# **Learnability for the Information Bottleneck**

Tailin Wu **MIT** tailin@mit.edu

Ian Fischer Google Research iansf@google.com

MIT ichuang@mit.edu

Isaac L. Chuang

Max Tegmark MIT

tegmark@mit.edu

## **Abstract**

The Information Bottleneck (IB) method (Tishby et al. (2000)) 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  $\beta$  to tune this trade-off. However, in practice, not only is  $\beta$  chosen empirically without theoretical guidance, there is also a lack of theoretical understanding between  $\beta$ , learnability, the intrinsic nature of the dataset and model capacity. In this paper, we show that if  $\beta$  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  $\beta$  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  $\beta$ . 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  $\beta$  for a given dataset. We also empirically demonstrate our theoretical conditions with analyses of synthetic datasets, MNIST, and CIFAR10.

#### INTRODUCTION

Tishby et al. (2000) introduced the *Information Bottle*neck (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:

$$\min \mathbf{IB}_{\beta}(X, Y; Z) = \min I(X; Z) - \beta I(Y; Z) \quad (1)$$

 $I(\cdot)$  is the mutual information. The hyperparameter  $\beta$ controls the trade-off between compression and prediction, in the same spirit as Rate-Distortion Theory (Shannon, 1948), but with a learned representation function  $\overline{P(Z|X)}$  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 (Chechik et al. (2005)), meta-Gaussians (Rey and Roth (2012)), continuous variables via variational methods (Alemi et al. (2016); Chalk et al. (2016); Fischer (2018)), deterministic scenarios (Strouse and Schwab (2017a); Kolchinsky et al. (2019)), geometric clustering (Strouse and Schwab (2017b)), and is used for learning invariant and disentangled representations in deep neural nets (Achille and Soatto (2018a,b)). However, a core issue remains: how should we set a good  $\beta$ ? In the original work, the authors recommend sweeping  $\beta > 1$ , which can be prohibitively expensive in practice, but also leaves open interesting theoretical questions around the relationship between  $\beta$ , P(Z|X), and the observed data, P(X,Y).

This work begins to answer some of those questions by characterizing the *onset* of learning. Specifically:

- We show that improperly chosen  $\beta$  may result in a failure to learn: the trivial solution P(Z|X) =P(Z) becomes the global minimum of the IB objective, even for  $\beta \gg 1$ .
- We introduce the concept of IB-Learnability, and show that when we vary  $\beta$ , the IB objective will un-

dergo a phase transition from the inability to learn to the ability to learn.

- Using the second-order variation, we derive sufficient conditions for IB-Learnability, which provide theoretical guidance for choosing a good β.
- 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.
- We additionally prove a deep relationship between IB-Learnability, the hypercontractivity coefficient, the contraction coefficient, and the maximum correlation.

We use our main results to demonstrate on synthetic datasets, MNIST (LeCun et al.) [1998) and CIFAR10 (Krizhevsky and Hinton, 2009) that the theoretical prediction for IB-Learnability closely matches experiment. We also present an algorithm for estimating the onset of IB-Learnability and the conspicuous subset, and demonstrate that it does a good job of approximating both the theoretical predictions and the empirical results.

# 2 RELATED WORK

The original IB work (Tishby et al., 2000) provides a tabular method for exactly computing the optimal encoder distribution P(Z|X) for a given  $\beta$  and cardinality of the discrete representation, |Z|. Thus, the search for the desired model involves not only sweeping  $\beta$ , but also considering different representation dimensionalities. These restrictions were lifted somewhat by Chechik et al. (2005), which 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. They also note the presence of the trivial solution not only when  $\beta \leq 1$ , but also depending on the eigenspectrum of the observed variables. However, the restriction to multivariate Gaussian datasets limits the generality of the analysis. Another analytic treatment of IB is given in Rey and Roth (2012), 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 (2017a) presents 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  $\beta$ . 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. (2016). 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) (Fischer) 2018). 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. (2019) presents analytic and empirical results about trivial solutions in the particular setting of Y being a deterministic function of X in the observed sample. However, their use of the term "trivial solution" is distinct from ours. They are referring to the observation that  $\beta$  will demonstrate trivial interpolation 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  $\beta$ .

Achille and Soatto (2018b) applies the IB Lagrangian to the weights of a neural network, yielding InfoDropout. In Achille and Soatto (2018a), 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. (2018) 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 (Anantharam et al.) (2013); Polyanskiy and Wu (2017)), defined as  $\sup_{Z-X-Y} \frac{I(Y;Z)}{I(X;Z)}$ , which by definition equals the inverse of  $\beta_0$ , our IB-learnability threshold. In Anantharam et al. (2013), the authors prove that the hypercontractivity cofficient equals the contraction coefficient  $\eta_{KL}(P_{Y|X}, P_X)$ , and Kim et al. (2017) propose a practical algorithm to estimate  $\eta_{KL}(P_{Y|X}, P_X)$ , 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  $(x,y) \in \mathcal{X} \times \mathcal{Y}$  drawn from a distribution with probability (density) P(X,Y), where unless otherwise stated, both X and Y can be discrete or continuous variables. (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  $\beta$  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  $Z \leftarrow X \leftrightarrow Y$ . Eq. 1 above gives the IB objective with Lagrange multiplier  $\beta$ ,  $\mathrm{IB}_{\beta}(X,Y;Z)$ , 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  $\beta$ , the more the objective favors making a good prediction for Y. Conversely, the smaller  $\beta$ , the more the objective favors learning a concise representation.

How can we select  $\beta$  such that the IB objective learns a useful representation? In practice, the selection of  $\beta$  is done empirically. Indeed, Tishby et al. (2000) recommends "sweeping  $\beta$ ". In this paper, we provide theoretical guidance for choosing  $\beta$  by introducing the concept of IB-Learnability and providing a series of IB-learnable conditions.

**Definition 1.** (X,Y) is  $IB_{\beta}$ -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  $\mathrm{IB}_{\beta}$ -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  $\mathrm{IB}_{\beta}$ -learnable, then when  $\mathrm{IB}_{\beta}(X,Y;Z)$  is globally minimized, it may learn a trivial representation.

**Trivial solutions.** Definition  $\blacksquare$  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.

Necessary condition for IB-Learnability. From Definition II we can see that  $\mathrm{IB}_{\beta}$ -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  $\mathrm{IB}_{\beta}$ -learnable for  $\beta\leq 1$ .

Due to the reparameterization invariance of mutual information, we have the following theorem for  $IB_{\beta}$ -Learnability:

**Theorem 1.** Let X' = g(X) be an uniquely invertible map (if X is a continuous variable, g is additionally required to be continuous). Then (X,Y) and (X',Y) have the same  $IB_{\beta}$ -Learnability.

The proof for Theorem [I] is in Appendix [B]. Theorem [I] 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  $\mathrm{IB}_{\beta}$ -learnable? To answer this question, we derive a series of sufficient conditions for  $\mathrm{IB}_{\beta}$ -Learnability, starting from its definition. The conditions are in increasing order of

<sup>&</sup>lt;sup>1</sup>We use capital letters X,Y,Z for 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.

practicality, while sacrificing as little generality as possible.

Firstly, Theorem 2 characterizes the  $IB_{\beta}$ -Learnability range for  $\beta$ , with proof in Appendix C

**Theorem 2.** If (X, Y) is  $IB_{\beta_1}$ -learnable, then for any  $\beta_2 > \beta_1$ , it is  $IB_{\beta_2}$ -learnable.

Based on Theorem 2 the range of  $\beta$  such that (X,Y) is  $\mathrm{IB}_{\beta}$ -learnable has the form  $\beta \in (\beta_0,+\infty)$ . Thus,  $\beta_0$  is the *threshold* of IB-Learnability.

**Lemma 2.1.** p(z|x) = p(z) is a stationary solution for  $IB_{\beta}(X, Y; Z)$ .

The proof in Appendix F 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  $\delta \mathrm{IB}_{\beta}[p(z|x)]=0$  at the trivial representation.

Lemma [2.1] 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  $\mathrm{IB}_{\beta}[p(z|x)]$ . Based on the necessary condition for the minimum (Appendix  $\boxed{\mathsf{D}}$ ), we have the following theorem

**Theorem 3 (Suff. Cond. 1).** A sufficient condition for (X,Y) to be  $IB_{\beta}$ -learnable is that there exists a perturbation function h(z|x) with  $\int_{-\infty}^{3} h(z|x)dz = 0$ , 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 3 is given in Appendix 1 Intuitively, if  $\delta^2 \mathrm{IB}_\beta[p(z|x)]\Big|_{p(z|x)=p(z)}<0$ , we can always find a p'(z|x)=p(z|x)+h(z|x) in the neighborhood of the trivial representation p(z|x)=p(z), such that  $\mathrm{IB}_\beta[p'(z|x)]<\mathrm{IB}_\beta[p(z|x)]$ , thus satisfying the definition for  $\mathrm{IB}_\beta$ -Learnability.

To make Theorem 3 more practical, we perturb p(z|x) around the trivial solution  $p'(z|x) = p(z|x) + \epsilon h(z|x)$ , and expand  $\mathrm{IB}_\beta[p(z|x) + h(z|x)] - \mathrm{IB}_\beta[p(z|x)]$  to the second order of  $\epsilon$ . We can then prove Theorem 4:

**Theorem 4 (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 G, which also gives a construction for h(z|x) for Theorem G for any h(x) satisfying Theorem G and shows that the converse is also true: if there exists h(z|x) suth that the condition in Theorem G is true, then we can find h(x) satisfying the the condition in Theorem G

Theorem 4 suggests a method to estimate  $\beta_0$ : we can parameterize h(x) 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  $\beta_0$ , and h(x) provides a soft clustering of the examples.

Alternatively, based on the property of  $\beta_0[h(x)]$ , we can also use a specific functional form for h(x) in Eq. (2), and obtain a stronger sufficient condition for  $IB_{\beta}$ -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 Eq. (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:

$$\beta > \inf_{h(x),\Omega_x \in \mathcal{X}} \frac{\frac{\mathbb{E}_{x \sim p(x),x \in \Omega_x}[h(x)^2]}{\left(\mathbb{E}_{x \sim p(x),x \in \Omega_x}[h(x)]\right)^2} - 1}{\int \frac{dy}{p(y)} \left(\frac{\mathbb{E}_{x \sim p(x),x \in \Omega_x}[p(y|x)h(x)]}{\mathbb{E}_{x \sim p(x),x \in \Omega_x}[h(x)]}\right)^2 - 1}$$
(3)

• The numerator of the R.H.S. of Eq. (3) attains its minimum when h(x) is a constant within  $\Omega_x$ . 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 Eq. (3)  $\geq \frac{1}{\int_{x\in\Omega_x}p(x)}-1$ , 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  $\mathrm{IB}_{\beta}$ -Learnability, which is a key result of this paper:

 $<sup>^2</sup>$ 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 from the lens of the model.

<sup>&</sup>lt;sup>3</sup>Whenever a variable W is discrete, we can simply replace the integral  $(\int \cdot dw)$  by summation  $(\sum_{w} \cdot)$ .

**Theorem 5 (Conspicuous Subset Suff. Cond.).** A sufficient condition for (X, Y) to be  $IB_{\beta}$ -learnable is X and Y are not independent, and

$$\beta > \inf_{\Omega_x \subset \mathcal{X}} \beta_0(\Omega_x) \tag{4}$$

where

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

 $\Omega_x$  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 H. In the proof we also show that this condition is invariant to invertible mappings of X.

### 4.1 Discussion

The conspicuous subset determines  $\beta_0$ . From Eq. (4), we see that three characteristics of the subset  $\Omega_x \subset \mathcal{X}$  lead to low  $\beta_0$ : (1) confidence:  $p(y|\Omega_x)$  is large; (2) typicality and size: the number of elements in  $\Omega_x$  is large, or the elements in  $\Omega_x$  are typical, leading to a large probability of  $p(\Omega_x)$ ; (3) imbalance: p(y) is small for the subset  $\Omega_x$ , but large for its complement. In summary,  $\beta_0$  will be determined by the largest *confident*, *typical* and *imbalanced subset* of examples, or an equilibrium of those characteristics. We term  $\Omega_x$  at the minimization of  $\beta_0(\Omega_x)$  the *conspicuous subset*.

Multiple phase transitions. Based on this characterization of  $\Omega_x$ , we can hypothesize datasets with multiple learnability phase transitions. Specifically, consider a region  $\Omega_{x0}$  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  $\Omega_{x0}$ will dominate the infimum in Eq. (4), resulting in a small value of  $\beta_0$ . However, the remaining  $\mathcal{X} - \Omega_{x0}$  effectively form a new dataset,  $\mathcal{X}_1$ . At exactly  $\beta_0$ , we may have that the current encoder,  $p_0(z|x)$ , has no mutual information with the remaining classes in  $\mathcal{X}_1$ ; i.e.,  $I(Y_1; Z_0) = 0$ . In this case, Definition  $\boxed{1}$  applies to  $p_0(z|x)$  with respect to  $I(X_1; Z_1)$ . We might expect to see that, at  $\beta_0$ , 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  $\mathcal{X}_{i-1}$ . The end of Appendix G gives a more detailed analysis on multiple phase transitions.

Similarity to information measures. The denominator of Eq. (4) is closely related to mutual information. Using the inequality  $x-1 \ge \log(x)$  for x > 0, it becomes:

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

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 Eq. (4) 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" (Lin and Tegmark (2016)) at  $\Omega_x$ .

Similarly, the numerator is related to the self-information of  $\Omega_x$ :

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

so we can estimate the phase transition as:

$$\beta \gtrsim \inf_{\Omega_x \subset \mathcal{X}} \frac{h(\Omega_x)}{\tilde{I}(\Omega_x; Y)} \tag{5}$$

Since Eq. (5) uses upper bounds on both the numerator and the denominator, it does not give us a bound on  $\beta_0$ .

**Estimating model capacity.** The observation that a model can't 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.

Learnability and the Information Plane. Many of our results can be interpreted in terms of the geometry of the Pareto frontier illustrated in Fig. 1, which describes the trade-off between increasing I(Y; Z) and decreasing I(X; Z). At any point on this frontier that minimizes  $IB_{\beta}^{min} \equiv \min I(X; Z) - \beta I(Y; Z)$ , the frontier will have slope  $\beta^{-1}$  if it is differentiable. If the frontier is also concave (has negative second derivative), then this slope  $\beta^{-1}$  will take its maximum  $\beta_0^{-1}$  at the origin, which implies  $IB_{\beta}$ -Learnability for  $\beta > \beta_0$ , so that the threshold for IB<sub>\beta</sub>-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<sub>β</sub>learnability is the inverse of its maximum slope. Indeed, Theorem 4 and Theorem 5 give lower bounds of the slope of the Pareto frontier at the origin.

**IB-Learnability, hypercontractivity, and maximum correlation.** In Appendix [J] we prove the following re-

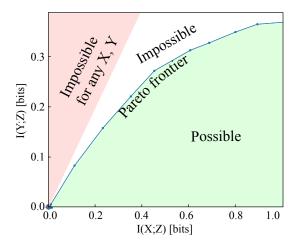


Figure 1: The Pareto frontier of mutual information that Z can have with X and Y is shown for a binary classification of MNIST digits 0 and 1 with 20% label noise. On this problem, no learning happens for models trained at  $\beta < 3.25$ . H(Y) = 1 bit since only the two of ten digits are used, and  $I(Y;Z) \leq I(X;Y) \approx 0.5$  bits < H(Y) because of the 20% label swapping. 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.

lationships:

$$\frac{1}{\beta_0} = \xi(X;Y) = \eta_{KL} \ge \sup_{h(x)} \frac{1}{\beta_0[h(x)]} = \rho_m^2(X;Y)$$
(6)

where  $\rho_m(X;Y) \equiv \max_{f,g} \mathbb{E}[f(X)g(Y)]$  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 (Hirschfeld, 1935). Gebelein, 1941),  $\xi(X;Y) \equiv \sup_{Z-X-Y} \frac{f(Y;Z)}{f(X;Z)}$  is the hypercontractivity coefficient, and  $\eta_{\text{KL}}(p(y|x), p(x)) \equiv \sup_{r(x) \neq p(x)} \frac{\mathbb{D}_{\text{KL}}(r(x)||p(x))}{\mathbb{D}_{\text{KL}}(r(x)||p(x))}$  is the contraction coefficient. Our proof relies on Anantharam et al. (2013)'s proof  $\xi(X;Y) = \eta_{\text{KL}}$ . 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).

# 5 ESTIMATING THE IB-LEARNABILITY CONDITION

Theorem 5 not only reveals the relationship between the learnability threshold for  $\beta$  and the least noisy region of P(Y|X), but also provides a way to practically estimate

 $\beta_0$ , both in the general classification case, and in more structured settings.

### 5.1 Estimation Algorithm

Based on Theorem 5 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  $\beta_0$ .

We approximate the probability of each example  $p(x_i)$  by its empirical probability,  $\hat{p}(x_i)$ . E.g., 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 e.g. 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  $\Omega$  such that the  $\tilde{\beta}_0$  is as small as possible, by previous analysis we want to find a *confident* subset such that its p(y|x) is large for a certain class j (to make the denominator of Eq. (1) 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  $\Omega$  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  $\varepsilon$ . 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  $\beta_0$  can be very large.

#### 5.2 Special Cases for Estimating $\beta_0$

Theorem 5 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.

**Class-conditional label noise.** Classification with noisy labels is a common practical scenario. An important noise model is that the labels are randomly flipped

# Algorithm 1 Estimating the upper bound for $\beta_0$ and identifying the conspicuous subset

**Require**: Dataset  $\mathcal{D} = \{(x_i, y_i)\}, i = 1, 2, ...N$ . The number of classes is C.

**Require**  $\varepsilon$ : tolerance for estimating  $\beta_0$ 

- 1: Learn a maximum likelihood model  $p_{\theta}(y|x)$  using the dataset  $\mathcal{D}$ .
- 2: Construct matrix  $(P_{y|x})$  such that
- $(P_{y|x})_{ij}=p_{\theta}(y=j|x=x_i).$ 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:  $P_{y|x}^{(\text{sort}j)} \leftarrow \text{Sort the rows of } P_{y|x} \text{ in decreasing}$ values of  $(P_{y|x})_{ij}$ .
- $$\begin{split} \tilde{\beta}_0^{(j)}, \Omega^{(j)} \leftarrow & \text{Search } i_{\text{left}}, i_{\text{right}} \text{ until } \tilde{\beta}_0^{(j)} = \\ & \textbf{Get} \boldsymbol{\beta}(P_{y|x}, p_y, \Omega) \text{ is minimal with tolerance } \varepsilon, \end{split}$$
  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**  $\tilde{\beta}_0, P_{u|x}^{(\tilde{\beta}_0)}$

with some hidden class-conditional probabilities and we only observe the corrupted labels. This problem has been studied extensively (Angluin and Laird, 1988; Natarajan et al., 2013; Liu and Tao, 2016; Xiao et al., 2015; Northcutt et al., 2017). If IB is applied to this scenario, how large  $\beta$  do we need? The following corollary provides a simple formula.

**Corollary 5.1.** 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 classconditional 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:

$$\beta > \inf_{y^*} \frac{\frac{1}{p(y^*)} - 1}{\sum_{y} \frac{p(y|y^*)^2}{p(y)} - 1} \tag{7}$$

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 e.g. Northcutt et al. (2017). Additionally, if we know that the noise is class-conditional, but the observed  $\beta_0$  is greater than the R.H.S. of Eq. (7), we can deduce that there is overlap between the true classes. The proof of Corollary 5.1 is provided in Appendix I

# Algorithm 2 Get 3

**Require**  $P_{y|x}$ : matrix for p(y|x)

**Require**  $p_y$ : vector for p(y)

**Require**  $\Omega$ : set of indices for the rows of  $P_{y|x}$ 

1:  $N \leftarrow$  number of rows of  $P_{y|x}$ .

2:  $C \leftarrow$  number of columns of  $P_{y|x}$ .

3:  $n \leftarrow$  number of elements of  $\Omega$ .

5: 
$$\tilde{\beta}_0 \leftarrow \frac{\frac{N}{n} - 1}{\sum_{j} \left[ \frac{(p_{y|\Omega_x})_j^2}{p_{y,j}} - 1 \right]}$$

6: return  $\beta_0$ 

**Deterministic relationships.** Theorem 5 also reveals that  $\beta_0$  relates closely to whether Y is a deterministic function of X, as shown by Corollary 5.2:

**Corollary 5.2.** 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 5.2 is satisfied by classification, and certain regression problems. Combined with the necessary condition  $\beta > 1$  for any dataset (X,Y) to be  $\mathrm{IB}_{\beta}$ -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  $\beta > 1$ ; i.e., its  $\beta_0$  is 1. The proof of Corollary 5.2 is provided in Appendix I

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  $\beta_0$ , even when learning deterministic functions.

# **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  $\beta_0$  found experimentally. We also compare with the algorithm in Kim et al. (2017) for estimating the hypercontractivity coefficient  $(=1/\beta_0)$  via the contraction coefficient  $\eta_{KL}$ . Experiment details are in Section K.

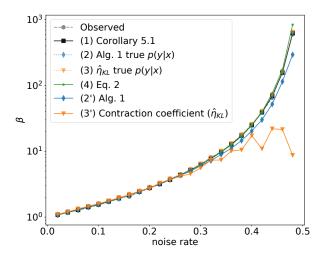


Figure 2: Predicted vs. experimentally identified  $\beta_0$ , for mixture of Gaussians with varying class-conditional noise rates.

## **6.1** 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  $\beta$  with exponential steps, and observe I(X;Z) and I(Y;Z). We then compare the empirical  $\beta_0$  indicated by the onset of above-zero I(X;Z) with predicted values for  $\beta_0$ .

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  $\Sigma = \text{diag}(0.25, 0.25)$ . 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  $\beta$ , and observe the onset of learning via random I(X; Z) (Observed). To test how different methods perform in estimating  $\beta_0$ , we apply the following methods: (1) Corollary 5.1, since this is classification with class-conditional noise, and the two true classes have virtually no overlap; (2) Alg. 1 with true p(y|x); (3) The algorithm in Kim et al. (2017) that estimates  $\hat{\eta}_{KL}$ , provided with true p(y|x); (4)  $\beta_0[h(x)]$  in Eq. [2]; (2') Alg. [1] with p(y|x) estimated by a neural net;

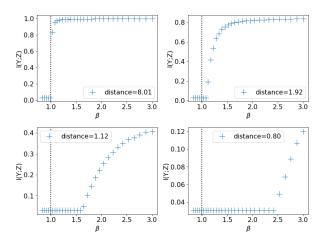


Figure 3: I(Y;Z) vs.  $\beta$ , for mixture of Gaussian datasets with different distances between the two mixture components. The vertical lines are  $\beta_{0,\text{predicted}}$  by the R.H.S. of Eq. (7). As Eq. (7) does not make predictions w.r.t. class overlap, the vertical lines are always just above  $\beta_{0,\text{predicted}} = 1$ . However, as expected, decreasing the distance between the classes in X space also increases the true  $\beta_0$ .

(3')  $\hat{\eta}_{KL}$  with the same p(y|x) as in (2'). The results are shown in Fig. 2 and in Appendix K.1.

From Fig. 2 we see the following. (A) When using the true p(y|x), both Alg. 1 and  $\hat{\eta}_{KL}$  generally upper bound the empirical  $\beta_0$ , and Alg. 1 is generally tighter. (B) When using the true p(y|x), Alg. 1 and Corollary 5.1 give the same result. (C) Comparing Alg. 1 and  $\hat{\eta}_{KL}$  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, Alg. I gives more precise values than  $\hat{\eta}_{KL}$ , indicating that Alg. 1 is more robust to the estimation error of p(y|x). (**D**) Eq. 2 empirically upper bounds the experimentally observed  $\beta_0$ , and gives almost the same result as theoretical estimation in Corollary 5.1 and Alg. 1 with the true p(y|x). In the classification setting, this approach doesn't 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,  $\beta_0$  can be very high. Instead of blindly sweeping  $\beta$ , our result can provide guidance for setting  $\beta$  so learning can happen.

Classification with class overlap. In this experiment, we test how different amounts of overlap among classes influence  $\beta_0$ . We use the mixture of Gaussians with two components, each of which is a 2D Gaussian with diagonal covariance matrix  $\Sigma = \text{diag}(0.25, 0.25)$ . The two

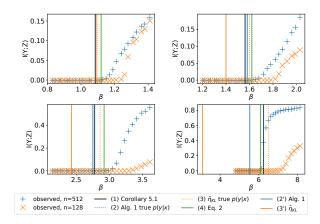


Figure 4: I(Y;Z) vs.  $\beta$  for the MNIST binary classification with different hidden units per layer n and noise rates  $\rho$ : (upper left)  $\rho=0.02$ , (upper right)  $\rho=0.1$ , (lower left)  $\rho=0.2$ , (lower right)  $\rho=0.3$ . The vertical lines are  $\beta_0$  estimated by different methods. n=128 has insufficient capacity for the problem, so its observed learnability onset is pushed higher, similar to the class overlap case.

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 don't 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 5.2. The more overlap between the two classes, the more uncertain Y is given X. By Eq. 4 we expect  $\beta_0$  to be larger, which is corroborated in Fig. 3

# **6.2** 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  $\rho$ . 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  $\beta_0$  for the n=128 model to be larger. This result is confirmed by the experiment (Fig. 4). Also, in Fig. 4 we plot the estimated  $\beta_0$  by different methods. We see that the observations (A), (B), (C) and (D) in Section 6.1 still hold.

## 6.3 CIFAR10 Forgetting Experiments

For CIFAR10 (Krizhevsky and Hinton, 2009), we study how *forgetting* varies with  $\beta$ . In other words, given a

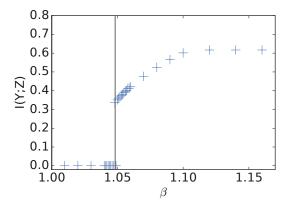


Figure 5: CIFAR10 plot of I(Y;Z) vs.  $\beta$  on the training set with 20% label noise. Each blue cross corresponds to a fully-converged model starting with independent initialization. The vertical black line corresponds to the predicted  $\beta_0=1.0483$  using Alg. []. The empirical  $\beta_0=1.048$ .

VIB model trained at some high  $\beta_2$ , if we anneal it down to some much lower  $\beta_1$ , what accuracy does the model converge to? We estimated  $\beta_0=1.0483$  on a version of CIFAR10 with 20% label noise using Alg. II. The lowest  $\beta$  with performance above chance was  $\beta=1.048$ , a very tight match with the estimate from Alg. II. See Appendix K.3 for details.

### 7 CONCLUSION

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 believe these results will provide theoretical and practical guidance for choosing  $\beta$  in the IB framework for balancing prediction and compression. 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.

## Acknowledgements

Tailin Wu'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.

## References

- Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1): 1947–1980, 2018a.
- Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018b.
- Alessandro Achille, Glen Mbeng, and Stefano Soatto. The Dynamics of Differential Learning I: Information-Dynamics and Task Reachability. *arXiv preprint arXiv:1810.02440*, 2018.
- Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.
- Venkat Anantharam, Amin Gohari, Sudeep Kamath, and Chandra Nair. On maximal correlation, hypercontractivity, and the data processing inequality studied by erkip and cover. *arXiv preprint arXiv:1304.6133*, 2013.
- Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.
- Matthew Chalk, Olivier Marre, and Gasper Tkacik. Relevant sparse codes with variational information bottleneck. In *Advances in Neural Information Processing Systems*, pages 1957–1965, 2016.
- Gal Chechik, Amir Globerson, Naftali Tishby, and Yair Weiss. Information bottleneck for gaussian variables. *Journal of machine learning research*, 6(Jan):165–188, 2005.
- Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.
- Elza Erkip and Thomas M Cover. The efficiency of investment information. *IEEE Transactions on Information Theory*, 44(3):1026–1040, 1998.
- Ian Fischer. The conditional entropy bottleneck,
  2018. URL | openreview.net/forum?id=
  | rkVOXhAqY7|.
- Hans Gebelein. Das statistische problem der korrelation als variations-und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik, 21(6):364–379, 1941.
- Izrail Moiseevitch Gelfand, Richard A Silverman, et al. *Calculus of variations*. Courier Corporation, 2000.

- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- Hermann O Hirschfeld. A connection between correlation and contingency. In *Mathematical Proceedings* of the Cambridge Philosophical Society, volume 31, pages 520–524. Cambridge University Press, 1935.
- Hyeji Kim, Weihao Gao, Sreeram Kannan, Sewoong Oh, and Pramod Viswanath. Discovering potential correlations via hypercontractivity. In *Advances in Neural Information Processing Systems*, pages 4577–4587, 2017.
- Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Artemy Kolchinsky, Brendan D Tracey, and Steven Van Kuyk. Caveats for information bottleneck in deterministic scenarios. *ICLR*, 2019.
- Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.
- Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.
- Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- Henry W Lin and Max Tegmark. Criticality in formal languages and statistical physics. *arXiv* preprint *arXiv*:1606.06737, 2016.
- Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2016.
- Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in neural information processing systems*, pages 1196–1204, 2013.
- Curtis G Northcutt, Tailin Wu, and Isaac L Chuang. Learning with confident examples: Rank pruning for robust classification with noisy labels. *arXiv* preprint *arXiv*:1705.01936, 2017.
- Yury Polyanskiy and Yihong Wu. Strong data-processing inequalities for channels and bayesian networks. In *Convexity and Concentration*, pages 211–249. Springer, 2017.
- Alfréd Rényi. On measures of dependence. *Acta mathematica hungarica*, 10(3-4):441–451, 1959.

- Mélanie Rey and Volker Roth. Meta-gaussian information bottleneck. In *Advances in Neural Information Processing Systems*, pages 1916–1924, 2012.
- Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P. Kingma. PixelCNN++: A PixelCNN Implementation with Discretized Logistic Mixture Likelihood and Other Modifications. In *ICLR*, 2017.
- Claude Elwood Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27:379–423, 1948.
- DJ Strouse and David J Schwab. The deterministic information bottleneck. *Neural computation*, 29(6):1611–1630, 2017a.
- DJ Strouse and David J Schwab. The information bottleneck and geometric clustering. *arXiv* preprint *arXiv*:1712.09657, 2017b.
- Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.
- Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, koray kavukcuoglu, Oriol Vinyals, and Alex Graves.
  Conditional Image Generation with PixelCNN Decoders. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, Advances in Neural Information Processing Systems 29, pages 4790–4798.
  Curran Associates, Inc., 2016.
- Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2691–2699, 2015.
- S. Zagoruyko and N. Komodakis. Wide Residual Networks. *arXiv:* 1605.07146, 2016.