The Variational Deficiency Bottleneck

Pradeep Kr. Banerjee
MPI MiS
Email: pradeep@mis.mpg.de

Guido Montúfar
UCLA and MPI MiS
Email: montufar@math.ucla.edu

Abstract—We introduce a bottleneck method for learning data representations based on information deficiency, rather than the more traditional information sufficiency. A variational upper bound allows us to implement this method efficiently. The bound itself is bounded above by the variational information bottleneck objective, and the two methods coincide in the regime of single-shot Monte Carlo approximations. The notion of deficiency provides a principled way of approximating complicated channels by relatively simpler ones. We show that the deficiency of one channel with respect to another has an operational interpretation in terms of the optimal risk gap of decision problems, capturing classification as a special case. Experiments demonstrate that the deficiency bottleneck can provide advantages in terms of minimal sufficiency as measured by information bottleneck curves, while retaining robust test performance in classification tasks.

Index Terms—Blackwell sufficiency, deficiency, information bottleneck, synergy, robustness

I. INTRODUCTION

The information bottleneck (IB) is an approach to learning data representations based on a notion of minimal sufficiency. The general idea is to map an input source to an intermediate representation that retains as little information as possible about the input (*minimality*), but preserves as much information as possible in relation to a target variable of interest (*sufficiency*). See Fig. 1. For example, in a classification problem, the target variable could be the class label of the input data. In a reconstruction problem, the target variable could be a denoised reconstruction of the input. Intuitively, a representation which is minimal in relation to a given task, will discard nuisances in the inputs that are irrelevant to the task, and hence distill more meaningful information and allow for a better generalization. The IB methods [1]–[4] have found numerous applications in representation learning, clustering, classification, generative modeling, model selection and analysis in deep neural networks, among others (see, e.g., [5]–[10]).

In the traditional IB paradigm, minimality and sufficiency are measured in terms of the mutual information. Computing the mutual information can be challenging in practice. Recent works have formulated more tractable functions by way of variational bounds on the mutual information [11]–[14]. Instead of maximizing the sufficiency term of the IB, we formulate a new bottleneck method that minimizes deficiency. Deficiencies provide a principled way of approximating complex channels
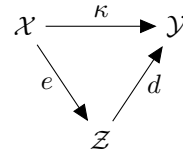
Fig. 1: The bottleneck paradigm: The general idea of a bottleneck method is to first map an input $X \in \mathcal{X}$ to an intermediate representation $Z \in \mathcal{Z}$, and then map Z to an output $Y \in \mathcal{Y}$. We call the mappings, resp., an encoder (e) and a decoder (d). In general, the true channel κ is unknown, and only accessible through a set of training examples. We would like to obtain an approximation of κ.

by relatively simpler ones and have a rich heritage in the theory of comparison of statistical experiments [15]–[17]. From this angle, the formalism of deficiencies has been used to obtain bounds on optimal risk gaps of statistical decision problems. As we show, the deficiency bottleneck minimizes a regularized risk gap. Moreover, the proposed method has an immediate variational formulation that can be easily implemented as a modification of the variational information bottleneck (VIB) [13]. In fact, both methods coincide in the limit of single-shot Monte Carlo approximations. We call our method the variational deficiency bottleneck (VDB).

Experiments on basic data sets show that the VDB is able to obtain more compressed representations than the VIB while retaining the same level of sufficiency. Training with the VDB also improves out-of-distribution robustness over the VIB as we demonstrate on two benchmark datasets, the MNIST-C [18] and the CIFAR-10-C [19].

We describe the details of our method in Section II. We elaborate on the theory of deficiencies in Section III. Experimental results with the VDB are presented in Section IV. We use notation that is standard in information theory [20].

II. THE VARIATIONAL DEFICIENCY BOTTLENECK

Let X denote an observation or *input* variable and Y an *output* variable of interest and let $\mathcal{X}$, $\mathcal{Y}$ denote, resp., the space of possible inputs and outputs. Let $p(x, y) = \pi(x)\kappa(y|x)$ be the true joint distribution, where the conditional distribution or *channel* $\kappa(y|x)$ describes how the output depends on the input. We consider the situation where the true channel is unknown, but we are given a set of N independent and identically distributed (i.i.d.) samples $(x^{(i)}, y^{(i)})_{i=1}^{N}$ from p. Our goal

is to use this data to learn a more structured version of the channel κ, by first "compressing" the input X to an intermediate *representation* variable Z and subsequently mapping the representation back to the output Y. The presence of an intermediate representation can be regarded as a bottleneck, a model selection problem, or as a regularization strategy.

We define an encoder and a decoder model using two parameterized families of channels $e(z|x)$ and $d(y|z)$, respectively. The encoder-decoder pair induces a model $\widehat{\kappa}(y|x) = \int d(y|z)e(z|x)\, dz$. Equivalently, we write $\widehat{\kappa} = d \circ e$. Given a representation, we want the decoder to be as powerful as the original channel κ in terms of ability to recover the output. The deficiency of a decoder d w.r.t. κ quantifies the extent to which any pre-processing at the input (by way of randomized encodings) fails to approximate κ. Let $\mathsf{M}(\mathcal{X};\mathcal{Y})$ denote the space of all channels from $\mathcal{X}$ to $\mathcal{Y}$. We define the deficiency of d w.r.t. κ as follows:

Definition 1 (Deficiency). Given a channel $\kappa \in \mathsf{M}(\mathcal{X};\mathcal{Y})$ from $\mathcal{X}$ to $\mathcal{Y}$, and a decoder $d \in \mathsf{M}(\mathcal{Z};\mathcal{Y})$ from some $\mathcal{Z}$ to $\mathcal{Y}$, the *deficiency of d w.r.t. κ* is defined as

$$\delta^\pi(d,\kappa) = \min_{e \in \mathsf{M}(\mathcal{X};\mathcal{Z})} D_{\mathrm{KL}}(\kappa \| d \circ e | \pi). \quad (1)$$

Here $D_{\mathrm{KL}}(\cdot \| \cdot | \cdot)$ is the conditional KL divergence [20], and π is an input distribution over $\mathcal{X}$. The definition is similar in spirit to Lucien Le Cam's notion of weighted deficiencies of one channel w.r.t. another [16], [17, Section 6.2] and its recent generalizations [21].

We propose to train the model by minimizing the deficiency of d w.r.t. κ subject to a regularization that limits the *rate* $I(Z;X)$, i.e., the mutual information between the representation and the raw inputs. We call our method the *deficiency bottleneck* (DB). The DB minimizes the following objective over all tuples $(e \in \mathsf{M}(\mathcal{X};\mathcal{Z}), d \in \mathsf{M}(\mathcal{Z};\mathcal{Y}))$:

$$\mathcal{L}_{DB}^\beta(e,d) := \delta^\pi(d,\kappa) + \beta I(Z;X). \quad (2)$$

The parameter $\beta \geq 0$ allows us to adjust the level of regularization.

For any distribution $r(z)$, the rate term admits a simple variational upper bound [20]:

$$I(Z;X) \leq \int p(x,z) \log \frac{e(z|x)}{r(z)} \, dx\, dz. \quad (3)$$

Let $\hat{p}_{\mathrm{data}}$ be the empirical distribution of the data (input-output pairs). By noting that $\delta^\pi(d,\kappa) \leq D_{\mathrm{KL}}(\kappa \| d \circ e | \pi)$ for any $e \in \mathsf{M}(\mathcal{X};\mathcal{Z})$, and ignoring data-dependent constants, we obtain the following optimization objective which we call the *variational deficiency bottleneck* (VDB) objective:

$$\mathcal{L}_{VDB}^\beta(e,d) := \mathbb{E}_{(x,y)\sim \hat{p}_{\mathrm{data}}} \left[-\log \int d(y|z)e(z|x)dz \right.$$
$$\left. + \beta D_{\mathrm{KL}}(e(Z|x)\|r(Z)) \right]. \quad (4)$$

The computation is simplified by defining $r(z)$ to be a standard multivariate Gaussian distribution $\mathcal{N}(0,I)$, and using an encoder of the form $e(z|x) = \mathcal{N}(z|f_\phi(x))$, where f_ϕ is a neural network that outputs the parameters of a Gaussian

distribution. Using the reparameterization trick [22], [23], we then write $e(z|x)dz = p(\epsilon)d\epsilon$, where $z = f(x,\varepsilon)$ is a function of x and the realization ϵ of a standard normal distribution. This allows us to do stochastic backpropagation through a single sample z. The KL term in (4) admits an analytic expression when $r(z)$ and the encoder are Gaussian. We train the model by minimizing the following empirical objective over all tuples $(e \in \mathsf{M}(\mathcal{X};\mathcal{Z}), d \in \mathsf{M}(\mathcal{Z};\mathcal{Y}))$:

$$\frac{1}{N}\sum_{i=1}^N \left[-\log\left(\frac{1}{M}\sum_{j=1}^M [d(y^{(i)}|f(x^{(i)},\epsilon^{(j)}))]\right) \right.$$
$$\left. + \beta D_{\mathrm{KL}}(e(Z|x^{(i)})\|r(Z)) \right]. \quad (5)$$

For training, we choose a mini-batch size of $N = 100$. For estimating the expectation inside the log, we use $M = 1, 3, 6, 12$ Monte Carlo samples from the encoding distribution.

We note that the Variational Information Bottleneck (VIB) [13] leads to a similar-looking objective function, with the only difference that the sum over j in (5) is *outside* of the logarithm. By Jensen's inequality, the VIB loss is an upper bound to our loss. If one uses a single sample from the encoding distribution (i.e., $M = 1$), the VDB and the VIB empirical objective functions coincide. For a large enough mini-batch size, e.g., $N = 100$, taking $M = 1$ is sufficient to estimate the VIB objective [13]. This is the standard setting as presented in [13] that we want to compare with. In the case of the VDB, on the other hand, the mini-batch size N and M are *not* exchangeable, since the expectation is inside the log function.

To better understand the behavior of the VDB optimization (5), we adopt two training strategies:

- a *oneshot* strategy where the encoder and decoder networks are updated simultaneously, and
- a *sequential* strategy where the encoder network is updated for k steps before alternating to one decoder update. We choose $k = 5, 10, 20$.

The idea of using the sequential strategy is to better approximate the deficiency which involves an optimization over the encoder (see Definition 1).

III. TWO PERSPECTIVES ON THE DEFICIENCY BOTTLENECK

In this section, we present two different perspectives on the deficiency bottleneck, namely, a decision-theoretic perspective and an information decomposition perspective.

In Section III-A, we review the notions of information sufficiency and deficiency through the lens of Blackwell-Le Cam decision theory [15]–[17]. We formulate the learning task as a decision problem and give an operational characterization of the deficiency $\delta^\pi(d,\kappa)$ as the gap in the expected losses of optimal decision rules when using the channel d rather than κ.

In Section III-B, we review the classical IB and our DB objective through the lens of nonnegative mutual information decompositions [24]–[26]. This leads us to a new interpretation

of the IB as a *Unique Information Bottleneck* and also sheds light on the difference between the IB and DB formulations.

A. A decision-theoretic perspective

1) Blackwell sufficiency and channel deficiency: In a seminal paper [15], David Blackwell asked the following question: Suppose that a learner has a finite set of possible actions and she wishes to make an optimal decision to minimize a loss depending on the value of some random variable Y and her chosen action. If the learner cannot observe Y directly before choosing her action and has to pick between two channels with the *common input* Y, which one should she prefer? Blackwell introduced an ordering that compares channels by the minimal expected loss or risk that a learner incurs when her decisions are based on the channel outputs. He then showed that such an ordering can be equivalently characterized in terms of a purely probabilistic relation between the channels: The learner will always prefer one channel over another if and only if the latter is an *output-degraded* version of the former, in the sense that she can simulate a single use of the latter by randomizing at the output of the former.

Very recently, Nasser [27] asked the same question, only now the learner has to choose between two channels with a *common output* alphabet. Nasser introduced the *input-degraded* ordering and gave a characterization of input-degradedness that is similar to Blackwell's ordering [15].

Definition 2 (Blackwell sufficiency). Given two channels, $\kappa \in \mathsf{M}(\mathcal{X}; \mathcal{Y})$ and $d \in \mathsf{M}(\mathcal{Z}; \mathcal{Y})$, κ is *input-degraded* from d, denoted $d \succeq_{\mathcal{Y}} \kappa$, if $\kappa = d \circ e$ for some $e \in \mathsf{M}(\mathcal{X}; \mathcal{Z})$. We say that d is *input Blackwell sufficient* for κ if $d \succeq_{\mathcal{Y}} \kappa$.

Stated in another way, d is input Blackwell sufficient for κ if d can be reduced to κ by applying a randomization e at its input so that $d \circ e = \kappa$. Blackwell sufficiency induces only a preorder on the set of all channels with a common output alphabet. In practice, most channels are uncomparable, i.e., one cannot be reduced to another by a randomization. When such is the case, the deficiency quantifies how far the true channel κ is from being a randomization (by way of any input encoding e) of the decoder d.

2) Deficiency as an optimal risk gap: We formulate a learning task as a decision problem and show that the deficiency quantifies the gap in the optimal risks when using the channel d rather than κ.

Let $\mathbb{P}_Y$ be the set of all distributions on $\mathcal{Y}$. In the following, we assume that $\mathcal{X}$ and $\mathcal{Y}$ are finite. For every $x \in \mathcal{X}$, define $\kappa_x \in \mathbb{P}_Y$ as $\kappa_x(y) = \kappa(y|x)$, $\forall y \in \mathcal{Y}$. Consider the following decision problem between a learner and Nature: Nature draws $x \sim \pi$ and $y \sim \kappa_x$. The learner observes x and proposes a distribution $q_x \in \mathbb{P}_Y$ that expresses her uncertainty about the true value y. The quality of a prediction q_x in relation to y is measured by the *log-loss* function $\varsigma(y, q_x) := -\log q_x(y)$. The log-loss is an instance of a "strictly proper" loss function that enjoys nice properties such as the uniqueness of the optimum; see, e.g., [28].

Ideally, the prediction q_x should be as close as possible to the true conditional distribution κ_x. This is achieved by minimizing the expected loss $L(\kappa_x, q_x) := \mathbb{E}_{y \sim \kappa_x} \varsigma(y, q_x)$, for all $x \in \mathcal{X}$. Define the *Bayes act* against κ_x as the optimal prediction $q_x^* := \arg\min_{q_x \in \mathbb{P}_Y} L(\kappa_x, q_x)$, and the *Bayes risk* for the distribution $P_{XY} = \pi \times \kappa$ as $R(P_{XY}, \varsigma) := \mathbb{E}_{x \sim \pi} L(\kappa_x, q_x^*)$. For the log-loss, the Bayes act is $q_x^* = \kappa_x$ and hence, the Bayes risk is

$$R(P_{XY}, \varsigma) = \mathbb{E}_{x \sim \pi} \mathbb{E}_{y \sim \kappa_x} \left[-\log \kappa_x(y) \right] = H(Y|X).$$

Given a channel $d \in \mathsf{M}(\mathcal{Z}; \mathcal{Y})$, we want a representation $z \in \mathcal{Z}$ of x (output by some encoder), so that the outputs of d match those of the true channel κ. Let $C = \mathrm{conv}(\{d_z : z \in \mathcal{Z}\}) \subset \mathbb{P}_Y$ be the convex hull of the points $\{d_z\}_{z \in \mathcal{Z}} \in \mathbb{P}_Y$. The Bayes act against κ_x is $q_{x_d}^* := \arg\min_{q_x \in C} \mathbb{E}_{y \sim \kappa_x} \left[-\log q_x(y) \right]$. $q_{x_d}^*$ is the rI-projection of κ_x to the convex set $C \subset \mathbb{P}_Y$ [29]. Such a projection exists but is not necessarily unique. If non-unique, we arbitrarily select one of the minimizers as the Bayes act. The associated Bayes risk is

$$R_d(P_{XY}, \varsigma) := \mathbb{E}_{x \sim \pi} \mathbb{E}_{y \sim \kappa_x} \left[-\log q_{x_d}^*(y) \right].$$

The next Proposition 3 states that the gap in the Bayes risks, $\Delta R := R_d(P_{XY}, \varsigma) - R(P_{XY}, \varsigma)$, when making a decision based on Z vs. X is just the deficiency.

Proposition 3 (Deficiency quantifies the optimal risk gap for the log-loss). $\delta^\pi(d, \kappa) = \Delta R$.

Proof. The proof follows from noting that

$$\begin{aligned}
\Delta R &= \sum_{x \in \mathcal{X}} \pi(x) \min_{q_x \in C \subset \mathbb{P}_Y} D_{\mathrm{KL}}(\kappa_x \| q_x) \\
&= \min_{e \in \mathsf{M}(\mathcal{X}; \mathcal{Z})} \sum_{x \in \mathcal{X}} \pi(x) D_{\mathrm{KL}}(\kappa_x \| d \circ e_x) \\
&= \min_{e \in \mathsf{M}(\mathcal{X}; \mathcal{Z})} D_{\mathrm{KL}}(\kappa \| d \circ e | \pi) = \delta^\pi(d, \kappa).
\end{aligned}$$

∎

B. An information decomposition perspective

1) IB as Unique Information Bottleneck: A quantity that is similar in spirit to the deficiency is the Unique Information (UI) [24]:

Definition 4 (Unique information). Let $(Y, X, Z) \sim P$. The unique information that X conveys about Y w.r.t. Z is

$$UI(Y; X \backslash Z) := \min_{Q \in \Delta_P} I_Q(Y; X | Z), \tag{6}$$

where the subscript Q denotes the joint distribution on which the mutual information is evaluated, and

$$\begin{aligned}
\Delta_P := \big\{ Q \in \mathbb{P}_{\mathcal{Y} \times \mathcal{X} \times \mathcal{Z}} : &\, Q_{YX}(y, x) = P_{YX}(y, x), \\
&\, Q_{YZ}(y, z) = P_{YZ}(y, z) \big\} \tag{7}
\end{aligned}$$

is the set of joint distributions of (Y, X, Z) that have the same marginals on (Y, X) and (Y, Z) as P.

While the deficiency quantifies a deviation from the *input-degraded* order, the UI quantifies a deviation from the *output-degraded* order [25]. Note, however, that the vanishing sets

of $\delta^\pi(d,\kappa)$ and $UI(Y;X\backslash Z)$ are not equivalent as the next example shows.

Example 5. Let $\mathcal{Y} = \{0,1,\mathsf{e}\}$, and $\mathcal{X} = \mathcal{Z} = \{0,1,\mathsf{e}\}$. Let $P = P_Y \times P_{X|Y} \times P_{Z|X}$ where $P_Y \sim \text{Bernoulli}(\frac{1}{2})$ and $P_{X|Y}$ and $P_{Z|X}$ are symmetric erasure channels with erasure probabilities $\frac{1}{6}$ and $\frac{1}{5}$, resp. Recall that a symmetric erasure channel from $\mathcal{Y}$ to $\mathcal{X}$ with erasure probability $\epsilon \in [0,1]$ has transition probabilities: $P_{X|Y}(\mathsf{e}|0) = P_{X|Y}(\mathsf{e}|1) = \epsilon$, $P_{X|Y}(0|0) = P_{X|Y}(1|1) = 1 - \epsilon$. For the distribution P, we have $UI(Y;X\backslash Z) = I(Y;X|Z) = \frac{1}{6} > 0$. On the other hand, the induced "reverse" erasure channels $P_{Y|X} = \kappa$ and $P_{Y|Z} = d$ are identical. Thus, $\delta^\pi(d,\kappa) = 0$.

In [24], the value $UI(Y;X\backslash Z)$ is interpreted as the information about Y that is known to X but unknown to Z. This interpretation is motivated by Blackwell's result [15]: whenever $UI(Y;X\backslash Z) > 0$, there exists a decision problem in which it is better to know X than to know Z. Moreover, this induces a decomposition of the mutual information between Y and (X,Z) into four terms:

$$I(Y;XZ) = UI(Y;X\backslash Z) + SI(Y;X,Z)$$
$$+ UI(Y;Z\backslash X) + CI(Y;X,Z). \tag{8}$$

The quantity

$$SI(Y;X,Z) := I(Y;X) - UI(Y;X\backslash Z) \tag{9}$$

is interpreted as *shared* or *redundant* information, i.e., information about Y that is known in common to both X and Z, and the quantity

$$CI(Y;X,Z) := I(Y;X|Z) - UI(Y;X\backslash Z) \tag{10}$$

is interpreted as *complementary* or *synergistic* information, i.e., the information about Y that materializes only when X and Z act jointly.

Example 6. If X and Z are independent binary random variables, and $Y = \text{XOR}(X,Z)$, then $CI(Y;X,Z) = 1$, while $SI(Y;X,Z) = UI(Y;X\backslash Z) = UI(Y;Z\backslash X) = 0$. This is an instance of a purely *synergistic* interaction.

If Y, X, Z are uniformly distributed binary random variables with $Y = X = Z$, then $SI(Y;X,Z) = 1$, while $CI(Y;X,Z) = UI(Y;X\backslash Z) = UI(Y;Z\backslash X) = 0$. This is an instance of a purely *redundant* interaction.

If X, Z are independent binary random variables, and $Y = (X,Z)$, then $UI(Y;X\backslash Z) = UI(Y;Z\backslash X) = 1$, while $SI(Y;X,Z) = CI(Y;X,Z) = 0$. This is an instance of a purely *unique* interaction.

For some probability distributions P with special structure, the decomposition (8) can be computed analytically [24].

Lemma 7. Let $Q^0 := P_{YX} \times e_{Z|X}$ for some $e \in \mathsf{M}(\mathcal{X};\mathcal{Z})$. Then $UI_{Q^0}(Y;X\backslash Z) = I_{Q^0}(Y;X|Z)$, $SI_{Q^0}(Y;X,Z) = I(Y;Z)$ and $UI_{Q^0}(Y;Z\backslash X) = CI_{Q^0}(Y;X,Z) = 0$.

Proof. The distribution Q^0 defines a Markov chain $Y - X - Z$, which implies that

$$I_{Q^0}(Y;Z|X) = 0 \le \min_{Q \in \Delta_P} I_Q(Y;Z|X).$$

Hence, Q^0 solves the optimization problem (6). ∎

In the setting of the Lemma, $CI_{Q^0}(Y;X,Z) = 0$ and therefore Q^0 is a *zero-synergy distribution*.

The information decomposition leads us to a new interpretation of the IB as a *Unique Information Bottleneck*. To see this we first make the following definition.

Definition 8 (IB curve). The IB curve is defined as follows [1], [3], [30]:

$$B(r) := \max\{I(Z;Y) : I(Z;X) \le r, \ r \ge 0\}. \tag{11}$$

Here the maximization is over all random variables Z satisfying the Markov condition $Y - X - Z$ with fixed $(X,Y) \sim P$ and it suffices to restrict the size of $\mathcal{Z}$ to $|\mathcal{X}|$.

The IB curve is concave and monotonically nondecreasing [1], [30]. We can explore the IB curve by solving the following optimization problem:

$$\min_{e \in \mathsf{M}(\mathcal{X};\mathcal{Z})} \left[UI_{Q^0}(Y;X\backslash Z) + \beta I_{Q^0}(Z;X) \right]. \tag{12}$$

Here Q^0 is a zero-synergy distribution and $\beta \in [0,1]$ is a Lagrange multiplier. Equation (12) has the flavor of a rate-distortion problem [3] where the term $UI_{Q^0}(Y;X\backslash Z) = I_{Q^0}(Y;X|Z)$ is interpreted as the average *distortion* and $I_{Q^0}(Z;X)$ as the *rate*. Classically, the IB is formulated as the maximization of $I_{Q^0}(Y;Z) - \beta I_{Q^0}(Z;X)$ [2]. By Lemma 7, we have that $UI_{Q^0}(Y;X\backslash Z) = I_{Q^0}(Y;X|Z) = I_{Q^0}(Y;X) - I_{Q^0}(Y;Z)$. Since $I_{Q^0}(Y;X) = I_P(X;Y)$ is constant, it follows that the Unique Information Bottleneck defined by (12) is equivalent to the classical IB. Each point on the IB curve satisfies the Markov condition $Y - X - Z$ which implies that the solution is always constrained to have *zero synergy* about the output Y.

Like the UI, the deficiency also induces an information decomposition as we show next.

2) Deficiency induces an information decomposition: We first propose a general construction that forms the basis of an information decomposition satisfying (9)-(8) (proved in the Appendix):

Proposition 9. Let $(Y,X,Z) \sim P$ and let δ^X be a nonnegative function defined on the simplex $\mathbb{P}_{\mathcal{Y}\times\mathcal{X}\times\mathcal{Z}}$ that satisfies the bound:

$$0 \le \delta^X(P) \le \min\{I(Y;X), I(Y;X|Z)\}. \tag{13}$$

Let $\mathcal{X}' = \mathcal{Z}$, $\mathcal{Z}' = \mathcal{X}$, and define a function $\tau : \mathbb{P}_{\mathcal{Y}\times\mathcal{X}\times\mathcal{Z}} \to \mathbb{P}_{\mathcal{Y}\times\mathcal{X}'\times\mathcal{Z}'}$ such that $\tau(P_{YXZ}(y,x,z)) = P_{YX'Z'}(y,z,x)$. Let $\delta^Z(P) := \delta^X(\tau(P))$. Then the following functions define a nonnegative information decomposition satisfying (9)-(8):

$$\widetilde{UI}(Y;X\backslash Z) = \max\{\delta^X, \delta^Z + I(Y;X) - I(Y;Z)\},$$
$$\widetilde{UI}(Y;Z\backslash X) = \max\{\delta^Z, \delta^X + I(Y;Z) - I(Y;X)\},$$
$$\widetilde{SI}(Y;X,Z) = \min\{I(Y;X) - \delta^X, I(Y;Z) - \delta^Z\},$$
$$\widetilde{CI}(Y;X,Z) = \min\{I(Y;X|Z) - \delta^X, I(Y;Z|X) - \delta^Z\}.$$

We now apply the construction in Proposition 9 to derive an information decomposition based on the deficiency. The

following proposition is proved in the Appendix.

Proposition 10. Let $(Y, X, Z) \sim P$, and let $\kappa \in \mathsf{M}(\mathcal{X}; \mathcal{Y})$ and $d \in \mathsf{M}(\mathcal{Z}; \mathcal{Y})$ be two channels representing, resp., the conditional distributions $P_{Y|X}$ and $P_{Y|Z}$. Define $\delta^X = \delta^\pi(d, \kappa)$. Then the functions $\widetilde{UI}$, $\widetilde{SI}$, and $\widetilde{CI}$ in Proposition 9 define a nonnegative information decomposition.

The next proposition shows the relationship between the decompositions induced by the deficiency (see Proposition 10) and that induced by the UI (see (6)–(8)).

Proposition 11 ([24])**.**

$$\widetilde{UI}(Y; X \backslash Z) \leq UI(Y; X \backslash Z),$$
$$\widetilde{UI}(Y; Z \backslash X) \leq UI(Y; Z \backslash X),$$
$$\widetilde{SI}(Y; X, Z) \geq SI(Y; X, Z),$$
$$\widetilde{CI}(Y; X, Z) \geq CI(Y; X, Z),$$

with equality if and only if there exists $Q \in \Delta_P$ such that $\widetilde{CI}_Q(Y; X, Z) = 0$.

IV. EXPERIMENTS

A. Experiments on MNIST

We present experiments on the MNIST dataset. Classification on MNIST is a very well-studied problem. The main objective of these experiments is to evaluate the information-theoretic properties of the representations learned by the VDB model and to compare the classification accuracy for different values of M, the number of encoder output samples used in the training objective (5) when using the *oneshot* strategy. As mentioned in Section II, when $M = 1$, we recover the VIB model [13].

Settings. For the encoder, we use a fully connected feedforward network with 784 input units–1024 ReLUs–1024 ReLUs–512 linear output units. The deterministic output of this network is interpreted as the vector of means and variances of a 256-dimensional Gaussian distribution. The decoder is simply a softmax with 10 classes. These are the same settings of the model used by [13]. At test time, the classifier is evaluated using L encoder samples (i.e., we use $\frac{1}{L} \sum_{j=1}^{L} d(y|z^{(j)})$ where $z^{(j)} \sim e(z|x)$). We implement the algorithm in TensorFlow and train for 200 epochs using the Adam optimizer.

Test accuracy. The resulting test accuracy for different values of β and M is reported in Fig. 2(a) and Table I. As can be seen from Fig. 2(a), the test accuracy is stable with increasing M. From Table I we see that choosing M larger than one can in fact slightly improve test accuracy.

Information curve. The IB curve traces the mutual information $I(Z; Y)$ of representation and output (sufficiency) vs. the mutual information $I(Z; X)$ of representation and input (minimality), for different values of the regularization parameter β. In our method, "more sufficient" is replaced by "less deficient". The term corresponding to sufficiency is

$$J(Z; Y) := H(Y) - \mathbb{E}_{(x,y) \sim \hat{p}_{\text{data}}} \left[-\log \left(\int d(y|z) e(z|x) \, dz \right) \right].$$

Here $H(Y)$ is the entropy of the output, which for MNIST is $\log_2(10)$. Fig. 2(b) shows the VDB curve which traces $J(Z; Y)$ vs. $I(Z; X)$ for different values of β at the end of training. Note that the curve corresponding to $M = 1$ is just the VIB curve which traces $I(Z; Y)$ vs. $I(Z; X)$ for different values of β. For orientation, lower values of β have higher values of $I(Z; X)$ (towards the right of the plot). For small values of β, when the effect of the regularization is negligible, the bottleneck allows more information from the input through the representation. In this case, $J(Z; Y)$ increases on the training set, but not necessarily on the test set. This is manifest in the gap between the train and test curves indicative of a degradation in generalization. For intermediate values of β, *the gap is smaller for larger values of M (our method)*. Fig. 2(c) plots the attained mutual information values $I(Z; Y)$ vs. $I(Z; X)$ after training with the VDB objective for different values of β, while Fig. 2(c) plots the minimality term $I(Z; X)$ vs. β. Evidently, the levels of compression vary depending on M. For good values of β, *higher values of M (our method) lead to a more compressed representation while retaining the same level of sufficiency*. For example, for $\beta = 10^{-5}$, setting $M = 12$ requires storing ~ 50 less bits of information about the input when compared to the setting $M = 1$, while retaining the same mutual information about the output.

Training dynamics. The dynamics of the information quantities during training are also interesting, and shown in Fig. 3. At early epochs, training mainly effects fitting of the input-output relationship and an *increase of $J(Z; Y)$*. At later epochs, training mainly effects a *decrease of $I(Z; X)$* leading to a better generalization. An exception is when the regularization parameter β is very small, in which case the representation captures more information about the input, and longer training decreases $J(Z; Y)$, which is indicative of overfitting. Higher values of M (our method) lead to the representation capturing more information about the output, while at the same time discarding more information about the input. $M = 1$ corresponds to the VIB.

Low dimensional representations. To better understand the behavior of our method, we also visualize a 2-dimensional Gaussian representations after training the VDB with the *oneshot* strategy for $M = 1$ (when the VDB and VIB objectives are the same), and the *sequential* strategy for $M = 1$, and different values of the regularization parameter β. We

TABLE I: Test accuracy on MNIST for different values of β and M, bottleneck size K, and $L = 12$.

β	K	M			
		1	3	6	12
10^{-5}	256	0.9869	0.9873	**0.9885**	0.9878
	2	0.9575	0.9678	**0.9696**	0.9687
10^{-3}	256	0.9872	0.9879	0.9875	**0.9882**
	2	0.9632	0.9726	**0.9790**	0.9702

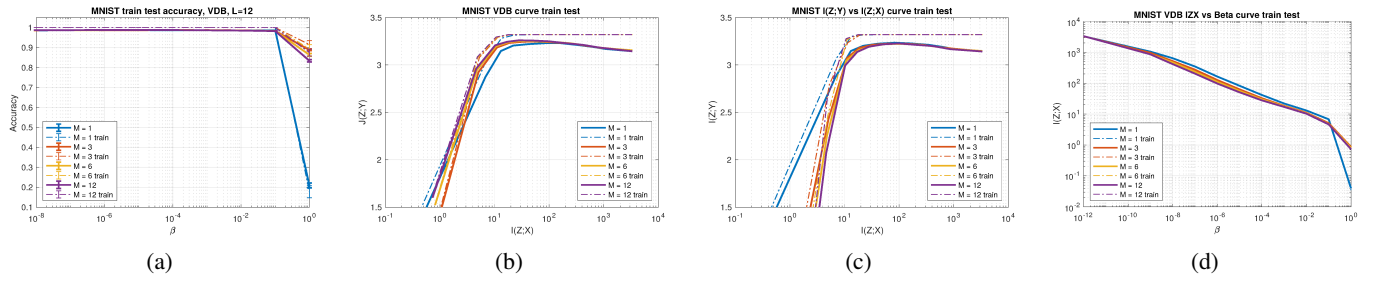(a)   (b)   (c)   (d)

Fig. 2: Effect of the regularization parameter β: (a) Accuracy on train and test data for the MNIST after training the VDB for different values of M. Here M is the number of encoder samples used in the training objective, and $L = 12$ is the number of encoder samples used for evaluating the classifier. (b) The VDB curve for different values of β. The curves are averages over 5 repetitions of the experiment. Each curve corresponds to one value of $M = 1, 3, 6, 12$. (c) The attained mutual information values $I(Z; Y)$ vs. $I(Z; X)$ for different values of β. (d) $I(X; Z)$ vs. β. For $M = 1$, the VDB and the VIB models coincide.
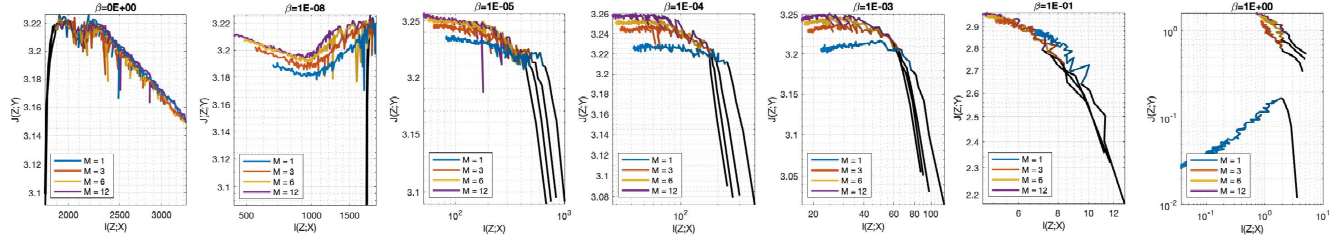
Fig. 3: Evolution of the sufficiency and minimality terms (values farther up and to the left are better) over 200 training epochs (dark to light color) on MNIST with a 256-dimensional representation for different values of β. The curves are averages over 20 repetitions of the experiment. $M = 1$ corresponds to the VIB model.

use the same settings as before, with the only difference that the dimension of the output layer of the encoder is 4, with two coordinates representing the mean, and two a diagonal covariance matrix. The results are shown in Fig. 4. We see that for $\beta = 10^{-4}$, representations of the different classes are well separated. We also observe that as the frequency of the encoder updates is increased relative to the decoder (our method), individual clusters tend to be more spread out in latent space. This translates into a better discriminative performance when compared to the *oneshot* (VIB) strategy as we show next with our robustness experiments.

B. Classification robustness under distributional shift

We demonstrate that the VDB generalizes well across distributional shifts, i.e., when the train and test distributions are different. We use the MNIST-C [18] and CIFAR-10-C [19] benchmarks to evaluate the classifier's robustness to common corruptions when trained with the VDB objective. These datasets are constructed by applying 15 common corruptions (at 5 different severity levels) to the MNIST and CIFAR-10 test sets. The corruptions comprise of four different categories, namely, noise, blur, weather, and digital.

To evaluate a classifier's robustness to common corruptions, we use the metrics proposed in [19]. Given a classifier f and a baseline classifier b, the corruption error (CE) on a certain corruption type c is computed as the ratio E_c^f / E_c^b, where E_c^f and E_c^b are resp. the errors of f and b on c, aggregated over five different severity levels. A more nuanced measure is the relative CE that measures corruption robustness relative to the *Clean error*, the usual classification error on the uncorrupted test set. The relative CE is computed as the ratio $(E_c^f - E_{clean}^f)/(E_c^b - E_{clean}^b)$, where E_{clean}^f and E_{clean}^b are resp. the clean errors of f and b. Averaging the CE and relative CE across all 15 corruption types yields the *mean CE (mCE)* and the *Relative mCE* values.

Results of computation of the robustness metrics for the MNIST-C and CIFAR-10-C datasets using different training strategies for different values of β are shown, resp., in Tables II and III. The statistics are averages over 4 independent runs. The keys "oneshot/M1" and "oneshot/M6" refer to a *oneshot* training strategy with resp., $M = 1$ and $M = 6$ encoder samples used for evaluating the training objective (5). The keys "seq:k:10/M1" and "seq:k:10/M6" refer to a *sequential* training strategy with resp., $M = 1$ and $M = 6$ encoder

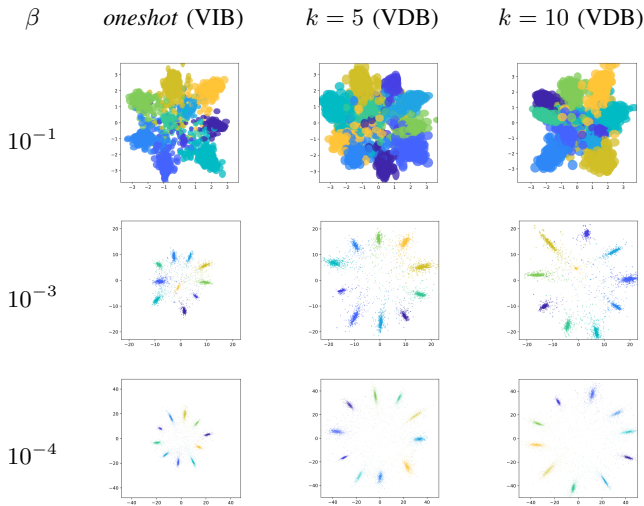|	*oneshot* (VIB)	$k = 5$ (VDB)	$k = 10$ (VDB)
10^{-1}			
10^{-3}			
10^{-4}			

Fig. 4: Posterior Gaussian distributions of 5000 test images from MNIST in a 2-dimensional latent space after training with $\beta = 10^{-1}, 10^{-3}, 10^{-4}$ and $M = 1$ with the *oneshot* strategy and the *sequential* strategy with $k = 5, 10$ encoder update steps per decoder update. Color corresponds to the class label. Boxes in each row have the same dimension.

TABLE II: *Clean Error*, *mCE*, and *Relative mCE* values for the MNIST-C dataset using a MLP of size 784-1024-1024-512 trained using various strategies for different values of β. Lower values are better.

β	Train strategy	*Clean Error*	*mCE*	*Relative mCE*
10^{-3}	oneshot/M1	1.53±0.10	100.00	100.00
	seq/k:10/M1	1.42±0.13	93.88	93.56
	oneshot/M6	1.38±0.04	86.84	84.46
	seq/k:10/M6	1.39±0.04	85.95	**83.39**
10^{-4}	oneshot/M1	1.46±0.11	100.00	100.00
	seq/k:10/M1	1.39±0.08	89.64	87.80
	oneshot/M6	1.30±0.08	91.13	90.95
	seq/k:10/M6	1.32±0.07	88.25	**86.85**
10^{-5}	oneshot/M1	1.63±0.44	100.00	100.00
	seq/k:10/M1	1.32±0.11	84.64	**85.98**
	oneshot/M6	1.29±0.02	84.51	87.84
	seq/k:10/M6	1.30±0.10	84.85	88.28

samples used for evaluating the training objective (5), and with $k = 10$ encoder update steps per decoder update. We choose the baseline classifier as the VIB model ("oneshot/M1").

For MNIST-C, we used the same encoder with a 256-dimensional representation as before. We trained the VDB for 200 epochs using the Adam optimizer with a fixed learning rate of 10^{-4}. For CIFAR-10-C, we used the 20-layer residual network "ResNet20" from [31] for the encoder with a 20-dimensional Gaussian representation and a softmax layer for the decoder.

We see that for $M = 1$, the sequential training strategy achieves lower *mCE* and *Relative mCE* values than the VIB across different values of β for both the MNIST-C and

TABLE III: *Clean Error*, *mCE*, and *Relative mCE* values for the CIFAR-10-C dataset using the ResNet20 network [31] trained using various strategies for different values of β. Lower values are better.

β	Train strategy	*Clean Error*	*mCE*	*Relative mCE*
10^{-3}	oneshot/M1	19.23±1.74	100.00	100.00
	seq/k:10/M1	19.75±0.51	98.85	98.43
	oneshot/M6	20.08±1.04	96.91	**95.95**
	seq/k:10/M6	18.20±1.48	97.46	97.82
10^{-4}	oneshot/M1	20.69±1.79	100.00	100.00
	seq/k:10/M1	20.41±1.39	99.50	99.55
	oneshot/M6	19.55±0.44	100.51	101.14
	seq/k:10/M6	20.88±0.39	98.48	**98.12**
10^{-5}	oneshot/M1	20.42±1.05	100.00	100.00
	seq/k:10/M1	18.78±0.52	97.22	**97.76**
	oneshot/M6	19.10±1.02	97.58	98.07
	seq/k:10/M6	19.48±1.12	97.68	98.01

CIFAR-10-C datasets. Recall that the objective of using the sequential strategy is to better approximate the deficiency which involves an optimization over the encoder. The advantage of sampling the encoder multiple times ($M > 1$) for each input sample during training is also evident for both the oneshot and sequential strategy. The improved robustness in this case might be explained by way of data augmentation in latent space.

V. DISCUSSION

We have formulated a bottleneck method based on channel deficiencies. The deficiency of a decoder w.r.t. a given channel quantifies how well an input randomization of the decoder (by a stochastic encoder) can be used to approximate the given channel. The DB has a natural variational formulation which recovers the VIB in the limit of a single sample of the encoder output. Moreover, the resulting variational objective can be implemented as an easy modification of the VIB objective with little to no computational overhead. Experiments show that the VDB can provide advantages in terms of minimality while retaining the same discriminative capacity as the VIB. We demonstrated that training with the VDB improves out-of-distribution robustness over the VIB on two benchmark datasets, the MNIST-C and the CIFAR-10-C.

An unsupervised version of the VDB objective (5) (for $\beta = 1$) shares some superficial similarities with the Importance Weighted Autoencoder (IWAE) [32] which also features a sum inside a logarithm. Note, however, that the IWAE objective cannot be decomposed for $M > 1$. This implies that we cannot trade-off reconstruction fidelity for learning meaningful representations by incorporating bottleneck constraints. As M increases, while the posterior approximation gets better, the magnitude of the gradient w.r.t. the encoder parameters also decays to zero [33]. This potentially limits the IWAE's ability to learn useful representations. It is plausible that a similar bias-variance trade-off occurs with the VDB objective for high values of M. This is worth investigating.

REFERENCES

[1] H. S. Witsenhausen and A. D. Wyner, "A conditional entropy bound for a pair of discrete random variables," *IEEE Transactions on Information Theory*, vol. 21, no. 5, pp. 493–501, 1975.

[2] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, 1999, pp. 368–377.

[3] P. Harremoës and N. Tishby, "The information bottleneck revisited or how to choose a good distortion measure," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2007, pp. 566–570.

[4] H. Hsu, S. Asoodeh, S. Salamatian, and F. P. Calmon, "Generalizing bottleneck problems," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 531–535.

[5] O. Shamir, S. Sabato, and N. Tishby, "Learning and generalization with the information bottleneck," in *International Conference on Algorithmic Learning Theory*. Springer, 2008, pp. 92–107.

[6] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "β-VAE: Learning basic visual concepts with a constrained variational framework," 2017, ICLR 2017.

[7] A. A. Alemi, B. Poole, I. Fischer, J. Dillon, R. A. Saurous, and K. Murphy, "Fixing a broken ELBO," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 159–168.

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

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

[10] Z. Goldfeld, E. Van Den Berg, K. Greenewald, I. Melnyk, N. Nguyen, B. Kingsbury, and Y. Polyanskiy, "Estimating information flow in deep neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, 2019, pp. 2299–2308.

[11] M. Chalk, O. Marre, and G. Tkacik, "Relevant sparse codes with variational information bottleneck," in *Advances in Neural Information Processing Systems*, 2016, pp. 1957–1965.

[12] A. Achille and S. Soatto, "Information dropout: Learning optimal representations through noisy computation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[13] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," *arXiv preprint arXiv:1612.00410*, 2016, ICLR 2017.

[14] A. Kolchinsky, B. D. Tracey, and D. H. Wolpert, "Nonlinear information bottleneck," *arXiv preprint arXiv:1705.02436*, 2017.

[15] D. Blackwell, "Equivalent comparisons of experiments," *The Annals of Mathematical Statistics*, vol. 24, no. 2, pp. 265–272, 1953.

[16] L. Le Cam, "Sufficiency and approximate sufficiency," *The Annals of Mathematical Statistics*, pp. 1419–1455, 1964.

[17] E. Torgersen, *Comparison of statistical experiments*. Cambridge University Press, 1991, vol. 36.

[18] N. Mu and J. Gilmer, "Mnist-c: A robustness benchmark for computer vision," *arXiv preprint arXiv:1906.02337*, 2019.

[19] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *arXiv preprint arXiv:1903.12261*, 2019, ICLR 2018.

[20] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[21] M. Raginsky, "Shannon meets Blackwell and Le Cam: Channels, codes, and statistical experiments," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2011, pp. 1220–1224.

[22] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013, ICLR 2013.

[23] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proceedings of the 31st International Conference on Machine Learning*, 2014, pp. 1278–1286.

[24] N. Bertschinger, J. Rauh, E. Olbrich, J. Jost, and N. Ay, "Quantifying unique information," *Entropy*, vol. 16, no. 4, pp. 2161–2183, 2014.

[25] P. K. Banerjee, E. Olbrich, J. Jost, and J. Rauh, "Unique informations and deficiencies," in *Proceedings of the 56th Annual Allerton Conference on Communication, Control and Computing*, 2018, pp. 32–38.

[26] M. Harder, C. Salge, and D. Polani, "A bivariate measure of redundant information," *Physical Review E*, vol. 87, p. 012130, 2013.

[27] R. Nasser, "On the input-degradedness and input-equivalence between channels," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2017, pp. 2453–2457.

[28] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, 2007.

[29] I. Csiszár and F. Matúš, "Information projections revisited," *IEEE Transactions on Information Theory*, vol. 49, no. 6, pp. 1474–1490, 2003.

[30] 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.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[32] Y. Burda, R. Grosse, and R. Salakhutdinov, "Importance weighted autoencoders," *arXiv preprint arXiv:1509.00519*, 2015, ICLR 2016.

[33] T. Rainforth, A. R. Kosiorek, T. A. Le, C. J. Maddison, M. Igl, F. Wood, and Y. W. Teh, "Tighter variational bounds are not necessarily better," *arXiv preprint arXiv:1802.04537*, 2018, ICML 2018.