entropy

MDPI

Article

A Comparison of Variational Bounds for the Information Bottleneck Functional

Bernhard C. Geiger [1,*] **and Ian S. Fischer** [2]

[1] Know-Center GmbH, Inffeldgasse 13/6, 8010 Graz, Austria
[2] Google Research, Mountain View, CA 94043, USA; iansf@google.com
* Correspondence: geiger@ieee.org

Received: 24 September 2020; Accepted: 20 October 2020; Published: 29 October 2020

Abstract: In this short note, we relate the variational bounds proposed in Alemi et al. (2017) and Fischer (2020) for the information bottleneck (IB) and the conditional entropy bottleneck (CEB) functional, respectively. Although the two functionals were shown to be equivalent, it was empirically observed that optimizing bounds on the CEB functional achieves better generalization performance and adversarial robustness than optimizing those on the IB functional. This work tries to shed light on this issue by showing that, in the most general setting, no ordering can be established between these variational bounds, while such an ordering can be enforced by restricting the feasible sets over which the optimizations take place. The absence of such an ordering in the general setup suggests that the variational bound on the CEB functional is either more amenable to optimization or a relevant cost function for optimization in its own regard, i.e., without justification from the IB or CEB functionals.

Keywords: information bottleneck; deep learning; neural networks

1. Introduction

The celebrated information bottleneck (IB) functional [1] is a cost function for supervised lossy compression. More specifically, if X is an observation and Y a stochastically related random variable (RV) that we associate with relevance, then the IB problem aims to find an encoder $e_{Z|X}$, i.e., a conditional distribution of Z given X, that minimizes

$$\mathcal{L}_{\text{IB}} := I(X;Z) - \beta I(Y;Z). \tag{1}$$

In (1), $I(X;Z)$ and $I(Y;Z)$ denote the mutual information between observation X and representation Z and between relevant variable Y and representation Z, respectively, and β is a Lagrangian parameter. The aim is to obtain a representation Z that is simultaneously compressed (small $I(X;Z)$) and informative about the relevant variable Y (large $I(Y;Z)$), and the parameter β trades between these two goals.

Recently, Fischer proposed an equivalent formulation, termed the conditional entropy bottleneck (CEB) [2]. While the IB functional inherently assumes the Markov condition $Y - X - Z$, the CEB is motivated from the principle of Minimum Necessary Information, which lacks this Markov condition and which aims to find a representation Z that compresses a bi-variate dataset $(X;Y)$ while still being useful for a given task. Instantiating the principle of Minimum Necessary Information induces then a Markov condition. For example, the task of finding a representation Z that makes X and Y conditionally independent induces the Markov condition $X - Z - Y$, and the representation optimal w.r.t. the principle of Minimum Necessary Information turns out to be $\arg\inf_{X-Z-Y} I(X,Y;Z)$, i.e., it is related to Wyner's common information [3]. The task relevant in this work—estimating Y from a representation Z that is obtained exclusively from X—induces the Markov condition $Y - X - Z$ and

the constraint $I(Y;Z) \geq I(X;Y)$. A Lagrangian formulation of the constrained optimization problem $\inf_{I(Y;Z)\geq I(X;Y)} I(X;Z)$, where the infimum is taken over all encoders $e_{Z|X}$ that take only X as input, yields the CEB functional (see Section 2.3 of [2])

$$\mathcal{L}_{\text{CEB}} := I(X;Z|Y) - \gamma I(Y;Z). \tag{2}$$

Due to the chain rule of mutual information [4] (Theorem 2.5.2), (2) is equivalent to (1) for $\gamma = \beta - 1$. Nevertheless, (2) has additional appeals. To this end, note that $I(X;Z|Y)$ captures the information about X contained in the representation Z that is redundant for the task of predicting the class variable Y. In the language of [5], which essentially also proposed (2), $I(X;Z|Y)$ thus quantifies class-conditional compression. Minimizing this class-conditional compression term $I(X;Z|Y)$ is not in conflict with maximizing $I(Y;Z)$, whereas minimizing $I(X;Z)$ is (see Figure 2 in [2] and Section 2 in [5]). At the same time, as stated in [2] (p. 6), $I(X;Z|Y)$ allows to "measure in absolute terms how much more we could compress our representation at the same predictive performance", i.e., by how much $I(X;Z|Y)$ could potentially be further reduced without simultaneously reducing $I(Y;Z)$.

Aside from these theoretical considerations that make the CEB functional preferable over the equivalent IB functional, it has been shown that minimizing variational bounds on the former achieve better performance than minimizing variational bounds on the latter [2,6]. More specifically, it was shown that variational CEB (VCEB) achieves higher classification accuracy and better robustness against adversarial attacks than variational IB (VIB) proposed in [7].

The exact underlying reason why VCEB outperforms VIB is currently still being investigated. Comparing these two bounds at $\beta - 1 = \gamma = 1$, Fischer suggests that "we may expect VIB to converge to a looser approximation of $I(X;Z) = I(Y;Z) = I(X;Y)$", where the later equation corresponds to the Minimum Necessary Information point (see Section 2.5.1 of [2]). Furthermore, Fischer and Alemi claim that VCEB "can be thought of as a tighter variational approximation to the IB objective than VIB" (see Section 2.1 of [6]). Nevertheless, the following question remains: Does VCEB outperform VIB because the variational bound of VCEB is tighter, or because VCEB is more amenable to optimization than VIB?

To partly answer this question, we compare the optimization problems corresponding to VCEB and VIB. Rather than focusing on actual (commonly neural network-based) implementations of these problems, we keep an entirely mathematical perspective and discuss the problem of finding minimizers within well-defined feasible sets (see Section 3). Our main result in Section 4 shows that the optimization problems corresponding to VCEB and VIB are indeed ordered if additional constraints are added: If VCEB is constrained to use a consistent classifier-backward encoder pair (see Definition 1 below), then (unconstrained) VIB yields a tighter approximation of the IB functional. In contrast, if VIB is constrained to use a consistent classifier-marginal pair, then (constrained and unconstrained) VCEB yields a tighter approximation. If neither VCEB nor VIB are constrained, then no ordering can be shown between the resulting optimal variational bounds. Taken together, these results indicate that the superiority of VCEB over VIB observed in [2,6] cannot be due to VCEB better approximating the IB functional. Rather, we conclude in Section 5 that the variational bound provided in [2] is either more amenable to optimization, at least when the variational terms in VCEB and VIB are implemented using neural networks (NNs), or a successful cost function for optimization in its own regard, i.e., without justification from the IB or Minimum Necessary Information principles.

Related Work and Scope. Many variational bounds for mutual information have been proposed [8], and many of these bounds can be applied to the IB functional. Both the VIB and VCEB variational bounds belong to the class of Barber & Agakov bounds, cf. Section 2.1 of [8]. As an alternative example, the authors of [9] bounded the IB functional using the Donsker–Varadhan representation of mutual information. Aside from that, the IB functional has been used for NN training also without resorting to purely variational approaches. For example, the authors of [10] applied the Barber & Agakov bound to replace $I(Y;Z)$ by the standard cross-entropy loss of a trained classifier, but used a non-parametric estimator for $I(X;Z)$. Rather than comparing multiple variational bounds

with each other, in this work we focus exclusively on the VIB [7] and VCEB [2] bounds. The structural similarity of these bounds allows a direct comparison and still yields interesting insights that can potentially carry over to other variational approaches.

We finally want to mention two works that draw conclusions similar to ours. First, Achille and Soatto [11] pointed to the fact that their choice of injecting multiplicative noise to neuron activations is not only a restriction of the feasible set over which the optimization is performed, but it can also be interpreted as a means of regularization or as an approach to perform optimization. Thus, the authors claim, there is an intricate connection between regularization (i.e., the cost function), the feasible set, and the method of optimization (see Section 9 of [11]); this claim resonates with our Section 5. Second, Wieczorek and Roth [12] investigate the difference between IB and VIB: While IB implicitly assumes the Markov condition $Y - X - Z$, the variational approach taken in VIB assumes that an estimate of Y is obtained from the representation Z, i.e., $X - Z - Y$. Dropping the former assumption allows to express the difference between the VIB bound and the IB functional via mutual and lautum information, which, taken together, measure the violation of the condition $Y - X - Z$. The authors thus argue that dropping this condition enables VIB and similar variants to optimize over larger sets of joint distributions of X, Y, and Z. In this work, we take a slightly different approach and argue that the posterior distribution of Y given Z is approximated by a classifier with input Z that responds with a class estimate $\hat{Y}$. Thus, we stick to the Markov condition inherent to IB and extend it by an additional variable, resulting in $Y - X - Z - \hat{Y}$. As a consequence, our variational approach does not assume that $X - Z - Y$ holds, which also leads to a larger set of joint distributions of X, Y, and Z. Finally, while [12] compares the IB functional with the VIB bound, in our work we compare two variational bounds on the IB functional with each other.

Notation. We consider a classification task with a feature RV X on $\mathbb{R}^m$ and a class RV Y on the finite set $\mathcal{Y}$ of classes. We assume that the joint distribution of X and Y is denoted by p_{XY}. In this work we are interested in representations Z of the feature RV X. This (typically real-valued) representation Z is obtained by feeding X to a stochastic encoder $e_{Z|X}$, and the representation Z can be used to infer the class label by feeding it to a classifier $c_{\hat{Y}|Z}$. Note that this classifier yields a class *estimate* $\hat{Y}$ that need not coincide with the class RV Y. Thus, the setup of encoder, representation, and classifier yields the following Markov condition: $Y - X - Z - \hat{Y}$. We abuse notation and abbreviate the conditional probability (density) $p_{W|V=v}(\cdot)$ of a RV W given that another RV V assumes a certain value v as $p_{W|V}(\cdot|v)$. For example, the probability density of the representation Z for an input $X = x$ is induced by the encoder $e_{Z|X}$ and is given as $e_{Z|X}(\cdot|x)$.

We obtain encoder, classifier, and eventual variational distributions via solving a constrained optimization problem. For example, $\min_{e_{Z|X} \in \mathcal{E}} \mathcal{J}$ minimizes the objective $\mathcal{J}$ over all encoders $e_{Z|X}$ from a given family $\mathcal{E}$. In practice, encoder, classifier, and variational distributions are parameterized by (stochastic) feed-forward NNs. The chosen architecture has a certain influence on the feasible set; e.g., $\mathcal{E}$ may denote the set of encoders that can be parameterized by a NN of a given architecture.

We assume that the reader is familiar with information-theoretic quantities. More specifically, we let $I(\cdot;\cdot)$ and $D(\cdot\|\cdot)$ denote mutual information and Kullback–Leibler divergence, respectively. The expectation w.r.t. to a RV W drawn from a distribution p_W is denoted as $E_{W\sim p_W}[\cdot]$.

2. Variational Bounds on the Information Bottleneck Functional

We consider the IB principle for NN training. Specifically, we are interested in a (real-valued) representation Z, obtained directly from X, that minimizes the following functional:

$$\mathcal{L}_{\mathrm{IB}}(\beta) := I(X;Z) - \beta I(Y;Z) = I(X;Z|Y) - (\beta - 1) I(Y;Z) =: \mathcal{L}_{\mathrm{CEB}}(\beta - 1) \tag{3}$$

Rather than optimizing (3) directly (which was shown to be ill-advised at least for deterministic NNs in [13]), we rely on minimizing variational upper bounds. More specifically, the authors of [7] introduced the following variational bound on $\mathcal{L}_{\mathrm{IB}}$:

$$\mathcal{L}_{\text{VIB}}(\beta) := E_{X\sim p_X}\left[D\left(e_{Z|X}(\cdot|X)\|q_Z\right)\right] - \beta H(Y) - \beta E_{XYZ\sim p_{XY}e_{Z|X}}\left[\log c_{\hat{Y}|Z}(Y|Z)\right] \tag{4}$$

where $e_{Z|X}$, $c_{\hat{Y}|Z}$, and q_Z are called the encoder, classifier, and marginal. The classifier is used as a variational approximation to the distribution $p_{Y|Z}$. The marginal q_Z is a learned distribution that aims to marginalize out the encoder $e_{Z|X}$. As such, this distribution is conceptually different from a fixed (unlearned) prior distribution in a Bayesian framework as in, e.g., the variational auto-encoder [14].

As an alternative and motivated by the principle of Minimum Necessary Information, the author of [2] proposed the variational bound on the CEB functional:

$$\mathcal{L}_{\text{VCEB}}(\beta) := E_{XY\sim p_{XY}}\left[D\left(e_{Z|X}(\cdot|X)\|b_{Z|Y}(\cdot|Y)\right)\right] - \beta H(Y) - \beta E_{XYZ\sim p_{XY}e_{Z|X}}\left[\log c_{\hat{Y}|Z}(Y|Z)\right] \tag{5}$$

where $b_{Z|Y}$ is called the backward encoder, which is a variational approximation to the distribution $p_{Z|Y}$.

3. Variational IB and Variational CEB as Optimization Problems

While it is known that $\mathcal{L}_{\text{IB}}(\beta) \leq \mathcal{L}_{\text{VIB}}(\beta)$ and $\mathcal{L}_{\text{IB}}(\beta) \leq \mathcal{L}_{\text{VCEB}}(\beta - 1)$ for all possible p_{XY} and all choices of $e_{Z|X}$, $b_{Z|Y}$, $c_{\hat{Y}|Z}$, and q_Z, it is not obvious how $\mathcal{L}_{\text{IB}}(\beta)$ and $\mathcal{L}_{\text{VCEB}}(\beta - 1)$ compare during optimization. In other words, we are interested in determining whether there is an ordering between

$$\min_{e_{Z|X}, c_{\hat{Y}|Z}, q_Z} \mathcal{L}_{\text{VIB}}(\beta) \tag{6a}$$

and

$$\min_{e_{Z|X}, c_{\hat{Y}|Z}, b_{Z|Y}} \mathcal{L}_{\text{VCEB}}(\beta - 1). \tag{6b}$$

Since we will always compare variational bounds for equivalent parameterization, i.e., compare $\mathcal{L}_{\text{VIB}}(\beta)$ with $\mathcal{L}_{\text{VCEB}}(\beta - 1)$, we will drop the arguments β and $\beta - 1$ for the sake of readability.

For a fair comparison, we need to ensure that both cost functions are optimized over comparable feasible sets $\mathcal{E}$, $\mathcal{C}$, $\mathcal{B}$, and $\mathcal{Q}$ for the encoder, classifier, the backward encoder, and the marginal. We make this explicit in the following assumption.

Assumption 1. *The optimizations of VCEB and VIB are performed over equivalent feasible sets. Specifically, the families $\mathcal{E}$ and $\mathcal{C}$ from which VCEB and VIB can choose encoder $e_{Z|X}$ and classifier $c_{\hat{Y}|Z}$ shall be the same. Depending on the scenario, we may require that the optimization over the marginal q_Z is able to choose from the same mixture models as are induced by VCEB. I.e., if $b_{Z|Y}(\cdot|y)$ is a feasible solution of $\mathcal{L}_{\text{VCEB}}$, then $q_Z(\cdot) = \sum_y b_{Z|Y}(\cdot|y)p_Y(y)$ shall also be a feasible solution for $\mathcal{L}_{\text{VIB}}$; we thus require that $\mathcal{Q} \supseteq \{q_Z \colon q_Z(z) = \sum_y b_{Z|Y}(z|y)p_Y(y),\ b_{Z|Y} \in \mathcal{B}\}$. Depending on the scenario, we may require that every feasible solution for the marginal q_Z shall be achievable by selecting feasible backward encoders; we thus require that $\mathcal{B} \supseteq \{b_{Z|Y} \colon q_Z(z) = \sum_y b_{Z|Y}(z|y)p_Y(y),\ q_Z \in \mathcal{Q}\}$. If both conditions are fulfilled, then we write that $\mathcal{B} \leftrightarrow \mathcal{Q}$.*

We furthermore need the following definition:

Definition 1. *In the optimization of $\mathcal{L}_{\text{VCEB}}$, we say that backward encoder $b_{Z|Y}$ and classifier $c_{\hat{Y}|Z}$ are a* consistent pair *if*

$$c_{\hat{Y}|Z}(y|z) = \frac{p_Y(y)b_{Z|Y}(z|y)}{\sum_{y'} p_Y(y')b_{Z|Y}(z|y')} = \frac{p_Y(y)b_{Z|Y}(z|y)}{q'_Z(z)} \tag{7}$$

holds. In the optimization of $\mathcal{L}_{\mathrm{VIB}}$*, we say that marginal* q_Z *and classifier* $c_{\hat{Y}|Z}$ *are a* consistent pair *if*

$$p_Y(y) = \sum_z c_{\hat{Y}|Z}(y|z) q_Z(z) \tag{8}$$

holds.

The restriction to consistent pairs restricts the feasible sets. For example, for VCEB, if $\mathcal{C}$ is large enough to contain all classifiers consistent with backward encoders in $\mathcal{B}$, i.e., if $\mathcal{C} \supseteq \{c_{\hat{Y}|Z} \colon c_{\hat{Y}|Z}(y|z) \propto p_Y(y) b_{Z|Y}(z|y),\ b_{Z|Y}(\cdot|y) \in \mathcal{B}\}$, then the triple minimization

$$\min_{\substack{e_{Z|X} \in \mathcal{E}, c_{\hat{Y}|Z} \in \mathcal{C}, b_{Z|Y} \in \mathcal{B} \\ (c_{\hat{Y}|Z}, b_{Z|Y}) \text{ consistent}}} \mathcal{L}_{\mathrm{VCEB}} \tag{9}$$

is reduced to the double minimization

$$\min_{e_{Z|X} \in \mathcal{E}, b_{Z|Y} \in \mathcal{B}} \mathcal{L}_{\mathrm{VCEB}}. \tag{10}$$

Equivalently, one can write the joint triple minimization as a consecutive double minimization and a single minimization, where the inner minimization runs over all backwards encoders consistent with the classifier chosen in the outer minimization (where the minimization over an empty set returns infinity):

$$\min_{e_{Z|X} \in \mathcal{E}, c_{\hat{Y}|Z} \in \mathcal{C}} \left[\min_{b_{Z|Y} \in \mathcal{B} \cap \left\{ b'_{Z|Y} \colon \frac{p_Y(y) b'_{Z|Y}(z|y)}{\sum_{y'} p_Y(y') b'_{Z|Y}(z|y')} = c_{\hat{Y}|Z}(y|z) \right\}} \mathcal{L}_{\mathrm{VCEB}} \right]. \tag{11}$$

Similar considerations hold for VIB.

4. Main Results

Our first main result is negative in the sense that it shows $\mathcal{L}_{\mathrm{VIB}}$ and $\mathcal{L}_{\mathrm{VCEB}}$ cannot be ordered in general. To this end, consider the following two examples.

Example 1 (VIB < VCEB)**.** *In this example, let* $\mathcal{B} \leftrightarrow \mathcal{Q}$*, where* $\mathcal{B}$ *and* $\mathcal{Q}$ *are constrained, and let* $\mathcal{C}$ *be unconstrained, thus* $\min_{c_{\hat{Y}|Z} \in \mathcal{C}} -E_{XYZ \sim p_{XY} e_{Z|X}} \left[\log c_{\hat{Y}|Z}(Y|Z) \right] = H(Y|Z)$*. Suppose further that we have selected a fixed encoder* $e_{Z|X}$ *that induces the marginal and conditional distributions* p_Z *and* $p_{Z|Y}$*, respectively. With this, we can write*

$$E_{XY \sim p_{XY}} \left[D\left(e_{Z|X}(\cdot|X) \| b_{Z|Y}(\cdot|Y) \right) \right] = I(X;Z|Y) + E_{Y \sim p_Y} \left[D\left(p_{Z|Y}(\cdot|Y) \| b_{Z|Y}(\cdot|Y) \right) \right] \tag{12a}$$

and

$$E_{X \sim p_X} \left[D\left(e_{Z|X}(\cdot|X) \| q_Z \right) \right] = I(X;Z) + D\left(p_Z \| q_Z \right). \tag{12b}$$

Suppose that $b^{\mathrm{VCEB}}_{Z|Y}$ *is a minimizer of* (12a) *over* $\mathcal{B}$ *and that* $q^{\mathrm{VCEB}}_Z(z) = \sum_y p_Y(y) b^{\mathrm{VCEB}}_{Z|Y}(z|y)$*. By the chain rule of of Kullback–Leibler divergence [4] (Th. 2.5.3) and with* $b^{\mathrm{VCEB}}_{Y|Z}(y|z) = p_Y(y) b^{\mathrm{VCEB}}_{Z|Y}(z|y) / q^{\mathrm{VCEB}}_Z(z)$*, we can expand*

$$D\left(p_{ZY}\|b_{Z|Y}^{\mathrm{VCEB}}p_Y\right) = D\left(p_Z\|q_Z^{\mathrm{VCEB}}\right) + \underbrace{E_{Z\sim p_Z}\left[D\left(p_{Y|Z}(\cdot|Z)\|b_{Y|Z}^{\mathrm{VCEB}}(\cdot|Z)\right)\right]}_{\geq 0}$$
$$= \underbrace{D\left(p_Y\|p_Y\right)}_{=0} + E_{Y\sim p_Y}\left[D\left(p_{Z|Y}(\cdot|Y)\|b_{Z|Y}^{\mathrm{VCEB}}(\cdot|Y)\right)\right]$$

thus

$$E_{Y\sim p_Y}\left[D\left(p_{Z|Y}(\cdot|Y)\|b_{Z|Y}^{\mathrm{VCEB}}(\cdot|Y)\right)\right] \geq D\left(p_Z\|q_Z^{\mathrm{VCEB}}\right).$$

Suppose that $e_{Z|X}$ is such that the inequality above is strict. Then,

$$\begin{aligned}
&\min_{e_{Z|X}\in\mathcal{E}, c_{\hat{Y}|Z}\in\mathcal{C}, b_{Z|Y}\in\mathcal{B}} \mathcal{L}_{\mathrm{VCEB}}(\beta-1)\\
&= I(X;Z|Y) + E_{Y\sim p_Y}\left[D\left(p_{Z|Y}(\cdot|Y)\|b_{Z|Y}^{\mathrm{VCEB}}(\cdot|Y)\right)\right] - (\beta-1)I(Y;Z)\\
&> I(X;Z|Y) + D\left(p_Z\|q_Z^{\mathrm{VCEB}}\right) - (\beta-1)I(Y;Z)\\
&= I(X;Z) + D\left(p_Z\|q_Z^{\mathrm{VCEB}}\right) - \beta I(Y;Z)\\
&\geq \min_{e_{Z|X}\in\mathcal{E}, c_{\hat{Y}|Z}\in\mathcal{C}, q_Z\in\mathcal{Q}} \mathcal{L}_{\mathrm{VIB}}(\beta)
\end{aligned}$$

where the last inequality follows because q_Z^{VCEB} may not be optimal for the VIB cost function.

Example 2 (VIB > VCEB). *Let $\mathcal{B} \leftrightarrow \mathcal{Q}$, where $\mathcal{Q}$ and $\mathcal{B}$ are unconstrained, thus with (12) we have*

$$\min_{b_{Z|Y}\in\mathcal{B}} E_{XY\sim p_{XY}}\left[D\left(e_{Z|X}(\cdot|X)\|b_{Z|Y}(\cdot|Y)\right)\right] = I(X;Z|Y)$$
$$\text{and} \quad \min_{q_Z\in\mathcal{Q}} E_{X\sim p_X}\left[D\left(e_{Z|X}(\cdot|X)\|q_Z\right)\right] = I(X;Z).$$

Suppose further that $\mathcal{C}$ is such that $\min_{c_{\hat{Y}|Z}\in\mathcal{C}} -E_{XYZ\sim p_{XY}e_{Z|X}}\left[\log c_{\hat{Y}|Z}(Y|Z)\right] = H(Y|Z)+\varepsilon$, where $\varepsilon > 0$. It then follows that

$$\begin{aligned}
\min_{e_{Z|X}\in\mathcal{E}, c_{\hat{Y}|Z}\in\mathcal{C}, q_Z\in\mathcal{Q}} \mathcal{L}_{\mathrm{VIB}}(\beta) &= I(X;Z) - \beta I(Y;Z) + \beta\varepsilon\\
&> I(X;Z|Y) - (\beta-1)I(Y;Z) + (\beta-1)\varepsilon = \min_{e_{Z|X}\in\mathcal{E}, c_{\hat{Y}|Z}\in\mathcal{C}, b_{Z|Y}\in\mathcal{B}} \mathcal{L}_{\mathrm{VCEB}}(\beta-1).
\end{aligned}$$

In both of these examples we have ensured that the comparison is fair in the sense of Assumption 1. Aside from showing that VIB and VCEB in general allow no ordering, additional interesting insights can be gleaned from Examples 1 and 2. First, whether VIB or VCEB yield tighter approximations of the IB and CEB functionals for a fixed encoder depends largely on the feasible sets $\mathcal{C}$ and $\mathcal{B}$: Constraints on $\mathcal{C}$ cause disadvantages for VIB, while constraints on $\mathcal{B}$ lead to the VCEB bound becoming looser. Second, for fixed encoders, the tightness of the respective bounds and the question which of the bounds is tighter do not depend on how well the IB and CEB objectives are met: These objectives are functions only of the encoder $e_{Z|X}$, whereas the tightness of the variational bounds depends on $\mathcal{C}$, $\mathcal{B}$, and $\mathcal{Q}$. (Of course, the tightness of the respective bounds after the triple optimization in (6) depends also on $\mathcal{E}$, as the optimization over $\mathcal{B}$ and $\mathcal{Q}$ in Example 1 and over $\mathcal{C}$ in Example 2 interacts with the optimization over $\mathcal{E}$ in a non-trivial manner.)

Our second main result, in contrast, shows that the variational bounds can indeed be ordered if additional constraints are introduced. More specifically, if the variational bounds are restricted to

consistent pairs as in Definition 1, then the following ordering can be shown. The proof of Theorem 1 is deferred to Section 6.

Theorem 1. *If VCEB is constrained to a consistent classifier-backward encoder pair, and if* $\mathcal{Q} \supseteq \{q_Z\colon q_Z(z) = \sum_y b_{Z|Y}(z|y)p_Y(y),\ b_{Z|Y} \in \mathcal{B}\}$, *then*

$$\min_{e_{Z|X}\in\mathcal{E},c_{\hat{Y}|Z}\in\mathcal{C},q_Z\in\mathcal{Q}} \mathcal{L}_{\text{VIB}} \leq \min_{\substack{e_{Z|X}\in\mathcal{E},c_{\hat{Y}|Z}\in\mathcal{C},b_{Z|Y}\in\mathcal{B}\\ (c_{\hat{Y}|Z},b_{Z|Y})\ \textit{consistent}}} \mathcal{L}_{\text{VCEB}}. \tag{13a}$$

If VIB and VCEB are constrained to a consistent classifier–marginal and classifier-backward encoder pair, respectively, and if $\mathcal{B} \supseteq \{b_{Z|Y}\colon b_{Z|Y}(z|y) = c_{\hat{Y}|Z}(y|z)q_Z(z)/p_Y(y),\ q_Z \in \mathcal{Q}, c_{\hat{Y}|Z} \in \mathcal{C}\}$, *then*

$$\min_{\substack{e_{Z|X}\in\mathcal{E},c_{\hat{Y}|Z}\in\mathcal{C},q_Z\in\mathcal{Q}\\ (c_{\hat{Y}|Z},q_Z)\ \textit{consistent}}} \mathcal{L}_{\text{VIB}} \geq \min_{\substack{e_{Z|X}\in\mathcal{E},c_{\hat{Y}|Z}\in\mathcal{C},b_{Z|Y}\in\mathcal{B}\\ (c_{\hat{Y}|Z},b_{Z|Y})\ \textit{consistent}}} \mathcal{L}_{\text{VCEB}}. \tag{13b}$$

A fortiori, (13b) *continues to hold if VCEB is not constrained to a consistent classifier-backward encoder pair.*

Theorem 1 thus relates the cost functions of VIB and VCEB in certain well-defined scenarios, contingent on the size of the feasible sets $\mathcal{B}$ and $\mathcal{Q}$. If the variational approximations are implemented using NNs, then these bounds are thus contingent on the capacity of the NNs trained to represent the backward encoder in case of VCEB and the marginal in the case of VIB. A few clarifying statements are now in order.

First, it is easy to imagine scenarios in which the inequalities are strict. Trivially, this is the case for (13a) if $\mathcal{C}$ and $\mathcal{B}$, and for (13b) if $\mathcal{C}$ and $\mathcal{Q}$ do not contain a consistent pair. Furthermore, if the set relations in the respective conditions do not hold with equality, the optimization over the strictly larger set of, e.g., marginals in (13a), may yield strictly smaller values for the cost function $\mathcal{L}_{\text{VIB}}$.

Second, the condition that $\mathcal{B} \supseteq \{b_{Z|Y}\colon b_{Z|Y}(z|y) = c_{\hat{Y}|Z}(y|z)q_Z(z)/p_Y(y),\ q_Z \in \mathcal{Q}, c_{\hat{Y}|Z} \in \mathcal{C}\}$ is less restrictive than the condition stated in Assumption 1. This is because every backward encoder that is written as $b_{Z|Y}(z|y) = c'_{\hat{Y}|Z}(y|z)q'_Z(z)/p_Y(y)$ for $q'_Z \in \mathcal{Q}$ and $c'_{\hat{Y}|Z} \in \mathcal{C}$ satisfies trivially that $\sum_y b'_{Z|Y}(z|y)p_Y(y) = q'_Z(z)$. Thus, if one accepts Assumption 1 as reasonable for a fair comparison between VCEB and VIB, then one must also accept that the ordering provided in the theorem is mainly a consequence of the restriction to consistent pairs, and not to one of the optimization problems having access to a significantly larger feasible set.

Finally, if $\mathcal{C}$, $\mathcal{B}$, and $\mathcal{Q}$ are sufficiently large, i.e., if the NNs implementing the classifier, backward encoder, and marginal are sufficiently powerful, then both VCEB and VIB can be assumed to yield equally good approximations of the IB functional. To see this, let p_Z, $p_{Z|Y}$, and $p_{Y|Z}$ denote the marginal and conditional distributions induced by $e_{Z|X}$ and note that with (12) we get

$$\mathcal{L}_{\text{VIB}}(\beta) = \mathcal{L}_{\text{IB}}(\beta) + D\left(p_Z \| q_Z\right) + \beta E_{Z\sim p_Z}\left[D\left(p_{Y|Z}(\cdot|Z)\|c_{\hat{Y}|Z}(\cdot|Z)\right)\right] \tag{14a}$$

and

$$\begin{aligned}&\mathcal{L}_{\text{VCEB}}(\beta-1)\\ &= \mathcal{L}_{\text{IB}}(\beta) + E_{Y\sim p_Y}\left[D\left(p_{Z|Y}(\cdot|Y)\|b_{Z|Y}(\cdot|Y)\right)\right] + (\beta-1)E_{Z\sim p_Z}\left[D\left(p_{Y|Z}(\cdot|Z)\|c_{\hat{Y}|Z}(\cdot|Z)\right)\right].\end{aligned} \tag{14b}$$

Large $\mathcal{B}$ and $\mathcal{Q}$ render the second terms in both equations close to zero for all choices of $e_{Z|X}$ (see Example 2), while large $\mathcal{C}$ renders the last terms close to zero (see Example 1). Thus, in this case not only do we have $\mathcal{L}_{\text{VIB}}(\beta) \approx \mathcal{L}_{\text{VCEB}}(\beta-1) \approx \mathcal{L}_{\text{IB}}(\beta)$, but we also have that VCEB employs a consistent classifier-backward encoder pair by the fact that $b_{Z|Y} \approx p_{Z|Y}$ and $c_{\hat{Y}|Z} \approx p_{Y|Z}$. Thus, one may

argue that if the feasible sets are sufficiently large, the restriction to consistent pairs may not lead to significantly looser bounds.

5. Discussion

In this note we have compared the IB and CEB functionals and their respective variational approximations. While IB and CEB are shown to be equivalent, the variational approximations VIB and VCEB yield different results after optimization. Specifically, it was observed that using VCEB as a training objective for stochastic NNs outperforms VIB in terms of accuracy, adversarial robustness, and out-of-distribution detection (see Section 3.1 of [2]). In our analysis we have observed that, although in general there is no ordering between VIB and VCEB (Examples 1 and 2), the optimal values of the cost functions can be ordered if additional restrictions are imposed (Theorem 1). Specifically, if VCEB is constrained to a consistent classifier-backward encoder pair, then its optimal value cannot fall below the optimal value of VIB. If, in contrast, VIB is constrained, then the optimal value of VIB cannot fall below the optimal value of VCEB (constrained or unconstrained). Thus, as expected, adding restrictions weakens the optimization problem w.r.t. the unconstrained counterpart.

These results imply that the superiority of VCEB is not caused by enabling a tighter bound on the IB functional than VIB does. Furthermore, it was shown in Table 1 of [6] that VCEB, constrained to a consistent classifier-backward encoder pair, yields better classification accuracy and robustness against corruptions than the unconstrained VCEB objective. Since obviously

$$\min_{\substack{e_{Z|X}\in\mathcal{E},c_{\hat{Y}|Z}\in\mathcal{C},b_{Z|Y}\in\mathcal{B}\\(c_{\hat{Y}|Z},b_{Z|Y})\text{ consistent}}}\mathcal{L}_{\text{VCEB}}\geq\min_{e_{Z|X}\in\mathcal{E},c_{\hat{Y}|Z}\in\mathcal{C},b_{Z|Y}\in\mathcal{B}}\mathcal{L}_{\text{VCEB}}\tag{15}$$

the achievable tightness of a variational bound on the IB functional appears to be even negatively correlated with generalization performance in this set of experiments. (We note that [6] only reports constrained VCEB results for the largest NN models, and the constrained models perform slightly worse on robustness to adversarial examples than the unconstrained VCEB models of the same size.)

One may hypothesize, though, that VCEB is more amenable to optimization, in the sense that it achieves a tighter bound on the IB functional when encoder, classifier, and variational distributions are implemented and optimized using NNs. However, optimizing VCEB and VIB was shown to yield very similar results in terms of a lower bound on $I(X;Z)$ for several values of β, cf. Figure 4 of [2], which seems not to support above hypothesis.

We therefore conclude that the superiority of (constrained) VCEB is not due to it better approximating the IB functional. While the hypothesis that the optimized VCEB functional approximates the optimized IB functional better cannot be ruled out, we will now formulate an alternative hypothesis. Namely, that the VCEB cost function itself instills desirable properties in the encoder that would otherwise not be instilled when relying exclusively on the IB functional, cf. Section 5.4 of [13]. For example, neither IB nor the Minimum Necessary Information principle include a classifier $c_{\hat{Y}|Z}$ in their formulations. Thus, by the invariance of mutual information under bijections, there may be many encoders $e_{Z|X}$ in the feasible set $\mathcal{E}$ that lead to representations Z equivalent in terms of (1) and (2). Only few of these representations are useful in the sense that the information about the class Y can be extracted "easily". The variational approach of using a classifier to approximate $I(Y;Z)$, however, ensures that, among all encoders $e_{Z|X}$ that are equivalent under the IB principle, one is chosen such that there exists a classifier $c_{\hat{Y}|Z}$ in $\mathcal{C}$ that allows inferring the class variable Y from Z with low entropy: While the IB and Minimum Necessary Information principles ensure that Z is informative about Y, the variational approaches of VIB and VCEB ensure that this information can be accessed in practice. Regarding the observed superiority of VCEB over VIB, one may argue that a variational bound relying on a backward encoder instills properties in the latent representation Z that are preferable over those that are achieved by optimizing a variational bound relying on a marginal only.

In other words, VCEB and VIB are justified as cost functions for NN training even without recourse to the IB and Minimum Necessary Information principles. This does not say that the concept of compression, inherent in both of these principles, is not a useful guidance—whether compression and generalization are causally related is the topic of an ongoing debate to which we do not want to contribute in this work. Rather, we claim that variational approaches may yield desirable properties that go beyond compression and that may be overlooked when too much focus is put on the functionals that are approximated with these variational bounds.

In combination with the variational approach, the selection of feasible sets can also have profound impact on the properties of the representation Z. A representation Z is called disentangled if its distribution p_Z factorizes. Disentanglement can thus be measured by total correlation, i.e., the Kullback–Leibler divergence between p_Z and the product of its marginals Section 5 of the [11]. Achille and Soatto have shown that selecting $\mathcal{Q}$ in the optimization of VIB as a family of factorized marginals is equivalent to adding a total correlation term to the IB functional, effectively encouraging disentanglement, cf. Proposition 1 in [11]. Similarly, Amjad and Geiger note that selecting $\mathcal{B}$ in the optimization of VCEB as a family of factorized backward encoders encourages class-conditional disentanglement; i.e., it enforces a Naive Bayes structure on the representation Z, cf. Corollary 1 & Section 3.1 of [5]. To understand the implications of these observations, it is important to note that neither disentanglement nor class-conditional disentanglement are encouraged by the IB or CEB functionals. However, by appropriately selecting the feasible sets of VIB or VCEB, disentanglement and class-conditional disentanglement can be achieved. While we leave it to the discretion of the reader to decide whether disentanglement is desirable or not, we believe that it is vital to understand that disentanglement is an achievement of optimizing a variational bound over an appropriately selected feasible set, and not one of the principles based on which these variational approaches are motivated.

6. Proof of Theorem 1

We start with the first assertion. Assume that $e^{\text{VCEB}}_{Z|X}$, $b^{\text{VCEB}}_{Z|Y}$, and $c^{\text{VCEB}}_{\hat{Y}|Z}$ are the optimal encoder, backward encoder, and classifier in terms of the VCEB cost function under the assumption of consistency, i.e.,

$$\min_{\substack{e_{Z|X}\in\mathcal{E}, c_{\hat{Y}|Z}\in\mathcal{C}, b_{Z|Y}\in\mathcal{B}\\ (c_{\hat{Y}|Z}, b_{Z|Y}) \text{ consistent}}} \mathcal{L}_{\text{VCEB}} = E_{XY\sim p_{XY}}\left[D\left(e^{\text{VCEB}}_{Z|X}(\cdot|X)\|b^{\text{VCEB}}_{Z|Y}(\cdot|Y)\right)\right] - (\beta-1)H(Y) - (\beta-1)E_{XYZ\sim p_{XY}e^{\text{VCEB}}_{Z|X}}\left[\log c^{\text{VCEB}}_{\hat{Y}|Z}(Y|Z)\right] \quad (16)$$

where

$$c^{\text{VCEB}}_{\hat{Y}|Z}(y|z) = \frac{p_Y(y)b^{\text{VCEB}}_{Z|Y}(z|y)}{\sum_{y'} p_Y(y')b^{\text{VCEB}}_{Z|Y}(z|y')} = \frac{p_Y(y)b^{\text{VCEB}}_{Z|Y}(z|y)}{q'_Z(z)}. \quad (17)$$

Certainly, if $\mathcal{C}$ and $\mathcal{B}$ are such that they do not admit a consistent pair, then this minimum is infinity and the inequality holds trivially.

For the VIB optimization problem, we obtain

$$
\begin{aligned}
&\min_{e_{Z|X}\in\mathcal{E},c_{\hat{Y}|Z}\in\mathcal{C},q_Z\in\mathcal{Q}} \mathcal{L}_{\mathrm{VIB}} \\
&= \min_{e_{Z|X}\in\mathcal{E},c_{\hat{Y}|Z}\in\mathcal{C},q_Z\in\mathcal{Q}} E_{X\sim p_X}\left[D\left(e_{Z|X}(\cdot|X)\|q_Z\right)\right] - \beta H(Y) - \beta E_{XYZ\sim p_{XY}e_{Z|X}}\left[\log c_{\hat{Y}|Z}(Y|Z)\right] \\
&\overset{(a)}{=} \min_{e_{Z|X}\in\mathcal{E},c_{\hat{Y}|Z}\in\mathcal{C},q_Z\in\mathcal{Q}} E_{XZ\sim p_X e_{Z|X}}\left[\log\frac{e_{Z|X}(Z|X)}{q_Z(Z)}\right] - \beta H(Y) - \beta E_{XYZ\sim p_{XY}e_{Z|X}}\left[\log c_{\hat{Y}|Z}(Y|Z)\right] \\
&\overset{(b)}{=} \min_{e_{Z|X}\in\mathcal{E},c_{\hat{Y}|Z}\in\mathcal{C},q_Z\in\mathcal{Q}} E_{XYZ\sim p_{XY}e_{Z|X}}\left[\log\frac{e_{Z|X}(\cdot|X)c^{\mathrm{VCEB}}_{\hat{Y}|Z}(Y|Z)}{q_Z(Z)c^{\mathrm{VCEB}}_{\hat{Y}|Z}(Y|Z)}\right] - \beta H(Y) \\
&\quad - \beta E_{XYZ\sim p_{XY}e_{Z|X}}\left[\log c_{\hat{Y}|Z}(Y|Z)\right] \\
&\overset{(c)}{\leq} \min_{e_{Z|X}\in\mathcal{E},q_Z\in\mathcal{Q}} E_{XYZ\sim p_{XY}e_{Z|X}}\left[\log\frac{e_{Z|X}(\cdot|X)c^{\mathrm{VCEB}}_{\hat{Y}|Z}(Y|Z)}{q_Z(Z)c^{\mathrm{VCEB}}_{\hat{Y}|Z}(Y|Z)}\right] - \beta H(Y) - \beta E_{XYZ\sim p_{XY}e_{Z|X}}\left[\log c^{\mathrm{VCEB}}_{\hat{Y}|Z}(Y|Z)\right] \\
&\overset{(d)}{\leq} \min_{e_{Z|X}\in\mathcal{E}} E_{XYZ\sim p_{XY}e_{Z|X}}\left[\log\frac{e_{Z|X}(Z|X)c^{\mathrm{VCEB}}_{\hat{Y}|Z}(Y|Z)}{p_Y(Y)b^{\mathrm{VCEB}}_{Z|Y}(Z|Y)}\right] - \beta H(Y) - \beta E_{XYZ\sim p_{XY}e_{Z|X}}\left[\log c^{\mathrm{VCEB}}_{\hat{Y}|Z}(Y|Z)\right] \\
&\overset{(e)}{=} \min_{e_{Z|X}\in\mathcal{E}} E_{XY\sim p_{XY}}\left[D\left(e_{Z|X}(\cdot|X)\|b^{\mathrm{VCEB}}_{Z|Y}(\cdot|Y)\right)\right] + E_{XYZ\sim p_{XY}e_{Z|X}}\left[\log\frac{c^{\mathrm{VCEB}}_{\hat{Y}|Z}(Y|Z)}{p_Y(Y)}\right] \\
&\quad - \beta H(Y) - \beta E_{XYZ\sim p_{XY}e_{Z|X}}\left[\log c^{\mathrm{VCEB}}_{\hat{Y}|Z}(Y|Z)\right] \\
&\overset{(f)}{=} \min_{e_{Z|X}\in\mathcal{E}} E_{XY\sim p_{XY}}\left[D\left(e_{Z|X}(\cdot|X)\|b^{\mathrm{VCEB}}_{Z|Y}(\cdot|Y)\right)\right] - (\beta-1)H(Y) \\
&\quad - (\beta-1)E_{XYZ\sim p_{XY}e_{Z|X}}\left[\log c^{\mathrm{VCEB}}_{\hat{Y}|Z}(Y|Z)\right] \\
&\overset{(g)}{\leq} E_{XY\sim p_{XY}}\left[D\left(e^{\mathrm{VCEB}}_{Z|X}(\cdot|X)\|b^{\mathrm{VCEB}}_{Z|Y}(\cdot|Y)\right)\right] - (\beta-1)H(Y) - (\beta-1)E_{XYZ\sim p_{XY}e^{\mathrm{VCEB}}_{Z|X}}\left[\log c^{\mathrm{VCEB}}_{\hat{Y}|Z}(Y|Z)\right] \\
&= \min_{e_{Z|X}\in\mathcal{E},c_{\hat{Y}|Z}\in\mathcal{C},b_{Z|Y}\in\mathcal{B}} \mathcal{L}_{\mathrm{CEB}}
\end{aligned}
$$

where

- (a) follows by writing the KL divergence as an expectation of the logarithm of a ratio;
- (b) follows by multiplying both numerator and denominator in the first term with $c^{\mathrm{VCEB}}_{\hat{Y}|Z}$;
- (c) is because of the (potential) suboptimality of $c^{\mathrm{VCEB}}_{\hat{Y}|Z}$ for the VIB cost function;
- (d) is because $\mathcal{Q} \supseteq \{q_Z\colon q_Z(z) = \sum_y b_{Z|Y}(z|y)p_Y(y),\ b_{Z|Y} \in \mathcal{B}\}$, thus we may choose $q_Z = q'_Z$ where q'_Z is defined in (17); and because this particular choice may be suboptimal for the VIB cost function;
- (e) follows by splitting the logarithm
- (f) follows by noticing that $E_{XYZ\sim p_{XY}e_{Z|X}}\left[\log p_Y(Y)\right] = -H(Y)$
- (g) follows because $e^{\mathrm{VCEB}}_{Z|X}$ may be suboptimal for the VIB cost function.

Comparing the last line with (16) completes the proof of the first assertion.

We next turn to the second assertion. Assume that $e^{\mathrm{VIB}}_{Z|X}$, $c^{\mathrm{VIB}}_{\hat{Y}|Z}$, and q^{VIB}_Z are the optimal encoder, classifier, and marginal in terms of the VIB cost function under the assumption of consistency, i.e.,

$$
\min_{\substack{e_{Z|X}\in\mathcal{E},c_{\hat{Y}|Z}\in\mathcal{C},q_Z\in\mathcal{Q}\\ (c_{\hat{Y}|Z},q_Z)\text{ consistent}}} \mathcal{L}_{\mathrm{VIB}} := E_{X\sim p_X}\left[D\left(e^{\mathrm{VIB}}_{Z|X}(\cdot|X)\|q^{\mathrm{VIB}}_Z\right)\right] - \beta H(Y) - \beta E_{XYZ\sim p_{XY}e^{\mathrm{VIB}}_{Z|X}}\left[\log c^{\mathrm{VIB}}_{\hat{Y}|Z}(Y|Z)\right] \quad (18)
$$

where

$$p_Y(y) = \sum_z c^{\mathrm{VIB}}_{\hat{Y}|Z}(y|z) q^{\mathrm{VIB}}_Z(z). \tag{19}$$

Again, if $\mathcal{C}$ and $\mathcal{Q}$ are such that they do not admit a consistent pair, then this minimum is infinity and the inequality holds trivially.

For the VCEB optimization problem, we obtain

$$\begin{aligned}
&\min_{\substack{e_{Z|X}\in\mathcal{E}, c_{\hat{Y}|Z}\in\mathcal{C}, b_{Z|Y}\in\mathcal{B} \\ (c_{\hat{Y}|Z}, b_{Z|Y}) \text{ consistent}}} \mathcal{L}_{\mathrm{VCEB}} + (\beta-1)H(Y) \\
&= \min_{\substack{e_{Z|X}\in\mathcal{E}, c_{\hat{Y}|Z}\in\mathcal{C}, b_{Z|Y}\in\mathcal{B} \\ (c_{\hat{Y}|Z}, b_{Z|Y}) \text{ consistent}}} E_{XY\sim p_{XY}}\left[D\left(e_{Z|X}(\cdot|X)\|b_{Z|Y}(\cdot|Y)\right)\right] - (\beta-1)E_{XYZ\sim p_{XY}e_{Z|X}}\left[\log c_{\hat{Y}|Z}(Y|Z)\right] \\
&\overset{(a)}{=} \min_{\substack{e_{Z|X}\in\mathcal{E}, c_{\hat{Y}|Z}\in\mathcal{C}, b_{Z|Y}\in\mathcal{B} \\ (c_{\hat{Y}|Z}, b_{Z|Y}) \text{ consistent}}} E_{XYZ\sim p_{XY}e_{Z|X}}\left[\log \frac{e_{Z|X}(Z|X)}{b_{Z|Y}(Y|Z)}\right] - (\beta-1)E_{XYZ\sim p_{XY}e_{Z|X}}\left[\log c_{\hat{Y}|Z}(Y|Z)\right] \\
&\overset{(b)}{=} \min_{e_{Z|X}\in\mathcal{E}, c_{\hat{Y}|Z}\in\mathcal{C}} \min_{b_{Z|Y}\in\mathcal{B}\cap\left\{b'_{Z|Y}:\ \frac{p_Y(y)b'_{Z|Y}(z|y)}{\sum_{y'}p_Y(y')b'_{Z|Y}(z|y')}=c_{\hat{Y}|Z}(y|z)\right\}} E_{XYZ\sim p_{XY}e_{Z|X}}\left[\log \frac{e_{Z|X}(Z|X)}{b_{Z|Y}(Y|Z)}\right] \\
&\qquad - (\beta-1)E_{XYZ\sim p_{XY}e_{Z|X}}\left[\log c_{\hat{Y}|Z}(Y|Z)\right] \\
&\overset{(c)}{\leq} \min_{e_{Z|X}\in\mathcal{E}} \min_{b_{Z|Y}\in\mathcal{B}\cap\left\{b'_{Z|Y}:\ \frac{p_Y(y)b'_{Z|Y}(z|y)}{\sum_{y'}p_Y(y')b'_{Z|Y}(z|y')}=c^{\mathrm{VIB}}_{\hat{Y}|Z}(y|z)\right\}} E_{XYZ\sim p_{XY}e_{Z|X}}\left[\log \frac{e_{Z|X}(Z|X)}{b_{Z|Y}(Y|Z)}\right] \\
&\qquad - (\beta-1)E_{XYZ\sim p_{XY}e_{Z|X}}\left[\log c^{\mathrm{VIB}}_{\hat{Y}|Z}(Y|Z)\right] \\
&\overset{(d)}{\leq} \min_{e_{Z|X}\in\mathcal{E}} E_{XYZ\sim p_{XY}e_{Z|X}}\left[\log \frac{e_{Z|X}(Z|X)}{c^{\mathrm{VIB}}_{\hat{Y}|Z}(Y|Z)q^{\mathrm{VIB}}_Z(Z)}\right] - H(Y) - (\beta-1)E_{XYZ\sim p_{XY}e_{Z|X}}\left[\log c^{\mathrm{VIB}}_{\hat{Y}|Z}(Y|Z)\right] \\
&= \min_{e_{Z|X}\in\mathcal{E}} E_{XZ\sim p_X e_{Z|X}}\left[\log \frac{e_{Z|X}(Z|X)}{q^{\mathrm{VIB}}_Z(Z)}\right] - H(Y) - \beta E_{XYZ\sim p_{XY}e_{Z|X}}\left[\log c^{\mathrm{VIB}}_{\hat{Y}|Z}(Y|Z)\right] \\
&\overset{(e)}{\leq} E_{XZ\sim p_X e^{\mathrm{VIB}}_{Z|X}}\left[\log \frac{e^{\mathrm{VIB}}_{Z|X}(Z|X)}{q^{\mathrm{VIB}}_Z(Z)}\right] - H(Y) - \beta E_{XYZ\sim p_{XY}e^{\mathrm{VIB}}_{Z|X}}\left[\log c^{\mathrm{VIB}}_{\hat{Y}|Z}(Y|Z)\right] \\
&= \min_{\substack{e_{Z|X}\in\mathcal{E}, c_{\hat{Y}|Z}\in\mathcal{C}, q_Z\in\mathcal{Q} \\ (c_{\hat{Y}|Z}, q_Z) \text{ consistent}}} \mathcal{L}_{\mathrm{VIB}} + (\beta-1)H(Y)
\end{aligned}$$

where

- (a) follows by writing the KL divergence as an expectation of the logarithm of a ratio;
- (b) follows by the assumption that the VCEB problem is constrained to a consistent classifier-backward encoder pair, and from (11);
- (c) is because of the (potential) suboptimality of $c^{\mathrm{VIB}}_{\hat{Y}|Z}$ for the VCEB cost function;
- (d) follows by adding and subtracting $H(Y)$; by choosing $b^{\mathrm{VIB}}_{Z|Y} = c^{\mathrm{VIB}}_{\hat{Y}|Z} q^{\mathrm{VIB}}_Z / p_Y$, which is possible because $\mathcal{B} \supseteq \{b_{Z|Y}\colon b_{Z|Y}(z|y) = c_{\hat{Y}|Z}(y|z)q_Z(z)/p_Y(y),\ q_Z \in \mathcal{Q}, c_{\hat{Y}|Z} \in \mathcal{C}\}$; and by the fact that this particular choice may be suboptimal for the VCEB cost function;
- (e) follows because $e^{\mathrm{VIB}}_{Z|X}$ may be suboptimal for the VCEB cost function.

This completes the proof. □

Author Contributions: Conceptualization, formal analysis, validation, writing: B.C.G. and I.S.F.; Proof of Theorem 1: B.C.G. Both authors have read and agreed to the published version of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: The work of Bernhard C. Geiger was supported by the iDev40 project. The iDev40 project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 783163. The JU receives support from the European Union's Horizon 2020 research and innovation programme. It is co-funded by the consortium members, grants from Austria, Germany, Belgium, Italy, Spain and Romania. The information and results set out in this publication are those of the authors and do not necessarily reflect the opinion of the ECSEL Joint Undertaking. The Know-Center is funded within the Austrian COMET Program-Competence Centers for Excellent Technologies - under the auspices of the Austrian Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology, the Austrian Federal Ministry of Digital and Economic Affairs, and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tishby, N.; Pereira, F.C.; Bialek, W. The Information Bottleneck Method. In Proceedings of the Allerton Conference on Communication, Control, and Computing, Monticello, IL, USA, 22–24 September 1999; pp. 368–377.
2. Fischer, I. The Conditional Entropy Bottleneck. *Entropy* **2020**, *22*, 999. [CrossRef]
3. Wyner, A. The common information of two dependent random variables. *IEEE Trans. Inf. Theory* **1975**, *21*, 163–179. [CrossRef]
4. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 1st ed.; John Wiley & Sons, Inc.: New York, NY, USA, 1991.
5. Amjad, R.A.; Geiger, B.C. Class-Conditional Compression and Disentanglement: Bridging the Gap between Neural Networks and Naive Bayes Classifiers. *arXiv* **2019**, arXiv:1906.02576.
6. Fischer, I.; Alemi, A.A. CEB Improves Model Robustness. *Entropy* **2020**, *22*, 1081. [CrossRef]
7. Alemi, A.A.; Fischer, I.; Dillon, J.V.; Murphy, K. Deep Variational Information Bottleneck. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
8. Poole, B.; Ozari, S.; van den Oord, A.; Alemi, A.A.; Tucker, G. On Variational Bounds of Mutual Information. In Proceedings of the International Conference on Machine Learning (ICML), Long Beach, CA, USA, 10–15 June 2019; pp. 5171–5180.
9. Belghazi, M.I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; Hjelm, D. Mutual Information Neural Estimation. In Proceedings of the International Conference on Machine Learning (ICML), Stockholm, Sweden, 10–15 July 2018; pp. 531–540.
10. Kolchinsky, A.; Tracey, B.D.; Wolpert, D.H. Nonlinear Information Bottleneck. *Entropy* **2019**, *21*, 1181. [CrossRef]
11. Achille, A.; Soatto, S. Information Dropout: Learning Optimal Representations Through Noisy Computation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 2897–2905. [CrossRef] [PubMed]
12. Wieczorek, A.; Roth, V. On the difference between the Information Bottleneck and the Deep Information Bottleneck. *Entropy* **2020**, *22*, 131. [CrossRef]
13. Amjad, R.A.; Geiger, B.C. Learning Representations for Neural Network-Based Classification Using the Information Bottleneck Principle. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2225–2239. [CrossRef] [PubMed]
14. Kingma, D.P.; Welling, M. Auto-encoding variational Bayes. In Proceedings of the International Conference on Learning Representations (ICLR), Banff, AB, Canada, 14–16 April 2014.