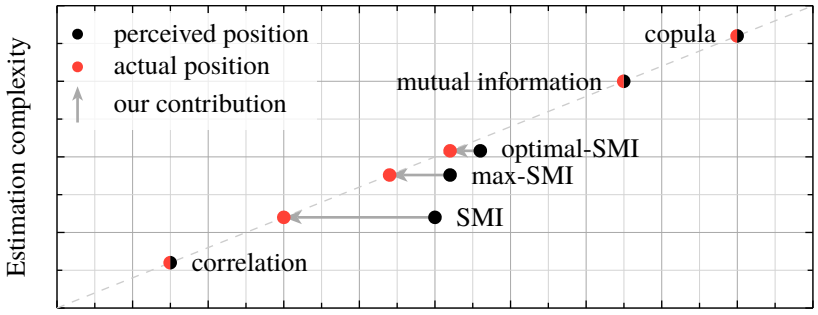
Curse of Slicing: Why Sliced Mutual Information is a Deceptive Measure of Statistical Dependence

Anonymous Author(s)
Affilation, Address
anon.email@example.org

Abstract

Sliced Mutual Information (SMI) is widely used as a scalable alternative to mutual information for measuring non-linear statistical dependence. Despite its advantages, such as faster convergence, robustness to high dimensionality, and nullification only under statistical independence, we demonstrate that SMI is highly susceptible to data manipulation and exhibits counterintuitive behavior. Through extensive benchmarking and theoretical analysis, we show that SMI saturates easily, fails to detect increases in statistical dependence (even under linear transformations designed to enhance the extraction of information), prioritizes redundancy over informative content, and in some cases, performs worse than simpler dependence measures like the correlation coefficient.



1 Introduction

Mutual information (MI) is a fundamental and invariant measure of nonlinear statistical dependence between two random vectors, defined as the Kullback-Leibler divergence between the joint distribution and the product of marginals [1]:

$$\mathsf{I}(X;Y) = \mathrm{D}_{\mathsf{KL}}\big(\mathbb{P}_{X,Y} \,\big\|\, \mathbb{P}_X \otimes \mathbb{P}_Y\big).$$

Due to several outstanding properties, such as nullification only under statistical independence, invariance to invertible transformations, and ability to capture non-linear dependencies, MI is used extensively for theoretical analysis of overfitting [2], [3], hypothesis testing [4], feature selection [5], [6], [7], representation learning [8], [9], [10], [11], [12], [13], and studying the mechanisms behind generalization in deep neural networks (DNNs) [14], [15], [16], [17].

In practical scenarios, $\mathbb{P}_{X,Y}$ and $\mathbb{P}_X \otimes \mathbb{P}_Y$ are unknown, requiring MI to be estimated from finite samples. Despite all the aforementioned merits, this reliance on empirical estimates leads to the curse

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

of dimensionality: the sample complexity of MI grows exponentially with the number of dimensions [18], [19]. A common strategy to mitigate this issue is to use alternative measures of statistical dependence that are more stable in high dimensions. However, such measures usually offer only a fraction of MI capabilities. Therefore, it is crucial to maintain a balance between robustness to the curse of dimensionality and the ability to detect complex dependency structures.

To strike this balance, popular techniques often retain MI as a backbone statistical measure but employ dimensionality reduction before estimation. While some studies explore sophisticated nonlinear compression methods [17], [20], others favor more scalable linear projection approaches [21], [22], [23], [24], [25]. Among the latter group, the *Sliced Mutual Information* (SMI) [22], [23] stands out, leveraging random projections to cover all directions uniformly:

$$\mathsf{SI}(X;Y) = \frac{1}{\oint_{\mathbb{S}^{d_x-1}} \mathrm{d}\theta} \frac{1}{\oint_{\mathbb{S}^{d_y-1}} \mathrm{d}\phi} \oint_{\mathbb{S}^{d_x-1}} \oint_{\mathbb{S}^{d_y-1}} \mathsf{I}(\theta^\mathsf{T} X; \phi^\mathsf{T} Y)\, \mathrm{d}\theta\, \mathrm{d}\phi. \tag{1}$$

Uniform slicing allows SMI to maintain some crucial properties of MI (e.g., being zero if and only if X and Y are independent), while remaining completely free from additional optimization problems (e.g., from finding optimal projections, as in [24], [25]). Combined with fast convergence rates, this has established SMI as a scalable alternative to MI. Consequently, it has been widely adopted for studying DNNs [26], [27], [28], [29], [30], deriving generalization bounds [31], independence testing [32] and auditing differential privacy [33]. It was also proposed to use SMI for feature selection [22] and preventing mode collapse in generative models [23].

Despite its popularity, the research community has largely overlooked potential shortcomings of SMI. Some studies prematurely attribute their results to underlying phenomena without rigorously investigating whether they stem from artifacts introduced by random projections. Furthermore, existing works fail to comprehensively address issues related to random slicing, focusing primarily on suboptimality of random projections for information preservation [24], [25].

Contribution. In this article, we address this gap by systematically analyzing SMI across diverse settings, demonstrating that it frequently exhibits counterintuitive behavior and fails to accurately capture statistical dependence dynamics. Our key contributions are:

1. **Saturation and Sensitivity Analysis.** Through theoretical analysis and extensive benchmarking, we show that SMI saturates prematurely, even for low-dimensional synthetic problems, and fails to detect significant increases in statistical dependence.
2. **Redundancy Bias.** We refute the prevailing assumption that SMI favors linearly extractable information by constructing an explicit example where introducing such structure increases MI and even linear correlation, but decreases SMI. In fact, we show that SMI prioritizes information *redundancy* over information content. We argue that this bias can lead to catastrophic failures in some applications, e.g. collapses in representation learning.
3. **Curse of Dimensionality.** We revisit the dynamics of SMI for increasing dimensionality and argue that SMI is, in fact, cursed, with the curse of dimensionality manifesting itself not through sample complexity, but via asymptotic decay to zero in high-dimensional regimes due to diminishing redundancy.
4. **Reestablishing the Trade-off.** Finally, we discuss to which extent the aforementioned problems can be solved by using non-uniform/non-random slicing strategies, and how they affect the trade-off between scalability and utility of different measures of statistical dependence.

Our paper is structured as follows. In Section 2, we provide the mathematical background that is necessary for our analysis. In Section 3, we discuss previous findings which are related to the research topic of this work. Section 4 consists of our main theoretical results, with the complete proofs being provided in Section B. In Section 5, we employ synthetic benchmarks to show the disconnection between dynamics of MI and SMI. Section 6 illustrates that tasks related to SMI maximization may yield degenerate solutions, contrary to MI maximization. Finally, we discuss our results in Section 7.

2 Preliminaries

Elements of Information Theory. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with sample space Ω, σ -algebra $\mathcal{F}$, and probability measure $\mathbb{P}$ defined on $\mathcal{F}$. Consider random vectors $X : \Omega \to \mathbb{R}^{d_x}$ and $Y : \Omega \to \mathbb{R}^{d_y}$ with joint distribution $\mathbb{P}_{X,Y}$ and marginals $\mathbb{P}_X$ and $\mathbb{P}_Y$, respectively. Wherever it is needed, we assume the relevant Radon-Nikodym derivatives exist. For any probability measure $\mathbb{Q} \ll \mathbb{P}$, the Kullback-Leibler (KL) divergence is $\mathrm{D}_{\mathsf{KL}}(\mathbb{Q} \,\|\, \mathbb{P}) = \mathbb{E}_{\mathbb{Q}}\left[\log \frac{\mathrm{d}\,\mathbb{Q}}{\mathrm{d}\,\mathbb{P}}\right]$, which is non-negative and vanishes if and only if (iff) $\mathbb{P} = \mathbb{Q}$. The mutual information (MI) between X and Y quantifies the divergence between the joint distribution and the product of marginals:

$$\mathsf{I}(X;Y) = \mathbb{E} \log \frac{\mathrm{d}\,\mathbb{P}_{X,Y}}{\mathrm{d}\,\mathbb{P}_X \otimes \mathbb{P}_Y} = \mathrm{D}_{\mathsf{KL}}\big(\mathbb{P}_{X,Y} \,\big\|\, \mathbb{P}_X \otimes \mathbb{P}_Y\big).$$

When $\mathbb{P}_X$ admits a probability density function (PDF) $p(X)$ with respect to (w.r.t.) the Lebesgue measure, the differential entropy is defined as $\mathsf{h}(X) = -\mathbb{E}[\log p(X)]$, where $\log(\,\cdot\,)$ denotes the natural logarithm. Likewise, the joint entropy $\mathsf{h}(X,Y)$ is defined via the joint density $p(X,Y)$, and conditional entropy is $\mathsf{h}(X \,|\, Y) = -\mathbb{E}[\log p(X \,|\, Y)] = -\mathbb{E}_Y\left[\mathbb{E}_{X \,|\, Y} \log p(X \,|\, Y)\right]$. Under the existence of PDFs, MI satisfies the identities

$$\mathsf{I}(X;Y) = \mathsf{h}(X) - \mathsf{h}(X \,|\, Y) = \mathsf{h}(Y) - \mathsf{h}(Y \,|\, X) = \mathsf{h}(X) + \mathsf{h}(Y) - \mathsf{h}(X,Y). \tag{2}$$

In this work, we denote by μ_{M} the normalized Haar (uniform) probability measure on a compact manifold M, i.e., the unique bi-invariant measure satisfying $\mu_{\mathrm{M}}(\mathrm{M}) = 1$. Hence, to sample uniformly from specific spaces we write $\mathrm{W} \sim \mu_{\mathrm{O}(d)}, \theta \sim \mu_{\mathbb{S}^{d-1}}, \mathrm{A} \sim \mu_{\mathrm{St}(k,d)}$, indicating draws from the Haar measures on orthogonal group $\mathrm{O}(d) = \big\{\mathrm{Q} \in \mathbb{R}^{d\times d} : \mathrm{Q}^{\mathsf{T}}\mathrm{Q} = \mathrm{Q}\mathrm{Q}^{\mathsf{T}} = \mathrm{I}\big\}$, the unit sphere $\mathbb{S}^{d-1} = \big\{X \in \mathbb{R}^d : \|X\|_2 = 1\big\}$, and the Stiefel manifold $\mathrm{St}(k,d) = \big\{\mathrm{Q} \in \mathbb{R}^{d\times k} : \mathrm{Q}^{\mathsf{T}}\mathrm{Q} = \mathrm{I}\big\}$, respectively.

Sliced Mutual Information. To mitigate the curse of dimensionality, one may average MI over all k-dimensional projections. The k-sliced mutual information (k-SMI) [23] between X and Y is defined as

$$\mathsf{SI}_k(X;Y) = \int_{\mathrm{St}(k,d_x)} \int_{\mathrm{St}(k,d_y)} \mathsf{I}(\Theta^{\mathsf{T}}X; \Phi^{\mathsf{T}}Y)\, \mathrm{d}\mu_{\mathrm{St}(k,d_x)}(\Theta)\, \mathrm{d}\mu_{\mathrm{St}(k,d_y)}(\Phi),$$

which can be efficiently estimated. Setting $k = 1$ recovers the standard sliced mutual information (1).

3 Background

Merits of SMI are straightforward and have been investigated thoroughly in [22], [23]. We remind the reader of the two most important of them:

1. **Scalability** (i.e., fast convergence in high dimensions), enabled by low-dimensional projections.
2. **Nullification Property** (i.e., $\mathsf{SI}_k(X;Y) = 0$ iff X and Y are independent), which stems from the projections being random and independent.

In contrast, demerits of SMI are not very obvious and not well-covered in the literature. In this section, we recapitulate and analyze previous works which address the shortcomings of SMI. To facilitate the analysis, we divide them into three main categories.

Suboptimality of random slicing. In [24] and [25], it is argued that a uniform slicing strategy can produce suboptimal projections, impairing SMI's ability to capture dependencies in the presence of noisy or non-informative components. To address this issue, [24] proposed max-sliced MI (mSMI), which selects non-random projectors that maximize the MI between projected representations. This approach is also claimed to improve interpretability and convergence rates.

However, deterministic slicing may overlook dependencies captured by non-optimal components. To mitigate this, [25] extends the max-sliced approach by optimizing SMI over probability distributions of projectors, with regularization to maintain slice diversity. While the authors emphasize that optimization should occur over *joint* distributions, their motivation primarily addresses the issue of non-optimal *marginal* distributions of θ and ϕ — specifically, the presence of non-informative

components in X and Y. We contend that this represents only a partial understanding of the problem, as many SMI artifacts arise from other factors. Needless to say that optimization over probability distributions is also a heavy burden, which does not align with the slicing philosophy.

Data Processing Inequality violation. A fundamental property of MI is that it cannot be increased by deterministic processing or, more generally, by Markov kernels. Furthermore, MI is preserved under invertible transformations. This is formalized by the *data processing inequality* (DPI).

Theorem 3.1. (Theorem 3.7 in [1]) For a Markov chain $X \to Y \to Z$, $\mathsf{I}(X;Y) \geq \mathsf{I}(X;Z)$. Additionally, if $Z = f(Y)$ where f is measurably invertible, then equality holds.

In contrast to MI, SMI violates the DPI (see Section 3.2 in [22] for an example). While the intuition behind DPI is clear (raw data already contains full information, and processing can only destroy it), the implications of DPI violation are less straightforward.

Existing works suggest that SMI's violation of DPI can reflect a preference for linearly extractable features, framing this as a useful property that aligns with the informal understanding of "practically available" (i.e., easily accessible) information [22], [26], [30]. However, this interpretation can be misleading if the factors behind SMI increases are misidentified. Our analysis reveals that this is indeed the case, as SMI exhibits more inherent biases than previously recognized.

Asymptotics in high-dimensional regime. Convergence analysis suggests that the sample complexity of SMI estimation is far less sensitive to data dimensionality compared to that of MI. In fact, it has been argued that the estimation error may even decrease with dimensionality in some cases (see Remark 4 in [23]). However, an analysis of SMI itself reveals that this behavior may result from the fact that SMI can decrease as dimensionality grows. Specifically, Theorem 3 in [23] provides an asymptotic expression (as $d \to \infty$) for SMI in the case of jointly normal X and Y, which decays hyperbolically with d under some circumstances.

To date, no explanation for this phenomenon has been provided in the literature. We therefore elaborate on this finding by deriving non-asymptotic expressions, along with experimental results for non-Gaussian data, which reveal further nuances behind the decay.

4 Theoretical analysis

We start our analysis with considering a simple example, which (a) admits closed-form expression for SMI and (b) is capable of illustrating severe problems of the quantity in question.

Lemma 4.1. Consider the following pair of jointly Gaussian d-dimensional random vectors:

$$(X, Y) \sim \mathcal{N}\left(0, \begin{pmatrix} \mathrm{I} & \rho\mathrm{I} \\ \rho\mathrm{I} & \mathrm{I} \end{pmatrix}\right), \quad \rho \in (-1; 1).$$

In this setup, MI and SMI can be calculated analytically:

$$\mathsf{I}(X;Y) = -\frac{d}{2}\log(1-\rho^2), \qquad \mathsf{SI}(X;Y) = \frac{\rho^2}{2d}\,{}_3F_2\left(1, 1, \frac{3}{2}; \frac{d}{2}+1, 2; \rho^2\right),$$

where ${}_3F_2$ is the *generalized hypergeometric function*. Additionally, the following limits hold:

$$\lim_{d\to\infty} \mathsf{I}(X;Y) = +\infty \qquad \lim_{d\to\infty} \mathsf{SI}(X;Y) = 0$$

$$\lim_{\rho^2\to 1} \mathsf{I}(X;Y) = +\infty \qquad \lim_{\rho^2\to 1} \mathsf{SI}(X;Y) = \psi(d-1) - \psi\left(\frac{d-1}{2}\right) - \log 2 \leq \frac{3}{d-1},$$

with ψ being the *digamma function*.

Note that while MI correctly captures the growing statistical dependence as $d \to \infty$ (since additional components contribute shared information), SMI drops to zero, exposing a fundamental problem. This issue was briefly noted in [23], but only through providing an asymptotic expression without further discussion. We interpret this behavior as a distinct manifestation of the **curse of dimensionality**: as d grows, SMI uniformly decays to zero and becomes ineffective for statistical analysis.

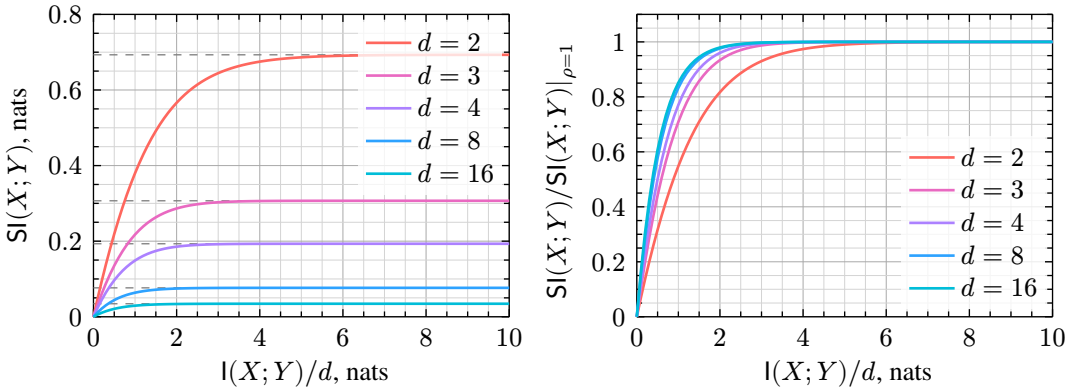


Figure 2: Saturation of $\mathsf{SI}(X;Y)$ as function of $\mathsf{I}(X;Y)/d$ for the example from Lemma 4.1, non-normalized (left) and normalized (right) versions. Note that the problem becomes more prominent in higher dimensions, both because of lower plateau and faster saturation.

The second pair of limits reveals another critical flaw of SMI. When $\rho^2 \to 1$, the X-Y relationship becomes deterministic — a property MI reflects successfully. In stark contrast, SMI remains bounded by a dimension-dependent factor that decays hyperbolically. Furthermore, plotting SMI against MI shows this bound is reached prematurely, demonstrating SMI's **rapid saturation** with increasing dependence (Figure 2). In this saturated regime, SMI becomes effectively insensitive to further growth in shared information. Moreover, this renders estimates of SMI for different dimensionalities fundamentally incomparable, as they are theoretically bounded by factors depending on d.

These phenomena can not be explained by suboptimality of individual projections. In fact, each individual projection is optimal, as $\mathsf{I}(\theta^\mathsf{T} X; Y)$ does not depend on θ in this particular example. The proof of Lemma 4.1 suggests that the problem arises from the majority of *pairs* of projectors being suboptimal, yielding near-independent $\theta^\mathsf{T} X$ and $\phi^\mathsf{T} Y$ in the most outcomes, even for $d = 2$. Although similar analysis for k-SMI is extremely challenging, we argue that the problems in question prevail even when employing k-rank projectors.

Proposition 4.2. Under the setup of Lemma 4.1, k-SMI has the following integral representation

$$\mathsf{SI}_k(X;Y) = -\frac{1}{2} \int_{[0,1]^k} \sum_{i=1}^{k} \log(1 - \rho^2 \lambda_i)\, p(\boldsymbol{\lambda})\, \mathrm{d}\boldsymbol{\lambda},$$

where $p(\boldsymbol{\lambda}) \propto \prod_{i<j} |\lambda_j - \lambda_i| \underbrace{\prod_{i=1}^{k} (1 - \lambda_i)^{(d-2k-1)/2}}_{(\star)}$.

***Remark*. 4.3.** As the dimension d grows, the term $(\star)$ asymptotically concentrates the eigenvalues λ_i near zero, leading to the decay of SI_k to zero.

We argue that the limitations we uncovered can be attributed to a strong bias of SMI toward **information redundancy**. That is, SMI favors repetition of information across different axes, and suffers from the curse of dimensionality if X and Y have high entropy. The following proposition and remark present a simple example to clarify this bias.

Proposition 4.4. Let X and Y be d_x, d_y-dimensional random vectors correspondingly, with $d_x, d_y < k$. Let $\mathrm{A} \in \mathbb{R}^{m_x \times d_x}$ and $\mathrm{B} \in \mathbb{R}^{m_y \times d_y}$ be matrices of ranks d_x, d_y. Then $\mathsf{SI}_k(\mathrm{A}X; \mathrm{B}Y) = \mathsf{I}(X;Y)$.

Corollary 4.5. Consider the following pair of jointly Gaussian d-dimensional random vectors:

$$(X, Y) \sim \mathcal{N}\left(0, \begin{pmatrix} \mathrm{J} & \rho\mathrm{J} \\ \rho\mathrm{J} & \mathrm{J} \end{pmatrix}\right), \quad \rho \in (-1; 1),$$

where $\mathrm{J} = \mathbf{1} \cdot \mathbf{1}^\mathsf{T}$ with $\mathbf{1}^\mathsf{T} = (1, ..., 1)$. Then $\mathsf{SI}_k(X;Y) = \mathsf{I}(X;Y) = -\frac{1}{2}\log(1 - \rho^2)$.

***Remark*. 4.6.** Applying $\mathbf{1} \cdot e_1^\mathsf{T}$ to the random vectors from Lemma 4.1 individually yields the example from Corollary 4.5. Therefore, this linear transform increases SMI despite decreasing MI.

4.1 Extension to optimal slicing

Although our work primarily focuses on conventional (average) sliced mutual information (SMI), as it is the most widely used variant, we also provide some intuition regarding the limitations of its "optimal" counterparts: max-sliced MI (mSMI) [24] and *optimal-sliced* MI (oSMI) [25]. Since mSMI is a special case of oSMI without regularization constraints, we restrict our discussion to mSMI, though our reasoning extends to oSMI as well. The k-mSMI is defined as:

$$\overline{\mathsf{SI}}_k(X;Y) = \sup_{\substack{\Theta \in \mathrm{St}(d_x,k) \\ \Phi \in \mathrm{St}(d_y,k)}} \mathsf{I}(\Theta^\mathsf{T} X; \Phi^\mathsf{T} Y) \tag{3}$$

To highlight the shortcomings of linear compression, we revisit a Gaussian example. The following proposition demonstrates that even in this simple setting, mSMI captures only a subset of dependencies and can exhibit opposite trends to MI. This occurs, for instance, when dependencies become more evenly distributed across components, which again returns us to the **redundancy bias**.

Proposition 4.7. (Proposition 2 in [24]) Let $(X, Y) \sim \mathcal{N}(\mu, \Sigma)$, with marginal covariances Σ_X, Σ_Y and cross-covariance Σ_{XY}. Suppose the matrix $\Sigma_X^{-\frac{1}{2}} \Sigma_{XY} \Sigma_Y^{-\frac{1}{2}}$ exists, and let $\{\rho_i\}_{i=1}^d$ denote its singular values in descending order, where $d = \min\big(d_x, d_y\big)$. Then

$$\mathsf{I}(X;Y) = -\frac{1}{2}\sum_{i=1}^{d} \log(1-\rho_i^2), \qquad \overline{\mathsf{SI}}_k(X;Y) = -\frac{1}{2}\sum_{i=1}^{k} \log(1-\rho_i^2).$$

5 Synthetic Experiments

To complement the theoretical analysis from the previous section and address complex, non-Gaussian cases, we conduct an extensive benchmarking of SMI using synthetic tests from [34], based on the works of [35], [36]. This benchmark suite is used to evaluate MI estimators. However, we do not assess whether SMI estimates converge to ground-truth MI values. SMI is a *distinct measure of statistical dependance*, and should not be viewed as an approximation of MI. Instead, our analysis focuses on the relationship between the two measures: since MI captures the true degree of statistical dependence, opposing trends in MI and SMI reveal problems with the latter quantity.

For the experiments, we use *correlated normal*, *correlated uniform*, *smoothed uniform* and *log-gamma-exponential* distributions, for which the ground-truth value of MI is available. To increase the dimensionality, we use independent components with equally distributed per-component MI. These setups will be referred to as "randomized" and "non-randomized" correpsondingly. For each distribution, we vary both the data dimensionality (d) and the projection dimensionality ($k < d$).

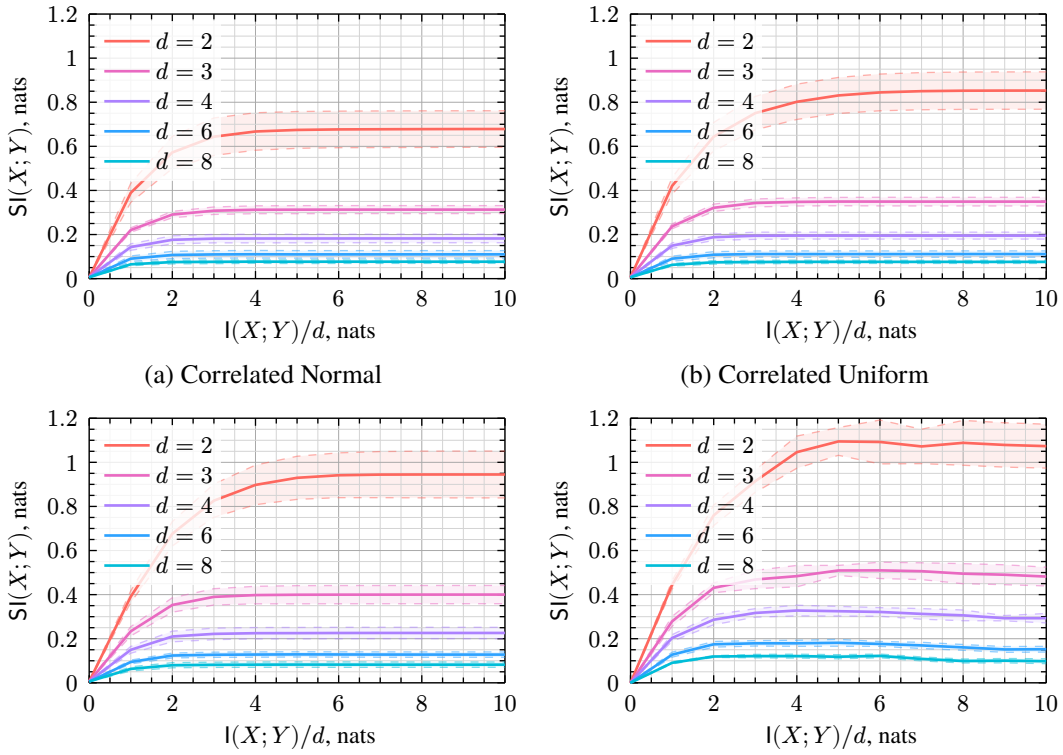
To estimate MI between projections, we use the KSG estimator [35] with the number of neighbors fixed at 1. For each configuration, we conduct 10 independent runs with different random seeds to compute means and standard deviations. Our experiments use 10^4 samples for (X, Y) and 128 samples for (Θ, Φ).

To experimentally verify saturation, we plot SMI against MI normalized by dimensionality d in Figure 3. The plots clearly show that SMI reaches a plateau relatively early for all the featured distributions. The results for the normal distribution also align well with those from Lemma 4.1. We further confirm the saturation of k-SMI for $k \in \{2, 3\}$ experimentally in Section C. Finally, we plot the saturated values against d on a log-log scale, demonstrating that the $1/d$ trend from Lemma 4.1 also holds for non-Gaussian distributions.

6 SMI for InfoMax-like tasks

Since mutual information is interpretable and captures non-linear dependencies, it is widely used as a training objective. Many applications involve maximizing MI (InfoMax) for feature selection [5], [6], [7] and self-supervised representation learning [8], [9], [10], [11], [12], [13]. However, due to the curse of dimensionality, alternative objectives have been proposed, with some works using sliced mutual information maximization for feature extraction [22] and disentanglement in InfoGAN [23].

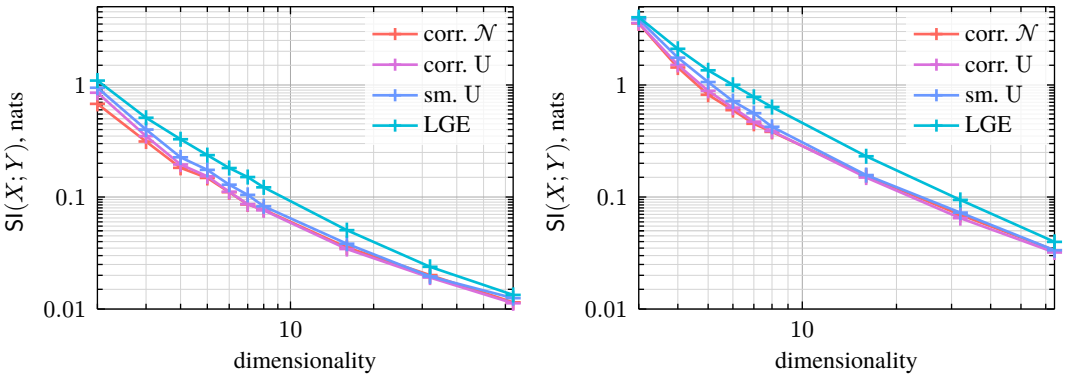
In this section, we argue that SMI is not a suitable alternative to MI for InfoMax tasks. Since SMI exhibits a strong preference for redundancy, SMI maximization may lead to collapsed (high-redundancy) solutions. We demonstrate this through two experiments. Firstly, we revisit the Gaussian noisy channel to demonstrate that SMI favors linear mappings which decrease robustness to noise. Then, we consider a self-supervised representation learning task and show that using SMI immediately leads to collapsed representations.



(a) Correlated Normal (b) Correlated Uniform

(c) Smoothed Uniform (d) Log-Gamma-Exponential

Figure 3: Results of synthetic experiments with different distributions. We report mean values and standard deviations computed across 10 runs, with 10^4 samples used for MI estimation and 128 for averaging across projections.



(a) $k = 1$ (b) $k = 2$

Figure 4: Decaying trends of k-SMI for *correlated normal* (corr. $\mathcal{N}$), *correlated uniform* (corr. U), *smoothed uniform* (sm. U) and *log-gamma-exponential* (LGE). We plot saturated values of k-SMI against data dimensionality d. Log scale is used to illustrate the $1/d$ trend predicted in Lemma 4.1.

6.1 Gaussian Channel

Let X be a zero-mean d-dimensional random vector, and let $Z \sim \mathcal{N}(0, \sigma \mathrm{I})$ be an independent noise. Additive white noise Gaussian (AWGN) channel is defined as $X \to X + Z$. Maximization of $\mathsf{I}(X; X + Z)$ w.r.t. the distribution of X is a classical information transmission problem, which arises in many fields under the Gaussian noise assumption. Given energy constraints, it admits an analytical solution [37]:

$$\sup_{\mathbb{E}\, X_i^2 = 1} \mathsf{I}(X; X + Z) = \frac{d}{2} \log\left(1 + \frac{1}{\sigma^2}\right), \qquad X_{\text{opt}} \sim \mathcal{N}(0, \mathrm{I}) \tag{4}$$

It is somewhat intuitive that unit covariance matrix allows for more information to be transmitted, as all the components of X are utilized to full extent. However, due to the redundancy bias, SMI prefers less robust distributions. To demonstrate this, we consider two linear normalization mappings which impose energy constraints on a vector X with zero mean and covariance Σ:

1. *Whitening*: $\Sigma^{-1/2} X$;
2. *Standardization*: $\mathrm{D}^{-1/2} X$, where $\mathrm{D} = \operatorname{diag}(\Sigma)$.

We conduct numerical experiments for $\sigma = 0.1$, $X' \sim \mathrm{A} \cdot \mathrm{U}([-1; 1]^5)$ and $X'' \sim \mathrm{A} \cdot \mathcal{N}(0, \mathrm{I}_5)$, where $\mathrm{A} = 10^{-2} \cdot \mathrm{I} + \mathbf{1} \cdot \mathbf{1}^{\mathsf{T}}$ is an ill-conditioned matrix. We employ the same estimators and hyperparameters as in Section 5. The results are presented in Table 1.

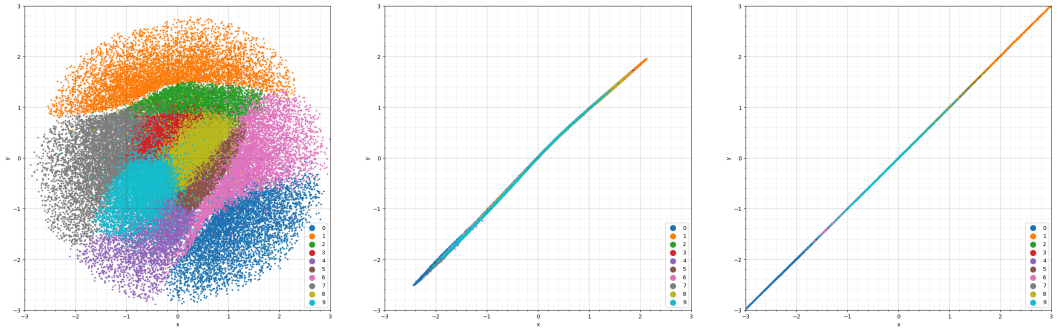
Table 1: Results for additive white Gaussian noise channel ($\sigma = 0.1$), mean and std for 10 runs.

	MI		SMI		2-SMI	
	$\Sigma^{-1/2}$	$\mathrm{D}^{-1/2}$	$\Sigma^{-1/2}$	$\mathrm{D}^{-1/2}$	$\Sigma^{-1/2}$	$\mathrm{D}^{-1/2}$
X'	7.48 ± 0.01	3.04 ± 0.01	0.17 ± 0.02	1.82 ± 0.04	0.96 ± 0.04	2.46 ± 0.03
X''	7.49 ± 0.02	3.04 ± 0.01	0.14 ± 0.02	1.83 ± 0.04	0.82 ± 0.05	2.49 ± 0.05

6.2 Representation Learning

To further demonstrate SMI's sensitivity to information redundancy, we examine its performance in learning compressed representations through mutual information maximization (*Deep InfoMax*) [8]. This approach is known to be equivalent to many popular contrastive self-supervised learning methods [13].

In Deep InfoMax, an encoder network f is trained to maximize a lower bound on $\mathsf{I}(X; f(X))$, where X represents input data and $f(X)$ its compressed representation. This method is theoretically sound, as maximizing MI ensures the most informative embeddings under the latent space dimensionality



(a) MI $\to$ max, 2000 epochs. (b) SMI $\to$ max, 10 epochs. (c) SMI $\to$ max, 2000 epochs.

Figure 5: Visualizations of embeddings from the representation learning experiments, with points colored by class. Note that mutual information maximization (left) produces clustered low-redundancy representations, while SMI maximization results in immediate (after 10 epochs) collapse.

constraint. For our study, we replace MI with SMI in this framework. This substitution is straightforward since both MI and SMI admit Donsker-Varadhan variational lower bounds [38]:

$$\begin{aligned} \mathsf{I}(X;Y) &= \sup_{T:\Omega\to\mathbb{R}} \Big[\mathbb{E}_{\mathbb{P}_{X,Y}} T(X,Y) - \log\Big(\mathbb{E}_{\mathbb{P}_X\otimes\mathbb{P}_Y} e^{T(X,Y)}\Big)\Big], \\ \mathsf{SI}_k(X;Y) &= \sup_{T:\Omega\to\mathbb{R}} \mathbb{E}_{\Theta,\Phi}\Big[\mathbb{E}_{\mathbb{P}_{X,Y}} T(\Theta^\mathsf{T}X, \Phi^\mathsf{T}Y, \Theta, \Phi) - \log\Big(\mathbb{E}_{\mathbb{P}_X\otimes\mathbb{P}_Y} e^{T(\Theta^\mathsf{T}X,\Phi^\mathsf{T}Y,\Theta,\Phi)}\Big)\Big], \end{aligned} \tag{5}$$

where T is a critic function, which is also approximated in practice by a neural network. For detailed derivations of these bounds, we refer the reader to [39] (MI) and [22], [23] (SMI).

We strictly follow the experimental protocol from [13]. In particular, we use MNIST handwritten digits dataset [40], employ InfoNCE loss [41] to approximate (5), use convolutional network for f and fully-connected network for T. Latent space dimensionality is fixed at $d = 2$ for visualization purposes. Small Gaussian noise is added to the outlet of the encoder to combat representation collapse [13]. More details are provided in Section D. We focus on this simple setup because our objective is to show that SMI produces degenerate results even in elementary tasks, making more complex configurations unnecessary for this demonstration.

Results are presented in Figure 5. As our theory predicts, maximization of SMI immediately leads to collapsed representations, while conventional InfoMax yields embeddings with low or even zero redundancy (components are close to $\mathcal{N}(0, \mathrm{I})$). This behavior is consistent across different runs.

7 Discussion

Results. Sliced mutual information (SMI) has been proposed as a scalable alternative to Shannon's mutual information. While SMI enables efficient computation in high-dimensional settings and satisfies the nullification property, our findings reveal critical deficiencies that undermine its reliability for feature extraction and related tasks.

We demonstrate that SMI saturates rapidly, failing to capture variations in statistical dependence. This makes it difficult to distinguish between intrinsic SMI fluctuations and genuine changes in dependence structure. Furthermore, we invalidate the common hypothesis that SMI favors linear features through a counterexample where even correlation coefficients reflect dependence more faithfully than SMI, which exhibits inverted behavior.

In high-dimensional spaces, SMI decays with increasing dimensionality, contrary to MI's monotonic behavior. This is established analytically for Gaussian cases and validated empirically across diverse synthetic experiments. Consequently, SMI variations may reflect redundancy, dependence changes, or high-dimensional artifacts without a principled way to disentangle these factors.

Impact. Thanks to fast convergence rates and the absence of additional optimization problems, SMI has been widely applied across various fields of statistics and machine learning. Given our findings, it is therefore crucial to recognize how the inherent biases of SMI affect practical applications.

The works [22] and [23] propose using SMI in a Deep InfoMax setting. However, we demonstrate that maximizing SMI can lead to collapsed solutions due to redundancy bias. Meanwhile, [26], [27], [28], [30] study deep neural networks by measuring SMI between intermediate layers. Yet, as our analysis reveals, changes in SMI do not always reflect true shifts in statistical dependence; they may instead result from differences in layer dimensionality, redundancy in intermediate representations, low sensitivity in saturated regimes, or other factors. Finally, [33] suggests using SMI for independence testing in differential privacy tasks. We contend that this approach poses critical issues, as SMI estimates can become statistically indistinguishable from zero in high-dimensional or low-redundancy settings.

Limitations. While we support our claims with both theoretical analysis and experimental evidence, we were able to derive analytical expressions for the Gaussian case only. Furthermore, our synthetic tests do not feature complex, highly non-linear distributions (such as structured image data used in [17]). Nevertheless, we demonstrate that our findings are more than sufficient to expose fundamental limitations of SMI, and to support all the claims we made.

References

[1] Y. Polyanskiy and Y. Wu, *Information Theory: From Coding to Learning*. Cambridge University Press, 2024. [Online]. Available: https://books.google.ru/books?id=CySo0AEACAAJ

[2] A. Asadi, E. Abbe, and S. Verdu, "Chaining Mutual Information and Tightening Generalization Bounds," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., Curran Associates, Inc., 2018, p. . [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/8d7628dd7a710c8638dbd22d4421ee46-Paper.pdf

[3] J. Negrea, M. Haghifam, G. K. Dziugaite, A. Khisti, and D. M. Roy, "Information-Theoretic Generalization Bounds for SGLD via Data-Dependent Estimates," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., Curran Associates, Inc., 2019, p. . [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/05ae14d7ae387b93370d142d82220f1b-Paper.pdf

[4] B. Duong and T. Nguyen, "Conditional Independence Testing via Latent Representation Learning," in *2022 IEEE International Conference on Data Mining (ICDM)*, Los Alamitos, CA, USA: IEEE Computer Society, Dec. 2022, pp. 121–130. doi: 10.1109/ICDM54844.2022.00022.

[5] S. Yang and J. Gu, "Feature selection based on mutual information and redundancy-synergy coefficient," *J. Zhejiang Univ. Sci.*, vol. 5, no. 11, pp. 1382–1391, Nov. 2004.

[6] N. Kwak and C.-H. Choi, "Input feature selection by mutual information based on Parzen window," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1667–1671, 2002, doi: 10.1109/TPAMI.2002.1114861.

[7] M. A. Sulaiman and J. Labadin, "Feature selection based on mutual information," in *2015 9th International Conference on IT in Asia (CITA)*, 2015, pp. 1–6. doi: 10.1109/CITA.2015.7349827.

[8] R. D. Hjelm *et al.*, "Learning deep representations by mutual information estimation and maximization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bklr3j0cKX

[9] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning Representations by Maximizing Mutual Information Across Views," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., Curran Associates, Inc., 2019, p. . [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/ddf354219aac374f1d40b7e760ee5bb7-Paper.pdf

[10] P. Veličković, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm, "Deep Graph Infomax," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=rklz9iAcKQ

[11] 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. [Online]. Available: https://openreview.net/forum?id=rkxoh24FPH

[12] X. Yu, "Leveraging Superfluous Information in Contrastive Representation Learning." [Online]. Available: https://arxiv.org/abs/2408.10292

[13] I. Butakov, A. Semenenko, A. Tolmachev, A. Gladkov, M. Munkhoeva, and A. Frolov, "Efficient Distribution Matching of Representations via Noise-Injected Deep InfoMax," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=mAmCdASmJ5

[14] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," *2015 IEEE Information Theory Workshop (ITW)*, pp. 1–5, 2015.

[15] R. Shwartz-Ziv and N. Tishby, "Opening the Black Box of Deep Neural Networks via Information." 2017.

[16] Z. Goldfeld *et al.*, "Estimating Information Flow in Deep Neural Networks," in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., in Proceedings of Machine Learning Research, vol. 97. PMLR, 2019, pp. 2299–2308. [Online]. Available: https://proceedings.mlr.press/v97/goldfeld19a.html

[17] I. Butakov, A. Tolmachev, S. Malanchuk, A. Neopryatnaya, A. Frolov, and K. Andreev, "Information Bottleneck Analysis of Deep Neural Networks via Lossy Compression," in *The Twelfth International*

Conference on Learning Representations, 2024. [Online]. Available: https://openreview.net/forum?id=huGECz8dPp

[18] Z. Goldfeld, K. Greenewald, J. Niles-Weed, and Y. Polyanskiy, "Convergence of Smoothed Empirical Measures With Applications to Entropy Estimation," *IEEE Transactions on Information Theory*, vol. 66, no. 7, pp. 4368–4391, 2020, doi: 10.1109/TIT.2020.2975480.

[19] D. McAllester and K. Stratos, "Formal Limitations on the Measurement of Mutual Information," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, S. Chiappa and R. Calandra, Eds., in Proceedings of Machine Learning Research, vol. 108. PMLR, 2020, pp. 875–884. [Online]. Available: https://proceedings.mlr.press/v108/mcallester20a.html

[20] G. Gowri, X. Lun, A. M. Klein, and P. Yin, "Approximating mutual information of high-dimensional variables using learned representations," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. [Online]. Available: https://openreview.net/forum?id=HN05DQxyLl

[21] K. H. Greenewald, B. Kingsbury, and Y. Yu, "High-Dimensional Smoothed Entropy Estimation via Dimensionality Reduction," in *IEEE International Symposium on Information Theory, ISIT 2023, Taipei, Taiwan, June 25-30, 2023*, IEEE, 2023, pp. 2613–2618. doi: 10.1109/ISIT54713.2023.10206641.

[22] Z. Goldfeld and K. Greenewald, "Sliced Mutual Information: A Scalable Measure of Statistical Dependence," in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: https://openreview.net/forum?id=27qon5Ut4PSl

[23] Z. Goldfeld, K. Greenewald, T. Nuradha, and G. Reeves, "k-Sliced Mutual Information: A Quantitative Study of Scalability with Dimension," in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022. [Online]. Available: https://openreview.net/forum?id=L-ceBdl2DPb

[24] D. Tsur, Z. Goldfeld, and K. Greenewald, "Max-Sliced Mutual Information," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: https://openreview.net/forum?id=ce9B2x3zQa

[25] A. Fayad and M. Ibrahim, "On Slicing Optimality for Mutual Information," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: https://openreview.net/forum?id=JMuKfZx2xU

[26] S. Wongso, R. Ghosh, and M. Motani, "Understanding Deep Neural Networks Using Sliced Mutual Information," in *2022 IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 133–138. doi: 10.1109/ISIT50566.2022.9834357.

[27] S. Wongso, R. Ghosh, and M. Motani, "Using Sliced Mutual Information to Study Memorization and Generalization in Deep Neural Networks," in *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, F. Ruiz, J. Dy, and J.-W. van de Meent, Eds., in Proceedings of Machine Learning Research, vol. 206. PMLR, 2023, pp. 11608–11629. [Online]. Available: https://proceedings.mlr.press/v206/wongso23a.html

[28] S. Wongso, R. Ghosh, and M. Motani, "Pointwise Sliced Mutual Information for Neural Network Explainability," in *2023 IEEE International Symposium on Information Theory (ISIT)*, 2023, pp. 1776–1781. doi: 10.1109/ISIT54713.2023.10207010.

[29] J. Dentan, D. Buscaldi, A. Shabou, and S. Vanier, "Predicting and analyzing memorization within fine-tuned Large Language Models." [Online]. Available: https://arxiv.org/abs/2409.18858

[30] S. Wongso, R. Ghosh, and M. Motani, "Sliced Information Plane for Analysis of Deep Neural Networks," Jan. 2025, doi: 10.36227/techrxiv.173833980.08812687/v1.

[31] K. Nadjahi, K. Greenewald, R. B. Gabrielsson, and J. Solomon, "Slicing Mutual Information Generalization Bounds for Neural Networks," in *ICML 2023 Workshop Neural Compression: From Information Theory to Applications*, 2023. [Online]. Available: https://openreview.net/forum?id=cbLcwK3SZi

[32] Z. Hu, S. Kang, Q. Zeng, K. Huang, and Y. Yang, "InfoNet: Neural Estimation of Mutual Information without Test-Time Optimization," in *Forty-first International Conference on Machine Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=40hCy8n5XH

[33] T. Nuradha and Z. Goldfeld, "Pufferfish Privacy: An Information-Theoretic Study," *IEEE Trans. Inf. Theor.*, vol. 69, no. 11, pp. 7336–7356, Nov. 2023, doi: 10.1109/TIT.2023.3296288.

[34] I. Butakov *et al.*, "MUTINFO." [Online]. Available: https://github.com/VanessB/mutinfo

[35] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E*, vol. 69, no. 6, p. 66138, Jun. 2004, doi: 10.1103/PhysRevE.69.066138.

[36] F. Czyż Paweland Grabowski, J. Vogt, N. Beerenwinkel, and A. Marx, "Beyond Normal: On the Evaluation of Mutual Information Estimators," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., Curran Associates, Inc., 2023, pp. 16957–16990. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/36b80eae70ff629d667f210e13497edf-Paper-Conference.pdf

[37] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. USA: Wiley-Interscience, 2006.

[38] M. D. Donsker and S. R. Varadhan, "Asymptotic evaluation of certain markov process expectations for large time. IV," *Communications on Pure and Applied Mathematics*, vol. 36, no. 2, pp. 183–212, Mar. 1983, doi: 10.1002/cpa.3160360204.

[39] M. I. Belghazi *et al.*, "Mutual Information Neural Estimation," in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., in Proceedings of Machine Learning Research, vol. 80. PMLR, 2018, pp. 531–540. [Online]. Available: https://proceedings.mlr.press/v80/belghazi18a.html

[40] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[41] A. van den Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding." [Online]. Available: https://arxiv.org/abs/1807.03748

[42] A. Edelman and B. D. Sutton, "The beta-Jacobi matrix model, the CS decomposition, and generalized singular value problems," *Foundations of Computational Mathematics*, vol. 8, no. 2, pp. 259–285, 2008.

[43] A. McBride, "Special functions, by George E. Andrews, Richard Askey and Ranjan Roy. Pp. 664.£ 60. 1999. ISBN 0 521 62321 9 (Cambridge University Press.)," *The Mathematical Gazette*, vol. 83, no. 497, pp. 355–357, 1999.

[44] N. Elezovic, C. Giordano, and J. Pecaric, "The best bounds in Gautschi's inequality," *Math. Inequal. Appl*, vol. 3, no. 2, pp. 239–252, 2000.

[45] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization." 2017.

A Supplementary theory

Lemma A.1. (Example 2.4 in [1]) $\mathsf{h}(\mathcal{N}(\mu, \Sigma)) = \frac{1}{2}\log\big((2\pi e)^d \det \Sigma\big)$.

Corollary A.2. For $(X, Y) \sim \mathcal{N}(\mu, \Sigma)$ with non-singular Σ

$$\begin{aligned}\mathsf{I}(X;Y) &= \frac{1}{2}\log\det\Sigma_X + \frac{1}{2}\log\det\Sigma_Y - \frac{1}{2}\log\det\Sigma \\ &= -\frac{1}{2}\sum_{i=1}^{d}\log(1-\rho_i^2),\end{aligned}$$

where Σ_X, Σ_Y are marginal covariances, Σ_{XY} is cross-covariance, $d = \min\big(d_x, d_y\big)$, and $\{\rho_i\}_{i=1}^{d}$ are singular values of $\Sigma_X^{-\frac{1}{2}}\Sigma_{XY}\Sigma_Y^{-\frac{1}{2}}$.

Proof of Corollary A.2. Combining Lemma A.1 and (2) yields the first result. Now note that

$$\mathsf{I}(X;Y) = \mathsf{I}\Big(\Sigma_X^{-\frac{1}{2}}X; \Sigma_Y^{-\frac{1}{2}}Y\Big) = \mathsf{I}\Big(\mathrm{U}^\mathsf{T}\Sigma_X^{-\frac{1}{2}}X; \mathrm{V}\Sigma_Y^{-\frac{1}{2}}Y\Big),$$

where $\mathrm{U}\,\mathrm{diag}(\rho_i)\mathrm{V}^\mathsf{T}$ is the SVD of $\Sigma_X^{-\frac{1}{2}}\Sigma_{XY}\Sigma_Y^{-\frac{1}{2}}$. However,

$$\Big(\mathrm{U}^\mathsf{T}\Sigma_X^{-\frac{1}{2}}X, \mathrm{V}\Sigma_Y^{-\frac{1}{2}}Y\Big) \sim \mathcal{N}\left(\mu', \begin{pmatrix}\mathrm{I} & \mathrm{diag}(\rho_i) \\ \mathrm{diag}(\rho_i) & \mathrm{I}\end{pmatrix}\right),$$

from which we arrive at the second expression. □

Lemma A.3. Let $\mathrm{A} \in \mathbb{R}^{n\times m}$ be full column-rank matrix and $\Theta \sim \mu_{\mathrm{St}(n,k)}$ Then $\Theta^\mathsf{T}\mathrm{A}$ is full-rank with probability one.

Proof of Lemma A.3. Performing QR decomposition of A yields $\Theta^\mathsf{T}\mathrm{A} = \Theta^\mathsf{T}\mathrm{QR} \overset{\mathrm{d}}{=} \Theta^\mathsf{T}\begin{pmatrix}\mathrm{I}_m \\ 0\end{pmatrix}\mathrm{R}$. Since A is full-rank, R is invertible and $\operatorname{rank}\Theta^\mathsf{T}\mathrm{A} = \operatorname{rank}\Theta^\mathsf{T}\begin{pmatrix}\mathrm{I}_m \\ 0\end{pmatrix}$. Therefore,

$$\mathbb{P}\{\Theta^\mathsf{T}\mathrm{A} \text{ is full-rank}\} = 1 - \mathbb{P}\left\{\Theta^\mathsf{T}\begin{pmatrix}\mathrm{I}_m \\ 0\end{pmatrix} \text{ is not full-rank}\right\} = 1 - 0 = 1.$$

□

Lemma A.4. (Theorem 1.5 in [42]) Let $\mathrm{W} \sim \mu_{\mathrm{O}(d)}$ and partition

$$\mathrm{W} = \begin{pmatrix}\mathrm{W}_{11} & \mathrm{W}_{12} \\ \mathrm{W}_{21} & \mathrm{W}_{22}\end{pmatrix}.$$

with W_{11} of size k by k. Then the eigenvalues $\{\lambda_i\}_{i=1}^{k}$ of $\mathrm{W}_{11}\mathrm{W}_{11}^\mathsf{T}$ follow the Jacobi ensemble

$$p(\boldsymbol{\lambda}) \propto \prod_{i<j}|\lambda_i - \lambda_j|^\beta \prod_{i=1}^{k}\lambda_i^{\frac{\beta}{2}(a+1)-1}(1-\lambda_i)^{\frac{\beta}{2}(b+1)-1}$$

with parameters $a = 0, b = d - 2k$, and $\beta = 1$ (over $\mathbb{R}$).

Proof of Lemma A.3. Let $\mathrm{A}_1 \in \mathbb{R}^{k\times d}$ and $\mathrm{A}_2 \in \mathbb{R}^{(d-k)\times d}$ be independent matrices with i.i.d. entries from $\mathcal{N}(0,1)$. By stacking A_1 atop A_2 and then performing a block QR decomposition on the resulting Gaussian matrix, the orthogonal invariance of the Gaussian law implies that the two Q-blocks are independent of the upper-triangular factor R, with Q_1 and Q_2 uniformly distributed on $\mathrm{O}(k)$ and $\mathrm{St}(k, d-k)$, respectively. Finally, computing the SVD of the block rows together with R yields the generalized singular value decomposition (GSVD) of the pair $(\mathrm{A}_1, \mathrm{A}_2)$:

$$\begin{pmatrix}\mathrm{A}_1 \\ \mathrm{A}_2\end{pmatrix} = \begin{pmatrix}\mathrm{Q}_1 \\ \mathrm{Q}_2\end{pmatrix}\mathrm{R} = \begin{pmatrix}\mathrm{U}_1 & \\ & \mathrm{U}_2\end{pmatrix}\begin{pmatrix}\tilde{\mathrm{C}} \\ 0 \\ \hline -\tilde{\mathrm{S}} \\ 0\end{pmatrix}\tilde{\mathrm{V}}^\mathsf{T}\mathrm{R},$$

where $\mathrm{U}_1 \in \mathrm{O}(k), \mathrm{U}_2 \in \mathrm{O}(d-k), \tilde{\mathrm{V}} \in \mathrm{O}(k)$, and $\mathrm{C} = \mathrm{diag}(c_i)$, $\mathrm{S} = \mathrm{diag}(s_i)$ with $c_i \geq 0$, $s_i \geq 0$, and $c_i^2 + s_i^2 = 1$ for all i. The diagonal entries of $\tilde{\mathrm{C}}$ are known as the generalized singular values of the pair $(\mathrm{A}_1, \mathrm{A}_2)$.

For a matrix $\mathrm{P} = \mathrm{diag}(p_1, ..., p_k)$ with i.i.d. p_i sampled uniformly from $\{-1, 1\}$, we have $\mathrm{Q}_1 P \overset{\mathrm{d}}{=} \mathrm{W}_{11}$. Let $\mathrm{W}_{11} = \mathrm{UCV}^\mathsf{T}$ be the SVD of W_{11}, then one has

$$\mathrm{U}_1 \begin{pmatrix} \tilde{\mathrm{C}} \\ 0 \end{pmatrix} \tilde{\mathrm{V}}^\mathsf{T} \mathrm{P} \overset{\mathrm{d}}{=} \mathrm{UCV}^\mathsf{T}.$$

Since $\mathrm{U}_1, \tilde{\mathrm{V}}$, and U, V are uniformly distributed and independent of $\tilde{\mathrm{C}}, \mathrm{C}$, we have $\tilde{\mathrm{C}} \overset{\mathrm{d}}{=} \mathrm{C}$ by the invariance of the Haar measure under orthogonal transformations. On the other hand, the generalized singular values $\tilde{\mathrm{C}}$ of a pair $(\mathrm{A}_1, \mathrm{A}_2)$ follow the law of the Jacobi ensemble with parameters $a = 0, b = d - 2k$, and $\beta = 1$ (Proposition 1.2 in [42]). Therefore, the squared singular values of W_{11} follow the Jacobi ensemble with the same parameters. □

Corollary A.5. The squared inner product $|\theta^\mathsf{T}\phi|^2$ between two independent random vectors $\theta, \phi \sim \mu_{\mathbb{S}^{d-1}}$ follows $\mathrm{Beta}\left(\frac{1}{2}, \frac{d-1}{2}\right)$. Moreover, the shifted inner product $(1 + \theta^\mathsf{T}\phi)/2$ is symmetrically distributed as $\mathrm{Beta}\left(\frac{d-1}{2}, \frac{d-1}{2}\right)$.

Proof of Corollary A.5. Setting Jacobi parameters $k = 1, a = 0, b = d - 2$ and $\beta = 1$, the density is proportional to $x^{-1/2}(1-x)^{(d-3)/2}$ on $[0, 1]$, which matches the $\mathrm{Beta}\left(\frac{1}{2}, \frac{d-1}{2}\right)$ distribution.

Next, observe that $\theta^\mathsf{T}\phi$ has a density proportional to $(1-t)^{\frac{d-3}{2}}$ for $t \in [-1, 1]$. Under the change of variables $\eta \sim \mathrm{Beta}\left(\frac{d-1}{2}, \frac{d-1}{2}\right)$.

□

B Complete proofs

Proof of Lemma 4.1. One can acquire $\mathsf{I}(X;Y) = -\frac{d}{2}\log(1-\rho^2)$ from a general expression for MI of two jointly Gaussian random vectors (see Corollary A.2).

Recall that $(\theta^\mathsf{T}X, \phi^\mathsf{T}Y)$ is also Gaussian with cross-covariance $\rho\,\theta^\mathsf{T}\phi$. Therefore, by Corollary A.2 we have

$$\mathsf{SI}(X;Y) = \mathbb{E}\left[\mathsf{I}(\theta^\mathsf{T}X; \phi^\mathsf{T}Y) \,\middle|\, \theta, \varphi\right] = -\frac{1}{2}\,\mathbb{E}\left[\log(1 - \rho^2\ |\theta^\mathsf{T}\phi|^2)\right].$$

From Corollary A.5, we note that $|\theta^\mathsf{T}\phi|^2 \sim \mathrm{Beta}\left(\frac{1}{2}, \frac{d-1}{2}\right)$, so

$$\begin{aligned} \mathsf{SI}(X;Y) &= -\frac{1}{2\mathrm{B}\left(\frac{1}{2}, \frac{d-1}{2}\right)} \int_0^1 \log(1-\rho^2 x)(1-x)^{\frac{d-3}{2}} x^{-\frac{1}{2}}\,\mathrm{d}x \\ &= \frac{\rho^2}{2} \frac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{d-1}{2}\right)} \int_0^1 x^{\frac{1}{2}}(1-x)^{\frac{d-3}{2}}\ {}_2F_1\left(1, 1; 2; \rho^2 x\right)\mathrm{d}x, \end{aligned} \tag{6}$$

where the last equality follows from the identity $\log(1-z) = -z\ {}_2F_1(1, 1; 2; z)$ with hypergeometric function ${}_2F_1$. Appling Euler's integral transform ([43], Eq. (2.2.3)) gives

$$\begin{aligned} \mathsf{SI}(X;Y) &= \frac{\rho^2}{2d} \frac{\Gamma\left(\frac{d}{2}+1\right)}{\Gamma\left(\frac{3}{2}\right)\Gamma\left(\frac{d-2}{2}\right)} \int_0^1 x^{\frac{3}{2}-1}(1-x)^{\left(\frac{d}{2}+1\right)-\frac{3}{2}-1}\ {}_2F_1\left(1, 1; 2; \rho^2 x\right)\mathrm{d}x \\ &= \frac{\rho^2}{2d}\ {}_3F_2\left(1, 1, \frac{3}{2}; \frac{d}{2}+1, 2; \rho^2\right). \end{aligned}$$

Here ${}_3F_2$ denotes the generalized hypergeometric function.

Finally, we calculate the limit of $\mathsf{SI}(X;Y)$ as $\rho^2 \to 1$ using properties of beta-distribution. Denoting $\eta = (1 + \theta^\mathsf{T}\phi)/2 \sim \mathrm{Beta}\left(\frac{d-1}{2}, \frac{d-1}{2}\right)$ (see Corollary A.5), we get

$$\mathsf{SI}(X;Y) = -\log 2 - \mathbb{E}\log(1-\eta) = -\log 2 - \mathbb{E}\log\eta = \psi(d-1) - \psi\left(\frac{d-1}{2}\right) - \log 2,$$

where ψ is the digamma function. Using the bounds on digamma function [44]

$$\log\left(x+\frac{1}{2}\right) - \frac{1}{x} \le \psi(x) \le \log\left(x + e^{\psi(1)}\right) - \frac{1}{x},$$

we derive an upper bound on this expression:

$$\psi(d-1) - \psi\left(\frac{d-1}{2}\right) - \log 2 \le \frac{1}{d-1} + \log\left(1 + \frac{1+e^{\psi(1)}}{d}\right)$$

To simplify the bound, one can note that $1 + e^{\psi(0)} < 2$, $\log(1+x) < x$ and $\frac{1}{d} < \frac{1}{d-1}$.

□

Proof of Proposition 4.2.

Let $\mathrm{Q_X}, \mathrm{Q_Y} \sim \mu_{\mathrm{St}(k,d)}$. Then $\left[\mathrm{Q_X^\mathsf{T}} X, \mathrm{Q_Y^\mathsf{T}} Y\right] \sim \mathcal{N}(0, \Sigma)$, where Σ is a $2k \times 2k$ covariance matrix with the following block structure

$$\Sigma = \begin{pmatrix} \mathrm{I}_k & \rho\, \mathrm{Q_X^\mathsf{T}} \mathrm{Q_Y} \\ \rho\, \mathrm{Q_Y^\mathsf{T}} \mathrm{Q_X} & \mathrm{I}_k \end{pmatrix}.$$

Using the formula for the determinant of a block matrix Σ yields

$$\mathsf{SI}_k(X;Y) = -\frac{1}{2}\,\mathbb{E}[\log\det(\Sigma)] = -\frac{1}{2}\,\mathbb{E}\left[\log\det\left(\mathrm{I} - \rho^2 \left(\mathrm{Q_X^\mathsf{T}} \mathrm{Q_Y}\right)\left(\mathrm{Q_X^\mathsf{T}} \mathrm{Q_Y}\right)^\mathsf{T}\right)\right].$$

By the invariance of the Haar measure under left and right multiplication, $\mathrm{Q_X^\mathsf{T}} \mathrm{Q_Y} \overset{d}{=} \mathrm{W}_{11}$, where W_{11} is a k by k left upper block of the matrix $\mathrm{W} \sim \mu_{\mathrm{O}(d)}$. According to Lemma A.4, the eigenvalues of $\mathrm{W}_{11}\mathrm{W}_{11}^\mathsf{T}$ follow Jacobi ensemble with parameters $a = 0, b = d - 2k$ and $\beta = 1$:

$$p(\lambda) \propto \prod_{i<j} |\lambda_j - \lambda_i| \prod_{i=1}^{k} (1-\lambda_i)^{\frac{d-2k-1}{2}}.$$

Thus, we get a general expresion for k-SMI

$$\mathsf{SI}_k(X;Y) = -\frac{1}{2}\int_{[0,1]^k} \sum_{i=1}^{k} \log(1-\rho^2\lambda_i) p(\lambda)\, \mathrm{d}\lambda.$$

□

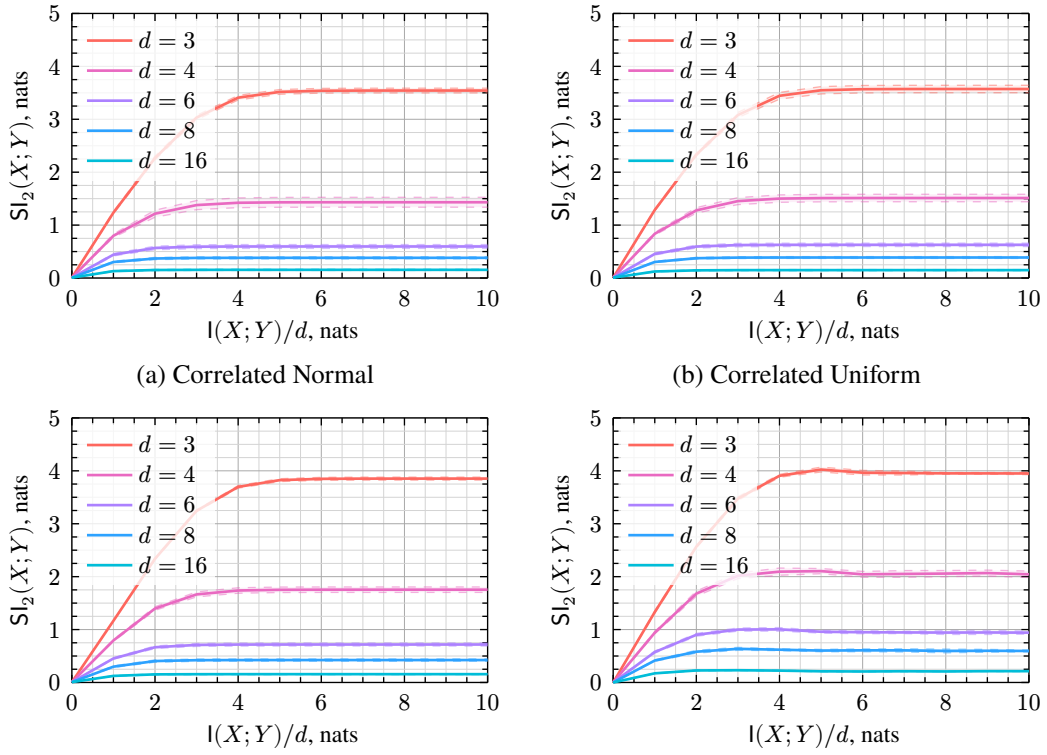
Proof of Proposition 4.4. Using Lemma A.3 and $d_x, d_y < k$, we get that $\Theta^\mathsf{T}\mathrm{A}$ and $\Phi^\mathsf{T}\mathrm{B}$ are injective with probability one for independent Θ, Φ distributed uniformly on $\mathrm{St}(d_x, k)$ and $\mathrm{St}(d_y, k)$. Therefore, according to Theorem 3.1, $\left[\mathsf{I}(\Theta^\mathsf{T}\mathrm{A}X; \Phi^\mathsf{T}\mathrm{B}Y) \,\middle|\, \Theta, \Phi\right] = \mathsf{I}(X;Y)$ almost sure. As a result, $\mathsf{SI}_k(\mathrm{A}X; \mathrm{B}Y) = \mathsf{I}(\Theta^\mathsf{T}\mathrm{A}X; \Phi^\mathsf{T}\mathrm{B}Y \,|\, \Theta, \Phi) = \mathsf{I}(X;Y)$. □

Proof of Proposition 4.7. Direct corollary of Corollary A.2. □

C Additional experiments

In this section, we conduct supplementary experiments to evaluate SMI under a broader range of setups. We begin by assessing k-SMI on the same set of benchmarks from Section 5. The results for $k = 1, 2, 3$ are presented in Figure 3, Figure 6, and Figure 7, respectively. Notably, saturation remains consistent even for $k = d - 1$ (i.e., when only one component is discarded).

Next, we examine a setup involving randomized distribution parameters, following the methodology of [34]. Among other adjustments, this includes randomizing per-component mutual information (e.g., assigning interactions unevenly in this experiment). In some cases (e.g., the log-gamma-



(a) Correlated Normal

(b) Correlated Uniform

(c) Smoothed Uniform

(d) Log-Gamma-Exponential

Figure 6: Results of synthetic experiments with different distributions for 2-SMI. We report mean values and standard deviations computed across 10 runs, with 10^4 samples used for MI estimation and 128 for averaging across projections.

exponential distribution), this increases linear redundancy, as component pairs with higher mutual information also exhibit higher variance in this particular scenario. Our results are displayed in Figure 8.

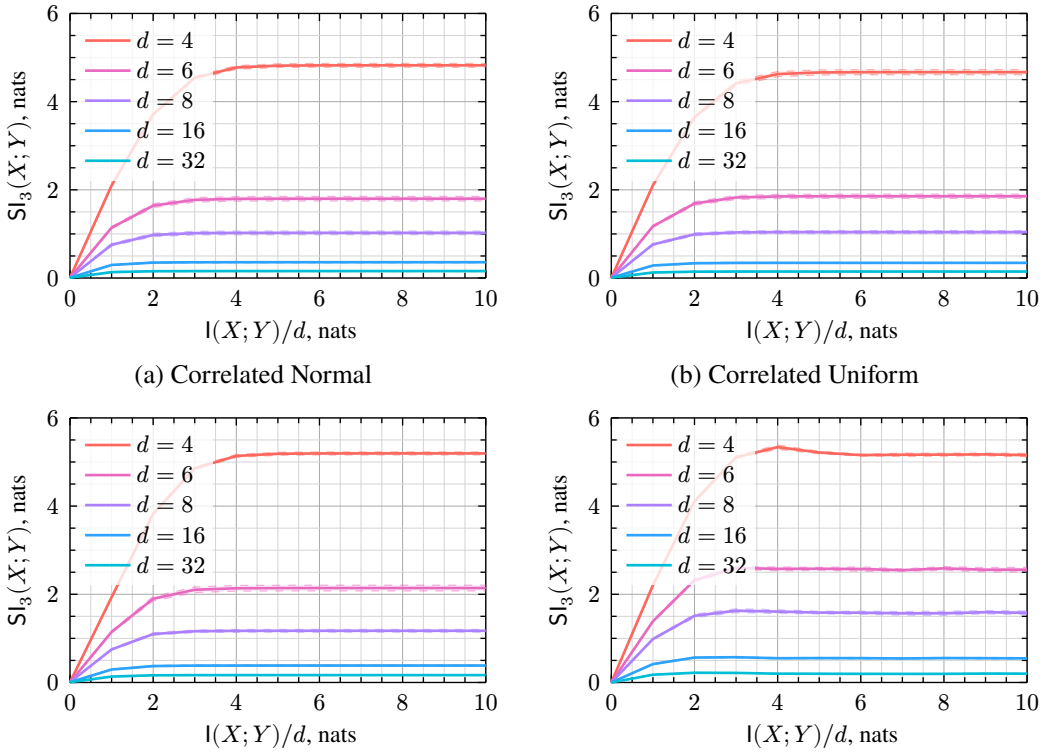
Due to numerical constraints, we do not track $\mathsf{I}(X;Y)/d$, instead plotting the results against the total mutual information. While this makes saturation slightly less evident, the general trend of SMI decreasing with d remains observable. We also highlight the log-gamma-exponential distribution (Figure 8d), where SMI is less prone to saturation under parameter randomization due to the reasons mentioned earlier.

D Implementation details

D.1 Synthetic experiments

For the experiments from Section 5 and Section 6.1, we use implementation of Kraskov-Stoegbauer-Grassberger (KSG) [35] mutual information estimator and random slicing from [34]. The number of neighbors is set to $k_{\text{NN}} = 1$ for the KSG estimator. For each configuration, we conduct 10 independent runs with different random seeds to compute means and standard deviations. Our experiments use 10^4 samples for (X, Y) and 128 samples for (Θ, Φ).

For the experiments from Section 5, we use independent components with equally distributed per-component MI. For the supplementary experiments from Figure 8, parameters of each distribution (e.g., covariance matrices) are randomized via the algorithm implemented in [34]. This includes randomization of per-component MI (which is done using a uniform distribution over a $(d-1)$-dimensional simplex).



(a) Correlated Normal

(b) Correlated Uniform

(c) Smoothed Uniform

(d) Log-Gamma-Exponential

Figure 7: Results of synthetic experiments with different distributions for 3-SMI. We report mean values and standard deviations computed across 10 runs, with 10^4 samples used for MI estimation and 128 for averaging across projections.

For the experiments, we used AMD EPYC 7543 CPU, one core per distribution. Each experiment (fixed k, varying d) took no longer then 3 days to compute.

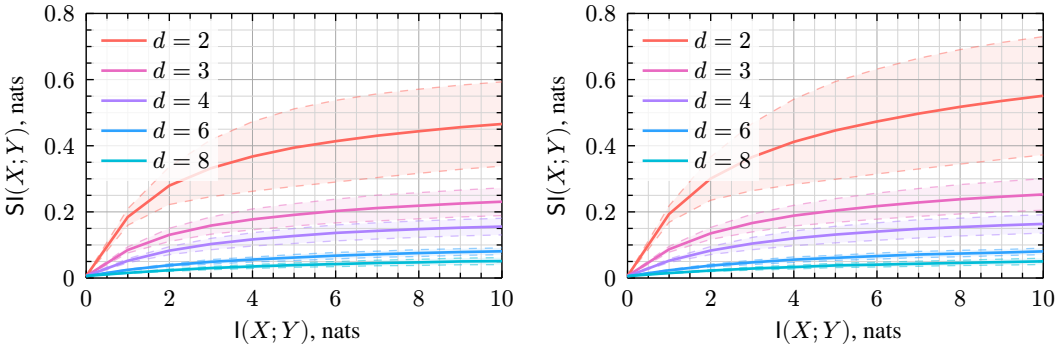
D.2 Representation learning experiments

For experiments on MNIST dataset, we use a simple ConvNet with three convolutional and two fully connected layers. A three-layer fully-connected perceptron serves as a critic network for the InfoNCE loss. We provide the details in Table 2. We use additive Gaussian noise with $\sigma = 0.2$ as an input augmentation. Training hyperparameters are as follows: batch size = 512, 2000 epochs, Adam optimizer [45] with learning rate 10^{-3}.

For the experiments, we used Nvidia A100 GPUs. Each experiment took no longer then 1 day to compute.

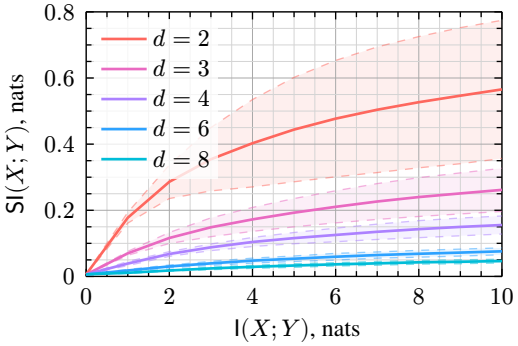
Table 2: The NN architectures used to conduct the tests on MNIST images in Section 6.2.

NN	Architecture
ConvNet, 24×24 images	$\times$ 1: Conv2d(1, 32, ks=3), MaxPool2d(2), BatchNorm2d, LeakyReLU(0.01) $\times$ 1: Conv2d(32, 64, ks=3), MaxPool2d(2), BatchNorm2d, LeakyReLU(0.01) $\times$ 1: Conv2d(64, 128, ks=3), MaxPool2d(2), BatchNorm2d, LeakyReLU(0.01) $\times$ 1: Dense(128, 128), LeakyReLU(0.01), Dense(128, dim)
Critic NN, pairs of vectors	$\times$ 1: Dense(dim + dim, 256), LeakyReLU(0.01) $\times$ 1: Dense(256, 256), LeakyReLU(0.01), Dense(256, 1)

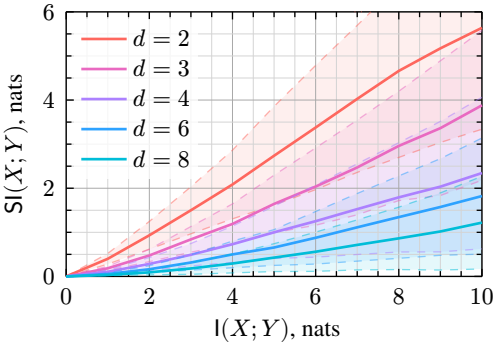


(a) Correlated Normal

(b) Correlated Uniform



(c) Smoothed Uniform



(d) Log-Gamma-Exponential

Figure 8: Results of synthetic experiments with different distributions. We report mean values and standard deviations computed across 10 runs, with 10^4 samples used for MI estimation and 128 for averaging across projections.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [YES]

   Justification: We state our claims clearly in the abstract and introduction. The claims are supported by theoretical analysis and various experiments.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [YES]

   Justification: We discuss limitations in Section 7.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [YES]

Justification: We provide comprehensive statements for theorems and lemmas. We also provide complete proofs in Section B.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [YES]

Justification: We provide complete setup descriptions for the experiments in corresponding sections.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [YES]

Justification: We use openly accessible data only for our experiments. Community-provided code has been used for our experiments, and we reference its origin. Finally, we include additional source code in the submission.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [YES]

Justification: We provide all the necessary details in the corresponding sections, or in Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [YES]

Justification: We report mean and standard deviation.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [YES]

Justification: We describe our setup and computational load of the experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines

Answer: [YES]

Justification:

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: Although the paper may question methodology of some other works, there are no broader impacts of the research conducted.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [YES]

    Justification: For all the assets, we use citations that the original authors provided.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `https://paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.