PHASE TRANSITIONS FOR THE INFORMATION BOTTLENECK IN REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In the Information Bottleneck (IB) (Tishby et al., 2000), when tuning the relative strength between compression and prediction terms, how do the two terms behave, and what's their relationship with the dataset and the learned representation? In this paper, we set out to answer this question by studying multiple phase transitions in the IB objective: $\text{IB}_\beta[p(z|x)] = I(X;Z) - \beta I(Y;Z)$, where sudden jumps of $\frac{dI(Y;Z)}{d\beta}$ and prediction accuracy are observed with increasing β. We introduce a definition for IB phase transitions as a qualitative change of the IB loss landscape, and show that the transitions correspond to the onset of learning new classes. Using second-order calculus of variations, we derive a formula that provides a practical condition for IB phase transitions, and draw its connection with the Fisher information matrix for parameterized models. We provide two perspectives to understand the formula, revealing that each IB phase transition is finding a component of maximum (nonlinear) correlation between X and Y orthogonal to the learned representation, in close analogy with canonical-correlation analysis (CCA) in linear settings. Based on the theory, we present an algorithm for discovering phase transition points. Finally, we verify that our theory and algorithm accurately predict phase transitions in categorical datasets, predict the onset of learning new classes and class difficulty in MNIST, and predict prominent phase transitions in CIFAR10 experiments.

1 INTRODUCTION

The Information Bottleneck (IB) objective (Tishby et al., 2000):

$$\text{IB}_\beta[p(z|x)] = I(X;Z) - \beta I(Y;Z) \tag{1}$$

explicitly trades off model compression ($I(X;Z)$) with predictive performance ($I(Y;Z)$) using the Lagrange multiplier β, where X, Y are observed random variables, and Z is a learned representation of X. The IB method has proved effective in a variety of scenarios, including improving the robustness against adversarial attacks (Alemi et al., 2016; Fischer, 2018), learning invariant and disentangled representations (Achille & Soatto, 2018a;b), underlying information-based geometric clustering (Strouse & Schwab, 2017b), improving the training and performance in adversarial learning (Peng et al., 2018), and facilitating skill discovery (Sharma et al., 2019) and learning goal-conditioned policy (Goyal et al., 2019) in reinforcement learning.

From Eq. (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[1] that does not constrain the information flow. Between these two extremes, how will the IB objective behave? Will prediction and compression performance change smoothly, or do there exist interesting transitions in between? In Wu et al. (2019), the authors observe and study the learnability transition, i.e. the β value such that the IB objective transitions from a trivial global minimum to learning a nontrivial representation. They also show how this first phase transition relates to the structure of the dataset. However, to answer the full question, we need to consider the full range of β.

Motivation. To get a sense of how $I(Y;Z)$ and $I(X;Z)$ vary with β, we train Variational Information Bottleneck (VIB) models (Alemi et al., 2016) on the CIFAR10 dataset (Krizhevsky & Hinton,

[1] For example, in classification, it reduces to cross-entropy loss.

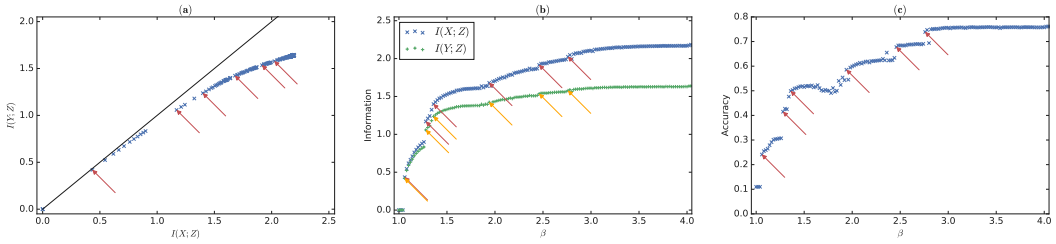


Figure 1: CIFAR10 plots **(a)** showing the information plane, as well as β vs **(b)** $I(X;Z)$ and $I(Y;Z)$, and **(c)** accuracy, all on the training set with 20% label noise. The arrows point to empirically-observed phase transitions. The vertical lines correspond to phase transitions found with Alg. 1.

2009), where each experiment is at a different β and random initialization of the model. Fig. 1 shows the $I(X;Z)$, $I(Y;Z)$ and accuracy vs. β, as well as $I(Y;Z)$ vs. $I(X;Z)$ for CIFAR10 with 20% label noise (see Appendix J for details).

From Fig. 1 **(b)(c)**, we see that as we increase β, instead of going up smoothly, both $I(X;Z)$ and $I(Y;Z)$ show multiple phase transitions, where the slopes $\frac{dI(X;Z)}{d\beta}$ and $\frac{dI(Y;Z)}{d\beta}$ are discontinuous and the accuracy has discrete jumps. The observation lets us refine our question: When do the phase transitions occur, and how do they depend on the structure of the dataset? These questions are important, since answering them will help us gain a better understanding of the IB objective and its close interplay with the dataset and the learned representation.

Moreover, the IB objective belongs to a general form of two-term trade-offs in many machine learning objectives: $L = \text{Prediction-loss} + \beta \cdot \text{Complexity}$, where the complexity term generally takes the form of regularization. Usually, learning is set at a specific β. But much more insights can be gained if we understand the behavior of the prediction loss and complexity with varying β, and how it depends on the dataset. The techniques developed to address the question in IB settings may also help us understand the two-term tradeoff in other learning objectives.

Contributions. In this work, we begin to address the above question in IB settings. Specifically:

- We identify a *qualitative* change of the IB loss landscape w.r.t. $p(z|x)$ for varying β as IB phase transitions (Section 3).
- Based on the definition, we introduce a quantity $G[p(z|x)]$ and use it to prove a theorem giving a practical condition for IB phase transitions. We further reveal the connection between $G[p(z|x)]$ and the Fisher information matrix when $p(z|x)$ is parameterized by $\boldsymbol{\theta}$ (Section 3).
- We reveal the close interplay between the IB objective, the dataset and the learned representation, by showing that in IB, each phase transition corresponds to learning a new nonlinear component of maximum correlation between X and Y, orthogonal to the previously-learned Z, and each with decreasing strength (Section 4).

We also present an algorithm for iteratively finding the IB phase transition points (Section 5). We show that our theory and algorithm give tight matches with the observed phase transitions in categorical datasets, predict the onset of learning new classes and class difficulty in MNIST, and predict prominent transitions in CIFAR10 experiments (Section 6).

2 Related Work

The Information Bottleneck Method (Tishby et al., 2000) provides a tabular method based on the Blahut-Arimoto (BA) Algorithm (Blahut, 1972) to numerically solve the IB functional for the optimal encoder distribution $P(Z|X)$, given the trade-off parameter β and the cardinality of the representation variable Z. This work has been extended in a variety of directions, including to the case where all three variables X, Y, Z are multivariate Gaussians (Chechik et al., 2005), cases of variational bounds

on the IB and related functionals for amortized learning (Alemi et al., 2016; Achille & Soatto, 2018a; Fischer, 2018), and a more generalized interpretation of the constraint on model complexity as a Kolmogorov Structure Function (Achille et al., 2018). Previous theoretical analyses of IB include Rey & Roth (2012), which looks at IB through the lens of copula functions, and Shamir et al. (2010), which starts to tackle the question of how to bound generalization with IB. We will make practical use of the original IB algorithm, as well as the amortized bounds of the Variational Informormation Bottleneck (Alemi et al., 2016) and the Conditional Entropy Bottleneck (Fischer, 2018).

Phase transitions, where key quantities change discontinuously with varying relative strength in the two-term trade-off, have been observed in many different learning domains, for multiple learning objectives. In Rezende & Viola (2018), the authors observe phase transitions in the latent representation of β-VAE for varying β. Strouse & Schwab (2017b) utilize the kink angle of the phase transitions in the Deterministic Information Bottleneck (DIB) (Strouse & Schwab, 2017a) to determine the optimal number of clusters for geometric clustering. Tegmark & Wu (2019) explicitly considers critical points in binary classification tasks using a discrete information bottleneck with a non-convex Pareto-optimal frontier. In Achille & Soatto (2018a), the authors observe a transition on the tradeoff of $I(\theta; X, Y)$ vs. $H(Y|X, \theta)$ in InfoDropout. Under IB settings, Chechik et al. (2005) study the Gaussian Information Bottleneck, and analytically solve the critical values $\beta_i^c = \frac{1}{1-\lambda_i}$, where λ_i are eigenvalues of the matrix $\Sigma_{x|y}\Sigma_x^{-1}$, and Σ_x is the covariance matrix. This work provides valuable insights for IB, but is limited to the special case that X, Y and Z are jointly Gaussian. Phase transitions in the general IB setting have also been observed, which Tishby (2018) describes as "information bifurcation". In Wu et al. (2019), the authors study the first phase transition, i.e. the learnability phase transition, and provide insights on how the the learnability depends on the dataset. Our work is the first work that addresses all the IB phase transitions in the most general setting, and provides theoretical insights on the interplay between the IB objective, its phase transitions, the dataset, and the learned representation.

3 Formula for IB phase transitions

To predict when the phase transitions will occur, a definition of IB phase transitions is needed. We can view the IB objective $\text{IB}_\beta[p(z|x)]$ (Eq. 1) as a functional of the encoding distribution $p(z|x)$. We define the IB phase transition as a change in the local curvature on the global minima of $\text{IB}_\beta[p(z|x)]$.

Definition 1. *The phase transitions β_i^c are the β values satisfying the following two conditions: (1) for all relative perturbation functions*[2] *$r(z|x)$, $\delta^2 IB_\beta[p(z|x)]\big|_{p^*_\beta(z|x)} \geq 0$; (2) For any positive η with $\eta \ll \beta$, there exists $r(z|x)$ such that $\delta^2 IB_{\beta+\eta}[p(z|x)]\big|_{p^*_\beta(z|x)} < 0$. Here $p^*_\beta(z|x)$ denotes the optimal solution of $IB_\beta[p(z|x)]$ at β. The relative perturbation function $r(z|x)$ is such that the perturbed probability (density) is $p'(z|x) = p(z|x)\,(1 + \epsilon \cdot r(z|x))$ for some positive ϵ with $|\epsilon| \ll 1$, $r(z|x)$ is defined on the support of $p(z|x)$ and bounded by some constant M, and satisfies $\mathbb{E}_{z\sim p(z|x)}[r(z|x)] = 0$ due to the normalization of probability.*

We can understand the $\delta^2\text{IB}_\beta[p(z|x)]$ as a local "curvature" of the IB objective IB_β (Eq. 1) w.r.t. $p(z|x)$, along some relative perturbation $r(z|x)$. A phase transition occurs when the convexity of $\text{IB}_\beta[p(z|x)]$ w.r.t. $p(z|x)$ changes from a minimum to a saddle point in the neighborhood of its optimal solution $p^*_\beta(z|x)$ as β increases from β_c to $\beta_c + \eta$. This means that there exists a perturbation to go downhill and find a better minimum. We validate this definition empirically below.

Definition 1 indicates the important role $\delta^2\text{IB}_\beta[p(z|x)]$ plays on the optimal solution in providing the condition for phase transitions. To concretize it and prepare for a more practical condition for IB phase transitions, we expand $\text{IB}_\beta[p(z|x)(1 + \epsilon \cdot r(z|x))]$ to the variational second order of ϵ, giving:

[2]The definition of second-order variation $\delta^2 F[f(x)]$ is provided in Appendix A.

Lemma 0.1. *For* $\mathrm{IB}_\beta[p(z|x)]$*, the condition of* $\forall r(z|x)$*,* $\delta^2\mathrm{IB}_\beta[p(z|x)] \geq 0$ *is equivalent to* $\beta \leq G[p(z|x)]$*. The threshold function* $G[p(z|x)]$ *is given by:*

$$\begin{aligned} G[p(z|x)] &\equiv \inf_{r(z|x)} \mathcal{G}[r(z|x); p(z|x)] \\ &\equiv \inf_{r(z|x)} \frac{\mathbb{E}_{x,z\sim p(x,z)}[r^2(z|x)] - \mathbb{E}_{z\sim p(z)}\left[\left(\mathbb{E}_{x\sim p(x|z)}[r(z|x)]\right)^2\right]}{\mathbb{E}_{y,z\sim p(y,z)}\left[\left(\mathbb{E}_{x\sim p(x|y,z)}[r(z|x)]\right)^2\right] - \mathbb{E}_{z\sim p(z)}\left[\left(\mathbb{E}_{x\sim p(x|z)}[r(z|x)]\right)^2\right]} \end{aligned} \tag{2}$$

where $r(z|x)$ *is the relative perturbation function satisfying* $\mathbb{E}_{z\sim p(z|x)}[r(z|x)] = 0$ *for any* x*.*

The proof is given in Appendix C, in which we also give Eq. (20) for empirical estimation. Note that Lemma 0.1 is very general and can be applied to any $p(z|x)$, not only at the optimal solution $p^*_\beta(z|x)$.

The Fisher Information matrix. In practice, the encoder $p_{\boldsymbol{\theta}}(z|x)$ is usually parameterized by some parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, ...\theta_k)^T \in \Theta$, e.g. weights and biases in a neural net. A small change of $\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta} + \Delta\boldsymbol{\theta}$ induces a relative perturbation $\epsilon \cdot r(z|x) \simeq \Delta\boldsymbol{\theta}^T \frac{\partial \log p_{\boldsymbol{\theta}}(z|x)}{\partial \boldsymbol{\theta}}$ on $p_{\boldsymbol{\theta}}(z|x)$, from which we can compute the threshold function $G_\Theta[p_{\boldsymbol{\theta}}(z|x)]$:

Lemma 0.2. *For* $\mathrm{IB}_\beta[p_{\boldsymbol{\theta}}(z|x)]$ *objective, the condition of* $\forall\Delta\boldsymbol{\theta}$*,* $\delta^2\mathrm{IB}_\beta[p_{\boldsymbol{\theta}}(z|x)] \geq 0$ *is equivalent to* $\beta \leq G_\Theta[p_{\boldsymbol{\theta}}(z|x)]$*, where*

$$G_\Theta[p_{\boldsymbol{\theta}}(z|x)] = \inf_{\Delta\boldsymbol{\theta}} \frac{\Delta\boldsymbol{\theta}^T\left(\mathcal{I}_{Z|X}(\boldsymbol{\theta}) - \mathcal{I}_Z(\boldsymbol{\theta})\right)\Delta\boldsymbol{\theta}}{\Delta\boldsymbol{\theta}^T\left(\mathcal{I}_{Z|Y}(\boldsymbol{\theta}) - \mathcal{I}_Z(\boldsymbol{\theta})\right)\Delta\boldsymbol{\theta}} = \lambda_{\max}^{-1} \tag{3}$$

where $\mathcal{I}_Z(\boldsymbol{\theta}) \equiv \int dz p_{\boldsymbol{\theta}}(z)\left(\frac{\partial \log p_{\boldsymbol{\theta}}(z)}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial \log p_{\boldsymbol{\theta}}(z)}{\partial\boldsymbol{\theta}}\right)^T$ *is the Fisher information matrix of* $\boldsymbol{\theta}$ *for* Z*,* $\mathcal{I}_{Z|X}(\boldsymbol{\theta}) \equiv \int dxdz p(x)p_{\boldsymbol{\theta}}(z|x)\left(\frac{\partial \log p_{\boldsymbol{\theta}}(z|x)}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial \log p_{\boldsymbol{\theta}}(z|x)}{\partial\boldsymbol{\theta}}\right)^T$*,* $\mathcal{I}_{Z|Y}(\boldsymbol{\theta}) \equiv \int dydz p(y)p_{\boldsymbol{\theta}}(z|y)\left(\frac{\partial \log p_{\boldsymbol{\theta}}(z|y)}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial \log p_{\boldsymbol{\theta}}(z|y)}{\partial\boldsymbol{\theta}}\right)^T$ *are the conditional Fisher information matrix (Zegers, 2015) of* $\boldsymbol{\theta}$ *for* Z *conditioned on* X *and* Y*, respectively.* $\lambda_{\max}$ *is the largest eigenvalue of* $C^{-1}\left(\mathcal{I}_{Z|Y}(\boldsymbol{\theta}) - \mathcal{I}_Z(\boldsymbol{\theta})\right)(C^T)^{-1}$ *with* $v_{\max}$ *the corresponding eigenvector, where* CC^T *is the Cholesky decomposition of the matrix* $\mathcal{I}_{Z|X}(\boldsymbol{\theta}) - \mathcal{I}_Z(\boldsymbol{\theta})$*, and* $v_{\max}$ *is the eigenvector for* $\lambda_{\max}$*. The infimum is attained at* $\Delta\boldsymbol{\theta} = (C^T)^{-1}v_{\max}$*.*

We see that for parameterized encoders $p_{\boldsymbol{\theta}}(z|x)$, each term of $G[p(z|x)]$ in Eq. (2) can be replaced by a bilinear form with the Fisher information matrix of the respective variables. The proof for Lemma 0.2 is in appendix D.

Phase Transitions. Now we introduce Theorem 1 that gives a concrete and practical condition for IB phase transitions, which is the core result of the paper:

Theorem 1. *The phase transition points* $\{\beta_i^c\}$ *are given by the roots of the following equation:*

$$G[p^*_\beta(z|x)] = \beta \tag{4}$$

where $G[p(z|x)]$ *is given by Eq. (2) and* $p^*_\beta(z|x)$ *is the optimal solution of* $IB_\beta[p(z|x)]$ *at* β*.*

We can understand Eq. (4) as the condition when $\delta^2\mathrm{IB}_\beta[p(z|x)]$ is *about* to be able to be negative at the optimal solution $p^*_\beta(z|x)$ for a given β. The proof for Theorem 1 is given in Appendix E.

4 Understanding the formula for IB phase transitions

In this section we set out to understand $G[p(z|x)]$ as given by Eq. (2) and the phase transition condition as given by Theorem 1, from the perspectives of Jensen's inequality and representational maximum correlation.

4.1 JENSEN'S INEQUALITY

The condition for IB phase transitions given by Theorem 1 involves $G[p(z|x)] = \inf_{r(z|x)} \mathcal{G}[r(z|x); p(z|x)]$ which is in itself an optimization problem. We can understand $G[p(z|x)] = \inf_{r(z|x)} \frac{A-C}{B-C}$ in Eq. (2) using Jensen's inequality:

$$\underbrace{\mathbb{E}_{x,z\sim p(x,z)}[r^2(z|x)]}_{A} \geq \underbrace{\mathbb{E}_{y,z\sim p(y,z)}\left[\left(\mathbb{E}_{x\sim p(x|y,z)}[r(z|x)]\right)^2\right]}_{B} \geq \underbrace{\mathbb{E}_{z\sim p(z)}\left[\left(\mathbb{E}_{x\sim p(x|z)}[r(z|x)]\right)^2\right]}_{C} \tag{5}$$

The equality between A and B holds when the perturbation $r(z|x)$ is constant w.r.t. x for any z; the equality between B and C holds when $\mathbb{E}_{x\sim p(x|y,z)}[r(z|x)]$ is constant w.r.t. y for any z. Therefore, the minimization of $\frac{A-C}{B-C}$ encourages the relative perturbation $r(z|x)$ to be as constant w.r.t. x as possible (minimizing intra-class difference), but as different w.r.t. different y as possible (maximizing inter-class difference), resulting in a *clustering* of the values of $r(z|x)$ for different examples x according to their class y.

4.2 REPRESENTATIONAL MAXIMUM CORRELATION

Under certain conditions we can further simplify $G[p(z|x)]$ and gain a deeper understanding of it. Firstly, inspired by maximum correlation (Anantharam et al., 2013), we introduce two new concepts, *representational maximum correlation* and *conditional maximum correlation*, as follows.

Definition 2. *The representational maximum correlation $\rho_r(X, Y; Z)$ is*

$$\rho_r(X, Y; Z) = \sup_{f(X,Z), g(Y,Z)} \mathbb{E}[f(X, Z)g(Y, Z)] \tag{6}$$

such that $\mathbb{E}_{x\sim p(x|z)}[f(x, z)] = \mathbb{E}_{y\sim p(y|z)}[g(y, z)] = 0$ and $\mathbb{E}[f^2(X, Z)] = \mathbb{E}[g^2(Y, Z)] = 1$, where f and g are real-valued functions.

The conditional maximum correlation $\rho_m(X, Y|Z)$ is:

$$\rho_m(X, Y|Z) \equiv \sup_{f,g} \mathbb{E}_{X,Y\sim p(X,Y|Z)}[f(X)g(Y)] \tag{7}$$

with $\mathbb{E}_{X\sim p(X|Z)}[f(X)] = \mathbb{E}_{Y\sim p(Y|Z)}[g(Y)] = 0$, $\mathbb{E}_{X\sim p(X|Z)}[f^2(X)] = \mathbb{E}_{Y\sim p(Y|Z)}[g^2(Y)] = 1$.

We prove the following Theorem 2, which expresses $G[p(z|x)]$ in terms of representational maximum correlation and related quantities, with proof given in Appendix G.

Theorem 2. *If $\forall f(x, z)$, it can be decomposed*[3] *as $f(x, z) = h(x, z) + s(z)$ such that $\mathbb{E}_{p(z|x)}[h(x, z)] = 0$, then we have:*

(i) The representation maximum correlation and G:

$$G[p(z|x)] = \frac{1}{\rho_r^2(X, Y; Z)} \tag{8}$$

(ii) The representational maximum correlation and conditional maximum correlation:

$$\rho_r(X, Y; Z) = \sup_Z [\rho_m(X, Y|Z)] \tag{9}$$

(iii) When Z is continuous, an optimal relative perturbation function $r(z|x)$ for $G[p(z|x)]$ is given by

$$r^*(z|x) = h^*(x)\sqrt{\frac{\delta(z - z^*)}{p(z)}} \tag{10}$$

[3]For discrete X, Z such that the cardinality $|Z| \geq |X|$, this is generally true since in this scenario, $h(x, z)$ and $s(z)$ have $|X||Z| + |Z|$ unknown variables, but the condition has only $|X||Z| + |X|$ linear equations.

where $z^* = \arg\max_z \rho_m(X, Y|Z = z)$, and $h^*(x)$ is the optimal solution for the learnability threshold function $h^*(x) = \arg\min_{h(x)} \beta_0[h(x)]$ with $p(X, Y|Z = z^*)$ ($\beta_0[h(x)]$ is given in Theorem 4 of Wu et al. (2019)).

(iv) For discrete X, Y and Z, we have

$$\rho_r(X, Y; Z) = \max_Z \sigma_2(Z) \tag{11}$$

where $\sigma_2(Z)$ is the second largest singular value of the matrix $Q_{X,Y|Z} \equiv \left(\frac{p(x,y|z)}{\sqrt{p(x|z)p(y|z)}}\right)_{x,y} = \left(\frac{p(x,y)}{\sqrt{p(x)p(y)}}\sqrt{\frac{p(z|x)}{p(z|y)}}\right)_{x,y}$.

Theorem 2 furthers our understanding of $G[p(z|x)]$ and the phase transition condition (Theorem 1), which we elaborate as follows.

Discovering maximum correlation in the orthogonal space of a learned representation: Intuitively, the representational maximum correlation measures the maximum linear correlation between $f(X, Z)$ and $g(Y, Z)$ among all real-valued functions f, g, under the constraint that $f(X, Z)$ is "orthogonal" to $p(X|Z)$ and $g(Y, Z)$ is "orthogonal" to $p(Y|Z)$. Theorem 2 (i) reveals that $G[p(z|x)]$ is the inverse square of this representational maximum correlation. Theorem 2 (ii) further shows that $G[p(z|x)]$ is finding a specific z^* on which maximum (nonlinear) correlation between X and Y conditioned on Z can be found. Combined with Theorem 1, we have that when we continuously increase β, for the optimal representation Z^*_β given by $p^*_\beta(z|x)$ at β, $\rho_r(X, Y; Z^*_\beta)$ shall monotonically decrease due to that X and Y has to find their maximum correlation on the orthogonal space of an increasingly better representation Z^*_β that captures more information about X. A phase transition occurs when $\rho_r(X, Y; Z^*_\beta)$ reduces to $\frac{1}{\sqrt{\beta}}$, after which as β continues to increase, $\rho_r(X, Y; Z^*_\beta)$ will try to find maximum correlation between X and Y orthogonal to the full previously learned representation. This is reminiscent of canonical-correlation analysis (CCA) (Hotelling, 1992) in linear settings, where components with decreasing linear maximum correlation that are orthogonal to previous components are found one by one. In comparison, we show that in IB, each phase transition corresponds to learning a new *nonlinear* component of maximum correlation between X and Y in Z, orthogonal to the previously-learned Z. In the case of classification where different classes may have different difficulty (e.g. due to label noise or support overlap), we should expect that classes that are less difficult as measured by a larger maximum correlation between X and Y are learned earlier.

Conspicuous subset conditioned on a single z: Furthermore, we show in (iii) that an optimal relative perturbation function $r(z|x)$ can be decomposed into a product of two factors, a $\sqrt{\frac{\delta(z-z^*)}{p(z)}}$ factor that only focus on perturbing a specific point z^* in the representation space, and an $h^*(x)$ factor that is finding the "conspicuous subset" (Wu et al., 2019), i.e. the most confident, large, typical, and imbalanced subset in the X space for the distribution $(X, Y) \sim p(X, Y|z^*)$.

Singular values In categorical settings, (iv) reveals a connection between $G[p(z|x)]$ and the singular value of the $Q_{X,Y|Z}$ matrix. Due to the property of SVD, we know that the square of the singular values of $Q_{X,Y|Z}$ equals the non-negative eigenvalue of the matrix $Q^T_{X,Y|Z}Q_{X,Y|Z}$. Then the phase transition condition in Theorem 1 is equivalent to a (nonlinear) eigenvalue problem. This is resonant with previous analogy with CCA in linear settings, and is also reminiscent of the linear eigenvalue problem in Gaussian IB (Chechik et al., 2005).

5 Algorithm for Phase Transitions Discovery in Classification

Typically, classification involves high-dimensional inputs X. Without sweeping the full range of β where at each β it is a full learning problem, it is in general a difficult task to estimate the phase transitions. In Algorithm 1, we present a two-stage approach.

In the first stage, we train a single maximum likelihood neural network $f_{\boldsymbol{\theta}}$ with the same encoder architecture as in the (variational) IB to estimate $p(y|x)$, and obtain an $N \times C$ matrix $p(y|x)$, where

Algorithm 1 Phase transitions discovery for IB

Require (X, Y): the dataset
Require $f_{\boldsymbol{\theta}}$: a neural net with the same encoder architecture as the (variational) IB
Require K: patience
Require δ: precision floor
Require R: maximum ratio between $\beta^{(\text{th})}$ and β.
// First stage: fit $p(y|x)$ using neural net $f_{\boldsymbol{\theta}}$:
1: $p(y|x) \leftarrow$ fitting (X, Y) using $f_{\boldsymbol{\theta}}$ via maximum likelihood.
2: $p(x) \leftarrow \frac{1}{N}$

// Second stage: coordinate descent using $G[p(z|x)]$ and IB algorithm:
3: $\beta_0^c \leftarrow \beta^{(\text{th})}(1)$
4: $Q \leftarrow \{\beta_0^c\}$
5: $(\beta^{(\text{new})}, \beta, \text{count}) \leftarrow (\beta_0^c, 1, 0)$
6: **while** $\frac{\beta^{(\text{new})}}{\beta} < R$ **do**:
7: $\quad \beta \leftarrow \beta^{(\text{new})}$
8: $\quad \beta^{(\text{new})} \leftarrow \beta^{(\text{th})}(\beta)$
9: $\quad$ **if** $\beta^{(\text{new})} - \beta < \delta$ **do**:
10: $\quad\quad$ count $\leftarrow$ count $+ 1$
11: $\quad\quad$ **if** count $> K$ **do**:
12: $\quad\quad\quad Q \leftarrow Q \cup \{\beta^{(\text{new})}\}$
13: $\quad\quad\quad \beta^{(\text{new})} \leftarrow \beta^{(\text{new})} + \delta$
14: $\quad\quad$ **end if**
15: $\quad$ **else**: count $\leftarrow 0$
16: $\quad$ **end if**
17: **end while**
18: **return** Q

subroutine $\beta^{(\text{th})}(\beta)$:
s1: Compute $p^*_\beta(z|x)$ using the IB algorithm (Tishby et al., 2000).
s2: $\beta^{(\text{new})} \leftarrow G[p^*_\beta(z|x)]$ using SVD (Eq. 8 and 11).
s3: **return** $\beta^{(\text{new})}$

N is the number of examples in the dataset and C is the number of classes. In the second stage, we perform an iterative algorithm w.r.t. G and β, alternatively, to converge to a phase transition point.

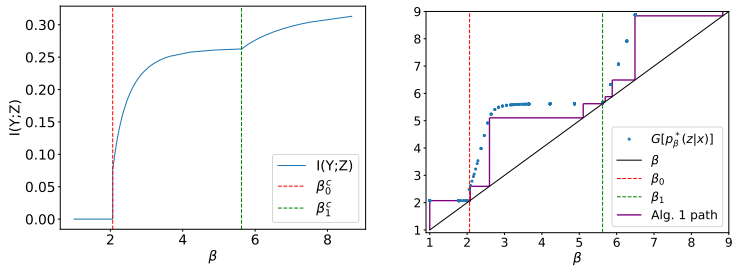
Specifically, for a given β, we use a Blahut-Arimoto type IB algorithm (Tishby et al., 2000) to efficiently reach IB optimal $p^*_\beta(z|x)$ at β, then use SVD (with the formula given in Theorem 2 (iv)) to efficiently estimate $G[p^*_\beta(z|x)]$ at β (step 8). We then use the $G[p^*_\beta(z|x)]$ value as the new β and do it again (step 7 in the next iteration). At convergence, we will reach the phase transition point given by $G[p^*_\beta(z|x)] = \beta$ (Theorem 1). After convergence as measured by patience parameter K, we slightly increase β by δ (step 13), so that the algorithm can discover the subsequent phase transitions.

6 Empirical study

We quantitatively and qualitatively test the ability of our theory and Algorithm 1 to provide good predictions for IB phase transitions. We first verify them in fully categorical settings, where X, Y, Z are all discrete, and we show that the phase transitions can correspond to learning new classes as we increase β. We then test our algorithm on versions of the MNIST and CIFAR10 datasets with added label noise.

6.1 Categorical dataset

For categorical datasets, X and Y are discrete, and $p(X)$ and $p(Y|X)$ are given. To test Theorem 1, we use the Blahut-Arimoto IB algorithm to compute the optimal $p^*_\beta(z|x)$ for each β. $I(Y; Z^*)$



(a) (b)

Figure 2: (a) $I(Y;Z^*)$ vs. β for a categorical dataset with $|X| = |Y| = |Z| = 3$, where Z^* is given by $p^*_\beta(z|x)$, and the vertical lines are the experimentally discovered phase transition points β_0^c and β_1^c. (b) $G[p^*_\beta(z|x)]$ vs. β for the same dataset, and the path for Alg. 1, with β_0^c and β_1^c in (a) also plotted. The dataset is given in Fig. 5.

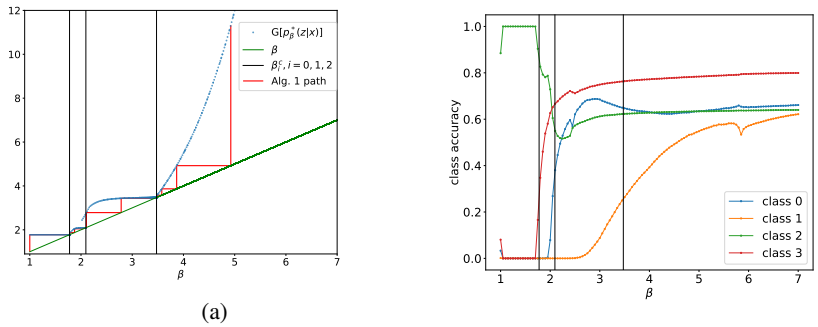
vs. β is plotted in Fig. 2 (a). There are two phase transitions at β_0^c and β_1^c. For each β and the corresponding $p^*_\beta(z|x)$, we use the SVD formula (Theorem 2) to compute $G[p^*_\beta(z|x)]$, shown in Fig. 2 (b). We see that $G[p^*_\beta(z|x)] = \beta$ at exactly the observed phase transition points β_0^c and β_1^c. Moreover, starting at $\beta = 1$, Alg. 1 converges to each phase transition points within few iterations. Our other experiments with random categorical datasets show similarly tight matches.

Furthermore, in Appendix H we show that the phase transitions correspond to the onset of separation of $p(z|x)$ for subsets of X that correspond to different classes. Connecting with the previous analysis of representational maximum correlation, we see that in IB, the different components of maximum correlation (as separated by phase transitions) correspond to the learning of new classes. Thus, we conjecture that in IB for classification, the number of phase transitions should be at most $C - 1$, where C is the number of classes.

6.2 MNIST DATASET

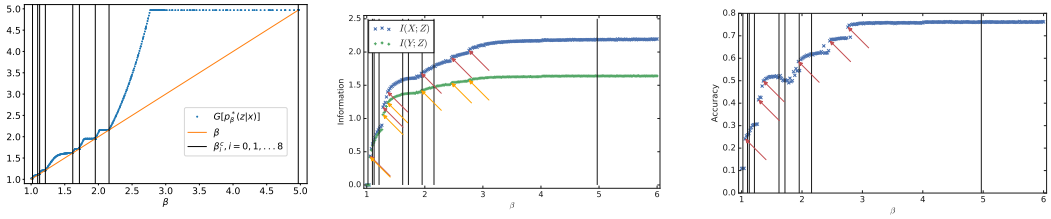
For continuous X, how does our algorithm perform, and will it reveal aspects of the dataset? We first test our algorithm in a 4-class MNIST with noisy labels[4], whose confusion matrix and experimental settings are given in Appendix I. Fig. 3 (a) shows the path Alg. 1 takes. We see again that in each phase Alg. 1 converges to the phase transition points within a few iterations, and it discovers in total 3 phase transition points. Similar to the categorical case, we expect that each phase transition corresponds to the onset of learning a new class, and that the last class is much harder to learn due to a larger separation of β. Therefore, this class should have a much larger label noise so that it is hard to capture this component of maximum correlation between X and Y, as analyzed in representational maximum correlation (Section 4.2). Fig. 3 (b) plots the per-class accuracy with increasing β for running the Conditional Entropy Bottleneck (Fischer, 2018) (another variational bound on IB). We see that the first two predicted phase transition points β_0^c, β_1^c closely match the observed onset of learning class 3 and class 0. Class 1 is observed to learn earlier than expected, possibly due to the gap between the variational IB objective and the true IB objective in continuous settings. By looking at the confusion matrix for the label noise (Fig. 7), we see that the ordering of onset of learning: class 2, 3, 0, 1, corresponds exactly to the decreasing diagonal element $p(\tilde{y} = 1|y = 1)$ (increasing noise) of the classes, and as predicted, class 1 has a much smaller diagonal element $p(\tilde{y} = 1|y = 1)$ than the other three classes, which makes it much more difficult to learn. This ordering of classes by difficulty is what our representational maximum correlation predicts.

[4]We use 4 classes since it is simpler than the full 10 classes, but still potentially possesses phase transitions. We use noisy label to mimic realistic settings where the data may be noisy and also to have controllable difficulty for different classes.



(a) (b)

Figure 3: (a) Path of Alg. 1 starting with $\beta = 1$, where the maximum likelihood model $f_{\boldsymbol{\theta}}$ is using the same encoder architecture as in the CEB model. This stairstep path shows that Alg. 1 is able to ignore very large regions of β, while quickly and precisely finding the phase transition points. Also plotted is an accumulation of $G[p^*_\beta(z|x)]$ vs. β by running Alg. 1 with varying starting β (blue dots). (b) Per-class accuracy vs. β, where the accuracy at each β is from training an independent CEB model on the dataset. The per-class accuracy denotes the fraction of correctly predicted labels by the CEB model for the observed label $\tilde{y}$.



(a) (b) (c)

Figure 4: **(a)** Accumulated $G[p^*_\beta(z|x)]$ vs. β by running Alg. 1 with varying starting β (blue dots). Also plotted are predicted phase transition points. **(b)** $I(X;Z)$ and $I(Y;Z)$ vs. β. The manually-identified phase transition points are labelled with arrows. The vertical black lines are the phase transitions identified by Alg. 1, denoted as β^c_0, β^c_1,... β^c_8, from left to right. **(c)** Accuracy vs. β with the same sets of points identified. The most interesting region is right before $\beta = 2$, where accuracy *decreases* with β. Alg. 1 identifies both sides of that region, as well as points at or near all of the early obvious phase transitions. It also seems to miss later transitions, possibly due to the gap between the variational IB objective and the true IB objective in continuous settings.

6.3 CIFAR10 DATASET

Finally, we investigate the CIFAR10 experiment from Section 1. The details of the experimental setup are described in Appendix J. This experiment stretches the current limits of our discrete approximation to the underlying continuous representation being learned by the models. Nevertheless, we can see in Fig. 4 that many of the visible empirical phase transitions are tightly identified by Alg. 1. Particularly, the onset of learning is predicted quite accurately; the large interval between the predicted $\beta_3 = 1.21$ and $\beta_4 = 1.61$ corresponds well to the continuous increase of $I(X;Z)$ and $I(Y;Z)$ at the same interval. And Alg. 1 is able to identify many dense transitions not obviously seen by just looking at $I(Y;Z)$ vs. β curve alone. Alg. 1 predicts 9 phase transitions, exactly equal to $C - 1$ for CIFAR10.

7 CONCLUSION

In this work, we observe and study the phase transitions in IB as we vary β. We introduce the definition for IB phase transitions, and based on it derive a formula that gives a practical condition for IB phase transitions. We further understand the formula via Jensen's inequality and representational

maximum correlation. We reveal the close interplay between the IB objective, the dataset and the learned representation, as each phase transition is learning a nonlinear maximum correlation component in the orthogonal space of the learned representation. We present an algorithm for finding the phase transitions, and show that it gives tight matches with observed phase transitions in categorical datasets, predicts onset of learning new classes and class difficulty in MNIST, and predicts prominent transitions in CIFAR10 experiments. This work is a first theoretical step towards a deeper understanding of the phenomenon of phase transitions in the Information Bottleneck. We believe our approach will be applicable to other "trade-off" objectives, like β-VAE (Higgins et al., 2017) and InfoDropout (Achille & Soatto, 2018a), where the model's ability to predict is balanced against a measure of complexity.

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.

Richard Blahut. Computation of channel capacity and rate-distortion functions. *IEEE transactions on Information Theory*, 18(4):460–473, 1972.

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.

Ian Fischer. The conditional entropy bottleneck, 2018. URL `openreview.net/forum?id=rkVOXhAqY7`.

Anirudh Goyal, Riashat Islam, Daniel Strouse, Zafarali Ahmed, Matthew Botvinick, Hugo Larochelle, Sergey Levine, and Yoshua Bengio. Infobot: Transfer and exploration via the information bottleneck. *arXiv preprint arXiv:1901.10902*, 2019.

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.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=Sy2fzU9gl`.

Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics*, pp. 162–190. Springer, 1992.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL `https://arxiv.org/abs/1412.6980`.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, CIFAR, 2009.

Xue Bin Peng, Angjoo Kanazawa, Sam Toyer, Pieter Abbeel, and Sergey Levine. Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow. *arXiv preprint arXiv:1810.00821*, 2018.

Mélanie Rey and Volker Roth. Meta-gaussian information bottleneck. In *Advances in Neural Information Processing Systems*, pp. 1916–1924, 2012.

Danilo Jimenez Rezende and Fabio Viola. Taming VAEs. *arXiv preprint arXiv:1810.00597*, 2018.

Ohad Shamir, Sivan Sabato, and Naftali Tishby. Learning and generalization with the information bottleneck. *Theoretical Computer Science*, 411(29-30):2696–2711, 2010.

Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-aware unsupervised discovery of skills. *arXiv preprint arXiv:1907.01657*, 2019.

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.

Max Tegmark and Tailin Wu. Pareto-optimal data compression for binary classification tasks. *arXiv preprint arXiv:1908.08961*, 2019.

Naftali Tishby. Lecture: the information theory of deep neural networks: the statistical physics aspects. `https://www.perimeterinstitute.ca/videos/information-theory-deep-neural-networks-statistical-physics-aspects/`, 2018.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Tailin Wu, Ian Fischer, Isaac Chuang, and Max Tegmark. Learnability for the information bottleneck. *arXiv preprint arXiv:1907.07331*, 2019.

S. Zagoruyko and N. Komodakis. Wide Residual Networks. *arXiv: 1605.07146*, 2016.

Pablo Zegers. Fisher information properties. *Entropy*, 17(7):4918–4939, 2015.

Appendix

In this paper, we implicitly assume that the integral (or summing if X, Y or Z are discrete random variables) is on the support of $P(X, Y, Z)$, i.e. $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. We assume that the relative perturbation function $r(z|x)$ is continuous on $\mathcal{X} \times \mathcal{Z}$, and there exists a constant M such that $\left|r(z|x)\right| < M$, $\forall (x, z) \in \mathcal{X} \times \mathcal{Z}$.

A 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 $F[\cdot]$.

B CALCULUS OF VARIATIONS AT ANY ORDER OF $\mathrm{IB}_\beta[p(z|x)]$

Here we prove the following Lemma, which will be crucial in the lemmas and theorems in this paper that follows.

Lemma 2.1. *For a relative perturbation function $r(z|x)$ on a $p(z|x)$, where $r(z|x)$ satisfies $\mathbb{E}_{z\sim p(z|x)}[r(z|x)] = 0$, we have that the IB objective can be expanded as*

$$\begin{aligned}
&\mathrm{IB}_\beta[p(z|x)(1 + \epsilon \cdot r(z|x))] \\
=&\mathrm{IB}_\beta[p(z|x)] + \epsilon \cdot \left(\mathbb{E}_{x,z\sim p(x,z)} \left[r(z|x) \log \frac{p(z|x)}{p(z)} \right] - \beta \cdot \mathbb{E}_{y,z\sim p(y,z)} \left[r(z|y) \log \frac{p(z|y)}{p(z)} \right] \right) \\
&+ \sum_{n=2}^{\infty} \frac{(-1)^n \epsilon^n}{n(n-1)} \left\{ (\mathbb{E}[r^n(z|x)] - \mathbb{E}[r^n(z)]) - \beta \cdot (\mathbb{E}[r^n(z|y)] - \mathbb{E}[r^n(z)]) \right\} \\
=&\mathrm{IB}_\beta[p(z|x)] + \epsilon \cdot \left(\mathbb{E}_{x,z\sim p(x,z)} \left[r(z|x) \log \frac{p(z|x)}{p(z)} \right] - \beta \cdot \mathbb{E}_{y,z\sim p(y,z)} \left[r(z|y) \log \frac{p(z|y)}{p(z)} \right] \right) \\
&+ \frac{\epsilon^2}{1 \cdot 2} \left\{ \left(\mathbb{E}[r^2(z|x)] - \mathbb{E}[r^2(z)] \right) - \beta \cdot \left(\mathbb{E}[r^2(z|y)] - \mathbb{E}[r^2(z)] \right) \right\} \\
&- \frac{\epsilon^3}{2 \cdot 3} \left\{ \left(\mathbb{E}[r^3(z|x)] - \mathbb{E}[r^3(z)] \right) - \beta \cdot \left(\mathbb{E}[r^3(z|y)] - \mathbb{E}[r^3(z)] \right) \right\} \\
&+ \frac{\epsilon^4}{3 \cdot 4} \left\{ \left(\mathbb{E}[r^4(z|x)] - \mathbb{E}[r^4(z)] \right) - \beta \cdot \left(\mathbb{E}[r^4(z|y)] - \mathbb{E}[r^4(z)] \right) \right\} \\
&- \ldots
\end{aligned} \tag{12}$$

where $r(z|y) = \mathbb{E}_{x\sim p(x|y,z)}[r(z|x)]$ and $r(z) = \mathbb{E}_{x\sim p(x|z)}[r(z|x)]$. The expectations in the equations are all w.r.t. all variables. For example $\mathbb{E}[r^2(z|x)] = \mathbb{E}_{x,z\sim p(x,z)}[r^2(z|x)]$.

Proof. Suppose that we perform a relative perturbation $r(z|x)$ on $p(z|x)$ such that the perturbed conditional probability is $p'(z|x) = p(z|x)\,(1 + \epsilon \cdot r(z|x))$, then we have

$$p'(z) = \int p(x)p'(z|x)dx = \int dxp(x)p(z|x)\,(1 + \epsilon \cdot r(z|x)) = p(z) + \epsilon \cdot \int dxp(x)p(z|x)r(z|x)$$

Therefore, we can denote the corresponding relative perturbation $r(z)$ on $p(z)$ as

$$r(z) \equiv \frac{1}{\epsilon}\frac{p'(z) - p(z)}{p(z)} = \frac{1}{p(z)}\int dxp(x)p(z|x)r(z|x) = \mathbb{E}_{x\sim p(x|z)}[r(z|x)]$$

Similarly, we have

$$p'(z|y) = \frac{p'(y,z)}{p(y)} = \frac{1}{p(y)}\int dxp(x,y)p(z|x)\,(1 + \epsilon \cdot r(z|x)) = p(z|y) + \epsilon \cdot \frac{1}{p(y)}\int dxp(x,y)p(z|x)r(z|x)$$

And we can denote the corresponding relative perturbation $r(z|y)$ on $p(z|y)$ as

$$r(z|y) \equiv \frac{1}{\epsilon}\frac{p'(z|y) - p(z|y)}{p(z|y)} = \frac{1}{p(z|y)p(y)}\int dxp(x,y)p(z|x)r(z|x) = \mathbb{E}_{x\sim p(x|y,z)}[r(z|x)]$$

Since

$$\mathrm{IB}_\beta[p(z|x)] = I(X;Z) - \beta \cdot I(Y;Z) = \int dxdzp(x,z)\log\frac{p(z|x)}{p(z)} - \beta \cdot \int dydzp(y,z)\log\frac{p(z|y)}{p(z)}$$

We have

$$\begin{aligned}
&\mathrm{IB}_\beta[p'(z|x)] = \mathrm{IB}_\beta[p(z|x)(1 + \epsilon \cdot r(z|x))] \\
&= \int dxdzp(x)p'(z|x)\log\frac{p'(z|x)}{p'(z)} - \beta \cdot \int dydzp(y)p'(z|y)\log\frac{p'(z|y)}{p'(z)} \\
&= \int dxdzp(x)p(z|x)(1 + \epsilon \cdot r(z|x))\log\frac{p(z|x)(1 + \epsilon \cdot r(z|x))}{p(z)(1 + \epsilon \cdot r(z))} \\
&\quad - \beta \cdot \int dydzp(y)p(z|y)(1 + \epsilon \cdot r(z|y))\log\frac{p(z|y)(1 + \epsilon \cdot r(z|y))}{p(z)(1 + \epsilon \cdot r(z))} \\
&= \int dxdzp(x)p(z|x)(1 + \epsilon \cdot r(z|x))\left[\log\frac{p(z|x)}{p(z)} + \log\left(1 + \epsilon \cdot r(z|x)\right) - \log\left(1 + \epsilon \cdot r(z)\right)\right] \\
&\quad - \beta \cdot \int dydzp(y)p(z|y)(1 + \epsilon \cdot r(z|y))\left[\log\frac{p(z|y)}{p(z)} + \log\left(1 + \epsilon \cdot r(z|y)\right) - \log\left(1 + \epsilon \cdot r(z)\right)\right] \\
&= \int dxdzp(x)p(z|x)(1 + \epsilon \cdot r(z|x))\left[\log\frac{p(z|x)}{p(z)} + \sum_{n=1}^{\infty}(-1)^{n-1}\frac{\epsilon^n}{n}\left(r(z|x) - r(z)\right)\right] \\
&\quad - \beta \cdot \int dydzp(y)p(z|y)(1 + \epsilon \cdot r(z|y))\left[\log\frac{p(z|y)}{p(z)} + \sum_{n=1}^{\infty}(-1)^{n-1}\frac{\epsilon^n}{n}\left(r(z|y) - r(z)\right)\right]
\end{aligned}$$

The 0th-order term is simply $\mathrm{IB}_\beta[p(z|x)]$. The first order term is

$$\delta\mathrm{IB}_\beta[p(z|x)] = \epsilon \cdot \left(\mathbb{E}_{x,z\sim p(x,z)}\left[r(z|x)\log\frac{p(z|x)}{p(z)}\right] - \beta \cdot \mathbb{E}_{y,z\sim p(y,z)}\left[r(z|y)\log\frac{p(z|y)}{p(z)}\right]\right)$$

The nth-order term for $n \geq 2$ is

$$
\begin{aligned}
&\delta^n \mathrm{IB}_\beta[p(z|x)] \\
=&(-1)^n \epsilon^n \int dxdzp(x)p(z|x) \left(-\frac{1}{n} \left[r^n(z|x) - r^n(z) \right] + r(z|x) \frac{1}{n-1} \left[r^{n-1}(z|x) - r^n(z) \right] \right) \\
&- \beta \cdot (-1)^n \epsilon^n \int dydzp(y)p(z|y) \left(-\frac{1}{n} \left[r^n(z|y) - r^n(z) \right] + r(z|y) \frac{1}{n-1} \left[r^{n-1}(z|y) - r^n(z) \right] \right) \\
=&\frac{(-1)^n \epsilon^n}{n(n-1)} \left(\mathbb{E}_{x,z\sim p(x,z)}[r^n(z|x)] - n\mathbb{E}_{x,z\sim p(x,z)}[r(z|x)r^{n-1}(z)] - (n-1)\mathbb{E}_{z\sim p(z)}[r^n(z)] \right) \\
&- \beta \cdot \frac{(-1)^n \epsilon^n}{n(n-1)} \left(\mathbb{E}_{y,z\sim p(y,z)}[r^n(z|y)] - n\mathbb{E}_{y,z\sim p(y,z)}[r(z|y)r^{n-1}(z)] - (n-1)\mathbb{E}_{z\sim p(z)}[r^n(z)] \right) \\
=&\frac{(-1)^n \epsilon^n}{n(n-1)} \left\{ (\mathbb{E}[r^n(z|x)] - \mathbb{E}[r^n(z)]) - \beta \cdot (\mathbb{E}[r^n(z|y)] - \mathbb{E}[r^n(z)]) \right\}
\end{aligned}
$$

In the last equality we have used

$$
\mathbb{E}_{x,z\sim p(x,z)}[r(z|x)r^{n-1}(z)] = \mathbb{E}_{z\sim p(z)}[r^{n-1}(z)\mathbb{E}_{x\sim p(x|z)}[r(z|x)]] = \mathbb{E}_{z\sim p(z)}[r^{n-1}(z)r(z)] = \mathbb{E}_{z\sim p(z)}[r^n(z)]
$$

Combining the terms with all orders, we have

$$
\begin{aligned}
&\mathrm{IB}_\beta[p(z|x)(1+\epsilon \cdot r(z|x))] \\
=&\mathrm{IB}_\beta[p(z|x)] + \epsilon \cdot \left(\mathbb{E}_{x,z\sim p(x,z)} \left[r(z|x) \log \frac{p(z|x)}{p(z)} \right] - \beta \cdot \mathbb{E}_{y,z\sim p(y,z)} \left[r(z|y) \log \frac{p(z|y)}{p(z)} \right] \right) \\
&+ \sum_{n=2}^{\infty} \frac{(-1)^n \epsilon^n}{n(n-1)} \left\{ (\mathbb{E}[r^n(z|x)] - \mathbb{E}[r^n(z)]) - \beta \cdot (\mathbb{E}[r^n(z|y)] - \mathbb{E}[r^n(z)]) \right\}
\end{aligned}
$$

□

As a side note, the KL-divergence between $p'(z|x) = p(z|x)(1+\epsilon \cdot r(z|x))$ and $p(z|x)$ is

$$
\begin{aligned}
\mathrm{KL}\left(p'(z|x)||p(z|x)\right) &= \int dzp(z|x)(1+\epsilon \cdot r(z|x)) \log \frac{p(z|x)(1+\epsilon \cdot r(z|x))}{p(z|x)} \\
&= \int dzp(z|x)(1+\epsilon \cdot r(z|x)) \left(\epsilon \cdot r(z|x) - \frac{\epsilon^2}{2} \cdot r^2(z|x)) + O(\epsilon^3) \right) \\
&= \epsilon \cdot \int dzp(z|x)r(z|x) + \frac{\epsilon^2}{2} \int dzp(z|x)r^2(z|x) + O(\epsilon^3) \\
&= \frac{\epsilon^2}{2} \mathbb{E}_{z\sim p(z|x)}[r^2(z|x)] + O(\epsilon^3)
\end{aligned}
$$

Therefore, to the second order, we have

$$
\mathbb{E}_{x\sim p(x)} \left[\mathrm{KL}\left(p'(z|x)||p(z|x)\right) \right] = \frac{\epsilon^2}{2} \mathbb{E}[r^2(z|x)] \tag{13}
$$

Similarly, we have $\mathbb{E}_{x\sim p(x)} \left[\mathrm{KL}\left(p(z|x)||p'(z|x)\right) \right] = \frac{\epsilon^2}{2} \mathbb{E}[r^2(z|x)]$ up to the second order. Using similar procedure, we have up to the second-order,

$$
\begin{aligned}
&\mathbb{E}_{y\sim p(y)} \left[\mathrm{KL}\left(p'(z|y)||p(z|y)\right) \right] = \mathbb{E}_{y\sim p(y)} \left[\mathrm{KL}\left(p(z|y)||p'(z|y)\right) \right] = \frac{\epsilon^2}{2} \mathbb{E}[r^2(z|y)] \\
&\mathrm{KL}\left(p'(z)||p(z)\right) = \mathrm{KL}\left(p(z)||p'(z)\right) = \frac{\epsilon^2}{2} \mathbb{E}[r^2(z)]
\end{aligned}
$$

C PROOF OF LEMMA 0.1

Proof. From Lemma 2.1, we have

$$\delta^2 \mathrm{IB}_\beta[p(z|x)] = \frac{\epsilon^2}{2}\left\{\left(\mathbb{E}[r^2(z|x)] - \mathbb{E}[r^2(z)]\right) - \beta \cdot \left(\mathbb{E}[r^2(z|y)] - \mathbb{E}[r^2(z)]\right)\right\} \tag{14}$$

The condition of

$$\forall r(z|x), \delta^2 \mathrm{IB}_\beta[p(z|x)] \geq 0 \tag{15}$$

is equivalent to

$$\forall r(z|x), \beta \cdot \left(\mathbb{E}[r^2(z|y)] - \mathbb{E}[r^2(z)]\right) \leq \mathbb{E}[r^2(z|x)] - \mathbb{E}[r^2(z)] \tag{16}$$

Using Jensen's inequality and the convexity of the square function, we have

$$\begin{aligned}
\mathbb{E}[r^2(z|y)] &= \mathbb{E}_{y,z\sim p(y,z)}\left[\left(\mathbb{E}_{x\sim p(x|y,z)}[r(z|x)]\right)^2\right] \\
&= \mathbb{E}_{z\sim p(z)}\left[\mathbb{E}_{y\sim p(y|z)}\left[\left(\mathbb{E}_{x\sim p(x|y,z)}[r(z|x)]\right)^2\right]\right] \\
&\geq \mathbb{E}_{z\sim p(z)}\left[\left(\mathbb{E}_{y\sim p(y|z)}\left[\mathbb{E}_{x\sim p(x|y,z)}[r(z|x)]\right]\right)^2\right] \\
&= \mathbb{E}_{z\sim p(z)}\left[\left(\mathbb{E}_{x\sim p(x|z)}[r(z|x)]\right)^2\right] \\
&= \mathbb{E}[r^2(z)]
\end{aligned}$$

The equality holds iff $r(z|y) = \mathbb{E}_{x\sim p(x|y,z)}[r(z|x)]$ is constant w.r.t. y, for any z.

Using Jensen's inequality on $\mathbb{E}[r^2(z)]$, we have $\mathbb{E}[r^2(z)] = \mathbb{E}_{z\sim p(z)}\left[\left(\mathbb{E}_{x\sim p(x|z)}[r(z|x)]\right)^2\right] \leq \mathbb{E}_{z\sim p(z)}\left[\mathbb{E}_{x\sim p(x|z)}[r^2(z|x)]\right] = \mathbb{E}[r^2(z|x)]$, where the equality holds iff $r(z|x)$ is constant w.r.t. x for any z.

When $\mathbb{E}[r^2(z|y)] - \mathbb{E}[r^2(z)] > 0$, we have that the condition Eq. (16) is equivalent to $\forall r(z|x)$, $\beta \leq \frac{\mathbb{E}[r^2(z|x)]-\mathbb{E}[r^2(z)]}{\mathbb{E}[r^2(z|y)]-\mathbb{E}[r^2(z)]}$, i.e.

$$\beta \leq G[p(z|x)] \equiv \inf_{r(z|x)} \frac{\mathbb{E}[r^2(z|x)] - \mathbb{E}[r^2(z)]}{\mathbb{E}[r^2(z|y)] - \mathbb{E}[r^2(z)]} \tag{17}$$

where $r(z|y) = \mathbb{E}_{x\sim p(x|y,z)}[r(z|x)]$ and $r(z) = \mathbb{E}_{x\sim p(x|z)}[r(z|x)]$.

If $\mathbb{E}[r^2(z|y)] - \mathbb{E}[r^2(z)] = 0$, substituting into Eq. (16), we have

$$\beta \cdot 0 \leq \mathbb{E}[r^2(z|x)] - \mathbb{E}[r^2(z)] \tag{18}$$

which is always true due to that $\mathbb{E}[r^2(z|x)] \geq \mathbb{E}[r^2(z)]$, and will be a looser condition than Eq. (17) above. Above all, we have Eq. (17).

□

Empirical estimate of $G[p(z|x)]$ To empirically estimate $G[p(z|x)]$ from a minibatch of $\{(x_i, y_i)\}, i = 1, 2, ...N$ and the encoder $p(z|x)$, we can make the following Monte Carlo importance sampling estimation, where we use the samples $\{x_j\} \sim p(x)$ and also get samples of $\{z_i\} \sim p(z) = p(x)p(z|x)$, and have:

$$
\begin{aligned}
\mathbb{E}_{x,z\sim p(x,z)}[r^2(z|x)] =& \int dxdzp(x)p(z)\frac{p(x,z)}{p(x)p(z)}r^2(z|x) \\
\simeq& \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{p(x_j,z_i)}{p(x_j)p(z_i)}r^2(z_i|x_j) \\
\mathbb{E}_{z\sim p(z)}[r^2(z)] =& \mathbb{E}_{z\sim p(z)}\left[\left(\mathbb{E}_{x\sim p(x|z)}[r(z|x)]\right)^2\right] \\
\simeq& \frac{1}{N}\sum_{i=1}^{N}\left(\int dxp(x|z_i)r(z_i|x)\right)^2 \\
=& \frac{1}{N}\sum_{i=1}^{N}\left(\int dxp(x)\frac{p(z_i|x)}{p(z_i)}r(z_i|x)\right)^2 \\
\simeq& \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{N}\sum_{j=1}^{N}\frac{p(z_i|x_j)}{p(z_i)}r(z_i|x_j)\right)^2 \\
\simeq& \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{N}\sum_{j=1}^{N}\frac{p(z_i|x_j)}{\frac{1}{N}\sum_{k=1}^{N}p(z_i|x_k)}r(z_i|x_j)\right)^2 \\
=& \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\sum_{j=1}^{N}p(z_i|x_j)r(z_i|x_j)}{\sum_{j=1}^{N}p(z_i|x_j)}\right)^2
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}_{y,z\sim p(y,z)}[r^2(z|y)] =& \mathbb{E}_{y,z\sim p(y,z)}\left[\left(\mathbb{E}_{x\sim p(x|y,z)}[r(z|x)]\right)^2\right] \\
\simeq& \frac{1}{N}\sum_{i=1}^{N}\left(\int dxp(x|y_i,z_i)r(z_i|x)\right)^2 \\
=& \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{p(y_i,z_i)}\int dxp(y_i)p(x|y_i)p(z_i|x)r(z_i|x)\right)^2 \\
=& \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\int dxp(y_i)p(x|y_i)p(z_i|x)r(z_i|x)}{\int dxp(y_i)p(x|y_i)p(z_i|x)}\right)^2 \\
\simeq& \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\sum_{x_j\in\Omega_x(y_i)}p(z_i|x_j)r(z_i|x_j)}{\sum_{x_j\in\Omega_x(y_i)}p(z_i|x_j)}\right)^2 \\
=& \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\sum_{j=1}^{N}p(z_i|x_j)r(z_i|x_j)\mathbb{1}\left[y_i=y_j\right]}{\sum_{j=1}^{N}p(z_i|x_j)\mathbb{1}\left[y_i=y_j\right]}\right)^2
\end{aligned}
$$

Here $\Omega_x(y_i)$ denotes the set of x examples that has label of y_i, and $\mathbb{1}[\cdot]$ is an indicator function that takes value 1 if its argument is true, 0 otherwise.

The requirement of $\mathbb{E}_{z\sim p(z|x)}[r(z|x)] = 0$ yields

$$
0 = \mathbb{E}_{z\sim p(z|x)}[r(z|x)] = \int dzp(z)\frac{p(z|x)}{p(z)}r(z|x) \simeq \frac{1}{N}\sum_{i=1}^{N}\frac{p(z_i|x_j)}{p(z_i)}r(z_i|x_j) \tag{19}
$$

for any x_j.

Combining all terms, we have that the empirical $\hat{G}[p(z|x)]$ is given by

$$\hat{G}[p(z|x)] = \inf_{r(z|x)} \frac{\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{p(x_j,z_i)}{p(x_j)p(z_i)}r^2(z_i|x_j) - \sum_{i=1}^{N}\left(\frac{\sum_{j=1}^{N}p(z_i|x_j)r(z_i|x_j)}{\sum_{j=1}^{N}p(z_i|x_j)}\right)^2}{\sum_{i=1}^{N}\left(\frac{\sum_{j=1}^{N}p(z_i|x_j)r(z_i|x_j)\mathbb{1}[y_i=y_j]}{\sum_{j=1}^{N}p(z_i|x_j)\mathbb{1}[y_i=y_j]}\right)^2 - \sum_{i=1}^{N}\left(\frac{\sum_{j=1}^{N}p(z_i|x_j)r(z_i|x_j)}{\sum_{j=1}^{N}p(z_i|x_j)}\right)^2} \tag{20}$$

where $\{z_i\} \sim p(z)$ and $\{x_i\} \sim p(x)$. It is also possible to use different distributions for importance sampling, which will results in different formulas for empirical estimation of $G[p(z|x)]$.

D $G_\Theta[p_{\boldsymbol{\theta}}(z|x)]$ FOR PARAMETERIZED DISTRIBUTION $p_{\boldsymbol{\theta}}(z|x)$

Proof. For the parameterized[5] $p_{\boldsymbol{\theta}}(z|x)$, after $\boldsymbol{\theta}' \leftarrow \boldsymbol{\theta} + \Delta\boldsymbol{\theta}$ where $|\Delta\boldsymbol{\theta}| \ll |\boldsymbol{\theta}|$, we have that the distribution changes from $p_{\boldsymbol{\theta}}(z|x)$ to $p_{\boldsymbol{\theta}+\Delta\boldsymbol{\theta}}(z|x)$, and thus the relative perturbation on $p_{\boldsymbol{\theta}}(z|x)$ is

$$\begin{aligned}
\epsilon \cdot r(z|x) &= \frac{p_{\boldsymbol{\theta}+\Delta\boldsymbol{\theta}}(z|x) - p_{\boldsymbol{\theta}}(z|x)}{p_{\boldsymbol{\theta}}(z|x)} \\
&= \frac{1}{p_{\boldsymbol{\theta}}(z|x)}\left(p_{\boldsymbol{\theta}}(z|x) + \Delta\boldsymbol{\theta}^T\frac{\partial p_{\boldsymbol{\theta}}(z|x)}{\partial\boldsymbol{\theta}} + \frac{1}{2}\Delta\boldsymbol{\theta}^T\frac{\partial^2 p_{\boldsymbol{\theta}}(z|x)}{\partial\boldsymbol{\theta}^2}\Delta\boldsymbol{\theta} + O(|\Delta\boldsymbol{\theta}|^3) - p_{\boldsymbol{\theta}}(z|x)\right) \\
&\simeq \Delta\boldsymbol{\theta}^T\frac{\partial}{\partial\boldsymbol{\theta}}\log p_{\boldsymbol{\theta}}(z|x) + \frac{1}{2}\Delta\boldsymbol{\theta}^T\frac{1}{p_{\boldsymbol{\theta}}(z|x)}\frac{\partial^2 p_{\boldsymbol{\theta}}(z|x)}{\partial\boldsymbol{\theta}^2}\Delta\boldsymbol{\theta} + O(|\Delta\boldsymbol{\theta}|^3)
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\epsilon \cdot r(z|y) &= \Delta\boldsymbol{\theta}^T\frac{\partial}{\partial\boldsymbol{\theta}}\log p_{\boldsymbol{\theta}}(z|y) + \frac{1}{2}\Delta\boldsymbol{\theta}^T\frac{1}{p_{\boldsymbol{\theta}}(z|y)}\frac{\partial^2 p_{\boldsymbol{\theta}}(z|y)}{\partial\boldsymbol{\theta}^2}\Delta\boldsymbol{\theta} + O(|\Delta\boldsymbol{\theta}|^3) \\
\epsilon \cdot r(z) &= \Delta\boldsymbol{\theta}^T\frac{\partial}{\partial\boldsymbol{\theta}}\log p_{\boldsymbol{\theta}}(z) + \frac{1}{2}\Delta\boldsymbol{\theta}^T\frac{1}{p_{\boldsymbol{\theta}}(z)}\frac{\partial^2 p_{\boldsymbol{\theta}}(z)}{\partial\boldsymbol{\theta}^2}\Delta\boldsymbol{\theta} + O(|\Delta\boldsymbol{\theta}|^3)
\end{aligned}$$

Substituting the above expressions into the expansion of $\mathrm{IB}_\beta[p(z|x)]$ in Eq. (12), and preserving to the second order of $|\Delta\boldsymbol{\theta}|^2$, we have

[5] In this paper, $\boldsymbol{\theta} = (\theta_1, \theta_2, ...\theta_k)^T$ and $\frac{\partial p_{\boldsymbol{\theta}}(z|x)}{\partial\boldsymbol{\theta}} = \left(\frac{\partial p_{\boldsymbol{\theta}}(z|x)}{\partial\theta_1}, \frac{\partial p_{\boldsymbol{\theta}}(z|x)}{\partial\theta_2}, ... \frac{\partial p_{\boldsymbol{\theta}}(z|x)}{\partial\theta_k}\right)^T$ are all column vectors. $\frac{\partial^2 p_{\boldsymbol{\theta}}(z|x)}{\partial\boldsymbol{\theta}^2}$ is a $k \times k$ matrix with (i,j) element of $\frac{\partial^2 p_{\boldsymbol{\theta}}(z|x)}{\partial\theta_i\partial\theta_j}$.

$$
\begin{aligned}
&\mathrm{IB}_\beta[p_{\boldsymbol{\theta}}(z|x)(1+\epsilon\cdot r(z|x))] \\
=&\mathrm{IB}_\beta[p_{\boldsymbol{\theta}}(z|x)] + \epsilon\cdot\left(\mathbb{E}_{x,z\sim p_{\boldsymbol{\theta}}(x,z)}\left[r(z|x)\log\frac{p_{\boldsymbol{\theta}}(z|x)}{p_{\boldsymbol{\theta}}(z)}\right] - \beta\cdot\mathbb{E}_{y,z\sim p_{\boldsymbol{\theta}}(y,z)}\left[r(z|y)\log\frac{p_{\boldsymbol{\theta}}(z|y)}{p_{\boldsymbol{\theta}}(z)}\right]\right) \\
&+\frac{\epsilon^2}{1\cdot 2}\left\{\left(\mathbb{E}_{x,z\sim p_{\boldsymbol{\theta}}(x,z)}[r^2(z|x)] - \mathbb{E}_{z\sim p_{\boldsymbol{\theta}}(z)}[r^2(z)]\right) - \beta\cdot\left(\mathbb{E}_{y,z\sim p_{\boldsymbol{\theta}}(y,z)}[r^2(z|y)] - \mathbb{E}_{z\sim p_{\boldsymbol{\theta}}(z)}[r^2(z)]\right)\right\} \\
=&\mathrm{IB}_\beta[p_{\boldsymbol{\theta}}(z|x)] + \mathbb{E}_{x,z\sim p_{\boldsymbol{\theta}}(x,z)}\left[\left(\Delta\boldsymbol{\theta}^T\frac{\partial}{\partial\boldsymbol{\theta}}\log p_{\boldsymbol{\theta}}(z|x) + \frac{1}{2}\Delta\boldsymbol{\theta}^T\frac{1}{p_{\boldsymbol{\theta}}(z|x)}\frac{\partial^2 p_{\boldsymbol{\theta}}(z|x)}{\partial\boldsymbol{\theta}^2}\Delta\boldsymbol{\theta}\right)\log\frac{p_{\boldsymbol{\theta}}(z|x)}{p_{\boldsymbol{\theta}}(z)}\right] \\
&-\beta\cdot\mathbb{E}_{y,z\sim p_{\boldsymbol{\theta}}(y,z)}\left[\left(\Delta\boldsymbol{\theta}^T\frac{\partial}{\partial\boldsymbol{\theta}}\log p_{\boldsymbol{\theta}}(z|y) + \frac{1}{2}\Delta\boldsymbol{\theta}^T\frac{1}{p_{\boldsymbol{\theta}}(z|y)}\frac{\partial^2 p_{\boldsymbol{\theta}}(z|y)}{\partial\boldsymbol{\theta}^2}\Delta\boldsymbol{\theta}\right)\log\frac{p_{\boldsymbol{\theta}}(z|y)}{p_{\boldsymbol{\theta}}(z)}\right] \\
&+\frac{1}{2}\left(\mathbb{E}_{x,z\sim p_{\boldsymbol{\theta}}(x,z)}\left[\left(\Delta\boldsymbol{\theta}^T\frac{\partial}{\partial\boldsymbol{\theta}}\log p_{\boldsymbol{\theta}}(z|x)\right)^2\right] - \mathbb{E}_{z\sim p_{\boldsymbol{\theta}}(z)}\left[\left(\Delta\boldsymbol{\theta}^T\frac{\partial}{\partial\boldsymbol{\theta}}\log p_{\boldsymbol{\theta}}(z)\right)^2\right]\right) \\
&-\frac{\beta}{2}\left(\mathbb{E}_{y,z\sim p_{\boldsymbol{\theta}}(y,z)}\left[\left(\Delta\boldsymbol{\theta}^T\frac{\partial}{\partial\boldsymbol{\theta}}\log p_{\boldsymbol{\theta}}(z|y)\right)^2\right] - \mathbb{E}_{z\sim p_{\boldsymbol{\theta}}(z)}\left[\left(\Delta\boldsymbol{\theta}^T\frac{\partial}{\partial\boldsymbol{\theta}}\log p_{\boldsymbol{\theta}}(z)\right)^2\right]\right) \\
=&\mathrm{IB}_\beta[p_{\boldsymbol{\theta}}(z|x)] + \Delta\boldsymbol{\theta}^T\left\{\mathbb{E}_{x,z\sim p_{\boldsymbol{\theta}}(x,z)}\left[\log\frac{p_{\boldsymbol{\theta}}(z|x)}{p_{\boldsymbol{\theta}}(z)}\frac{\partial}{\partial\boldsymbol{\theta}}\log p_{\boldsymbol{\theta}}(z|x)\right] - \beta\cdot\mathbb{E}_{x,z\sim p_{\boldsymbol{\theta}}(x,z)}\left[\log\frac{p_{\boldsymbol{\theta}}(z|x)}{p_{\boldsymbol{\theta}}(z)}\frac{\partial}{\partial\boldsymbol{\theta}}\log p_{\boldsymbol{\theta}}(z|x)\right]\right\} \\
&+\frac{1}{2}\Delta\boldsymbol{\theta}^T\left\{\left(\mathcal{I}_{Z|X}(\boldsymbol{\theta}) - \mathcal{I}_Z(\boldsymbol{\theta})\right) - \beta\left(\mathcal{I}_{Z|X}(\boldsymbol{\theta}) - \mathcal{I}_Z(\boldsymbol{\theta})\right)\right\}\Delta\boldsymbol{\theta}
\end{aligned}
$$

In the last equality we have used $\mathbb{E}_{x,z\sim p_{\boldsymbol{\theta}}(x,z)}[\frac{1}{p_{\boldsymbol{\theta}}(z|x)}\frac{\partial^2 p_{\boldsymbol{\theta}}(z|x)}{\partial\boldsymbol{\theta}^2}] = \int dxp(x)\frac{\partial^2}{\partial\boldsymbol{\theta}^2}\int dzp_{\boldsymbol{\theta}}(z|x) = \int dxp(x)\frac{\partial^2}{\partial\boldsymbol{\theta}^2}1 = \mathbf{0}$, and similarly $\mathbb{E}_{y,z\sim p_{\boldsymbol{\theta}}(y,z)}[\frac{1}{p_{\boldsymbol{\theta}}(z|y)}\frac{\partial^2 p_{\boldsymbol{\theta}}(z|y)}{\partial\boldsymbol{\theta}^2}] = \mathbf{0}$. In other words, the $|\Delta\boldsymbol{\theta}|^2$ terms in the first-order variation $\delta\mathrm{IB}_\beta[p_{\boldsymbol{\theta}}(z|x)]$ vanish, and the remaining $|\Delta\boldsymbol{\theta}|^2$ are all in $\delta^2\mathrm{IB}_\beta[p_{\boldsymbol{\theta}}(z|x)]$. Also in the last expression, $\mathcal{I}_Z(\boldsymbol{\theta}) \equiv \int dzp_{\boldsymbol{\theta}}(z)\left(\frac{\partial\log p_{\boldsymbol{\theta}}(z)}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial\log p_{\boldsymbol{\theta}}(z)}{\partial\boldsymbol{\theta}}\right)^T$ is the Fisher information matrix of $\boldsymbol{\theta}$ for Z, $\mathcal{I}_{Z|X}(\boldsymbol{\theta}) \equiv \int dxdzp(x)p_{\boldsymbol{\theta}}(z|x)\left(\frac{\partial\log p_{\boldsymbol{\theta}}(z|x)}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial\log p_{\boldsymbol{\theta}}(z|x)}{\partial\boldsymbol{\theta}}\right)^T$, $\mathcal{I}_{Z|Y}(\boldsymbol{\theta}) \equiv \int dydzp(y)p_{\boldsymbol{\theta}}(z|y)\left(\frac{\partial\log p_{\boldsymbol{\theta}}(z|y)}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial\log p_{\boldsymbol{\theta}}(z|y)}{\partial\boldsymbol{\theta}}\right)^T$ are the conditional Fisher information matrix (Zegers, 2015) of $\boldsymbol{\theta}$ for Z conditioned on X and Y, respectively.

Let us look at

$$
\delta^2\mathrm{IB}_\beta[p_{\boldsymbol{\theta}}(z|x)] = \frac{1}{2}\Delta\boldsymbol{\theta}^T\left\{\left(\mathcal{I}_{Z|X}(\boldsymbol{\theta}) - \mathcal{I}_Z(\boldsymbol{\theta})\right) - \beta\left(\mathcal{I}_{Z|X}(\boldsymbol{\theta}) - \mathcal{I}_Z(\boldsymbol{\theta})\right)\right\}\Delta\boldsymbol{\theta} \tag{21}
$$

It is a special case of Eq. (14) with $\epsilon\cdot r(z|x) = \Delta\boldsymbol{\theta}^T\frac{\partial}{\partial\boldsymbol{\theta}}\log p_{\boldsymbol{\theta}}(z|x)$. Therefore, The inequalities due to Jensen still hold: $\epsilon^2\left(\mathbb{E}[r^2(z|x)] - \mathbb{E}[r^2(z)]\right) = \Delta\boldsymbol{\theta}^T\left(\mathcal{I}_{Z|X}(\boldsymbol{\theta}) - \mathcal{I}_Z(\boldsymbol{\theta})\right)\Delta\boldsymbol{\theta} \geq 0$, $\epsilon^2\left(\mathbb{E}[r^2(z|y)] - \mathbb{E}[r^2(z)]\right) = \Delta\boldsymbol{\theta}^T\left(\mathcal{I}_{Z|Y}(\boldsymbol{\theta}) - \mathcal{I}_Z(\boldsymbol{\theta})\right)\Delta\boldsymbol{\theta} \geq 0$. If $\Delta\boldsymbol{\theta}^T\left(\mathcal{I}_{Z|Y}(\boldsymbol{\theta}) - \mathcal{I}_Z(\boldsymbol{\theta})\right)\Delta\boldsymbol{\theta} > 0$, then the condition of $\forall\Delta\boldsymbol{\theta}$, $\delta^2\mathrm{IB}_\beta[p_{\boldsymbol{\theta}}(z|x)] \geq 0$ is equivalent to $\forall\Delta\boldsymbol{\theta}$,

$$
\beta \leq \frac{\Delta\boldsymbol{\theta}^T\left(\mathcal{I}_{Z|X}(\boldsymbol{\theta}) - \mathcal{I}_Z(\boldsymbol{\theta})\right)\Delta\boldsymbol{\theta}}{\Delta\boldsymbol{\theta}^T\left(\mathcal{I}_{Z|Y}(\boldsymbol{\theta}) - \mathcal{I}_Z(\boldsymbol{\theta})\right)\Delta\boldsymbol{\theta}}
$$

i.e.

$$
\beta \leq G_\Theta[p_{\boldsymbol{\theta}}(z|x)] \equiv \inf_{\boldsymbol{\theta}}\frac{\Delta\boldsymbol{\theta}^T\left(\mathcal{I}_{Z|X}(\boldsymbol{\theta}) - \mathcal{I}_Z(\boldsymbol{\theta})\right)\Delta\boldsymbol{\theta}}{\Delta\boldsymbol{\theta}^T\left(\mathcal{I}_{Z|Y}(\boldsymbol{\theta}) - \mathcal{I}_Z(\boldsymbol{\theta})\right)\Delta\boldsymbol{\theta}} \tag{22}
$$

If $\Delta\boldsymbol{\theta}^T\left(\mathcal{I}_{Z|Y}(\boldsymbol{\theta}) - \mathcal{I}_Z(\boldsymbol{\theta})\right)\Delta\boldsymbol{\theta} = 0$, we have that Eq. (21) always holds, which is a looser condition than Eq. (22). Above all, we have that the condition of $\forall\Delta\boldsymbol{\theta}$, $\delta^2\mathrm{IB}_\beta[p_{\boldsymbol{\theta}}(z|x)]$ is equivalent to $\beta \leq G_\Theta[p_{\boldsymbol{\theta}}(z|x)]$.

Moreover, $(G_\Theta[p_{\boldsymbol{\theta}}(z|x)])^{-1}$ given by Eq. (22) has the format of a generalized Rayleigh quotient $R(A, B; x) \equiv \frac{\Delta\boldsymbol{\theta}^T A \Delta\boldsymbol{\theta}}{\Delta\boldsymbol{\theta}^T B \Delta\boldsymbol{\theta}}$ where $A = \mathcal{I}_{Z|Y}(\boldsymbol{\theta}) - \mathcal{I}_Z(\boldsymbol{\theta})$ and $B = \mathcal{I}_{Z|X}(\boldsymbol{\theta}) - \mathcal{I}_Z(\boldsymbol{\theta})$ are both Hermitian matrices[6], which can be reduced to Rayleigh quotient $R(D, C^T\Delta\boldsymbol{\theta}) = \frac{(C^T\Delta\boldsymbol{\theta})^T D (C^T\Delta\boldsymbol{\theta})}{(C^T\Delta\boldsymbol{\theta})^T (C^T\Delta\boldsymbol{\theta})}$, with the transformation $D = C^{-1}A(C^T)^{-1}$ where CC^T is the Cholesky decomposition of $B = \mathcal{I}_{Z|X}(\boldsymbol{\theta}) - \mathcal{I}_Z(\boldsymbol{\theta})$. Moreover, we have that when $G_\Theta[p_{\boldsymbol{\theta}}(z|x)]$ attains its minimum value, the Reyleigh quotient $R(D, C^T\Delta\boldsymbol{\theta})$ attains its maximum value of λ_{max} with $C^T\Delta\boldsymbol{\theta} = v_{\text{max}}$, i.e. $\Delta\boldsymbol{\theta} = (C^T)^{-1}v_{\text{max}}$, where λ_{max} is the largest eigenvalue of D and v_{max} the corresponding eigenvector.

□

E Proof of Theorem 1

Proof. Using the condition in Definition 1 and Lemma 2.1, we have that phase transition corresponds to

$$\begin{cases} \forall r(z|x), \big(\mathbb{E}_\beta[r^2(z|x)] - \mathbb{E}_\beta[r^2(z)]\big) - \beta \cdot \big(\mathbb{E}_\beta[r^2(z|y)] - \mathbb{E}_\beta[r^2(z)]\big) \geq 0 \\ \exists r(z|x), \big(\mathbb{E}_\beta[r^2(z|x)] - \mathbb{E}_\beta[r^2(z)]\big) - (\beta+\eta) \cdot \big(\mathbb{E}_\beta[r^2(z|y)] - \mathbb{E}_\beta[r^2(z)]\big) < 0 \end{cases}$$

for any $0 < \eta \ll \beta$. Here $\mathbb{E}_\beta[\cdot]$ denotes taking expectation w.r.t. the optimal solution $p^*_\beta(x, y, z) = p(x, y)p^*_\beta(z|x)$ at β. Using the boundedness of $r(z|x)$ s.t. $|r(z|x)| < M$ for some constant M, we have that all expectation terms $\mathbb{E}_\beta[\cdot]$ above are less than M^2. Let $\eta \to 0^+$, we have that the gap between the L.H.S. of the two formulas are less than ηM^2. As $\eta \to 0^+$, the gap should also tend to 0^+. Thus we have $\inf_{r(z|x)} \big[\big(\mathbb{E}_\beta[r^2(z|x)] - \mathbb{E}_\beta[r^2(z)]\big) - \beta \cdot \big(\mathbb{E}_\beta[r^2(z|y)] - \mathbb{E}_\beta[r^2(z)]\big)\big] = 0$. After simple manipulation we have $G[p^*_\beta(z|x)] = \beta$. □

F Invariance of $\mathcal{G}[r(z|x); p(z|x)]$ to addition of a global representation

Here we prove the following lemma:

Lemma 2.2. *$\mathcal{G}[r(z|x); p(z|x)]$ defined in Lemma 0.1 is invariant to the transformation $r'(z|x) \leftarrow r(z|x) + s(z)$.*

Proof. When we $r(z|x)$ is shifted by a global transformation $r'(z|x) \leftarrow r(z|x) + s(z)$, we have $r'(z) \leftarrow \mathbb{E}_{x\sim p(x|z)}[r(z|x) + s(z)] = \mathbb{E}_{x\sim p(x|z)}[r(z|x)] + s(z)\mathbb{E}_{x\sim p(x|z)}[1] = r(z) + s(z)$, and similarly $r'(z|y) \leftarrow r(z|y) + s(z)$.

The numerator of $\mathcal{G}[r(z|x); p(z|x)]$ is then

$$\begin{aligned}
&\mathbb{E}_{x,z\sim p(x,z)}\left[(r'(z|x))^2\right] - \mathbb{E}_{z\sim p(z)}\left[(r'(z))^2\right] \\
=&\mathbb{E}_{x,z\sim p(x,z)}\left[(r(z|x) + s(z))^2\right] - \mathbb{E}_{z\sim p(z)}\left[(r(z) + s(z))^2\right] \\
=&\left(\mathbb{E}_{x,z\sim p(x,z)}\left[r^2(z|x)\right] + 2\mathbb{E}_{x,z\sim p(x,z)}\left[r(z|x)s(z)\right] + \mathbb{E}_{x,z\sim p(x,z)}\left[s^2(z)\right]\right) \\
&- \left(\mathbb{E}_{z\sim p(z)}\left[r^2(z)\right] + 2\mathbb{E}_{z\sim p(z)}\left[r(z)s(z)\right] + \mathbb{E}_{z\sim p(z)}\left[s^2(z)\right]\right) \\
=&\left(\mathbb{E}_{x,z\sim p(x,z)}\left[r^2(z|x)\right] + 2\mathbb{E}_{z\sim p(z)}\left[s(z)\mathbb{E}_{x\sim p(x|z)}\left[r(z|x)\right]\right] + \mathbb{E}_{z\sim p(z)}\left[s^2(z)\right]\right) \\
&- \left(\mathbb{E}_{z\sim p(z)}\left[r^2(z)\right] + 2\mathbb{E}_{z\sim p(z)}\left[r(z)s(z)\right] + \mathbb{E}_{z\sim p(z)}\left[s^2(z)\right]\right) \\
=&\left(\mathbb{E}_{x,z\sim p(x,z)}\left[r^2(z|x)\right] + 2\mathbb{E}_{z\sim p(z)}\left[r(z)s(z)\right] + \mathbb{E}_{z\sim p(z)}\left[s^2(z)\right]\right) \\
&- \left(\mathbb{E}_{z\sim p(z)}\left[r^2(z)\right] + 2\mathbb{E}_{z\sim p(z)}\left[r(z)s(z)\right] + \mathbb{E}_{z\sim p(z)}\left[s^2(z)\right]\right) \\
=&\mathbb{E}_{x,z\sim p(x,z)}\left[r^2(z|x)\right] - \mathbb{E}_{z\sim p(z)}\left[r^2(z)\right]
\end{aligned}$$

[6]Here all the Fisher information matrices are real symmetric, thus Hermitian

Symmetrically, we have

$$\mathbb{E}_{y,z\sim p(y,z)}\left[\left(r'(z|y)\right)^2\right]-\mathbb{E}_{z\sim p(z)}\left[\left(r'(z)\right)^2\right]=\mathbb{E}_{y,z\sim p(y,z)}\left[r^2(z|y)\right]-\mathbb{E}_{z\sim p(z)}\left[r^2(z)\right]$$

Therefore, $\mathcal{G}[r(z|x);p(z|x)]=\frac{\mathbb{E}_{x,z\sim p(x,z)}\left[r^2(z|x)\right]-\mathbb{E}_{z\sim p(z)}\left[r^2(z)\right]}{\mathbb{E}_{y,z\sim p(y,z)}\left[r^2(z|y)\right]-\mathbb{E}_{z\sim p(z)}\left[r^2(z)\right]}$ is invariant to $r'(z|x)\leftarrow r(z|x)+s(z)$. □

G PROOF OF THEOREM 2

Proof. Since $\forall f(x,z)$, it can be decomposed as $f(x,z)=h(x,z)+s(z)$ such that $\mathbb{E}_{p(z|x)}[h(x,z)]=0$, using Lemma 2.2, we have that we can drop the requirement of $\mathbb{E}_{z\sim p(z|x)}[r(z|x)]=0$ in the definition of $G[p(z|x)]$ in Lemma 0.1. This is because any arbitrary $r(z|x)$ can be decomposed into $r(z|x)=r_1(z|x)+s(z)$ such that $\mathbb{E}_{p(z|x)}[r_1(z|x)]=0$, and using lemma 2.2 we have

$$\inf_{r(z|x)}\mathcal{G}[r(z|x);p(z|x)]=\inf_{r_1(z|x):\mathbb{E}_{z\sim p(z|x)}[r_1(z|x)]=0}\mathcal{G}[r_1(z|x);p(z|x)]=G[p(z|x)]$$

where $r(z|x)$ doesn't have constraints.

After dropping the constraint of $\mathbb{E}_{z\sim p(z|x)}[r(z|x)]=0$, again using Lemma 2.2, we can let $r(z)=\mathbb{E}_{x\sim p(x|z)}[r(z|x)]=0$ (since we can perform the transformation $r'(z|x)\leftarrow r(z|x)-r(z)$, so that the new $r'(z)\equiv 0$). Now we get a simpler formula for $G[p(z|x)]$, as follows:

$$G[p(z|x)]=\inf_{r(z|x)}\frac{\mathbb{E}_{x,z\sim p(x,z)}[r^2(z|x)]}{\mathbb{E}_{y,z\sim p(y,z)}\left[\left(\mathbb{E}_{x\sim p(x|y,z)}[r(z|x)]\right)^2\right]} \tag{23}$$

with $\mathbb{E}_{x\sim p(x|z)}[r(z|x)]=0$.

From Eq. (23), we can further require that $\mathbb{E}_{x,z\sim p(x,z)}[r^2(z|x)]=1$. Define

$$\rho_s^2(X,Y;Z)\equiv\sup_{f(X,Z)}\mathbb{E}[(\mathbb{E}[f(X,Z)|Y,Z])^2]=\sup_{f(x,z)}\mathbb{E}_{y,z\sim p(y,z)}\left[\left(\mathbb{E}_{x\sim p(x|y,z)}[f(x,z)]\right)^2\right] \tag{24}$$

such that[7] $\mathbb{E}_{x\sim p(x|z)}[f(x,z)]\equiv 0$ and $\mathbb{E}[f^2(X,Z)]=1$. Comparing with Eq. (23), it immediately follows that

$$G[p(z|x)]=\frac{1}{\rho_s^2(X,Y;Z)}$$

(i) We only have to prove that $\rho_s(X,Y;Z)=\rho_r(X,Y;Z)$, where $\rho_r(X,Y;Z)$ is defined in Definition 2.

We have

$$\begin{aligned}
&\mathbb{E}[f(X,Z)g(Y,Z)]\\
=&\int dxdydzp(x,y,z)f(x,z)g(y,z)\\
=&\int dydzp(y,z)g(y,z)\int dxp(x|y,z)f(x,z)\\
\equiv&\int dydzp(y,z)g(y,z)F(y,z)\\
\leq&\sqrt{\int dydzp(y,z)g^2(y,z)}\cdot\sqrt{\int dydzp(y,z)F^2(y,z)}
\end{aligned}$$

[7] In the definition of $\rho_r(X,Y;Z)$, we have used an equivalent format $f(x,z)$ instead of $r(z|x)$.

where $F(y,z) \equiv \int dx p(x|y,z) f(x,z)$. We have used Cauchy-Schwarz inequality, where the equality holds when $g(y,z) = \alpha F(y,z)$ for some α. Since $\mathbb{E}[g^2(y,z)] = 1$, we have $\alpha^2 \mathbb{E}[F^2(y,z)] = 1$. Taking the supremum of $(\mathbb{E}[f(X,Z)g(Y,Z)])^2$ w.r.t. f and g, we have

$$\begin{aligned}
\rho_r^2(X,Y;Z) &= \sup_{f(X,Z),g(Y,Z)} (\mathbb{E}[f(X,Z)g(Y,Z)])^2 \\
&= \sup_{f(x,z),g(y,z)} \int dydzp(y,z)g^2(y,z) \cdot \int dydzp(y,z)F^2(y,z) \\
&= \sup_{f(x,z)} \int dydzp(y,z)F^2(y,z) \\
&= \sup_{f(x,z)} \int dydzp(y,z) \left(\int dxp(x|y,z)f(x,z) \right)^2 \\
&= \sup_{f(X,Z)} \mathbb{E}[(\mathbb{E}[f(X,Z)|Y,Z])^2] \\
&\equiv \rho_s^2(X,Y;Z)
\end{aligned}$$

By definition both $\rho_r(X,Y;Z)$ and $\rho_s(X,Y;Z)$ take non-negative values. Therefore,

$$\rho_s(X,Y;Z) = \rho_r(X,Y;Z) \tag{25}$$

(ii) Using the definition of $\rho_r(X,Y;Z)$, we have

$$\begin{aligned}
&\rho_r^2(X,Y;Z) \\
&\equiv \sup_{f(x,z)} \int dydzp(y,z) \left(\int dxp(x|y,z)f(x,z) \right)^2 \\
&= \sup_{f(x,z)} \int dzp(z) \int dyp(y|z) \left(\int dxp(x|y,z)f(x,z) \right)^2 \\
&\equiv \sup_{f(x,z)} \int dzp(z) W[f(x,z)]
\end{aligned}$$

where $W[f(x,z)] \equiv \int dyp(y|z) \left(\int dxp(x|y,z)f(x,z) \right)^2$, and $f(x,z)$ satisfies $\mathbb{E}_{x\sim p(x|z)}[f(x,z)] = 0$, $\mathbb{E}_{x,z\sim p(x,z)}[f^2(x,z)] = 1$.

Denote $c(z) \equiv p(z)\mathbb{E}_{x\sim p(x|z)}[f^2(x,z)]$, we have $\int c(z)dz = \mathbb{E}_{x,z\sim p(x,z)}[f^2(x,z)] = 1$. Then the supremum $\rho_r^2(X,Y;Z) = \sup_{f(x,z)} \int dzp(z)W[f(x,z)]$ is equivalent to the following two-stage supremum:

$$\rho_r^2(X,Y;Z) = \sup_{c(z):\int c(z)dz=1} \int dzp(z) \sup_{f(x,z)} W[f(x,z)] \tag{26}$$

with $\mathbb{E}_{x\sim p(x|z)}[f^2(x,z)] = \frac{c(z)}{p(z)}$ and $\mathbb{E}_{x\sim p(x|z)}[f(x,z)] = 0$. We can think of the inner supremum $\sup_{f(x,z)} W[f(x,z)]$ as only w.r.t. x, for some given z.

Now let's consider another supremum:

$$\sup_{h(x)} \int dyp(y|z) \left(\int dxp(x|y,z)h(x) \right)^2 \tag{27}$$

with $\mathbb{E}_{p(x|z)}[h(x)] = 0$ and $\mathbb{E}_{p(x|z)}[h^2(x)] = 1$. Using similar technique in (ii), it is easy to prove that it equals $\rho_m^2(X,Y|Z)$ as defined in Theorem 2.

Comparing Eq. (27) and the supremum:

$$\sup_{f(x,z)} W[f(x,z)]$$

with $\mathbb{E}_{x\sim p(x|z)}[f^2(x,z)] = \frac{c(z)}{p(z)}$ and $\mathbb{E}_{x\sim p(x|z)}[f(x,z)] = 0$, we see that the only difference is that in the latter $\mathbb{E}_{x\sim p(x|z)}[f^2(x,z)]$ equals $\frac{c(z)}{p(z)}$ instead of 1. Since $W[f(x,z)]$ is a quadratic functional of $f(x,z)$, we have

$$\sup_{f(x,z)} W[f(x,z)] = \frac{c(z)}{p(z)}\rho_m^2(X,Y|Z)$$

Therefore,

$$\begin{aligned}
\rho_r(X,Y;Z) &= \sup_{c(z):\int c(z)dz=1} \int dz\, p(z) \sup_{f(x,z)} W[f(x,z)] \\
&= \sup_{c(z):\int c(z)dz=1} \int dz\, p(z) \frac{c(z)}{p(z)} \rho_m^2(X,Y|Z) \\
&= \sup_{c(z):\int c(z)dz=1} \int dz\, c(z) \rho_m^2(X,Y|Z=z) \\
&= \sup_{Z} \rho_m^2(X,Y|Z)
\end{aligned}$$

where in the last equality we have let $c(z)$ have "mass" only on the place where $\rho_m^2(X,Y|Z=z)$ attains supremum w.r.t. z.

(iii) When Z is a continuous variable, let $f(x,z) = f_X(x)\sqrt{\frac{\delta(z-z_0)}{p(z)}}$, where $\delta(\cdot)$ is the Dirac-delta function, z_0 is a parameter, $f_X(x)$ satisfies $\mathbb{E}_{X\sim p(X|Z)}[f_X(x)] = 0$ and $\mathbb{E}_{X\sim p(X|Z)}[f_X^2(x)] = 1$, we have

$$\begin{aligned}
\mathbb{E}_{x\sim p(x|z)}[f(x,z)] &= \int p(x|z) f(x,z) dx \\
&= \sqrt{\frac{\delta(z-z_0)}{p(z)}} \int p(x|z) f_X(x) dx \\
&= \sqrt{\frac{\delta(z-z_0)}{p(z)}} \cdot 0 \\
&= 0
\end{aligned}$$

And

$$\begin{aligned}
\mathbb{E}[f^2(X,Z)] &= \int p(x,z) f^2(x,z) dx dz \\
&= \int p(x,z) f_X^2(x) \frac{\delta(z-z_0)}{p(z)} dx dz \\
&= \int dz \delta(z-z_0) \int dx p(x|z) f_X^2(x) dx \\
&= \int dz \delta(z-z_0) \cdot 1 \\
&= 1
\end{aligned}$$

Therefore, such constructed $f(x,z) = f_X(x)\sqrt{\frac{\delta(z-z_0)}{p(z)}}$ with $\mathbb{E}_{X\sim p(X|Z)}[f_X(x)] = 0$ and $\mathbb{E}_{X\sim p(X|Z)}[f_X^2(x)] = 1$ satisfies the requirement for $\rho_s(X,Y;Z)$ (which equals $\rho_r(X,Y;Z)$ by Eq. 25).

Substituting in the special form of $f(x,z)$ into the expression of $\rho_s(X,Y;Z)$ in Eq. (24), we have

$$
\begin{aligned}
&\sup_{f(x,z)} \int dz p(z) \int dy p(y|z) \left(\int dx p(x|y,z) f(x,z) \right)^2 \\
=& \sup_{f_X(x),z_0} \int dz p(z) \int dy p(y|z) \left(\int dx p(x|y,z) f_X(x) \sqrt{\frac{\delta(z-z_0)}{p(z)}} \right)^2 \\
=& \sup_{z_0} \int dz p(z) \frac{\delta(z-z_0)}{p(z)} \sup_{f_X(x)} \int dy p(y|z) \left(\int dx p(x|y,z) f_X(x) \right)^2 \\
=& \sup_{z_0} \int dz \delta(z-z_0) \sup_{f_X(X)} \mathbb{E}[(\mathbb{E}[f_X(X)|Y,Z=z])^2|Z=z] \\
=& \sup_{z_0} \int dz \delta(z-z_0) \rho_m^2(X,Y|Z=z) \\
=& \sup_{z_0} \rho_m^2(X,Y|Z=z_0) \\
=& \sup_{Z} \rho_m^2(X,Y|Z)
\end{aligned}
$$

We can identify $\sup_{f_X(X)} \mathbb{E}[(\mathbb{E}[f_X(X)|Y,Z=z])^2|Z=z]$ with $\rho_m^2(X,Y|Z=z)$ because $f_X(x)$ satisfies the requirement for conditional maximum correlation that $\mathbb{E}_{p(x|z)}[f_X(x)]=0$ and $\mathbb{E}_{p(x|z)}[f_X^2(x)]=1$, for any z, and using the same technique in (i), it is straightforward to prove that $\sup_{f_X(X)} \mathbb{E}[(\mathbb{E}[f_X(X)|Y,Z=z])^2|Z=z]$ with the condition $\mathbb{E}_{p(x|z)}[f_X(x)]=0$ and $\mathbb{E}_{p(x|z)}[f_X^2(x)]=1$ for any z, equals the conditional maximum correlation as defined in Definition 2.

Since the conditional maximum correlation can be viewed as the maximum correlation between X and Y, where $X,Y \sim p(X,Y|Z)$, using the equality of $(\beta_0[h(x)])^{-1} = \rho_m^2(X;Y)$ (Eq. 7 in Wu et al. (2019)), we can identify the $h(x)$ in $\beta_0[h(x)]$ with the $f_X(X)$ here, and an optimal $f_X^*(X)$ that maximizes $\rho_m^2(X,Y|Z)$ is also an optimal $h^*(x)$ that minimizes $\beta_0[h(x)]$.

(iv) For discrete X, Y and Z and a given $Z=z$, let $Q_{X,Y|Z} \equiv \left(\frac{p(x,y|z)}{\sqrt{p(x|z)p(y|z)}} \right)_{x,y} = \left(\frac{p(x,y)}{\sqrt{p(x)p(y)}} \sqrt{\frac{p(z|x)}{p(z|y)}} \right)_{x,y}$, we first prove that its second largest singular value is $\rho_m^2(X,Y|Z) = \sup_{f,g} \mathbb{E}_{x,y\sim p(x,y|z)}[f(x)g(y)]$ (with $\mathbb{E}_{x\sim p(x|z)}[f(x)] = \mathbb{E}_{y\sim p(y|z)}[g(y)] = 0$ and $\mathbb{E}_{x\sim p(x|z)}[f^2(x)] = \mathbb{E}_{y\sim p(y|z)}[g^2(y)] = 1$).

Let column vectors $u_1 = \sqrt{p(x|z)}$ and $v_1 = \sqrt{p(y|z)}$ (note that z is given and fixed). Also let $u_2 = f(x)\sqrt{p(x|z)}$ and $v_2 = g(y)\sqrt{p(y|z)}$. Denote inner product $\langle u,v \rangle \equiv \sum_i u_i v_i$, and the length of a vector as $||u|| = \sqrt{\langle u,u \rangle}$. We have $||u_1|| = ||v_1|| = 1$ due to the normalization of probability, $||u_2|| = ||v_2|| = 1$ due to $\mathbb{E}_{x\sim p(x|z)}[f^2(x)] = \mathbb{E}_{y\sim p(y|z)}[g^2(y)] = 1$, and $\langle u_1,u_2 \rangle = \langle v_1,v_2 \rangle = 0$ due to $\mathbb{E}_{x\sim p(x|z)}[f(x)] = \mathbb{E}_{y\sim p(y|z)}[g(y)] = 0$. Furthermore, we have

$$
\sup_{f,g} \mathbb{E}_{x,y\sim p(x,y|z)}[f(x)g(y)] = \max_{u,v} u^T Q_{X,Y|Z} v
$$

which is exactly the second largest singular value $\sigma_2(Z)$ of the matrix $Q_{X,Y|Z}$. Using the result in **(ii)**, we have that $\rho_r(X,Y;Z) = \max_Z \sigma_2(Z)$.

□

H SUBSET SEPARATION AT PHASE TRANSITIONS

In this section we study the behavior of $p(z|x)$ on the phase transitions. We use the same categorical dataset (where $|X| = |Y| = |Z| = 3$ and $p(x)$ is uniform, and $p(y|x)$ is given in Fig. 5). In Fig. 6

we show the $p(z|x)$ on the simplex before and after each phase transition. We see that the first phase transition corresponds to the separation of $x = 2$ (belonging to $y = 2$) w.r.t. $x \in \{0, 1\}$ (belonging to classes $y \in \{0, 1\}$), on the $p(z|x)$ simplex. The second phase transition corresponds to the separation of $x = 0$ with $x = 1$. Therefore, each phase transition corresponds to the ability to distinguish subset of examples, and learning of new classes.

I MNIST EXPERIMENT DETAILS

We use the MNIST training examples with class $0, 1, 2, 3$, with a hidden label-noise matrix as given in Fig. 7, based on which at each minibatch we dynamically sample the observed label. We use conditional entropy bottleneck (CEB) (Fischer, 2018) as the variational IB objective, and run multiple independent instances with different the target β. We jump start learning by started training at $\beta = 100$ for 100 epochs, annealing β from 100 down to the target β over 600 epochs, and continue to train at the target epoch for another 800 epochs. The encoder is a three-layer neural net, where each hidden layer has 512 neurons and leakyReLU activation, and the last layer has linear activation. The classifier $p(y|z)$ is a 2-layer neural net with a 128-neuron ReLU hidden layer. The backward encoder $p(z|y)$ is also a 2-layer neural net with a 128-neuron ReLU hidden layer. We trained with Adam (Kingma & Welling, 2013) at learning rate of 10^{-3}, and anneal down with factor $1/(1 + 0.01 \cdot \text{epoch})$. For Alg. 1, for the $f_{\boldsymbol{\theta}}$ we use the same architecture as the encoder of CEB, and use $|Z| = 50$ in Alg. 1.

J CIFAR10 EXPERIMENT DETAILS

We use the same CIFAR10 class confusion matrix provided in Wu et al. (2019) to generate noisy labels with about 20% label noise on average (reproduced in Table 1). We trained 28×1 Wide ResNet (He et al., 2016; Zagoruyko & Komodakis, 2016) models using the open source implementation from Cubuk et al. (2018) as encoders for the Variational Information Bottleneck (VIB) (Alemi et al., 2016). The 10 dimensional output of the encoder parameterized a mean-field Gaussian with unit covariance. Samples from the encoder were passed to the classifier, a 2 layer MLP. The marginal distributions were mixtures of 500 fully covariate 10-dimensional Gaussians, all parameters of which are trained.

With this standard model, we trained 251 different models at β from 1.0 to 6.0 with step size of 0.02. As in Wu et al. (2019), we jump-start learning by annealing β from 100 down to the target β. We do this over the first 4000 steps of training. The models continued to train for another 56,000 gradient steps after that, a total of 600 epochs. We trained with Adam (Kingma & Ba, 2015) at a base learning rate of 10^{-3}, and reduced the learning rate by a factor of 0.5 at 300, 400, and 500 epochs. The models converged to essentially their final accuracy within 40,000 gradient steps, and then remained stable. The accuracies reported in Figure 4 are averaged across five passes over the training set. We use $|Z| = 50$ in Alg. 1.

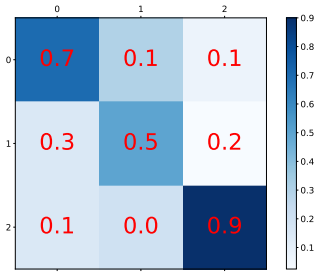
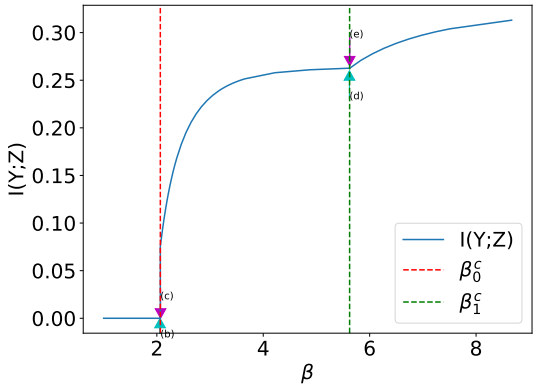
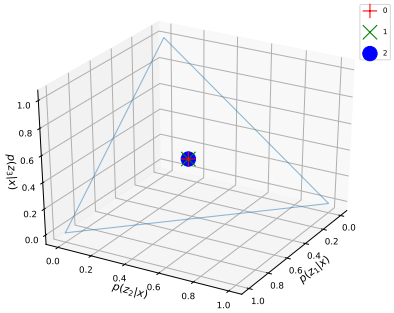


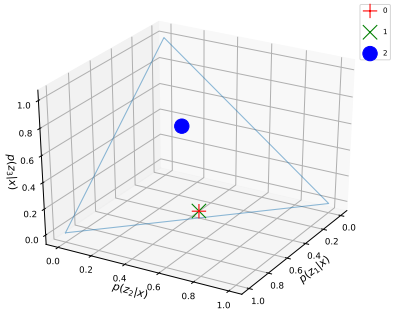
Figure 5: $p(y|x)$ for the categorical dataset in Fig. 2 and Fig. 6. The value in i^{th} row and j^{th} column denotes $p(y = j|x = i)$. $p(x)$ is uniform.



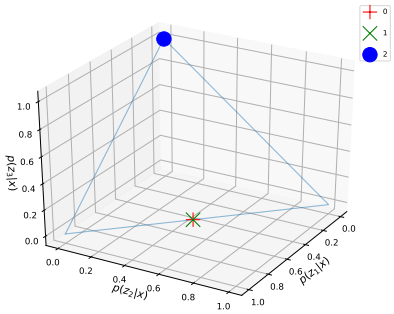
(a)



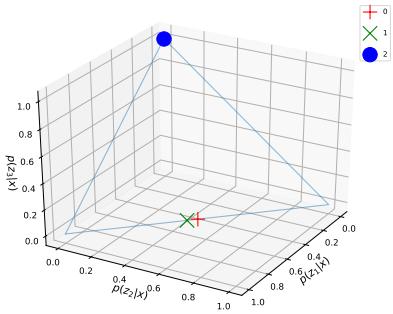
(b)



(c)



(d)



(e)

Figure 6: (a) $I(Y;Z)$ vs. β for the dataset given in Fig. 5. The phase transitions are marked with vertical dashed line, with $\beta_0^c = 2.065571$ and $\beta_1^c = 5.623333$. (b)-(e) Optimal $p^*_\beta(z|x)$ for four values of β, i.e. (b) $\beta = 2.060$, (c) $\beta = 2.070$, (d) $\beta = 5.620$ (e) $\beta = 5.625$ (their β values are also marked in (a)), where each marker denotes $p(z|x = i)$ for a given $i \in \{0, 1, 2\}$.

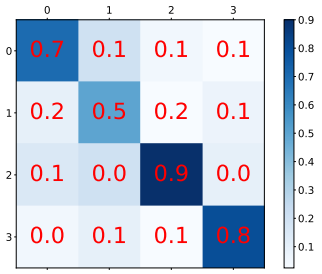


Figure 7: Confusion matrix for MNIST experiment. The value in i^{th} row and j^{th} column denotes $p(\tilde{y} = j|y = i)$ for the label noise.

Table 1: Class confusion matrix used in CIFAR10 experiments, reproduced from (Wu et al., 2019). 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