Learnability for the Information Bottleneck

entropy

MDPI

Article

Learnability for the Information Bottleneck

Tailin Wu [1,*], Ian Fischer [2], Isaac L. Chuang [1] and Max Tegmark [1]

[1] Department of Physics, MIT, 77 Massachusetts Ave, Cambridge, MA 02139, USA; ichuang@mit.edu (I.L.C.); tegmark@mit.edu (M.T.)
[2] Google Research, 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA; iansf@google.com
* Correspondence: tailin@mit.edu

Received: 1 August 2019; Accepted: 12 September 2019; Published: 23 September 2019

Abstract: The Information Bottleneck (IB) method provides an insightful and principled approach for balancing compression and prediction for representation learning. The IB objective $I(X;Z) - \beta I(Y;Z)$ employs a Lagrange multiplier β to tune this trade-off. However, in practice, not only is β chosen empirically without theoretical guidance, there is also a lack of theoretical understanding between β, learnability, the intrinsic nature of the dataset and model capacity. In this paper, we show that if β is improperly chosen, learning cannot happen—the trivial representation $P(Z|X) = P(Z)$ becomes the global minimum of the IB objective. We show how this can be avoided, by identifying a sharp phase transition between the unlearnable and the learnable which arises as β is varied. This phase transition defines the concept of IB-Learnability. We prove several sufficient conditions for IB-Learnability, which provides theoretical guidance for choosing a good β. We further show that IB-learnability is determined by the largest *confident*, *typical* and *imbalanced subset* of the examples (the *conspicuous subset*), and discuss its relation with model capacity. We give practical algorithms to estimate the minimum β for a given dataset. We also empirically demonstrate our theoretical conditions with analyses of synthetic datasets, MNIST and CIFAR10.

Keywords: learnability; information bottleneck; representation learning; conspicuous subset

1. Introduction

Tishby et al. [1] introduced the *Information Bottleneck* (IB) objective function which learns a representation Z of observed variables (X, Y) that retains as little information about X as possible but simultaneously captures as much information about Y as possible:

$$\min \mathrm{IB}_\beta(X, Y; Z) = \min[I(X;Z) - \beta I(Y;Z)] \tag{1}$$

$I(\cdot)$ is the mutual information. The hyperparameter β controls the trade-off between compression and prediction, in the same spirit as Rate-Distortion Theory [2] but with a learned representation function $P(Z|X)$ that automatically captures some part of the "semantically meaningful" information, where the semantics are determined by the observed relationship between X and Y. The IB framework has been extended to and extensively studied in a variety of scenarios, including Gaussian variables [3], meta-Gaussians [4], continuous variables via variational methods [5–7], deterministic scenarios [8,9], geometric clustering [10] and is used for learning invariant and disentangled representations in deep neural nets [11,12].

From the IB objective (Equation (1)) we see that when $\beta \to 0$ it will encourage $I(X;Z) = 0$ which leads to a trivial representation Z that is independent of X, while when $\beta \to +\infty$, it reduces to a maximum likelihood objective (e.g., in classification, it reduces to cross-entropy loss). Therefore, as we vary β from 0 to $+\infty$, there must exist a point β_0 at which IB starts to learn a nontrivial representation where Z contains information about X.

As an example, we train multiple variational information bottleneck (VIB) models on binary classification of MNIST [13] digits 0 and 1 with 20% label noise at different β. The accuracy vs. β is shown in Figure 1. We see that when $\beta < 3.25$, no learning happens and the accuracy is the same as random guessing. Beginning with $\beta > 3.25$, there is a clear phase transition where the accuracy sharply increases, indicating the objective is able to learn a nontrivial representation. In general, we observe that different datasets and model capacity will result in different β_0 at which IB starts to learn a nontrivial representation. How does β_0 depend on the aspects of the dataset and model capacity and how can we estimate it? What does an IB model learn at the onset of learning? Answering these questions may provide a deeper understanding of IB in particular and learning on two observed variables in general.

In this work, we begin to answer the above questions. Specifically:

- We introduce the concept of *IB-Learnability* and show that when we vary β, the IB objective will undergo a phase transition from the inability to learn to the ability to learn (Section 3).
- Using the second-order variation, we derive sufficient conditions for IB-Learnability, which provide upper bounds for the learnability threshold β_0 (Section 4).
- We show that IB-Learnability is determined by the largest *confident*, *typical* and *imbalanced subset* of the examples (the *conspicuous subset*), reveal its relationship with the slope of the Pareto frontier at the origin on the information plane $I(X;Z)$ vs. $I(Y;Z)$ and discuss its relation to model capacity (Section 5).
- We prove a deep relationship between IB-Learnability, our upper bounds on β_0, the hypercontractivity coefficient, the contraction coefficient and the maximum correlation (Section 5).

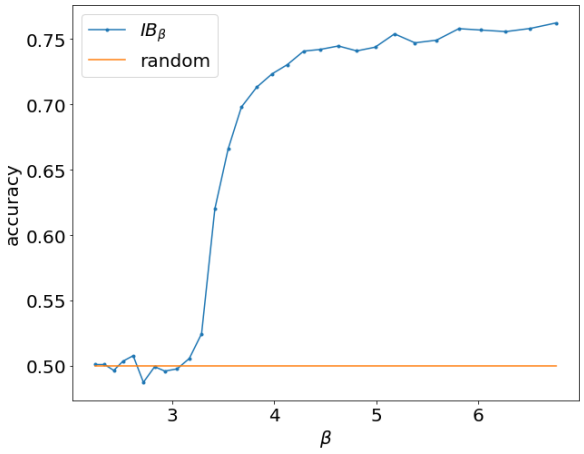

Figure 1. Accuracy for binary classification of MNIST digits 0 and 1 with 20% label noise and varying β. No learning happens for models trained at $\beta < 3.25$.

We also present an algorithm for estimating the onset of IB-Learnability and the conspicuous subset, which provide us with a tool for understanding a key aspect of the learning problem (X, Y) (Section 6). Finally, we use our main results to demonstrate on synthetic datasets, MNIST [13] and CIFAR10 [14] that the theoretical prediction for IB-Learnability closely matches experiment, and show the conspicuous subset our algorithm discovers (Section 7).

2. Related Work

The seminal IB work [1] provides a tabular method for exactly computing the optimal encoder distribution $P(Z|X)$ for a given β and cardinality of the discrete representation, $|Z|$. They did not consider the IB learnability problem as addressed in this work. Chechik et al. [3] presents the Gaussian Information Bottleneck (GIB) for learning a multivariate Gaussian representation Z of (X, Y), assuming

that both X and Y are also multivariate Gaussians. Under GIB, they derive analytic formula for the optimal representation as a noisy linear projection to eigenvectors of the normalized regression matrix $\Sigma_{x|y}\Sigma_x^{-1}$ and the learnability threshold β_0 is then given by $\beta_0 = \frac{1}{1-\lambda_1}$ where λ_1 is the largest eigenvalue of the matrix $\Sigma_{x|y}\Sigma_x^{-1}$. This work provides deep insights about relations between the dataset, β_0 and optimal representations in the Gaussian scenario but the restriction to multivariate Gaussian datasets limits the generality of the analysis Another analytic treatment of IB is given in [4], which reformulates the objective in terms of the copula functions. As with the GIB approach, this formulation restricts the form of the data distributions—the copula functions for the joint distribution (X, Y) are assumed to be known, which is unlikely in practice.

Strouse and Schwab [8] present the Deterministic Information Bottleneck (DIB), which minimizes the coding cost of the representation, $H(Z)$, rather than the transmission cost, $I(X;Z)$ as in IB. This approach learns hard clusterings with different code entropies that vary with β. In this case, it is clear that a hard clustering with minimal $H(Z)$ will result in a single cluster for all of the data, which is the DIB trivial solution. No analysis is given beyond this fact to predict the actual onset of learnability, however.

The first amortized IB objective is in the Variational Information Bottleneck (VIB) of Alemi et al. [5]. VIB replaces the exact, tabular approach of IB with variational approximations of the classifier distribution ($P(Y|Z)$) and marginal distribution ($P(Z)$). This approach cleanly permits learning a stochastic encoder, $P(Z|X)$, that is applicable to any $x \in \mathcal{X}$, rather than just the particular X seen at training time. The cost of this flexibility is the use of variational approximations that may be less expressive than the tabular method. Nevertheless, in practice, VIB learns easily and is simple to implement, so we rely on VIB models for our experimental confirmation.

Closely related to IB is the recently proposed Conditional Entropy Bottleneck (CEB) [7]. CEB attempts to explicitly learn the Minimum Necessary Information (MNI), defined as the point in the information plane where $I(X;Y) = I(X;Z) = I(Y;Z)$. The MNI point may not be achievable even in principle for a particular dataset. However, the CEB objective provides an explicit estimate of how closely the model is approaching the MNI point by observing that a necessary condition for reaching the MNI point occurs when $I(X;Z|Y) = 0$. The CEB objective $I(X;Z|Y) - \gamma I(Y;Z)$ is equivalent to IB at $\gamma = \beta + 1$, so our analysis of IB-Learnability applies equally to CEB.

Kolchinsky et al. [9] show that when Y is a deterministic function of X, the "corner point" of the IB curve (where $I(X;Y) = I(X;Z) = I(Y;Z)$) is the unique optimizer of the IB objective for all $0 < \beta' < 1$ (with the parameterization of Kolchinsky et al. [9], $\beta' = 1/\beta$), which they consider to be a "trivial solution". However, their use of the term "trivial solution" is distinct from ours. They are referring to the observation that all points on the IB curve contain uninteresting interpolations between two different but valid solutions on the optimal frontier, rather than demonstrating a non-trivial trade-off between compression and prediction as expected when varying the IB Lagrangian. Our use of "trivial" refers to whether IB is capable of learning at all given a certain dataset and value of β.

Achille and Soatto [12] apply the IB Lagrangian to the weights of a neural network, yielding InfoDropout. In Achille and Soatto [11], the authors give a deep and compelling analysis of how the IB Lagrangian can yield invariant and disentangled representations. They do not, however, consider the question of the onset of learning, although they are aware that not all models will learn a non-trivial representation. More recently, Achille et al. [15] repurpose the InfoDropout IB Lagrangian as a Kolmogorov Structure Function to analyze the ease with which a previously-trained network can be fine-tuned for a new task. While that work is tangentially related to learnability, the question it addresses is substantially different from our investigation of the onset of learning.

Our work is also closely related to the hypercontractivity coefficient [16,17], defined as $\sup_{Z-X-Y} \frac{I(Y;Z)}{I(X;Z)}$, which by definition equals the inverse of β_0, our IB-learnability threshold. In [16], the authors prove that the hypercontractivity cofficient equals the contraction coefficient $\eta_{\mathrm{KL}}(P_{Y|X}, P_X)$ and Kim et al. [18] propose a practical algorithm to estimate $\eta_{\mathrm{KL}}(P_{Y|X}, P_X)$, which provides a measure

for potential influence in the data. Although our goal is different, the sufficient conditions we provide for IB-Learnability are also lower bounds for the hypercontractivity coefficient.

3. IB-Learnability

We are given instances of (x, y) drawn from a distribution with probability (density) $P(X, Y)$ with support of $\mathcal{X} \times \mathcal{Y}$, where unless otherwise stated, both X and Y can be discrete or continuous variables. We use capital letters X, Y, Z for random variables and lowercase x, y, z to denote the instance of variables, with $P(\cdot)$ and $p(\cdot)$ denoting their probability or probability density, respectively. (X, Y) is our *training data* and may be characterized by different types of noise. The nature of this training data and the choice of β will be sufficient to predict the transition from unlearnable to learnable.

We can learn a representation Z of X with conditional probability $p(z|x)$, such that X, Y, Z obey the Markov chain $Z \leftarrow X \leftrightarrow Y$. Equation (1) above gives the IB objective with Lagrange multiplier β, $\mathrm{IB}_\beta(X, Y; Z)$, which is a functional of $p(z|x)$: $\mathrm{IB}_\beta(X, Y; Z) = \mathrm{IB}_\beta[p(z|x)]$. The IB learning task is to find a conditional probability $p(z|x)$ that minimizes $\mathrm{IB}_\beta(X, Y; Z)$. The larger β, the more the objective favors making a good prediction for Y. Conversely, the smaller β, the more the objective favors learning a concise representation.

How can we select β such that the IB objective learns a useful representation? In practice, the selection of β is done empirically. Indeed, Tishby et al. [1] recommends "sweeping β". In this paper, we provide theoretical guidance for choosing β by introducing the concept of IB-Learnability and providing a series of IB-learnable conditions.

Definition 1. *(X, Y) is IB_β-learnable if there exists a Z given by some $p_1(z|x)$, such that $IB_\beta(X, Y; Z)|_{p_1(z|x)} < IB_\beta(X, Y; Z)|_{p(z|x)=p(z)}$, where $p(z|x) = p(z)$ characterizes the trivial representation where $Z = Z_{trivial}$ is independent of X.*

If $(X; Y)$ is IB_β-learnable, then when $\mathrm{IB}_\beta(X, Y; Z)$ is globally minimized, it will *not* learn a trivial representation. On the other hand, if $(X; Y)$ is not IB_β-learnable, then when $\mathrm{IB}_\beta(X, Y; Z)$ is globally minimized, it may learn a trivial representation.

3.1. Trivial Solutions

Definition 1 defines trivial solutions in terms of representations where $I(X; Z) = I(Y; Z) = 0$. Another type of trivial solution occurs when $I(X; Z) > 0$ but $I(Y; Z) = 0$. This type of trivial solution is not directly achievable by the IB objective, as $I(X; Z)$ is minimized but it can be achieved by construction or by chance. It is possible that starting learning from $I(X; Z) > 0, I(Y; Z) = 0$ could result in access to non-trivial solutions not available from $I(X; Z) = 0$. We do not attempt to investigate this type of trivial solution in this work.

3.2. Necessary Condition for IB-Learnability

From Definition 1, we can see that IB_β-Learnability for any dataset $(X; Y)$ requires $\beta > 1$. In fact, from the Markov chain $Z \leftarrow X \leftrightarrow Y$, we have $I(Y; Z) \leq I(X; Z)$ via the data-processing inequality. If $\beta \leq 1$, then since $I(X; Z) \geq 0$ and $I(Y; Z) \geq 0$, we have that $\min(I(X; Z) - \beta I(Y; Z)) = 0 = \mathrm{IB}_\beta(X, Y; Z_{trivial})$. Hence (X, Y) is not IB_β-learnable for $\beta \leq 1$.

Due to the reparameterization invariance of mutual information, we have the following theorem for IB_β-Learnability:

Lemma 1. *Let $X' = g(X)$ be an invertible map (if X is a continuous variable, g is additionally required to be continuous). Then (X, Y) and (X', Y) have the same IB_β-Learnability.*

The proof for Lemma 1 is in Appendix A.2. Lemma 1 implies a favorable property for any condition for IB$_\beta$-Learnability: the condition should be invariant to invertible mappings of X. We will inspect this invariance in the conditions we derive in the following sections.

4. Sufficient Conditions for IB-Learnability

Given (X, Y), how can we determine whether it is IB$_\beta$-learnable? To answer this question, we derive a series of sufficient conditions for IB$_\beta$-Learnability, starting from its definition. The conditions are in increasing order of practicality, while sacrificing as little generality as possible.

Firstly, Theorem 1 characterizes the IB$_\beta$-Learnability range for β, with proof in Appendix A.3:

Theorem 1. *If (X, Y) is IB$_{\beta_1}$-learnable, then for any $\beta_2 > \beta_1$, it is IB$_{\beta_2}$-learnable.*

Based on Theorem 1, the range of β such that (X, Y) is IB$_\beta$-learnable has the form $\beta \in (\beta_0, +\infty)$. Thus, β_0 is the *threshold* of IB-Learnability.

Lemma 2. *$p(z|x) = p(z)$ is a stationary solution for IB$_\beta(X, Y; Z)$.*

The proof in Appendix A.6 shows that both first-order variations $\delta I(X; Z) = 0$ and $\delta I(Y; Z) = 0$ vanish at the trivial representation $p(z|x) = p(z)$, so $\delta \text{IB}_\beta[p(z|x)] = 0$ at the trivial representation.

Lemma 2 yields our strategy for finding sufficient conditions for learnability: find conditions such that $p(z|x) = p(z)$ is not a local minimum for the functional IB$_\beta[p(z|x)]$. Based on the necessary condition for the minimum (Appendix A.4), we have the following theorem (The theorems in this paper deal with learnability w.r.t. true mutual information. If parameterized models are used to approximate the mutual information, the limitation of the model capacity will translate into more uncertainty of Y given X, viewed through the lens of the model.):

Theorem 2 (Suff. Cond. 1). *A sufficient condition for (X, Y) to be IB$_\beta$-learnable is that there exists a perturbation function $h(z|x)$ (so that the perturbed probability (density) is $p'(z|x) = p(z|x) + \epsilon \cdot h(z|x)$) with $\int h(z|x)dz = 0$, such that the second-order variation $\delta^2 IB_\beta[p(z|x)] < 0$ at the trivial representation $p(z|x) = p(z)$.*

The proof for Theorem 2 is given in Appendix A.4. Intuitively, if $\delta^2 \text{IB}_\beta[p(z|x)]\big|_{p(z|x)=p(z)} < 0$, we can always find a $p'(z|x) = p(z|x) + \epsilon \cdot h(z|x)$ in the neighborhood of the trivial representation $p(z|x) = p(z)$, such that IB$_\beta[p'(z|x)] <$ IB$_\beta[p(z|x)]$, thus satisfying the definition for IB$_\beta$-Learnability.

To make Theorem 2 more practical, we perturb $p(z|x)$ around the trivial solution $p'(z|x) = p(z|x) + \epsilon \cdot h(z|x)$ and expand IB$_\beta[p(z|x) + \epsilon \cdot h(z|x)] -$ IB$_\beta[p(z|x)]$ to the second order of ϵ. We can then prove Theorem 3:

Theorem 3 (Suff. Cond. 2). *A sufficient condition for (X, Y) to be IB$_\beta$-learnable is X and Y are not independent and*

$$\beta > \inf_{h(x)} \beta_0[h(x)] \tag{2}$$

where the functional $\beta_0[h(x)]$ is given by

$$\beta_0[h(x)] = \frac{\mathbb{E}_{x\sim p(x)}[h(x)^2] - \left(\mathbb{E}_{x\sim p(x)}[h(x)]\right)^2}{\mathbb{E}_{y\sim p(y)}\left[\left(\mathbb{E}_{x\sim p(x|y)}[h(x)]\right)^2\right] - \left(\mathbb{E}_{x\sim p(x)}[h(x)]\right)^2}$$

Moreover, we have that $\left(\inf_{h(x)} \beta[h(x)]\right)^{-1}$ is a lower bound of the slope of the Pareto frontier in the information plane $I(Y; Z)$ vs. $I(X; Z)$ at the origin.

The proof is given in Appendix A.7, which also shows that if $\beta > \inf_{h(x)} \beta_0[h(x)]$ in Theorem 3 is satisfied, we can construct a perturbation function $h(z|x) = h^*(x)h_2(z)$ with $h^*(x) = \arg\min_{h(x)} \beta_0[h(x)]$, $\int h_2(z)dz = 0$, $\int \frac{h_2^2(z)}{p(z)}dz > 0$ for some $h_2(z)$, such that $h(z|x)$ satisfies Theorem 2. It also shows that the converse is true: if there exists $h(z|x)$ such that the condition in Theorem 2 is true, then Theorem 3 is satisfied, that is, $\beta > \inf_{h(x)} \beta_0[h(x)]$. (We do not claim that any $h(z|x)$ satisfying Theorem 2 can be decomposed to $h^*(x)h_2(z)$ at the onset of learning. But from the equivalence of Theorems 2 and 3 as explained above, when there exists an $h(z|x)$ such that Theorem 2 is satisfied, we can always construct an $h'(z|x) = h^*(x)h_2(z)$ that also satisfies Theorem 2.) Moreover, letting the perturbation function $h(z|x) = h^*(x)h_2(z)$ at the trivial solution, we have

$$p_\beta(y|x) = p(y) + \epsilon^2 C_z(h^*(x) - \overline{h}_x^*) \int p(x,y)(h^*(x) - \overline{h}_x^*)dx \tag{3}$$

where $p_\beta(y|x)$ is the estimated $p(y|x)$ by IB for a certain β, $\overline{h}_x^* = \int h^*(x)p(x)dx$ and $C_z = \int \frac{h_2^2(z)}{p(z)}dz > 0$ is a constant. This shows how the $p_\beta(y|x)$ by IB explicitly depends on $h^*(x)$ at the onset of learning. The proof is provided in Appendix A.8.

Theorem 3 suggests a method to estimate β_0: we can parameterize $h(x)$ for example, by a neural network, with the objective of minimizing $\beta_0[h(x)]$. At its minimization, $\beta_0[h(x)]$ provides an upper bound for β_0, and $h(x)$ provides a *soft clustering* of the examples corresponding to a nontrivial perturbation of $p(z|x)$ at $p(z|x) = p(z)$ that minimizes $\text{IB}_\beta[p(z|x)]$.

Alternatively, based on the property of $\beta_0[h(x)]$, we can also use a specific functional form for $h(x)$ in Equation (2) and obtain a stronger sufficient condition for IB_β-Learnability. But we want to choose $h(x)$ as near to the infimum as possible. To do this, we note the following characteristics for the R.H.S of Equation (2):

- We can set $h(x)$ to be nonzero if $x \in \Omega_x$ for some region $\Omega_x \subset \mathcal{X}$ and 0 otherwise. Then we obtain the following sufficient condition:

$$\beta > \inf_{h(x), \Omega_x \subset \mathcal{X}} \frac{\frac{\mathbb{E}_{x\sim p(x), x\in\Omega_x}[h(x)^2]}{\left(\mathbb{E}_{x\sim p(x), x\in\Omega_x}[h(x)]\right)^2} - 1}{\int \frac{dy}{p(y)} \left(\frac{\mathbb{E}_{x\sim p(x), x\in\Omega_x}[p(y|x)h(x)]}{\mathbb{E}_{x\sim p(x), x\in\Omega_x}[h(x)]}\right)^2 - 1} \tag{4}$$

- The numerator of the R.H.S. of Equation (4) attains its minimum when $h(x)$ is a constant within Ω_x. This can be proved using the Cauchy-Schwarz inequality: $\langle u,u\rangle\langle v,v\rangle \geq \langle u,v\rangle^2$, setting $u(x) = h(x)\sqrt{p(x)}$, $v(x) = \sqrt{p(x)}$ and defining the inner product as $\langle u,v\rangle = \int u(x)v(x)dx$. Therefore, the numerator of the R.H.S. of Equation (4) $\geq \frac{1}{\int_{x\in\Omega_x} p(x)} - 1$ and attains equality when $\frac{u(x)}{v(x)} = h(x)$ is constant.

Based on these observations, we can let $h(x)$ be a nonzero constant inside some region $\Omega_x \subset \mathcal{X}$ and 0 otherwise and the infimum over an arbitrary function $h(x)$ is simplified to infimum over $\Omega_x \subset \mathcal{X}$ and we obtain a sufficient condition for IB_β-Learnability, which is a key result of this paper:

Theorem 4 (Conspicuous Subset Suff. Cond.). *A sufficient condition for (X,Y) to be IB_β-learnable is X and Y are not independent and*

$$\beta > \inf_{\Omega_x \subset \mathcal{X}} \beta_0(\Omega_x) \tag{5}$$

where

$$\beta_0(\Omega_x) = \frac{\frac{1}{p(\Omega_x)} - 1}{\mathbb{E}_{y\sim p(y|\Omega_x)}\left[\frac{p(y|\Omega_x)}{p(y)} - 1\right]}$$

Ω_x denotes the event that $x \in \Omega_x$, with probability $p(\Omega_x)$.

$(\inf_{\Omega_x \subset \mathcal{X}} \beta_0(\Omega_x))^{-1}$ *gives a lower bound of the slope of the Pareto frontier in the information plane* $I(Y;Z)$ *vs.* $I(X;Z)$ *at the origin.*

The proof is given in Appendix A.9. In the proof we also show that this condition is invariant to invertible mappings of X.

5. Discussion

5.1. The Conspicuous Subset Determines β_0

From Equation (5), we see that three characteristics of the subset $\Omega_x \subset \mathcal{X}$ lead to low β_0: **(1) confidence:** $p(y|\Omega_x)$ is large; **(2) typicality and size:** the number of elements in Ω_x is large or the elements in Ω_x are typical, leading to a large probability of $p(\Omega_x)$; **(3) imbalance:** $p(y)$ is small for the subset Ω_x but large for its complement. In summary, β_0 will be determined by the largest *confident*, *typical* and *imbalanced subset* of examples or an equilibrium of those characteristics. We term Ω_x at the minimization of $\beta_0(\Omega_x)$ the *conspicuous subset*.

5.2. Multiple Phase Transitions

Based on this characterization of Ω_x, we can hypothesize datasets with multiple learnability phase transitions. Specifically, consider a region Ω_{x0} that is small but "typical", consists of all elements confidently predicted as y_0 by $p(y|x)$ and where y_0 is the least common class. By construction, this Ω_{x0} will dominate the infimum in Equation (5), resulting in a small value of β_0. However, the remaining $\mathcal{X} - \Omega_{x0}$ effectively form a new dataset, $\mathcal{X}_1$. At exactly β_0, we may have that the current encoder, $p_0(z|x)$, has no mutual information with the remaining classes in $\mathcal{X}_1$; that is, $I(Y_1;Z_0) = 0$. In this case, Definition 1 applies to $p_0(z|x)$ with respect to $I(X_1;Z_1)$. We might expect to see that, at β_0, learning will plateau until we get to some $\beta_1 > \beta_0$ that defines the phase transition for $\mathcal{X}_1$. Clearly this process could repeat many times, with each new dataset $\mathcal{X}_i$ being distinctly more difficult to learn than $\mathcal{X}_{i-1}$.

5.3. Similarity to Information Measures

The denominator of $\beta_0(\Omega_x)$ in Equation (5) is closely related to mutual information. Using the inequality $x - 1 \geq \log(x)$ for $x > 0$, it becomes:

$$\mathbb{E}_{y\sim p(y|\Omega_x)}\left[\frac{p(y|\Omega_x)}{p(y)} - 1\right] \geq \mathbb{E}_{y\sim p(y|\Omega_x)}\left[\log\frac{p(y|\Omega_x)}{p(y)}\right] = \tilde{I}(\Omega_x;Y)$$

where $\tilde{I}(\Omega_x;Y)$ is the mutual information "density" at $\Omega_x \subset \mathcal{X}$. Of course, this quantity is also $\mathbb{D}_{\mathrm{KL}}[p(y|\Omega_x)||p(y)]$, so we know that the denominator of Equation (5) is non-negative. Incidentally, $\mathbb{E}_{y\sim p(y|\Omega_x)}\left[\frac{p(y|\Omega_x)}{p(y)} - 1\right]$ is the density of "rational mutual information" [19] at Ω_x.

Similarly, the numerator of $\beta_0(\Omega_x)$ is related to the self-information of Ω_x:

$$\frac{1}{p(\Omega_x)} - 1 \geq \log\frac{1}{p(\Omega_x)} = -\log p(\Omega_x) = h(\Omega_x)$$

so we can estimate β_0 as:

$$\beta_0 \simeq \inf_{\Omega_x \subset \mathcal{X}} \frac{h(\Omega_x)}{\tilde{I}(\Omega_x;Y)} \tag{6}$$

Since Equation (6) uses upper bounds on both the numerator and the denominator, it does not give us a bound on β_0, only an estimate.

5.4. Estimating Model Capacity

The observation that a model cannot distinguish between cluster overlap in the data and its own lack of capacity gives an interesting way to use IB-Learnability to measure the capacity of a set of

models relative to the task they are being used to solve. For example, for a classification task, we can use different model classes to estimate $p(y|x)$. For each such trained model, we can estimate the corresponding IB-learnability threshold β_0. A model with smaller capacity than the task needs will translate to more uncertainty in $p(y|\Omega_x)$, resulting in a larger β_0. On the other hand, models that give the same β_0 as each other all have the same capacity relative to the task, even if we would otherwise expect them to have very different capacities. For example, if two deep models have the same core architecture but one has twice the number of parameters at each layer and they both yield the same β_0, their capacities are equivalent with respect to the task. Thus, β_0 provides a way to measure model capacity in a task-specific manner.

5.5. Learnability and the Information Plane

Many of our results can be interpreted in terms of the geometry of the Pareto frontier illustrated in Figure 2, which describes the trade-off between increasing $I(Y;Z)$ and decreasing $I(X;Z)$. At any point on this frontier that minimizes $\text{IB}_\beta^{\min} \equiv \min I(X;Z) - \beta I(Y;Z)$, the frontier will have slope β^{-1} if it is differentiable. If the frontier is also concave (has negative second derivative), then this slope β^{-1} will take its maximum β_0^{-1} at the origin, which implies IB_β-Learnability for $\beta > \beta_0$, so that the threshold for IB_β-Learnability is simply the inverse slope of the frontier at the origin. More generally, as long as the Pareto frontier is differentiable, the threshold for IB_β-learnability is the inverse of its maximum slope. Indeed, Theorem 3 and Theorem 4 give lower bounds of the slope of the Pareto frontier at the origin.

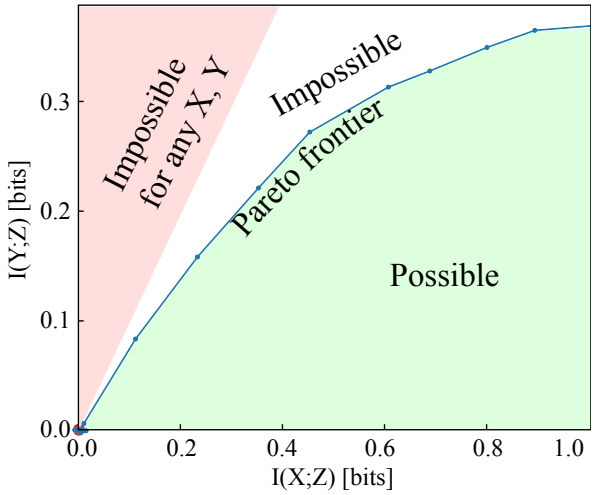

Figure 2. The Pareto frontier of the information plane, $I(X;Z)$ vs. $I(Y;Z)$, for the binary classification of MNIST digits 0 and 1 with 20% label noise described in Section 1 and Figure 1. For this problem, learning happens for models trained at $\beta > 3.25$. $H(Y) = 1$ bit since only two of ten digits are used and $I(Y;Z) \leq I(X;Y) \approx 0.5$ bits $< H(Y)$ because of the 20% label noise. The true frontier is differentiable; the figure shows a variational approximation that places an upper bound on both informations, horizontally offset to pass through the origin.

5.6. IB-Learnability, Hypercontractivity and Maximum Correlation

IB-Learnability and its sufficient conditions we provide harbor a deep connection with hypercontractivity and maximum correlation:

$$\frac{1}{\beta_0} = \xi(X;Y) = \eta_{\text{KL}} \geq \sup_{h(x)} \frac{1}{\beta_0[h(x)]} = \rho_m^2(X;Y) \tag{7}$$

which we prove in Appendix A.11. Here $\rho_m(X;Y) \equiv \max_{f,g} \mathbb{E}[f(X)g(Y)]$ s.t. $\mathbb{E}[f(X)] = \mathbb{E}[g(Y)] = 0$ and $\mathbb{E}[f^2(X)] = \mathbb{E}[g^2(Y)] = 1$ is the *maximum correlation* [20,21], $\xi(X;Y) \equiv \sup_{Z-X-Y} \frac{I(Y;Z)}{I(X;Z)}$ is the *hypercontractivity coefficient* and $\eta_{\mathrm{KL}}(p(y|x), p(x)) \equiv \sup_{r(x)\neq p(x)} \frac{\mathbb{D}_{\mathrm{KL}}(r(y)||p(y))}{\mathbb{D}_{\mathrm{KL}}(r(x)||p(x))}$ is the *contraction coefficient*. Our proof relies on Anantharam et al. [16]'s proof $\xi(X;Y) = \eta_{\mathrm{KL}}$. Our work reveals the deep relationship between IB-Learnability and these earlier concepts and provides additional insights about what aspects of a dataset give rise to high maximum correlation and hypercontractivity: the most confident, typical, imbalanced subset of (X, Y).

6. Estimating the IB-Learnability Condition

Theorem 4 not only reveals the relationship between the learnability threshold for β and the least noisy region of $P(Y|X)$ but also provides a way to practically estimate β_0, both in the general classification case and in more structured settings.

6.1. Estimation Algorithm

Based on Theorem 4, for general classification tasks we suggest Algorithm 1 to empirically estimate an upper-bound $\tilde{\beta}_0 \geq \beta_0$, as well as discovering the conspicuous subset that determines β_0.

We approximate the probability of each example $p(x_i)$ by its empirical probability, $\hat{p}(x_i)$. For example, for MNIST, $p(x_i) = \frac{1}{N}$, where N is the number of examples in the dataset. The algorithm starts by first learning a maximum likelihood model of $p_\theta(y|x)$, using for example, feed-forward neural networks. It then constructs a matrix $P_{y|x}$ and a vector p_y to store the estimated $p(y|x)$ and $p(y)$ for all the examples in the dataset. To find the subset Ω such that the $\tilde{\beta}_0$ is as small as possible, by previous analysis we want to find a *conspicuous* subset such that its $p(y|x)$ is large for a certain class j (to make the denominator of Equation (5) large) and containing as many elements as possible (to make the numerator small).

We suggest the following heuristics to discover such a conspicuous subset. For each class j, we sort the rows of $(P_{y|x})$ according to its probability for the pivot class j by decreasing order and then perform a search over $i_{\text{left}}, i_{\text{right}}$ for $\Omega = \{i_{\text{left}}, i_{\text{left}} + 1, ..., i_{\text{right}}\}$. Since $\tilde{\beta}_0$ is large when Ω contains too few or too many elements, the minimum of $\tilde{\beta}_0^{(j)}$ for class j will typically be reached with some intermediate-sized subset and we can use binary search or other discrete search algorithm for the optimization. The algorithm stops when $\tilde{\beta}_0^{(j)}$ does not improve by tolerance ε. The algorithm then returns the $\tilde{\beta}_0$ as the minimum over all the classes $\tilde{\beta}_0^{(1)}, ...\tilde{\beta}_0^{(N)}$, as well as the conspicuous subset that determines this $\tilde{\beta}_0$.

After estimating $\tilde{\beta}_0$, we can then use it for learning with IB, either directly or as an anchor for a region where we can perform a much smaller sweep than we otherwise would have. This may be particularly important for very noisy datasets, where β_0 can be very large.

Algorithm 1 Estimating the upper bound for β_0 and identifying the conspicuous subset

Require: Dataset $\mathcal{D} = \{(x_i, y_i)\}, i = 1, 2, ...N$. The number of classes is C.
Require ε: tolerance for estimating β_0
1: Learn a maximum likelihood model $p_\theta(y|x)$ using the dataset $\mathcal{D}$.
2: Construct matrix $(P_{y|x})$ such that $(P_{y|x})_{ij} = p_\theta(y = y_j|x = x_i)$.
3: Construct vector $p_y = (p_{y1}, .., p_{yC})$ such that $p_{yj} = \frac{1}{N}\sum_{i=1}^{N}(P_{y|x})_{ij}$.
4: **for** j **in** $\{1, 2, ...C\}$:
5: $\quad P_{y|x}^{(\text{sort}j)} \leftarrow$Sort the rows of $P_{y|x}$ in decreasing values of $(P_{y|x})_{ij}$.
6: $\quad \tilde{\beta}_0^{(j)}, \Omega^{(j)} \leftarrow$Search $i_{\text{left}}, i_{\text{right}}$ until $\tilde{\beta}_0^{(j)} = \textbf{Get}\boldsymbol{\beta}(P_{y|x}, p_y, \Omega)$ is minimal with tolerance ε, where $\Omega = \{i_{\text{left}}, i_{\text{left}} + 1, ...i_{\text{right}}\}$.
7: **end for**
8: $j^* \leftarrow \arg\min_j\{\tilde{\beta}_0^{(j)}\}, j = 1, 2, ...N$.
9: $\tilde{\beta}_0 \leftarrow \tilde{\beta}_0^{(j^*)}$.
10: $P_{y|x}^{(\tilde{\beta}_0)} \leftarrow$ the rows of $P_{y|x}^{(\text{sort}j^*)}$ indexed by $\Omega^{(j^*)}$.
11: **return** $\tilde{\beta}_0, P_{y|x}^{(\tilde{\beta}_0)}$

subroutine Get$\boldsymbol{\beta}$ $(P_{y|x}, p_y, \Omega)$:
s1: $N \leftarrow$ number of rows of $P_{y|x}$.
s2: $C \leftarrow$ number of columns of $P_{y|x}$.
s3: $n \leftarrow$ number of elements of Ω.
s4: $(p_{y|\Omega})_j \leftarrow \frac{1}{n}\sum_{i\in\Omega}(P_{y|x})_{ij}, j = 1, 2, ..., C$.
s5: $\tilde{\beta}_0 \leftarrow \dfrac{\frac{N}{n} - 1}{\sum_j\left[\frac{(p_{y|\Omega_x})_j^2}{p_{yj}} - 1\right]}$
s6: **return** $\tilde{\beta}_0$

6.2. Special Cases for Estimating β_0

Theorem 4 may still be challenging to estimate, due to the difficulty of making accurate estimates of $p(\Omega_x)$ and searching over $\Omega_x \subset \mathcal{X}$. However, if the learning problem is more structured, we may be able to obtain a simpler formula for the sufficient condition.

6.2.1. Class-Conditional Label Noise

Classification with noisy labels is a common practical scenario. An important noise model is that the labels are randomly flipped with some hidden class-conditional probabilities and we only observe the corrupted labels. This problem has been studied extensively [22–26]. If IB is applied to this scenario, how large β do we need? The following corollary provides a simple formula.

Corollary 1. *Suppose that the true class labels are y^* and the input space belonging to each y^* has no overlap. We only observe the corrupted labels y with class-conditional noise $p(y|x, y^*) = p(y|y^*)$ and Y is not independent of X. We have that a sufficient condition for IB$_\beta$-Learnability is:*

$$\beta > \inf_{y^*} \frac{\frac{1}{p(y^*)} - 1}{\sum_y \frac{p(y|y^*)^2}{p(y)} - 1} \tag{8}$$

We see that under class-conditional noise, the sufficient condition reduces to a discrete formula which only depends on the noise rates $p(y|y^*)$ and the true class probability $p(y^*)$, which can be accurately estimated via, for example, Northcutt et al. [26]. Additionally, if we know that the noise is

class-conditional but the observed β_0 is greater than the R.H.S. of Equation (8), we can deduce that there is overlap between the true classes. The proof of Corollary 1 is provided in Appendix A.10.

6.2.2. Deterministic Relationships

Theorem 4 also reveals that β_0 relates closely to whether Y is a deterministic function of X, as shown by Corollary 2:

Corollary 2. *Assume that Y contains at least one value y such that its probability $p(y) > 0$. If Y is a deterministic function of X and not independent of X, then a sufficient condition for IB$_\beta$-Learnability is $\beta > 1$.*

The assumption in the Corollary 2 is satisfied by classification and certain regression problems. (The following scenario does not satisfy this assumption: for certain regression problems where Y is a continuous random variable and the probability density function $p_Y(y)$ is bounded, then for any y, the *probability* $P(Y = y)$ has measure 0.)This corollary generalizes the result in Reference [9] which only proves it for classification problems. Combined with the necessary condition $\beta > 1$ for any dataset (X, Y) to be IB$_\beta$-learnable (Section 3), we have that under the assumption, if Y is a deterministic function of X, then a necessary and sufficient condition for IB$_\beta$-learnability is $\beta > 1$; that is, its β_0 is 1. The proof of Corollary 2 is provided in Appendix A.10.

Therefore, in practice, if we find that $\beta_0 > 1$, we may infer that Y is not a deterministic function of X. For a classification task, we may infer that either some classes have overlap or the labels are noisy. However, recall that finite models may add effective class overlap if they have insufficient capacity for the learning task, as mentioned in Section 4. This may translate into a higher observed β_0, even when learning deterministic functions.

7. Experiments

To test how the theoretical conditions for IB$_\beta$-learnability match with experiment, we apply them to synthetic data with varying noise rates and class overlap, MNIST binary classification with varying noise rates and CIFAR10 classification, comparing with the β_0 found experimentally. We also compare with the algorithm in Kim et al. [18] for estimating the hypercontractivity coefficient (=$1/\beta_0$) via the contraction coefficient η_{KL}. Experiment details are in Section A.12.

7.1. Synthetic Dataset Experiments

We construct a set of datasets from 2D mixtures of 2 Gaussians as X and the identity of the mixture component as Y. We simulate two practical scenarios with these datasets: (1) noisy labels with class-conditional noise and (2) class overlap. For (1), we vary the class-conditional noise rates. For (2), we vary class overlap by tuning the distance between the Gaussians. For each experiment, we sweep β with exponential steps and observe $I(X; Z)$ and $I(Y; Z)$. We then compare the empirical β_0 indicated by the onset of above-zero $I(X; Z)$ with predicted values for β_0.

7.1.1. Classification with Class-Conditional Noise

In this experiment, we have a mixture of Gaussian distribution with 2 components, each of which is a 2D Gaussian with diagonal covariance matrix $\Sigma = \mathrm{diag}(0.25, 0.25)$. The two components have distance 16 (hence virtually no overlap) and equal mixture weight. For each x, the label $y \in \{0, 1\}$ is the identity of which component it belongs to. We create multiple datasets by randomly flipping the labels y with a certain noise rate $\rho = P(y = 0|y^* = 1) = P(y = 1|y^* = 0)$. For each dataset, we train VIB models across a range of β and observe the onset of learning via random $I(X; Z)$ (Observed). To test how different methods perform in estimating β_0, we apply the following methods: **(1)** Corollary 1, since this is classification with class-conditional noise and the two true classes have virtually no overlap; **(2)** Algorithm 1 with true $p(y|x)$; **(3)** The algorithm in Kim et al. [18] that estimates $\hat{\eta}_{\mathrm{KL}}$, provided with

true $p(y|x)$; **(4)** $\beta_0[h(x)]$ in Equation (2); **(2′)** Algorithm 1 with $p(y|x)$ estimated by a neural net; **(3′)** $\hat{\eta}_{\mathrm{KL}}$ with the same $p(y|x)$ as in (2′). The results are shown in Figure 3 and in Table 1.

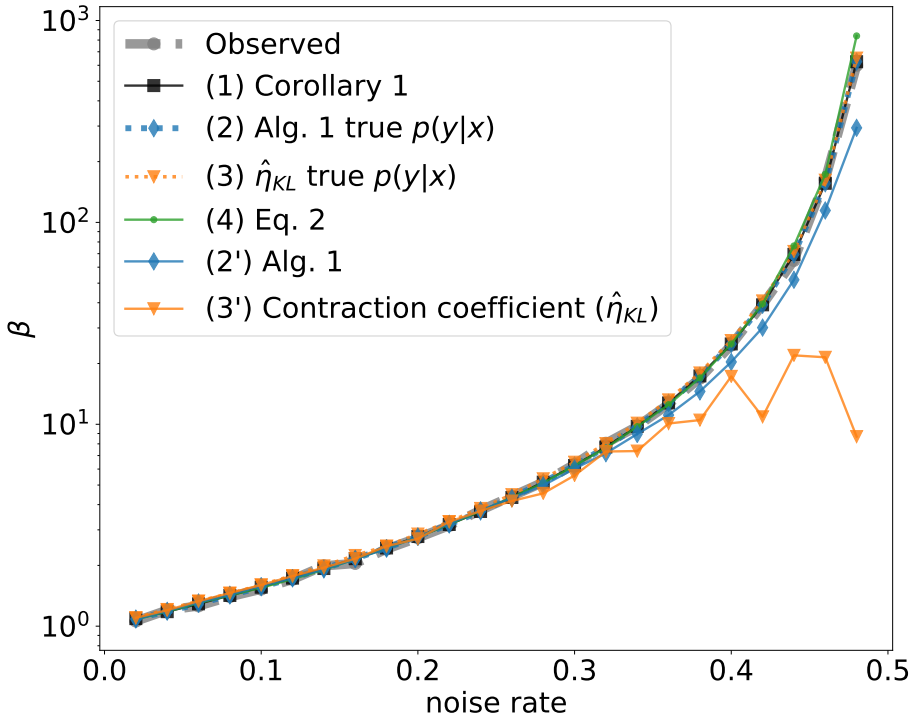

Figure 3. Predicted vs. experimentally identified β_0, for mixture of Gaussians with varying class-conditional noise rates.

Table 1. Full table of values used to generate Figure 3.

			(2) Algorithm 1	(3) $\hat{\eta}_{\mathrm{KL}}$			
Noise Rate	**Observed**	**(1) Corollary 1**	**True $p(y\|x)$**	**True $p(y\|x)$**	**(4) Equation (2)**	**(2′) Algorithm 1**	**(3′) $\hat{\eta}_{\mathrm{KL}}$**
0.02	1.06	1.09	1.09	1.10	1.08	1.08	1.10
0.04	1.20	1.18	1.18	1.21	1.18	1.19	1.20
0.06	1.26	1.29	1.29	1.33	1.30	1.31	1.33
0.08	1.40	1.42	1.42	1.45	1.42	1.43	1.46
0.10	1.52	1.56	1.56	1.60	1.55	1.58	1.60
0.12	1.70	1.73	1.73	1.78	1.71	1.73	1.77
0.14	1.99	1.93	1.93	1.99	1.90	1.91	1.95
0.16	2.04	2.16	2.16	2.24	2.15	2.15	2.16
0.18	2.41	2.44	2.44	2.49	2.43	2.42	2.49
0.20	2.74	2.78	2.78	2.86	2.76	2.77	2.71
0.22	3.15	3.19	3.19	3.29	3.19	3.21	3.29
0.24	3.75	3.70	3.70	3.83	3.71	3.75	3.72
0.26	4.40	4.34	4.34	4.48	4.35	4.31	4.17
0.28	5.16	5.17	5.17	5.37	5.12	4.98	4.55
0.30	6.34	6.25	6.25	6.49	6.24	6.03	5.58
0.32	8.06	7.72	7.72	8.02	7.63	7.19	7.33
0.34	9.77	9.77	9.77	10.13	9.74	8.95	7.37
0.36	12.58	12.76	12.76	13.21	12.51	11.11	10.09
0.38	16.91	17.36	17.36	17.96	16.97	14.55	10.49
0.40	24.66	25.00	25.00	25.99	25.01	20.36	17.27
0.42	39.08	39.06	39.06	40.85	39.48	30.12	10.89
0.44	64.82	69.44	69.44	71.80	76.48	51.95	21.95
0.46	163.07	156.25	156.26	161.88	173.15	114.57	21.47
0.48	599.45	625.00	625.00	651.47	838.90	293.90	8.69

From Figure 3 and Table 1 we see the following. **(A)** When using the true $p(y|x)$, both Algorithm 1 and $\hat{\eta}_{\mathrm{KL}}$ generally upper bound the empirical β_0 and Algorithm 1 is generally tighter. **(B)** When using the true $p(y|x)$, Algorithm 1 and Corollary 1 give the same result. **(C)** Comparing Algorithm 1 and $\hat{\eta}_{\mathrm{KL}}$ both of which use the same empirically estimated $p(y|x)$, both approaches provide good estimation in the low-noise region; however, in the high-noise region, Algorithm 1 gives more precise values than $\hat{\eta}_{\mathrm{KL}}$, indicating that Algorithm 1 is more robust to the estimation error of $p(y|x)$. **(D)** Equation (2) empirically upper bounds the experimentally observed β_0 and gives almost the same result as theoretical estimation in Corollary 1 and Algorithm 1 with the true $p(y|x)$. In the classification setting, this approach does not require any learned estimate of $p(y|x)$, as we can directly use the empirical $p(y)$ and $p(x|y)$ from SGD mini-batches.

This experiment also shows that for dataset where the signal-to-noise is small, β_0 can be very high. Instead of blindly sweeping β, our result can provide guidance for setting β so learning can happen.

7.1.2. Classification with Class Overlap

In this experiment, we test how different amounts of overlap among classes influence β_0. We use the mixture of Gaussians with two components, each of which is a 2D Gaussian with diagonal covariance matrix $\Sigma = \mathrm{diag}(0.25, 0.25)$. The two components have weights 0.6 and 0.4. We vary the distance between the Gaussians from 8.0 down to 0.8 and observe the $\beta_{0,exp}$. Since we do not add noise to the labels, if there were no overlap and a deterministic map from X to Y, we would have $\beta_0 = 1$ by Corollary 2. The more overlap between the two classes, the more uncertain Y is given X. By Equation (5) we expect β_0 to be larger, which is corroborated in Figure 4.

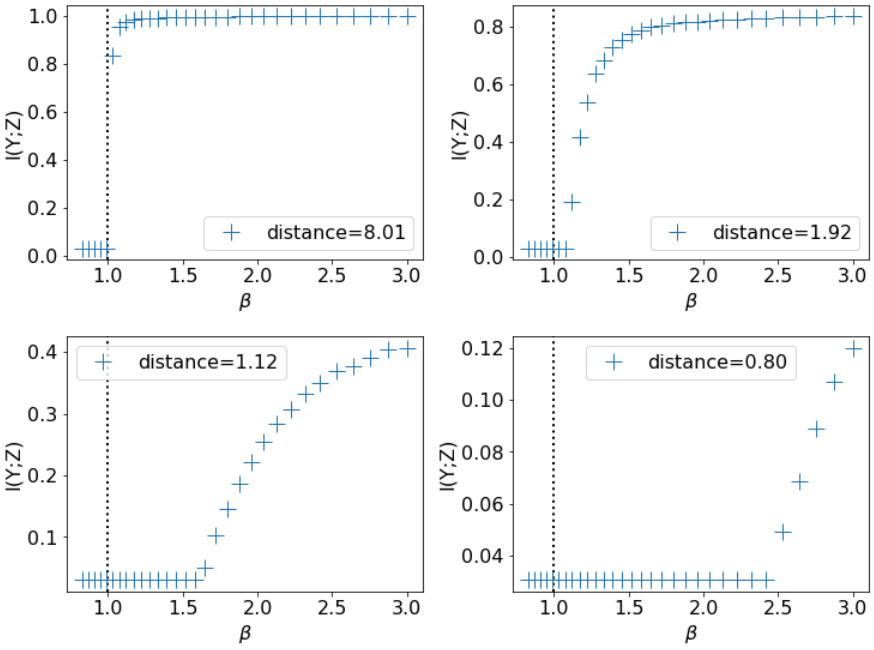

Figure 4. $I(Y;Z)$ vs. β, for mixture of Gaussian datasets with different distances between the two mixture components. The vertical lines are $\beta_{0,\mathrm{predicted}}$ computed by the R.H.S. of Equation (8). As Equation (8) does not make predictions w.r.t. class overlap, the vertical lines are always just above $\beta_{0,\mathrm{predicted}} = 1$. However, as expected, decreasing the distance between the classes in X space also increases the true β_0.

7.2. MNIST Experiments

We perform binary classification with digits 0 and 1 and as before, add class-conditional noise to the labels with varying noise rates ρ. To explore how the model capacity influences the onset of

learning, for each dataset we train two sets of VIB models differing only by the number of neurons in their hidden layers of the encoder: one with $n = 512$ neurons, the other with $n = 128$ neurons. As we describe in Section 4, insufficient capacity will result in more uncertainty of Y given X from the point of view of the model, so we expect the observed β_0 for the $n = 128$ model to be larger. This result is confirmed by the experiment (Figure 5). Also, in Figure 5 we plot β_0 given by different estimation methods. We see that the observations (A), (B), (C) and (D) in Section 7.1 still hold.

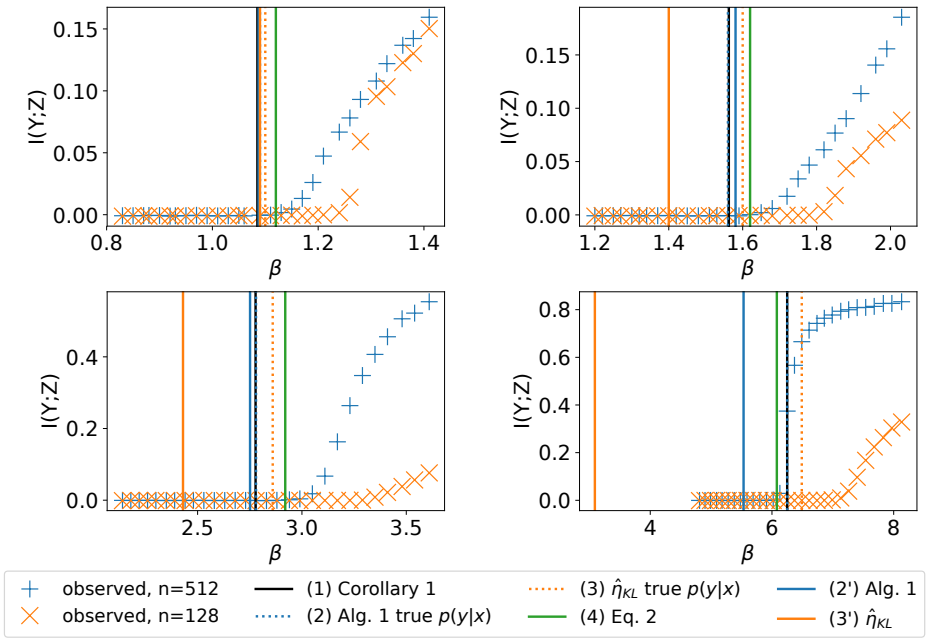

Figure 5. $I(Y;Z)$ vs. β for the MNIST binary classification with different hidden units per layer n and noise rates ρ: (upper left) $\rho = 0.02$, (upper right) $\rho = 0.1$, (lower left) $\rho = 0.2$, (lower right) $\rho = 0.3$. The vertical lines are β_0 estimated by different methods. $n = 128$ has insufficient capacity for the problem, so its observed learnability onset is pushed higher, similar to the class overlap case.

7.3. MNIST Experiments Using Equation (2)

To see what IB learns at its onset of learning for the full MNIST dataset, we optimize Equation (2) w.r.t. the full MNIST dataset and visualize the clustering of digits by $h(x)$. Equation (2) can be optimized using SGD using any differentiable parameterized mapping $h(x) : \mathcal{X} \to \mathbb{R}$. In this case, we chose to parameterize $h(x)$ with a PixelCNN++ architecture [27,28], as PixelCNN++ is a powerful autoregressive model for images that gives a scalar output (normally interpreted as $\log p(x)$). Equation (2) should generally give two clusters in the output space, as discussed in Section 4. In this setup, smaller values of $h(x)$ correspond to the subset of the data that is easiest to learn. Figure 6 shows two strongly separated clusters, as well as the threshold we choose to divide them. Figure 7 shows the first 5776 MNIST training examples as sorted by our learned $h(x)$, with the examples above the threshold highlighted in red. We can clearly see that our learned $h(x)$ has separated the "easy" one (1) digits from the rest of the MNIST training set.

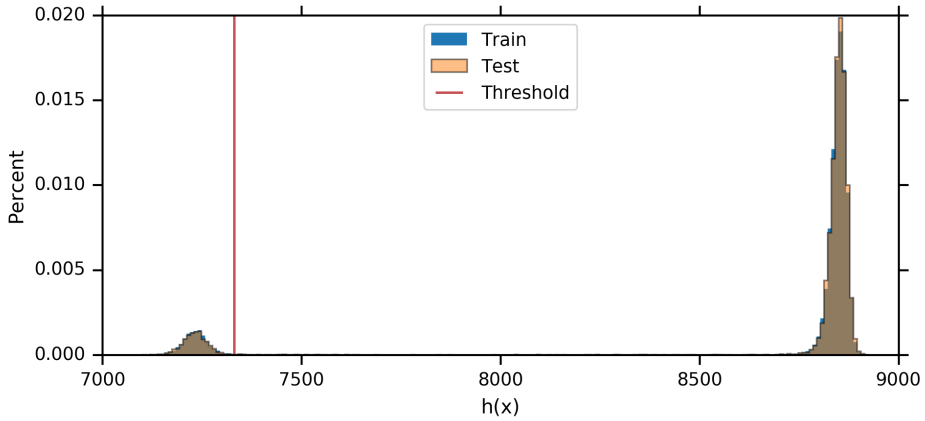

Figure 6. Histograms of the full MNIST training and validation sets according to $h(X)$. Note that both are bimodal and the histograms are indistinguishable. In both cases, $h(x)$ has learned to separate most of the ones into the smaller mode but difficult ones are in the wide valley between the two modes. See Figure 7 for all of the training images to the left of the red threshold line, as well as the first few images to the right of the threshold.

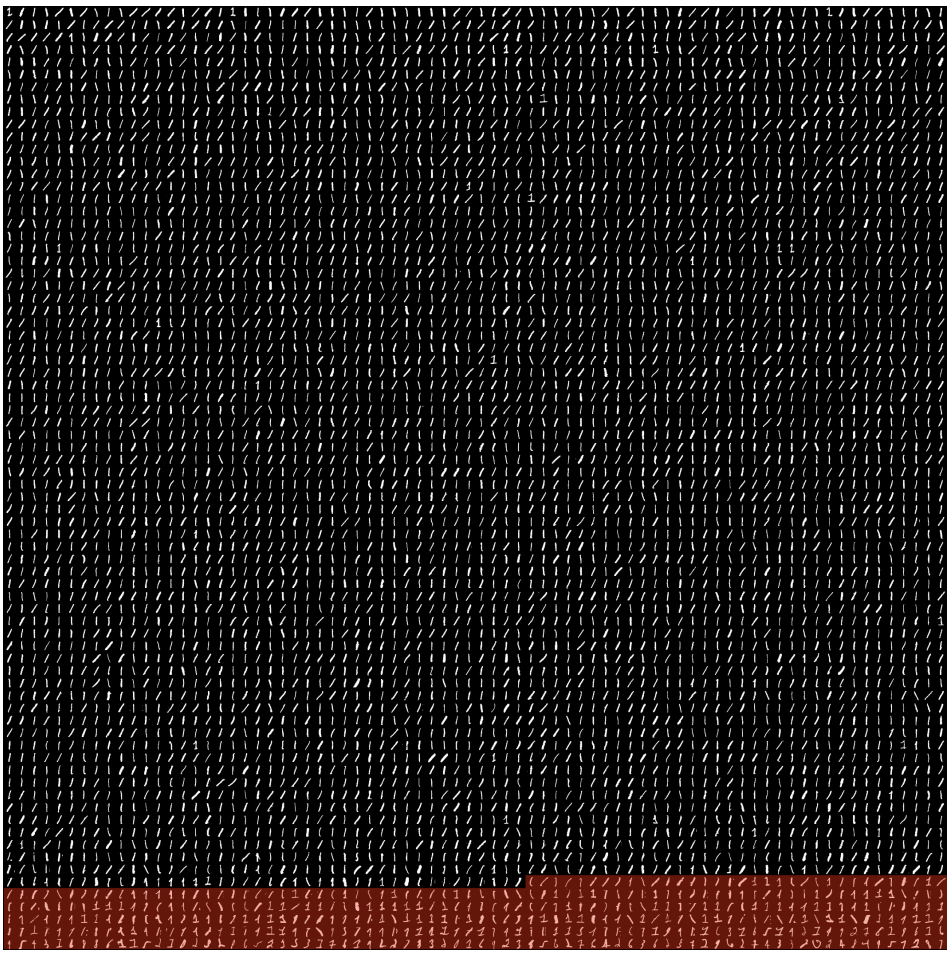

Figure 7. The first 5776 MNIST training set digits when sorted by $h(x)$. The digits highlighted in red are above the threshold drawn in Figure 6.

7.4. CIFAR10 Forgetting Experiments

For CIFAR10 [14], we study how *forgetting* varies with β. In other words, given a VIB model trained at some high β_2, if we anneal it down to some much lower β_1, what $I(Y;Z)$ does the model

converge to? Using Algorithm 1, we estimated $\beta_0 = 1.0483$ on a version of CIFAR10 with 20% label noise, where the $P_{y|x}$ is estimated by maximum likelihood training with the same encoder and classifier architectures as used for VIB. For the VIB models, the lowest β with performance above chance was $\beta = 1.048$ (Figure 8), a very tight match with the estimate from Algorithm 1. See Appendix A.12 for details.

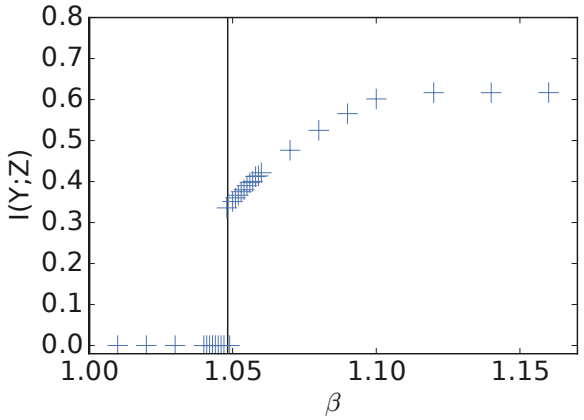

Figure 8. Plot of $I(Y;Z)$ vs. β for CIFAR10 training set with 20% label noise. Each blue cross corresponds to a fully-converged model starting with independent initialization. The vertical black line corresponds to the predicted $\beta_0 = 1.0483$ using Algorithm 1. The empirical $\beta_0 = 1.048$.

8. Conclusions

In this paper, we have presented theoretical results for predicting the onset of learning and have shown that it is determined by the conspicuous subset of the training examples. We gave a practical algorithm for predicting the transition as well as discovering this subset and showed that those predictions are accurate, even in cases of extreme label noise. We proved a deep connection between IB-learnability, our upper bounds on β_0, the hypercontractivity coefficient, the contraction coefficient and the maximum correlation. We believe that these results provide a deeper understanding of IB, as well as a tool for analyzing a dataset by discovering its conspicuous subset and a tool for measuring model capacity in a task-specific manner. Our work also raises other questions, such as whether there are other phase transitions in learnability that might be identified. We hope to address some of those questions in future work.

Author Contributions: Conceptualization, T.W. and I.F.; methodology, T.W., I.F., I.L.C. and M.T.; software, T.W. and I.F.; validation, T.W. and I.F.; formal analysis, T.W. and I.F.; investigation, T.W. and I.F.; resources, T.W., I.F., I.L.C. and M.T.; data curation, T.W. and I.F.; writing–original draft preparation, T.W., I.F., I.L.C. and M.T.; writing–review and editing, T.W., I.F., I.L.C. and M.T.; visualization, T.W. and I.F.; supervision, I.F., I.L.C. and M.T.; project administration, I.F., I.L.C. and M.T.; funding acquisition, M.T.

Funding: T.W.'s work was supported by the The Casey and Family Foundation, the Foundational Questions Institute and the Rothberg Family Fund for Cognitive Science. He thanks the Center for Brains, Minds and Machines (CBMM) for hospitality.

Acknowledgments: The authors would like to thank the anonymous reviewers for their constructive comments that contributed to improving the paper.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A

The structure of the Appendix is as follows. In Appendix A.1, we provide preliminaries for the first-order and second-order variations on functionals. We prove Theorem 1 and Theorem 1 in Appendixes A.2 and A.3, respectively. In Appendix A.4, we prove Theorem 2, the sufficient condition

1 for IB-Learnability. In Appendix A.5, we calculate the first and second variations of $\mathrm{IB}_\beta[p(z|x)]$ at the trivial representation $p(z|x) = p(z)$, which is used in proving Lemma 2 (Appendix A.6) and the Sufficient Condition 2 for IB_β-learnability (Appendix A.7). In Appendix A.8, we prove Equation (3) at the onset of learning. After these preparations, we prove the key result of this paper, Theorem 4, in Section A.9. Then two important Corollaries 1, 2 are proved in Appendix A.10. In Appendix A.11 we explore the deep relation between β_0, $\beta_0[h(x)]$, the hypercontractivity coefficient, contraction coefficient and maximum correlation. Finally in Appendix A.12, we provide details for the experiments.

Below are some implicit conventions of the paper: for integrals, whenever a variable W is discrete, we can simply replace the integral $(\int \cdot dw)$ by summation $(\sum_w \cdot)$.

Appendix A.1. Preliminaries: First-Order and Second-Order Variations

Let functional $F[f(x)]$ be defined on some normed linear space $\mathscr{R}$. Let us add a perturbative function $\epsilon \cdot h(x)$ to $f(x)$, and now the functional $F[f(x) + \epsilon \cdot h(x)]$ can be expanded as

$$\begin{aligned}\Delta F[f(x)] &= F[f(x) + \epsilon \cdot h(x)] - F[f(x)] \\ &= \varphi_1[f(x)] + \varphi_2[f(x)] + \mathcal{O}(\epsilon^3 ||h||^2)\end{aligned}$$

where $||h||$ denotes the norm of h, $\varphi_1[f(x)] = \epsilon \frac{dF[f(x)]}{d\epsilon}$ is a linear functional of $\epsilon \cdot h(x)$, and is called the *first-order variation*, denoted as $\delta F[f(x)]$. $\varphi_2[f(x)] = \frac{1}{2}\epsilon^2 \frac{d^2F[f(x)]}{d\epsilon^2}$ is a quadratic functional of $\epsilon \cdot h(x)$, and is called the *second-order variation*, denoted as $\delta^2 F[f(x)]$.

If $\delta F[f(x)] = 0$, we call $f(x)$ a stationary solution for the functional $F[\cdot]$.

If $\Delta F[f(x)] \geq 0$ for all $h(x)$ such that $f(x) + \epsilon \cdot h(x)$ is at the neighborhood of $f(x)$, we call $f(x)$ a (local) minimum of $F[\cdot]$.

Appendix A.2. Proof of Lemma 1

Proof. If (X, Y) is IB_β-learnable, then there exists $Z \in \mathcal{Z}$ given by some $p_1(z|x)$ such that $\mathrm{IB}_\beta(X, Y; Z) < \mathrm{IB}(X, Y; Z_{trivial}) = 0$, where $Z_{trivial}$ satisfies $p(z|x) = p(z)$. Since $X' = g(X)$ is a invertible map (if X is continuous variable, g is additionally required to be continuous), and mutual information is invariant under such an invertible map [29], we have that $\mathrm{IB}_\beta(X', Y; Z) = I(X'; Z) - \beta I(Y; Z) = I(X; Z) - \beta I(Y; Z) = \mathrm{IB}_\beta(X, Y; Z) < 0 = \mathrm{IB}(X', Y; Z_{trivial})$, so (X', Y) is IB_β-learnable. On the other hand, if (X, Y) is not IB_β-learnable, then $\forall Z$, we have $\mathrm{IB}_\beta(X, Y; Z) \geq \mathrm{IB}(X, Y; Z_{trivial}) = 0$. Again using mutual information's invariance under g, we have for all Z, $\mathrm{IB}_\beta(X', Y; Z) = \mathrm{IB}_\beta(X, Y; Z) \geq \mathrm{IB}(X, Y; Z_{trivial}) = 0$, leading to that (X', Y) is not IB_β-learnable. Therefore, we have that (X, Y) and (X', Y) have the same IB_β-learnability. □

Appendix A.3. Proof of Theorem 1

Proof. At the trivial representation $p(z|x) = p(z)$, we have $I(X; Z) = 0$, and $I(Y; Z) = 0$ due to the Markov chain, so $\mathrm{IB}_\beta(X, Y; Z)|_{p(z|x)=p(z)} = 0$ for any β. Since (X, Y) is IB_{β_1}-learnable, there exists a Z given by a $p_1(z|x)$ such that $\mathrm{IB}_{\beta_1}(X, Y; Z)|_{p_1(z|x)} < 0$. Since $\beta_2 > \beta_1$, and $I(Y; Z) \geq 0$, we have $\mathrm{IB}_{\beta_2}(X, Y; Z)|_{p_1(z|x)} \leq \mathrm{IB}_{\beta_1}(X, Y; Z)|_{p_1(z|x)} < 0 = \mathrm{IB}_{\beta_2}(X, Y; Z)|_{p(z|x)=p(z)}$. Therefore, (X, Y) is IB_{β_2}-learnable. □

Appendix A.4. Proof of Theorem 2

Proof. To prove Theorem 2, we use the Theorem 1 of Chapter 5 of Gelfand et al. [30] which gives a necessary condition for $F[f(x)]$ to have a minimum at $f_0(x)$. Adapting to our notation, we have:

Theorem A1 ([30]). *A necessary condition for the functional $F[f(x)]$ to have a minimum at $f(x) = f_0(x)$ is that for $f(x) = f_0(x)$ and all admissible $\epsilon \cdot h(x)$,*

$$\delta^2 F[f(x)] \geq 0.$$

Applying to our functional $\mathrm{IB}_\beta[p(z|x)]$, an immediate result of Theorem A1 is that, if at $p(z|x) = p(z)$, there exists an $\epsilon \cdot h(z|x)$ such that $\delta^2 \mathrm{IB}_\beta[p(z|x)] < 0$, then $p(z|x) = p(z)$ is not a minimum for $\mathrm{IB}_\beta[p(z|x)]$. Using the definition of IB_β learnability, we have that (X, Y) is IB_β-learnable. □

Appendix A.5. First- and Second-Order Variations of $IB_\beta[p(z|x)]$

In this section, we derive the first- and second-order variations of $\mathrm{IB}_\beta[p(z|x)]$, which are needed for proving Lemma 2 and Theorem 3.

Lemma A1. *Using perturbative function $h(z|x)$, we have*

$$\delta IB_\beta[p(z|x)] = \int dxdzp(x)h(z|x)\log\frac{p(z|x)}{p(z)} - \beta\int dxdydzp(x,y)h(z|x)\log\frac{p(z|y)}{p(z)}$$

$$\delta^2 IB_\beta[p(z|x)] = \frac{1}{2}\Big[\int dxdz\frac{p(x)^2}{p(x,z)}h(z|x)^2 - \beta\int dxdx'dydz\frac{p(x,y)p(x',y)}{p(y,z)}h(z|x)h(z|x') + (\beta-1)\int dxdx'dz\frac{p(x)p(x')}{p(z)}h(z|x)h(z|x')\Big]$$

Proof. Since $\mathrm{IB}_\beta[p(z|x)] = I(X;Z) - \beta I(Y;Z)$, let us calculate the first and second-order variation of $I(X;Z)$ and $I(Y;Z)$ w.r.t. $p(z|x)$, respectively. Through this derivation, we use $\epsilon \cdot h(z|x)$ as a perturbative function, for ease of deciding different orders of variations. We assume that $h(z|x)$ is continuous, and there exists a constant M such that $|\frac{h(z|x)}{p(z|x)}| < M$, $\forall(x,z) \in \mathcal{X}\times\mathcal{Z}$. We will finally absorb ϵ into $h(z|x)$.

Denote $I(X;Z) = F_1[p(z|x)]$. We have

$$F_1[p(z|x)] = I(X;Z) = \int dxdzp(z|x)p(x)\log\frac{p(z|x)}{p(z)}$$

In this paper, we implicitly assume that the integral (or summing) are only on the support of $p(x,y,z)$.

Since

$$p(z) = \int p(z|x)p(x)dx$$

We have

$$p(z)|_{p(z|x)+\epsilon h(z|x)} = p(z)|_{p(z|x)} + \epsilon\int h(z|x)p(x)dx$$

Expanding $F_1[p(z|x)+\epsilon h(z|x)]$ to the second order of ϵ, we have

$$
\begin{aligned}
&F_1[p(z|x)+\epsilon h(z|x)]\\
&=\int dxdzp(x)[p(z|x)+\epsilon h(z|x)]\log\frac{p(z|x)+\epsilon h(z|x)}{p(z)+\epsilon\int h(z|x')p(x')dx'}\\
&=\int dxdzp(x)p(z|x)\left(1+\epsilon\frac{h(z|x)}{p(z|x)}\right)\log\frac{p(z|x)\left(1+\epsilon\frac{h(z|x)}{p(z|x)}\right)}{p(z)\left(1+\epsilon\frac{\int h(z|x')p(x')dx'}{p(z)}\right)}\\
&=\int dxdzp(x)p(z|x)\left(1+\epsilon\frac{h(z|x)}{p(z|x)}\right)\log\left[\frac{p(z|x)}{p(z)}\left(1+\epsilon\frac{h(z|x)}{p(z|x)}\right)\left(1-\epsilon\frac{\int h(z|x')p(x')dx'}{p(z)}\right.\right.\\
&\left.\left.+\epsilon^2\left(\frac{\int h(z|x')p(x')dx'}{p(z)}\right)^2\right)\right]+\mathcal{O}(\epsilon^3)\\
&=\int dxdzp(x)p(z|x)\left(1+\epsilon\frac{h(z|x)}{p(z|x)}\right)\log\left[\frac{p(z|x)}{p(z)}\left(1+\epsilon\left(\frac{h(z|x)}{p(z|x)}-\frac{\int h(z|x')p(x')dx'}{p(z)}\right)\right.\right.\\
&\left.\left.+\epsilon^2\left(\frac{\int h(z|x')p(x')dx'}{p(z)}\right)^2-\epsilon^2\frac{h(z|x)}{p(z|x)}\frac{\int h(z|x')p(x')dx'}{p(z)}\right)\right]+\mathcal{O}(\epsilon^3)\\
&=\int dxdzp(x)p(z|x)\left(1+\epsilon\frac{h(z|x)}{p(z|x)}\right)\left[\log\frac{p(z|x)}{p(z)}+\epsilon\left(\frac{h(z|x)}{p(z|x)}-\frac{\int h(z|x')p(x')dx'}{p(z)}\right)\right.\\
&\left.+\epsilon^2\left(\frac{\int h(z|x')p(x')dx'}{p(z)}\right)^2-\epsilon^2\frac{h(z|x)}{p(z|x)}\frac{\int h(z|x')p(x')dx'}{p(z)}-\frac{1}{2}\epsilon^2\left(\frac{h(z|x)}{p(z|x)}-\frac{\int h(z|x')p(x')dx'}{p(z)}\right)^2\right]+\mathcal{O}(\epsilon^3)
\end{aligned}
$$

Collecting the first order terms of ϵ, we have

$$
\begin{aligned}
&\delta F_1[p(z|x)]\\
&=\epsilon\int dxdzp(x)p(z|x)\left(\frac{h(z|x)}{p(z|x)}-\frac{\int h(z|x')p(x')dx'}{p(z)}\right)+\epsilon\int dxdzp(x)p(z|x)\frac{h(z|x)}{p(z|x)}\log\frac{p(z|x)}{p(z)}\\
&=\epsilon\int dxdzp(x)h(z|x)-\epsilon\int dx'dzp(x')h(z|x')+\epsilon\int dxdzp(x)h(z|x)\log\frac{p(z|x)}{p(z)}\\
&=\epsilon\int dxdzp(x)h(z|x)\log\frac{p(z|x)}{p(z)}
\end{aligned}
$$

Collecting the second order terms of ϵ^2, we have

$$
\begin{aligned}
&\delta^2F_1[p(z|x)]\\
&=\epsilon^2\int dxdzp(x)p(z|x)\left[\left(\frac{\int h(z|x')p(x')dx'}{p(z)}\right)^2-\frac{h(z|x)}{p(z|x)}\frac{\int h(z|x')p(x')dx'}{p(z)}-\frac{1}{2}\left(\frac{h(z|x)}{p(z|x)}-\frac{\int h(z|x')p(x')dx'}{p(z)}\right)^2\right]\\
&+\epsilon^2\int dxdzp(x)p(z|x)\frac{h(z|x)}{p(z|x)}\left(\frac{h(z|x)}{p(z|x)}-\frac{\int h(z|x')p(x')dx'}{p(z)}\right)\\
&=\frac{\epsilon^2}{2}\int dxdz\frac{p(x)^2}{p(x,z)}h(z|x)^2-\frac{\epsilon^2}{2}\int dxdx'dz\frac{p(x)p(x')}{p(z)}h(z|x)h(z|x')
\end{aligned}
$$

Now let us calculate the first and second-order variation of $F_2[p(z|x)]=I(Z;Y)$. We have

$$F_2[p(z|x)]=I(Y;Z)=\int dydzp(z|y)p(y)\log\frac{p(y,z)}{p(y)p(z)}=\int dxdydzp(z|y)p(x,y)\log\frac{p(y,z)}{p(y)p(z)}$$

Using the Markov chain $Z\leftarrow X\leftrightarrow Y$, we have

$$p(y,z)=\int p(z|x)p(x,y)dx$$

Hence

$$p(y,z)|_{p(z|x)+\epsilon h(z|x)}=p(y,z)|_{p(z|x)}+\epsilon\int h(z|x)p(x,y)dx$$

Then expanding $F_2[p(z|x)+\epsilon h(z|x)]$ to the second order of ϵ, we have

$$
\begin{aligned}
&F_2[p(z|x)+\epsilon h(z|x)]\\
&=\int dxdydzp(x,y)p(z|x)\left(1+\epsilon\frac{h(z|x)}{p(z|x)}\right)\log\frac{p(y,z)\left(1+\epsilon\frac{\int h(z|x')p(x',y)dx'}{p(y,z)}\right)}{p(y)p(z)(1+\epsilon\frac{\int h(z|x'')p(x'')dx''}{p(z)})}\\
&=\int dxdydzp(x,y)p(z|x)\left(1+\epsilon\frac{h(z|x)}{p(z|x)}\right)\left[\log\frac{p(y,z)}{p(y)p(z)}+\epsilon\left(\frac{\int h(z|x')p(x',y)dx'}{p(y,z)}-\frac{\int h(z|x')p(x')dx'}{p(z)}\right)\right.\\
&\left.+\epsilon^2\left[\left(\frac{\int h(z|x')p(x')dx'}{p(z)}\right)^2-\frac{\int h(z|x')p(x',y)dx'}{p(y,z)}\frac{\int h(z|x'')p(x'')dx''}{p(z)}-\frac{1}{2}\left(\frac{\int h(z|x')p(x',y)dx'}{p(y,z)}-\frac{\int h(z|x')p(x')dx'}{p(z)}\right)^2\right]\right.\\
&+\mathcal{O}(\epsilon^3)
\end{aligned}
$$

Collecting the first order terms of ϵ, we have

$$
\begin{aligned}
&\delta F_2[p(z|x)]\\
&=\epsilon\int dxdydzp(x,y)h(z|x)\log\frac{p(y,z)}{p(y)p(z)}+\epsilon\int dxdydzp(x,y)p(z|x)\frac{\int h(z|x')p(x',y)dx'}{p(y,z)}\\
&-\epsilon\int dxdydzp(x,y)p(z|x)\frac{\int h(z|x')p(x')dx'}{p(z)}\\
&=\epsilon\int dxdydzp(x,y)h(z|x)\log\frac{p(y,z)}{p(y)p(z)}+\epsilon\int dx'dydzh(z|x')p(x',y)-\epsilon\int dzh(z|x')p(x')dx'\\
&=\epsilon\int dxdydzp(x,y)h(z|x)\log\frac{p(z|y)}{p(z)}
\end{aligned}
$$

Collecting the second order terms of ϵ, we have

$$
\begin{aligned}
&\delta^2F_2[p(z|x)]\\
&=\epsilon^2\int dxdydzp(x,y)p(z|x)\left[\left(\frac{\int h(z|x')p(x')dx'}{p(z)}\right)^2-\frac{\int h(z|x')p(x',y)dx'}{p(y,z)}\frac{\int h(z|x'')p(x'')dx''}{p(z)}\right]\\
&-\frac{\epsilon^2}{2}\int dxdydzp(x,y)p(z|x)\left(\frac{\int h(z|x')p(x',y)dx'}{p(y,z)}-\frac{\int h(z|x')p(x')dx'}{p(z)}\right)^2\\
&+\epsilon^2\int dxdydzp(x,y)p(z|x)\frac{h(z|x)}{p(z|x)}\left(\frac{\int h(z|x')p(x',y)dx'}{p(y,z)}-\frac{\int h(z|x')p(x')dx'}{p(z)}\right)\\
&=\frac{\epsilon^2}{2}\int dxdx'dydz\frac{p(x,y)p(x',y)}{p(y,z)}h(z|x)h(z|x')-\frac{\epsilon^2}{2}\int dxdx'dz\frac{p(x)p(x')}{p(z)}h(z|x)h(z|x')
\end{aligned}
$$

Finally, we have

$$
\begin{aligned}
\delta\mathrm{IB}_\beta[p(z|x)]&=\delta F_1[p(z|x)]-\beta\cdot\delta F_2[p(z|x)]\\
&=\epsilon\left(\int dxdzp(x)h(z|x)\log\frac{p(z|x)}{p(z)}-\beta\int dxdydzp(x,y)h(z|x)\log\frac{p(z|y)}{p(z)}\right)
\end{aligned}
\tag{A1}
$$

$$\begin{aligned}\delta^2\mathrm{IB}_\beta[p(z|x)] =&\delta^2 F_1[p(z|x)] - \beta\cdot\delta^2 F_2[p(z|x)]\\ =&\frac{\epsilon^2}{2}\int dxdz\frac{p(x)^2}{p(x,z)}h(z|x)^2 - \frac{\epsilon^2}{2}\int dxdx'dz\frac{p(x)p(x')}{p(z)}h(z|x)h(z|x')\\ &-\beta\epsilon^2\left[\frac{1}{2}\int dxdx'dydz\frac{p(x,y)p(x',y)}{p(y,z)}h(z|x)h(z|x') - \frac{1}{2}\int dxdx'dz\frac{p(x)p(x')}{p(z)}h(z|x)h(z|x')\right]\\ =&\frac{\epsilon^2}{2}\left[\int dxdz\frac{p(x)^2}{p(x,z)}h(z|x)^2\right.\\ &\left.-\beta\int dxdx'dydz\frac{p(x,y)p(x',y)}{p(y,z)}h(z|x)h(z|x') + (\beta-1)\int dxdx'dz\frac{p(x)p(x')}{p(z)}h(z|x)h(z|x')\right]\end{aligned}$$

Absorb ϵ into $h(z|x)$, we get rid of the ϵ factor and obtain the final expression in Lemma A1. □

Appendix A.6. Proof of Lemma 2

Proof. Using Lemma A1, we have

$$\delta\mathrm{IB}_\beta[p(z|x)] = \int dxdzp(x)h(z|x)\log\frac{p(z|x)}{p(z)} - \beta\int dxdydzp(x,y)h(z|x)\log\frac{p(z|y)}{p(z)}$$

Let $p(z|x) = p(z)$ (the trivial representation), we have that $\log\frac{p(z|x)}{p(z)} \equiv 0$. Therefore, the two integrals are both 0. Hence,

$$\delta\mathrm{IB}_\beta[p(z|x)]\big|_{p(z|x)=p(z)} \equiv 0$$

Therefore, the $p(z|x) = p(z)$ is a stationary solution for $\mathrm{IB}_\beta[p(z|x)]$. □

Appendix A.7. Proof of Theorem 3

Proof. Firstly, from the necessary condition of $\beta > 1$ in Section 3, we have that any sufficient condition for IB_β-learnability should be able to deduce $\beta > 1$.

Now using Theorem 2, a sufficient condition for (X, Y) to be IB_β-learnable is that there exists $h(z|x)$ with $\int h(z|x)dx = 0$ such that $\delta^2\mathrm{IB}_\beta[p(z|x)] < 0$ at $p(z|x) = p(x)$.

At the trivial representation, $p(z|x) = p(z)$ and hence $p(x,z) = p(x)p(z)$. Due to the Markov chain $Z \leftarrow X \leftrightarrow Y$, we have $p(y,z) = p(y)p(z)$. Substituting them into the $\delta^2\mathrm{IB}_\beta[p(z|x)]$ in Lemma A1, the condition becomes: there exists $h(z|x)$ with $\int h(z|x)dz = 0$, such that

$$\begin{aligned}&0 > \delta^2\mathrm{IB}_\beta[p(z|x)] =\\ &\frac{1}{2}\left[\int dxdz\frac{p(x)^2}{p(x)p(z)}h(z|x)^2 - \beta\int dxdx'dydz\frac{p(x,y)p(x',y)}{p(y)p(z)}h(z|x)h(z|x') + (\beta-1)\int dxdx'dz\frac{p(x)p(x')}{p(z)}h(z|x)h(z|x')\right]\end{aligned} \tag{A2}$$

Rearranging terms and simplifying, we have

$$\int\frac{dz}{p(z)}G[h(z|x)] = \int\frac{dz}{p(z)}\left[\int dxh(z|x)^2p(x) - \beta\int\frac{dy}{p(y)}\left(\int dxh(z|x)p(x)p(y|x)\right)^2 + (\beta-1)\left(\int dxh(z|x)p(x)\right)^2\right] < 0$$

where

$$G[h(x)] = \int dxh(x)^2p(x) - \beta\int\frac{dy}{p(y)}\left(\int dxh(x)p(x)p(y|x)\right)^2 + (\beta-1)\left(\int dxh(x)p(x)\right)^2$$

Now we prove that the condition that $\exists h(z|x)$ s.t. $\int\frac{dz}{p(z)}G[h(z|x)] < 0$ is equivalent to the condition that $\exists h(x)$ s.t. $G[h(x)] < 0$.

If $\forall h(z|x)$, $G[h(z|x)] \geq 0$, then we have $\forall h(z|x)$, $\int\frac{dz}{p(z)}G[h(z|x)] \geq 0$. Therefore, if $\exists h(z|x)$ s.t. $\int\frac{dz}{p(z)}G[h(z|x)] < 0$, we have that $\exists h(z|x)$ s.t. $G[h(z|x)] < 0$. Since the functional $G[h(z|x)]$ does not

contain integration over z, we can treat the z in $G[h(z|x)]$ as a parameter and we have that $\exists h(x)$ s.t. $G[h(x)] < 0$.

Conversely, if there exists an certain function $h(x)$ such that $G[h(x)] < 0$, we can find some $h_2(z)$ such that $\int h_2(z)dz = 0$ and $\int \frac{h_2^2(z)}{p(z)}dz > 0$, and let $h_1(z|x) = h(x)h_2(z)$. Now we have

$$\int \frac{dz}{p(z)} G[h(z|x)] = \int \frac{h_2^2(z)dz}{p(z)} G[h(x)] = G[h(x)] \int \frac{h_2^2(z)dz}{p(z)} < 0$$

In other words, the condition Equation (A2) is equivalent to requiring that there exists an $h(x)$ such that $G[h(x)] < 0$. Hence, a sufficient condition for IB$_\beta$-learnability is that there exists an $h(x)$ such that

$$G[h(x)] = \int dx h(x)^2 p(x) - \beta \int \frac{dy}{p(y)} \left(\int dx h(x) p(x) p(y|x) \right)^2 + (\beta - 1) \left(\int dx h(x) p(x) \right)^2 < 0 \tag{A3}$$

When $h(x) = C =$ constant in the entire input space $\mathcal{X}$, Equation (A3) becomes:

$$C^2 - \beta C^2 + (\beta - 1)C^2 < 0$$

which cannot be true. Therefore, $h(x) =$ constant cannot satisfy Equation (A3).

Rearranging terms and simplifying, we have

$$\beta \left[\int \frac{dy}{p(y)} \left(\int dx h(x) p(x) p(y|x) \right)^2 - \left(\int dx h(x) p(x) \right)^2 \right] > \int dx h(x)^2 p(x) - \left(\int dx h(x) p(x) \right)^2 \tag{A4}$$

Written in the form of expectations, we have

$$\beta \cdot \left(\mathbb{E}_{y \sim p(y)} \left[\left(\mathbb{E}_{x \sim p(x|y)} [h(x)] \right)^2 \right] - \left(\mathbb{E}_{x \sim p(x)} [h(x)] \right)^2 \right) > \mathbb{E}_{x \sim p(x)} [h(x)^2] - \left(\mathbb{E}_{x \sim p(x)} [h(x)] \right)^2 \tag{A5}$$

Since the square function is convex, using Jensen's inequality on the L.H.S. of Equation (A5), we have

$$\mathbb{E}_{y \sim p(y)} \left[\left(\mathbb{E}_{x \sim p(x|y)} [h(x)] \right)^2 \right] \geq \left(\mathbb{E}_{y \sim p(y)} \left[\mathbb{E}_{x \sim p(x|y)} [h(x)] \right] \right)^2 = \left(\mathbb{E}_{x \sim p(x)} [h(x)] \right)^2$$

The equality holds iff $\mathbb{E}_{x \sim p(x|y)} [h(x)]$ is constant w.r.t. y, i.e., Y is independent of X. Therefore, in order for Equation (A5) to hold, we require that Y is not independent of X.

Using Jensen's inequality on the innter expectation on the L.H.S. of Equation (A5), we have

$$\mathbb{E}_{y \sim p(y)} \left[\left(\mathbb{E}_{x \sim p(x|y)} [h(x)] \right)^2 \right] \leq \mathbb{E}_{y \sim p(y)} \left[\mathbb{E}_{x \sim p(x|y)} [h(x)^2] \right] = \mathbb{E}_{x \sim p(x)} [h(x)^2] \tag{A6}$$

The equality holds when $h(x)$ is a constant. Since we require that $h(x)$ is not a constant, we have that the equality cannot be reached.

Similarly, using Jensen's inequality on the R.H.S. of Equation (A5), we have that

$$\mathbb{E}_{x \sim p(x)} [h(x)^2] > \left(\mathbb{E}_{x \sim p(x)} [h(x)] \right)^2$$

where we have used the requirement that $h(x)$ cannot be constant.

Under the constraint that Y is not independent of X, we can divide both sides of Equation (A5), and obtain the condition: there exists an $h(x)$ such that

$$\beta > \frac{\mathbb{E}_{x\sim p(x)}[h(x)^2] - \left(\mathbb{E}_{x\sim p(x)}[h(x)]\right)^2}{\mathbb{E}_{y\sim p(y)}\left[\left(\mathbb{E}_{x\sim p(x|y)}[h(x)]\right)^2\right] - \left(\mathbb{E}_{x\sim p(x)}[h(x)]\right)^2}$$

i.e.,

$$\beta > \inf_{h(x)} \frac{\mathbb{E}_{x\sim p(x)}[h(x)^2] - \left(\mathbb{E}_{x\sim p(x)}[h(x)]\right)^2}{\mathbb{E}_{y\sim p(y)}\left[\left(\mathbb{E}_{x\sim p(x|y)}[h(x)]\right)^2\right] - \left(\mathbb{E}_{x\sim p(x)}[h(x)]\right)^2}$$

which proves the condition of Theorem 3.

Furthermore, from Equation (A6) we have

$$\beta_0[h(x)] > 1$$

for $h(x) \not\equiv$ const, which satisfies the necessary condition of $\beta > 1$ in Section 3.

Proof of lower bound of slope of the Pareto frontier at the origin: Now we prove the second statement of Theorem 3. Since $\delta I(X;Z) = 0$ and $\delta I(Y;Z) = 0$ according to Lemma 2, we have $\left(\frac{\Delta I(Y;Z)}{\Delta I(X;Z)}\right)^{-1} = \left(\frac{\delta^2 I(Y;Z)}{\delta^2 I(X;Z)}\right)^{-1}$. Substituting into the expression of $\delta^2 I(Y;Z)$ and $\delta^2 I(X;Z)$ from Lemma A1, we have

$$\begin{aligned}
&\left(\frac{\Delta I(Y;Z)}{\Delta I(X;Z)}\right)^{-1} \\
&= \left(\frac{\delta^2 I(Y;Z)}{\delta^2 I(X;Z)}\right)^{-1} \\
&= \frac{\frac{\epsilon^2}{2}\int dxdz \frac{p(x)^2}{p(x)p(z)} h(z|x)^2 - \frac{\epsilon^2}{2}\int dxdx'dz \frac{p(x)p(x')}{p(z)} h(z|x)h(z|x')}{\frac{\epsilon^2}{2}\int dxdx'dydz \frac{p(x,y)p(x',y)}{p(y)p(z)} h(z|x)h(z|x') - \frac{\epsilon^2}{2}\int dxdx'dz \frac{p(x)p(x')}{p(z)} h(z|x)h(z|x')} \\
&= \frac{\left(\int dxp(x)h(x)^2 - \int dxdx'p(x)p(x')h(x)h(z|x')\right)\int \frac{h_2(z)^2}{p(z)}dz}{\left(\int dxdx'dy \frac{p(x,y)p(x',y)}{p(y)} h(x)h(z|x') - \int dxdx'p(x)p(x')h(x)h(z|x')\right)\int \frac{h_2(z)^2}{p(z)}dz} \\
&= \frac{\int dxp(x)h(x)^2 - \int dxdx'p(x)p(x')h(x)h(z|x')}{\int dxdx'dy \frac{p(x,y)p(x',y)}{p(y)} h(x)h(z|x') - \int dxdx'p(x)p(x')h(x)h(z|x')} \\
&= \frac{\mathbb{E}_{x\sim p(x)}[h(x)^2] - \left(\mathbb{E}_{x\sim p(x)}[h(x)]\right)^2}{\mathbb{E}_{y\sim p(y)}\left[\left(\mathbb{E}_{x\sim p(x|y)}[h(x)]\right)^2\right] - \left(\mathbb{E}_{x\sim p(x)}[h(x)]\right)^2} \\
&= \frac{\frac{\mathbb{E}_{x\sim p(x)}[h(x)^2]}{\left(\mathbb{E}_{x\sim p(x)}[h(x)]\right)^2} - 1}{\mathbb{E}_{y\sim p(y)}\left[\left(\frac{\mathbb{E}_{x\sim p(x|y)}[h(x)]}{\mathbb{E}_{x\sim p(x)}[h(x)]}\right)^2\right] - 1} \\
&= \beta_0[h(x)]
\end{aligned}$$

Therefore, $\left(\inf_{h(x)} \beta_0[h(x)]\right)^{-1}$ gives the largest slope of $\Delta I(Y;Z)$ vs. $\Delta I(X;Z)$ for perturbation function of the form $h_1(z|x) = h(x)h_2(z)$ satisfying $\int h_2(z)dz = 0$ and $\int \frac{h_2^2(z)}{p(z)}dz > 0$, which is a lower

bound of slope of $\Delta I(Y;Z)$ vs. $\Delta I(X;Z)$ for all possible perturbation function $h_1(z|x)$. The latter is the slope of the Pareto frontier of the $I(Y;Z)$ vs. $I(X;Z)$ curve at the origin.

Inflection point for general *Z*: If we *do not* assume that Z is at the origin of the information plane, but at some general stationary solution Z^* with $p(z|x)$, we define

$$\begin{aligned}
\beta^{(2)}[h(x)] &= \left(\frac{\delta^2 I(Y;Z)}{\delta^2 I(X;Z)}\right)^{-1} \\
&= \frac{\frac{\epsilon^2}{2}\int dxdz\frac{p(x)^2}{p(x,z)}h(z|x)^2 - \frac{\epsilon^2}{2}\int dxdx'dz\frac{p(x)p(x')}{p(z)}h(z|x)h(z|x')}{\frac{\epsilon^2}{2}\int dxdx'dydz\frac{p(x,y)p(x',y)}{p(y,z)}h(z|x)h(z|x') - \frac{\epsilon^2}{2}\int dxdx'dz\frac{p(x)p(x')}{p(z)}h(z|x)h(z|x')} \\
&= \frac{\int dxdz\frac{p(x)^2}{p(x,z)}h(z|x)^2 - \int dxdx'dz\frac{p(x)p(x')}{p(z)}h(z|x)h(z|x')}{\int dxdx'dydz\frac{p(x,y)p(x',y)}{p(y,z)}h(z|x)h(z|x') - \int dxdx'dz\frac{p(x)p(x')}{p(z)}h(z|x)h(z|x')} \\
&= \frac{\int \frac{dz}{p(z)}\left[\int dx\frac{p(x)^2}{p(x|z)}h(z|x)^2 - \left(\int dxp(x)h(z|x)\right)^2\right]}{\int \frac{dz}{p(z)}\left[\int \frac{dy}{p(y|z)}\left(\int dxp(x,y)h(z|x)\right)^2 - \left(\int dxp(x)h(z|x)\right)^2\right]} \\
&= \frac{\int \frac{dz}{p(z)}\left[\frac{\int dx\frac{p(x)^2}{p(x|z)}h(z|x)^2}{\left(\int dxp(x)h(z|x)\right)^2} - 1\right]}{\int \frac{dz}{p(z)}\left[\frac{\int \frac{dy}{p(y|z)}\left(\int dxp(x,y)h(z|x)\right)^2}{\left(\int dxp(x)h(z|x)\right)^2} - 1\right]} \\
&= \frac{\int dz\left[\frac{\int dx\frac{p(x)}{p(z|x)}h(z|x)^2}{\left(\int dxp(x)h(z|x)\right)^2} - \frac{1}{p(z)}\right]}{\int dz\left[\frac{\int \frac{dy}{p(z|y)p(y)}\left(\int dxp(x,y)h(z|x)\right)^2}{\left(\int dxp(x)h(z|x)\right)^2} - \frac{1}{p(z)}\right]} \\
&= \frac{\int dz\left[\int dx\frac{p(x)}{p(z|x)}h(z|x)^2 - \frac{1}{p(z)}\left(\int dxp(x)h(z|x)\right)^2\right]}{\int dz\left[\int \frac{dy}{p(z|y)p(y)}\left(\int dxp(x,y)h(z|x)\right)^2 - \frac{1}{p(z)}\left(\int dxp(x)h(z|x)\right)^2\right]}
\end{aligned}$$

which reduces to $\beta_0[h(x)]$ when $p(z|x) = p(z)$. When

$$\beta > \inf_{h(z|x)} \beta^{(2)}[h(z|x)] \tag{A7}$$

it becomes a non-stable solution (non-minimum), and we will have other Z that achieves a better $\mathrm{IB}_\beta(X,Y;Z)$ than the current Z^*. □

Appendix A.8. What IB First Learns at Its Onset of Learning

In this section, we prove that at the onset of learning, if letting $h(z|x) = h^*(x)h_2(z)$, we have

$$p_\beta(y|x) = p(y) + \epsilon^2 C_z(h^*(x) - \overline{h}_x^*)\int p(x,y)(h^*(x) - \overline{h}_x^*)dx \tag{A8}$$

where $p_\beta(y|x)$ is the estimated $p(y|x)$ by IB for a certain β, $h^*(x) = \inf_{h(x)} \beta_0[h(x)]$, $\overline{h}_x^* = \int h^*(x)p(x)dx$, $C_z = \int \frac{h_2^2(z)}{p(z)}dz$ is a constant.

Proof. In IB, we use $p_\beta(z|x)$ to obtain Z from X, then obtain the prediction of Y from Z using $p_\beta(y|z)$. Here we use subscript β to denote the probability (density) at the optimum of $\mathrm{IB}_\beta[p(z|x)]$ at a specific β. We have

$$\begin{aligned} p_\beta(y|x) &= \int p_\beta(y|z)p_\beta(z|x)dz \\ &= \int dz \frac{p_\beta(y,z)p_\beta(z|x)}{p_\beta(z)} \\ &= \int dz \frac{p_\beta(z|x)}{p_\beta(z)} \int p(x',y)p_\beta(z|x')dx' \end{aligned}$$

When we have a small perturbation $\epsilon \cdot h(z|x)$ at the trivial representation, $p_\beta(z|x) = p_{\beta_0}(z) + \epsilon \cdot h(z|x)$, we have $p_\beta(z) = p_{\beta_0}(z) + \epsilon \cdot \int h(z|x'')p(x'')dx''$. Substituting, we have

$$\begin{aligned} p_\beta(y|x) &= \int dz \frac{p_{\beta_0}(z)\left(1+\epsilon\cdot\frac{h(z|x)}{p_{\beta_0}(z)}\right)}{p_{\beta_0}(z)\left(1+\epsilon\cdot\frac{\int h(z|x'')p(x'')dx''}{p_{\beta_0}(z)}\right)} \int p(x',y)p_{\beta_0}(z)\left(1+\epsilon\cdot\frac{h(z|x')}{p_{\beta_0}(z)}\right)dx' \\ &= \int dz \frac{1+\epsilon\cdot\frac{h(z|x)}{p_{\beta_0}(z)}}{1+\epsilon\cdot\frac{\int h(z|x'')p(x'')dx''}{p_{\beta_0}(z)}} \int p(x',y)p_{\beta_0}(z)\left(1+\epsilon\cdot\frac{h(z|x')}{p_{\beta_0}(z)}\right)dx' \end{aligned}$$

The 0th-order term is $\int dzdx'p(x',y)p_{\beta_0}(z) = p(y)$. The first-order term is

$$\begin{aligned} \delta p_\beta(z|x) =& \epsilon \cdot \int dzdx' \left(h(z|x) + h(z|x') - \int h(z|x'')p(x'')dx''\right)p(x',y) \\ =& \epsilon \cdot \int dx' \left(\int dz h(z|x) + \int dz h(z|x')\right) - \epsilon \cdot \int dx'dx''p(x',y)p(x'')\int dz h(z|x'') \\ =& 0 - 0 \\ =& 0 \end{aligned}$$

since we have $\int h(z|x)dz = 0$ for any x.

For the second-order term, using $h(z|x) = h^*(x)h_2(z)$ and $C_z = \int \frac{dz}{p_{\beta_0}(z)}h_2^2(z)$, it is

$$\begin{aligned} \delta^2 p_\beta(y|x) =& \epsilon^2 \cdot \int dz \left(\frac{\int h(z|x'')p(x'')dx''}{p_{\beta_0}(z)}\right)^2 \int p(x',y)p_{\beta_0}(z)dx' \\ & - \epsilon^2 \cdot \int dz \frac{h(z|x)\int h(z|x'')p(x'')dx''}{(p_{\beta_0}(z))^2} \int p(x',y)p_{\beta_0}(z)dx' \\ & + \epsilon^2 \int dz \left(h(z|x) - \int h(z|x'')p(x'')dx\right) \int p(x',y)\frac{h(z|x')}{p_{\beta_0}(z)}dx' \\ =& \epsilon^2 C_z \cdot \left(\int h^*(x'')p(x'')dx''\right)^2 p(y) \\ & - \epsilon^2 C_z \cdot h^*(x) \int h^*(x'')p(x'')dx''p(y) \\ & + \epsilon^2 C_z \cdot h^*(x) \int p(x',y)h^*(x')dx' \\ & - \epsilon^2 C_z \cdot \int h^*(x'')p(x'')dx \int p(x',y)h^*(x')dx' \\ =& \epsilon^2 C_z (h^*(x) - \overline{h}_x^*) \left[\left(\int p(x',y)h^*(x')dx'\right) - \overline{h}_x^* p(y)\right] \\ =& \epsilon^2 C_z (h^*(x) - \overline{h}_x^*) \int p(x',y)\left(h^*(x') - \overline{h}_x^*\right)dx' \end{aligned}$$

where $\overline{h}_x^* = \int h^*(x)p(x)dx$. Combining everything, we have up to the second order,

$$p_\beta(y|x) = p(y) + \epsilon^2 C_z(h^*(x) - \overline{h}_x^*) \int p(x,y)(h^*(x) - \overline{h}_x^*)dx$$

□

Appendix A.9. Proof of Theorem 4

Proof. According to Theorem 3, a sufficient condition for (X, Y) to be IB$_\beta$-learnable is that X and Y are not independent, and

$$\beta > \inf_{h(x)} \frac{\frac{\mathbb{E}_{x\sim p(x)}[h(x)^2]}{\left(\mathbb{E}_{x\sim p(x)}[h(x)]\right)^2} - 1}{\mathbb{E}_{y\sim p(y)}\left[\left(\frac{\mathbb{E}_{x\sim p(x|y)}[h(x)]}{\mathbb{E}_{x\sim p(x)}[h(x)]}\right)^2\right] - 1} \tag{A9}$$

We can assume a specific form of $h(x)$, and obtain a (potentially stronger) sufficient condition. Specifically, we let

$$h(x) = \begin{cases} 1, & x \in \Omega_x \\ 0, & \text{otherwise} \end{cases} \tag{A10}$$

for certain $\Omega_x \subset \mathcal{X}$. Substituting into Equation (A10), we have that a sufficient condition for (X, Y) to be IB$_\beta$-learnable is

$$\beta > \inf_{\Omega_x \subset \mathcal{X}} \frac{\frac{p(\Omega_x)}{p(\Omega_x)^2} - 1}{\int dy p(y) \left(\frac{\int_{x\in\Omega_x} dx p(x|y)dx}{p(\Omega_x)}\right)^2 - 1} > 0 \tag{A11}$$

where $p(\Omega_x) = \int_{x\in\Omega_x} p(x)dx$.

The denominator of Equation (A11) is

$$\begin{aligned} &\int dy p(y) \left(\frac{\int_{x\in\Omega_x} dx p(x|y)dx}{p(\Omega_x)}\right)^2 - 1 \\ &= \int dy p(y) \left(\frac{p(\Omega_x|y)}{p(\Omega_x)}\right)^2 - 1 \\ &= \int dy \frac{p(y|\Omega_x)^2}{p(y)} - 1 \\ &= \mathbb{E}_{y\sim p(y|\Omega_x)}\left[\frac{p(y|\Omega_x)}{p(y)} - 1\right] \end{aligned}$$

Using the inequality $x - 1 \geq \log x$, we have

$$\mathbb{E}_{y\sim p(y|\Omega_x)}\left[\frac{p(y|\Omega_x)}{p(y)} - 1\right] \geq \mathbb{E}_{y\sim p(y|\Omega_x)}\left[\log\frac{p(y|\Omega_x)}{p(y)}\right] \geq 0$$

Both equalities hold iff $p(y|\Omega_x) \equiv p(y)$, at which the denominator of Equation (A11) is equal to 0 and the expression inside the infimum diverge, which will not contribute to the infimum. Except this scenario, the denominator is greater than 0. Substituting into Equation (A11), we have that a sufficient condition for (X, Y) to be IB$_\beta$-learnable is

$$\beta > \inf_{\Omega_x \subset \mathcal{X}} \frac{\frac{p(\Omega_x)}{p(\Omega_x)^2} - 1}{\mathbb{E}_{y\sim p(y|\Omega_x)}\left[\frac{p(y|\Omega_x)}{p(y)} - 1\right]} \tag{A12}$$

Since Ω_x is a subset of $\mathcal{X}$, by the definition of $h(x)$ in Equation (A10), $h(x)$ is not a constant in the entire $\mathcal{X}$. Hence the numerator of Equation (A12) is positive. Since its denominator is also positive, we can then neglect the "> 0", and obtain the condition in Theorem 4.

Since the $h(x)$ used in this theorem is a subset of the $h(x)$ used in Theorem 3, the infimum for Equation (5) is greater than or equal to the infimum in Equation (2). Therefore, according to the second statement of Theorem 3, we have that the $\left(\inf_{\Omega_x \subset \mathcal{X}} \beta_0(\Omega_x)\right)^{-1}$ is also a lower bound of the slope for the Pareto frontier of $I(Y;Z)$ vs. $I(X;Z)$ curve.

Now we prove that the condition Equation (5) is invariant to invertible mappings of X. In fact, if $X' = g(X)$ is a uniquely invertible map (if X is continuous, g is additionally required to be continuous), let $\mathcal{X}' = \{g(x)|x \in \Omega_x\}$, and denote $g(\Omega_x) \equiv \{g(x)|x \in \Omega_x\}$ for any $\Omega_x \subset \mathcal{X}$, we have $p(g(\Omega_x)) = p(\Omega_x)$, and $p(y|g(\Omega_x)) = p(y|\Omega_x)$. Then for dataset (X, Y), let $\Omega'_x = g(\Omega_x)$, we have

$$\frac{\frac{1}{p(\Omega'_x)} - 1}{\mathbb{E}_{y\sim p(y|\Omega'_x)}\left[\frac{p(y|\Omega'_x)}{p(y)} - 1\right]} = \frac{\frac{1}{p(\Omega_x)} - 1}{\mathbb{E}_{y\sim p(y|\Omega_x)}\left[\frac{p(y|\Omega_x)}{p(y)} - 1\right]} \tag{A13}$$

Additionally we have $\mathcal{X}' = g(\mathcal{X})$. Then

$$\inf_{\Omega'_x \subset \mathcal{X}'} \frac{\frac{1}{p(\Omega'_x)} - 1}{\mathbb{E}_{y\sim p(y|\Omega'_x)}\left[\frac{p(y|\Omega'_x)}{p(y)} - 1\right]} = \inf_{\Omega_x \subset \mathcal{X}} \frac{\frac{1}{p(\Omega_x)} - 1}{\mathbb{E}_{y\sim p(y|\Omega_x)}\left[\frac{p(y|\Omega_x)}{p(y)} - 1\right]} \tag{A14}$$

For dataset $(X', Y) = (g(X), Y)$, applying Theorem 4 we have that a sufficient condition for it to be IB$_\beta$-learnable is

$$\beta > \inf_{\Omega'_x \subset \mathcal{X}'} \frac{\frac{1}{p(\Omega'_x)} - 1}{\mathbb{E}_{y\sim p(y|\Omega'_x)}\left[\frac{p(y|\Omega'_x)}{p(y)} - 1\right]} = \inf_{\Omega_x \subset \mathcal{X}} \frac{\frac{1}{p(\Omega_x)} - 1}{\mathbb{E}_{y\sim p(y|\Omega_x)}\left[\frac{p(y|\Omega_x)}{p(y)} - 1\right]} \tag{A15}$$

where the equality is due to Equation (A14). Comparing with the condition for IB$_\beta$-learnability for (X, Y) (Equation (5)), we see that they are the same. Therefore, the condition given by Theorem 4 is invariant to invertible mapping of X. □

Appendix A.10. Proof of Corollary 1 and Corollary 2

Appendix A.10.1. Proof of Corollary 1

Proof. We use Theorem 4. Let Ω_x contain all elements x whose true class is y^* for some certain y^*, and 0 otherwise. Then we obtain a (potentially stronger) sufficient condition. Since the probability $p(y|y^*, x) = p(y|y^*)$ is class-conditional, we have

$$\begin{aligned}&\inf_{\Omega_x \subset \mathcal{X}} \frac{\frac{1}{p(\Omega_x)} - 1}{\mathbb{E}_{y\sim p(y|\Omega_x)}\left[\frac{p(y|\Omega_x)}{p(y)} - 1\right]}\\ =&\inf_{y^*} \frac{\frac{1}{p(y^*)} - 1}{\mathbb{E}_{y\sim p(y|y^*)}\left[\frac{p(y|y^*)}{p(y)} - 1\right]}\end{aligned}$$

By requiring $\beta > \inf_{y^*} \frac{\frac{1}{p(y^*)} - 1}{\mathbb{E}_{y\sim p(y|y^*)}\left[\frac{p(y|y^*)}{p(y)} - 1\right]}$, we obtain a sufficient condition for IB$_\beta$ learnability. □

Appendix A.10.2. Proof of Corollary 2

Proof. We again use Theorem 4. Since Y is a deterministic function of X, let $Y = f(X)$. By the assumption that Y contains at least one value y such that its probability $p(y) > 0$, we let Ω_x contain only x such that $f(x) = y$. Substituting into Equation (5), we have

$$\begin{aligned}
&\frac{\frac{1}{p(\Omega_x)} - 1}{\mathbb{E}_{y\sim p(y|\Omega_x)}\left[\frac{p(y|\Omega_x)}{p(y)} - 1\right]} \\
=&\frac{\frac{1}{p(y)} - 1}{\mathbb{E}_{y\sim p(y|\Omega_x)}\left[\frac{1}{p(y)} - 1\right]} \\
=&\frac{\frac{1}{p(y)} - 1}{\frac{1}{p(y)} - 1} \\
=&1
\end{aligned}$$

□

Therefore, the sufficient condition becomes $\beta > 1$.

Appendix A.11. β_0, Hypercontractivity Coefficient, Contraction Coefficient, $\beta_0[h(x)]$, and Maximum Correlation

In this section, we prove the relations between the IB-Learnability threshold β_0, the hypercontractivity coefficient $\xi(X;Y)$, the contraction coefficient $\eta_{\mathrm{KL}}(p(y|x), p(x))$, $\beta_0[h(x)]$ in Equation (2), and maximum correlation $\rho_m(X,Y)$, as follows:

$$\frac{1}{\beta_0} = \xi(X;Y) = \eta_{\mathrm{KL}}(p(y|x), p(x)) \geq \sup_{h(x)} \frac{1}{\beta_0[h(x)]} = \rho_m^2(X;Y) \tag{A16}$$

Proof. The hypercontractivity coefficient ξ is defined as [16]:

$$\xi(X;Y) \equiv \sup_{Z-X-Y} \frac{I(Y;Z)}{I(X;Z)}$$

By our definition of IB-learnability, (X, Y) is IB-Learnable iff there exists Z obeying the Markov chain $Z - X - Y$, such that

$$I(X;Z) - \beta \cdot I(Y;Z) < 0 = IB_\beta(X,Y;Z)|_{p(z|x)=p(z)}$$

Or equivalently there exists Z obeying the Markov chain $Z - X - Y$ such that

$$0 < \frac{1}{\beta} < \frac{I(Y;Z)}{I(X;Z)} \tag{A17}$$

By Theorem 1, the IB-Learnability region for β is $(\beta_0, +\infty)$, or equivalently the IB-Learnability region for $1/\beta$ is

$$0 < \frac{1}{\beta} < \frac{1}{\beta_0} \tag{A18}$$

Comparing Equations (A17) and (A18), we have that

$$\frac{1}{\beta_0} = \sup_{Z-X-Y} \frac{I(Y;Z)}{I(X;Z)} = \xi(X;Y) \tag{A19}$$

In Anantharam et al. [16], the authors prove that

$$\xi(X;Y) = \eta_{\mathrm{KL}}(p(y|x), p(x)) \tag{A20}$$

where the contraction coefficient $\eta_{\mathrm{KL}}(p(y|x), p(x))$ is defined as

$$\eta_{\mathrm{KL}}(p(y|x), p(x)) = \sup_{r(x) \neq p(x)} \frac{\mathbb{D}_{\mathrm{KL}}(r(y)||p(y))}{\mathbb{D}_{\mathrm{KL}}(r(x)||p(x))}$$

where $p(y) = \mathbb{E}_{x\sim p(x)}[p(y|x)]$ and $r(y) = \mathbb{E}_{x\sim r(x)}[p(y|x)]$. Treating $p(y|x)$ as a channel, the contraction coefficient measures how much the two distributions $r(x)$ and $p(x)$ becomes "nearer" (as measured by the KL-divergence) after passing through the channel.

In Anantharam et al. [16], the authors also provide a counterexample to an earlier result by Erkip and Cover [31] that incorrectly proved $\xi(X;Y) = \rho_m^2(X;Y)$. In the specific counterexample Anantharam et al. [16] design, $\xi(X;Y) > \rho_m^2(X;Y)$.

The maximum correlation is defined as $\rho_m(X;Y) \equiv \max_{f,g} \mathbb{E}[f(X)g(Y)]$ where $f(X)$ and $g(Y)$ are real-valued random variables such that $\mathbb{E}[f(X)] = \mathbb{E}[g(Y)] = 0$ and $\mathbb{E}[f^2(X)] = \mathbb{E}[g^2(Y)] = 1$ [20,21].

Now we prove $\xi(X;Y) \geq \rho_m^2(X;Y)$, based on Theorem 3. To see this, we use the alternate characterization of $\rho_m(X;Y)$ by Rényi [32]:

$$\rho_m^2(X;Y) = \max_{f(X):\mathbb{E}[f(X)]=0,\mathbb{E}[f^2(X)]=1} \mathbb{E}[(\mathbb{E}[f(X)|Y])^2] \tag{A21}$$

Denoting $\overline{h} = \mathbb{E}_{p(x)}[h(x)]$, we can transform $\beta_0[h(x)]$ in Equation (2) as follows:

$$\begin{aligned}
\beta_0[h(x)] &= \frac{\mathbb{E}_{x\sim p(x)}[h(x)^2] - \left(\mathbb{E}_{x\sim p(x)}[h(x)]\right)^2}{\mathbb{E}_{y\sim p(y)}\left[\left(\mathbb{E}_{x\sim p(x|y)}[h(x)]\right)^2\right] - \left(\mathbb{E}_{x\sim p(x)}[h(x)]\right)^2} \\
&= \frac{\mathbb{E}_{x\sim p(x)}[h(x)^2] - \overline{h}^2}{\mathbb{E}_{y\sim p(y)}\left[\left(\mathbb{E}_{x\sim p(x|y)}[h(x)]\right)^2\right] - \overline{h}^2} \\
&= \frac{\mathbb{E}_{x\sim p(x)}[(h(x) - \overline{h})^2]}{\mathbb{E}_{y\sim p(y)}\left[\left(\mathbb{E}_{x\sim p(x|y)}[h(x) - \overline{h}]\right)^2\right]} \\
&= \frac{1}{\mathbb{E}_{y\sim p(y)}\left[\left(\mathbb{E}_{x\sim p(x|y)}[f(x)]\right)^2\right]} \\
&= \frac{1}{\mathbb{E}[(\mathbb{E}[f(X)|Y])^2]}
\end{aligned}$$

where we denote $f(x) = \frac{h(x)-\overline{h}}{\left(\mathbb{E}_{x\sim p(x)}[(h(x)-\overline{h})^2]\right)^{1/2}}$, so that $\mathbb{E}[f(X)] = 0$ and $\mathbb{E}[f^2(X)] = 1$.

Combined with Equation (A21), we have

$$\sup_{h(x)} \frac{1}{\beta_0[h(x)]} = \rho_m^2(X;Y) \tag{A22}$$

Our Theorem 3 states that

$$\sup_{h(x)} \frac{1}{\beta_0[h(x)]} \leq \frac{1}{\beta_0} \tag{A23}$$

Combining Equations (A18), (A22) and Equation (A23), we have

$$\rho_m^2(X;Y) \leq \xi(X;Y) \tag{A24}$$

In summary, the relations among the quantities are:

$$\frac{1}{\beta_0} = \xi(X;Y) = \eta_{\mathrm{KL}}(p(y|x), p(x)) \geq \sup_{h(x)} \frac{1}{\beta_0[h(x)]} = \rho_m^2(X;Y) \tag{A25}$$

□

Appendix A.12. Experiment Details

We use the Variational Information Bottleneck (VIB) objective from [5]. For the synthetic experiment, the latent Z has dimension of 2. The encoder is a neural net with 2 hidden layers, each of which has 128 neurons with ReLU activation. The last layer has linear activation and 4 output neurons; the first two parameterize the mean of a Gaussian and the last two parameterize the log variance. The decoder is a neural net with 1 hidden layer with 128 neurons and ReLU activation. Its last layer has linear activation and outputs the logit for the class labels. It uses a mixture of Gaussian prior with 500 components (for the experiment with class overlap, 256 components), each of which is a 2D Gaussian with learnable mean and log variance, and the weights for the components are also learnable. For the MNIST experiment, the architecture is mostly the same, except the following: (1) for Z, we let it have dimension of 256. (2) For the prior, we use standard Gaussian with diagonal covariance matrix.

For all experiments, we use Adam [33] optimizer with default parameters. We do not add any explicit regularization. We use learning rate of 10^{-4} and have a learning rate decay of $\frac{1}{1+0.01\times \mathrm{epoch}}$. We train in total 2000 epochs with mini-batch size of 500.

For estimation of the observed β_0 in Figure 3, in the $I(X;Z)$ vs. β_i curve (β_i denotes the i-th β), we take the mean and standard deviation of $I(X;Z)$ for the lowest 5 β_i values, denoting as μ_β, σ_β ($I(Y;Z)$ has similar behavior, but since we are minimizing $I(X;Z) - \beta \cdot I(Y;Z)$, the onset of nonzero $I(X;Z)$ is less prone to noise). When $I(X;Z)$ is greater than $\mu_\beta + 3\sigma_\beta$, we regard it as learning a non-trivial representation, and take the average of β_i and β_{i-1} as the experimentally estimated onset of learning. We also inspect manually and confirm that it is consistent with human intuition.

For estimating β_0 using Algorithm 1, at step 6 we use the following discrete search algorithm. We fix $i_{\mathrm{left}} = 1$ and gradually narrow down the range $[a, b]$ of i_{right}, starting from $[1, N]$. At each iteration, we set a tentative new range $[a', b']$, where $a' = 0.8a + 0.2b$, $b' = 0.2a + 0.8b$, and calculate $\tilde{\beta}_{0,a'} = \mathbf{Get\beta}(P_{y|x}, p_y, \Omega_{a'})$, $\tilde{\beta}_{0,b'} = \mathbf{Get\beta}(P_{y|x}, p_y, \Omega_{b'})$ where $\Omega_{a'} = \{1, 2, ...a'\}$ and $\Omega_{b'} = \{1, 2, ...b'\}$. If $\tilde{\beta}_{0,a'} < \tilde{\beta}_{0,a}$, let $a \leftarrow a'$. If $\tilde{\beta}_{0,b'} < \tilde{\beta}_{0,b}$, let $b \leftarrow b'$. In other words, we narrow down the range of i_{right} if we find that the Ω given by the left or right boundary gives a lower $\tilde{\beta}_0$ value. The process stops when both $\tilde{\beta}_{0,a'}$ and $\tilde{\beta}_{0,b'}$ stop improving (which we find always happens when $b' = a' + 1$), and we return the smaller of the final $\tilde{\beta}_{0,a'}$ and $\tilde{\beta}_{0,b'}$ as $\tilde{\beta}_0$.

For estimation of $p(y|x)$ for (2′) Algorithm 1 and (3′) $\hat{\eta}_{\mathrm{KL}}$ for both synthetic and MNIST experiments, we use a 3-layer neuron net where each hidden layer has 128 neurons and ReLU activation. The last layer has linear activation. The objective is cross-entropy loss. We use Adam [33] optimizer with a learning rate of 10^{-4}, and train for 100 epochs (after which the validation loss does not go down).

For estimating β_0 via (3′) $\hat{\eta}_{\mathrm{KL}}$ by the algorithm in [18], we use the code from the GitHub repository provided by the paper (At https://github.com/wgao9/hypercontractivity), using the same $p(y|x)$ employed for (2′) Algorithm 1. Since our datasets are classification tasks, we use $A_{ij} = p(y_j|x_i)/p(y_j)$ instead of the kernel density for estimating matrix A; we take the maximum of 10 runs as estimation of μ.

CIFAR10 Details

We trained a deterministic 28 × 10 wide resnet [34,35], using the open source implementation from Cubuk et al. [36]. However, we extended the final 10 dimensional logits of that model through another 3 layer MLP classifier, in order to keep the inference network architecture identical between this model and the VIB models we describe below. During training, we dynamically added label noise according to the class confusion matrix in Table A1. The mean label noise averaged across the 10 classes is 20%. After that model had converged, we used it to estimate β_0 with Algorithm 1. Even with 20% label noise, β_0 was estimated to be 1.0483.

Table A1. Class confusion matrix used in CIFAR10 experiments. The value in row i, column j means for class i, the probability of labeling it as class j. The mean confusion across the classes is 20%.

	Plane	**Auto.**	**Bird**	**Cat**	**Deer**	**Dog**	**Frog**	**Horse**	**Ship**	**Truck**
Plane	0.82232	0.00238	0.021	0.00069	0.00108	0	0.00017	0.00019	0.1473	0.00489
Auto.	0.00233	0.83419	0.00009	0.00011	0	0.00001	0.00002	0	0.00946	0.15379
Bird	0.03139	0.00026	0.76082	0.0095	0.07764	0.01389	0.1031	0.00309	0.00031	0
Cat	0.00096	0.0001	0.00273	0.69325	0.00557	0.28067	0.01471	0.00191	0.00002	0.0001
Deer	0.00199	0	0.03866	0.00542	0.83435	0.01273	0.02567	0.08066	0.00052	0.00001
Dog	0	0.00004	0.00391	0.2498	0.00531	0.73191	0.00477	0.00423	0.00001	0
Frog	0.00067	0.00008	0.06303	0.05025	0.0337	0.00842	0.8433	0	0.00054	0
Horse	0.00157	0.00006	0.00649	0.00295	0.13058	0.02287	0	0.83328	0.00023	0.00196
Ship	0.1288	0.01668	0.00029	0.00002	0.00164	0.00006	0.00027	0.00017	0.83385	0.01822
Truck	0.01007	0.15107	0	0.00015	0.00001	0.00001	0	0.00048	0.02549	0.81273

We then trained 73 different VIB models using the same 28 × 10 wide resnet architecture for the encoder, parameterizing the mean of a 10-dimensional unit variance Gaussian. Samples from the encoder distribution were fed to the same 3 layer MLP classifier architecture used in the deterministic model. The marginal distributions were mixtures of 500 fully covariate 10-dimensional Gaussians, all parameters of which are trained. The VIB models had β ranging from 1.02 to 2.0 by steps of 0.02, plus an extra set ranging from 1.04 to 1.06 by steps of 0.001 to ensure we captured the empirical β_0 with high precision.

However, this particular VIB architecture does not start learning until $\beta > 2.5$, so none of these models would train as described. (A given architecture trained using maximum likelihood and with no stochastic layers will tend to have higher effective capacity than the same architecture with a stochastic layer that has a fixed but non-trivial variance, even though those two architectures have exactly the same number of learnable parameters.) Instead, we started them all at $\beta = 100$, and annealed β down to the corresponding target over 10,000 training gradient steps. The models continued to train for another 200,000 gradient steps after that. In all cases, the models converged to essentially their final accuracy within 20,000 additional gradient steps after annealing was completed. They were stable over the remaining ∼180,000 gradient steps.

References

1. Tishby, N.; Pereira, F.C.; Bialek, W. The information bottleneck method. *arXiv* **2000**, arXiv:physics/0004057.
2. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
3. Chechik, G.; Globerson, A.; Tishby, N.; Weiss, Y. Information bottleneck for Gaussian variables. *J. Mach. Learn. Res.* **2005**, *6*, 165–188.
4. Rey, M.; Roth, V. Meta-Gaussian information bottleneck. In *Advances in Neural Information Processing Systems*; INIPS: San Diego, CA, USA, 2012; pp. 1916–1924.
5. Alemi, A.A.; Fischer, I.; Dillon, J.V.; Murphy, K. Deep variational information bottleneck. *arXiv* **2016**, arXiv:1612.00410.

6. Chalk, M.; Marre, O.; Tkacik, G. Relevant sparse codes with variational information bottleneck. In *Advances in Neural Information Processing Systems*; NIPS: San Diego, CA, USA, 2016; pp. 1957–1965.
7. Fischer, I. The Conditional Entropy Bottleneck. 2018. Available online: https://openreview.net/forum?id=rkVOXhAqY7 (accessed on 20 September 2019).
8. Strouse, D.; Schwab, D.J. The deterministic information bottleneck. *Neural Comput.* **2017**, *29*, 1611–1630. [CrossRef] [PubMed]
9. Kolchinsky, A.; Tracey, B.D.; Van Kuyk, S. Caveats for information bottleneck in deterministic scenarios. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 30 April 2019.
10. Strouse, D.; Schwab, D.J. The information bottleneck and geometric clustering. *arXiv* **2017**, arXiv:1712.09657.
11. Achille, A.; Soatto, S. Emergence of invariance and disentanglement in deep representations. *J. Mach. Learn. Res.* **2018**, *19*, 1947–1980.
12. Achille, A.; Soatto, S. Information dropout: Learning optimal representations through noisy computation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**. [CrossRef]
13. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
14. Krizhevsky, A.; Hinton, G. *Learning Multiple Layers of Features from Tiny Images*; Technical Report; University of Toronto: Toronto, ON, Canada, 2009.
15. Achille, A.; Mbeng, G.; Soatto, S. The Dynamics of Differential Learning I: Information-Dynamics and Task Reachability. *arXiv* **2018**, arXiv:1810.02440.
16. Anantharam, V.; Gohari, A.; Kamath, S.; Nair, C. On maximal correlation, hypercontractivity, and the data processing inequality studied by Erkip and Cover. *arXiv* **2013**, arXiv:1304.6133.
17. Polyanskiy, Y.; Wu, Y. Strong data-processing inequalities for channels and Bayesian networks. In *Convexity and Concentration*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 211–249.
18. Kim, H.; Gao, W.; Kannan, S.; Oh, S.; Viswanath, P. Discovering potential correlations via hypercontractivity. In *Advances in Neural Information Processing Systems*; NIPS: San Diego, CA, USA, 2017; pp. 4577–4587.
19. Lin, H.W.; Tegmark, M. Criticality in formal languages and statistical physics. *arXiv* **2016**, arXiv:1606.06737.
20. Hirschfeld, H.O. A connection between correlation and contingency. In *Mathematical Proceedings of the Cambridge Philosophical Society*; Cambridge University Press: Cambridge, UK, 1935; Volume 31, pp. 520–524.
21. Gebelein, H. Das statistische Problem der Korrelation als Variations-und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung. *ZAMM-J. Appl. Math. Mech. Für Angew. Math. Und Mech.* **1941**, *21*, 364–379. [CrossRef]
22. Angluin, D.; Laird, P. Learning from noisy examples. *Mach. Learn.* **1988**, *2*, 343–370. [CrossRef]
23. Natarajan, N.; Dhillon, I.S.; Ravikumar, P.K.; Tewari, A. Learning with noisy labels. In *Advances in Neural Information Processing Systems*; NIPS: San Diego, CA, USA, 2013; pp. 1196–1204.
24. Liu, T.; Tao, D. Classification with noisy labels by importance reweighting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 447–461. [CrossRef] [PubMed]
25. Xiao, T.; Xia, T.; Yang, Y.; Huang, C.; Wang, X. Learning from massive noisy labeled data for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2691–2699.
26. Northcutt, C.G.; Wu, T.; Chuang, I.L. Learning with confident examples: Rank pruning for robust classification with noisy labels. *arXiv* **2017**, arXiv:1705.01936.
27. van den Oord, A.; Kalchbrenner, N.; Espeholt, L.; Kavukcuoglu, K.; Vinyals, O.; Graves, A. Conditional Image Generation with PixelCNN Decoders. In *Advances in Neural Information Processing Systems 29*; Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2016; pp. 4790–4798.
28. Salimans, T.; Karpathy, A.; Chen, X.; Kingma, D.P. PixelCNN++: A PixelCNN Implementation with Discretized Logistic Mixture Likelihood and Other Modifications. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
29. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E* **2004**, *69*, 066138. [CrossRef]
30. Gelfand, I.M.; Silverman, R.A. *Calculus of Variations*; Courier Corporation: North Chelmsford, MA, USA, 2000.

31. Erkip, E.; Cover, T.M. The efficiency of investment information. *IEEE Trans. Inf. Theory* **1998**, *44*, 1026–1040. [CrossRef]
32. Rényi, A. On measures of dependence. *Acta Math. Hung.* **1959**, *10*, 441–451. [CrossRef]
33. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
34. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
35. Zagoruyko, S.; Komodakis, N. Wide Residual Networks. *arXiv* **2016**, arXiv: 1605.07146.
36. Cubuk, E.D.; Zoph, B.; Mane, D.; Vasudevan, V.; Le, Q.V. Autoaugment: Learning augmentation policies from data. *arXiv* **2018**, arXiv:1805.09501.