y;X)$ for $\gamma > I(Y;X)$. Using the definition of $\bar{I}_{\alpha,M}$ and the concavity of Ω, we immediately obtain the following result.

Lemma 1. *$\bar{I}_{\alpha,M}(\gamma) \le \Omega(\gamma) \le \log_2 M$ for any $\gamma \ge 0$.*

Moreover, we can relate $\bar{I}_{1,M}$ to the IB function $\bar{I}_{\mathrm{IB}}$ in (3):

Lemma 2. *$\bar{I}_{1,M}(\gamma) \le \bar{I}_{\mathrm{IB}}(\gamma)$ for any $\gamma \ge 0$.*

Proof: For any $\gamma \ge 0$, since the constraint $H(W) \le \gamma$ implies that $I(X;W) \le \gamma$, we have that $\bar{I}_{\mathrm{IB}}(\gamma) \ge F_{\alpha,M}(\gamma)$. By invoking the concavity of $\bar{I}_{\mathrm{IB}}$ and the definition of $\bar{I}_{1,M}$, we immediately obtain the desired inequality. ∎

Using a similar argument, one can also deduce the following two statements whose proofs we omit for simplicity.

Lemma 3. *$\bar{I}_{\alpha,M}(\gamma) \le \bar{I}_{1,M}(\gamma)$ for any $\gamma \ge 0$.*

Lemma 4. *$\bar{I}_{\alpha,M}(\gamma) \le \bar{I}_{\alpha,M+1}(\gamma)$ for any $\gamma \ge 0$.*

A visualization of these bounds is given in Fig. 1(a). In the following, we present some properties of $\bar{I}_{\alpha,M}$. First, $\bar{I}_{\alpha,M}(\gamma)$ is non-decreasing in γ since $F_{\alpha,M}(\gamma)$ is non-decreasing. Moreover, $\bar{I}_{\alpha,M}(\gamma)$ is a continuous function in γ. To see this, we note that $\bar{I}_{\alpha,M}(\gamma)$ is continuous for $\gamma > 0$ due to the concavity [27, Theorem 10.2]. One can directly verify that $\lim_{\gamma\to 0^+} \bar{I}_{\alpha,M}(\gamma) = \bar{I}_{\alpha,M}(0)$ to complete the proof.

Furthermore, let $\mathcal{P}_M(\mathcal{W}|\mathcal{X}) = \{P_{W|X} : |\mathcal{W}| = M\}$. Due to the compactness of $\mathcal{P}_M(\mathcal{W}|\mathcal{X})$, the image S of $\mathcal{P}_M(\mathcal{W}|\mathcal{X})$ under the continuous mapping $P_{W|X} \mapsto (H_\alpha(W), I(Y;W))$ is also compact, which guarantees the existence of a $P_{W_M^*|X} \in \mathcal{P}_M(\mathcal{W}|\mathcal{X})$ that satisfies $I(Y;W_M^*) = \max(\eta : (\gamma,\eta) \in S) := J$ and $H_\alpha(W_M^*) = \min(\gamma : (\gamma, J) \in S)$. This result indicates that the graph of $\bar{I}_{\alpha,M}$ is flat with $\bar{I}_{\alpha,M}(\gamma) = J$ for $\gamma \ge H_\alpha(W_M^*)$. When $M \ge |\mathcal{X}|$, we have that $J = I(Y;X)$; a visualization of this case is given in Fig. 1(b). Note that the W_M^* does not necessarily equal to X (in distribution).

We next establish an operational characterization of $\bar{I}_{\alpha,M}$ by associating each pair $(\gamma, \bar{I}_{\alpha,M}(\gamma))$ with an optimal operational scheme. This result implies that $\bar{I}_{\alpha,M}(\gamma)$ is the maximal achievable relevance of a system when the representation cost is at most γ (bits/input symbol).

B. An Operational Characterization of $\bar{I}_{\alpha,M}(\gamma)$

We consider the system model in Section II. The goal of an operational scheme here is to transform X^n into W^n under a Rényi entropy constraint where $\alpha \in (0,1)$ while preserving the relevant information about Y^n as much as possible. For this purpose, we define Rényi entropy-relevance codes below.

Definition 3. *A length-n Rényi entropy-relevance code is a mapping $\Phi_{\alpha,M}^n : \mathcal{X}^n \to \mathcal{W}^n$.*

Let W_i denote the ith component of the output $\Phi_{\alpha,M}^n(X^n)$. The average output Rényi entropy γ_n associated with $\Phi_{\alpha,M}^n$ is then given by $\gamma_n = \frac{1}{n}\sum_{i=1}^n H_\alpha(W_i)$, and the associated average relevance level η_n is computed as $\eta_n = \frac{1}{n}\sum_{i=1}^n I(Y_i;W_i)$.

Definition 4. *A Rényi entropy-relevance pair (γ,η) is said to be achievable if there exists a sequence of codes $\{\Phi_{\alpha,M}^n\}$ such that $\limsup_{n\to\infty} \gamma_n \le \gamma$ and $\liminf_{n\to\infty} \eta_n \ge \eta$. The achievable Rényi entropy-relevance region $\mathcal{R}_{\alpha,M} \subseteq \mathbb{R}_{\ge 0}^2$ is defined as the closure of all achievable pairs.*

Similar to the IB result, we next define $I_{\alpha,M}(\gamma) = \max(\eta : (\gamma,\eta) \in \mathcal{R}_{\alpha,M})$. We obtain the following theorem.

Theorem 1. *$I_{\alpha,M}(\gamma) = \bar{I}_{\alpha,M}(\gamma)$ for $\gamma \ge 0$.*

Proof: (*Achievability*): It suffices to consider the situation where $\gamma \le H_\alpha(W_M^*)$ since by the flatness property of the function $\bar{I}_{\alpha,M}$ and Definition 4, the achievability of the pair $(H_\alpha(W_M^*), \bar{I}_{\alpha,M}(H_\alpha(W_M^*)))$ implies the achievability of $(\gamma, \bar{I}_{\alpha,M}(\gamma))$ for any $\gamma > H_\alpha(W_M^*)$. Now we show that any pair $(\gamma,\eta) = (\gamma, \bar{I}_{\alpha,M}(\gamma))$ with $\gamma \le H_\alpha(W_M^*)$ is achievable. By the definition of $\bar{I}_{\alpha,M}(\gamma)$ in (2), we can

write $(\gamma, \bar{I}_{\alpha,M}(\gamma)) = \sum_{k=1}^{2} \lambda_k(\gamma^{(k)}, F_{\alpha,M}(\gamma^{(k)}))$ for some pair $(\gamma^{(k)}, F_{\alpha,M}(\gamma^{(k)}))$, $\lambda_k \geq 0$, $k = 1, 2$, and $\lambda_1 + \lambda_2 = 1$. Suppose that the conditional probability $P_{W^{(k)}|X}$ determines $(\gamma^{(k)}, F_{\alpha,M}(\gamma^{(k)}))$. Note that such $P_{W^{(k)}|X}$ exists due to the definition of $F_{\alpha,M}$.

We next construct a code $\Phi^n_{\alpha,M}$ using time-sharing [33]. Specifically, the input sequence X^n is divided into two disjoint sub-blocks and the size of the kth sub-block is $n\alpha_k$. In the kth sub-block, $k = 1, 2$, our code $\Phi^n_{\alpha,M}$ maps each X_i into W_i symbol-wise according to $P_{W^{(k)}|X}$. Clearly, using this coding scheme, the average output Rényi and the average relevance of the kth sub-block will be $\gamma^{(k)}$ and $\eta^{(k)}$, respectively, implying that the pair (γ, η) is achievable.

(*Converse*): We claim that any achievable pair (γ, η) satisfies $\eta \leq \bar{I}_{\alpha,M}(\gamma)$. Given (γ, η)-achievable codes $\{\Phi^n_{\alpha,M}\}$, we proceed the following standard steps to prove the claim:

$$\begin{aligned}\eta_n = \frac{1}{n}\sum_{i=1}^{n} I(Y_i; W_i) &\leq \frac{1}{n}\sum_{i=1}^{n} \bar{I}_{\alpha,M}(H_\alpha(W_i)) \\ &\leq \bar{I}_{\alpha,M}\left(\frac{1}{n}\sum_{i=1}^{n} H_\alpha(W_i)\right) = \bar{I}_{\alpha,M}(\gamma_n),\end{aligned}$$

where the second inequality holds since $\bar{I}_{\alpha,M}$ is concave while others hold by definition. The claim is proved by noting that

$$\begin{aligned}\eta \leq \liminf_{n\to\infty} \eta_n \leq \liminf_{n\to\infty} \bar{I}_{\alpha,M}(\gamma_n) &= \bar{I}_{\alpha,M}\left(\liminf_{n\to\infty} \gamma_n\right) \\ \leq \bar{I}_{\alpha,M}\left(\limsup_{n\to\infty} \gamma_n\right) &\leq \bar{I}_{\alpha,M}(\gamma),\end{aligned}$$

where we have used the continuity and non-decreasing properties of $\bar{I}_{\alpha,M}$ and the definition of achievability. ∎

IV. Numerical Methods and Examples

Due to an entropy mismatch for W in the objective function $I(Y;W) = H(W) - H(W|Y)$ and the constraint $H_\alpha(W) \leq \gamma$ of (1), an analytical expression for $\bar{I}_{\alpha,M}$ is difficult to derive. This section discusses numerical methods for approximating $\bar{I}_{\alpha,M}$. Two special examples where $\bar{I}_{1,M}$ can be exactly determined are given and some numerical results are also presented. To approximate $\bar{I}_{\alpha,M}$, we consider the maximization problem

$$\max_{\substack{P_{W|X}: Y \multimap X \multimap W \\ |\mathcal{W}| = M}} \left[\beta I(Y;W) - H_\alpha(W)\right] \tag{6}$$

where $\beta \in [0, \infty)$ controls the trade-off between $I(Y;W)$ and $H_\alpha(W)$. For a fixed β, the maximizer $P_{W^*|X}$ of (6) will result in a Rényi entropy-relevance pair $(H_\alpha(W^*), I(Y;W^*))$.[3] When $\beta = 0$, one obtains the trivial pair $(0, 0)$. Moreover, the argument in Section II-B implies that the maximizer of (6) is deterministic. Thus, one can estimate $\bar{I}_{\alpha,M}(\gamma)$ by varying β and checking $|\mathcal{W}|^{|\mathcal{X}|}$ possible deterministic mappings for each fixed β. Specifically, let S' denote the set of all obtained Rényi entropy-relevance pairs. For each $\gamma \geq 0$, the estimation of $\bar{I}_{\alpha,M}(\gamma)$ is then given by $\max(\eta : (\gamma', \eta) \in S', \gamma' \leq \gamma)$. Clearly, the overall procedure is quite complex.

[3]We note that this pair $(H_\alpha(W^*), I(Y;W^*))$ is an extreme point of the hypograph of $\bar{I}_{\alpha,M}(\gamma)$ picked out by a support line of slope $\frac{1}{\beta}$.

Next we provide an iterative algorithm that avoids checking all $|\mathcal{W}|^{|\mathcal{X}|}$ possible mappings for a fixed β. The algorithm is derived using the first-order optimality condition [34] for the following modified Lagrangian of (6):

$$\begin{aligned}\mathcal{L}(\nu, \boldsymbol{\mu}, \beta, \alpha, P_{W|X}) = \beta I(Y;W) - \nu H(W|X) \\ - H_\alpha(W) - \sum_{x\in\mathcal{X}} \mu(x) \sum_{w\in\mathcal{W}} P_{W|X}(w|x)\end{aligned} \tag{7}$$

where $\nu \geq 0$ and $\boldsymbol{\mu}$ is a vector containing the Lagrangian multipliers $\mu(x)$, $x \in \mathcal{X}$, for the constraint $\sum_w P_{W|X}(w|x) = 1$. As $\nu \to 0$, the function $\mathcal{L}$ converges to the Lagrangian of (6). Note that the term $-\nu H(W|X)$ is added to obtain an explicit expression of $P_{W|X}$ in (8) below. Setting $\frac{\partial \mathcal{L}}{\partial P_{W|X}(w|x)} = 0$ for each pair w and x, we obtain the following consistency equation for the maximizer $P_{W|X}$ of $\mathcal{L}$:

$$\begin{aligned}P_{W|X}(w|x) = \frac{1}{Z}\exp\Bigg[\frac{-1}{\nu}\Bigg(\frac{\alpha(P_W(w))^{\alpha-1}}{(1-\alpha)\sum_{w\in\mathcal{W}}(P_W(w))^\alpha} \\ + \beta D(P_{Y|X=x}||P_{Y|W=w})\Bigg)\Bigg]\end{aligned} \tag{8}$$

where $Z := Z(x, \nu, \boldsymbol{\mu}, \alpha)$ is a normalization factor and $D(\cdot||\cdot)$ denotes the Kullback–Leibler divergence [33]. Letting $\nu \to 0$ in (8) for all w and fixed x, one easily observes that the optimal $P_{W|X}$ tends to be deterministic, which coincides with the result obtained by applying the convex optimization argument in Section II-B to (6). Using this observation and (8), we propose the following iterative algorithm to solve (6). Note that [16, Algorithm 2] can be employed to estimate $\bar{I}_{1,M}$.

Initialization: Randomly generate $P^{(0)}_{W|X}$ and obtain $P^{(0)}_W$ and $P^{(0)}_{Y|W}$ from the probability distribution $P^{(0)}_{W|X}P_{Y,X}$. Set $l = 1$.

Step 1: For each $x \in \mathcal{X}$, compute

$$\begin{aligned}w^*(x) = \operatorname*{argmax}_{w\in\mathcal{W}} \Bigg[\frac{\alpha(P^{(l-1)}_W(w))^{\alpha-1}}{(1-\alpha)\sum_{w\in\mathcal{W}}(P^{(l-1)}_W(w))^\alpha} \\ + \beta D(P_{Y|X=x}||P^{(l-1)}_{Y|W=w})\Bigg]\end{aligned}$$

and set $P^{(l)}_{W|X}(w|x) = 1$ for $w = w^*(x)$ and set $P^{(l)}_{W|X}(w|x) = 0$ for all $w \neq w^*(x)$.

Step 2: If $P^{(l)}_{W|X} = P^{(l-1)}_{W|X}$ or the maximum number of iterations is reached, then terminate the procedure and compute the pair $(H_\alpha(W), I(Y;W))$ using $P^{(l)}_{W|X}P_{Y,X}$. Otherwise, obtain the marginal probability distributions $P^{(l)}_W$ and $P^{(l)}_{Y|W}$ from $P^{(l)}_{W|X}P_{Y,X}$, set $l = l + 1$, and go back to Step 1.

We remark that the above iterative algorithm may only yield a local maximizer for (6). To alleviate this situation, one can initialize this algorithm with different $P^{(0)}_{W|X}$ and choose the best result. Next, we determine $\bar{I}_{1,M}$ for two special cases.

Example 1. Given a function f, consider $Y = f(X)$. Then, we have that $\bar{I}_{1,M}(\gamma) = \Omega(\gamma)$ for any finite $M \geq |\mathcal{X}|$; the function $\bar{I}_{1,M}$ is drawn in Fig. 1(c). Based on (5), it suffices to show that $\bar{I}_{1,M}(0) = 0$ and $\bar{I}_{1,M}(I(Y;X)) = I(Y;X)$, where $I(Y;X) = H(Y)$. The former case is apparent since $H(W) = 0$ implies that $I(Y;W) = 0$ and hence $\bar{I}_M(0) = 0$. For the latter case, we set $W = h(f(X))$ for some injective

function $h : \mathcal{Y} \to \mathcal{W}$. The desired result simply follows from the fact that the $P_{W|X}$ induced by the mapping $W = h(f(X))$ achieves the upper bound in (5a) since

$$H(W) = H(h(f(X))) = H(h(Y)) = H(Y),$$
$$I(Y;W) = H(Y) - H(Y|W) = H(Y),$$

where $H(Y|W) = H(Y|h(Y)) = 0$ since h is injective.

When comparing our result with the DIB result in Fig. 1(c), we observe that $\bar{I}_{1,M}$ attains a higher relevance level given the same Shannon entropy constraint. In fact, this increment of relevance level mainly comes from the convexification of the DIB curve. Such an operation is missing in the context of the DIB method due to the lack of an operational interpretation. Here, our Theorem 1 provides the rationale for this operation and indicates that one can use time-sharing between two DIB schemes to take a better Shannon entropy-relevance trade-off.

Example 2. Suppose that the given joint probability matrix $[P_{Y,X}(\cdot\,,\cdot\,)]$ can be arranged into a diagonal form as shown in Table I(a) with the maximum possible number K of non-zero blocks and the kth block contains identical probability mass p_k, $1 \le k \le K$. Choose $f_1 : \mathcal{X} \to \{1, 2, \ldots, K\}$ and $f_2 : \mathcal{Y} \to \{1, 2, \ldots, K\}$ such that $P_{f_1(X)}$ and $P_{f_2(Y)}$ are strictly positive and $f_1(x) = f_2(y)$ whenever $P_{X,Y}(x, y) > 0$ for all x and y. A choice of such f_1 and f_2 is given in Table I(b). Moreover, let $\mathcal{X}_k = \{x \in \mathcal{X} : f_1(x) = k\}$ and $\mathcal{Y}_k = \{y \in \mathcal{Y} : f_2(y) = k\}$, and set $s_k = \Pr(X \in \mathcal{X}_k, Y \in \mathcal{Y}_k) = |\mathcal{X}_k||\mathcal{Y}_k|p_k$. We can then show that $\bar{I}_{1,M}(\gamma) = \Omega(\gamma)$ for any finite $M \ge K$. The details are provided in the Appendix.

To end this section, we apply our iterative algorithm to estimate $\bar{I}_{\alpha,M}$ for the $P_{Y,X}$ in Table I(a) with $\alpha \in \{0.1, 0.5\}$ and $M \in \{2, 3\}$. Here, $H(X) = 2.25$ bits and $I(Y;X) = 1.5$ bits. Our estimation for the different $\bar{I}_{\alpha,M}$'s are depicted in Fig. 2. When $M = 2$ and $\beta \ge 1$, our algorithm produces a maximizer $P_{W|X}$ that induces a uniform distribution on $\mathcal{W}$, regardless of the value of α. This maximizer corresponds to the Rényi entropy-relevance pair $(1, 1)$. Together with the trivial pair $(0, 0)$, we obtain an estimate for $\bar{I}_{\alpha,2}$. In fact, Lemmas 1 and 3 imply that our estimation is exact in this case. Next, when considering $M = 3$, we observe that maximum relevance is attained. The required $H_\alpha(W^*_M)$ varies with α, but all of them are less than $H(X)$. This result shows that our bottleneck method can also effectively extract relevance information while minimizing the representation cost.

V. Conclusion

Unlike the IB method that characterizes the optimal complexity-relevance trade-off via a single optimization problem, our bottleneck problem needs additional convexification to describe the optimal Rényi entropy-relevance trade-off. Our optimal operational scheme consists of two symbol-wise transformations that operate in a time-sharing manner. Though not discussed here, our information-Rényi entropy method has been applied to geometric clustering and shows a robustness result when the probability distributions of the data model and the data sets are mismatched [35]. Still, it remains unclear how the optimal Rényi entropy-relevance trade-off vary with α and M, which we leave for future research. Other research topics include extending our approach to the variational IB problem [12] and applying our result to other tasks in machine learning.

$P_{Y,X}$	1	2	3	4	5
1	$\frac{1}{4}$	0	0	0	0
2	0	$\frac{1}{8}$	$\frac{1}{8}$	0	0
3	0	$\frac{1}{8}$	$\frac{1}{8}$	0	0
4	0	0	0	$\frac{1}{8}$	$\frac{1}{8}$

(a) $P_{Y,X}$

X	1	2	3	4	5
$f_1(X)$	1	2	2	3	3

Y	1	2	3	4
$f_2(Y)$	1	2	2	3

(b) f_1 and f_2

TABLE I: An illustration of Example 2. Here, $K = 3$, $p_1 = \frac{1}{4}$, $p_2 = p_3 = \frac{1}{8}$, $s_1 = \frac{1}{4}$, $s_2 = \frac{1}{2}$, and $s_3 = \frac{1}{4}$.

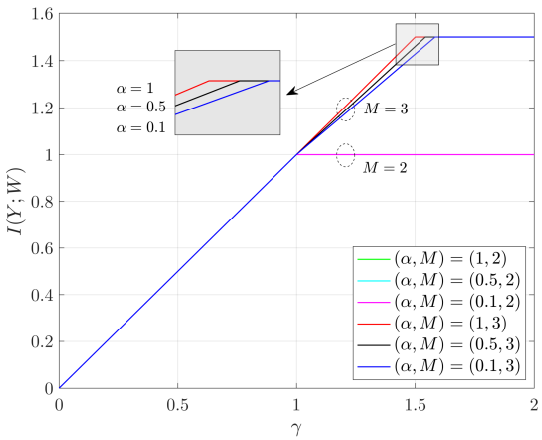

Fig. 2: The estimated $\bar{I}_{\alpha,M}$ for $P_{Y,X}$ given in Table 1(a), where $M = \{2, 3\}$ and $\alpha = \{0.1, 0.5, 1\}$. Our estimation for $\bar{I}_{\alpha,2}$ are tight, and the curves are overlapping. Our estimate of $\bar{I}_{\alpha,M}$ for $M \ge 3$ are identical for each given α for this $P_{Y,X}$.

Appendix
Supplementary Result for Example 2

Similar to Example 1, it suffices to show that $\bar{I}_{1,M}(0) = 0$ and $\bar{I}_{1,M}(I(Y;X)) = I(Y;X)$. The former case is clear and thus omitted, but for the latter case, we need the following result. Given any injective function $h : \{1, 2, \ldots, K\} \to \mathcal{W}$, the function $g : \mathcal{X} \to \mathcal{W}$ defined as $g(x) = h(f_1(x))$ yields:

$$\begin{aligned} I(Y;W) &= H(W) - H(W|Y) \\ &= H(W) - H(h(f_1(X))|Y) \\ &= H(W) - H(h(f_2(Y))|Y) = H(W). \end{aligned}$$

Without loss of generality, choose $g(x) = k$ for $x \in \mathcal{X}_k$. We prove explicitly that $H(W) = I(X;Y)$. First,

$$\begin{aligned} &I(Y;X) = H(Y) - H(Y|X) \\ &= -\sum_{k=1}^{K} \underbrace{|\mathcal{Y}_k|(|\mathcal{X}_k|p_k)}_{=s_k} \log_2(|\mathcal{X}_k|p_k) - \sum_{k=1}^{K} \underbrace{|\mathcal{X}_k||\mathcal{Y}_k|p_k}_{=s_k} \log_2 |\mathcal{Y}_k| \\ &= -\sum_{k=1}^{K} s_k \log_2 s_k. \end{aligned}$$

Moreover, since $w = g(x) = k$ for $x \in \mathcal{X}_k$, we obtain that

$$P_W(k) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} P_{Y,X}(y, x) P_{W|X}(k|x) = |\mathcal{Y}_k||\mathcal{X}_k|p_k = s_k.$$

and hence $I(Y;X) = H(W) = I(Y;W)$. Based on the bound in (5b), the equality $\bar{I}_{1,M}(I(Y;X)) = I(Y;X)$ clearly holds.

REFERENCES

[1] R. A. Amjad and B. C. Geiger, "Learning representations for neural network-based classification using the information bottleneck principle," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 9, pp. 2225–2239, Sep. 2019.

[2] D. Strouse and D. J. Schwab, "The information bottleneck and geometric clustering," *Neural Computation*, vol. 31, no. 3, pp. 596–612, 2019.

[3] Z. Goldfeld and Y. Polyanskiy, "The information bottleneck problem and its applications in machine learning," *IEEE J. Select. Areas Inf. Theory*, vol. 1, no. 1, pp. 19–38, May 2020.

[4] A. Zaidi, I. Estella-Aguerri, and S. Shamai (Shitz), "On the information bottleneck problems: models, connections, applications and information theoretic views," *Entropy*, vol. 22, no. 151, pp. 1–36, 2020.

[5] G. Zeitler, A. C. Singer, and G. Kramer, "Low-precision A/D conversion for maximum information rate in channels with memory," *IEEE Trans. Commun.*, vol. 60, no. 9, pp. 2511–2521, Sep. 2012.

[6] A. Winkelbauer, G. Matz, and A. Burg, "Channel-optimized vector quantization with mutual information as fidelity criterion," in *Proc. IEEE Asilomar Conf. Signals, Syst. and Comp.*, 2013, pp. 851–855.

[7] B. M. Kurkoski and H. Yagi, "Quantization of binary-input discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. 60, no. 8, pp. 4544–4552, Aug. 2014.

[8] M. Meidlinger, G. Matz, and A. Burg, "Design and decoding of irregular LDPC codes based on discrete message passing," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1329–1343, Mar. 2019.

[9] T. Nguyen and T. Nguyen, "On binary quantizer for maximizing mutual information," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5435–5445, Jun. 2020.

[10] M. Stark, L. Wang, G. Bauch, and R. D. Wesel, "Decoding rate-compatible 5G-LDPC codes with coarse quantization using the information bottleneck method," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 646–660, May 2020.

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

[12] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," in *Proc. 5th Int. Conf. Learning Representations (ICLR)*, 2017, pp. 1–17.

[13] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *arXiv preprint arXiv:1703.00810*, 2017.

[14] H. Hsu, S. Asoodeh, S. Salamatian, and F. P. Calmon, "Generalizing bottleneck problems," in *Proc. IEEE Int. Symp. Inf. Theory*, 2018, pp. 531–535.

[15] S. Asoodeh and F. Calmon, "Bottleneck problems: Information and estimation-theoretic view," *Entropy Special Issue on Information-Theoretic Methods for Deep Learning*, 2020.

[16] D. Strouse and D. J. Schwab, "The deterministic information bottleneck," *Neural Computation*, vol. 29, no. 6, pp. 1611–1630, 2017.

[17] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.

[18] L. L. Campbell, "A coding theorem and Rényi's entropy," *Inform. Contr.*, vol. 8, no. 4, pp. 423–429, 1965.

[19] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Stat. Probab.*, 1960, pp. 547–561.

[20] I. Csiszár, "Generalized cutoff rates and Rényi's information measures," *IEEE Trans. Inf. Theory*, vol. 41, no. 1, pp. 26–34, Jan. 1995.

[21] P.-N. Chen and F. Alajaji, "Csiszár's cutoff rates for arbitrary discrete sources," *IEEE Trans. Inf. Theory*, vol. 47, no. 1, pp. 330–338, Jan. 2001.

[22] W. Kreitmeier and T. Linder, "High-resolution scalar quantization with Rényi entropy constraint," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6837–6859, Oct. 2011.

[23] C. Bunte and A. Lapidoth, "Encoding tasks and Rényi entropy," *IEEE Trans. Inf. Theory*, vol. 60, no. 9, pp. 5065–5076, Sep. 2014.

[24] A. Bracher, E. Hof, and A. Lapidoth, "Guessing attacks on distributed-storage systems," *IEEE Trans. Inf. Theory*, vol. 65, no. 11, pp. 6975–6998, Nov. 2019.

[25] C. Hirche, "Rényi bounds on information combining," in *Proc. IEEE Int. Symp. Inf. Theory*, 2020, pp. 2297–2302.

[26] H. Bhatia, W. Paul, F. Alajaji, B. Gharesifard, and P. Burlina, "Least kth-order and Rényi generative adversarial networks," *arXiv preprint arXiv:2006.02479v2*, 2020.

[27] R. T. Rockafellar, *Convex Analysis*. Princeton University Press, 1970.

[28] R. Gilad-Bachrach, A. Navot, and N. Tishby, "An information theoretic tradeoff between complexity and accuracy," in *Learning Theory and Kernel Machines*. Springer, 2003, pp. 595–609.

[29] T. van Erven and P. Harremos, "Rényi divergence and Kullback-Leibler divergence," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 3797–3820, Jul. 2014.

[30] H. Witsenhausen and A. Wyner, "A conditional entropy bound for a pair of discrete random variables," *IEEE Trans. Inf. Theory*, vol. 21, no. 5, pp. 493–501, Sep. 1975.

[31] B. C. Geiger and R. A. Amjad, "Hard clusters maximize mutual information," in *ITG Conf. Syst., Commun. and Coding (SCC)*, 2016.

[32] A. Kolchinsky, B. D. Tracey, and S. Van Kuyk, "Caveats for information bottleneck in deterministic scenarios," in *Proc. 6th Int. Conf. Learning Representations (ICLR)*, 2018.

[33] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New Jersey: John Wiley & Sons, 2006.

[34] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, UK: Cambridge University Press, 2004.

[35] A. Moran-MacDonald, "The Rényi deterministic information bottleneck and geometric clustering," NSERC Summer Research Project, Queen's University, Tech. Rep., 2018.