A THEORY OF USABLE INFORMATION UNDER COMPUTATIONAL CONSTRAINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a new framework for reasoning about information in complex systems. Our foundation is based on a variational extension of Shannon's information theory that takes into account the modeling power and computational constraints of the observer. The resulting *predictive $\mathcal{F}$-information* encompasses mutual information and other notions of informativeness such as the coefficient of determination. Unlike Shannon's mutual information and in violation of the data processing inequality, $\mathcal{F}$-information can be created through computation. This is consistent with deep neural networks extracting hierarchies of progressively more informative features in representation learning. Additionally, we show that by incorporating computational constraints, $\mathcal{F}$-information can be reliably estimated from data even in high dimensions with PAC-style guarantees. Empirically, we demonstrate predictive $\mathcal{F}$-information is more effective than mutual information for structure learning and fair representation learning.

1 INTRODUCTION

Extracting actionable *information* from noisy, possibly redundant, and high-dimensional data sources is a key computational and statistical challenge at the core of AI and machine learning. Information theory, which lies at the foundation of AI and machine learning, provides a conceptual framework to characterize information in a mathematically rigorous sense (Shannon & Weaver, 1948; Cover & Thomas, 1991). However, important computational aspects are not considered in information theory. To illustrate this, consider a dataset of encrypted messages intercepted from an opponent. According to information theory, these encrypted messages have high mutual information with the opponent's plans. Indeed, with infinite computation, the messages can be decrypted and the plans revealed. Modern cryptography originated from this observation by Shannon that perfect secrecy is (essentially) impossible if the adversary is computationally unbounded (Shannon & Weaver, 1948). This motivated cryptographers to consider restricted classes of adversaries that have access to limited computational resources (Pass & Shelat, 2010). More generally, it is known that information theoretic quantities can be expressed in terms of betting games (Cover & Thomas, 1991). For example, the (conditional) entropy of a random variable X is directly related to how predictable X is in a certain betting game, where an agent is rewarded for correct guesses. Yet, the standard definition unrealistically assumes agents are computationally unbounded, i.e., they can employ arbitrarily complex prediction schemes.

Leveraging modern ideas from variational inference and learning (Ranganath et al., 2013; Kingma & Welling, 2013; LeCun et al., 2015), we propose an alternative formulation based on realistic computational constraints that is in many ways closer to our intuitive notion of information, which we term *predictive $\mathcal{F}$-information*. Without constraints, predictive $\mathcal{F}$-information specializes to classic mutual information. Under natural restrictions, $\mathcal{F}$-information specializes to other well-known notions of predictiveness, such as the coefficient of determination (R^2). A consequence of this new formulation is that computation can "create usable information" (e.g., by decrypting the intercepted messages), invalidating the famous data processing inequality. This generalizes the idea that clever feature extraction enables prediction with extremely simple (e.g., linear) classifiers, a key notion in modern representation and deep learning (LeCun et al., 2015).

As an additional benefit, we show that predictive $\mathcal{F}$-information can be estimated with statistical guarantees using the Probably Approximately Correct framework (Valiant, 1984). This is in sharp contrast with Shannon information, which is well known to be difficult to estimate for high dimensional or continuous random variables (Battiti, 1994). Theoretically we show that the statistical guarantees of estimating $\mathcal{F}$ information translate to statistical guarantees for a variant of the Chow-Liu algorithm for structure learning. In practice, when the observer employs deep neural networks as a prediction scheme, $\mathcal{F}$-information outperforms methods that approximate Shannon information in various applications, including Chow-Liu tree contruction in high dimension and gene regulatory network inference.

2 Definitions and Notations

To formally define the predictive $\mathcal{F}$-information, we begin with a formal model of a computationally bounded agent trying to predict the outcome of a real-valued random variable Y; the agent is either provided another real-valued random variable X as side information, or provided no side information $\varnothing$. We use $\mathcal{X}$ and $\mathcal{Y}$ to denote the samples spaces of X and Y respectively (while assuming they are separable), and use $\mathcal{P}(\mathcal{X})$ to denote the set of all probability measures over the Borel algebra on $\mathcal{X}$ ($\mathcal{P}(\mathcal{Y})$ similarly defined for $\mathcal{Y}$).

Definition 1 (Predictive Family [1])**.** *Let $\Omega = \{f : \mathcal{X} \cup \{\varnothing\} \to \mathcal{P}(\mathcal{Y})\}$. We say that $\mathcal{F} \subseteq \Omega$ is a predictive family if it satisfies*

$$\forall f \in \mathcal{F}, \forall P \in \text{range}(f), \quad \exists f' \in \mathcal{F}, \quad s.t. \quad \forall x \in \mathcal{X}, f'[x] = P, f'[\varnothing] = P \tag{1}$$

A predictive family is a set of predictive models the agent is allowed to use, e.g., due to computational or statistical constraints. We refer to the additional condition in Eq.(1) as *optional ignorance*. Intuitively, it means that the agent can, in the context of the prediction game we define next, ignore the side information if she chooses to.

Definition 2 (Predictive conditional $\mathcal{F}$-entropy)**.** *Let X, Y be two random variables taking values in $\mathcal{X} \times \mathcal{Y}$, and $\mathcal{F}$ be a predictive family. Then the predictive conditional $\mathcal{F}$-entropy is defined as*

$$H_{\mathcal{F}}(Y|X) = \inf_{f \in \mathcal{F}} \mathbb{E}_{x,y \sim X,Y} \left[-\log f[x](y)\right]$$
$$H_{\mathcal{F}}(Y|\varnothing) = \inf_{f \in \mathcal{F}} \mathbb{E}_{y \sim Y} \left[-\log f[\varnothing](y)\right]$$

We additionally call $H_{\mathcal{F}}(Y|\varnothing)$ the $\mathcal{F}$-entropy, and also denote it as $H_{\mathcal{F}}(Y)$

In our notation f is a function $\mathcal{X} \cup \{\varnothing\} \to \mathcal{P}(\mathcal{Y})$, so $f[x] \in \mathcal{P}(\mathcal{Y})$ is a probability measure on $\mathcal{Y}$ chosen based on the received side information x (we use $f[\cdot]$ instead of the more conventional $f(\cdot)$); and $f[x](y) \in \mathbb{R}$ is the value of the density evaluated at $y \in \mathcal{Y}$. Intuitively, $\mathcal{F}$ (conditional) entropy is the smallest expected negative log-likelihood that can be achieved predicting Y given observation (side information) X (or no side information $\varnothing$), using models from $\mathcal{F}$. Eq.(1) means that whenever the agent can use P to predict $\mathcal{Y}$'s outcomes, it has the option to ignore the input, and use P no matter whether X is observed or not.

Definition 2 generalizes several known definitions of uncertainty. For example, as shown in proposition 2, if the $\mathcal{F}$ is the largest possible predictive family that includes all possible models, i.e. $\mathcal{F} = \Omega$, then Definition 2 reduces to Shannon entropy: $H_{\Omega}(Y|X) = H(Y|X)$ and $H_{\mathcal{F}}(Y|\varnothing) = H_{\Omega}(Y) = H(Y)$. By choosing more restrictive families $\mathcal{F}$, we recover several other notions of uncertainty such as trace of covariance, as will be shown in Proposition 1.

Shannon mutual information is a measure of changes in entropy when conditioning on new variables:

$$I(X;Y) = H(Y) - H(Y|X) = H_{\Omega}(Y) - H_{\Omega}(Y|X) \tag{2}$$

Here, we will use predictive $\mathcal{F}$-entropy to define an analogous quantity, $I_{\mathcal{F}}(X \to Y)$, to represent the *change* in predictability of an output variable Y when given side information X.

Definition 3 (Predictive $\mathcal{F}$-information)**.** *Let X, Y be two random variables taking values in $\mathcal{X} \times \mathcal{Y}$, and $\mathcal{F}$ be a predictive family. The predictive $\mathcal{F}$-information from X to Y is defined as*

$$I_{\mathcal{F}}(X \to Y) = H_{\mathcal{F}}(Y|\varnothing) - H_{\mathcal{F}}(Y|X) \tag{3}$$

[1] Regularity Conditions: To minimize technical overhead we restrict out discussion only to distributions with probability density functions (PDF) or probability mass functions (PMF) with respect to the underlying measure. Also $\varnothing \notin \mathcal{X}$.

2.1 Important Special Cases

Several important notions of uncertainty and predictiveness are special cases of our definition. Note that when we are defining $\mathcal{F}$-entropy of a random variable Y in sample space $\mathcal{Y} \in \mathbb{R}^d$ (without side information), out of convenience we can assume $\mathcal{X}$ is empty $\mathcal{X} = \varnothing$ (this does not violate our requirement that $\varnothing \notin \mathcal{X}$.)

Proposition 1. *For $\mathcal{F}$-entropy and $\mathcal{F}$-information, we have:*

1. *Let Ω be as in Def. 2. Then $H_\Omega(Y)$ is the Shannon entropy, $H_\Omega(Y \mid X)$ is the Shannon conditional entropy, and $I_\Omega(Y \to X)$ is the mutual information.*
2. *Let $\mathcal{F} = \{f : \{\varnothing\} \to \frac{1}{Z} e^{-\|y-\mu\|_2} \mid \mu \in \mathbb{R}^d\}$, where $Z = \int e^{-\|y-\mu\|_2} dy$, then the $\mathcal{F}$-entropy of a random variable Y equals its median absolute deviation, up to a constant.*
3. *Let $\mathcal{F} = \{f : \{\varnothing\} \to \mathcal{N}(\mu, \Sigma) \mid \mu \in \mathbb{R}^d, \Sigma = 1/2 I_{d\times d}\}$, then the $\mathcal{F}$-entropy of a random vector Y equals to the trace of its covariance $\operatorname{tr}(\operatorname{Cov}(Y))$, up to a constant.*
4. *Let $\mathcal{F} = \{f : \{\varnothing\} \to h(y) \exp(\theta \cdot \mathbf{t}(y) - A(\theta)), \theta \in \Theta\}$, where the range is a set of minimal exponential families with sufficient statistics $\mathbf{t}(y) : \mathcal{Y} \to \mathbb{R}^d_t$ and $\Theta = \{\theta \in \mathbb{R}^d_t | A(\theta) < +\infty\}$. Then $\mathcal{F}$-entropy of Y is the maximum Shannon entropy for any random variable with expected sufficient statistics $\mu_Y = \mathbb{E}_{y\sim Y}[\mathbf{t}(y)]$.*
$$H_\mathcal{F}(Y) = \sup_{\hat{Y} | \mathbb{E}_{y\sim\hat{Y}}[\mathbf{t}(y)] = \mu_Y} H(\hat{Y})$$
More specifically it is equal to $\mathbb{E}_{y\sim Y}[\log h(y)] - A^*(\mu_Y)$, where and A^* is the Fenchel conjugate of A.
5. *Let $\mathcal{F} = \{f : x \mapsto \mathcal{N}(\mathbf{t}(x), \Sigma), x \in \mathcal{X}; \varnothing \mapsto \mathcal{N}(\mu, \Sigma) | \mu \in range(\mathbf{t}) = \mathbb{R}^d; \Sigma = \operatorname{tr}(\operatorname{Cov}(Y))/2I_{d\times d}\}$, then $\mathcal{F}$-information $I_\mathcal{F}(X \to Y)$ equals to the coefficient of determination , R^2, with function $\mathbf{t}(x)$ as the regression model.*

The trace of covariance represents a natural notion of uncertainty – for example, a random variable with zero variance (when $d = 1$,$\operatorname{tr}(\operatorname{Cov}(Y)) = \operatorname{Var}(Y)$)) is trivial to predict. Proposition 1.3 shows that the trace of covariance corresponds to a notion of surprise (in the Shannon sense) for an agent *restricted* to make predictions using certain Gaussian models. More broadly, a similar analogy can be drawn for other exponential families of distributions. In the same spirit, the coefficient of determination, also known as the fraction of variance explained, represents a natural notion of informativeness for computationally bounded agents. Also note that in the case of Proposition 1.4, the $\mathcal{F}$-entropy is invariant if the expected sufficient statistics remain the same.

3 Properties of $\mathcal{F}$-information

3.1 Elementary Properties

We first show several elementary properties of $\mathcal{F}$-entropy and $\mathcal{F}$-information. In particular, $\mathcal{F}$-information preserves many properties of Shannon information that are desirable in a machine learning context. For example, mutual information (and $\mathcal{F}$-information) should be non-negative as conditioning on additional side information X should not reduce an agent's ability to predict Y.

Proposition 2. *Let Y and X be any random variables on $\mathcal{Y}$ and $\mathcal{X}$, and $\mathcal{F}$ and $\mathcal{G}$ be any predictive families, then we have*

1. ***Monotonicity**: If $\mathcal{F} \subseteq \mathcal{G}$, then $H_\mathcal{F}(Y) \geq H_\mathcal{G}(Y)$, $H_\mathcal{F}(Y \mid X) \geq H_\mathcal{G}(Y \mid X)$.*
2. ***Non-Negativity**: $I_\mathcal{F}(X \to Y) \geq 0$.*
3. ***Independence**: If X is independent of Y, $I_\mathcal{F}(X \to Y) = I_\mathcal{F}(Y \to X) = 0$.*

The *optional ignorance* requirement in Eq.(1) is a technical condition needed for these properties to hold. Intuitively, it guarantees that conditioning on side information does not restrict the class of densities the agent can use to predict Y. This property is satisfied by many existing machine learning models, often by setting some weights to zero so that an input is effectively ignored.

3.2 On the production of information through preprocessing

The Data Processing Inequality guarantees that computing on data cannot increase its mutual information with other random variables. Formally, letting $t : \mathcal{X} \to \mathcal{X}$ be any function, $t(X)$ cannot have higher mutual information with Y than X: $I(t(X); Y) \leq I(X; Y)$. But is this property desirable? In analyzing optimal communication, yes - it demonstrates a fundamental limit to the number of bits that can be transmitted through a communication channel. However, we argue that in machine learning settings this property is less appropriate.

Consider an RSA encryption scheme where the public key is known. Given plain text and its corresponding encrypted text X, if we have infinite computation, we can perfectly compute one from the other. Therefore, the plain text and the encrypted text should have identical Shannon mutual information with respect to any label Y we want to predict. However, to any human (or machine learning algorithm), it is certainly easier to predict the label from the plain text than the encrypted text. In other words, decryption increases a human's ability to predict the label: processing increases the "usable information". More formally, denoting t as the decryption algorithm and $\mathcal{F}$ as a class of natural language processing functions, we have that: $I_\mathcal{F}(t(X) \to Y) > I_\mathcal{F}(X \to Y) \approx 0$.

As another example, consider the mutual information between an image's pixels and its label. Due to data processing inequality, we cannot expect to use a function to map raw pixels to "features" that have higher mutual information with the label. However, the fundamental principle of representation learning is precisely the ability to learn predictive features — functions of the raw inputs that enable predictions with higher accuracy. This is a direct contradiction to the guiding principle of Shannon information: processing does increase "usable information" ($\mathcal{F}$-information).

3.3 On the asymmetry of predictive $\mathcal{F}$-Information

$\mathcal{F}$-information also captures the intuition that sometimes, it is easy to predict Y from X but not vice versa. In fact, modern cryptography is founded on the assumption that certain functions $h : \mathcal{X} \to \mathcal{Y}$ are one-way, meaning that there exists an polynomial algorithm to compute $h(x)$ but no polynomial algorithm to compute $h^{-1}(y)$. This means that if $\mathcal{F}$ contains all polynomial-time computable functions, then $I_\mathcal{F}(X \to h(X)) \gg I_\mathcal{F}(h(X) \to X)$.

This property is also reasonable in the machine learning context. For example, several important methods for causal discovery (Peters et al., 2017) rely on this asymmetry: if X causes Y, then usually it is easier to predict Y from X than vice versa; another commonly used assumption is that $Y|X$ can be accurately modeled by a Gaussian distribution, while $X|Y$ cannot (Pearl, 2000).

4 Estimation of $\mathcal{F}$-information from data

For many practical applications of mutual information (e.g., structure learning), we do not know the joint distribution of X, Y, so cannot directly compute the mutual information. Instead we only have samples $\{(x_i, y_i)\}_{i=1}^N \sim X, Y$ and need to estimate mutual information from data.

Shannon information is notoriously difficult to estimate for high dimensional random variables. Although non-parametric estimators of mutual information exist (Kraskov et al., 2004; Darbellay & Vajda, 1999; Gao et al., 2017), these estimators do not scale to high dimensions. Several variational estimators for Shannon information have been recently proposed (van den Oord et al., 2018; Nguyen et al., 2010; Belghazi et al., 2018), but have two shortcomings: due to their variational assumptions, their bias/variance tradeoffs are poorly understood and they are still not efficient enough for high dimensional problems.

On the other hand, $\mathcal{F}$-information is explicit about the assumptions (as a feature instead of a bug). $\mathcal{F}$-information is also easy to estimate with guarantees if we can bound the complexity of $\mathcal{F}$ (such as its Radamacher or covering number complexity) As we will show, bounds on the complexity of $\mathcal{F}$ directly translate to PAC bounds for $\mathcal{F}$-information estimation. In practice, we can efficiently optimize over $\mathcal{F}$, e.g., via gradient descent.

4.1 Inefficiency of Approximate Estimators for Shannon Information

We theoretically show that there are inherent limitations for the two commonly used high dimension mutual information estimators due to the large bias or variance. We analysis the CPC (or InfoNCE in

Poole et al. (2018)) estimator (I_{CPC}) proposed by van den Oord et al. (2018) and the NWJ estimator (I_{NWJ}) proposed by Nguyen et al. (2010). The I_{CPC} estimator, as shown in van den Oord et al. (2018) and Poole et al. (2018), saturates at $\log N$, where N is typically the dataset size. For the I_{NWJ} estimator, we show in Theorem 5 (in Appendix) that its variance is at least $\frac{e^{I(X;Y)-3}}{N}$, i.e., *exponential* in the true amount of mutual information. The theoretical analyses are consistent with the empirical results in Poole et al. (2018). Formal statements and proofs are available in Appendix B.

4.2 PAC Guarantees for $\mathcal{F}$-Information Estimation

Contrary to the difficulty of estimating Shannon information (even with variational bounds), we can estimate $\mathcal{F}$ information with PAC guarantees. In this paper we will present the Rademacher complexity version; other complexity measures (such as covering number) can be derived similarly.

We define the function family $\mathcal{G}_{\mathcal{F}} = \{g|g(x,y) = f[x](y), f \in \mathcal{F}\}$, and denote the Rademacher complexity of $\mathcal{G}$ with sample number N as $\mathfrak{R}_N(\mathcal{G})$ (Bartlett & Mendelson, 2001). Let X, Y be two random variables taking values in $\mathcal{X}, \mathcal{Y}$ and $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N \sim X, Y$ denotes the set of samples drawn from the joint distribution over $\mathcal{X} \times \mathcal{Y}$. Take $\hat{f} = \arg\min_{f\in\mathcal{F}} \frac{1}{|\mathcal{D}|}\sum_{x_i,y_i\in\mathcal{D}} -\log f[x_i](y_i)$ and $\hat{f}^{\varnothing} = \arg\min_{f\in\mathcal{F}} \frac{1}{|\mathcal{D}|}\sum_{x_i,y_i\in\mathcal{D}} -\log f[\varnothing](y_i)$, we can estimate $\mathcal{F}$-information as

$$\hat{I}_{\mathcal{F}}(X \to Y; \mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{x_i,y_i\in\mathcal{D}} \log \hat{f}^{\varnothing}[\varnothing](y_i) - \log \hat{f}[x_i](y_i)$$

assuming the above optimization problems can be solved. Then we have the following:

Theorem 1. *Assume* $\forall f \in \mathcal{F}, x \in \mathcal{X}, y \in \mathcal{Y}, \log f[x](y) \in [-B, B]$. *Then for any* $\delta > 0$, *with probability at least* $1 - 2\delta$, *we have:*

$$\left|I_{\mathcal{F}}(X \to Y) - \hat{I}_{\mathcal{F}}(X \to Y; \mathcal{D})\right| \le 4\mathfrak{R}_{|\mathcal{D}|}(\log \circ \mathcal{G}_{\mathcal{F}}) + 2B\sqrt{\frac{\log\frac{1}{\delta}}{2|\mathcal{D}|}} \tag{4}$$

where the Rademacher complexity term can often be bounded by $\mathcal{O}(|\mathcal{D}|^{-\frac{1}{2}})$ (Bartlett & Mendelson, 2001; Gao & cheng Zhou, 2014). It's worth noticing that a complex function family $\mathcal{F}$ (i.e., with large Rademacher complexity) could lead to overfitting. On the other hand, an overly-simple $\mathcal{F}$ maybe not expressive enough to capture the relationship between X and Y. As an example of the theorem, we provide a concrete estimation bound when $\mathcal{F}$ is chosen to be linear functions mapping $\mathcal{X}$ to the mean of a Gaussian distribution. This was shown in Proposition 1 to be the coefficient of determination.

Corollary 1.1. *Assume* $\mathcal{X} = \{x \in \mathbb{R}^{d_x}, \|x\|_2 \le k_x\}$ *and* $\mathcal{Y} = \{y \in \mathbb{R}^{d_y}, \|y\|_2 \le k_y\}$. *If*

$$\mathcal{F} = \{f : f[x] = \mathcal{N}(Wx + b, I), f[\varnothing] = \mathcal{N}(c, I), W \in \mathbb{R}^{d_y \times d_x}, b, c \in \mathbb{R}^{d_y}, \|\begin{pmatrix} W \\ b \end{pmatrix}\|_2 \le 1\}$$

Denote $M = (k_x + k_y)^2 + d_y \log 2\pi/2$, *then* $\forall\delta > 0$, *with probability at least* $1 - 2\delta$*:*

$$\left|I_{\mathcal{F}}(X \to Y) - \hat{I}_{\mathcal{F}}(X \to Y; \mathcal{D})\right| \le \frac{M}{\sqrt{4|\mathcal{D}|}}\left(1 + 4\sqrt{\log\frac{1}{\delta}}\right)$$

Similar results can be obtained using other classes of machine learning models with known (Rademacher) complexity.

5 Structure learning with $\mathcal{F}$-information

Among many possible applications of $\mathcal{F}$-information, we show how to use it to perform structure learning with provable guarantees. The goal of structure learning is to learn a directed graphical model (Bayesian network) or undirected graphical model (Markov network) that best captures the (conditional) independence structure of an underlying data generating process. Structure learning is difficult in general, but if we restrict ourselves to certain set of graphs G, there are efficient algorithms.

In particular, the Chow-Liu algorithm (Chow & Liu, 1968) efficiently learns the tree graphs (i.e. G is the set of trees). Chow & Liu (1968) show that the problem can be reduced to:

$$g^* = \arg\max_{g \in G_{\text{tree}}} \sum_{(X_i, X_j) \in \text{edge}(g)} I(X_i, X_j) \tag{5}$$

where $I(X_i, X_j)$ is the Shannon mutual information between variables X_i and X_j. In other words, it suffices to construct the maximal weighted spanning tree where the weight between two vertices is their Shannon mutual information. Chow & Wagner (1973) show that the Chow-Liu algorithm is consistent , i.e, it recovers the true solution as the dataset size goes to infinity. However, the finite sample behavior of the Chow-Liu algorithm for high dimensional problems is much less studied, due to the difficulty of estimating mutual information. In fact, we show in our experiments that the empirical performance is often poor, even with state-of-the-art estimators. Also, methods based on mutual information cannot take advantage of intrinsically asymmetric relationships, which are common for example in gene regulatory networks (Meyer et al., 2007).

To address these two issues, we propose a new structure learning algorithm based on $\mathcal{F}$-information instead of Shannon information. The idea is that we can associate to each *directed* edge in G (i.e., each pair of variables) a suitable predictive family $\mathcal{F}_{i,j}$ (cf. Def 1). The main challenge is that we cannot simply replace mutual information with $\mathcal{F}$-information in Eq. 5 because $\mathcal{F}$-information is asymmetric – we now have to optimize over *directed* trees:

$$g^* = \arg\max_{g \in G_{\text{tree}}} \sum_{i=2}^{m} I_{\mathcal{F}_{t(g)(i),i}}(X_{t(g)(i)} \to X_i) \tag{6}$$

where G_{tree} is the set of directed trees, and $t(g) : \mathbb{N} \to \mathbb{N}$ is the function mapping each non-root node of directed tree g to its parent, and $\mathcal{F}_{i,j}$ is the predictive family for random variables X_i and X_j. After estimating $\mathcal{F}$-information on each edge, we use the Chu-Liu algorithm (Chu & Liu, 1965) to construct the *maximal directed spanning tree*. This allows us to solve (6) exactly, even though there is a combinatorially large number of trees to consider. Pseudocode is summarized in Algorithm 1 in Appendix. Denote $C(g) = \sum_{i=2}^{m} I_{\mathcal{F}_{t(g)(i),i}}(X_{t(g)(i)} \to X_i)$, we show in the following theorem that unlike the original Chow-Liu algorithm, our algorithm has guarantees in the finite samples regime, even in continuous settings:

Theorem 2. *Let $\{X_i\}_{i=1}^m$ be the set of m random variables, $\mathcal{D}_{i,j}$ (resp. $\mathcal{D}_j$) be the set of samples drawn from $P(X_i, X_j)$ (resp. $P(X_j)$). Denote the optimal tree with maximum $C(g)$ as g^* and the optimal tree constructed on the dataset $\mathcal{D}$ as g. Then with the assumption in theorem 1, for any $\delta > 0$, with probability at least $1 - 2m(m-1)\delta$, we have:*

$$C(g) \geq C(g^*) - 2(m-1) \max_{i,j} \left\{ 4\mathfrak{R}_{\mathcal{D}_{i,j}}(\log \circ \mathcal{G}_{\mathcal{F}_{i,j}}) + B\sqrt{\frac{\log \frac{1}{\delta}}{2}} (|\mathcal{D}_j|^{-\frac{1}{2}} + |\mathcal{D}_{i,j}|^{-\frac{1}{2}}) \right\} \tag{7}$$

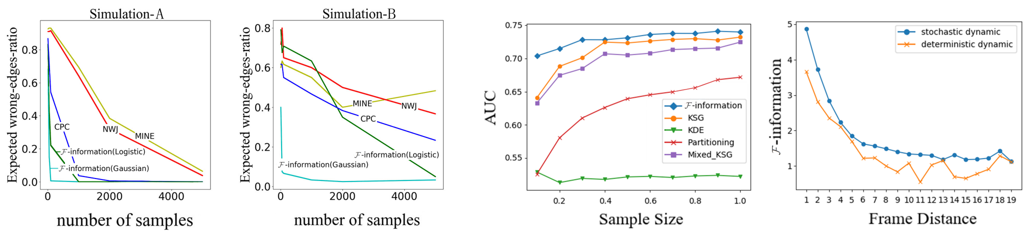
Theorem 2 shows that the total edge weights of the maximal directed spanning tree constructed by algorithm 1 would be close to the optimal total edge weights if the Rademacher term is small. Although larger $C(g)$ does not necessarily lead to better construction of Chow-Liu tree, empirically we find that the optimal tree in the sense of equation (6) is consistent with the optimal tree in equation (5) under commonly used $\mathcal{F}$.

6 Experimental results

6.1 Structure learning with continuous high-dimensional data

We generate synthetic data using various ground-truth tree structures g^* with between 7 and 20 variables, where each variable is 10-dimensional. We use Gaussians, Exponentials, and Uniforms as ground truth edge-conditionals. We use $\mathcal{F}$-information(Gaussian) and $\mathcal{F}$-information(Logistic) to denote Algorithm 1 with two different $\mathcal{F}$ families. Please refer to Appendix F.1 for more details. We compare with the original Chow-Liu algorithm equipped with state-of-the-art mutual information estimators: **CPC** (van den Oord et al., 2018), **NWJ** (Nguyen et al., 2010) and **MINE** (Belghazi et al., 2018), with the same neural network architecture as the $\mathcal{F}$-families for fair comparison. All the experiments are repeated for 10 times. As a performance metric, we use the wrong-edges-ratio (the ratio of edges that are different from ground truth) as a function of the amount of training data.

We show two illustrative experiments in figure 1a; please refer to Appendix F.1 for all simulations. We can see that *although the two $\mathcal{F}$-families used are misspecified with respect to the true underlying (conditional) distributions*, the estimated Chow-Liu trees are much more accurate across all data regimes, with **CPC** (blue) being the best alternative. Surprisingly, $\mathcal{F}$-information(Gaussian) works consistently well in all cases and only requires about 100 samples to recover the ground-truth Chow-Liu tree in simulation-A.



(a) Chow-Liu tree Construction (b) Gene network inference (c) $\mathcal{F}$-information of frames

Figure 1: (a) The expected wrong-edges-ratio of algorithm 1 with different $\mathcal{F}$ and other mutual information estimators-based algorithms from sample size 10 to 5×10^3. (b) AUC curve for gene regulatory network inference. (c) The predictive $\mathcal{F}$-information versus frame distance.

6.2 Gene regulatory network inference

Mutual information between pairs of gene expressions is often used to construct gene regulatory networks. We evaluate $\mathcal{F}$-information on the in-silico dataset from the DREAM5 challenge (Marbach et al., 2012) and use the setup of Gao et al. (2017), where 20 genes with 660 datapoints are utilized to evaluate all methods. We compare with state-of-the-art non-parametric Shannon mutual information estimators in this low dimensional setting: **KDE**, the traditional kernel density estimator; the **KSG** estimator (Kraskov et al., 2004); the **Mixed KSG** estimator (Gao et al., 2017) and **Partitioning**, an adaptive partitioning estimator (Darbellay & Vajda, 1999) implemented by Szabó (2014). For fair comparison with these low dimensional estimators, we select $\mathcal{F} = \{f : f[x] = \mathcal{N}(g(x), \frac{1}{2}), x \in \mathcal{X}; f[\varnothing] = \mathcal{N}(\mu, \frac{1}{2}) | \mu \in range(g)\}$, where g is a 3rd order polynomial.

The task is to predict whether a directed edge between genes exists in the ground-truth gene network. We use the estimated mutual information and $\mathcal{F}$-information for gene pairs as the test statistic to obtain the AUC for various methods. As shown in Figure 1b, our method outperforms all other methods in network inference under different fractions of data used for estimation. The natural information measure in this task is asymmetry since the goal is to find the pairs of genes (A_i, B_i)s in which A_i regulates B_i, thus $\mathcal{F}$-information is more suitable for such case than mutual information.

6.3 Recovering the order of video frames

Let $X_1, \cdots, X_{20}$ be random variables each representing a frame in videos from the Moving-MNIST dataset, which contains 10,000 sequences each of length 20 showing two digits moving with stochastic dynamics. Can Algorithm 1 be used to recover the natural (causal) order of the frames? Intuitively, predictability should be inversely related with frame distance, thus enabling structure learning. Using a conditional Pixelcnn++ (Salimans et al., 2017) as predictive family $\mathcal{F}$, we shown in Figure 1c that predictive $\mathcal{F}$-information does indeed decrease with frame distance, despite some fluctuations when the frame distances are large. Using algorithm 1 to construct a Chow-Liu tree, we find that *the tree perfectly recovers the relative order of the frames*. We also generate a Deterministic-Moving-MNIST dataset, where digits move according to *deterministic* dynamics. From the perspective of Shannon mutual information, every pair of frames has the same mutual information. Hence, standard Chow-Liu tree learning algorithm would fail to discover the natural ordering of the frames (causal structure). In contrast, once we constrain the observer to Pixelcnn++ models, algorithm 1 with predictive $\mathcal{F}$-information can still recover the order of different frames when the frame distances are relatively small (less than 9). Compared to the stochastic dynamics case, $\mathcal{F}$-information is more irregular with increasing frame distance, since the Pixelcnn++ tends to overfit.

6.4 Information theoretic approaches to fairness

The goal of fair representation learning is to map inputs $X \in \mathcal{X}$ to a feature space $Z \in \mathcal{Z}$ such that the mutual information between Z and some sensitive attribute $U \in \mathcal{U}$ (such as race or gender) is minimized. The motivation is that using Z (instead of X) as input we can no longer use the sensitive attributes U to make decisions, thus ensuring some notion of fairness. Existing methods obtain fair representations by optimizing against an "adversarial" discriminator so that the discriminator cannot predict U from Z (Edwards & Storkey, 2015; Louizos et al., 2015; Madras et al., 2018; Song et al., 2018). Under some assumptions on U and $\mathcal{F}$, we show in Appendix E that these works actually use $\mathcal{F}$-information minimization as part of their objective, where $\mathcal{F}$ depends on the functional form of the discriminator. However, it is clear from the $\mathcal{F}$-information perspective that features trained with $\mathcal{F}_A$-information minimization might not generalize to $\mathcal{F}_B$-information and vice versa. To illustrate this, we use a function family $\mathcal{F}_j$ as the attacker to extract information from features trained with $I_{\mathcal{F}_i}(Z \to U)$ minimization, where all the $\mathcal{F}$s are neural nets. On three datasets commonly used in the fairness literature (Adult, German, Heritage), previous methods work well at preventing information "leak" against the class of adversary they've been trained on, but fail when we consider different ones. As shown in Figure 3b in Appendix, the diagonal elements in the matrix are usually the smallest in rows, indicating that the attacker function family $\mathcal{F}_i$ extracts more information on featured trained with $\mathcal{F}_{j(j \neq i)}$-information minimization. This challenges the generalizability of fair representations in previous works. Please refer to Appendix F.2 for details.

7 Related work

Alternative definitions of Information Several alternative definitions of mutual information are available in the literature. Renyi entropy and Renyi mutual information (Lenzi et al., 2000) extend Shannon information by replacing KL divergence with f-divergences. However, they have the same difficulty when applied to high dimensional problems as Shannon information.

A line of works most related to ours is H entropy and H mutual information (DeGroot et al., 1962; Grünwald et al., 2004), which associate a definition of entropy to every prediction loss. However, there are two key differences. First, literatures in H entropy only consider a few special types of prediction functions that serve unique theoretical purposes; for example, (Duchi et al., 2018) considers the set of all functions on a feature space to prove surrogate risk consistency, and (Grünwald et al., 2004) considers distributions with fixed moments to prove the duality between maximum entropy and loss minimization. In contrast, our definition takes a completely different perspective — emphasizing bounded computation and intuitive properties of "usable" information. Furthermore H entropy still suffers from difficulty of estimation in high dimension because the definitions do not restrict to functions with small complexity (e.g. Radamacher complexity).

Mutual information estimation The estimation of mutual information in the machine learning field is often on the continuous underlying distribution. For non-parametric mutual information estimators, many methods have exploited the $3H$ principle to calculate the mutual information, such as the Kernel density estimator (Paninski & Yajima, 2008) ,k-Nearest-Neighbor estimator and the KSG estimator (Kraskov et al., 2004). However, these non-parametric estimators usually aren't scalable to high dimension. Recently, several works utilize the variational lower bounds of MI to design MI estimator based on deep neural network in order to estimate MI of high dimension continuous random variables, and typical works are NWJ (Nguyen et al., 2010), CPC (van den Oord et al., 2018), and MINE (Belghazi et al., 2018). In Appendix B we discuss the inefficiencies of theses estimators, which coincide with the experimental results in (Poole et al., 2018).

8 Conclusion

We defined and investigated $\mathcal{F}$-information, a variational extension to classic mutual information that incorporates computational constraints. Unlike Shannon mutual information, $\mathcal{F}$-information attempts to capture usable information, and has very different properties, such as invalidating the data processing inequality. In addition, $\mathcal{F}$-information can be provably estimated, and can thus be more effective for structure learning and fair representation learning.

REFERENCES

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, 2001.

Roberto Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on neural networks*, 5(4):537–550, 1994.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, R. Devon Hjelm, and Aaron C. Courville. Mutual information neural estimation. In *ICML*, 2018.

C Chow and Cong Liu. Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467, 1968.

C. K. Chow and Terry J. Wagner. Consistency of an estimate of tree-dependent probability distributions (corresp.). *IEEE Trans. Information Theory*, 19:369–371, 1973.

Yau Chu and T. Liu. On the shortest arborescence of a directed graph. *Scientia Sinica*, 14:1396–1400, 1965.

Thomas M. Cover and Joy A. Thomas. Elements of information theory. 1991.

Georges A. Darbellay and Igor Vajda. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Trans. Information Theory*, 45:1315–1321, 1999.

Morris H DeGroot et al. Uncertainty, information, and sequential experiments. *The Annals of Mathematical Statistics*, 33(2):404–419, 1962.

John Duchi, Khashayar Khosravi, Feng Ruan, et al. Multiclass classification, information, divergence and surrogate risk. *The Annals of Statistics*, 46(6B):3246–3275, 2018.

Harrison A Edwards and Amos J. Storkey. Censoring representations with an adversary. *CoRR*, abs/1511.05897, 2015.

Wei Gao and Zuo cheng Zhou. Dropout rademacher complexity of deep neural networks. *Science China Information Sciences*, 59:1–12, 2014.

Weihao Gao, Sreeram Kannan, Sewoong Oh, and Pramod Viswanath. Estimating mutual information for discrete-continuous mixtures. In *NIPS*, 2017.

Peter D Grünwald, A Philip Dawid, et al. Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *the Annals of Statistics*, 32(4):1367–1433, 2004.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Phys. Rev. E*, 69:066138, Jun 2004. doi: 10.1103/PhysRevE.69.066138. URL `https://link.aps.org/doi/10.1103/PhysRevE.69.066138`.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, 2015.

EK Lenzi, RS Mendes, and LR Da Silva. Statistical mechanics based on renyi entropy. *Physica A: Statistical Mechanics and its Applications*, 280(3-4):337–345, 2000.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel. The variational fair autoencoder. *CoRR*, abs/1511.00830, 2015.

David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. Learning adversarially fair and transferable representations. *ArXiv*, abs/1802.06309, 2018.

Daniel Marbach, James C. Costello, Robert Küffner, N. Vega, Robert J. Prill, Diogo M Camacho, Kyle R. Allison, Manolis Kellis, James J. Collins, and Gustavo Stolovitzky. Wisdom of crowds for robust gene network inference. In *Nature Methods*, 2012.

Patrick E. Meyer, Kevin Kontos, Frédéric Lafitte, and Gianluca Bontempi. Information-theoretic inference of large transcriptional regulatory networks. In *EURASIP J. Bioinformatics and Systems Biology*, 2007.

XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56:5847–5861, 2010.

Liam Paninski and Masanao Yajima. Undersmoothed kernel entropy estimators. *IEEE Transactions on Information Theory*, 54:4384–4388, 2008.

Rafael Pass and Abhi Shelat. A course in cryptography. 2010.

Judea Pearl. Causality: Models, reasoning, and inference. 2000.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

Ben Poole, Sherjil Ozair, Aäron van den Oord, Alexander A. Alemi, and George Tucker. On variational lower bounds of mutual information. 2018.

Rajesh Ranganath, Sean Gerrish, and David M. Blei. Black box variational inference. In *AISTATS*, 2013.

Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P. Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *ArXiv*, abs/1701.05517, 2017.

Claude E. Shannon and Warren Weaver. The mathematical theory of communication. 1948.

Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. In *AISTATS*, 2018.

Zoltán Szabó. Information theoretical estimators toolbox. *J. Mach. Learn. Res.*, 15:283–287, 2014.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018.

Laurens van der Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. 2008.

A Proofs

A.1 Proof of Proposition 1

Proof. (1)

$$\begin{aligned} H_\Omega(Y|X) &= \inf_{f\in\Omega} \mathbb{E}_{x\sim X, y\sim Y}\left[\log\frac{1}{f[x](y)}\right] = \inf_{f\in\Omega} \mathbb{E}_{x\sim X}\left[\int P_{Y|X}(y|x)\log\frac{1}{f[x](y)}\mathrm{d}\boldsymbol{y}\right] \\ &= \inf_{f\in\Omega} \mathbb{E}_{x\sim X}\left[KL(P_{Y|X}||f[x]) + H(P_{Y|X})\right] \\ &= \mathbb{E}_{x\sim X}\left[H(P_{Y|X})\right] \\ &= H(Y|X) \end{aligned} \tag{8}$$

where the optimal choice for f is $f[x] = P(Y|X=x)$. The same proof techniques for $H_\Omega(Y) = H(Y)$, with the optimal choice for f is $f[\varnothing](y) = P_Y(y)$. Hence we have

$$I_\Omega(Y \to X) = H_\Omega(Y) - H_\Omega(Y|X) = H(Y) - H(Y|X) = I(Y;X) \tag{9}$$

(2)Assume $Y \in \mathbb{R}^d$ and $\mathcal{F} = \{f : \{\varnothing\} \to \frac{1}{Z}e^{-\|y-\mu\|_2} \mid \mu \in \mathbb{R}^d\}$,where Z is a normalizing constant $\int e^{-\|y-\mu\|_2}dy$. In fact $Z = d\Gamma(d)B_d$, where B_d is the volume of d-dimensional unit ball. Then we have

$$\begin{aligned} H_{\mathcal{F}}(Y) &= \inf_{f\in\mathcal{F}} -\mathbb{E}_{y\sim Y}\left[\log \frac{1}{d\Gamma(d)B_d} e^{-\|y-\mu\|_2}\right] \\ &= \inf_{\mu\in\mathbb{R}^d} \mathbb{E}_{y\sim Y}\left[\| y-\mu \|_2\right] + \log\left(d\Gamma(d)B_d\right) \\ &= MAD(Y) + \log\left(d\Gamma(d)B_d\right) \end{aligned} \tag{10}$$

(3) Assume $\mathcal{F} = \{f : \{\varnothing\} \to \mathcal{N}(\mu, \Sigma) \mid \mu \in \mathbb{R}^d, \Sigma = 1/2I_{d\times d}\}$, then we have

$$\begin{aligned} H_{\mathcal{F}}(Y) &= \inf_{f\in\mathcal{F}} -\mathbb{E}_{y\sim Y}\left[\log \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)}\right] \\ &= \inf_{\mu\in\mathbb{R}^d} \mathbb{E}_{y\sim Y}[(y-\mu)^T(y-\mu)] + d\log\pi \\ &= \mathrm{tr}\left(\mathrm{Cov}(Y)\right) + d\log\pi \end{aligned} \tag{11}$$

(4)Assume $Y \in \mathbb{R}^d$ and $\mathcal{F} = \{f : X \to h(y)\exp\left(\theta\cdot\mathbf{t}(y) - A(\theta)\right), \theta \in \Theta\}$, where $\Theta = \{\theta \in \mathbb{R}^d | A(\theta) < +\infty\}$ then

$$\begin{aligned} H_{\mathcal{F}}(Y) &= \inf_{f\in\mathcal{F}} -\mathbb{E}_{y\sim Y}[\log h(y)] - \mathbb{E}_{y\sim Y}\left[\log\exp\left(\theta\cdot\mathbf{t}(y) - A(\theta)\right)\right] \\ &= -\mathbb{E}_{y\sim Y}[\log h(y)] - \sup_{\theta\in\Theta}\left(\theta\cdot\mathbb{E}_{y\sim Y}[\mathbf{t}(y)] - A(\theta)\right)] \\ &= -\mathbb{E}_{y\sim Y}[\log h(y)] - A^*(\mathbb{E}_{y\sim Y}[\mathbf{t}(y)]) \end{aligned} \tag{12}$$

where A^* is the Fenchel dual of the log-partition function $A(\theta)$. Under mild conditions

$$A^*(\mu) = -H(p_{\theta(\mu)}) - \mathbb{E}_{y\sim Y}[\log h(y)]$$

where $\theta(\mu)$ is the unique parameter satisfying $\mathbb{E}_{y\sim p_{\theta(\mu)}}[\mathbf{t}(y)] = \mathbb{E}_{y\sim Y}[\mathbf{t}(y)]$ and $H()$ is the Shannon entropy. Therefore

$$\begin{aligned} H_{\mathcal{F}}(Y) &= -\mathbb{E}_{y\sim Y}[\log h(y)] - A^*(\mathbb{E}_{y\sim Y}[\mathbf{t}(y)]) \\ &= H(p_{\theta(\mathbb{E}_{y\sim Y}[\mathbf{t}(y)])}) \end{aligned} \tag{13}$$

where $H(p_{\theta(\mu_Y)})$ is the entropy of the maximum entropy distribution with expected sufficient statistics μ_Y.

(5) Assume random variable $X \in \mathbb{R}^{d_x}, Y \in \mathbb{R}^{d_y}$, $\mathcal{F} = \{f : x \mapsto \mathcal{N}(\mathbf{t}(x), \Sigma), x \in \mathcal{X}; \varnothing \mapsto \mathcal{N}(\mu, \Sigma) | \mu \in range(\mathbf{t}); \Sigma = \frac{\mathrm{tr}(\mathrm{Cov}(Y))}{2} I_{d_y\times d_y}\}$. Then the $\mathcal{F}$-information from X to Y is

$$\begin{aligned} &I_{\mathcal{F}}(X \to Y) = H_{\mathcal{F}}(Y) - H_{\mathcal{F}}(Y|X) \\ &= \inf_{\mu\in\mathbb{R}^{d_y}} \mathbb{E}_{y\sim Y}\left[-\log\frac{1}{(2\pi)^{\frac{d_y}{2}}|\Sigma|^{\frac{1}{2}}} e^{\frac{-\|y-\mu\|_2^2}{\mathrm{tr}(\mathrm{Cov}(Y))}}\right] - \inf_{\mathbf{t}} \mathbb{E}_{x,y\sim X,Y}\left[-\log\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} e^{\frac{-\|y-\mathbf{t}(x)\|_2^2}{\mathrm{tr}(\mathrm{Cov}(Y))}}\right] \\ &= 1 - \inf_{\mathbf{t}} \mathbb{E}_{x,y\sim X,Y}\left[\frac{\| y-\mathbf{t}(x) \|_2^2}{\mathrm{tr}\left(\mathrm{Cov}(Y)\right)}\right] \\ &= R^2 \end{aligned} \tag{14}$$

□

A.2 Proof of Proposition 2

Proof. (1)

$$H_{\mathcal{F}}(Y) = \inf_{f\in\mathcal{F}} \mathbb{E}_{y\sim Y}\left[\log\frac{1}{f[\varnothing](y)}\right] \geq \inf_{f\in\mathcal{G}} \mathbb{E}_{y\sim Y}\left[\log\frac{1}{f[\varnothing](y)}\right] = H_{\mathcal{G}}(Y) \tag{15}$$

$$H_{\mathcal{F}}(Y|X) = \inf_{f\in\mathcal{F}} \mathbb{E}_{x,y\sim X,Y}\left[\log\frac{1}{f[x](y)}\right] \geq \inf_{f\in\mathcal{G}} \mathbb{E}_{x,y\sim X,Y}\left[\log\frac{1}{f[x](y)}\right] = H_{\mathcal{G}}(Y) \tag{16}$$

The inequalities (15) and (16) are due to take infimum over a larger set.

(2)

Define $\mathcal{F}_{\mathcal{X}\to\mathcal{P}(\mathcal{Y})} = \{g : \mathcal{X} \to \mathcal{P}(\mathcal{Y}) | \exists f \in \mathcal{F}, \forall x \in \mathcal{X}, g[x] = f[x]\}$. Similarly define $\mathcal{F}_{\varnothing\to\mathcal{P}(\mathcal{Y})} = \{g : \varnothing \to \mathcal{P}(\mathcal{Y}) | \exists f \in \mathcal{F}, \forall x \in \mathcal{X}, g[\varnothing] = f[\varnothing]\}$. Intuitively, $\mathcal{F}_{\mathcal{X}\to\mathcal{P}(\mathcal{Y})}$ (resp. $\mathcal{F}_{\varnothing\to\mathcal{P}(\mathcal{Y})}$) restricts the domain of functions in $\mathcal{F}$ to $\mathcal{X}$ (resp. $\varnothing$).

Let $\eta : (\varnothing \to \mathcal{P}(\mathcal{Y})) \to (\mathcal{X} \to \mathcal{P}(\mathcal{Y}))$ be the function matching each $f \in \mathcal{F}_{\varnothing\to\mathcal{P}(\mathcal{Y})}$ with the $g \in \mathcal{F}_{\mathcal{X}\to\mathcal{P}(\mathcal{Y})}$ that ignores X and returns $g[x] = f[\varnothing]$ for each $x \in X$. Then

$$\begin{aligned}
H_{\mathcal{F}}(Y|X) &= \inf_{g\in\mathcal{F}_{X\to\mathcal{P}(\mathcal{Y})}} \mathbb{E}_{x\sim X, y\sim Y}[-\log \frac{1}{g[x](y)]}] \\
&\le \inf_{f\in\mathcal{F}_{\varnothing\to\mathcal{P}(\mathcal{Y})}} \mathbb{E}_{x\sim X, y\sim Y}[-\log \frac{1}{\eta(f)([x](y)}] && (\eta(\mathcal{F}_{\varnothing\to\mathcal{P}(\mathcal{Y})}) \subset \mathcal{F}_{\mathcal{X}\to\mathcal{P}(\mathcal{Y})}) \\
&= \inf_{f\in\mathcal{F}_{\varnothing\to\mathcal{P}(\mathcal{Y})}} \mathbb{E}_{y\sim Y}[\mathbb{E}_{x\sim X|Y}[-\log \frac{1}{f[\varnothing](y)}]] \\
&= \inf_{f\in\mathcal{F}_{\varnothing\to\mathcal{P}(\mathcal{Y})}} \mathbb{E}_{y\sim Y}[-\log \frac{1}{f[\varnothing](y)}] \\
&= H_{\mathcal{F}}(Y) && (17)
\end{aligned}$$

The $\eta(\mathcal{F}_{\varnothing\to\mathcal{P}(\mathcal{Y})}) \subset \mathcal{F}_{\mathcal{X}\to\mathcal{P}(\mathcal{Y})}$ holds by the definition 1. Hence we have

$$I_{\mathcal{F}}(Y \to X) = H_{\mathcal{F}} - H_{\mathcal{F}}(Y|X) \ge 0$$

(3)

Define $\mathcal{F}_{\mathcal{X}\to\mathcal{P}(\mathcal{Y})} = \{g : \mathcal{X} \to \mathcal{P}(\mathcal{Y}) | \exists f \in \mathcal{F}, \forall x \in \mathcal{X}, g[x] = f[x]\}$. Similarly define $\mathcal{F}_{\varnothing\to\mathcal{P}(\mathcal{Y})} = \{g : \varnothing \to \mathcal{P}(\mathcal{Y}) | \exists f \in \mathcal{F}, \forall x \in \mathcal{X}, g[\varnothing] = f[\varnothing]\}$. Intuitively, $\mathcal{F}_{\mathcal{X}\to\mathcal{P}(\mathcal{Y})}$ (resp. $\mathcal{F}_{\varnothing\to\mathcal{P}(\mathcal{Y})}$) restricts the domain of functions in $\mathcal{F}$ to $\mathcal{X}$ (resp. $\varnothing$).

Let $\psi : (\mathcal{X} \to \mathcal{P}(\mathcal{Y})) \times \mathcal{X} \to (\varnothing \to \mathcal{P}(\mathcal{Y}))$ be the function matching each $g \in \mathcal{F}_{\mathcal{X}\to\mathcal{P}(\mathcal{Y})}$ and $x \in \mathcal{X}$ with the $f \in \mathcal{F}_{\varnothing\to P(Y)}$ such that $f[\varnothing] = g[x]$. Then

$$\begin{aligned}
H_{\mathcal{F}}(Y|X) &= \inf_{g\in\mathcal{F}_{\mathcal{X}\to\mathcal{P}(\mathcal{Y})}} \mathbb{E}_{x\sim X, y\sim Y}\left[-\log \frac{1}{g[x](y)}\right] \\
&= \inf_{g\in\mathcal{F}_{\mathcal{X}\to\mathcal{P}(\mathcal{Y})}} \mathbb{E}_{x\sim X}\left[\mathbb{E}_{y\sim Y|X}\left[-\log \frac{1}{\psi(g,x)[\varnothing](y)}\right]\right] \\
&\ge \mathbb{E}_{x\sim X}\left[\inf_{g\in\mathcal{F}_{\mathcal{X}\to\mathcal{P}(\mathcal{Y})}} \mathbb{E}_{y\sim Y|X}\left[-\log \frac{1}{\psi(g,x)[\varnothing](y)}\right]\right] \\
&\ge \mathbb{E}_{x\sim X}\left[\inf_{f\in\mathcal{F}_{\varnothing\to\mathcal{P}(\mathcal{Y})}} \mathbb{E}_{y\sim Y|X}\left[-\log \frac{1}{f[\varnothing](y)}\right]\right] && (\mathcal{F}_{\varnothing\to\mathcal{P}(\mathcal{Y})} \supset \psi(\mathcal{F}_{\mathcal{X}\to\mathcal{P}(\mathcal{Y})} \times \mathcal{X})) \\
&= \mathbb{E}_{x\sim X}\left[\inf_{f\in\mathcal{F}_{\varnothing\to\mathcal{P}(\mathcal{Y})}} \mathbb{E}_{y\sim Y}\left[-\log \frac{1}{f[\varnothing](y)}\right]\right] && (\text{Independence of } X \text{ and } Y) \\
&= \mathbb{E}_{x\sim X}\left[H_{\mathcal{F}}(Y)\right] \\
&= H_{\mathcal{F}}(Y) && (18)
\end{aligned}$$

The $\mathcal{F}_{\varnothing\to\mathcal{P}(\mathcal{Y})} \supset \psi(\mathcal{F}_{\mathcal{X}\to\mathcal{P}(\mathcal{Y})} \times \mathcal{X})$ holds by the definition 1. It was previously established via optional-ignorance that for any $\mathcal{F}$, $H_{\mathcal{F}}(Y|X) \le H_{\mathcal{F}}(Y)$, and thus $H_{\mathcal{F}}(Y|X) = H_{\mathcal{F}}(Y)$. We therefore observe that when X is independent of Y, $I_{\mathcal{F}}(X \to Y) = 0$. □

A.3 Proof of Theorem 1

Before proofing theorem 1, we introduce two lemmas:

Lemma 3. *Let X, Y be two random variables taking values in $\mathcal{X}, \mathcal{Y}$ and $\mathcal{D}$ denotes the set of samples drawn from the joint distribution over $\mathcal{X} \times \mathcal{Y}$. Assume $\forall f \in \mathcal{F}, x \in \mathcal{X}, y \in \mathcal{Y}, \log f[x](y) \in [-B, B]$.*

Take $\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{|\mathcal{D}|} \sum_{x_i, y_i \in \mathcal{D}} - \log f[x_i](y_i)$*, then* $\forall \delta > 0$*, with probability at least* $1 - \delta$*, we have:*

$$\left| H_{\mathcal{F}}(Y|X) - \frac{1}{|\mathcal{D}|} \sum_{x_i, y_i \in \mathcal{D}} - \log \hat{f}[x_i](y_i) \right| \leq 2\mathfrak{R}_{|\mathcal{D}|}(\log \circ \mathcal{G}_{\mathcal{F}}) + B\sqrt{\frac{\log \frac{2}{\delta}}{2|\mathcal{D}|}} \tag{19}$$

Proof. Applying McDiarmid's inequality to function Φ defined for any sample $\mathcal{D}$ by

$$\Phi(D) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{x,y} \left[- \log f[x](y) \right] - \frac{1}{|\mathcal{D}|} \sum_{x_i, y_i \in \mathcal{D}} - \log f[x_i](y_i) \right| \tag{20}$$

Let $\mathcal{D}$ and $\mathcal{D}'$ be two samples differing by exactly one point, then since the difference of suprema does not exceed the supremum of the difference and $\forall f \in \mathcal{F}, x \in \mathcal{X}, y \in \mathcal{Y}, \log f[x](y) \in [-B, 0]$, we have:

$$\begin{aligned}
&\Phi(\mathcal{D}) - \Phi(\mathcal{D}') \\
&\leq \sup_{f \in \mathcal{F}} \left[\left| \frac{1}{|\mathcal{D}|} \sum_{x_i, y_i \in \mathcal{D}} \log f[x_i](y_i) - \mathbb{E}_{x,y} \left[\log f[x](y) \right] \right| - \left| \frac{1}{|\mathcal{D}'|} \sum_{x_i, y_i \in \mathcal{D}'} \log f[x_i](y_i) - \mathbb{E}_{x,y} \left[\log f[x](y) \right] \right| \right] \\
&\leq \sup_{f \in \mathcal{F}} \left| \frac{1}{|\mathcal{D}|} \sum_{x_i, y_i \in \mathcal{D}} - \log f[x_i](y_i) | - \frac{1}{|\mathcal{D}'|} \sum_{x_i, y_i \in \mathcal{D}'} - \log f[x_i](y_i) \right| \\
&\leq \frac{B}{|D|}
\end{aligned}$$

then by McDiarmid's inequality, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds:

$$\Phi(D) \leq \mathbb{E}_{\mathcal{D}}[\Phi(D)] + B\sqrt{\frac{\log \frac{1}{\delta}}{2|\mathcal{D}|}} \tag{21}$$

Then we bound the $\mathbb{E}_{\mathcal{D}}[\Phi(D)]$ term:

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}}[\Phi(D)] &= \mathbb{E}_{\mathcal{D}}\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}_{x,y}\left[-\log f[x](y)\right] - \frac{1}{|\mathcal{D}|}\sum_{x_i,y_i\in\mathcal{D}} -\log f[x_i](y_i)\right|\right] && (22)\\
&= \mathbb{E}_{\mathcal{D}}\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}_{\mathcal{D}'}\left[\frac{1}{|\mathcal{D}'|}\sum_{x_i',y_i'\in\mathcal{D}'}\log f[x_i'](y_i')\right] - \frac{1}{|\mathcal{D}|}\sum_{x_i,y_i\in\mathcal{D}}\log f[x_i](y_i)\right|\right] && (23)\\
&\le \mathbb{E}_{\mathcal{D}}\left[\sup_{f\in\mathcal{F}}\mathbb{E}_{\mathcal{D}'}\left|\frac{1}{|\mathcal{D}'|}\sum_{x_i',y_i'\in\mathcal{D}'}\log f[x_i'](y_i')| - \frac{1}{|\mathcal{D}|}\sum_{x_i,y_i\in\mathcal{D}}\log f[x_i](y_i)\right|\right] && (24)\\
&\le \mathbb{E}_{\mathcal{D},\mathcal{D}'}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{|\mathcal{D}'|}\sum_{x_i',y_i'\in\mathcal{D}'}\log f[x_i'](y_i')| - \frac{1}{|\mathcal{D}|}\sum_{x_i,y_i\in\mathcal{D}}\log f[x_i](y_i)\right|\right] && (25)\\
&= \mathbb{E}_{\mathcal{D},\mathcal{D}'}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}(\log f[x_i'](y_i') - \log f[x_i](y_i))\right|\right] && (26)\\
&= \mathbb{E}_{\mathcal{D},\mathcal{D}',\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\sigma_i(\log f[x_i'](y_i') - \log f[x_i](y_i))\right|\right] && (27)\\
&\le \mathbb{E}_{\mathcal{D},\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\sigma_i\log f[x_i](y_i)\right|\right] + \mathbb{E}_{\mathcal{D}',\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\sigma_i\log f[x_i'](y_i')\right|\right] && (28)\\
&= 2\mathbb{E}_{\mathcal{D},\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\sigma_i\log f[x_i](y_i)\right|\right] && (29)\\
&= 2\mathbb{E}_{\mathcal{D},\sigma}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\sigma_i\log g(x_i,y_i)\right|\right] = 2\mathfrak{R}_{|\mathcal{D}|}(\log\circ\mathcal{G}_{\mathcal{F}}) && (30)
\end{aligned}
$$

where σ_is are Rademacher variables that is uniform in $\{-1,+1\}$. Inequality (25) follows from the convexity of sup, inequality (24) follows from the convexity of $|x-c|$. (30) follow from the definition of $\mathcal{G}$ and Rademacher complexity.

Finally, combining inequality (21) and (30) yields for all $f\in\mathcal{F}$, with probability at least $1-\delta$

$$
\left|\mathbb{E}_{x,y}[-\log f[x](y)] - \frac{1}{|\mathcal{D}|}\sum_{x_i,y_i\in\mathcal{D}} -\log f[x_i](y_i)\right| \le 2\mathfrak{R}_{|\mathcal{D}|}(\log\circ\mathcal{G}_{\mathcal{F}}) + B\sqrt{\frac{\log\frac{2}{\delta}}{2|\mathcal{D}|}} \tag{31}
$$

The results holds for $\hat{f} = \arg\min_{f\in\mathcal{F}} \frac{1}{|\mathcal{D}|}\sum_{x_i,y_i\in\mathcal{D}} -\log f[x_i](y_i)$. Then the bound (19) can be derived easily by the definition of $H_{\mathcal{F}}(Y|X)$. □

Similar bounds can be derived for $H_{\mathcal{F}}(Y)$ if we choose the domain of x to be $\mathcal{X}=\{\varnothing\}$. Define $\mathcal{G}_{\mathcal{F}^{\varnothing}} = \{g|g(y) = f[\varnothing](y), f\in\mathcal{F}\}$, then we have following lemma:

Lemma 4. *Let Y be random variable taking values in $\mathcal{Y}$ and $\mathcal{D}$ denotes the set of samples drawn from the underlying distribution $P(Y)$. Assume $\forall f\in\mathcal{F}, y\in\mathcal{Y}, \log f[\varnothing](y)\in[-B,B]$. Take*

$\hat{f} = \arg\min_{f\in\mathcal{F}} \frac{1}{|\mathcal{D}|} \sum_{x_i,y_i\in\mathcal{D}} -\log f[\varnothing](y_i)$, *then for any* $\delta > 0$, *with probability at least* $1-\delta$, *we have:*

$$\left| H_{\mathcal{F}}(Y) - \frac{1}{|\mathcal{D}|} \sum_{y_i\in\mathcal{D}} -\log \hat{f}[\varnothing](y_i) \right| \leq 2\mathfrak{R}_{|\mathcal{D}|}(\log\circ\mathcal{G}_{\mathcal{F}^{\varnothing}}) + B\sqrt{\frac{\log\frac{2}{\delta}}{2|\mathcal{D}|}} \tag{32}$$

$$\leq 2\mathfrak{R}_{|\mathcal{D}|}(\log\circ\mathcal{G}_{\mathcal{F}}) + B\sqrt{\frac{\log\frac{2}{\delta}}{2|\mathcal{D}|}} \tag{33}$$

Proof. The first inequality (32) can be similarly derived as Lemma 3. Since $\mathcal{F}$ belongs to predictive family, hence there exits function $h : \mathcal{F} \to \mathcal{F}$, such that $h(f) = f'$ and $\forall x \in \mathcal{X}, f'[x] = f[\varnothing]$.

$$\begin{aligned}
\mathfrak{R}_{|\mathcal{D}|}(\log\circ\mathcal{G}_{\mathcal{F}^{\varnothing}}) &= \mathbb{E}_{\mathcal{D},\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\sigma_i \log f[\varnothing](y_i)\right|\right] \\
&= \mathbb{E}_{\mathcal{D},\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\sigma_i \log h(f)[x_i](y_i)\right|\right] \\
&\leq \mathbb{E}_{\mathcal{D},\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\sigma_i \log f[x_i](y_i)\right|\right] \qquad (34) \\
&= \mathfrak{R}_{|\mathcal{D}|}(\log\circ\mathcal{G}_{\mathcal{F}})
\end{aligned}$$

The inequality (34) holds because of $\forall f \in \mathcal{F}, h(f) \in \mathcal{F}$. $\square$

Now we begin to prove the theorem 1:

Theorem 1. Assume $\forall f \in \mathcal{F}, x \in \mathcal{X}, y \in \mathcal{Y}, \log f[x](y) \in [-B, B]$, for any $\delta > 0$, with probability at least $1 - 2\delta$, we have:

$$\left| I_{\mathcal{F}}(X \to Y) - \hat{I}_{\mathcal{F}}(X \to Y; \mathcal{D}) \right| \leq 4\mathfrak{R}_{|\mathcal{D}|}(\log\circ\mathcal{G}_{\mathcal{F}}) + 2B\sqrt{\frac{\log\frac{2}{\delta}}{2|\mathcal{D}|}}$$

Proof. Using the triangular inequality we have:

$$\left| I_{\mathcal{F}}(X \to Y) - \hat{I}_{\mathcal{F}}(X \to Y; \mathcal{D}) \right| \leq \left| H_{\mathcal{F}}(Y|X) - \frac{1}{|\mathcal{D}|}\sum_{x_i,y_i\in\mathcal{D}} -\log \hat{f}[x_i](y_i) \right| + \left| H_{\mathcal{F}}(Y) - \frac{1}{|\mathcal{D}|}\sum_{y_i\in\mathcal{D}} -\log \hat{f}^{\varnothing}[\varnothing](y_i) \right| \tag{35}$$

For simplicity let $D_{Y|X} = \left| H_{\mathcal{F}}(Y|X) - \frac{1}{|\mathcal{D}|}\sum_{x_i, y_i \in \mathcal{D}} - \log \hat{f}[x_i](y_i) \right|$ and $D_Y = \left| H_{\mathcal{F}}(Y) - \frac{1}{|\mathcal{D}|}\sum_{y_i \in \mathcal{D}} - \log \hat{f}^{\varnothing}[\varnothing](y_i) \right|$. With inequality (35), Lemma 3 and Lemma 4, we have:

$$
\begin{aligned}
&\Pr\left(\left| I_{\mathcal{F}}(X \to Y) - \hat{I}_{\mathcal{F}}(X \to Y; \mathcal{D}) \right| > 4\mathfrak{R}_{|\mathcal{D}|}(\log \circ \mathcal{G}_{\mathcal{F}}) + 2B\sqrt{\frac{\log \frac{2}{\delta}}{2|\mathcal{D}|}} \right) \\
&\le \Pr\left(D_{Y|X} + D_Y > 4\mathfrak{R}_{|\mathcal{D}|}(\log \circ \mathcal{G}_{\mathcal{F}}) + 2B\sqrt{\frac{\log \frac{2}{\delta}}{2|\mathcal{D}|}} \right) && \text{(Inequality (35))} \\
&\le \Pr\left(\left(D_{Y|X} > 2\mathfrak{R}_{|\mathcal{D}|}(\log \circ \mathcal{G}_{\mathcal{F}}) + B\sqrt{\frac{\log \frac{2}{\delta}}{2|\mathcal{D}|}} \right) \vee \left(D_Y > 2\mathfrak{R}_{|\mathcal{D}|}(\log \circ \mathcal{G}_{\mathcal{F}}) + B\sqrt{\frac{\log \frac{2}{\delta}}{2|\mathcal{D}|}} \right) \right) \\
&\le \Pr\left(D_{Y|X} > 2\mathfrak{R}_{|\mathcal{D}|}(\log \circ \mathcal{G}_{\mathcal{F}}) + B\sqrt{\frac{\log \frac{2}{\delta}}{2|\mathcal{D}|}} \right) + \Pr\left(D_Y > 2\mathfrak{R}_{|\mathcal{D}|}(\log \circ \mathcal{G}_{\mathcal{F}}) + B\sqrt{\frac{\log \frac{2}{\delta}}{2|\mathcal{D}|}} \right) \\
& && \text{(Union bound)} \\
&\le 2\delta && \text{(Lemma 3 and Lemma 4)}
\end{aligned}
$$

Hence we have:

$$
\Pr\left(\left| I_{\mathcal{F}}(X \to Y) - \hat{I}_{\mathcal{F}}(X \to Y; \mathcal{D}) \right| \le 4\mathfrak{R}_{|\mathcal{D}|}(\log \circ \mathcal{G}_{\mathcal{F}}) + 2B\sqrt{\frac{\log \frac{2}{\delta}}{2|\mathcal{D}|}} \right) \ge 1 - 2\delta
$$

□

A.4 Proof of Corollary 1.1

Proof. From theorem 1 we have:

$$
\left| I_{\mathcal{F}}(X \to Y) - \hat{I}_{\mathcal{F}}(X \to Y; \mathcal{D}) \right| \le 4\mathfrak{R}_{|\mathcal{D}|}(\log \circ \mathcal{G}_{\mathcal{F}}) + 2B\sqrt{\frac{\log \frac{1}{\delta}}{2|\mathcal{D}|}}
$$

In the following $\|\begin{pmatrix} W \\ b \end{pmatrix}\|_2$ is the matrix 2-norm of $\begin{pmatrix} W \\ b \end{pmatrix}$, then the Rademacher term can be bounded as following:

$$
\begin{aligned}
\mathfrak{R}_{|\mathcal{D}|}(\log \circ \mathcal{G}_{\mathcal{F}}) &= \frac{1}{|\mathcal{D}|}\mathbb{E}_{\sigma}\left[\sup_{W,b,\|(W,b)\|_2 \le 1} \left| \sum_{i=1}^{|\mathcal{D}|} \sigma_i \left(\log \frac{1}{\sqrt{2\pi}} - \frac{1}{2}\|\begin{pmatrix} W \\ b \end{pmatrix}\|_2 \le 1 \right) \right| \right] \\
&\le \frac{1}{|\mathcal{D}|}\mathbb{E}_{\sigma}\left[\sup_{W,b,\|\begin{pmatrix} W \\ b \end{pmatrix}\|_2 \le 1} \left| \sum_{i=1}^{|\mathcal{D}|} \sigma_i \left(-\frac{1}{2}\|y_i - Wx_i - b\|^2 \right) \right| \right] + \frac{1}{|\mathcal{D}|}\mathbb{E}_{\sigma}\left[\left| \sum_{i=1}^{|\mathcal{D}|} \sigma_i \log \frac{1}{\sqrt{2\pi}} \right| \right]
\end{aligned} \tag{36}
$$

The second term in RHS can be bounded as following:

$$\frac{1}{|\mathcal{D}|}\mathbb{E}_\sigma\left[\left|\sum_{i=1}^{|\mathcal{D}|}\sigma_i \log\frac{1}{\sqrt{2\pi}}\right|\right] \leq \frac{1}{|\mathcal{D}|}\sqrt{\mathbb{E}_\sigma\left[\left(\sum_{i=1}^{|\mathcal{D}|}\sigma_i \log\frac{1}{\sqrt{2\pi}}\right)^2\right]} \qquad \text{(concavity of } x^{\frac{1}{2}}\text{)}$$

$$= \frac{1}{|\mathcal{D}|}\sqrt{|\mathcal{D}| * (\log\frac{1}{\sqrt{2\pi}})^2} \qquad \text{(Independence of } \sigma_i\text{s)}$$

$$= \sqrt{\frac{(\log\frac{1}{\sqrt{2\pi}})^2}{|\mathcal{D}|}} \tag{37}$$

The first term in RHS can be bounded as following:

$$\frac{1}{|\mathcal{D}|}\mathbb{E}_{\mathcal{D},\sigma}\left[\sup_{W,b,\|\binom{W}{b}\|_2\leq 1}\left|\sum_{i=1}^{|\mathcal{D}|}\sigma_i\left(-\frac{1}{2}\|y_i - Wx_i - b\|^2\right)\right|\right]$$

$$= \frac{1}{2|\mathcal{D}|}\mathbb{E}_{\mathcal{D},\sigma}\left[\sup_{W,b,\|\binom{W}{b}\|_2\leq 1}\left|\sum_{i=1}^{|\mathcal{D}|}\sigma_i\left(\|y_i - Wx_i - b\|^2\right)\right|\right]$$

$$\leq \frac{\max_i\|y_i\|_2^2}{2}\sqrt{\frac{1}{|\mathcal{D}|}} + \max_i\|x_i\|_2\sqrt{\frac{\max_i\|y_i\|^2}{|\mathcal{D}|}} + \frac{1}{2|\mathcal{D}|}\mathbb{E}_{\mathcal{D},\sigma}\left[\sup_{W,b,\|\binom{W}{b}\|_2\leq 1}\left|\sum_{i=1}^{|\mathcal{D}|}\sigma_i\left(\|Wx_i + b\|^2\right)\right|\right] \tag{38}$$

$$\leq \frac{\max_i\|y_i\|_2^2}{2}\sqrt{\frac{1}{|\mathcal{D}|}} + \max_i\|x_i\|_2\sqrt{\frac{\max_i\|y_i\|^2}{|\mathcal{D}|}} + \frac{\max_i\|x_i\|_2}{2|\mathcal{D}|}\mathbb{E}_{\mathcal{D},\sigma}\left[\sup_{W,b,\|\binom{W}{b}\|_2\leq 1}\left|\sum_{i=1}^{|\mathcal{D}|}\sigma_i\left(\|Wx_i + b\|\right)\right|\right] \tag{39}$$

$$\leq \frac{\max_i\|y_i\|_2^2}{2}\sqrt{\frac{1}{|\mathcal{D}|}} + \max_i\|x_i\|_2\sqrt{\frac{\max_i\|y_i\|_2^2}{|\mathcal{D}|}} + \frac{\max_i\|x_i\|_2}{2}\sqrt{\frac{\max_i\|x_i\|_2^2}{|\mathcal{D}|}} \tag{40}$$

□

The inequalities (39) and (38) follow the same proof in (37).

Hence we have:

$$\mathfrak{R}_{|\mathcal{D}|}(\log\circ\mathcal{G}_\mathcal{F}) \leq \frac{M}{\sqrt{4|\mathcal{D}|}} \tag{41}$$

Moreover, it's easy to prove that $B \leq 2M\sqrt{\log\frac{1}{\delta}}$, combine inequality (41) we arrive at the theorem.

A.5 Proof of Theorem 2

Proof. Let $C_\mathcal{D}(g^*)$ be the estimated sum of edge weights on dataset $\mathcal{D}$ of the optimal directed tree g^*. The same notation for tree g that is optimal on dataset $\mathcal{D}$. Let $\epsilon = \max_{i,j}\left\{\left|I_\mathcal{F}(X_i \to X_j) - \hat{I}_\mathcal{F}(X_i \to X_j; \mathcal{D})\right|\right\}$ which is the maximum absolute estimation error of single edge weight, then we have:

$$C(g) + (m-1)\epsilon \geq C_\mathcal{D}(g) \geq C_\mathcal{D}(g^*) \geq C(g^*) - (m-1)\epsilon \tag{42}$$

By the definition of ϵ we have $\forall g, |C(g) - C_D(g)| \leq (m-1)\epsilon$, hence the first and the third inequality in (42) hold. From theorem 1 we have:

$$\Pr\left(\epsilon > \max_{i,j}\left\{4\mathfrak{R}_{\mathcal{D}_{i,j}}(\log\circ\mathcal{G}_{i,j}) + B\sqrt{\frac{\log\frac{1}{\delta}}{2}}(|\mathcal{D}_j|^{-\frac{1}{2}} + |\mathcal{D}_{i,j}|^{-\frac{1}{2}})\right\}\right)$$

$$\leq \Pr\left(\exists i,j, \left|I_{\mathcal{F}}(X_i \to X_j) - \hat{I}_{\mathcal{F}}(X_i \to X_j;\mathcal{D})\right| > 4\mathfrak{R}_{\mathcal{D}_{i,j}}(\log\circ\mathcal{G}_{i,j}) + B\sqrt{\frac{\log\frac{1}{\delta}}{2}}(|\mathcal{D}_j|^{-\frac{1}{2}} + |\mathcal{D}_{i,j}|^{-\frac{1}{2}})\right)$$

$$\leq m(m-1)2\delta \qquad \text{(Theorem 1 and Union bound)}$$

Hence

$$\Pr\left(\epsilon \leq \max_{i,j}\left\{4\mathfrak{R}_{\mathcal{D}_{i,j}}(\log\circ\mathcal{G}_{i,j}) + B\sqrt{\frac{\log\frac{1}{\delta}}{2}}(|\mathcal{D}_j|^{-\frac{1}{2}} + |\mathcal{D}_{i,j}|^{-\frac{1}{2}})\right\}\right) \geq 1 - m(m-1)2\delta \tag{43}$$

Then using inequality (42) and (43) we arrive at the result. □

B Analysis of approximate estimators for Shannon information

Firstly we show that there are inherent limitations for the two commonly used high dimension mutual information estimators. The first is the CPC (or InfoNCE in Poole et al. (2018)) estimator (I_{CPC}) proposed by van den Oord et al. (2018):

$$I_{\text{CPC}} = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\log\frac{f_\theta(x_i,y_i)}{\frac{1}{N}\sum\limits_{j=1}^{N} f_\theta(x_i,y_j)}\right] \leq I(X;Y) \tag{44}$$

where the expectation is over N independent samples form the joint distribution $\prod\limits_i p(x_i,y_i)$, and the second is the NWJ estimator (I_{NWJ}) proposed by Nguyen et al. (2010):

$$I_{\text{NWJ}} = \mathbb{E}_{x,y\sim P(x,y)}\left[f_\theta(x,y)\right] - e^{-1}\mathbb{E}_{x,y\sim P(x)P(y)}\left[e^{f_\theta(x,y)}\right] \leq I(X;Y) \tag{45}$$

In both case, f_θ is a parameterized function. The objection of these work is to optimize θ to approximate mutual information. van den Oord et al. (2018) shows I_{CPC} tends to underestimate the mutual information and have estimates that saturate at $\log(N)$, typically N is the batch size. It's straightforward to show that $I_{\text{CPC}} \leq \log N$ (Appendix), which coincide with the empirical result. This means the I_{CPC} estimator will incur large bias when $I(X;Y) \geq \log N$.

Poole et al. (2018) also shows that I_{NWJ} suffers from high variance. Note that the I_{NWJ} involves the $\frac{1}{e}\mathbb{E}_{x,y\sim p(x)p(y)}\left[e^{f_\theta(x,y)}\right]$ term, which is commonly used in large deviation theory can often be dominated by rare datapoint. These phenomenons make it a poor mutual information estimator by optimizing θ. The optimal value for I_{NWJ} achieves when $f_\theta(x,y) = 1 + \log\frac{p(x,y)}{p(x)p(y)}$. Given $\{(x_i,y_i)\}_{i=1}^N$ (resp. $\{(\bar{x}_i,\bar{y}_i)\}_{i=1}^N$) as N datapoints independently sampled from the distribution $p(x,y)$ (resp. $p(x)p(y)$), denote the empirical estimation of I_{NWJ} as $\hat{I}_{NWJ} = \frac{1}{N}\sum_{i=1}^N\left[f_\theta(x_i,y_i)\right] - \frac{e^{-1}}{N}\sum_{i=1}^N\left[e^{f_\theta(\bar{x}_i,\bar{y}_i)}\right]$. Below we show that when even $f_\theta(x,y) = 1 + \log\frac{p(x,y)}{p(x)p(y)}$, $\hat{I}_{NWJ}$ suffers from high variance:

Theorem 5. *Assume f_θ achieves the optimum, which is $f_\theta(x,y) = 1 + \log\frac{p(x,y)}{p(x)p(y)}$, and $\mathbb{E}_{p(x,y)}\left[\frac{p(x,y)}{p(x)p(y)}\right] \geq \frac{1}{1-e^{-1}}$. Then we have* $\text{Var}\left(\hat{I}_{\text{NWJ}}\right) \geq \frac{e^{I(X;Y)-3}}{N}$.

Proof. Let $x_i, y_i \sim p(x)p(y)$, denote $z_i = \frac{p(x_i,y_i)}{p(x_i)p(y_i)}$. Apparently $\mathbb{E}_{p(x)p(y)}\left[z_i\right] = 1$. By central limit theorem we have:

$$\begin{aligned}
\mathrm{Var}(z_1) &= \mathbb{E}_{p(x)p(y)}\left[z_i^2\right] - \left(\mathbb{E}_{p(x)p(y)}\left[z_1\right]\right)^2 \\
&= \mathbb{E}_{p(x)p(y)}\left[z_1^2\right] - 1 \\
&= \mathbb{E}_{p(x)p(y)}\left[\left(\frac{p(x_1,y_1)}{p(x_1)p(y_1)}\right)^2\right] - 1 \\
&= \mathbb{E}_{p(x,y)}\left[\frac{p(x_1,y_1)}{p(x_1)p(y_1)}\right] - 1 \qquad (46) \\
& \qquad (47)
\end{aligned}$$

Hence, by assumption that

$$\begin{aligned}
\log \mathrm{Var}(z_1) &= \log\left(\mathbb{E}_{p(x,y)}\left[\frac{p(x_1,y_1)}{p(x_1)p(y_1)}\right] - 1\right) \\
&\geq \log\left(\mathbb{E}_{p(x,y)}\left[\frac{p(x_1,y_1)}{p(x_1)p(y_1)}\right]\right) - 1 \qquad (48) \\
&\geq \left(\mathbb{E}_{p(x,y)} \log\left[\frac{p(x_1,y_1)}{p(x_1)p(y_1)}\right]\right) - 1 \qquad (49) \\
&= I(X;Y) - 1 \qquad (50)
\end{aligned}$$

(48) holds by assumption $\mathbb{E}_{p(x,y)}\left[\frac{p(x,y)}{p(x)p(y)}\right] \geq \frac{1}{1-e^{-1}}$ and (49) follows the Jensen's inequality. Then we have the variance of the estimation satisfy

$$\begin{aligned}
\mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N} z_i\right) &= \frac{\mathrm{Var}\left(z_1\right)}{N} \\
&\geq \frac{e^{I(X;Y)-1}}{N} \qquad (51)
\end{aligned}$$

Since $\{(x_i,y_i)\}_{i=1}^{N}$ (resp. $\{(\bar{x}_i,\bar{y}_i)\}_{i=1}^{N}$) are N datapoints independently sampled from the distribution $p(x,y)$ (resp. $p(x)p(y)$), we have

$$\begin{aligned}
\mathrm{Var}\left(\hat{I}_{NWJ}\right) &= \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N}\left[f_\theta(x_i,y_i)\right] - \frac{e^{-1}}{N}\sum_{i=1}^{N}\left[e^{f_\theta(\bar{x}_i,\bar{y}_i)}\right]\right) \\
&\geq \mathrm{Var}\left(\frac{e^{-1}}{N}\sum_{i=1}^{N}\left[e^{f_\theta(\bar{x}_i,\bar{y}_i)}\right]\right) \\
&= \mathrm{Var}\left(\frac{e^{-1}}{N}\sum_{i=1}^{N} z_i\right) \geq \frac{e^{I(X;Y)-3}}{N} \qquad (52)
\end{aligned}$$

□

The lower bound of the variance highly coincides with the empirical results in Poole et al. (2018), which shows the variance increases in exponential to mutual information when sample number N is finite. Similar to I_{NWJ}, the mutual information estimator I_{MINE} proposed by Belghazi et al. (2018) is also a unnormalized mutual information estimator. The only difference between I_{MINE} and I_{NWJ} is the second term of I_{MINE} is the log partition function, which is notoriously difficult to be estimated.

C THE NEW ALGORITHM FOR CHOW-LIU TREE CONSTRUCTION

Denote $\hat{I}_{\mathcal{F}}(X_i \to X_j; \mathcal{D})$ as the estimation of $I_{\mathcal{F}}(X_i \to X_j)$ on the dataset $\mathcal{D}$,

Algorithm 1 Construct Chow-Liu Trees with $\mathcal{F}$-Information

Require: $\mathcal{D} = \{\hat{X}_i\}_{i=1}^m$, with each $\hat{X}_i$ being a set of datapoints sampled from the underlying distribution of random variable X_i. The set of function families $\{\mathcal{F}_{i,j}\}_{i,j=1,i\neq j}^m$

for $i = 1, \ldots, m$ **do**
  for $j = 1, \ldots, m$ **do**
    if $i \neq j$ **then**
      Calculate the edge weight: $e_{i\to j} = \hat{I}_{\mathcal{F}_{i,j}}(X_i \to X_j; \{\hat{X}_i, \hat{X}_j\})$.
    end if
  end for
end for
Construct the fully connected graph $G = (V, E)$, with node set $V = (X_1, \ldots, X_m)$ and edge set $E = \{e_{i\to j}\}_{i,j=1,i\neq j}^m$.
Construct the maximal directed spanning tree g on G by Chu-Liu algorithm.
return g

D Proofs for connecting $\mathcal{F}$-information to MLE

If we constrain the joint distribution on a directed tree structure and only use functions in $\mathcal{F}$ to approximate the relationship between random variables. Denote $\mathcal{F}_{t(g)(i)\to i}$ as the function family for $H_{\mathcal{F}}(X_i|X_{t(g)(i)})$ and we assume $\mathcal{F}_i$ is the same as $\mathcal{F}_{j,i}$. Then the maximum log-likelihood under $\mathcal{F}$ and g is:

$$
\begin{aligned}
\mathbb{E}_{\vec{X}}[\log P_{g,\mathcal{F}}(\vec{X})] &= \mathbb{E}_{\vec{X}}\left[\log \inf_{f\in\mathcal{F}_1} f[\varnothing](X_1) * \prod_{i=2}^{N} \inf_{f\in\mathcal{F}_{t(g)(i),i}} P(X_i \mid X_{t(g)(i)})\right] \\
&= -\sum_{i=2}^{N} H_{\mathcal{F}_{t(g)(i),i}}(X_i|X_{t(g)(i)}) - H_{\mathcal{F}_1}(X_1) \\
&= \sum_{i=2}^{N} H_{\mathcal{F}_i}(X_i) - H_{\mathcal{F}_{t(g),i}}(X_i|X_{t(g)(i)}) - \sum_{i=1}^{N} H_{\mathcal{F}_i}(X_i) \\
&= \sum_{i=2}^{N} I_{\mathcal{F}_{t(g)(i),i}}(X_{t(g)(i)} \to X_i) - \underbrace{\sum_{i=1}^{N} H_{\mathcal{F}_i}(X_i)}_{\text{independent of g}}
\end{aligned}
\tag{53}
$$

E Connection to existing methods for fairness

We can adapt the $\mathcal{F}$-information to fairness. Denote the sensitive data as U and the representation as Z. Assume U is discrete and $\mathcal{F}$ belongs to preditive family 1. Then we have $H_{\mathcal{F}}(U) = H(U)$ as long as $\mathcal{F}$ has $\mathrm{softmax}$ on the top and belongs to predictive family. In this case, minimizing $I_{\mathcal{F}}(Z \to U)$ equals to minimize $-H_{\mathcal{F}}(Y|X)$. Let joint distribution of Z and U be paramterized by ϕ. Hence the final objective is:

$$
\min_{\phi}\{I_{\mathcal{F}}(u; z)\} = \min_{\phi}\left(\sup_{f\in\mathcal{F}} \mathrm{E}_{z,u\sim q_\phi(z,u)}[\log P_f(z|u)]\right)
$$

In Edwards & Storkey (2015); Madras et al. (2018); Louizos et al. (2015); Song et al. (2018), functions in $\mathcal{F}$ are parameterized by a discriminator.

F Detailed Experiments setup

F.1 Chow-Liu tree construction

Figure 2 shows Simulation-1∼Simulation-6 of Chow-Liu tree construction. The Simulation-A and Simulation-B in the main body correspond to Simulation-1 and Simulation-4.

Simulation-1 $\sim$ Simulation-3 :

The true Chow-Liu tree is a star tree (i.e. all random variables are conditionally independent given X_1). We conduct all experiments for 10 times, each time with random simulated orthogonal matrices $\{W_i\}_{i=2}^{20}$. The ground-truth Chow-Liu is a star tree. Simulation-1: $X_1 \sim \mathcal{U}(0, 10)$ and $X_i \mid X_1 \sim \mathcal{N}(W_i X_1, 6I), (2 \leq i \leq 20)$; Simulation-2: $X_1 \sim \mathcal{U}(0, 10)$ and $X_i \mid X_1 \sim W_i \mathcal{E}(X_1 + \epsilon_i), (2 \leq i \leq 20)$, $\epsilon_i \sim \mathcal{E}(0.1)$; Simulation-3 is a mixed version:$X_1 \sim \mathcal{U}(0, 10), X_i \mid X_1 \sim \frac{1}{2}\mathcal{N}(W_i X_1, 6I) + \frac{1}{2} W_i \mathcal{E}(X_1 + \epsilon_1), (2 \leq i \leq 20)$.

Simulation-4 $\sim$ Simulation-6 :

The true Chow-Liu tree is a tree of depth two. We conduct all experiments for 10 times, each time with random simulated orthogonal matrices $\{W_i\}_{i=2}^{7}$. The ground-truth Chow-Liu is a tree of depth two. Simulation-4: $X_1 \sim \mathcal{U}(0, 10)$, $X_i \mid X_1 \sim \mathcal{N}(W_i X_1, 2I)(i = 2, 3)$, $X_i \mid X_2 \sim \mathcal{N}(W_i X_2, 2I)(i = 4, 5)$, $X_i \mid X_3 \sim \mathcal{N}(W_i X_3, 2I)(i = 6, 7)$; Simulation-5: $X_1 \sim \mathcal{U}(0, 10)$, $X_i \mid X_1 \sim \mathcal{E}(X_1 + \epsilon_i)(i = 2, 3)$, $X_i \mid X_2 \sim W_i \mathcal{E}(X_2 + \epsilon_i)(i = 4, 5)$, $X_i \mid X_3 \sim W_i \mathcal{E}(X_3 + \epsilon_i)(i = 6, 7)$, $\epsilon_i \sim \mathcal{E}(0.1)$; Simulation-6 is a mixed version: $X_1 \sim \mathcal{U}(0, 10)$, $X_i \mid X_1 \sim W_i \mathcal{E}(X_1 + \epsilon_i)(i = 2, 3)$, $X_i \mid X_2 \sim \mathcal{N}(W_i X_2, 2I)(i = 4, 5)$, $X_i \mid X_3 \sim \mathcal{N}(W_i X_3, 2I)(i = 6, 7)$,$\epsilon_i \sim \mathcal{E}(0.1)$.

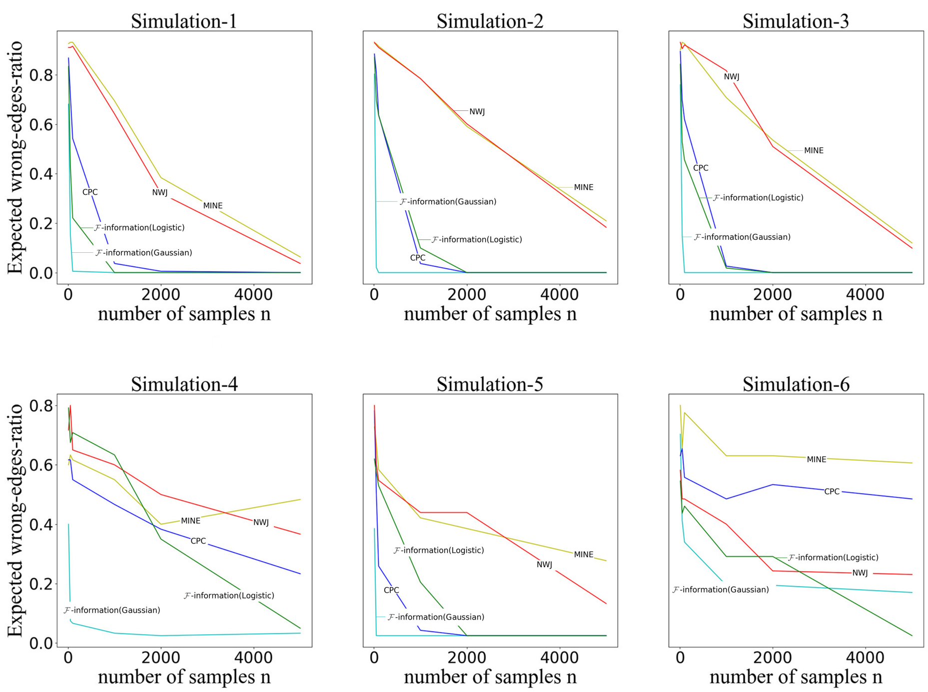


Figure 2: Chow-Liu Tree Construction: The expected wrong-edges-ratio of algorithm 1 with different $\mathcal{F}$ and other mutual information estimators-based algorithms from sample size 10 to 5×10^3.

F.2 Fairness

For the (F_i, F_j) elements described in the main body, please refer to figure 3b. The three datasets are: the UCI Adult dataset[2] has gender as the sensitive attribute; the UCI German credit dataset[3] has age as the sensitive attribute and the Heritage Health dataset[4] has the 18 configurations of ages and gender as the sensitive attribute.

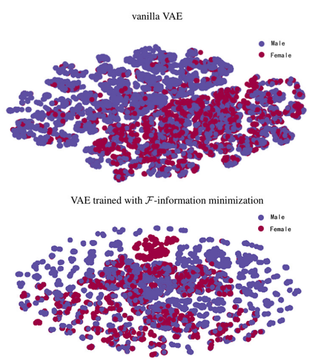
The models in the figure are:

[2] https://archive.ics.uci.edu/ml/datasets/adult
[3] https://archive.ics.uci.edu/ml/datasets
[4] https://www.kaggle.com/c/hhp

$\mathcal{F}_A = \{f : \mathcal{Z} \to \mathcal{P}(\mathcal{U}) | f[z](u) = \sum_{(z_i,u_i)\in\mathcal{D}} \frac{e^{\|z_i - z\|_2^2/h}}{\sum_{(z_i,u_i)\in\mathcal{D}} e^{\|z_i - z\|_2^2/h}} * \mathbb{I}(u_i = u), h \in \mathbb{R}\}$,where the $\mathcal{D}$ is the training set.

$\mathcal{F}_B = \{f : f[z] = \mathrm{softmax}(g(z))\}$,where g is a two layers MLP with Relu as the activation function.

$\mathcal{F}_C = \{f : f[z] = \mathrm{softmax}(g(z))\}$, where g is a three layers MLP with LeakyRelu as the activation function. We further visualize a special case of the $(\mathcal{F}_i, \mathcal{F}_j)$ pair in figure 3a, where the $\mathcal{F}_i = \{f : \mathcal{Z} \to \mathcal{P}(\mathcal{U}) | f[z](u) = \sum_{(z_i,u_i)\in\mathcal{D}} \frac{e^{\|z_i - z\|_2^2/h}}{\sum_{(z_i,u_i)\in\mathcal{D}} e^{\|z_i - z\|_2^2/h}} * \mathbb{I}(u_i = u), h \in \mathbb{R}\}$ explicitly makes the features of different sensitivity attributes more evenly spread, and functions in $\mathcal{F}_j$ is a simple two layers MLP with softmax at the top. The leaned features by $\mathcal{F}_i$-information minimization appear more evenly spread as expected, however, the attacker functions in $\mathcal{F}_j$ can still achieve a high AUC of 0.857.



(a) t-sne visualization

Adult

	$\mathcal{F}_A$	$\mathcal{F}_B$	$\mathcal{F}_C$
$\mathcal{F}_A$	**0.667**	0.678	0.857
$\mathcal{F}_B$	0.675	**0.671**	0.842
$\mathcal{F}_C$	0.673	0.682	**0.668**

German

	$\mathcal{F}_A$	$\mathcal{F}_B$	$\mathcal{F}_C$
$\mathcal{F}_A$	**0.669**	0.705	0.711
$\mathcal{F}_B$	0.591	**0.583**	0.601
$\mathcal{F}_C$	0.600	0.607	**0.587**

Health

	$\mathcal{F}_A$	$\mathcal{F}_B$	$\mathcal{F}_C$
$\mathcal{F}_A$	**0.764**	0.776	0.773
$\mathcal{F}_B$	**0.628**	0.630	0.635
$\mathcal{F}_C$	0.647	0.635	**0.634**

(b) AUC of predicted sensitive attribute

Figure 3: (a) The t-sne visualization (van der Maaten & Hinton, 2008) of latent features of models trained on UCI Adult dataset with/without $\mathcal{F}$-information minimization, where the $\mathcal{F}$ is specified above. (b) The AUC of predicted sensitive attribute on different dataset. The ($\mathcal{F}_i$,$\mathcal{F}_j$) element is the test AUC of the functions in $\mathcal{F}_j$ can achieve on the features that are obtained by $\mathcal{F}_i$-information minimization.

G MINIMALITY OF PREDICTIVE FAMILY

Define $\mathcal{F}_{\mathcal{X}\to\mathcal{P}(\mathcal{Y})} = \{g : \mathcal{X} \to \mathcal{P}(\mathcal{Y}) | \exists f \in \mathcal{F}, \forall x \in \mathcal{X}, g[x] = f[x]\}$. Similarly define $\mathcal{F}_{\varnothing\to\mathcal{P}(\mathcal{Y})} = \{g : \varnothing \to \mathcal{P}(\mathcal{Y}) | \exists f \in \mathcal{F}, \forall x \in \mathcal{X}, g[\varnothing] = f[\varnothing]\}$. Intuitively, $\mathcal{F}_{\mathcal{X}\to\mathcal{P}(\mathcal{Y})}$ (resp. $\mathcal{F}_{\varnothing\to\mathcal{P}(\mathcal{Y})}$) restricts the domain of functions in $\mathcal{F}$ to $\mathcal{X}$ (resp. $\varnothing$).

Non-Negativity As we demonstrated in Proposition 2, optional-ignorance guarantees that information will be non-negativity for any X and Y. Conversely, given any discrete X, Z, $\mathcal{F}_{\varnothing\to\mathcal{P}(\mathcal{Y})}$, $\mathcal{F}_{\mathcal{X}\to\mathcal{P}(\mathcal{Y})}$ not satisfying optional-ignorance, there exists distribution X, Y such that $I_{\mathcal{F}}(X \to Y) < 0$. Choose $Y \sim f^*[\varnothing]$ where f^* is the function that has no correspondent $g \in \mathcal{F}_{\mathcal{X}\to\mathcal{P}(\mathcal{Y})}$ that can ignore it's inputs. Pick X as the uniform distribution, and note that for all $g \in G$, there exists some measurable subset $X' \subset X$ on which g will produce a distribution unequal to $f^*[\varnothing]$, and therefore having higher crossentropy. The expected crossentropy expressed in $H_{\mathcal{F}_{\mathcal{X}\to\mathcal{P}(\mathcal{Y})}}(Y|X)$ is thus higher than in $H_{\mathcal{F}_{\varnothing\to\mathcal{P}(\mathcal{Y})}}(Y)$, and $I_{\mathcal{F}}(X \to Y) < 0$.

Independence Given any discrete $X, Y, \mathcal{F}_{\varnothing\to\mathcal{P}(\mathcal{Y})}, \mathcal{F}_{\mathcal{X}\to\mathcal{P}(\mathcal{Y})}$ not satisfying optional-ignorance, there exists an independent X, Y such that $I_{\mathcal{F}}(X \to Y) > 0$. Choose Y to be the distribution that can be expressed as $g[x]$ for some $x \in X, g \in \mathcal{F}_{\mathcal{X}\to\mathcal{P}(\mathcal{Y})}$, but which cannot be expressed by any $f \in \mathcal{F}_{\varnothing\to\mathcal{P}(\mathcal{Y})}$. Let X be the distribution with all it's mass on x, and note that the crossentropy of Y with $g[x]$ will be less than that of the function $f[\varnothing]$, which differs on a measurable subset. Thus, the optional-ignorance is the least we could ask of function classes in order for the $\mathcal{F}$-Information to be both nonnegative and take the value 0 when two distributions are independent.