Ordered $\mathcal{V}$-information Growth: A Fresh Perspective on Shared Information

Rohan Ghosh
National University of Singapore

Mehul Motani
National University of Singapore

Abstract

Mutual information (MI) is widely employed as a measure of shared information between random variables. However, MI assumes unbounded computational resources—a condition rarely met in practice, where predicting a random variable Y from X must rely on finite resources. $\mathcal{V}$-information addresses this limitation by employing a predictive family $\mathcal{V}$ to emulate computational constraints, yielding a directed measure of shared information. Focusing on the mixed setting (continuous X and discrete Y), here we highlight the upward bias of empirical $\mathcal{V}$-information, $\widehat{I}_{\mathcal{V}}(X \to Y)$, even when $\mathcal{V}$ is low-complexity (e.g., shallow neural networks). To mitigate this bias, we introduce $\mathcal{V}$-Information Growth (VI-Growth), defined as $\widehat{I}_{\mathcal{V}}(X \to Y) - \widehat{I}_{\mathcal{V}}(X' \to Y')$, where $X', Y' \sim P_X P_Y$ represent independent variables. While VI-Growth effectively counters over-estimation, more complex predictive families may lead to under-estimation. To address this, we construct a sequence of predictive families $\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}$ of increasing complexity and compute the maximum of VI-Growth across these families, yielding the ordered VI-Growth (O-VIG). We provide theoretical results that justify this approach, showing that O-VIG is a provably tighter lower bound for the true $\mathcal{V}$-Information than empirical $\mathcal{V}$-Information itself, and exhibits stronger convergence properties than $\mathcal{V}$-Information. Empirically, O-VIG alleviates bias and consistently outperforms state-of-the-art methods in both MI estimation and dataset complexity estimation, demonstrating its practical utility.

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

1 INTRODUCTION

Given two random variables (RVs) X and Y, and a sampled dataset $S = \{(X_1, Y_1), ..., (X_m, Y_m)\}$ from $P(X, Y)$ (or P_{XY}), a measure of particular interest is the amount of information shared between them. Throughout the years, many different measures of shared information have been proposed for both discrete and continuous RVs. Shared information has been of extensive interest across many fields (Graham et al., 2013; David et al., 2004; Church and Hanks, 1990; Tschannen et al., 2020; Belghazi et al., 2018; Kraskov et al., 2004; Yu et al., 2021), and has found substantial interest in machine learning (Shwartz-Ziv and Tishby, 2017; Saxe et al., 2019; Hjelm et al., 2019; Kleinman et al., 2023), particularly in the context of deep neural networks. An example of this is in the context of feature representation learning, by maximizing measures of shared information between the feature and the input or the output during training (Hjelm et al., 2019; Chalk et al., 2016). They find that controlling the level of information in the features can yield positive improvements in many scenarios.

Measuring shared information in the case when either or both variables are continuous has been more challenging, due to a few specific reasons. First, continuous RVs are of infinite cardinality, and thus unlike their finite counterparts, one cannot simply estimate the joint probabilities for all combinations of X and Y and obtain shared information. Second, in order to then successfully estimate a measure of shared information for continuous RVs, one will need to guess the joint probability $P(x, y)$ by extrapolating beyond the samples in the training dataset S, which represents only a finite slice of the potentially infinite possible valid co-occurrences of X and Y. This would require imposing some assumptions on the joint distribution, such as its level of smoothness (e.g., via Lipschitz constants) and other similar constraints. Without such constraints, it will not be possible to generate measures of shared information for continuous RVs.

For continuous RVs, a well-studied measure of shared information is the classical mutual information (MI) proposed by Shannon, which has seen many different approaches of estimation. Some of them include neural network based approaches (Belghazi et al., 2018; Oord et al., 2018; Song and Ermon, 2020), nearest neighbor based (Kraskov et al., 2004; Ross, 2014), kernel methods (Moon et al., 1995) and binning and its variants (Endres and Foldiak, 2005). Each approach essentially imposes its own set of inductive biases on the joint distribution P_{XY}. For some approaches, these biases are much more explicit, e.g. nearest neighbor approaches and binning approaches where the distribution is assumed to be locally smooth. For others, the inductive biases cannot be clearly outlined and is rather implicit, e.g., neural network based approaches.

Recent work in (Xu et al., 2020) generalizes the notion of constraints imposed in the process of estimating any shared information measure, by incorporating the constraints in an explicit manner, via a function class $\mathcal{V}$. Their first observation was that all inductive biases and constraints involved in the estimation of MI are artificial, i.e. classical MI actually assumes no constraints on the joint distribution. They use this observation to propose a new, constraint-aware measure of information, by constructing a function class $\mathcal{V}$ which predicts $P(Y|X)$ given an input X. They note that when $\mathcal{V}$ represents the universal set, it yields MI itself. Via this interpretation, MI appears to be a measure of shared information which imposes no *computational constraints* on $\mathcal{V}$. Intuitively, this is reasonable, as true MI does not have any *prior* preference of any conditional distribution $P_1(Y|X)$ over another $P_2(Y|X)$. Using this, they define the notion of $\mathcal{V}$-information, which estimates the shared information when the means to predict $P(Y|X)$ from X is restricted to the functions within $\mathcal{V}$. The authors then motivate the need for explicitly quantifying the computational constraints via the function class $\mathcal{V}$, and argue that in most practical scenarios, the complexity of $\mathcal{V}$ should be bounded.

Thus, as most approaches impose some assumptions on the joint probability distribution $P(X, Y)$ implicitly, $\mathcal{V}$-information ($\mathcal{V}$-I) represents a framework where these assumptions are imposed on $P(Y|X)$ explicitly via the computational function class $\mathcal{V}$. It is an elegant way of representing the shared information between two variables, and has desirable properties expected of such a measure. The authors in Xu et al. (2020) also propose an empirical extension of $\mathcal{V}$-Information called empirical $\mathcal{V}$-information, and they find that the higher the complexity of $\mathcal{V}$, the harder it is to accurately estimate the true $\mathcal{V}$-information.

Here, we identify two potential biases in the estimation of $\mathcal{V}$-Information via empirical $\mathcal{V}$-Information. First, we note that empirical $\mathcal{V}$-Information will likely overestimate the true $\mathcal{V}$-Information, especially for neural network based $\mathcal{V}$. One of the reasons for this is that neural networks have the ability to overfit, and even shallow neural networks can usually fit noisy labels perfectly (Zhang et al., 2017). To address this *over-estimation* bias, we first propose a new measure called $\mathcal{V}$-Information Growth (VI-Growth or VIG), which de-biases $\mathcal{V}$-information by subtracting the $\mathcal{V}$-Information between X' and Y' when $X', Y' \sim P_X P_Y$, thus $X' \perp Y'$. This ensures that VI-Growth does not confound the ability of $\mathcal{V}$ to predict noisy outputs in its estimate. However, we find that VI-Growth itself can suffer from an *underestimation* bias for high complexity $\mathcal{V}$, as high complexity $\mathcal{V}$ can explain noise almost as equally well as the given data, leading to underestimated measures. To address this, we propose a construction of an ordered set of predictive families $\mathcal{V}_1, \mathcal{V}_2, .., \mathcal{V}$ which are of ascending complexity. We then estimate the maximum VI-Growth among all $\mathcal{V}_i$ in this ordered set, and denote the subsequent measure as ordered VI-Growth (O-VIG). We find that ordered VI-Growth has many interesting and useful properties, and is a robust measure for both MI estimation and dataset complexity estimation.

Contributions: These are the specific contributions of our work:

1. We propose a novel measure of shared information called ordered $\mathcal{V}$-Information growth , which is motivated from $\mathcal{V}$-information. O-VIG is motivated from a two-step bias reduction of empirical $\mathcal{V}$-Information.
2. We find that O-VIG shares many of the desirable properties of $\mathcal{V}$-Information.
3. We find that O-VIG relates to MI and $\mathcal{V}$-Information, and can be used as an estimate for either, depending on the complexity of $\mathcal{V}$.
4. For the independent variable case, we find that O-VIG has provably stronger convergence than $\mathcal{V}$-Information.
5. Extensive experiments across synthetic, vision and tabular datasets showcase the benefits of the proposed approach. For MI estimation and dataset difficulty estimation, O-VIG improves on all other compared measures in almost all scenarios.

2 BACKGROUND

2.1 $\mathcal{V}$-Information

We provide the definitions for the predictive family $\mathcal{V}$, conditional $\mathcal{V}$-entropy and $\mathcal{V}$-information.

Definition 1 (**Predictive Family**). *Let* $\Omega = \{f : \mathcal{X} \cup \varnothing \rightarrow P(\mathcal{Y})\}$. *Here* $\varnothing$ *represents a null input*

that provides no information about Y. The predictive family $\mathcal{V} \subseteq \Omega$ is defined such that $\forall f \in \mathcal{V}$ and $\forall P \in \text{range}(f)$, $\exists f' \in \mathcal{V}$ such that $\forall x \in \mathcal{X}, f'[x] = P, f'[\varnothing] = P$.

Definition 2 (**Conditional $\mathcal{V}$-entropy**). *Let $(X,Y) \sim P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. We are given predictive family $\mathcal{V} \subseteq \Omega = \{f : \mathcal{X} \cup \varnothing \rightarrow P(\mathcal{Y})\}$. The predictive conditional $\mathcal{V}$-entropy is defined as:*

$$H_{\mathcal{V}}(Y|X) = \inf_{f \in \mathcal{V}} \mathbb{E}_{(X,Y)\sim P_{XY}}[-\log f[X](Y)] \quad (1)$$

$$H_{\mathcal{V}}(Y|\varnothing) = \inf_{f \in \mathcal{V}} \mathbb{E}_{Y\sim P_{Y}}[-\log f[\varnothing](Y)] \quad (2)$$

$H_{\mathcal{V}}(Y|\varnothing)$ is also called the $\mathcal{V}$-entropy and denoted as $H_{\mathcal{V}}(Y)$ for simplicity.

Definition 3. ***($\mathcal{V}$-information)*** *We are given predictive family $\mathcal{V}$. Then the $\mathcal{V}$-information from X to Y is defined as:*

$$I_{\mathcal{V}}(X \rightarrow Y) := H_{\mathcal{V}}(Y) - H_{\mathcal{V}}(Y|X) \quad (3)$$

Definition 4 (**Empirical Conditional $\mathcal{V}$-entropy and Empirical $\mathcal{V}$-Information**). *Let $(X,Y) \sim P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. We are given predictive family $\mathcal{V} \subseteq \Omega = \{f : \mathcal{X} \cup \varnothing \rightarrow P(\mathcal{Y})\}$. Let $S = \{(X_1,Y_1),...,(X_m,Y_m)\}$ represent m samples independently drawn from P_{XY}. The empirical conditional $\mathcal{V}$-entropy is defined as: $\hat{H}_{\mathcal{V}}(Y|X;S) = \inf_{f\in\mathcal{V}} \frac{1}{m}\sum_{i=1}^{m}[-\log f[X_i](Y_i)]$. Let us define $\hat{H}_{\mathcal{V}}(Y|\varnothing;S) = \inf_{f\in\mathcal{V}} \frac{1}{m}\sum_{i=1}^{m}[-\log f[\varnothing](Y_i)]$. The empirical $\mathcal{V}$-Information is then defined as: $\hat{I}_{\mathcal{V}}(X \rightarrow Y;S) = \hat{H}_{\mathcal{V}}(Y|\varnothing;S) - \hat{H}_{\mathcal{V}}(Y|X;S)$.*

Remark 1. *Note that the computational constraints that are embodied in $\mathcal{V}$ are primarily constraints on the set of conditional probability functions $P(Y|X)$ that can be modelled by $\mathcal{V}$. The overall constraints imposed on the estimated measure is therefore primarily decided by how flexible this set is. Also, note that when $\mathcal{V}$ can model any conditional distribution $P(Y|X)$ (i.e., is the universal set), the expression simply yields MI itself (Xu et al., 2020).*

2.2 Assumptions

In our work, we only consider mixed settings, i.e., X is continuous and Y is discrete. Also, we only consider $\mathcal{V}$ which are: (i) **Null-complete:** $\mathcal{V}$ can predict any probability distribution with null input $\varnothing$, and (ii) **P-Progressive:** $\mathcal{V}$ is a P-progressive predictive family, where P denotes the input distribution. We provide the formal definition of P-progressive predictive families below.

Definition 5. ***(P-Progressive Predictive Family)*** *We are given random variables $(X,Y) \sim P(X,Y)$. Let $S_m \sim P^m(X,Y)$ and $S'_m \sim (P(X)P(Y))^m$, and similarly $S_{m-1} \sim P^{m-1}(X,Y)$ and $S'_{m-1} \sim (P(X)P(Y))^{m-1}$, where $P^m(X,Y)$ and $(P(X)P(Y))^m$ denote m independently sampled instances from the respective distributions $P(X,Y)$ and $P(X)P(Y)$. $\mathcal{V}$ is P-progressive, if $\forall m$ we have*

$$\mathbb{E}_{S'_m}[\hat{H}_{\mathcal{V}}(Y|X;S'_m) - \hat{H}_{\mathcal{V}}(Y|X;S_m)] \geq$$
$$\mathbb{E}_{S_m}[\hat{H}_{\mathcal{V}}(Y|X;S'_{m-1}) - \hat{H}_{\mathcal{V}}(Y|X;S_{m-1})] \quad (4)$$

Note that all neural network $\mathcal{V}$ which have weights and biases in its layers are null-complete.

3 THE PROPOSED MEASURE

In this section, we first provide the motivation for ordered VI-Growth. After that, we provide the definition of our proposed VI-Growth measure.

3.1 Motivation

Given two RVs $X, Y \sim P_{XY}$, we start by considering the problem of estimating the $\mathcal{V}$-Information $I_{\mathcal{V}}(X \rightarrow Y)$ given the dataset S, where $S \sim P_{XY}^m$. We denote the empirical $\mathcal{V}$-information proposed in (Xu et al., 2020) by $\widehat{I}_{\mathcal{V}}(X \rightarrow Y;S)$. With this, we outline the following two-step motivation that ultimately yields the notion of ordered VI-Growth.

1. **Noise-Fitting Bias:** If X and Y are independent, the estimated amount of shared information between them must be as small as possible. This should be the case irrespective of the computational constraints imposed by $\mathcal{V}$. However, we find that there is a concrete bias in $\widehat{I}_{\mathcal{V}}(X \rightarrow Y;S)$ when X is independent of Y. When there are fewer computational constraints, and thus $\mathcal{V}$ is of high complexity, $\widehat{I}_{\mathcal{V}}(X \rightarrow Y;S)$ can become significantly larger than the ground truth $I_{\mathcal{V}}(X \rightarrow Y)$, and vice versa (see Figure 4 top row). We call this the *noise-fitting bias*. This is because with fewer computational constraints and thus more computational flexibility, one can still learn functions within the predictive family $\mathcal{V}$ which can relate independent variables within the finite dataset S. A known example of this is the ability of classifiers to fit independent, random labels equally well when compared to the given data, even with shallow low-complexity classifiers (Zhang et al., 2017). Thus, in the independent variable case, $\widehat{I}_{\mathcal{V}}(X;Y)$ will be primarily reflective of the complexity of $\mathcal{V}$, thus yielding a biased estimate. To address this, we first measure the $\mathcal{V}$-information between X and Y. Then, we simulate a scenario where X and Y are independent and generate $X', Y' \sim P_X P_Y$, according to the product of their marginal distributions. We then obtain VI-Growth by subtracting $\widehat{I}_{\mathcal{V}}(X';Y')$ from $\widehat{I}_{\mathcal{V}}(X;Y)$, i.e., $\widehat{IG}_{\mathcal{V}}(X \rightarrow$

$Y; S, S') = \widehat{I}_{\mathcal{V}}(X \to Y; S) - \widehat{I}_{\mathcal{V}}(X' \to Y'; S')$, where $X', Y' \sim P_X P_Y$ and $S' \sim (P_X P_Y)^m$. In practice the independent distribution X', Y' can be simulated by randomly shuffling the labels in (X, Y), yielding independent variables. VI-Growth measures the degree to which the shared $\mathcal{V}$-information between X and Y increases, when compared with the case when Y and X are independent and follow their marginal distributions. As we show later, by doing so, we obtain a measure that avoids the noise-fitting bias of $\mathcal{V}$-information.

2. **Under-estimation bias:** When there are very few computational constraints (i.e., high complexity $\mathcal{V}$), we expect the ability of $\mathcal{V}$ to relate both the independent X' to Y' and X to Y to be high. This yields a very low VI-Growth, as $\widehat{I}_{\mathcal{V}}(X \to Y; S)$ and $\widehat{I}_{\mathcal{V}}(X' \to Y'; S')$ will both be large. To avoid this under-estimation bias, we first realize that the true $\mathcal{V}$-Information can be lower bounded by VI-Growth (see property **P4**). Thus, we have $IG^m_{\mathcal{V}}(X \to Y) \leq I_{\mathcal{V}}(X \to Y)$, where $IG^m_{\mathcal{V}}(X \to Y)$ denotes the aggregate VI-Growth between X and Y. Now, we construct a sequence of predictive families $\boldsymbol{\mathcal{V}} = \{\mathcal{V}_1, \mathcal{V}_2, ..., \mathcal{V}_k\}$, such that $\mathcal{V}_1 \subset \mathcal{V}_2 \subset ... \subset \mathcal{V}_k$ and $\mathcal{V}_k = \mathcal{V}$. Now, let

$$M = \max\{IG^m_{\mathcal{V}_1}(X \to Y), .., IG^m_{\mathcal{V}_k}(X \to Y)\}.$$

Following the previous argument, we can express

$$\begin{aligned} M &\leq \max\{I_{\mathcal{V}_1}(X \to Y), .., I_{\mathcal{V}_k}(X \to Y)\} \\ &\leq I_{\mathcal{V}_k}(X \to Y) = I_{\mathcal{V}}(X \to Y). \end{aligned} \tag{5}$$

Thus, $\max\{IG^m_{\mathcal{V}_1}(X \to Y), ..., IG^m_{\mathcal{V}_k}(X \to Y)\}$ can be construed as a lower bound for the true $\mathcal{V}$-Information, and can be a potentially tighter estimate than the VI-Growth of a single predictive family $\mathcal{V}$. Furthermore, this measure can avoid the underestimation bias of VI-Growth as well. We call this measure the ordered $\mathcal{V}$-Information growth.

3.2 $\mathcal{V}$-Information Growth

We are given RVs $X \sim P(X) \in \mathbb{R}^d$ and $Y \in \{0, 1, 2, ..., c\}$, and $X, Y \sim P(X, Y)$. Samples from $P(X, Y)$ are represented as $S = \{(X_1, Y_1), .., (X_m, Y_m)\} \sim P(X, Y)$ and from $P(X)P(Y)$ are represented as $S' = \{(X'_1, Y'_1), .., (X'_m, Y'_m)\} \sim P(X)P(Y)$.

First, we define the aggregate and empirical variants of VI-Growth.

Definition 6. *($\mathcal{V}$-Information Growth) Let $X, Y \sim P_{XY}$ and $X', Y' \sim P_X P_Y$. Let $S \sim P^m_{XY}$ and $S' \sim (P_X P_Y)^m$. The aggregate VI-Growth between X and Y is defined as $IG^m_{\mathcal{V}}(X \to Y) = \mathbb{E}_{S \sim P_{XY}, S' \sim P_X P_Y}\left[\widehat{I}_{\mathcal{V}}(X \to Y; S) - \widehat{I}_{\mathcal{V}}(X' \to Y'; S')\right]$. The empirical VI-Growth is $\widehat{IG}_{\mathcal{V}}(X \to Y; S, S') = \widehat{I}_{\mathcal{V}}(X \to Y; S) - \widehat{I}_{\mathcal{V}}(X' \to Y'; S')$.*

Next, we define the notion of a $\mathcal{V}$-ordered set as follows, which is subsequently used in defining ordered VI-Growth.

Definition 7. *($\mathcal{V}$-ordered set) Given a predictive family $\mathcal{V}$, the set $\boldsymbol{\mathcal{V}} = \{\mathcal{V}_1 \subset \mathcal{V}_2 \subset ... \subset \mathcal{V}\}$, where each $\mathcal{V}_i$ is a predictive family, is a $\mathcal{V}$-ordered set.*

With this, we can define O-VIG variants as follows.

Definition 8. *(Ordered $\mathcal{V}$-Information Growth) The aggregate ordered VI-Growth between X and Y is defined as $IG^m_{\{\boldsymbol{\mathcal{V}}\}}(X \to Y) = \max_{\mathcal{V}_i \in \boldsymbol{\mathcal{V}}} IG^m_{\mathcal{V}_i}(X \to Y)$. The empirical ordered VI-Growth between them is estimated as $\widehat{IG}_{\{\boldsymbol{\mathcal{V}}\}}(X \to Y; S, S') = \max_{\mathcal{V}_i \in \boldsymbol{\mathcal{V}}} \widehat{IG}_{\mathcal{V}_i}(X \to Y; S, S')$ where $S \sim P^m_{XY}$ and $S' \sim (P_X P_Y)^m$.*

4 THEORETICAL RESULTS

4.1 Properties of VI-Growth and Ordered VI-Growth

First, we outline some properties of VIG and O-VIG as follows. For all following results, $S = \{(X_1, Y_1), ..., (X_m, Y_m)\} \sim P^m_{XY}$ and $S' = \{(X'_1, Y'_1), ..., (X'_m, Y'_m)\} \sim (P_X P_Y)^m$.

P1 (Independence) If X is independent of Y, then $IG^m_{\mathcal{V}}(X \to Y) = IG^m_{\mathcal{V}}(Y \to X) = 0$

P2 (Limits) $\lim_{|S|, |S'| \to \infty} \widehat{IG}_{\mathcal{V}}(X \to Y; S, S') = \lim_{|S|, |S'| \to \infty} \widehat{IG}_{\{\boldsymbol{\mathcal{V}}\}}(X \to Y; S, S') = I_{\mathcal{V}}(X \to Y)$

P3 (Non-negativity) $IG^m_{\mathcal{V}}(X \to Y) \geq 0$ and $IG^m_{\{\boldsymbol{\mathcal{V}}\}}(X \to Y) \geq 0$.

P4 (Upper Bounds) $IG^m_{\mathcal{V}}(X \to Y) \leq I_{\mathcal{V}}(X \to Y)$, $\widehat{IG}_{\mathcal{V}}(X \to Y; S, S') \leq \widehat{H}(Y; S)$, $IG^m_{\{\boldsymbol{\mathcal{V}}\}}(X \to Y) \leq I_{\mathcal{V}}(X \to Y)$ and $\widehat{IG}_{\{\boldsymbol{\mathcal{V}}\}}(Y \to X; S, S') \leq \widehat{H}(Y; S)$

P5 (Relation to Conditional $\mathcal{V}$-entropy) If S' is chosen such that $X'_i = X_i$ and $Y'_i = Y_{\sigma(i)}$, where $\sigma : \{1, ..., m\} \to \{1, ..., m\}$ represents a random permutation function, then $\widehat{IG}_{\mathcal{V}}(X \to Y; S, S') = \widehat{H}_{\mathcal{V}}(Y'|X; S') - \widehat{H}_{\mathcal{V}}(Y|X; S)$

Remark 2. *As **P5** shows, when S' is generated from S by simply random permuting the labels across samples, then the estimation of $\widehat{IG}_{\mathcal{V}}(X \to Y; S, S')$ only requires estimating two conditional $\mathcal{V}$-Information measures. As such, in our experiments, we estimate $\widehat{IG}_{\mathcal{V}}(X \to Y; S, S')$ this way. In what follows, we may refer to the randomly permuted Y by Y_σ.*

4.2 Relation to MI

Here, we provide our main theoretical results that eventually relate VI-Growth to mutual information.

Theorem 1. *Assume that* $\forall f \in \mathcal{V}$,$x \in \mathcal{X}$,$y \in \mathcal{Y}$ *we have* $\log f[x](y) \in [-B, B]$. *Then with probability* $p \geq 1-\delta$ *over the draw of* S *and* S', *we have*

$$I_{\mathcal{V}}\left(X \to Y\right) \geq \widehat{IG}_{\mathcal{V}}(X \to Y; S, S') - B\sqrt{\frac{4\log\frac{1}{\delta}}{|S|}} \quad (6)$$

With this, we have the following corollary that relates O-VIG to mutual information and $\mathcal{V}$-information itself, via a lower bound.

Corollary 1.1. *We are given a function class* $\mathcal{V}$. *Using this, we construct a set of function classes* $\boldsymbol{\mathcal{V}} = \{\mathcal{V}_1, \mathcal{V}_2, ..., \mathcal{V}\}$, *such that* $\{\mathcal{V}_1 \subset \mathcal{V}_2 \subset ... \subset \mathcal{V}\}$. *Assume that* $\forall f \in \mathcal{V}$,$x \in \mathcal{X}$,$y \in \mathcal{Y}$ *we have* $\log f[x](y) \in [-B, B]$. *Then with probability* $p \geq (1-\delta)$ *over the draw of* S *and* S', *we have*

$$I_{\mathcal{V}}(X;Y) \geq \widehat{IG}_{\boldsymbol{\{\mathcal{V}\}}}(Y \to X; S, S') - B\sqrt{\frac{4\log\frac{|\boldsymbol{\mathcal{V}}|}{\delta}}{|S|}} \quad (7)$$

Another consequence of Theorem 1 is regarding the rates of convergence of both VIG and O-VIG, when X is independent of Y.

Corollary 1.2. *We consider the setting of Theorem 1, where we assume that* X *is independent of* Y. *Then with probability* $p \geq (1-\delta)$ *over the draw of* S *and* S', *we have* $\widehat{IG}_{\mathcal{V}}(X \to Y; S, S') \leq B\sqrt{\frac{4\log\frac{1}{\delta}}{|S|}}$ & $\widehat{IG}_{\boldsymbol{\{\mathcal{V}\}}}(Y \to X; S, S') \leq B\sqrt{\frac{4\log\frac{|\boldsymbol{\mathcal{V}}|}{\delta}}{|S|}}$

Note that the convergence rates of $\mathcal{V}$-Information in the independent variable case follows directly from Theorem 1 of (Xu et al., 2020).

Corollary 1.3. ***(From Theorem 1 of (Xu et al., 2020))*** *In the same setting as Theorem 1, with probability at least* $1-2\delta$, *when* X *is independent of* Y, *we have* $|\widehat{I}_{\mathcal{V}}(X \to Y; S)| \leq 4\mathcal{R}_{|S|}(\mathcal{G}_{\mathcal{V}}) + B\sqrt{\frac{8\log\frac{2}{\delta}}{|S|}}$, *where* $\mathcal{G}_{\mathcal{V}} = \{g | g(x,y) = \log f[x](y), f \in \mathcal{V}\}$, *and* $\mathcal{R}_m(\mathcal{G})$ *denotes the Rademacher complexity of* $\mathcal{G}$ *with sample number* m.

Remark 3. *In the independent variable case, note that VIG and O-VIG have a smaller upper bound which does not contain the complexity term* $\mathcal{R}_m(\mathcal{G})$, *unlike* $\mathcal{V}$*-Information. This is mainly because the subtracted* $\widehat{I}_{\mathcal{V}}(X' \to Y'; S')$ *term in VIG effectively subsumes the Rademacher complexity term.*

5 EXPERIMENTS

We conduct various experiments to analyze different aspects of VIG and O-VIG, and compare with other information measures and dataset complexity measures. Note that for each experiment, the details regarding network configuration, training parameters, etc. are provided in the supplementary materials. In our experiments, we denote the sample size by m, and we refer to empirical $\mathcal{V}$-Information as $\mathcal{V}$-Information itself. For all O-VIG estimates, the size of the $\mathcal{V}$-ordered set $|\mathcal{V}|$ was fixed to 10. Code is available at https://github.com/kentridgeai/OVIG. More details on each experiment is provided in the appendix.

Datasets: We use the following datasets:

1. Real Datasets: MNIST (LeCun et al., 2010), CIFAR-10 (Krizhevsky et al., a), CIFAR-100 (Krizhevsky et al., b)
2. Synthetic Datasets: Our synthetic datasets are of the form (X, Y), where $X \in \mathbb{R}^d$ and $Y \in \{0, 1\}$. In what follows, we create 100 datasets (each of sample size m), from each configuration:

D1 ($d = 3$, $m = 200$) $P(X|Y=0) \sim \mathcal{N}(\mu_0, \Sigma_0)$ and $P(X|Y=1) \sim \mathcal{N}(\mu_1, \Sigma_1)$. Σ_0 and Σ_1 are randomly chosen. μ_0 and μ_1 are randomly generated.

D2 ($d = 3$, $m = 200$) $P(X|Y=0) \sim \mathcal{N}(\mu_0, \Sigma)$ and $P(X|Y=1) \sim \mathcal{N}(\mu_1, \Sigma)$, where Σ is randomly chosen. Rest the same as in **D1**.

D3 ($d = 10$, $m = 1000$) Same as **D2**.

D4 ($d = 100$, $m = 1000$) Same as **D2**, except μ_0, μ_1 are chosen such that $I(X;Y) = 0$ or $I(X;Y) = 1$.

VI-Growth Estimation Details: We construct the $\mathcal{V}$-ordered set required for the computation of VI-Growth by taking the base network and scaling the width of each layer by a factor of $\alpha \in \{0.1, 0.2, .., 1.0\}$. α is referred to as the **$\mathcal{V}$-complexity factor**. Note that the complexity of the networks corresponding to the $\mathcal{V}$-ordered set $\mathcal{V}_1, \mathcal{V}_2, ..\mathcal{V}_k$ grows from $\mathcal{V}_1$ to $\mathcal{V}_k$.

5.1 Notes on P-Progressiveness

First, we discuss the P-Progressive constraint in more detail, and verify whether it holds for neural networks for synthetic data distributions.

Note that $\mathbb{E}_{S'_m}\left[\widehat{H}_{\mathcal{V}}(Y|X; S'_m) - \widehat{H}_{\mathcal{V}}(Y|X; S_m)\right]$ is the aggregate VI-Growth $IG^m_{\mathcal{V}}(X \to Y)$ itself. So, alternatively, the P-Progressive constraint simply states that $IG^m_{\mathcal{V}}(X \to Y) \geq IG^{m-1}_{\mathcal{V}}(X \to Y)$, i.e. a monotonically increasing function of the sample size m. First, note that when $Y \perp X$, $IG^m_{\mathcal{V}}(X \to Y) = 0$ (Property **P1**), and thus the P-Progressive constraint holds. We only discuss the case when $Y \not\perp X$. We provide some intuitive arguments below why this should hold for most predictive families $\mathcal{V}$ and distributions P.

1. When $m = 1$, aggregate VIG will be 0, as $\mathbb{E}_{S'_1}\widehat{H}_{\mathcal{V}}(Y|X; S'_1) = \mathbb{E}_{S_1}\widehat{H}_{\mathcal{V}}(Y|X; S_1)$. This is because when minimizing the log-loss on a single sample, there is no difference between P_{XY} and $P_X P_Y$.

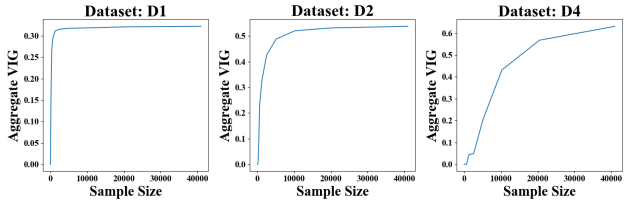

Figure 1: **Testing P-Progressiveness:** Figure showing the dependency of aggregate VIG on the sample size m, for three different datasets randomly chosen from the **D1**,**D2** and **D4** dataset collection respectively. The predictive family $\mathcal{V}$ here is a neural network with two hidden layers. Note that the dimensionality of these datasets are all different, being 3, 10 and 100 respectively. The aggregate VIG seems to always either grow or stay close to being level, which verifies the P-Progressiveness of $\mathcal{V}$ for these three cases.

2. It is easy to see that both $\mathbb{E}_{S'_m}\widehat{H}_{\mathcal{V}}(Y|X;S'_m)$ and $\mathbb{E}_{S_m}\widehat{H}_{\mathcal{V}}(Y|X;S_m)$ are individually monotonically increasing functions of m. Furthermore, the growth of $\mathbb{E}_{S'_m}\widehat{H}_{\mathcal{V}}(Y|X;S'_m)$ from $m = 1$ to $m = \infty$ is always larger than the growth of $\mathbb{E}_{S_m}\widehat{H}_{\mathcal{V}}(Y|X;S_m)$ from $m = 1$ to $m = \infty$. This is because, $\lim_{m\to\infty}\mathbb{E}_{S'_m}\widehat{H}_{\mathcal{V}}(Y|X;S'_m) = H(Y)$ and as $\mathbb{E}_{S_m}\widehat{H}_{\mathcal{V}}(Y|X;S_m) \leq H(Y|X) \leq H(Y)$, we will have that $\lim_{m\to\infty}\mathbb{E}_{S_m}\widehat{H}_{\mathcal{V}}(Y|X;S_m) \leq \lim_{m\to\infty}\mathbb{E}_{S'_m}\widehat{H}_{\mathcal{V}}(Y|X;S'_m)$. Furthermore, as shown in the previous point, $\mathbb{E}_{S_1}\widehat{H}_{\mathcal{V}}(Y|X;S_1) = \mathbb{E}_{S'_1}\widehat{H}_{\mathcal{V}}(Y|X;S'_1)$. Thus, the overall rate of growth of $\mathbb{E}_{S'_m}\widehat{H}_{\mathcal{V}}(Y|X;S'_m)$ is greater than or equal to the rate of growth of $\mathbb{E}_{S_m}\widehat{H}_{\mathcal{V}}(Y|X;S_m)$ w.r.t m.
3. When sampling from $X', Y' \sim P_X P_Y$, every new sample will be independent of all the previous samples as $Y' \perp X'$. Thus, from the perspective of the predictive family $\mathcal{V}$, it needs to do more to minimize the predictive log-loss $\widehat{H}_{\mathcal{V}}(Y|X;S'_m)$ with every additional sample. Whereas, when sampling from $X, Y \sim P_{XY}$, even though the new sample will be independent, the new label Y will share some functional similarities with the previous samples as $Y \not\perp X$, and thus most $\mathcal{V}$ should have less additional difficulty in minimizing the predictive log-loss $\widehat{H}_{\mathcal{V}}(Y|X;S_m)$ with every additional sample. Thus intuitively, $IG^m_{\mathcal{V}}(X \to Y)$ should increase with m.

Experiments: Next, we verify whether neural networks can be P-Progressive for the synthetic data distributions used here. We pick a random configuration from three scenarios outlined in section 6 of the main paper: **D1**, **D2** and **D4** dataset collections. Note that they are all different in dimensionality. We fix $\mathcal{V}$ to be a neural network configuration with two hidden layers containing 25 hidden neurons each.

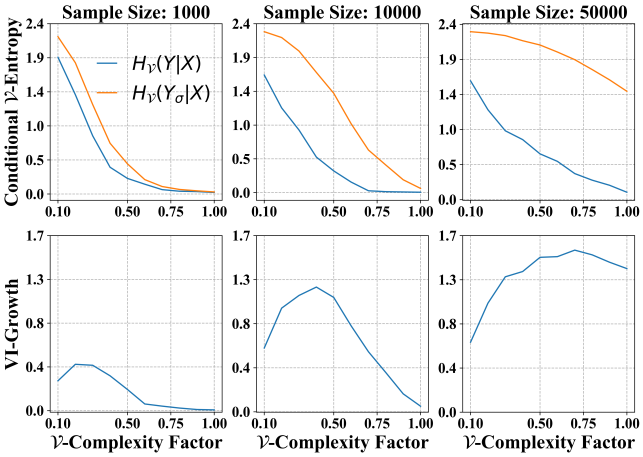

Figure 2: **(VIG-Trends, CIFAR-10)** Figure showing the trends of the various terms involved in the estimation of VIG and O-VIG, for the CIFAR-10 dataset, across three different sample sizes $m = 1000, 10000, 50000$. The top row shows the conditional $\mathcal{V}$-entropy measures, and the bottom row shows the corresponding estimated VI-Growth measures. The x-axis denotes the complexity factor of $\mathcal{V}_i$.

For each configuration, we estimate $IG^m_{\mathcal{V}}(X \to Y)$ for $m = 10 \times 2^i$, for $i \in \{1, 2, .., 12\}$, thus the sample size is within the range $10 \leq m \leq 40960$. For each sample size m, to estimate the aggregate VIG $IG^m_{\mathcal{V}}(X \to Y)$, we sample S_m and S'_m 20 times to estimate the average measures $\widehat{H}_{\mathcal{V}}(Y|X;S'_m)$ and $\widehat{H}_{\mathcal{V}}(Y|X;S_m)$. The results for each case is shown are Figure 1.

5.2 VI-Growth Trends

First, following property **P5**, we analyze the behaviour of various terms involved in the estimation of VIG and O-VIG. Y_σ here denotes the randomly permuted Y in **P5**. The analysis on CIFAR-10 is shown in Figure 2. There are three overall observations:

1. As expected, $H_{\mathcal{V}}(Y_\sigma|X)$ is always greater than or equal to $H_{\mathcal{V}}(Y|X)$. Thus, in this case, the estimated VI-Growth values are always positive.
2. $H_{\mathcal{V}}(Y_\sigma|X)$ increases with the sample size. This implies that the ability of the predictive family $\mathcal{V}$ to explain noisy labels reduces significantly with more data. The same is also observed to a certain extent for $H_{\mathcal{V}}(Y|X)$, but the effect is less pronounced.
3. VI-Growth is smaller at the extreme ends of complexity. For $\mathcal{V}_i$ of low complexity as well as $\mathcal{V}_i$ of very high complexity, it seems the difference between $H_{\mathcal{V}_i}(Y_\sigma|X)$ and $H_{\mathcal{V}_i}(Y|X)$ is small. Furthermore, the complexity-factor which leads to the maximum VI-Growth seems to increase with sample size. In fact, from **P3**, it follows that when $m \to \infty$, VI-Growth must be largest for the maximum complexity factor.

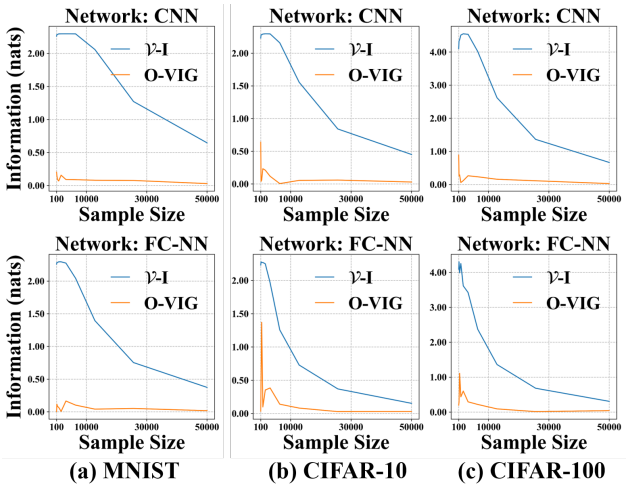

Figure 3: **(Convergence for Independent RVs)** Convergence Properties for Independent RVs ($I(X;Y) = 0$).

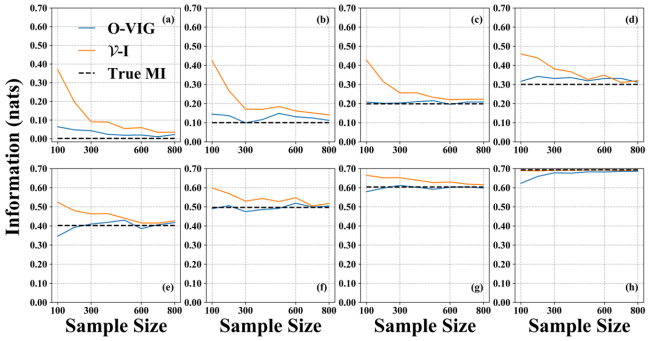

Figure 4: **(Convergence, D1 Dataset Collection)** Convergence behavior of $\mathcal{V}$-I and O-VIG biases on **D1** dataset. Values are reported after five-fold averaging, to yield the bias of the respective measures. The dotted line in each case represents the ground truth MI.

5.3 Convergence Analysis

Here we analyze the convergence properties of O-VIG and compare with empirical $\mathcal{V}$-Information. **P2** finds that VI-Growth converges to $\mathcal{V}$-information. We compare the convergence of O-VIG and empirical $\mathcal{V}$-information, to verify Corollary 1.2 and 1.3).

5.3.1 Independent Random Variables

To generate independent random variables, we take the datasets MNIST, CIFAR-10 and CIFAR-100, and randomly permute the labels to preserve the label probabilities. Then we measure the O-VIG between the input and the permuted labels. This is repeated across several sample sizes. The results are shown in Figure 3. We have two sets of results, for the case when $\mathcal{V}$ is a fully connected neural network (FC-NN) and when $\mathcal{V}$ is a convolutional neural network (CNN). In both cases, O-VIG converges to values very close to zero significantly earlier than $\mathcal{V}$-Information. In most cases $\mathcal{V}$-information does not reach zero, which highlights one of its biases as discussed in our motivation.

5.3.2 Dependent Random Variables

To generate random variables sharing varying degrees of dependency, we refer to the **D1** dataset collection. We pick eight distribution configurations from the 100 datasets in the collection, which have different ground truth MI, ranging from 0 to 0.69 (maximum in terms of nats). For each dataset, we then slowly increase the sample size and measure O-VIG and $\mathcal{V}$-Information, and plot the results in Figure 4. Results are five-fold averages to show the bias of tested metrics. We see that except for the single case when the ground truth MI is the highest, in all other cases O-VIG converges significantly earlier than $\mathcal{V}$-Information. Also, note that as our results are five-fold averages, they indicate the inherent bias in $\mathcal{V}$-Information, and the ability of O-VIG to cancel that bias to a significant degree.

5.4 MI Estimation Experiments

In this section, we analyze the accuracy of O-VIG as an MI estimator (using Neural Network predictive families), targeted towards small sample size scenarios. We compare the estimates with a suite of other MI estimators, which include both modern approaches, such as MINE (Belghazi et al., 2018), InfoNCE (Oord et al., 2018), SMILE (Song and Ermon, 2020), KSG-Mixed (Gao et al., 2017) and classical approaches such as KSG (Kraskov et al., 2004), JS and NWJ (Nguyen et al., 2010). We also compare with $\mathcal{V}$-information itself. Experiments are performed on the **D1**-**D4** dataset collections. Average error is reported over 100 trials. We report the root mean squared error (rmse), mean absolute error (mae) and the Spearman rank correlation with the ground truth MI in each case. The Spearman correlation measure is added to see if the estimated MI correlates with the ground truth MI even if the underlying estimation accuracy isn't good. We added this measure because many estimators are biased to output lower values for high dimensional input (e.g. KSG variants) and in those cases the Spearman correlation can reveal more information. As seen from Table 1, we see that O-VIG showcases significantly better MI estimation accuracy than all other compared measures in most cases. Notably, we find that the accuracy of estimators suffer significantly with more data dimensionality, but less so for O-VIG. Even for the challenging 100-dimensional dataset collection

Table 1: MI Estimation Experiments (d = dimension, m = Sample Size)

Dataset	Error Measure	KSG	KSG-Mixed	MINE	JS	Info-NCE	NWJ	SMILE	$\mathcal{V}$-I	O-VIG
D1	rmse	0.13	0.11	0.14	0.15	0.10	0.13	0.16	0.07	**0.05**
d=3	mae	0.10	0.08	0.11	0.14	0.08	0.10	0.14	0.05	**0.04**
m=200	spearman	0.77	0.77	0.30	0.75	0.75	0.86	0.71	**0.91**	**0.92**
D2	rmse	0.20	0.13	0.25	0.18	0.12	0.15	0.18	0.15	**0.09**
d=3	mae	0.12	0.08	0.22	0.14	**0.08**	**0.08**	0.14	0.12	**0.07**
m=200	spearman	0.79	0.83	0.74	0.81	0.89	0.84	0.89	**0.93**	**0.94**
D3	rmse	0.31	0.25	0.23	0.23	0.21	0.22	0.23	0.36	**0.15**
d=10	mae	0.22	0.21	0.21	0.20	0.17	0.20	0.20	0.32	**0.09**
m=1000	spearman	-0.01	0.12	0.36	0.62	0.57	0.62	0.65	0.82	**0.89**
D4	rmse	0.49	1.01	0.98	0.30	0.25	0.36	0.30	0.48	**0.18**
d=100	mae	0.34	2.1	0.50	0.27	0.21	0.26	0.27	0.34	**0.15**
m=1000	spearman	-0.19	-0.75	0.31	0.82	**0.85**	**0.84**	**0.85**	0.79	**0.84**

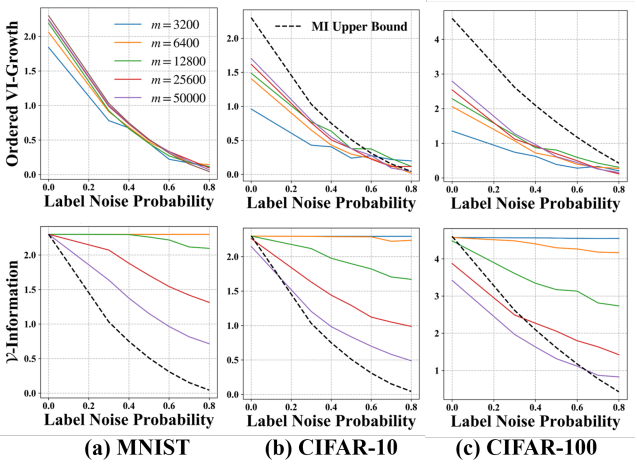

Figure 5: **(Label Noise)** Impact of Label Noise on $\mathcal{V}$-I and O-VIG.

in **D4**, we find that O-VIG performs reasonably well as an MI-Estimator.

5.5 Noise Addition Experiments

This section explores experiments where noise is added to the input X or targets Y in real datasets. Shared information measures are expected to decrease with added noise, as it reduces the shared information between X and Y. The experiments are discussed below.

Label Noise: Here, we note the trends of O-VIG and $\mathcal{V}$-I in response to label noise, where we randomly corrupt a proportion of the labels in the given sample. Let us denoted the corrupted labels by Y'. The results are shown in Figure 5. We also show the true upper bound on the mutual information $I(X;Y')$, which decreases with more noise probability as expected. We find that in all cases, O-VIG follows a similar trend to the MI upper bound. However, we find that, for smaller sample sizes $\mathcal{V}$-I does not always reduce with more label noise. This showcases that for smaller sample sizes, $\mathcal{V}$-I is unable to differentiate between the original and noisy data. Furthermore, the $\mathcal{V}$-I values in most cases turn out to be larger than the upper bound on MI, which again highlights its biases.

Gaussian-added Input Noise: We add Gaussian noise to the input and measure the O-VIG and $\mathcal{V}$-I values between $X' = X + \mathcal{N}(0, \sigma^2)$ and Y in each case. Similarly to the label noise experiments, we verify if the measures decrease with more noise. In this case, the amount of noise is controlled via the standard deviation σ. Results are shown in Figure 6. We note that O-VIG always shows a decreasing pattern, even for smaller sample size m. Interestingly, we find that $\mathcal{V}$-I increases for CIFAR-10 and CIFAR-100 datasets with more input noise. This indicates that in high dimensional spaces, neural networks may find it easier to fit datapoints which are further apart. Thus, our results show that $\mathcal{V}$-I doesn't respond desirably in this particular problem, for CIFAR-10 and CIFAR-100.

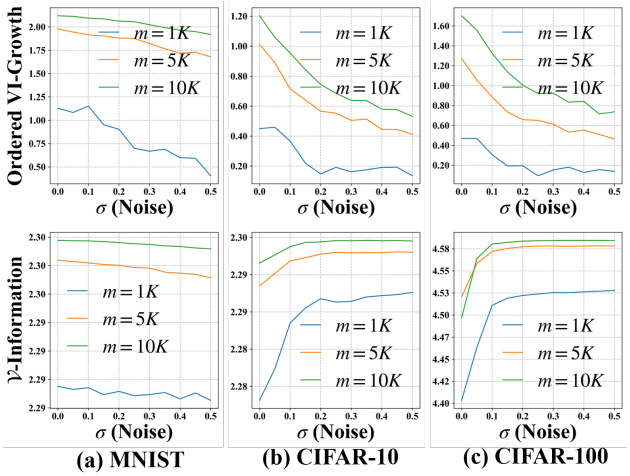

Figure 6: **(Input Gaussian Noise)** Impact of Gaussian Noise addition to Input, on $\mathcal{V}$-I and O-VIG.

Table 2: Spearman Correlation of Measures with Test Accuracy on Vision Datasets (CNN)

# Classes	Measure	MNIST	CIF-10	CIF-100	Combined
2	$\mathcal{V}$-I	-0.11	0.90	0.37	0.41
	O-VIG	**0.48**	**0.92**	**0.80**	**0.94**
	AOM	**0.48**	0.41	0.75	0.81
4	$\mathcal{V}$-I	-0.14	0.76	0.77	0.37
	O-VIG	0.01	**0.94**	**0.88**	**0.88**
	AOM	**0.52**	0.26	0.39	0.81
6	$\mathcal{V}$-I	-0.28	0.62	0.73	0.58
	O-VIG	0.07	**0.96**	**0.85**	**0.89**
	AOM	**0.24**	0.33	0.20	0.80

(a) Label-Subsets

# Classes	Measure	MNIST	CIF-10	CIF-100	Combined
2	$\mathcal{V}$-I	0.13	0.58	0.22	0.43
	O-VIG	**0.59**	**0.86**	**0.74**	**0.90**
	AOM	0.51	0.44	0.52	0.82
4	$\mathcal{V}$-I	0.05	0.65	0.57	0.55
	O-VIG	**0.71**	**0.92**	**0.83**	**0.88**
	AOM	0.66	0.77	0.57	0.76
6	$\mathcal{V}$-I	0.47	0.77	0.59	0.65
	O-VIG	**0.85**	**0.97**	**0.68**	**0.84**
	AOM	0.78	0.50	0.63	0.77

(b) Label-Subsets + Data-Size Variation

5.6 Dataset Complexity

Vision: We take random subsets of labels from MNIST, CIFAR-10 and CIFAR-100, and train CNNs of a fixed configuration. We record their test accuracies in each case. For each dataset, this is repeated for 30 runs. In each run, a different label subset is randomly chosen. We estimate O-VIG and $\mathcal{V}$-I with the same configuration of $\mathcal{V}$ as the neural network. We also include the best performance among eight benchmark data complexity measures from (Komorniczak and Ksieniewicz, 2023). The best performance in each case is represented by AOM (all other methods). We compute the Spearman correlation of the information measures with the observed test accuracy.

When no additional change is performed to the training data in each case, we show the correlation results in Table 2 (a). Except for the #4 and #6 classes cases for MNIST, we note that O-VIG better correlates to test accuracy than other approaches. For MNIST, in the #4 and #6 classes cases, we find that the test accuracy difference was significantly small ($< 0.5\%$) which made the problem harder. To get a larger range of test accuracy values, we additionally prune a random percentage of the training data and undergo the same experiment on the three datasets (Table 2 (b)). Here, O-VIG shows better correlation in all cases.

Tabular Data (Resnets): We choose 15 Openml (Vanschoren et al., 2014) datasets (denoted by OpenML-15) and train Resnet-MLPs on each and record the test accuracy in each case. The $\mathcal{V}$-ordered set is generated by scaling the width of the Resnet.

Table 3: Spearman Correlation of Measures with Test Accuracy: Across 15 OpenML Tabular Datasets

Spearman Correlation			
Datasets	**$\mathcal{V}$-I**	**O-VIG**	**AOM**
OpenML-15	0.60	**0.81**	0.76

We compare performance with a set of ten measures from (Komorniczak and Ksieniewicz, 2023), in terms of Spearman correlation with test accuracy. To enable comparisons with a larger set of approaches, the 15 datasets are chosen such that they have lower sample size. Overall, we see that O-VIG performs better than other compared approaches (Table 3).

6 REFLECTIONS

Motivated from $\mathcal{V}$-Information, ordered $\mathcal{V}$-Information growth was proposed. The proposed measure results from two key, unique ideas: subtraction of independent-variable $\mathcal{V}$-information (VI-Growth), and the construction of the ordered set of predictive families from low to high complexity. Both these steps have been motivated appropriately in writing and theory, as we find that VI-Growth serves as an unbiased lower-bound for $\mathcal{V}$-Information (Theorem 1 and Corollary 1.3), whereas the maximum over the ordered set is justified from the fact that VI-Growth is a lower bound of the true $\mathcal{V}$-Information (Corollary 1.1). Empirical results successfully demonstrate that O-VIG can circumvent the over-estimation of empirical $\mathcal{V}$-Information to a good degree. Furthermore, O-VIG can be used both as an MI-Estimator, and also as a measure of dataset difficulty, and in most cases outperforms other state-of-the-art measures.

However, there are a few current limitations. VI-Growth is a measure that, like $\mathcal{V}$-information, is biased positively with the number of classes in the problem. This indicates that if the number of classes across different datasets vary significantly, then it becomes important to consider normalization approaches which cancel the class variation. In our tabular experiments, the number of classes in the problem did vary, but over a relatively smaller scale (between 2 and 10 classes). Another limitation of our proposed approach is the added time complexity involved in the estimation of O-VIG ($\approx 2|\boldsymbol{\mathcal{V}}|$ times more than $\mathcal{V}$-information, where $|\boldsymbol{\mathcal{V}}|$ is the size of the $\mathcal{V}$-ordered set). However, we believe that early-stopping can be used to make estimation significantly faster. These approaches could be explored in future work. Given the other benefits from a performance perspective, we believe that O-VIG represents an exciting new direction for shared information measures.

Acknowledgements

This research is supported by A*STAR, CISCO Systems (USA) Pte. Ltd and the National University of Singapore under its Cisco-NUS Accelerated Digital Economy Corporate Laboratory (Award I21001E0002). We also thank the Kent-Ridge AI research group at the National University of Singapore for helpful discussions.

References

M. I. Belghazi et al. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR, 2018.

K. H. Brodersen et al. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, pages 3121–3124. IEEE, 2010.

M. Chalk, O. Marre, and G. Tkacik. Relevant sparse codes with variational information bottleneck. *Advances in Neural Information Processing Systems*, 29, 2016.

K. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

O. David et al. Evaluation of different measures of functional connectivity using a neural mass model. *Neuroimage*, 21(2):659–673, 2004.

D. Endres and P. Foldiak. Bayesian bin distribution inference and mutual information. *IEEE Transactions on Information Theory*, 51(11):3766–3779, 2005.

W. Gao et al. Estimating mutual information for discrete-continuous mixtures. *Advances in neural information processing systems*, 30, 2017.

M. J. Graham, S. Djorgovski, A. A. Mahabal, C. Donalek, and A. J. Drake. Machine-assisted discovery of relationships in astronomy. *Monthly Notices of the Royal Astronomical Society*, 431(3): 2371–2384, 2013.

R. D. Hjelm et al. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019.

A. Kadra, M. Lindauer, F. Hutter, and J. Grabocka. Well-tuned simple nets excel on tabular datasets. *Advances in neural information processing systems*, 34:23928–23941, 2021.

M. Kleinman et al. Gacs-korner common information variational autoencoder. *Advances in Neural Information Processing Systems*, 36:66020–66043, 2023.

J. Komorniczak and P. Ksieniewicz. problexity—an open-source python library for supervised learning problem complexity assessment. *Neurocomputing*, 521:126–136, 2023.

A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69(6):066138, 2004.

A. Krizhevsky, V. Nair, and G. Hinton. Cifar-10 (canadian institute for advanced research). a. URL `http://www.cs.toronto.edu/ kriz/cifar.html`.

A. Krizhevsky, V. Nair, and G. Hinton. Cifar-100 (canadian institute for advanced research). b. URL `http://www.cs.toronto.edu/ kriz/cifar.html`.

Y. LeCun et al. MNIST handwritten digit database. `http://yann.lecun.com/exdb/mnist/`, 2010.

Y.-I. Moon, B. Rajagopalan, and U. Lall. Estimation of mutual information using kernel density estimators. *Physical Review E*, 52(3):2318, 1995.

X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

B. C. Ross. Mutual information between discrete and continuous data sets. *PloS one*, 9(2):e87357, 2014.

A. M. Saxe et al. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124020, 2019.

R. Shwartz-Ziv and N. Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

J. Song and S. Ermon. Understanding the limitations of variational mutual information estimators. In *International Conference on Learning Representations*, 2020.

M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic. On mutual information maximization for representation learning. In *International Conference on Learning Representations*, 2020.

J. Vanschoren et al. Openml: networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014.

Y. Xu et al. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020.

S. Yu et al. Measuring dependence with matrix-based entropy functional. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10781–10789, 2021.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

A Summary of Supplementary Materials

In the appendices, we first outline the details of the experiments conducted in the main paper. Next, we provide the proofs of all theoretical results in the main paper. Lastly, we include additional experiments that explore different aspects of VI-Growth, such as its response to class additions and network depth changes.

B Empirical Details

General Details: We only test neural network $\mathcal{V}$ in our experiments. For every problem, we define a base network $\mathcal{V}$. To generate the estimate of O-VIG, $\widehat{IG}_{\{\boldsymbol{\mathcal{V}}\}}(X \rightarrow Y; S, S')$, given the dataset $S = \{(X_1, Y_1), ..., (X_m, Y_m)\}$, we generate S' by fixing $X_i' = X_i$ and permuting the labels such that $Y_i' = Y_{\sigma(i)}$ where $\sigma : \{1, ..., m\} \rightarrow \{1, ..., m\}$ represents a random permutation function. Note that via the property **P5** of the main paper, the VIG then can be estimated as $\widehat{IG}_{\mathcal{V}}(X \rightarrow Y; S, S') = \widehat{H}_{\mathcal{V}}(Y'|X; S') - \widehat{H}_{\mathcal{V}}(Y|X; S)$. Subsequently, we estimate O-VIG as $\widehat{IG}_{\{\boldsymbol{\mathcal{V}}\}}(X \rightarrow Y; S, S') = \max_{\mathcal{V}_i \in \boldsymbol{\mathcal{V}}} \widehat{IG}_{\mathcal{V}_i}(X \rightarrow Y; S, S')$, where $\boldsymbol{\mathcal{V}} = \{\mathcal{V}_1 \subset \mathcal{V}_2 \subset ... \subset \mathcal{V}\}$ denotes the $\mathcal{V}$-ordered set. Also note that the empirical conditional $\mathcal{V}$-entropy measure is defined as:

$$\widehat{H}_{\mathcal{V}}(Y|X; S) = \inf_{f \in \mathcal{V}} \mathbb{E}_{(X_i, Y_i) \in S}[-\log f[X_i](Y_i)] \tag{8}$$

We note that the above loss essentially becomes the cross-entropy loss when f is a neural network, and $f[X_i](Y_i)$ denotes the probability of the class Y_i given the input X_i, which is just the softmax output of the network. Therefore, the terms $\widehat{H}_{\mathcal{V}_i}(Y'|X; S')$ and $\widehat{H}_{\mathcal{V}_i}(Y|X; S)$ in the estimation of empirical VIG are computed by optimizing a cross-entropy loss function on the supervised prediction problems $X \rightarrow Y$ and $X \rightarrow Y'$, the hyperparameters for which we subsequently will outline for each experiment. For every experiment, given the base $\mathcal{V}$, we generate the $\mathcal{V}$-ordered set by scaling each hidden layer of $\mathcal{V}$ using a scaling factor α.

Let the network $\mathcal{V}$ be represented as $d - H_1 - H_2 - ... - H_l - C$, where d is the number of input channel dimensions and C is the number of output nodes (the number of classes), and $H_1, H_2, ., H_l$ denote the number of hidden neurons corresponding to each layer. Then, each network in the $\mathcal{V}$-ordered will have an architecture of the form $d - \alpha H_1 - \alpha H_2 - ... - \alpha H_l - C$, where α denotes a scaling factor. We also denote α as the $\mathcal{V}$-complexity factor. As $\{\mathcal{V}_1 \subset \mathcal{V}_2 \subset ... \subset \mathcal{V}_k\}$ should be in increasing order of complexity, for each experiment we pre-define a list of factors $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, ..., \alpha_k\}$, such that $\alpha_1 \leq \alpha_2 \leq ... \leq \alpha_k$. Note that α_k is always set to 1, unless otherwise specified. We also denote these as the $\mathcal{V}$-ordered factors. Note that any estimates of O-VIG that result from this $\mathcal{V}$-ordered set will essentially be an estimate of $\mathcal{V}$-information for most complex predictive family in the set, which is $\mathcal{V}_k$. This is a hyperparameter in our experiments. In what follows, we always refer to d as the number of input channel dimensions. For FC-NNs, this is simply the dimensionality of the input, and for CNNs, this is the number of input channels, which depends on whether the input is grayscale ($d = 1$) or RGB ($d = 3$).

Lastly, note that all our networks have batch normalization layers after each hidden layer, to facilitate faster training and convergence. We use the ReLU activation function in all our networks. Also, we found that the Adam optimizer yielded slightly more stable results for both $\mathcal{V}$-I and O-VIG later on, so part of our results are with SGD as the optimizer and the others are with Adam. We conduct all our neural network training using Pytorch.

Synthetic Dataset Details: For generating the means μ_0 and μ_1, we sample both $\mu_0 \sim \mathcal{N}(0, \tau I_d)$ and $\mu_1 \sim \mathcal{N}(0, \tau I_d)$, where I_d is an identity matrix of size d, and τ represents a fixed scaling factor. For generating the random covariance matrices, first sample a random matrix $A \sim \mathbb{R}^{d \times d}$, where each entry of A, $A(i, j) \sim U(0, 1)$, where $U(a, b)$ denotes the uniform distribution sampled over the real line interval [a,b). Next, we set $\Sigma_0 = AA^T$. For some of our experiments, we just set $\Sigma_1 = \Sigma_0$ (D2,D3,D4), whereas for others we repeat the random sampling process for Σ_1 by randomly sampling another A (D1). With this, we outline the specific details pertaining to the four configurations mentioned in the paper: D1: $d = 3, \tau = 0.002$ D2: $d = 3, \tau = 0.03$ D3: $d = 10, \tau = 0.005$ D4: $d = 100$, and out of the 100 generated configurations of μ_0, μ_1, Σ_0 and Σ_1, the first fifty are chosen such that $\tau = 0$ (same mean, zero $I(X; Y)$) and the next fifty are chosen such that $\tau = 0.1$ (which is large enough to always yield maximum $I(X; Y)$).

VI-Growth Trends: The base network has a structure of d-60-30-C, and the $\mathcal{V}$-ordered factors are $\boldsymbol{\alpha} = \{0.1, 0.2, .., 2.0\}$. For estimating VIG and $\mathcal{V}$-I, we used a batch size of 400, SGD as the optimizer with an initial learning rate of 0.05 and a momentum of 0.9, and 100 epochs.

Table 4: Full Results for Spearman Correlation of Measures with Test Accuracy: Label-Subsets

Dataset	# Classes	c1	c2	f2	f3	n4	t2	t3	t4	$\mathcal{V}$-I	O-VIG
MNIST	2	-0.29	-0.29	nan	0.02	0.48	0.23	0.25	0.14	-0.11	0.48
	4	-0.31	-0.32	nan	0.12	0.52	0.05	0.22	0.18	-0.14	0.01
	6	-0.11	-0.12	nan	0.11	0.24	0.21	0.24	0.19	-0.28	0.07
CIFAR-10	2	nan	nan	0.06	0.39	0.42	nan	0.05	-0.05	0.90	0.92
	4	nan	nan	0.21	0.23	0.26	nan	0.14	-0.14	0.76	0.94
	6	nan	nan	0.24	0.34	0.29	nan	0.32	-0.32	0.62	0.96
CIFAR-100	2	nan	nan	0.54	0.76	0.27	nan	0.23	0.23	0.37	0.80
	4	nan	nan	0.12	0.27	0.39	nan	0.03	-0.03	0.77	0.88
	6	nan	nan	0.18	0.21	0.11	nan	0.05	0.05	0.73	0.85
Combined	2	-0.81	-0.81	0.66	0.50	0.75	0.76	0.75	-0.70	0.41	0.94
	4	-0.81	-0.81	0.64	0.44	0.77	0.76	0.72	-0.74	0.37	0.84
	6	-0.80	-0.80	0.72	0.53	0.73	0.69	0.65	-0.68	0.58	0.89

Convergence Analysis (Independent RVs): For the FC-NN results, the base network has a structure of d-120-60-C, and the $\mathcal{V}$-ordered factors are $\boldsymbol{\alpha} = \{0.1, 0.2, .., 1.0\}$. For estimating VIG and $\mathcal{V}$-I, we used a batch size of 400, SGD as the optimizer with an initial learning rate of 0.05 and a momentum of 0.9, and 100 epochs. For the CNN results, the base network has a structure of d-60conv5-3MP-120conv3-GMP-C, where convp denotes convolutional layers of kernel size $p \times p$, kMP denotes max-pooling layers of size $k \times k$ and GMP denotes global max pooling layers. The rest of the parameters are the same as the FC-NN case.

Convergence Analysis (Dependent RVs): We pick distributions such as the ground truth MI is close to $\{0, 0.1, 0.2, ..., 0.6, 0.69\}$. This is done by first generating 100 configurations from **D1** and then picking the one with the true MI closest to each MI value. For the $\mathcal{V}$-I and O-VIG computations, we use the base network as an FC-NN with the architecture d-20-20-C and the $\mathcal{V}$-ordered factors for O-VIG computation are $\boldsymbol{\alpha} = \{0.1, 0.2, .., 1.0\}$. We use the SGD optimizer with an initial learning rate of 0.03 with a momentum of 0.9, 200 total epochs and a batch size of 400.

MI Estimation: For the $\mathcal{V}$-I and O-VIG computations, we use the base network as an FC-NN with the architecture d-20-20-C and the $\mathcal{V}$-ordered factors for O-VIG computation are $\boldsymbol{\alpha} = \{0.1, 0.2, .., 1.0\}$. We use the SGD optimizer with an initial learning rate of 0.03 with a momentum of 0.9, 200 total epochs and a batch size of 400. For every other compared approach in Table 1 of the main paper, we greedily optimize their parameters on one random dataset configuration separately generated for each dataset type, and then test them on the **D1-D4** dataset collection. We have compiled all approaches in Table 1 into a library which will be a part of our released code. The JS, Info-NCE, NWJ and SMILE estimators were adapted from the code in https://github.com/ermongroup/smile-mi-estimator. The KSG estimator is from (Gao et al., 2017)(https://github.com/wgao9/knnie/tree/master), and the KSG-Mixed estimator is from the NPEET toolbox (https://github.com/gregversteeg/NPEET). Lastly the MINE estimator is adapted from (https://github.com/gtegner/mine-pytorch).

Label Noise Addition: For the label noise experiments, we found that it was useful to reduce the complexity of the networks further to get more clearly defined and stable patterns across our working range of sample size. For the FC-NNs, the base network has a structure of d-15-8-C, and the $\mathcal{V}$-ordered factors are $\boldsymbol{\alpha} = \{0.1, 0.2, .., 1.0\}$. For estimating VIG and $\mathcal{V}$-I, we used a batch size of 400, SGD as the optimizer with an initial learning rate of 0.05 and a momentum of 0.9, and 100 epochs. For CNNs, the base network has a structure of d-15conv5-3MP-30conv3-GMP-C, and the rest of the hyperparameters are the same.

Gaussian Input Noise Addition: For the FC-NNs, the base network has a structure of d-120-60-C, and the $\mathcal{V}$-ordered factors are $\boldsymbol{\alpha} = \{0.1, 0.2, .., 1.0\}$. For estimating VIG and $\mathcal{V}$-I, we used a batch size of 400, SGD as the optimizer with an initial learning rate of 0.05 and a momentum of 0.9, and 100 epochs. For CNNs, the base network has a structure of d-60conv5-3MP-120conv3-GMP-C, and the rest of the hyperparameters are the same.

Dataset Complexity (Vision): The base network has a structure of d-60conv5-3MP-120conv3-GMP-C, and the $\mathcal{V}$-ordered factors are $\boldsymbol{\alpha} = \{0.1, 0.2, .., 1.0\}$. For estimating VIG and $\mathcal{V}$-I, we used a batch size of 400, Adam as the optimizer with an initial learning rate of 0.001, and 100 epochs. We also provide the full results for Tables 2 and 3 in the main paper in Table 4 and 5 respectively, which includes the Spearman correlation results for all

Table 5: Full Results for Spearman Correlation of Measures with Test Accuracy: Label-Subsets + Data-Size Variation

Dataset	# Classes	c1	c2	f2	f3	n4	t2	t3	t4	$\mathcal{V}$-I	O-VIG
MNIST	2	0.11	0.11	nan	0.14	0.51	0.25	0.23	-0.02	0.13	0.59
	4	-0.12	-0.15	nan	0.13	0.40	0.63	0.65	-0.04	0.05	0.71
	6	0.27	0.22	nan	0.52	0.59	0.77	0.78	-0.40	0.47	0.85
CIFAR-10	2	0.24	0.24	0.12	0.19	0.44	0.29	0.30	-0.17	0.58	0.86
	4	0.63	0.63	0.76	0.74	0.01	0.77	0.73	-0.65	0.65	0.92
	6	0.48	0.48	0.41	0.30	0.43	0.44	0.45	-0.50	0.77	0.97
CIFAR-100	2	0.05	0.05	0.10	0.16	0.52	0.40	0.46	-0.03	0.22	0.74
	4	0.38	0.38	0.22	0.11	0.36	0.58	0.48	-0.49	0.57	0.83
	6	0.33	0.34	0.43	0.37	0.01	0.61	0.63	-0.42	0.59	0.68
Combined	2	0.28	0.28	0.59	0.28	0.82	0.72	0.60	-0.77	0.43	0.90
	4	0.19	0.19	0.62	0.35	0.75	0.76	0.68	-0.76	0.55	0.88
	6	-0.14	-0.15	0.59	0.19	0.71	0.76	0.69	-0.77	0.65	0.84

Table 6: Full Spearman Correlation Results of Measures with Test Accuracy: OpenML-15

Datasets	c1	c2	f2	f3	l2	l3	n4	t2	t3	t4	**$\mathcal{V}$-I**	O-VIG
OpenML-15	0.41	0.65	0.58	0.77	0.67	0.72	0.71	-0.08	0.06	0.07	0.60	0.81

other measures compared from the problexity package. Note that some of the class imbalance measures return undefined values (nan) as in those cases there is no class imbalance across the datasets which yields undefined values of Spearman correlation as the measure remains fixed.

Dataset Complexity (Tabular): The 15 datasets in OpenML-15 can be described by the following task IDs in OpenML: 233088, 233090, 233091, 233093, 233088, 233094, 233096, 233103, 233108, 233109, 233115, 233116, 233117, 233134, 233135. Note that these task IDs are a part of the tests conducted in (Kadra et al., 2021) (Table 9 of that paper). For this experiment, we choose the base network as a resnet with 2 groups with 1 block per group and 512 units in each layer, similar in scale to the MLP trained in (Kadra et al., 2021). As even narrow resnets are very good at fitting the data, and as the OpenML-15 datasets are relatively low dimensional and of lower sample size compared to the vision datasets, we decide on a slightly different $\mathcal{V}$-ordered factors $\boldsymbol{\alpha}$. Specifically, we set $\boldsymbol{\alpha} = \{0.01, 0.02, .., 0.09, 1\}$, such that the majority of the $\mathcal{V}$-complexity factors are concentrated near to 0. For estimating VIG and $\mathcal{V}$-I, we set the batch size to 200, Adam as the optimizer with an initial learning rate of 0.001 and 200 epochs. Guided by (Kadra et al., 2021), to counter the heavy class imbalance in many of the datasets, we use a weighted cross-entropy measure to obtain the VIG and $\mathcal{V}$-I values, and estimate the class balanced accuracy on the test data (Brodersen et al., 2010). As the tabular datasets are of lower input dimensionality and have less sample size than the vision datasets, we can compare more dataset complexity measures from the problexity package, and the full Spearman correlation results for Table 4 in the main paper are provided in Table 6.

C Proofs of VI-Growth Properties

In this section we provide the proofs of the properties of VI-Growth mentioned in the main paper (section 5.1).

P1 (Independence) If X is independent of Y, then $IG^m_{\mathcal{V}}(X \to Y) = IG^m_{\mathcal{V}}(Y \to X) = 0$

Proof. When $X \perp Y$, we have that $P_{XY} = P_X P_Y$. Thus, X, Y and X', Y' have the same distribution. Thus, we can write:

$$\begin{aligned} IG^m_{\mathcal{V}}(X \to Y) &= \mathbb{E}_{S \sim P_{XY}, S' \sim P_X P_Y} \left[\widehat{I}_{\mathcal{V}}(X \to Y; S) - \widehat{I}_{\mathcal{V}}(X' \to Y'; S') \right] \\ &= \mathbb{E}_{S \sim P_X P_Y} \left[\widehat{I}_{\mathcal{V}}(X \to Y; S) \right] - \mathbb{E}_{S' \sim P_X P_Y} \left[\widehat{I}_{\mathcal{V}}(X' \to Y'; S') \right] = 0 \end{aligned} \tag{9}$$

The proof follows similarly for $IG^m_{\mathcal{V}}(Y \to X)$. □

P2 (Limit) $\lim_{|S|,|S'|\to\infty} \widehat{IG}_{\mathcal{V}}(X \to Y; S, S') = \lim_{|S|,|S'|\to\infty} \widehat{IG}_{\{\boldsymbol{\mathcal{V}}\}}(X \to Y; S, S') = I_{\mathcal{V}}(X \to Y)$

Proof. We have

$$\lim_{|S|,|S'|\to\infty} \widehat{IG}_{\mathcal{V}}(Y \to X; S, S') = \lim_{|S|\to\infty} \widehat{I}_{\mathcal{V}}(X \to Y; S) - \lim_{|S'|\to\infty} \widehat{I}_{\mathcal{V}}(X' \to Y'; S') \tag{10}$$

$$= I_{\mathcal{V}}(X \to Y) - I_{\mathcal{V}}(X' \to Y') = I_{\mathcal{V}}(X \to Y) \tag{11}$$

□

The last line follows from the fact that $I_{\mathcal{V}}(X' \to Y') = 0$ as $X' \perp Y'$ by definition. Next, we have that

$$\lim_{|S|,|S'|\to\infty} \widehat{IG}_{\{\boldsymbol{\mathcal{V}}\}}(X \to Y; S, S') = \lim_{|S|,|S'|\to\infty} \max_{\mathcal{V}_i \in \boldsymbol{\mathcal{V}}} \widehat{IG}_{\mathcal{V}_i}(X \to Y; S, S') \tag{12}$$

$$= \max_{\mathcal{V}_i \in \boldsymbol{\mathcal{V}}} I_{\mathcal{V}_i}(X \to Y) \tag{13}$$

As by its construction, we have that $\mathcal{V}_1 \subset \mathcal{V}_2 \subset ... \subset \mathcal{V}$. we have that $H_{\mathcal{V}}(Y|X) \leq H_{\mathcal{V}_{k-1}}(Y|X) \leq ... \leq H_{\mathcal{V}_1}(Y|X)$. As all $\mathcal{V}_1, \mathcal{V}_2, ..., \mathcal{V}$ are null-complete, we also have that $H_{\mathcal{V}_1}(Y) = H_{\mathcal{V}_2}(Y) = ... = H_{\mathcal{V}}(Y) = H(Y)$. This ultimately yields $I_{\mathcal{V}}(Y|X) \geq I_{\mathcal{V}_{k-1}}(Y|X) \geq ... \geq I_{\mathcal{V}_1}(Y|X)$, and thus $\max_{\mathcal{V}_i \in \boldsymbol{\mathcal{V}}} I_{\mathcal{V}_i}(X \to Y) = I_{\mathcal{V}}(Y|X)$, completing the proof.

P3 (Non-negativity) $IG^m_{\mathcal{V}}(X \to Y) \geq 0$ and $IG^m_{\{\boldsymbol{\mathcal{V}}\}}(X \to Y) \geq 0$.

Proof. First, note that:

$$\begin{aligned} IG^m_{\mathcal{V}}(X \to Y) &= \mathop{\mathbb{E}}_{S}\left[\widehat{I}_{\mathcal{V}}(X \to Y; S)\right] - \mathop{\mathbb{E}}_{S'}\left[\widehat{I}_{\mathcal{V}}(X' \to Y'; S')\right] \\ &= \mathop{\mathbb{E}}_{S}\left[\widehat{H}_{\mathcal{V}}(Y; S) - \widehat{H}_{\mathcal{V}}(Y|X; S)\right] - \mathop{\mathbb{E}}_{S'}\left[\widehat{H}_{\mathcal{V}}(Y'; S') - \widehat{H}_{\mathcal{V}}(Y'|X'; S')\right] \\ &= \mathop{\mathbb{E}}_{S'}\left[\widehat{H}_{\mathcal{V}}(Y'|X'; S')\right] - \mathop{\mathbb{E}}_{S}\left[\widehat{H}_{\mathcal{V}}(Y|X; S)\right], \end{aligned} \tag{14}$$

where the last step follows from the fact that $\mathbb{E}_S\left[\widehat{H}_{\mathcal{V}}(Y; S)\right] = \mathbb{E}_{S'}\left[\widehat{H}_{\mathcal{V}}(Y'; S')\right]$ as Y and Y' have the same distribution. Next, we use the P-Progressive constraint to prove the above as follows. Note that the P-Progressive constraint states that $\mathbb{E}_{S'_m}\left[\widehat{H}_{\mathcal{V}}(Y|X; S'_m) - \widehat{H}_{\mathcal{V}}(Y|X; S'_{m-1})\right] \geq \mathbb{E}_{S_m}\left[\widehat{H}_{\mathcal{V}}(Y|X; S_m) - \widehat{H}_{\mathcal{V}}(Y|X; S_{m-1})\right]$. We can write:

$$\begin{aligned} \sum_{i=2}^{m}\left[\mathbb{E}_{S'_i}\left[\widehat{H}_{\mathcal{V}}(Y|X; S'_i) - \widehat{H}_{\mathcal{V}}(Y|X; S'_{i-1})\right]\right] &\geq \sum_{i=2}^{m}\left[\mathbb{E}_{S_i}\left[\widehat{H}_{\mathcal{V}}(Y|X; S_i) - \widehat{H}_{\mathcal{V}}(Y|X; S_{i-1})\right]\right] \\ \mathbb{E}_{S'_m}\left[\widehat{H}_{\mathcal{V}}(Y|X; S'_m)\right] - \mathbb{E}_{S'_1}\left[\widehat{H}_{\mathcal{V}}(Y|X; S'_1)\right] &\geq \mathbb{E}_{S_m}\left[\widehat{H}_{\mathcal{V}}(Y|X; S_m)\right] - \mathbb{E}_{S_1}\left[\widehat{H}_{\mathcal{V}}(Y|X; S_1)\right] \\ \mathbb{E}_{S'_m}\left[\widehat{H}_{\mathcal{V}}(Y|X; S'_m)\right] - \mathbb{E}_{S_m}\left[\widehat{H}_{\mathcal{V}}(Y|X; S_m)\right] &\geq 0, \end{aligned} \tag{15}$$

where the last step follows from the fact that when S has a single sample, from a loss perspective, $(x, 1), .., (x, c)$ are all equivalent where $x \in \mathcal{X}$ and $y \in \{1, 2, .., c\}$. This ensures that from a single sample perspective, $(X, Y) \sim P_X P_Y$ and $(X, Y) \sim P_{XY}$ are equivalent. Thus, $\mathbb{E}_{S'_1}\left[\widehat{H}_{\mathcal{V}}(Y|X; S'_1)\right] = \mathbb{E}_{S_1}\left[\widehat{H}_{\mathcal{V}}(Y|X; S_1)\right]$. Lastly, coupled with (14), we have that $IG^m_{\mathcal{V}}(X \to Y) \geq 0$. As $IG^m_{\{\boldsymbol{\mathcal{V}}\}}(X \to Y) = \max_{\mathcal{V}_i \in \boldsymbol{\mathcal{V}}} IG^m_{\mathcal{V}_i}(X \to Y)$, it follows that $IG^m_{\{\boldsymbol{\mathcal{V}}\}}(X \to Y) \geq 0$.

□

P4 (Upper Bound) $IG^m_{\mathcal{V}}(X \to Y) \leq I_{\mathcal{V}}(X \to Y)$, $\widehat{IG}_{\mathcal{V}}(X \to Y; S, S') \leq \widehat{H}(Y; S)$, $IG^m_{\{\boldsymbol{\mathcal{V}}\}}(X \to Y) \leq I_{\mathcal{V}}(X \to Y)$ and $\widehat{IG}_{\{\boldsymbol{\mathcal{V}}\}}(Y \to X; S, S') \leq \widehat{H}(Y; S)$

Proof. Following the P-Progressive constraint, we can write:

$$\begin{aligned} \mathbb{E}_{S'_m}\left[\widehat{H}_{\mathcal{V}}(Y|X; S'_m)\right] - \mathbb{E}_{S_m}\left[\widehat{H}_{\mathcal{V}}(Y|X; S_m)\right] &\leq \lim_{m\to\infty}\left(\mathbb{E}_{S'_m}\left[\widehat{H}_{\mathcal{V}}(Y|X; S'_m)\right] - \mathbb{E}_{S_m}\left[\widehat{H}_{\mathcal{V}}(Y|X; S_m)\right]\right) \\ &= H_{\mathcal{V}}(Y'|X') - H_{\mathcal{V}}(Y|X) = H_{\mathcal{V}}(Y') - H_{\mathcal{V}}(Y|X) \\ &= H_{\mathcal{V}}(Y) - H_{\mathcal{V}}(Y|X) = I_{\mathcal{V}}(X \to Y), \end{aligned} \tag{16}$$

where $(X', Y') \sim P_X P_Y$ and $H_{\mathcal{V}}(Y'|X') = H_{\mathcal{V}}(Y')$ follows from the fact that $I_{\mathcal{V}}(X' \to Y') = 0$ as $X' \perp Y'$. As we have shown in the proof of P3, we have that $IG^m_{\mathcal{V}}(X \to Y) = \mathbb{E}_{S'_m}\left[\widehat{H}_{\mathcal{V}}(Y|X; S'_m)\right] - \mathbb{E}_{S_m}\left[\widehat{H}_{\mathcal{V}}(Y|X; S_m)\right]$, which yields the first result, $IG^m_{\mathcal{V}}(X \to Y) \leq I_{\mathcal{V}}(X \to Y)$.
Next, using P3's proof, we can write, $IG^m_{\{\boldsymbol{\mathcal{V}}\}}(X \to Y) = \max_{\mathcal{V}_i \in \boldsymbol{\mathcal{V}}} IG^m_{\mathcal{V}_i}(X \to Y) \leq \max_{\mathcal{V}_i \in \boldsymbol{\mathcal{V}}} I_{\mathcal{V}_i}(X \to Y) = I_{\mathcal{V}}(Y|X)$, yielding the upper bound for O-VIG.
For the empirical measures, we derive the upper bounds as follows. First, note that $\widehat{IG}_{\mathcal{V}}(X \to Y; S, S') = \widehat{I}_{\mathcal{V}}(X \to Y; S) - \widehat{I}_{\mathcal{V}}(X' \to Y'; S')$. We have $\widehat{I}_{\mathcal{V}}(X \to Y; S) = \widehat{H}_{\mathcal{V}}(Y; S) - \widehat{H}_{\mathcal{V}}(Y|X; S) \leq \widehat{H}_{\mathcal{V}}(Y; S)$. We also have that $\widehat{I}_{\mathcal{V}}(X' \to Y'; S') = \widehat{H}_{\mathcal{V}}(Y'; S') - \widehat{H}_{\mathcal{V}}(Y'|X'; S') \geq 0$. This yields $\widehat{IG}_{\mathcal{V}}(X \to Y; S, S') \leq \widehat{H}_{\mathcal{V}}(Y; S) = \widehat{H}(Y; S)$ as $\mathcal{V}$ is null-complete. For empirical O-VIG, we have $\widehat{IG}_{\{\boldsymbol{\mathcal{V}}\}}(X \to Y; S, S') = \max_{\mathcal{V}_i \in \boldsymbol{\mathcal{V}}} \widehat{IG}_{\mathcal{V}_i}(X \to Y; S, S') \leq \max_{\mathcal{V}_i \in \boldsymbol{\mathcal{V}}} \widehat{H}_{\mathcal{V}_i}(Y; S) = \widehat{H}(Y; S)$ as all predictive families are null complete. □

P5 (Relation to Conditional $\mathcal{V}$-entropy) If S' is chosen such that $X'_i = X_i$ and $Y'_i = Y_{\sigma(i)}$, where $\sigma : \{1, ..., m\} \to \{1, ..., m\}$ represents a random permutation function, then $\widehat{IG}_{\mathcal{V}}(X \to Y; S, S') = \widehat{H}_{\mathcal{V}}(Y'|X; S') - \widehat{H}_{\mathcal{V}}(Y|X; S)$

Proof. First, note that we can expand empirical VIG as $\widehat{IG}_{\mathcal{V}}(X \to Y; S, S') = \left(\widehat{H}_{\mathcal{V}}(Y; S) - \widehat{H}_{\mathcal{V}}(Y'; S')\right) + \left(\widehat{H}_{\mathcal{V}}(Y|X; S) - \widehat{H}_{\mathcal{V}}(Y'|X'; S')\right)$, where X', Y', S' are as in Definition 5 of the main paper. From the constraints in P5, we have that $X' = X$. Next, as $Y'_i = Y_{\sigma(i)}$ is a random permutation of the original labels, the empirical distributions of Y and Y' are exactly the same. That is, $P(Y = i|S) = P(Y' = i|S'), \forall i \in \{1, 2, .., c\}$. Thus, it follows that $\widehat{H}_{\mathcal{V}}(Y; S) = \widehat{H}_{\mathcal{V}}(Y'; S')$, and the result follows. □

D Proofs of Theoretical Results

Theorem 1. *Assume that $\forall f \in \mathcal{V}, x \in \mathcal{X}, y \in \mathcal{Y}$ we have $\log f[x](y) \in [-B, B]$. Then with probability $p \geq 1 - \delta$ over the draw of S and S', we have*

$$I_{\mathcal{V}}\left(X \to Y\right) \geq \widehat{IG}_{\mathcal{V}}(X \to Y; S, S') - B\sqrt{\frac{4\log\frac{1}{\delta}}{|S|}} \tag{17}$$

Proof. First, note that we can express aggregate VIG as $IG^m_{\mathcal{V}}(X \to Y) = \mathbb{E}_{SS'}\left[\widehat{I}_{\mathcal{V}}(X \to Y; S) - \widehat{I}_{\mathcal{V}}(X' \to Y'; S')\right] = \mathbb{E}_{S,S'}\left[\phi(S, S')\right]$, where ϕ is a function that takes in S and S' as its arguments. Next, we can use Mcdiarmid's inequality to bound $\phi(S, S')$ using its expected value $\mathbb{E}_{S,S'}[\phi(S, S')]$ as all datapoints in S and S' are independently generated according to their distributions.

To apply Mcdiarmid, we need to construct instances that differ by only one point. Given $D_1 = [S_1, S'_1]$, let us construct another $D_2 = [S_2, S'_2]$ which differs from D_1 by exactly one point. We then have:

$$\phi(D_1) - \phi(D_2) = [\widehat{I}_{\mathcal{V}}(X \to Y; S_1) - \widehat{I}_{\mathcal{V}}(X' \to Y'; S'_1)] - [\widehat{I}_{\mathcal{V}}(X \to Y; S_2) - \widehat{I}_{\mathcal{V}}(X' \to Y'; S'_2)]. \tag{18}$$

If the differing sample is in $S_1 - S_2$, then $\phi(D_1) - \phi(D_2) = \widehat{I}_{\mathcal{V}}(X \to Y; S_1) - \widehat{I}_{\mathcal{V}}(X' \to Y'; S_2) \leq 2B/m$, where the upper bound follows from the definition of empirical $\mathcal{V}$-I and the bounds on $\log f[x](y)$ where $f \in \mathcal{V}$.

Similarly, if the differing sample is in $S'_1 - S'_2$, then $\phi(D_1) - \phi(D_2) = \widehat{I}_{\mathcal{V}}(X \to Y; S'_1) - \widehat{I}_{\mathcal{V}}(X' \to Y'; S'_2) \leq 2B/m$ as well. Thus, $\phi(D_1) - \phi(D_2) \leq 2B/m$. Thus applying Mcdiarmid, we have with probability $p \geq 1 - \delta$,

$$\phi(S, S') \geq \mathop{\mathbb{E}}_{S,S'}\left[\phi(S, S')\right] - B\sqrt{\frac{4\log\frac{1}{\delta}}{|S|}} \tag{19}$$

As $IG^m_{\mathcal{V}}(X \to Y) = \mathbb{E}_{S,S'}\left[\phi(S, S')\right]$, and from property **P4** we have $IG^m_{\mathcal{V}}(X \to Y) \leq I_{\mathcal{V}}(X \to Y)$, which yields the intended result. □

Corollary 1.1. *We are given a function class $\mathcal{V}$. Using this, we construct a set of function classes $\boldsymbol{\mathcal{V}} = \{\mathcal{V}_1, \mathcal{V}_2, ..., \mathcal{V}\}$, such that $\{\mathcal{V}_1 \subset \mathcal{V}_2 \subset ... \subset \mathcal{V}\}$. Assume that $\forall f \in \mathcal{V}, x \in \mathcal{X}, y \in \mathcal{Y}$ we have $\log f[x](y) \in [-B, B]$.*

Then with probability $p \geq (1-\delta)$ over the draw of S and S', we have

$$I_{\mathcal{V}}(X;Y) \geq \widehat{IG}_{\{\boldsymbol{\mathcal{V}}\}}(Y \to X; S, S') - B\sqrt{\frac{4\log\frac{|\boldsymbol{\mathcal{V}}|}{\delta}}{|S|}} \tag{20}$$

Proof. The result follows almost directly from the Theorem, by applying it to all $\mathcal{V}_i$ in the $\mathcal{V}$-ordered set $\boldsymbol{\mathcal{V}} = \{\mathcal{V}_1, \mathcal{V}_2, ..., \mathcal{V}\}$. To show this, first let us denote the subset of S, S' combinations that yield $I_{\mathcal{V}_i}(X \to Y) \geq \widehat{IG}_{\mathcal{V}_i}(X \to Y; S, S') - B\sqrt{\frac{4\log\frac{1}{\delta}}{|S|}}$ by Γ_i. Note that from the Theorem, we have $P(\Gamma_i) \geq 1-\delta$. Next, we note that for all of the lower bounds to be true, i.e., $I_{\mathcal{V}_i}(X \to Y) \geq \widehat{IG}_{\mathcal{V}_i}(X \to Y; S, S') - B\sqrt{\frac{4\log\frac{1}{\delta}}{|S|}}\ \forall i \in \{1, 2, .., k\}$, the empirical sample S, S' has to be such that $S, S' \in \Gamma_1 \cap \Gamma_2 \cap ... \cap \Gamma_k$. Using Bonferroni's inequality, it follows that $P(\Gamma_1 \cap \Gamma_2 \cap ... \cap \Gamma_k) \geq 1 - \sum_i [1 - P(\Gamma_i)] \geq 1 - k\delta$. Lastly, note that $I_{\mathcal{V}_i}(X \to Y) \geq \widehat{IG}_{\mathcal{V}_i}(X \to Y; S, S') - B\sqrt{\frac{4\log\frac{1}{\delta}}{|S|}}\ \forall i \in \{1, 2, .., k\}$ implies that $I_{\mathcal{V}}(X \to Y) \geq \widehat{IG}_{\{\boldsymbol{\mathcal{V}}\}}(Y \to X; S, S') - B\sqrt{\frac{4\log\frac{1}{\delta}}{|S|}}$. Setting $\delta' = k\delta$, we arrive at the result, noting that $|\boldsymbol{\mathcal{V}}| = k$. □

Corollary 1.2. *We consider the setting of Theorem 1, where we assume that X is independent of Y. Then with probability $p \geq (1-\delta)$ over the draw of S and S', we have $\widehat{IG}_{\mathcal{V}}(X \to Y; S, S') \leq B\sqrt{\frac{4\log\frac{1}{\delta}}{|S|}}$ & $\widehat{IG}_{\{\boldsymbol{\mathcal{V}}\}}(Y \to X; S, S') \leq B\sqrt{\frac{4\log\frac{|\boldsymbol{\mathcal{V}}|}{\delta}}{|S|}}$*

Proof. The result follows from the observation that when $X \perp Y$, $I_{\mathcal{V}}(X;Y) = 0$. Subsequently, applying Theorem 1 and Corollary 1.1 yields the bounds. □

Corollary 1.3. ***(From Theorem 1 of (Xu et al., 2020))*** *In the same setting as Theorem 1, with probability at least $1-2\delta$, when X is independent of Y, we have $|\widehat{I}_{\mathcal{V}}(X \to Y; S)| \leq 4\mathcal{R}_{|S|}(\mathcal{G}_{\mathcal{V}}) + B\sqrt{\frac{8\log\frac{|\boldsymbol{\mathcal{V}}|}{\delta}}{|S|}}$, where $\mathcal{G}_{\mathcal{V}} = \{g | g(x,y) = \log f[x](y), f \in \mathcal{V}\}$, and $\mathcal{R}_m(\mathcal{G})$ denotes the Rademacher complexity of $\mathcal{G}$ with sample number m.*

Proof. The result follows directly from Theorem 1 of (Xu et al., 2020), by noting that when $X \perp Y$, $I_{\mathcal{V}}(X;Y) = 0$. Note that the rest of the definitions are the same as in Theorem 1, and thus we can directly apply the Theorem to obtain the result. □

E Additional Experiments

E.1 Dataset Complexity: Vision

For the dataset complexity experiments on vision datasets in the main paper, we note that only CNN architectures were used. Here, we extend the results to fully connected neural network (FC-NN) models, for the vision experiments in Table 4. We used the same FC-NN architecture that we used for the convergence analysis experiments for independent RVs: $d - 120 - 60 - C$, where d is the dimensionality of the input and C denotes the number of output classes. 120 and 60 are the number of hidden neurons in the first and second hidden layer respectively. 20 trials were conducted and the Spearman correlation between the measures and the test accuracy was estimated. The results are as follows (AOM represents the same set of eight methods tried out from the problexity package as used in Table 4):

Overall, we find that O-VIG is still showcasing good performance and improvements over the other baselines, including $\mathcal{V}$-Information itself.

E.2 VI-Growth Trends

We additionally report the VI-Growth trends on CIFAR-100 and MNIST here, as in the main paper we only discuss the trends on CIFAR-10 due to space constraints. To generate the $\mathcal{V}$-ordered set, we construct a fully connected neural network with two hidden layers, with the structure d-60-30-C, where d is the input dimensionality, and C represents the number of output nodes which is also the number of classes. In the case of MNIST

Table 7: Spearman Correlation of Measures with Test Accuracy: Label-Subsets (FC-NN)

# Classes	Measure	MNIST	CIF-10	CIF-100	Combined
2	$\mathcal{V}$-I	-0.23	0.93	0.56	0.64
	O-VIG	0.40	0.95	0.52	0.89
	AOM	**0.65**	0.26	**0.64**	0.81
4	$\mathcal{V}$-I	-0.15	0.92	0.78	0.85
	O-VIG	**0.65**	**0.92**	**0.84**	**0.91**
	AOM	0.63	0.23	0.21	0.82
6	$\mathcal{V}$-I	0.15	0.87	0.81	0.86
	O-VIG	**0.73**	**0.90**	**0.89**	**0.93**
	AOM	0.62	0.45	0.22	0.81

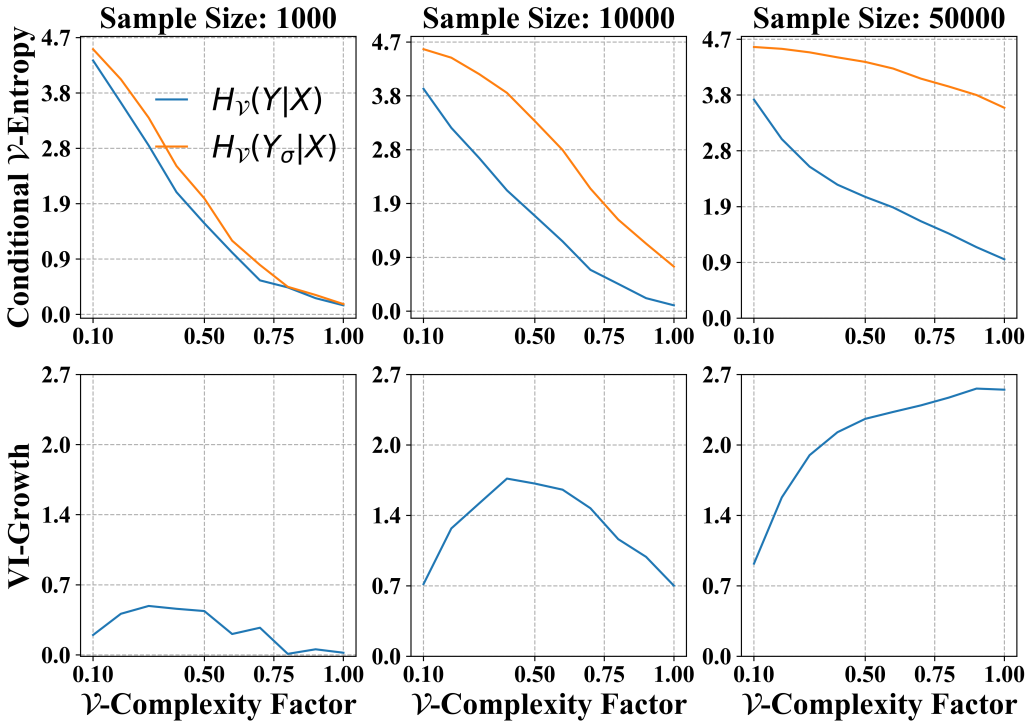

Figure 7: **(VIG-Trends, CIFAR-100)** Figure showing the trends of the various terms involved in the estimation of VIG and O-VIG, for the CIFAR-10 dataset, across three different sample sizes $m = 1000, 10000, 50000$. The top row shows the conditional $\mathcal{V}$-entropy measures, and the bottom row shows the corresponding estimated VI-Growth measures. The x-axis denotes the complexity factor of $\mathcal{V}_i$.

and CIFAR-10, $C = 10$ whereas for CIFAR-100, $C = 100$. To generate the $\mathcal{V}$-ordered set we scale the width of the network in steps, using $\boldsymbol{\alpha} = \{0.1, 0.2, .., 2.0\}$.

Figures 7 and 8 show the results for CIFAR-100 and MNIST respectively. Similar to the CIFAR-10 results in the main paper, we find that empirical VIG usually peaks for a certain complexity and reduces if $\mathcal{V}$ is of lower or higher complexity. The exception to that rule is seen for CIFAR-100 though, when the sample size is large (50000). However, this is consistent with our observations for CIFAR-10 and MNIST, which just show that the maximum VIG simply corresponds to a larger $\mathcal{V}$-complexity factor as the sample size increases. This is because as **P2** shows, in the limit when the sample size goes to infinity, the VIG becomes $\mathcal{V}$-information itself, and due to the nature of the construction of the $\mathcal{V}$-ordered set, the $\mathcal{V}$-information should always be an increasing function of the complexity factor. Thus, eventually when $m \to \infty$, we expect the VI-Growth trends to always be monotonically increasing and peaking at the highest $\mathcal{V}$-complexity factor.

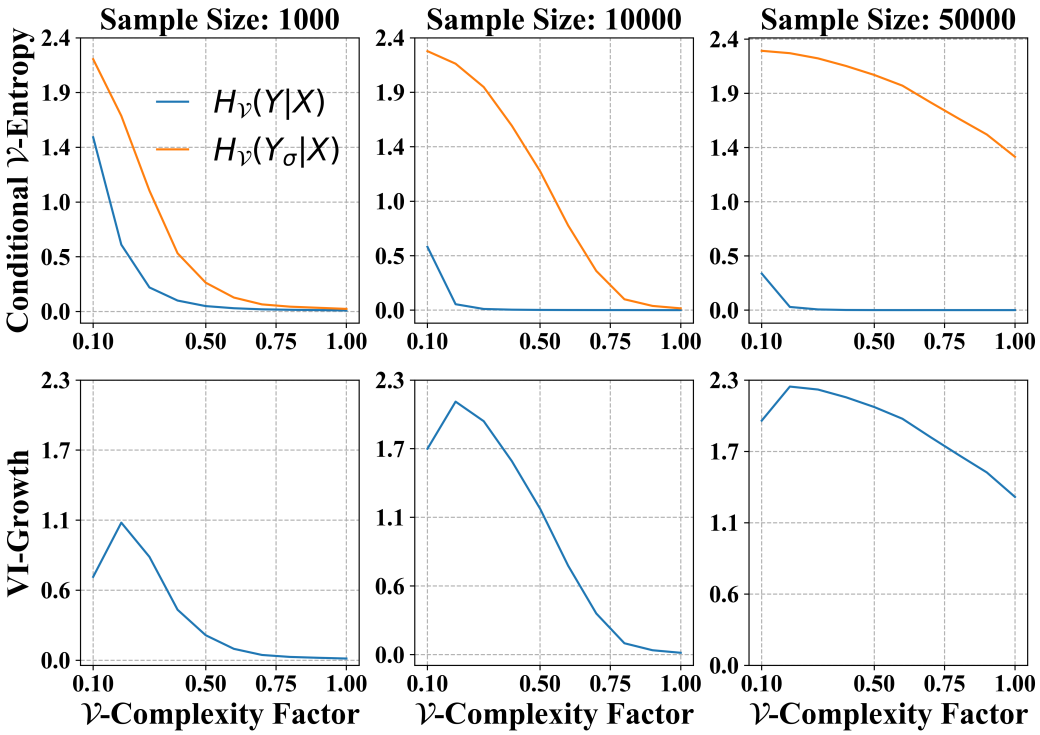

Figure 8: **(VIG-Trends, MNIST)** Figure showing the trends of the various terms involved in the estimation of VIG and O-VIG, for the MNIST dataset, across three different sample sizes $m = 1000, 10000, 50000$.

E.3 Verifying Consequences of Corollary 1.3

In this section, we verify the findings of Corollary 1.3, which finds that unlike VI-Growth and O-VIG, $\mathcal{V}$-Information between two independent variables can potentially track the complexity of the function class $\mathcal{G}_\mathcal{V}$ derived from $\mathcal{V}$ in the Corollary.

First, we note that estimating the Rademacher complexity (RC) of $\mathcal{G}_\mathcal{V}$ can be nontrivial, so we find upper bounds for the same. First, we consider the binary classification scenario with a single output. For this network, the final output has to be after the sigmoid operator to interpret these as probabilities to satisfy the constraints of a predictive family. Let $\mathcal{F}$ denote the function space of the neural network that corresponds to the predictive family $\mathcal{V}$, before the sigmoid output. Let $f_N \in \mathcal{F}$ be the network function. Then, we can relate the every function output from $\boldsymbol{g} \in \mathcal{G}_\mathcal{V}$ and $f_N \in \mathcal{F}$ as follows: $\boldsymbol{g}(\boldsymbol{x}) = \log \frac{e^{f_N(x)}}{1+e^{f_N(x)}}$. We then see that $\boldsymbol{g}(\boldsymbol{x})$ can be expressed as $g'(f_N(x))$ where $g'(\tau) = \log \frac{e^\tau}{1+e^\tau}$, and thus is differentiable. Thus, the Lipschitz constant of g' can then be estimated as $\max_{\tau\in\mathbb{R}} \frac{dg'(\tau)}{d\tau} = \max_{\tau\in\mathbb{R}} \frac{1}{1+e^\tau} = \mathbf{1}$. We can next apply Talagrand's Contraction Lemma [3] which states that $\mathcal{R}_\boldsymbol{m}(\phi \circ \mathcal{F}) \leq L\mathcal{R}_m(\mathcal{F})$, where $\phi \circ \mathcal{F}$ represents a function space derived from $\mathcal{F}$ where the function output from $\mathcal{F}$ passes through another function ϕ. Applying the lemma in this case, we obtain: $\mathcal{R}_\boldsymbol{m}(\mathcal{GV}) \leq \mathcal{R}_\boldsymbol{m}(\mathcal{F})$ as the Lipschitz constant of $\boldsymbol{g'}$ is 1 . Next, we can use the results from [1, 2] which bound $\mathcal{R}_\boldsymbol{m}(\mathcal{F})$ for single hidden layer relu-activated networks with h nodes. These results broadly find that $\mathcal{R}_\boldsymbol{m}(\mathcal{F})$ is upper bounded by a term which is proportional to $\sqrt{h}$.

Next, we setup an experiment where $\mathcal{F}$ is a single hidden layer relu-activated neural network that works with inputs X and output labels Y from a random subset of 5000 samples from the MNIST dataset. As our objective is to test Corollary 1.3 , where X has to be independent of Y, we randomly permute the labels in Y, yielding Y', to enforce this independence. We vary the number of hidden neurons h to create networks with different Rademacher Complexities. Subsequently, the statement of Corollary 1.3 states that the $\mathcal{V}$-Information from X to Y' must be dependent on the RC of $\mathcal{G}_\mathcal{V}$, which in our case is upper bounded by $\mathcal{R}_\boldsymbol{m}(\mathcal{F})$ where $\mathcal{F}$ represents the function space of the neural network before the sigmoid operator Similarly, we also aim to test whether VI-Growth from X to Y has such a dependence or not, as our results state that VI-Growth likely does not depend

Table 8: Tracking the dependency of information measures with the RC bound

	$h = 10$	$h = 20$	$h = 30$	$h = 40$	$h = 50$	$h = 60$	$h = 70$	$h = 80$	Corr with RC
V-I	0.36	0.56	0.71	1.09	1.22	1.45	1.56	1.60	0.99
VIG	0.04	-0.09	-0.04	-0.05	-0.10	0.03	-0.02	-0.06	-0.19
O-VIG	0.06	0.09	0.08	0.08	0.09	0.22	0.22	0.02	0.35

on the RC of the network. Additionally, we also test whether the ordered VI-Growth, which is the maximum VI-Growth across the ordered set constructed from the base $\mathcal{V}$, has a dependence on the RC of the base network. In what follows, $\boldsymbol{h}$ represents the number of hidden neurons that constitutes the predictive network of $\mathcal{V}$. We also report the overall Pearson's correlation of the information measures with the RC bound (i.e., $\sqrt{h}$).

As we can see in Table 8, $\mathcal{V}$-Information shows significantly high correlation with $\sqrt{h}$, which is also directly proportional to the RC bound. As expected, we don't see any significant correlation for VI-Growth, and while the correlation of O-VIG with $\sqrt{h}$ is slightly higher it is still not significant enough. This validates the result in Corollary 1.3 to a certain extent.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes
   (b) Complete proofs of all theoretical results. Yes
   (c) Clear explanations of any assumptions. Yes
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Not Applicable
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Yes
   (b) The license information of the assets, if applicable. Not Applicable
   (c) New assets either in the supplemental material or as a URL, if applicable. Yes
   (d) Information about consent from data providers/curators. Not Applicable
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable