7. Supplementary Material

7.1. Standard Definition of Minimal Sufficient Statistics

The more common phrasing of the definition of *minimal sufficient statistic* is:

Definition 3 (Minimal Sufficient Statistic)**.** A sufficient statistic $f(X)$ for Y is *minimal* if for any other sufficient statistic $h(X)$ there exists a measurable function g such that $f = g \circ h$ almost everywhere.

(Some references do not mention the "measurability" and "almost everywhere" conditions on g, but since we are in the probabilistic setting it is this definition of $f = g \circ h$ that is meaningful.)

Our preferred phrasing of the definition of *minimal sufficient statistic*, which we use in our Introduction is:

Definition 4 (Minimal Sufficient Statistic)**.** A sufficient statistic $f(X)$ for Y is *minimal* if for any measurable function g, $g(f(X))$ is no longer sufficient for Y unless g is invertible almost everywhere (i.e., there exist a measurable function g^{-1} and a set $\mathcal{A}$ such that $g^{-1}(g(x)) = x$ for all $x \in \mathcal{A}$ and the event $\{X \in \mathcal{A}^c\}$ has probability zero).

The equivalence of Definition 3 and Definition 4 is given by the following lemma:

Lemma 2. *Assume that there exists a minimal sufficient statistic $h(X)$ for Y by Definition 3. Then a sufficient statistic $f(X)$ is minimal in the sense of Definition 3 if and only if it is minimal in the sense of Definition 4.*

Proof. We first assume that $f(X)$ is minimal in the sense of Definition 3. Let g be any measurable function such that $g(f(X))$ is sufficient for Y. By the minimality (Def. 3) of f there must exist a measurable function $\tilde{g}$ such that $\tilde{g}(g(f(x))) = f(x)$ almost everywhere. This proves that f is minimal in the sense of Definition 4.

Now assume that $f(X)$ is minimal in the sense of Definition 4 and let $\tilde{f}(X)$ be another sufficient statistic. Because h is minimal (Def. 3), there exist g_1 such that $h = g_1 \circ \tilde{f}$ almost everywhere and g_2 such that $h = g_2 \circ f$ almost everywhere. Because f is minimal (Def. 4), g_2 must be one-to-one almost everywhere, i.e., there exists a $\tilde{g}_2$ such that $\tilde{g}_2 \circ h = \tilde{g}_2 \circ g_2 \circ f = f$ almost everywhere. In turn, we obtain that $\tilde{g}_2 \circ g_1 \circ \tilde{f} = f$ almost everywhere and as $\tilde{f}$ was arbitrary we have proven the minimality of f in the sense of Definition 3. □

7.2. The Mutual Information Between the Input and Output of a Deep Network is Infinite

Typically the mutual information between continuous random variables X and Y is given by

$$I(X,Y) = \int p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy.$$

But this quantity is only defined when the joint density $p(x,y)$ is integrable, which it is not in the case that $Y = f(X)$. (The technical term for $p(x,y)$ in this case is a "singular distribution".) Instead, to compute $I(X, f(X))$ we must refer to the "master definition" of mutual information (Cover & Thomas, 2006), which is

$$I(X,Y) = \sup_{\mathcal{P},\mathcal{Q}} I([X]_{\mathcal{P}}, [Y]_{\mathcal{Q}}), \tag{3}$$

where $\mathcal{P}$ and $\mathcal{Q}$ are finite partitions of the range of X and Y, respectively, and $[X]_{\mathcal{P}}$ is the random variable obtained by quantizing X using partition $\mathcal{P}$, and analogously for $[Y]_{\mathcal{Q}}$.

From this definition, we can prove the following Lemma:

Lemma 3. *If X and Y are continuous random variables, and there are open sets O_X and O_Y in the support of X and Y, respectively, such that $y = f(x)$ for $x \in O_X$, $y \in O_Y$, and some deterministic function f, then $I(X,Y) = \infty$.*

This includes all X and Y where $Y = f(X)$ for an f that is continuous somewhere on its domain, e.g., any deterministic deep network (considered as a function from an input vector to an output vector).

Proof. Suppose X and Y satisfy the conditions of the lemma. Let O_X and O_Y be open sets with $f(O_X) = O_Y$ and $\mathbb{P}[X \in O_X] =: \delta > 0$, which exist by the lemma's assumptions. Then let $\mathcal{P}^n_{O_Y}$ be a partition of O_Y into n disjoint sets. Because Y is continuous and hence does not have any atoms, we may assume that the probability of Y belonging to each element of $\mathcal{P}^n_{O_Y}$ is equal to the same nonzero value δ/n. Denote by $\mathcal{P}^n_{O_X}$ the partition of O_X into n disjoint sets, where

each set in $\mathcal{P}^n_{O_X}$ is the preimage of one of the sets in $\mathcal{P}^n_{O_Y}$. We can construct partitions of the whole domains of X and Y as $\mathcal{P}^n_{O_X} \cup O_X^c$ and $\mathcal{P}^n_{O_Y} \cup O_Y^c$, respectively. Using these partitions in (3), we obtain

$$\begin{aligned} I(X,Y) &\geq (1-\delta)\log(1-\delta) + \sum_{A \in [X]_{\mathcal{P}^n_{O_X}}} \mathbb{P}[X \in A, Y \in f(A)] \log \frac{\mathbb{P}[X \in A, Y \in f(A)]}{\mathbb{P}[X \in A]\mathbb{P}[Y \in f(A)]} \\ &= (1-\delta)\log(1-\delta) + n\frac{\delta}{n}\log\frac{\frac{\delta}{n}}{\frac{\delta}{n}\frac{\delta}{n}} \\ &= (1-\delta)\log(1-\delta) + \delta\log\frac{n}{\delta}. \end{aligned}$$

By letting n go to infinity, we can see that the supremum in Eq. 3 is infinity. □

7.3. Change of Variables Formula for Non-invertible Mappings

The change of variables formula is widely used in machine learning and is key to recent results in density estimation and generative modeling like normalizing flows (Rezende & Mohamed, 2015), NICE (Dinh et al., 2014), or Real NVP (Dinh et al., 2016). But all uses of the change of variables formula in the machine learning literature that we are aware of use it with respect to bijective mappings between random variables, despite the formula also being applicable to non-invertible mappings between random variables. To address this gap, we offer the following brief tutorial.

The familiar form of the change of variables formula for a random variable X with density $p(x)$ and a bijective, differentiable function $f\colon \mathbb{R}^d \to \mathbb{R}^d$ is

$$\int_{\mathbb{R}^d} p(x) J_f(x)\,\mathrm{d}x = \int_{\mathbb{R}^d} p(f^{-1}(y))\,\mathrm{d}y. \tag{4}$$

where $J_f(x) = \left|\det \frac{\partial f(x)}{\partial x^{\mathrm{T}}}\right|$.

A slightly more general phrasing of Equation 4 is

$$\int_{f^{-1}(\mathcal{B})} g(x) J_f(x)\,\mathrm{d}x = \int_{\mathcal{B}} g(f^{-1}(y))\,\mathrm{d}y. \tag{5}$$

where $g\colon \mathbb{R}^d \to \mathbb{R}$ is any non-negative measurable function, and $\mathcal{B} \subseteq \mathbb{R}^d$ is any measurable subset of $\mathbb{R}^d$.

We can extend Equation 5 to work in the case that f is not invertible. But to do this, we must address two issues. First, if f is not invertible, then $f^{-1}(y)$ is not a single point but rather a set. Second, if f is not invertible, then the Jacobian matrix $\frac{\partial f(x)}{\partial x^{\mathrm{T}}}$ may not be square, and thus has no well defined determinant. Both issues can be resolved and lead to the following change of variables theorem (Krantz & Parks, 2009), which is based on the so-called coarea formula (Federer, 1969).

Theorem 4. *Let $f\colon \mathbb{R}^d \to \mathbb{R}^r$ with $r \leq d$ be a differentiable function, $g\colon \mathbb{R}^d \to \mathbb{R}$ a non-negative measurable function, $\mathcal{B} \subseteq \mathbb{R}^d$ a measurable set, and $J_f(x) = \sqrt{\det\left(\frac{\partial f(x)}{\partial x^{\mathrm{T}}}\left(\frac{\partial f(x)}{\partial x^{\mathrm{T}}}\right)^{\mathrm{T}}\right)}$. Then*

$$\int_{f^{-1}(\mathcal{B})} g(x) J_f(x)\,\mathrm{d}x = \int_{\mathcal{B}} \int_{f^{-1}(y)} g(x)\,\mathrm{d}\mathscr{H}^{d-r}(x)\,\mathrm{d}y. \tag{6}$$

where $\mathscr{H}^{d-r}$ is the $(d-r)$-dimensional Hausdorff measure (one can think of this as a measure for lower-dimensional structures in high-dimensional space, e.g., the area of 2-dimensional surfaces in 3-dimensional space).[3]

We see in Theorem 4 that Equation 6 looks a lot like Equations 4 and 5, but with $f^{-1}(y)$ replaced by an integral over the set $f^{-1}(y)$, which for almost every y is a $(d-r)$-dimensional set. And if f in Equation 6 happens to be bijective, Equation 6 reduces to Equation 5.

[3] In what follows, we will sometimes replace g by g/J_f such that the Jacobian appears on the right-hand side. Furthermore, we will not only use non-negative g. This can be justified by splitting g into positive and negative parts provided that either part results in a finite integral.

We also see that the Jacobian determinant in Equation 5 was replaced by the so-called r-dimensional Jacobian

$$\sqrt{\det\left(\frac{\partial f(x)}{\partial x^{\mathrm{T}}}\left(\frac{\partial f(x)}{\partial x^{\mathrm{T}}}\right)^{\mathrm{T}}\right)}$$

in Equation 6. A word of caution is in order, as the r-dimensional Jacobian does not have the same nice properties for concatenated functions as does the Jacobian in the bijective case. In particular, we cannot calculate $J_{f_2\circ f_1}$ based on the values of J_{f_1} and J_{f_2} because the product $\frac{\partial f_2(x)}{\partial x^{\mathrm{T}}}\frac{\partial f_1(x)}{\partial x^{\mathrm{T}}}\left(\frac{\partial f_2(x)}{\partial x^{\mathrm{T}}}\frac{\partial f_1(x)}{\partial x^{\mathrm{T}}}\right)^{\mathrm{T}}$ does not decompose into a product of $\frac{\partial f_2(x)}{\partial x^{\mathrm{T}}}\left(\frac{\partial f_2(x)}{\partial x^{\mathrm{T}}}\right)^{\mathrm{T}}$ and $\frac{\partial f_1(x)}{\partial x^{\mathrm{T}}}\left(\frac{\partial f_1(x)}{\partial x^{\mathrm{T}}}\right)^{\mathrm{T}}$. In other words, the trick used in techniques like normalizing flows and NICE to compute determinants of deep networks for the change of variables formula by decomposing the network's Jacobian into the product of layerwise Jacobians does not work straightforwardly in the case of non-invertible mappings.

7.4. Conserved Differential Information Motivation

First, we present an alternative definition of conditional entropy that is meaningful for singular distributions (e.g., the joint distribution $p(X, f(X))$ for a function f). More information on this definition can be found in (Koliander et al., 2016).

7.4.1. Singular Conditional Entropy

Assume that the random variable X has a probability density function $p_X(x)$ on $\mathbb{R}^d$. For a given differentiable function $f\colon \mathbb{R}^d \to \mathbb{R}^r$ ($r \leq d$), we want to analyze the conditional differential entropy $H(X|f(X))$. Following (Koliander et al., 2016), we define this quantity as:

$$H(X|f(X)) = -\int_{\mathbb{R}^r} p_{f(X)}(y) \int_{f^{-1}(y)} \theta^{d-r}_{\Pr\{X\in\cdot|f(X)=y\}}(x) \log\left(\theta^{d-r}_{\Pr\{X\in\cdot|f(X)=y\}}(x)\right) \mathrm{d}\mathscr{H}^{d-r}(x)\,\mathrm{d}y \tag{7}$$

where $\mathscr{H}^{d-r}$ denotes $(d-r)$-dimensional Hausdorff measure. The function $p_{f(X)}$ is the probability density function of the random variable $f(X)$. Although $\theta^{d-r}_{\Pr\{X\in\cdot|f(X)=y\}}$ can also be interpreted as a probability density, it is not the commonly used density with respect to Lebesgue measure (which does not exist for $X|f(X)=y$) but a density with respect to a lower-dimensional Hausdorff measure. We will analyze the two functions $p_{f(X)}$ and $\theta^{d-r}_{\Pr\{X\in\cdot|f(X)=y\}}$ in more detail. The density $p_{f(X)}$ is defined by the relation

$$\int_{f^{-1}(\mathcal{B})} p_X(x)\,\mathrm{d}x = \int_{\mathcal{B}} p_{f(X)}(y)\,\mathrm{d}y\,, \tag{8}$$

which has to hold for every measurable set $\mathcal{B}\subseteq\mathbb{R}^r$. Using the coarea formula (or the related change-of-variables theorem), we see that

$$\int_{f^{-1}(\mathcal{B})} p_X(x)\,\mathrm{d}x = \int_{\mathcal{B}}\int_{f^{-1}(y)} \frac{p_X(x)}{J_f(x)}\,\mathrm{d}\mathscr{H}^{d-r}(x)\,\mathrm{d}y\,, \tag{9}$$

where $J_f(x) = \sqrt{\det\left(\frac{\partial f(x)}{\partial x^{\mathrm{T}}}\left(\frac{\partial f(x)}{\partial x^{\mathrm{T}}}\right)^{\mathrm{T}}\right)}$ is the r-dimensional Jacobian determinant. Thus, we identified

$$p_{f(X)}(y) = \int_{f^{-1}(y)} \frac{p_X(x)}{J_f(x)}\,\mathrm{d}\mathscr{H}^{d-r}(x)\,. \tag{10}$$

The second function, namely $\theta^{d-r}_{\Pr\{X\in\cdot|f(X)=y\}}$, is the Radon-Nikodym derivative of the conditional probability $\Pr\{X\in\cdot|f(X)=y\}$ with respect to $\mathscr{H}^{d-r}$ restricted to the set where $X|f(X)=y$ has positive probability (in the end, this will be the set $f^{-1}(y)$). To understand this function, we have to know something about the conditional distribution of X given $f(X)$. Formally, a (regular) conditional probability $\Pr\{X\in\cdot|f(X)=y\}$ has to satisfy three conditions:

- $\Pr\{X\in\cdot|f(X)=y\}$ is a probability measure for each fixed $y\in\mathbb{R}^r$.

- $\Pr\{X \in \mathcal{A} | f(X) = \cdot\}$ is measurable for each fixed measurable set $\mathcal{A} \subseteq \mathbb{R}^d$.
- For measurable sets $\mathcal{A} \subseteq \mathbb{R}^d$ and $\mathcal{B} \subseteq \mathbb{R}^r$, we have

$$\Pr\{(X, f(X)) \in \mathcal{A} \times \mathcal{B}\} = \int_{\mathcal{B}} \Pr\{X \in \mathcal{A} | f(X) = y\} p_{f(X)}(y) \, \mathrm{d}y \,. \tag{11}$$

In our setting, (11) becomes

$$\int_{\mathcal{A} \cap f^{-1}(\mathcal{B})} p_X(x) \, \mathrm{d}x = \int_{\mathcal{B}} \Pr\{X \in \mathcal{A} | f(X) = y\} p_{f(X)}(y) \, \mathrm{d}y \,. \tag{12}$$

Choosing

$$\Pr\{X \in \mathcal{A} | f(X) = y\} = \frac{1}{p_{f(X)}(y)} \int_{\mathcal{A} \cap f^{-1}(y)} \frac{p_X(x)}{J_f(x)} \, \mathrm{d}\mathscr{H}^{d-r}(x) \,, \tag{13}$$

the right-hand side in (12) becomes

$$\begin{aligned} \int_{\mathcal{B}} \Pr\{X \in \mathcal{A} | f(X) = y\} p_{f(X)}(y) \, \mathrm{d}y &= \int_{\mathcal{B}} \int_{\mathcal{A} \cap f^{-1}(y)} \frac{p_X(x)}{J_f(x)} \, \mathrm{d}\mathscr{H}^{d-r}(x) \, \mathrm{d}y \\ &= \int_{\mathcal{A} \cap f^{-1}(\mathcal{B})} p_X(x) \, \mathrm{d}x \,, \end{aligned} \tag{14}$$

where the final equality is again an application of the coarea formula. Thus, we identified

$$\theta^{d-r}_{\Pr\{X \in \cdot | f(X) = y\}}(x) = \frac{p_X(x)}{J_f(x) \, p_{f(X)}(y)} \,. \tag{15}$$

Although things might seem complicated up to this point, they simplify significantly once we put everything together. In particular, inserting (15) into (7), we obtain

$$\begin{aligned} H(X|f(X)) &= -\int_{\mathbb{R}^r} p_{f(X)}(y) \int_{f^{-1}(y)} \frac{p_X(x)}{J_f(x) \, p_{f(X)}(y)} \log\left(\frac{p_X(x)}{J_f(x) \, p_{f(X)}(y)}\right) \mathrm{d}\mathscr{H}^{d-r}(x) \, \mathrm{d}y \\ &= -\int_{\mathbb{R}^r} \int_{f^{-1}(y)} \frac{p_X(x)}{J_f(x)} \log\left(\frac{p_X(x)}{J_f(x) \, p_{f(X)}(y)}\right) \mathrm{d}\mathscr{H}^{d-r}(x) \, \mathrm{d}y \\ &= -\int_{\mathbb{R}^d} p_X(x) \log\left(\frac{p_X(x)}{J_f(x) \, p_{f(X)}(f(x))}\right) \mathrm{d}x \qquad (16) \\ &= H(X) + \int_{\mathbb{R}^d} p_X(x) \log\big(J_f(x) p_{f(X)}(f(x))\big) \, \mathrm{d}x \\ &= H(X) + \int_{\mathbb{R}^d} p_X(x) \log\big(p_{f(X)}(f(x))\big) \, \mathrm{d}x + \int_{\mathbb{R}^d} p_X(x) \log\big(J_f(x)\big) \, \mathrm{d}x \\ &= H(X) + \int_{\mathbb{R}^r} \int_{f^{-1}(y)} \frac{p_X(x)}{J_f(x)} \log\big(p_{f(X)}(f(x))\big) \, \mathrm{d}\mathscr{H}^{d-r}(x) \, \mathrm{d}y + \mathbb{E}\big[\log\big(J_f(X)\big)\big] \qquad (17) \\ &= H(X) + \int_{\mathbb{R}^r} \int_{f^{-1}(y)} \frac{p_X(x)}{J_f(x)} \, \mathrm{d}\mathscr{H}^{d-r}(x) \log\big(p_{f(X)}(y)\big) \, \mathrm{d}y + \mathbb{E}\big[\log\big(J_f(X)\big)\big] \\ &= H(X) + \int_{\mathbb{R}^r} p_{f(X)}(y) \log\big(p_{f(X)}(y)\big) \, \mathrm{d}y + \mathbb{E}\big[\log\big(J_f(X)\big)\big] \\ &= H(X) - H(f(X)) + \mathbb{E}\big[\log\big(J_f(X)\big)\big] \qquad (18) \end{aligned}$$

where (16) and (17) hold by the coarea formula.

So, altogether we have that for a random variable X and a function f, the singular conditional entropy between X and $f(X)$ is

$$H(X|f(X)) = H(X) - H(f(X)) + \mathbb{E}\big[\log\big(J_f(X)\big)\big]. \tag{19}$$

This quantity can loosely be interpreted as being the difference in differential entropies between X and $f(X)$ but with an additional term that corrects for any "uninformative" scaling that f does.

7.4.2. CONSERVED DIFFERENTIAL INFORMATION

For random variables that are not related by a deterministic function, mutual information can be expanded as

$$I(X,Y) = H(X) - H(X|Y) \tag{20}$$

where $H(X)$ and $H(X|Y)$ are differential entropy and conditional differential entropy, respectively. As we would like to measure information between random variables that are deterministically dependent, we can mimic this behavior by defining for a Lipschitz continuous mapping f:

$$C(X, f(X)) := H(X) - H(X|f(X))\,. \tag{21}$$

By (18), this can be simplified to

$$C(X, f(X)) = H(f(X)) - \mathbb{E}\big[\log\big(J_f(X)\big)\big] \tag{22}$$

yielding our definition of CDI.

7.5. Proof of CDI Data Processing Inequality

CDI Data Processing Inequality (Theorem 1)

For Lipschitz continuous functions f and g with the same output space,

$$C(X, f(X)) \geq C(X, g(f(X))$$

with equality if and only if g is one-to-one almost everywhere.

Proof. We calculate the difference between $C(X, f(X))$ and $C(X, g(f(X)))$.

$$\begin{aligned}
&C(X, f(X)) - C(X, g(f(X))) && (23)\\
&\quad = H(f(X)) - \mathbb{E}_X\big[\log J_f(X)\big] - H(g(f(X))) + \mathbb{E}_X\big[\log J_{g\circ f}(X)\big]\\
&\quad = H(f(X)) - H(g(f(X))) + \mathbb{E}_X\left[\log \frac{J_g(f(X))J_f(X)}{J_f(X)}\right] && (24)\\
&\quad = -\mathbb{E}_X[\log p_{f(X)}(f(X))] + \mathbb{E}_X\left[\log\left(\sum_{z\in g^{-1}(g(f(X)))} \frac{p_{f(X)}(f(z))}{J_g(f(z))}\right)\right] + \mathbb{E}_X[\log J_g(f(X))] && (25)\\
&\quad = \mathbb{E}_X\left[\log\left(\frac{\sum_{z\in g^{-1}(g(f(X)))} \frac{p_{f(X)}(f(z))}{J_g(f(z))}}{\frac{p_{f(X)}(f(X))}{J_g(f(X))}}\right)\right] && (26)
\end{aligned}$$

where (24) holds because the Jacobian determinant $J_{g\circ f}$ can be decomposed as g has the same domain and codomain and (25) holds because the probability density function of $g(f(X))$ can be calculated as $p_{g(f(X))}(z) = \sum_{z\in g^{-1}(g(f(X)))} \frac{p_{f(X)}(f(z))}{J_g(f(z))}$ using a change of variables argument. The resulting term in (26) is clearly always nonnegative which proves the inequality.

To prove the equality statement, we first assume that (26) is zero. In this case, $\sum_{z\in g^{-1}(g(f(x)))} \frac{p_{f(X)}(f(z))}{J_g(f(z))} = \frac{p_{f(X)}(f(x))}{J_g(f(x))}$ almost everywhere. Of course, we also have that $p_{f(X)}(f(x)) > 0$ almost everywhere. Thus, there exists a set $\mathcal{A}$ of probability one such that $\sum_{z\in g^{-1}(g(f(x)))} \frac{p_{f(X)}(f(z))}{J_g(f(z))} = \frac{p_{f(X)}(f(x))}{J_g(f(x))}$ and $p_{f(X)}(f(x)) > 0$ for all $x \in \mathcal{A}$. In particular, the set $g^{-1}(g(f(x))) \cap \mathcal{A} = \{f(x)\}$ and hence g is one-to-one almost everywhere.

For the other direction, assume that there exists $\tilde{g}$ such that $\tilde{g}(g(f(x))) = f(x)$ almost everywhere. We can assume without loss of generality that $p_{f(X)}(f(x)) = 0$ for all x that do not satisfy this equation. Restricting the expectation in (26) to the values that satisfy $\tilde{g}(g(f(x))) = f(x)$ does not change the expectation and gives the value zero. □

7.6. Theorem 3 Only Holds in the Reverse Direction for Continuous X

The specific claim we are making is as follows:

Theorem 5. *Let X be a continuous random variable drawn according to a distribution $p(X|Y)$ determined by the discrete random variable Y. Let $\mathcal{F}$ be the set of measurable functions of X to any target space. If $f(X)$ is a minimal sufficient statistic of X for Y then*

$$f \in \arg\min_{S \in \mathcal{F}} I(X, S(X)) \\ s.t.\ I(S(X), Y) = \max_{S' \in \mathcal{F}} I(S'(X), Y). \tag{27}$$

However, there may exist a function f satisfying (27) such that $f(X)$ is not a minimal sufficient statistic.

Proof. First, we prove the forward direction. According to Lemma 1, $Z = f(X)$ is a sufficient statistic for Y if and only if $I(Z, Y) = I(X, Y) = \max_{S'} I(S'(X), Y)$. To show the minimality condition in (27) for a minimal sufficient statistic, assume that there exists $S(X)$ such that $I(S(X), Y) = \max_{S' \in \mathcal{F}} I(S'(X), Y)$ and $I(X, S(X)) < I(X, f(X))$. Because f is assumed to be a minimal sufficient statistic, there exists g such that $f(X) = g(S(X))$ and by the data-processing inequality $I(X, S(X)) \geq I(X, f(X))$, a contradiction.

Next, we give an example of a function satisfying (27) such that $f(X)$ is not a minimal sufficient statistic. The example is the case when $I(X, f(X))$ is not finite, as is the case when f is a deterministic function and X is continuous. (See Lemma 3.) In this case, $I(X, S(X))$ is infinite for all deterministic, sufficient statistics S. Thus the set $\arg\min_S I(X, S(X))$ contains not only the minimal sufficient statistics, but all deterministic sufficient statistics. As a concrete example, consider two i.i.d. normally-distributed random variables with mean μ: $X = (X_1, X_2) \sim \mathcal{N}(\mu, 1)$. $T(X) = \frac{X_1 + X_2}{2}$ is a minimal sufficient statistic for μ. $T'(X) = (\frac{X_1 + X_2}{2}, X_1 \cdot X_2)$ is a non-minimal sufficient statistic for μ. However, both statistics satisfy $T, T' \in \arg\min_{S \in \mathcal{F}} I(X, S(X))$ since $\min_{S \in \mathcal{F}} I(X, S(X)) = \infty$ under the constraint $I(S(X), Y) = \max_{S' \in \mathcal{F}} I(S'(X), Y)$. □

7.7. Experiment Details

Code to reproduce all experiments is available online at `https://github.com/mwcvitkovic/MASS-Learning`.

7.7.1. Data

In all experiments above, the models were trained on the CIFAR-10 dataset (Krizhevsky, 2009). In the out-of-distribution detection experiments, the SVHN dataset (Netzer et al., 2011) was used as the out-of-distribution dataset. All channels in all datapoints were normalized to have zero mean and unit variance across their dataset. No data augmentation was used in any experiments.

7.7.2. Networks

The SmallMLP network is a 2-hidden-layer, fully-connected network with `elu` nonlinearities (Clevert et al., 2015). The first hidden layer contains 400 hidden units; the second contains 200 hidden units. Batch norm was applied after the linear mapping and before the nonlinearity of each hidden layer. Dropout, when used, was applied after the nonlinearity of each hidden layer. When used in VIB and MASS, the representation $f(X)$ was in $\mathbb{R}^{15}$. The marginal distribution in VIB and each component of the variational distribution q_ϕ (one component for each possible output class) in MASS were both mixtures of 10 full-covariance, 15-dimensional multivariate Gaussians.

The ResNet20 network is the 20-layer residual net of He et al. (2015). We adapted our implementation from `https://github.com/akamaster/pytorch_resnet_cifar10`, to whose authors we are very grateful. When used in VIB and MASS, the representation $f(X)$ was in $\mathbb{R}^{20}$. The marginal distribution in VIB and each component of the the variational distribution q_ϕ (one component for each possible output class) in MASS were both mixtures of 10 diagonal-covariance, 20-dimensional multivariate Gaussians.

7.7.3. Training

The SmallMLP network in all experiments and with all training methods was trained using the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 0.0005 for 60,000 iterations of stochastic gradient descent, using minibatches of size 256. All quantities we report in this paper were fully-converged to stable values by 60,000 iterations. When training VIB, 5 encoder samples per datapoint were used during training, and 10 during testing. When training MASS, the learning rate of the parameters of the variational distribution q_ϕ was set at $2.5e-5$ to aid numerical stability.

The ResNet20 network in all experiments and with all training methods was trained using SGD with a learning rate of 0.002, a momentum factor of 0.9, and minibatches of size 128. These values follow the training method reported in the original paper (He et al., 2015). However, unlike the original paper, we did not use a learning rate schedule during training. This, combined with the absence of any data augmentation and the smaller number of training points used, presumably explains the lower accuracy we observe on CIFAR-10 (around 68%) compared to the original paper (around 91%). We trained the network for 60,000 iterations. All quantities we report in this paper were fully-converged to stable values by 60,000 iterations. When training VIB, 5 encoder samples per datapoint were used during training, and 10 during testing. When training MASS, the learning rate of the parameters of the variational distribution q_ϕ was set at 0.002.

The values of β we chose for VIB and MASS were selected so that the largest β value used was much larger in magnitude than the remaining terms in the VIB or MASS training loss, and the smallest β value used was much smaller than the remaining terms. We made this choice in the hope of clearly observing the effect of the β parameter and more fairly comparing SoftmaxCE, VIB, and MASS. But we note that a finer-tuning of the β parameter would likely result in better performance for both VIB and MASS. We also note that the reason we omit a $\beta = 0$ run for VIB with the SmallMLP network was that we could not prevent training from failing with $\beta = 0$ with this network.