Learnability for the Information Bottleneck Supplemental Material

The structure of the Supplemental Material is as follows. In Section A, we provide preliminaries for the first-order and second-order variations on functionals. In Section C, we prove Theorem 3, the sufficient condition 1 for IB-Learnability. In Section D, we calculate the first and second variations of $\text{IB}_\beta[p(z|x)]$ at the trivial representation $p(z|x) = p(z)$, which is used in proving the Sufficient Condition 2 for IB_β-learnability (Section F). After these preparations, we prove the key result of this paper, Theorem 5, in Section G. Then two important corollaries 5.1, 5.2 are proved in Section H. In Section I we explore the deep relation between β_0, $\beta_0[h(x)]$, the hypercontractivity coefficient, contraction coefficient and maximum correlation. Finally in Section J, we provide details for the experiments.

A Preliminaries: first-order and second-order variation

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

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

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

B Proof of Theorem 1

Proof. If (X, Y) is IB_β-learnable, then there exists Z given by some $p_1(z|x)$ such that $\text{IB}_\beta(X, Y; Z) < \text{IB}(X, Y; Z_{trivial}) = 0$, where $Z_{trivial}$ satisfies $p(z|x) = p(z)$. Since $X' = g(X)$ is a uniquely invertible map (if X is continuous variable, g is additionally required to be continuous), and mutual information is invariant under such an invertible map (Kraskov et al. (2004)), we have that $\text{IB}_\beta(X', Y; Z) = I(X'; Z) - \beta I(Y; Z) = I(X; Z) - \beta I(Y; Z) = \text{IB}_\beta(X, Y; Z) < 0 = \text{IB}(X', Y; Z_{trivial})$, so (X', Y) is IB_β-learnable. On the other hand, if (X, Y) is not IB_β-learnable, then $\forall Z$, we have $\text{IB}_\beta(X, Y; Z) \geq \text{IB}(X, Y; Z_{trivial}) = 0$. Again using mutual information's invariance under g, we have for all Z, $\text{IB}_\beta(X', Y; Z) = \text{IB}_\beta(X, Y; Z) \geq \text{IB}(X, Y; Z_{trivial}) = 0$, leading to that (X', Y) is not IB_β-learnable. Therefore, we have that (X, Y) and (X', Y) have the same IB_β-learnability.

□

C Proof of Theorem 3

Proof. To prove Theorem 3, we use the Theorem 1 of Chapter 5 of Gelfand et al. (2000) which gives a necessary condition for $F[f(x)]$ to have a minimum at $f_0(x)$. Adapting to our notation, we have:

Theorem 6 (Gelfand et al. (2000)). *A necessary condition for the functional $F[f(x)]$ to have a minimum at $f(x) = f_0(x)$ is that for $f(x) = f_0(x)$ and all admissible $\epsilon h(x)$,*

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

.

Applying to our functional $\text{IB}_\beta[p(z|x)]$, an immediate result of Theorem 6 is that, if at $p(z|x) = p(z)$, there exists an $\epsilon h(z|x)$ such that $\delta^2\text{IB}_\beta[p(z|x)] < 0$, then $p(z|x) = p(z)$ is not a minimum for $\text{IB}_\beta[p(z|x)]$. Using the definition of IB_β learnability, we have that (X, Y) is IB_β-learnable.

□

D First- and second-order variations of $IB_\beta[p(z|x)]$

In this section, we derive the first- and second-order variations of $\text{IB}_\beta[p(z|x)]$, which are needed for proving Lemma 2.1 and Theorem 4.

Lemma 6.1. *Using perturbative function $h(z|x)$, we have*

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

$$\delta^2 IB_\beta[p(z|x)] =$$

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

Proof. Since $\text{IB}_\beta[p(z|x)] = I(X;Z) - \beta I(Y;Z)$, let us calculate the first and second-order variation of $I(X;Z)$ and $I(Y;Z)$ w.r.t. $p(z|x)$, respectively. Through this derivation, we use $\epsilon h(z|x)$ as a perturbative function, for ease of deciding different orders of variations. We will finally absorb ϵ into $h(z|x)$.

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

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

Since

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

We have

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

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

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

Collecting the first order terms of ϵ, we have

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

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

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

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

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

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

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

Hence

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

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

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

Collecting the first order terms of ϵ, we have

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

Collecting the second order terms of ϵ, we have

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

Finally, we have

$$\begin{aligned}
\delta\text{IB}_\beta[p(z|x)] &= \delta F_1[p(z|x)]-\beta\cdot\delta F_2[p(z|x)] \\
&= \epsilon\left(\int dxdzp(x)h(z|x)\log\frac{p(z|x)}{p(z)}-\beta\int dxdydzp(x,y)h(z|x)\log\frac{p(z|y)}{p(z)}\right)
\end{aligned} \tag{8}$$

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

Absorb ϵ into $h(z|x)$, we get rid of the ϵ factor and obtain the final expression in Lemma 6.1.

□

E Proof of Lemma 2.1

Proof. Using Lemma 6.1, we have

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

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

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

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

□

F Proof of Theorem 4

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

Now using Theorem 3, a sufficient condition for (X, Y) to be IB_β-learnable is that there exists $h(z|x)$ with $\int h(z|x)dx = 0$ such that $\delta^2\mathrm{IB}_\beta[p(z|x)] < 0$ at $p(z|x) = p(x)$.

At the trivial representation, $p(z|x) = p(z)$ and hence $p(x,z) = p(x)p(z)$. Due to the Markov chain $Z \leftarrow X \leftrightarrow Y$, we have $p(y,z) = p(y)p(z)$. Substituting them into the $\delta^2\mathrm{IB}_\beta[p(z|x)]$ in Lemma 6.1, the condition becomes: there exists $h(z|x)$ with $\int h(z|x)dz = 0$, such that

$$
\begin{aligned}
&0 > \delta^2\mathrm{IB}_\beta[p(z|x)] = \\
&\frac{1}{2}\left[\int dxdz \frac{p(x)^2}{p(x)p(z)} h(z|x)^2 - \beta\int dxdx'dydz \frac{p(x,y)p(x',y)}{p(y)p(z)} h(z|x)h(z|x') + (\beta-1)\int dxdx'dz \frac{p(x)p(x')}{p(z)} h(z|x)h(z|x')\right]
\end{aligned} \tag{9}
$$

Rearranging terms and simplifying, we have

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

where

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

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

If $\forall h(z|x), G[h(z|x)] \geq 0$, then we have $\forall h(z|x), \int \frac{dz}{p(z)} G[h(z|x)] \geq 0$. Therefore, if $\exists h(z|x)$ s.t. $\int \frac{dz}{p(z)} G[h(z|x)] < 0$, we have that $\exists h(z|x)$ s.t. $G[h(z|x)] < 0$. Since the functional $G[h(z|x)]$ does not contain integration over z, we can treat the z in $G[h(z|x)]$ as a parameter and we have that $\exists h(x)$ s.t. $G[h(x)] < 0$.

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

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

In other words, the condition Eq. (9) is equivalent to requiring that there exists an $h(x)$ such that $G[h(x)] < 0$. Hence, a sufficient condition for IB_β-learnability is that there exists an $h(x)$ such that

$$G[h(x)] = \int dx h(x)^2 p(x) - \beta \int \frac{dy}{p(y)} \Big(\int dx h(x) p(x) p(y|x) \Big)^2 + (\beta - 1) \Big(\int dx h(x) p(x) \Big)^2 < 0 \tag{10}$$

When $h(x) = C = \text{const}$ in the entire input space $\mathcal{X}$, Eq. (10) becomes:

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

which cannot be true. Therefore, $h(x) = \text{const}$ cannot satisfy Eq. (10).

Rearranging terms and simplifying, and note that $\big[\int dx h(x) p(x) \big]^2 > 0$ due to $h(x) \not\equiv 0 = \text{const}$, we have

$$\beta \left[\frac{\int \frac{dy}{p(y)} \big(\int dx h(x) p(x) p(y|x) \big)^2}{\big(\int dx h(x) p(x) \big)^2} - 1 \right] > \frac{\int dx h(x)^2 p(x)}{\big(\int dx h(x) p(x) \big)^2} - 1 \tag{11}$$

For the R.H.S. of Eq. (11), let us show that it is greater than 0. Using Cauchy-Schwarz inequality: $\langle u, u \rangle \langle v, v \rangle \geq \langle u, v \rangle^2$, and setting $u(x) = h(x)\sqrt{p(x)}$, $v(x) = \sqrt{p(x)}$, and defining the inner product as $\langle u, v \rangle = \int u(x) v(x) dx$. We have

$$\frac{\int dx h(x)^2 p(x)}{\big(\int dx h(x) p(x) \big)^2} \geq \frac{1}{\int p(x) dx} = 1$$

It attains equality when $\frac{u(x)}{v(x)} = h(x)$ is constant. Since $h(x)$ cannot be constant, we have that the R.H.S. of Eq. (11) is greater than 0.

For the L.H.S. of Eq. (11), due to the necessary condition that $\beta > 0$, if $\left[\frac{\int \frac{dy}{p(y)} \big(\int dx h(x) p(x) p(y|x) \big)^2}{\big(\int dx h(x) p(x) \big)^2} - 1 \right] \leq 0$, Eq. (11) cannot hold. Then the $h(x)$ such that Eq. (11) holds is for those that satisfies

$$\frac{\int \frac{dy}{p(y)} \big(\int dx h(x) p(x) p(y|x) \big)^2}{\big(\int dx h(x) p(x) \big)^2} - 1 > 0$$

i.e.

$$\int dy p(y) \Big(\int dx p(x|y) h(x) \Big)^2 > \Big(\int dx p(x) h(x) \Big)^2$$

We see this constraint contains the requirement that $h(x) \not\equiv \text{const}$.

Written in the form of expectations, we have

$$\mathbb{E}_{y\sim p(y)}\left[\left(\mathbb{E}_{x\sim p(x|y)}[h(x)]\right)^2\right] > \left(\mathbb{E}_{x\sim p(x)}[h(x)]\right)^2 \tag{12}$$

Since the square function is convex, using Jensen's inequality on the outer expectation on the L.H.S. of Eq. (12), we have

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

The equality holds iff $\mathbb{E}_{x\sim p(x|y)}[h(x)]$ is constant w.r.t. y, i.e. Y is independent of X. Therefore, in order for Eq. (12) to hold, we require that Y is not independent of X.

Using Jensen's inequality on the innter expectation on the L.H.S. of Eq. (12), we have

$$\mathbb{E}_{y\sim p(y)}\left[\left(\mathbb{E}_{x\sim p(x|y)}[h(x)]\right)^2\right] \leq \mathbb{E}_{y\sim p(y)}\left[\mathbb{E}_{x\sim p(x|y)}[h(x)^2]\right] = \mathbb{E}_{x\sim p(x)}[h(x)^2] \tag{13}$$

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

Under the constraint that Y is not independent of X, we can divide both sides of Eq. 10, and obtain the condition: there exists an $h(x)$ such that

$$\beta > \frac{\frac{\int dx h(x)^2 p(x)}{\left(\int dx h(x)p(x)\right)^2} - 1}{\frac{\int \frac{dy}{p(y)}\left(\int dx h(x)p(x)p(y|x)\right)^2}{\left(\int dx h(x)p(x)\right)^2} - 1}$$

i.e.

$$\beta > \inf_{h(x)} \frac{\frac{\int dx h(x)^2 p(x)}{\left(\int dx h(x)p(x)\right)^2} - 1}{\frac{\int \frac{dy}{p(y)}\left(\int dx h(x)p(x)p(y|x)\right)^2}{\left(\int dx h(x)p(x)\right)^2} - 1}$$

Written in the form of expectations, we have

$$\begin{aligned}\beta &> \inf_{h(x)} \frac{\frac{\mathbb{E}_{x\sim p(x)}[h(x)^2]}{\left(\mathbb{E}_{x\sim p(x)}[h(x)]\right)^2} - 1}{\int \frac{dy}{p(y)}\left(\frac{\mathbb{E}_{x\sim p(x)}[p(y|x)h(x)]}{\mathbb{E}_{x\sim p(x)}[h(x)]}\right)^2 - 1} \\ &= \inf_{h(x)} \frac{\frac{\mathbb{E}_{x\sim p(x)}[h(x)^2]}{\left(\mathbb{E}_{x\sim p(x)}[h(x)]\right)^2} - 1}{\mathbb{E}_{y\sim p(y)}\left[\left(\frac{\mathbb{E}_{x\sim p(x|y)}[h(x)]}{\mathbb{E}_{x\sim p(x)}[h(x)]}\right)^2\right] - 1}\end{aligned}$$

We can absorb the constraint Eq. (12) into the above formula, and get

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

where

$$\beta_0[h(x)] = \frac{\frac{\mathbb{E}_{x\sim p(x)}[h(x)^2]}{\left(\mathbb{E}_{x\sim p(x)}[h(x)]\right)^2} - 1}{\mathbb{E}_{y\sim p(y)}\left[\left(\frac{\mathbb{E}_{x\sim p(x|y)}[h(x)]}{\mathbb{E}_{x\sim p(x)}[h(x)]}\right)^2\right] - 1}$$

which proves the condition of Theorem 4.

Furthermore, from Eq. (13) we have

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

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

Proof of lower bound of slope of the Pareto frontier at the origin:

Now we prove the second statement of Theorem 4. Since $\delta I(X;Z) = 0$ and $\delta I(Y;Z) = 0$ according to Lemma 2.1, we have $\left(\frac{\Delta I(Y;Z)}{\Delta I(X;Z)}\right)^{-1} = \left(\frac{\delta^2 I(Y;Z)}{\delta^2 I(X;Z)}\right)^{-1}$. Substituting into the expression of $\delta^2 I(Y;Z)$ and $\delta^2 I(X;Z)$ from Lemma 6.1, we have

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

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

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

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

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

$$
\beta > \inf_{h(x)} \beta^{(2)}[h(x)] \tag{14}
$$

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

Multiple phase transitions To discuss multiple phase transitions, let us first obtain the $\beta^{(1)}$ for stationary solution for the IB objective. At a stationary solution for $\text{IB}_\beta[p(z|x)]$, for valid perturbation $h(z|x)$ satisfying $\int dz h(z|x) = 0$ for any x, we have $\delta\left[\text{IB}_\beta[p(z|x)] - \int dzdx\lambda(x)p(z|x)\right] = 0$ as a constraint optimization with $\lambda(x)$ as Lagrangian multipliers. Using Eq. (8), we have

$$
\begin{aligned}
&\delta\text{IB}_\beta[p(z|x)] - \delta\int dzdx\lambda(x)p(z|x) \\
&= \int dxdzp(x)h(z|x)\log\frac{p(z|x)}{p(z)} - \beta\int dxdydzp(x,y)h(z|x)\log\frac{p(z|y)}{p(z)} - \int dzdx\lambda(x)h(z|x) = 0
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
\beta^{(1)} &\equiv \frac{\int dxdzp(x)h(z|x)\log\frac{p(z|x)}{p(z)} - \int dzdx\lambda(x)h(z|x)}{\int dxdydzp(x,y)h(z|x)\log\frac{p(z|y)}{p(z)}} \\
&= \frac{p(x)\log\frac{p(z|x)}{p(z)} - \lambda(x)}{\int dyp(x,y)\log\frac{p(z|y)}{p(z)}}
\end{aligned} \tag{15}
$$

The last equality is due to that the first equality is always true for any function $h(z|x)$. So we can take out the $\int dxdzh(z|x)$ factor. $\lambda(x)$ is used for normalization of $p(z|x)$. Eq. (15) is equivalent to the result of the self-consistent equation in Tishby et al. (2000).

Eq. (15) and Eq. (14) provide us with an ideal tool to study multiple phase transitions. For each β, at the minimization of the IB objective, Eq. (15) is satisfied by some Z^* that is at the Pareto frontier on the $I(Y;Z)$ vs. $I(X;Z)$ plane. As we increase β, the $\inf_{h(x)} \beta^{(2)}[h(x)]$ may remain stable for a wide range of β, until β is greater than $\inf_{h(x)} \beta^{(2)}[h(x)]$, at which point we will have a phase transition where suddenly there is a better $Z = Z^{**}$ that achieves much lower $\text{IB}_\beta(X, Y; Z)$ value.

For example, we can rewrite Eq. (15) as

$$\log\frac{p(z|x)}{p(z)} = \beta^{(1)} \int dyp(y|x)\log\frac{p(z|y)}{p(z)} + \tilde{\lambda}(x) \tag{16}$$

where $\tilde{\lambda}(x) = \frac{\lambda(x)}{p(x)}$. By substituting into Eq. (14), we may proceed and get useful results. □

G Proof of Theorem 5

Proof. According to Theorem 4, a sufficient condition for (X, Y) to be IB_β-learnable is that X and Y are not independent, and

$$\beta > \inf_{h(x)} \frac{\frac{\mathbb{E}_{x\sim p(x)}[h(x)^2]}{\left(\mathbb{E}_{x\sim p(x)}[h(x)]\right)^2} - 1}{\mathbb{E}_{y\sim p(y)}\left[\left(\frac{\mathbb{E}_{x\sim p(x|y)}[h(x)]}{\mathbb{E}_{x\sim p(x)}[h(x)]}\right)^2\right] - 1} \tag{17}$$

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

$$h(x) = \begin{cases} 1, x \in \Omega_x \\ 0, \text{otherwise} \end{cases} \tag{18}$$

for certain $\Omega_x \subset \mathcal{X}$. Substituting into Eq. (18), we have that a sufficient condition for (X, Y) to be IB_β-learnable is

$$\beta > \inf_{\Omega_x \subset \mathcal{X}} \frac{\frac{p(\Omega_x)}{p(\Omega_x)^2} - 1}{\int dyp(y)\left(\frac{\int_{x\in\Omega_x} dxp(x|y)dx}{p(\Omega_x)}\right)^2 - 1} > 0 \tag{19}$$

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

The denominator of Eq. (19) is

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

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

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

Both equalities hold iff $p(y|\Omega_x) \equiv p(y)$, at which the denominator of Eq. (19) is equal to 0 and the expression inside the infimum diverge, which will not contribute to the infimum. Except this scenario, the denominator is greater than 0. Substituting into Eq. (19), we have that a sufficient condition for (X, Y) to be IB$_\beta$-learnable is

$$\beta > \inf_{\Omega_x \subset \mathcal{X}} \frac{\frac{p(\Omega_x)}{p(\Omega_x)^2} - 1}{\mathbb{E}_{y\sim p(y|\Omega_x)}\left[\frac{p(y|\Omega_x)}{p(y)} - 1\right]} \tag{20}$$

Since Ω_x is a subset of $\mathcal{X}$, by the definition of $h(x)$ in Eq. (18), $h(x)$ is not a constant in the entire $\mathcal{X}$. Hence the numerator of Eq. (20) is positive. Since its denominator is also positive, we can then neglect the "> 0", and obtain the condition in Theorem 5.

Since the $h(x)$ used in this theorem is a subset of the $h(x)$ used in Theorem 4, the infimum for Eq. (4) is greater than or equal to the infimum in Eq. (2). Therefore, according to the second statement of Theorem 4, we have that the $(\inf_{\Omega_x \subset \mathcal{X}} \beta_0(\Omega_x))^{-1}$ is also a lower bound of the slope for the Pareto frontier of $I(Y;Z)$ vs. $I(X;Z)$ curve.

Now we prove that the condition Eq. (4) is invariant to invertible mappings of X. In fact, if $X' = g(X)$ is a uniquely invertible map (if X is continuous, g is additionally required to be continuous), let $\mathcal{X}' = \{g(x)|x \in \Omega_x\}$, and denote $g(\Omega_x) \equiv \{g(x)|x \in \Omega_x\}$ for any $\Omega_x \subset \mathcal{X}$, we have $p(g(\Omega_x)) = p(\Omega_x)$, and $p(y|g(\Omega_x)) = p(y|\Omega_x)$. Then for dataset (X, Y), let $\Omega'_x = g(\Omega_x)$, we have

$$\frac{\frac{1}{p(\Omega'_x)} - 1}{\mathbb{E}_{y\sim p(y|\Omega'_x)}\left[\frac{p(y|\Omega'_x)}{p(y)} - 1\right]} = \frac{\frac{1}{p(\Omega_x)} - 1}{\mathbb{E}_{y\sim p(y|\Omega_x)}\left[\frac{p(y|\Omega_x)}{p(y)} - 1\right]} \tag{21}$$

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

$$\inf_{\Omega'_x \subset \mathcal{X}'} \frac{\frac{1}{p(\Omega'_x)} - 1}{\mathbb{E}_{y\sim p(y|\Omega'_x)}\left[\frac{p(y|\Omega'_x)}{p(y)} - 1\right]} = \inf_{\Omega_x \subset \mathcal{X}} \frac{\frac{1}{p(\Omega_x)} - 1}{\mathbb{E}_{y\sim p(y|\Omega_x)}\left[\frac{p(y|\Omega_x)}{p(y)} - 1\right]} \tag{22}$$

For dataset $(X', Y) = (g(X), Y)$, applying Theorem 5 we have that a sufficient condition for it to be IB$_\beta$-learnable is

$$\beta > \inf_{\Omega'_x \subset \mathcal{X}'} \frac{\frac{1}{p(\Omega'_x)} - 1}{\mathbb{E}_{y\sim p(y|\Omega'_x)}\left[\frac{p(y|\Omega'_x)}{p(y)} - 1\right]} = \inf_{\Omega_x \subset \mathcal{X}} \frac{\frac{1}{p(\Omega_x)} - 1}{\mathbb{E}_{y\sim p(y|\Omega_x)}\left[\frac{p(y|\Omega_x)}{p(y)} - 1\right]} \tag{23}$$

where the equality is due to Eq. (22). Comparing with the condition for IB$_\beta$-learnability for (X, Y) (Eq. (4)), we see that they are the same. Therefore, the condition given by Theorem 5 is invariant to invertible mapping of X.

□

H Proof of Corollary 5.1 and Corollary 5.2

H.1 Proof of Corollary 5.1

Proof. We use Theorem 5. Let Ω_x contain all elements x whose true class is y^* for some certain y^*, and 0 otherwise. Then we obtain a (potentially stronger) sufficient condition. Since the probability $p(y|y^*, x) = p(y|y^*)$ is class-conditional, we have

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

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

H.2 Proof of Corollary 5.2

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

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

□

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

I β_0, hypercontractivity coefficient, contraction coefficient, $\beta_0[h(x)]$, and maximum correlation

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

$$\frac{1}{\beta_0} = \xi(X;Y) = \eta_{\text{KL}}(p(y|x), p(x)) \geq \sup_{h(x)} \frac{1}{\beta_0[h(x)]} = \rho_m^2(X;Y) \tag{24}$$

Proof. The hypercontractivity coefficient ξ is defined as (Anantharam et al., 2013):

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

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

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

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

$$0 < \frac{1}{\beta} < \frac{I(Y;Z)}{I(X;Z)} \tag{25}$$

By Theorem 2, the IB-Learnability region for β is $(\beta_0, +\infty)$, or equivalently the IB-Learnability region for $1/\beta$ is

$$0 < \frac{1}{\beta} < \frac{1}{\beta_0} \tag{26}$$

Comparing Eq. (25) and Eq. (26), we have that

$$\frac{1}{\beta_0} = \sup_{Z-X-Y} \frac{I(Y;Z)}{I(X;Z)} = \xi(X;Y) \tag{27}$$

In Anantharam et al. (2013), the authors prove that

$$\xi(X;Y) = \eta_{\text{KL}}(p(y|x), p(x)) \tag{28}$$

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

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

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

In Anantharam et al. (2013), the authors also provide a counterexample to an earlier result by Erkip and Cover (1998) that incorrectly proved $\xi(X;Y) = \rho_m^2(X;Y)$. In the specific counterexample Anantharam et al. (2013) design, $\xi(X;Y) > \rho_m^2(X;Y)$.

The maximum correlation is defined as $\rho_m(X;Y) \equiv \max_{f,g} \mathbb{E}[f(X)g(Y)]$ where $f(X)$ and $g(Y)$ are real-valued random variables such that $\mathbb{E}[f(X)] = \mathbb{E}[g(Y)] = 0$ and $\mathbb{E}[f^2(X)] = \mathbb{E}[g^2(Y)] = 1$ (Hirschfeld, 1935; Gebelein, 1941).

Now we prove $\xi(X;Y) \geq \rho_m^2(X;Y)$, based on Theorem 4. To see this, we use the alternate characterization of $\rho_m(X;Y)$ by Rényi (1959):

$$\rho_m^2(X;Y) = \max_{f(X):\mathbb{E}[f(X)]=0,\mathbb{E}[f^2(X)]=1} \mathbb{E}[(\mathbb{E}[f(X)|Y])^2] \tag{29}$$

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

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

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

Combined with Eq. (29), we have

$$
\sup_{h(x)} \frac{1}{\beta_0[h(x)]} = \rho_m^2(X;Y) \tag{30}
$$

Our Theorem 4 states that

$$
\sup_{h(x)} \frac{1}{\beta_0[h(x)]} \leq \frac{1}{\beta_0} \tag{31}
$$

Combining Eqs. (26), (30) and Eq. (31), we have

$$
\rho_m^2(X;Y) \leq \xi(X;Y) \tag{32}
$$

In summary, the relations among the quantities are:

$$
\frac{1}{\beta_0} = \xi(X;Y) = \eta_{\text{KL}}(p(y|x), p(x)) \geq \sup_{h(x)} \frac{1}{\beta_0[h(x)]} = \rho_m^2(X;Y) \tag{33}
$$

□

J Experiment Details

We use the Variational Information Bottleneck (VIB) objective from Alemi et al. (2016). For the synthetic experiment, the latent Z has dimension of 2. The encoder is a neural net with 2 hidden layers, each of which has 128 neurons with ReLU activation. The last layer has linear activation and 4 output neurons; the first two parameterize the mean of a Gaussian and the last two parameterize the log variance. The decoder is a neural net with 1 hidden layer with 128 neurons and ReLU activation. Its last layer has linear activation and outputs the logit for the class labels. It uses a mixture of Gaussian prior with 500 components (for the experiment with class overlap, 256 components), each of which is a 2D Gaussian with learnable mean and log variance, and the weights for the components are also learnable. For the MNIST experiment, the architecture is mostly the same, except the following: (1) for Z, we let it have dimension of 256. For the prior, we use standard Gaussian with diagonal covariance matrix.

For all experiments, we use Adam (Kingma and Ba (2014)) optimizer with default parameters. We do not add any regularization. We use learning rate of 10^{-4} and have a learning rate decay of $\frac{1}{1+0.01\times \text{epoch}}$. We train in total 2000 epochs with batch size of 500.

For estimation of the observed β_0 in Fig. 2, in the $I(X;Z)$ vs. β_i curve (β_i denotes the i^{th} β), we take the mean and standard deviation of $I(X;Z)$ for the lowest 5 β_i values, denoting as μ_β, σ_β ($I(Y;Z)$ has similar behavior, but since we are minimizing $I(X;Z) - \beta \cdot I(Y;Z)$, the onset of nonzero $I(X;Z)$ is less prone to noise). When $I(X;Z)$ is greater than μ_β + $3\sigma_\beta$, we regard it as learning a non-trivial representation, and take the average of β_i and β_{i-1} as the experimentally estimated onset of learning. We also inspect manually and confirm that it is consistent with human intuition.

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

For estimation of $p(y|x)$ for (2′) Alg. 1 and (3′) $\hat{\eta}_{\text{KL}}$ for both synthetic and MNIST experiments, we use a 3-layer neuron net where each hidden layer has 128 neurons and ReLU activation. The last layer has linear activation. The objective is cross-entropy loss. We use Adam (Kingma and Ba, 2014) optimizer with a learning rate of 1e-4, and train for 100 epochs (at which the validation loss does not go down).

For estimating β_0 via (3′) $\hat{\eta}_{\text{KL}}$ by the algorithm in (Kim et al., 2017), we use the code from the GitHub repository provided by the paper[4], using the same $p(y|x)$ employed for (2′) Alg. 1. Since our datasets are classification tasks, we use $A_{ij} = p(y_j|x_i)/p(y_j)$ instead of the kernel density for estimating matrix A; we take the maximum of 10 runs as estimation of μ.

J.1 Detailed tables for classification with class-conditional noise

In Table J.1 we give the full set of values used for Fig. 2. As the noise rate increases, the true β_0 increases dramatically. Note that the Observed values are empirical estimates. Corollary 5.1 and Alg. 1 agree almost perfectly when using the true $p(y|x)$. $\hat{\eta}_{\text{KL}}$ is somewhat looser, but generally agrees well with the empirical estimates when using the true $p(y|x)$. However, its estimates become much less accurate when $p(y|x)$ is given by a learned neural network trained on the noisy dataset. In contrast, Alg. 1 generally gives much better predictions even when using the estimated $p(y|x)$. Directly optimizing Eq. 2 on the observed data is always an upper bound, although the bound becomes somewhat looser as the noise becomes extreme.

J.2 MNIST Experiments using Equation 2

Here we explore directly training $h(x)$ from Eq. 2 on the full MNIST training set. Eq. 2 can be optimized using SGD using any differentiable parameterized mapping $h(x) : \mathcal{X} \to \mathbb{R}$. In this case, we chose to parameterize $h(x)$ with a PixelCNN++ architecture (van den Oord et al., 2016; Salimans et al., 2017), as PixelCNN++ is a powerful autoregressive model for images that gives a scalar output (normally interpreted as log$p(x)$). Eq. 2 should generally give two clusters in the output space. In this setup, smaller values of $h(x)$ correspond to the subset of the data that is easiest to learn. Fig. 6 shows two strongly separated clusters, as well as the threshold we choose to divide them. Fig. 7 shows the first 5,776 MNIST training examples as sorted by our learned $h(x)$, with the examples above the threshold highlighted in red. We can clearly see that our learned $h(x)$ has separated "easy" ones from the rest of the MNIST training set, in addition to providing a tight bound on β_0.

[4] At https://github.com/wgao9/hypercontractivity

Table 1: Full table of values used to generate Fig. 2.

Noise rate	Observed	(1) Corollary 5.1	(2) Alg. 1 true $p(y\|x)$	(3) $\hat{\eta}_{\text{KL}}$ true $p(y\|x)$	(4) Eq. 2	(2′) Alg. 1	(3′) $\hat{\eta}_{\text{KL}}$
0.02	1.06	1.09	1.09	1.10	1.08	1.08	1.10
0.04	1.20	1.18	1.18	1.21	1.18	1.19	1.20
0.06	1.26	1.29	1.29	1.33	1.30	1.31	1.33
0.08	1.40	1.42	1.42	1.45	1.42	1.43	1.46
0.10	1.52	1.56	1.56	1.60	1.55	1.58	1.60
0.12	1.70	1.73	1.73	1.78	1.71	1.73	1.77
0.14	1.99	1.93	1.93	1.99	1.90	1.91	1.95
0.16	2.04	2.16	2.16	2.24	2.15	2.15	2.16
0.18	2.41	2.44	2.44	2.49	2.43	2.42	2.49
0.20	2.74	2.78	2.78	2.86	2.76	2.77	2.71
0.22	3.15	3.19	3.19	3.29	3.19	3.21	3.29
0.24	3.75	3.70	3.70	3.83	3.71	3.75	3.72
0.26	4.40	4.34	4.34	4.48	4.35	4.31	4.17
0.28	5.16	5.17	5.17	5.37	5.12	4.98	4.55
0.30	6.34	6.25	6.25	6.49	6.24	6.03	5.58
0.32	8.06	7.72	7.72	8.02	7.63	7.19	7.33
0.34	9.77	9.77	9.77	10.13	9.74	8.95	7.37
0.36	12.58	12.76	12.76	13.21	12.51	11.11	10.09
0.38	16.91	17.36	17.36	17.96	16.97	14.55	10.49
0.40	24.66	25.00	25.00	25.99	25.01	20.36	17.27
0.42	39.08	39.06	39.06	40.85	39.48	30.12	10.89
0.44	64.82	69.44	69.44	71.80	76.48	51.95	21.95
0.46	163.07	156.25	156.26	161.88	173.15	114.57	21.47
0.48	599.45	625.00	625.00	651.47	838.90	293.90	8.69

Table 2: Class confusion matrix used in CIFAR10 experiments. The value in row i, column j means for class i, the probability of labeling it as class j. The mean confusion across the classes is 20%.

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

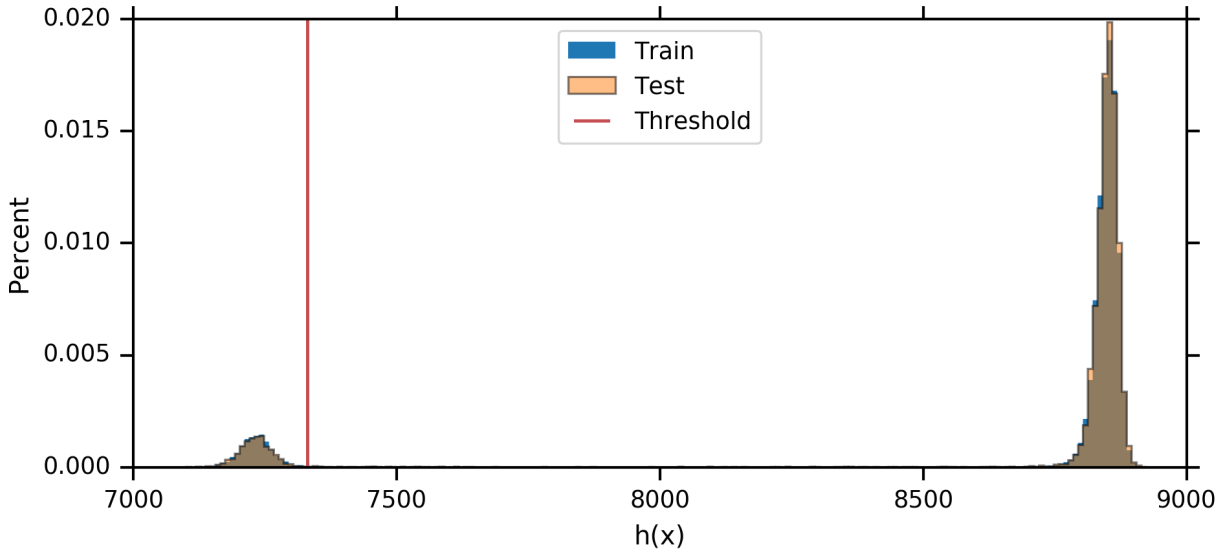

Figure 6: Histograms of the full MNIST training and validation sets according to $h(X)$. Note that both are bimodal, and the histograms are indistinguishable. In both cases, $h(x)$ has learned to separate most of the ones into the smaller mode, but difficult ones are in the wide valley between the two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

J.3 CIFAR10 Details

We trained a deterministic 28x10 wide resnet (He et al., 2016; Zagoruyko and Komodakis, 2016), using the open source implementation from Cubuk et al. (2018). However, we extended the final 10 dimensional logits of that model through another 3 layer MLP classifier, in order to keep the inference network architecture identical between this model and the VIB models we describe below. During training, we dynamically added label noise according to the class confusion matrix in Tab. J.1. The mean label noise averaged across the 10 classes is 20%. After that model had converged, we used it to estimate β_0 with Alg. 1. Even with 20% label noise, β_0 was estimated to be 1.0483.

We then trained 73 different VIB models using the same 28x10 wide resnet architecture for the encoder, parameterizing the mean of a 10-dimensional unit variance Gaussian. Samples from the encoder distribution were fed to the same 3 layer MLP classifier architecture used in the deterministic model. The marginal distributions were mixtures of 500 fully covariate 10-dimensional Gaussians, all parameters of which are trained. The VIB models had β ranging from 1.02 to 2.0 by steps of 0.02, plus an extra set ranging from 1.04 to 1.06 by steps of 0.001 to ensure we captured the empirical β_0 with high precision.

However, this particular VIB architecture does not start learning until $\beta > 2.5$, so none of these models would train as described.[5] Instead, we started them all at $\beta = 100$, and annealed β down to the corresponding target over 10,000 training gradient steps. The models continued to train for another 200,000 gradient steps after that. In all cases, the models converged to essentially their final accuracy within 20,000 additional gradient steps after annealing was completed. They were stable over the remaining $\sim 180,000$ gradient steps.

[5] A given architecture trained using maximum likelihood and with no stochastic layers will tend to have higher effective capacity than the same architecture with a stochastic layer that has a fixed but non-trivial variance, even though those two architectures have exactly the same number of learnable parameters.

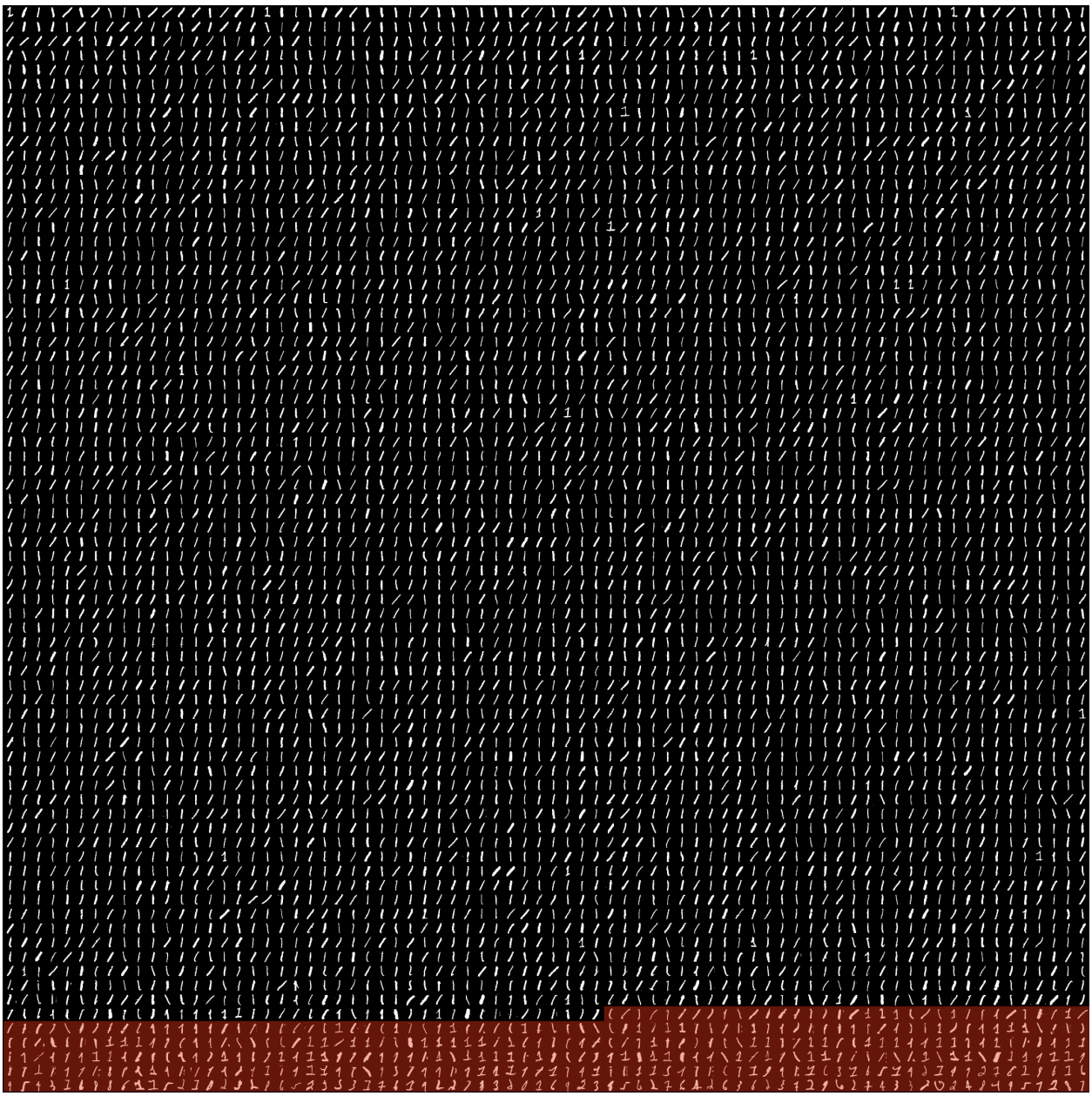

Figure 7: The first 5776 MNIST digits when sorted by $h(x)$. The digits highlighted in red are above the threshold drawn in Fig. 6.

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.