Inductive Mutual Information Estimation: A Convex Maximum-Entropy Copula Approach

Yves-Laurent Kom Samo
KXY Technologies, Inc.
✉ yl@kxy.ai ⧈ @Dr_YLKS 🐦 @Dr_YLKS
San Jose, California, USA

Abstract

We propose a novel estimator of the mutual information between two ordinal vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. Our approach is inductive (as opposed to deductive) in that it depends on the data generating distribution solely through some nonparametric properties revealing associations in the data, and does not require having enough data to fully characterize the true joint distributions $P_{\boldsymbol{x},\boldsymbol{y}}$. Specifically, our approach consists of (i) noting that $I(\boldsymbol{y};\boldsymbol{x}) = I(\boldsymbol{u}_y;\boldsymbol{u}_x)$ where $\boldsymbol{u}_y$ and $\boldsymbol{u}_x$ are the *copula-uniform dual representations* of $\boldsymbol{y}$ and $\boldsymbol{x}$ (i.e. their images under the probability integral transform), and (ii) estimating the copula entropies $h(\boldsymbol{u}_y)$, $h(\boldsymbol{u}_x)$ and $h(\boldsymbol{u}_y,\boldsymbol{u}_x)$ by solving a maximum-entropy problem over the space of copula densities under a constraint of the type $\boldsymbol{\alpha}_m = E[\phi_m(\boldsymbol{u}_y,\boldsymbol{u}_x)]$. We prove that, so long as the constraint is feasible, this problem admits a unique solution, it is in the exponential family, and it can be learned by solving a convex optimization problem. The resulting estimator, which we denote MIND, is marginal-invariant, always non-negative, unbounded for any sample size n, consistent, has MSE rate $O(1/n)$, and is more data-efficient than competing approaches.

1 Introduction

Mutual information plays a key role in statistical learning. It is directly related to the highest R^2, the highest true log-likelihood per observation, the lowest root mean square error and the highest classification accuracy that can be achieved by using explanatory variables $\boldsymbol{x}$ to predict categorical or continuous output(s) $\boldsymbol{y}$. It also plays an important role in representation learning (Brown et al. (1992); Bell and Sejnowski (1995); Tishby et al. (2000); Tishby and Zaslavsky (2015); Chen et al. (2016); Higgins et al. (2018)) and reinforcement learning (Pathak et al. (2017); Oord et al. (2018)).

Virtually every mutual information estimator in the litterature implicitly assumes that we have a number n of i.i.d. samples $(\boldsymbol{x}_1,\boldsymbol{y}_1),\ldots,(\boldsymbol{x}_n,\boldsymbol{y}_n)$ that is large enough to characterize the underlying distribution $P_{\boldsymbol{x},\boldsymbol{y}}$. We will refer to this scenario as the *deductive approach.* Examples include quantizing (Paninski (2003)) or hashing (Noshad et al. (2019)) $\boldsymbol{x}$ and $\boldsymbol{y}$ and computing the mutual information between the resulting discrete distributions using sample frequencies. Other approaches approximate the pdfs using kernel density estimators (Moon et al. (1995); Kwak and Choi (2002)), using local geometric properties based on k nearest neighbors (Kraskov et al. (2004); Gao et al. (2015)) and using Edgeworth approximation (Hulle (2005)). Another perspective has been to learn lower bounds based on variational characterizations of the mutual information (Nguyen et al. (2010); Belghazi et al. (2018)) using M-estimators.

The *deductive approach* is fraught with limitations. When relying on discrete approximations, the mutual information can never be greater than the mutual information between fully dependent uniform distributions, namely $\log n$ in the case of quantization (Paninski (2003)) and $\log F$ in the case of hashing (Noshad et al. (2019)), where $F < n$ is the number of distinct hashes and n the sample size. McAllester and Stratos (2020) extended this result empirically to various continuous mutual information estimators relying on the deductive approach, including variational estimators (Nguyen et al. (2010); Belghazi et al. (2018)). Their work touches on the core issue: if we require properly

characterizing the joint pdf nonparametrically from n i.i.d. samples in order to estimate a mutual information, then n ought to be large, otherwise we will not see enough tail events, and we will fail to account for tail dependency. Unfortunately, the alternative proposed by McAllester and Stratos (2020), namely approximating the density $p(\boldsymbol{x})$ and $p(\boldsymbol{x}|\boldsymbol{y})$ parametrically using deep neural networks is data inefficient and tends to overshoot on small sample sizes. Along the same line, Song and Ermon (2020) and Poole et al. (2019) reported that the variance of variational estimators MINE (Belghazi et al. (2018)) and NWJ (Nguyen et al. (2010)) could be very large, and even grow exponentially with the true mutual information (Song and Ermon (2020)), due to the need for a large number n of samples to accurately estimate an expectation of the form $E\left[e^{T(\boldsymbol{x}, \boldsymbol{y})}\right]$ under $P_{\boldsymbol{x}} \otimes P_{\boldsymbol{y}}$.

Another major limitation of the deductive approach is the unnecessary need to accurately model the marginals of $P_{\boldsymbol{x}, \boldsymbol{y}}$, directly or implicitly, as a prerequisite for estimating the mutual information, which could be data and compute intensive, even though the mutual information does not depend on marginal distributions.

The *inductive approach* we introduce in this paper is structured in two stages. First, we measure a few nonparametric properties of the data generating distribution that serve as marginal-invariant proxies revealing associations between coordinates of $\boldsymbol{x}$ and/or $\boldsymbol{y}$. Then we estimate the mutual information in the spirit of the maximum-entropy principle (Jaynes (1957a;b)), by being consistent with all observed properties, while remaining as uninformative as possible about any property we haven't observed. Intuitively, we would expect that the more expressive the properties we measure get, the closer we should get to the true mutual information. Indeed, we propose a family of nonparametric properties that give rise to a consistent estimator of the true mutual information.

The rest of the paper is structured as follows. In Section 2 we recall some results relating copulas and mutual information. In Section 3 we further motivate our inductive approach and we present core theoretical results pertaining to maximum-entropy inference of copulas. Our theoretical contribution builds on the study of the I-divergence geometry of probability distributions developed by Csiszár (1975), which generalizes the minimum discrimination information theorem of Kullback et al. (1966). In Section 4 we propose a convex pro gram for solving maximum-entropy copula problems under linear constraints, and we discuss practical considerations. Finally, in Section 5 we illustrate that our estimator outperforms the state of the art on large and small mutual information problems on synthetic data, we illustrate that our work can be used to mitigate mode collapse in GANs, and we apply our approach to the estimation of the highest performance achievable in a Kaggle competition.

2 Background

We begin by recalling that the *mutual information* between two random vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as

$$I(\boldsymbol{y}; \boldsymbol{x}) := \int_{\mathcal{X} \times \mathcal{Y}} \log \frac{dP_{\boldsymbol{x}, \boldsymbol{y}}}{dP_{\boldsymbol{x}} \otimes P_{\boldsymbol{y}}} dP_{\boldsymbol{x}, \boldsymbol{y}}$$

where $P_{\boldsymbol{x}, \boldsymbol{y}}$ (resp. $P_{\boldsymbol{x}}$, $P_{\boldsymbol{y}}$) is the (joint) probability measure of $(\boldsymbol{x}, \boldsymbol{y})$ (resp. $\boldsymbol{x}$, $\boldsymbol{y}$), and $\frac{dP_{\boldsymbol{x}, \boldsymbol{y}}}{dP_{\boldsymbol{x}} \otimes P_{\boldsymbol{y}}}$ is the Radon-Nikodym derivative of the joint probability measure with respect to the product measure of $P_{\boldsymbol{x}}$ and $P_{\boldsymbol{y}}$. Friendlier expressions depending on whether the random vectors have continuous and/or categorical coordinates are provided in Table 2 in the appendix.

While the mutual information is the canonical approach for *quantifying* associations betweeen random variables, copulas are the canonical tool for *modeling* associations between random variables. We recall some basic definitions and properties, and link the two notions.

Definition 2.1. A **copula distribution** is any probability distribution supported on $[0, 1]^d$ whose marginals are uniform. A **copula** (resp. **copula density**) is any function that is the cdf (resp. pdf) of a copula distribution.

The following theorem shows that every distribution with pdf is uniquely associated to a copula density that fully captures its dependence structure, independently from marginals.

Theorem 2.1. *(Sklar's Theorem) Any pdf $f : \mathcal{Z} \subset \mathbb{R}^d \to \mathbb{R}^+$ whose marginal pdfs are $f_1(z_1), \ldots, f_d(z_d)$ with associated cdfs $F_1(z_1), \ldots, F_d(z_d)$ can be uniquely decomposed as*

$$f(z_1, \ldots, z_d) = c(F_1(z_1), \ldots, F_d(z_d)) \prod_{i=1}^{d} f_i(z_i),$$

*where $c : [0, 1]^d \to \mathbb{R}^+$ is a copula density. We refer to $\boldsymbol{u}_z := (F_1(z_1), \ldots, F_d(z_d))$ as the **copula-uniform dual representation** of $\boldsymbol{z} := (z_1, \ldots, z_d)$, and to $\boldsymbol{z}$ as a **primal representation** of $\boldsymbol{u}_z$.*

Interestingly, the entropy of a copula-uniform dual representation is invariant by continuous 1-to-1 primal feature transformations.

Proposition 2.1. *If $\boldsymbol{u}_z$ is the copula-uniform dual representation of $\boldsymbol{z} := (z_1, \ldots, z_d)$ and $\boldsymbol{u}_{g(z)}$ is the copula-uniform dual representation of $g(\boldsymbol{z}) := (g_1(z_1), \ldots, g_d(z_d))$ where the functions $g_1, \ldots, g_d$ are continuous 1-to-1 functions, then*

$$h\left(\boldsymbol{u}_z\right) = h\left(\boldsymbol{u}_{g(\boldsymbol{z})}\right).$$

See Appendix B.1 for the proof.

The following entropy decomposition is a direct consequence of Sklar's theorem.

Proposition 2.2. *If $\boldsymbol{u}_z$ is the copula-uniform dual representation of $\boldsymbol{z} := (z_1, \ldots, z_d) \in \mathcal{Z} \subset \mathbb{R}^d$ and $\boldsymbol{z}$ admits a pdf, then the differential entropy of $\boldsymbol{z}$ can be decomposed as*

$$h\left(\boldsymbol{z}\right) = h\left(\boldsymbol{u}_z\right) + \sum_{i=1}^{d} h\left(z_i\right), \qquad (1)$$

*so long as all marginal entropies exist. $h\left(\boldsymbol{u}_z\right)$ is the entropy of the associated copula distribution, and we refer to it as the **copula entropy** of $\boldsymbol{z}$.*

This entropy decomposition implies that the mutual information between two continuous random vectors is the same as that of their copula-uniform dual representations:

$$\begin{aligned} I\left(\boldsymbol{y}; \boldsymbol{x}\right) &= h\left(\boldsymbol{y}\right) + h\left(\boldsymbol{x}\right) - h\left(\boldsymbol{y}, \boldsymbol{x}\right) \\ &= h\left(\boldsymbol{u}_y\right) + h\left(\boldsymbol{u}_x\right) - h\left(\boldsymbol{u}_y, \boldsymbol{u}_x\right) \\ &= I\left(\boldsymbol{u}_y; \boldsymbol{u}_x\right). \end{aligned}$$

The identity $I\left(\boldsymbol{y}; \boldsymbol{x}\right) = I\left(\boldsymbol{u}_y; \boldsymbol{u}_x\right)$ extends to all ordinal random vectors by noting that copula-uniform dual representations are well defined for ordinal random vectors, are in a 1-to-1 relationship with their primal representations, and that the mutual information is invariant by 1-to-1 maps. In general, we may use Table 2 to conclude that estimating any mutual information boils down to estimating copula entropies, and possibly one-dimensional primal entropies (when the problem involves categorical and non-ordinal coordinates that we choose not to ordinally encode).

Another important property of the mutual information we will rely on is that it is stable by addition of redundant information.

Proposition 2.3. *Let $\boldsymbol{y} \in \mathcal{Y}$ and $\boldsymbol{x} \in \mathcal{X}$ be two random vectors, and f a function defined on $\mathcal{X}$. Then we have:*

$$I\left(\boldsymbol{y}; \boldsymbol{x}, f\left(\boldsymbol{x}\right)\right) = I\left(\boldsymbol{y}; \boldsymbol{x}\right) + \underbrace{I\left(\boldsymbol{y}; f\left(\boldsymbol{x}\right) | \boldsymbol{x}\right)}_{=0} = I\left(\boldsymbol{y}; \boldsymbol{x}\right).$$

Going forward, and without loss of generality, we will focus on the estimation of the copula entropy of a continuous random vector $h\left(\boldsymbol{u}_z\right)$, with the understanding that the mutual information is recovered as $I\left(\boldsymbol{y}; \boldsymbol{x}\right) = h\left(\boldsymbol{u}_y\right) + h\left(\boldsymbol{u}_x\right) - h\left(\boldsymbol{u}_y, \boldsymbol{u}_x\right)$.

3 Inductive Mutual Information Estimation

We consider estimating the copula entropy $h\left(\boldsymbol{u}_z\right)$ of a continuous random vector $\boldsymbol{z} \in \mathcal{Z} \subset \mathbb{R}^d$, where $\boldsymbol{u}_z$ is the copula-uniform dual representation of $\boldsymbol{z}$, which we assume admits a pdf.

3.1 Motivation and Roadmap

The *deductive approach* to learning the copula entropy $h\left(\boldsymbol{u}_z\right)$ requires assuming that we have gathered enough samples to fully characterize the pdf $p(\boldsymbol{u}_z)$, estimating the pdf, and then estimating the copula entropy as the entropy of the estimated pdf. This is both data and compute inefficient. If we partition $[0, 1]^d$ into small hypercubes of side length δ, then we need to observe at least one sample per hypercube for a small enough δ to properly characterize the pdf non-parametrically. This requires $n \propto \delta^{-d}$ samples. How small δ needs to be depends on how quickly the true pdf varies on $[0, 1]^d$. Regardless, n would grow exponentially with the input dimension, and so would the associated compute requirement.

This inefficiency can be alleviated by assuming that the pdf belongs to a specific parametric family, at the expense of model misspecification. If the parametric family has sufficient statistics $T\left(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\right)$ for n i.i.d. observations, then pdfs in the family should be maximum-entropy among all pdfs with the same statistics. If this is not the case, then the parametric family would be violating Occam's razor as the learned pdf would be encoding more structure than evidenced by the data, and the estimated entropy would overshoot. We also note that the pdf $p(\boldsymbol{u}_z)$ should have uniform marginals, which makes finding an appropriate parametric family even more difficult.

In the absence of any empirical evidence, Occam's razor suggests that the most appropriate distribution for $\boldsymbol{u}_z$ is the uniform distribution on $[0, 1]^d$ as it is the least informative (or maximum-entropy) of all distributions supported on $[0, 1]^d$. Instead of choosing a rigid parametric family and inheriting its sufficient statistics, we could construct more expressive copula densities by first choosing how to reveal departure from the standard uniform distribution from the data, and then finding the least informative copula density among all copula densities satisfying the observed constraints.

The *inductive approach* we propose consists of revealing the dependence structure in $\boldsymbol{u}_z$ by estimating an

expectation of the form $\boldsymbol{\alpha}_m = E_{P_{\boldsymbol{u}_z}}[\phi_m(\boldsymbol{u}_z)]$, for a vector-valued statistics function $\phi_m : [0,1]^d \to \mathbb{R}^q$, and estimating $h(\boldsymbol{u}_z)$ as the highest copula entropy among all copulas satisfying the constraint $\boldsymbol{\alpha}_m = E_P[\phi_m(\boldsymbol{u}_z)]$. When ϕ_m is given and $E_{P_{\boldsymbol{u}_z}}[\phi_m(\boldsymbol{u}_z)]$ is all the data scientist can reliably observe about the structure of the data (e.g. we are only given pairwise Spearman rank correlations), our estimator is the only estimator consistent with Occam's razor. Crucially, if we may choose ϕ_m, then we may approximate the copula entropy with arbitrary precision. Specifically, we show that, so long as $(\phi_m)_m$ are universal approximators of continuous functions on $[0,1]^d$, the solution to this maximum-entropy problem is a consistent estimator of the true copula entropy $h(\boldsymbol{u}_z)$ (see Theorem 4.1). Equally important is Corollary 3.1 that states that we may perfectly recover the true mutual information using a finite dimensional statistics function that is not expressive enough to fully characterize the true copula distribution $P_{\boldsymbol{u}_{x,y}}$.

We note that $\boldsymbol{\alpha}_m$ can be efficiently estimated from n i.i.d. primal samples $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ as

$$\hat{\boldsymbol{\alpha}}_{m,n} = \frac{1}{n}\sum_{i=1}^{n} \phi_m\left(\frac{\mathrm{rg}(\boldsymbol{z}_i)}{n+1}\right), \qquad (2)$$

where $\mathrm{rg}(\boldsymbol{z}_i)$ is the vector of coordinatewise ranks of $\boldsymbol{z}_i$ among $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$. It follows from the weak convergence of the empirical copula process to the true copula that $\hat{\boldsymbol{\alpha}}_{m,n}$ is a consistent and asymptotically normal estimator of $\boldsymbol{\alpha}_m$ (Ruschendorf (1976)). Thus, our *inductive approach* to mutual information estimation truly does not require learning marginal distributions. We show that, so long as the constraint $E_P[\phi_m(\boldsymbol{u}_z)] = \hat{\boldsymbol{\alpha}}_{m,n}$ is feasible, the associated maximum-entropy problem admits a unique solution, and it is a consistent estimator[1] of $h(\boldsymbol{u}_z)$ when $(\phi_m)_m$ are universal approximators of continuous functions on $[0,1]^d$. Finally, we introduce a convex optimization problem whose minimizer is the maximizer of our maximum-entropy problem.

3.2 Maximum-Entropy Copulas

Let $\mathcal{C}_d$ be the space of all d-dimensional copula distributions with pdf, $h(P)$ the differential entropy of the probability distribution P, and $\phi_{m,d} : [0,1]^d \to \mathbb{R}^{q(d)}$ a vector-valued function whose coordinate functions are not linearly dependent.

We define the following properties of $\phi_{m,d}$: (P1) the first coordinate of $\phi_{m,d}$ is the constant 1, (P2) each coordinate of $E_P[\phi_{m,d}(\boldsymbol{u})]$ captures a way in which P departs from the uniform distribution, (P3) the fam-

---

[1]Consistency here is jointly in n and m.

ily $(\phi_{m,d})_m$ is a universal approximator of continuous functions defined on $[0,1]^d$ for any d, and (P4) $q(l) + q(k) \le q(l+k)$ for every $l, k > 0$ and all coordinates of $\phi_{m,l}$ and $\phi_{m,k}$ are also included in $\phi_{m,l+k}$. Going forward, we will use ϕ_m and q in-lieu-of $\phi_{m,d}$ and $q(d)$ for ease of notation when the input dimension is unambiguous.

We consider the following optimization problem:

$$\begin{cases} \max\limits_{P \in \mathcal{C}_d} \quad h(P) \\ \text{s.t. } E_P[\phi_m(\boldsymbol{u})] = \boldsymbol{\alpha}_m \end{cases} . \qquad \text{(MIND)}$$

The theorem below, which we prove in Appendix B.2, states that the solution of (MIND) is an exponential family distribution with sufficient statistics ϕ_m, and with base measure the product of d measures on $[0,1]$ that are absolutely continuous with respect to the standard uniform on $[0,1]$.

Theorem 3.1. *Let ϕ_m satisfy (P1). If there is any copula distribution P with finite differential entropy and satisfying $E_P[\phi_m(\boldsymbol{u})] = \boldsymbol{\alpha}_m$, then the maximum-entropy problem (MIND) admits a unique solution, and the maximizer P_M is the only copula distribution whose density takes the form*

$$p_M(\boldsymbol{u}; \phi_m, \boldsymbol{\alpha}_m) = e^{\boldsymbol{\theta}^T \phi_m(\boldsymbol{u})} \prod_{i=1}^{d} f_i(u_i), \qquad (3)$$

and that satisfies the constraint $E_{P_M}[\phi_m(\boldsymbol{u})] = \boldsymbol{\alpha}_m$ for some constant $\boldsymbol{\theta}$, and d non-negative univariate functions f_i, $\log f_i \in L_1([0,1])$. Moreover, for any copula distribution $P \in \mathcal{C}_d$ satisfying $E_P[\phi_m(\boldsymbol{u})] = \boldsymbol{\alpha}_m$,

$$h(P_M) - h(P) = KL(P||P_M). \qquad (4)$$

The practical challenge with applying Theorem 3.1 is the need to learn the free functions $f_1, \ldots, f_d$. These functions ensure that the maximizer is a copula distribution—i.e. has uniform marginals. We now consider relaxing this requirement and solving the maximum-entropy problem over the space $\mathcal{D}_d \supset \mathcal{C}_d$ of all continuous probability distributions supported on $[0,1]^d$:

$$\begin{cases} \max\limits_{P \in \mathcal{D}_d} \quad h(P) \\ \text{s.t. } E_P[\phi_m(\boldsymbol{u})] = \boldsymbol{\alpha}_m \end{cases} . \qquad \text{(A-MIND)}$$

As we later show in Theorem 4.1, (P3) guarantees that, despite this relaxation, the solution to (A-MIND) converges to the true copula entropy as m goes to infinity. For a given m, to control how close to uniform the maximizer's marginals are, we use the moment characterization of the standard uniform $\forall j$, $E(u^j) = 1/(1+j)$,

and we match the first k moments of marginals of candidate distributions to those of the standard uniform. To do so, we introduce the property (P5): ϕ_m has the form

$$\phi_m^k(\boldsymbol{u}) = \left(1, u_1, \ldots, u_1^k, \ldots, u_d, \ldots, u_d^k, \psi_m(\boldsymbol{u})\right),$$

with associated

$$\boldsymbol{\alpha}_m^k = (1, 1/2, \ldots, 1/(1+k), \ldots, 1/2, \ldots, 1/(1+k), \boldsymbol{\beta}_m).$$

The solution of (A-MIND) is provided by the following theorem, which we prove in Appendix B.3.

Theorem 3.2. *Let ϕ_m^k satisfy (P5). If there is any distribution P supported on $[0,1]^d$, with finite differential entropy and satisfying $E_P\left[\phi_m^k(\boldsymbol{u})\right] = \boldsymbol{\alpha}_m^k$, then the maximum-entropy problem (A-MIND) admits a unique solution of the form*

$$h_{AM}\left(\boldsymbol{u}; \phi_m^k, \boldsymbol{\beta}_m\right) = -\boldsymbol{\theta}^T \boldsymbol{\alpha}_m^k, \tag{5}$$

and the maximizer P_{AM} is the only distribution supported on $[0,1]^d$, whose pdf takes the form

$$p_{AM}\left(\boldsymbol{u}; \phi_m^k, \boldsymbol{\beta}_m\right) = e^{\boldsymbol{\theta}^T \phi_m^k(\boldsymbol{u})} \tag{6}$$

for some constant $\boldsymbol{\theta}$, and that satisfies the constraint $E_{P_{AM}}\left[\phi_m^k(\boldsymbol{u})\right] = \boldsymbol{\alpha}_m^k$. Moreover, for any distribution $P \in \mathcal{D}_d$ satisfying $E_P\left[\phi_m^k(\boldsymbol{u})\right] = \boldsymbol{\alpha}_m^k$,

$$h(P_{AM}) - h(P) = KL(P||P_{AM}). \tag{7}$$

Corollary 3.1. *Let $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y} \in \mathcal{Y}$ be two continuous random variables with mutual information $I(\boldsymbol{y}; \boldsymbol{x})$ and true individual and joint copula distributions $P_{\boldsymbol{u}_x}$, $P_{\boldsymbol{u}_y}$, and $P_{\boldsymbol{u}_{x,y}}$. If $P_{AM}(\boldsymbol{u}_x)$, $P_{AM}(\boldsymbol{u}_y)$ and $P_{AM}(\boldsymbol{u}_x, \boldsymbol{u}_y)$ are the solutions to three (A-MIND) problems whose constraints are satisfied by the true copula distributions, and*

$$\begin{aligned} I_{AM}(\boldsymbol{y}; \boldsymbol{x}) :=\ & h\left(P_{AM}(\boldsymbol{u}_x)\right) + h\left(P_{AM}(\boldsymbol{u}_y)\right) \\ & - h\left(P_{AM}(\boldsymbol{u}_x, \boldsymbol{u}_y)\right) \end{aligned}$$

is the associated mutual information estimator, then

$$I_{AM}(\boldsymbol{y}; \boldsymbol{x}) - I(\boldsymbol{y}; \boldsymbol{x}) = KL\left[P_{\boldsymbol{u}_x}||P_{AM}(\boldsymbol{u}_x)\right] \tag{8}$$
$$+ KL\left[P_{\boldsymbol{u}_y}||P_{AM}(\boldsymbol{u}_y)\right] - KL\left[P_{\boldsymbol{u}_{x,y}}||P_{AM}(\boldsymbol{u}_{x,y})\right].$$

Equation (8) in Corollary 3.1 lays out the theoretical ground for favoring our *inductive approach* over the traditional *deductive approach*. Indeed, it shows that it is not necessary to accurately learn the true data generating distributions $P_{\boldsymbol{x}}$, $P_{\boldsymbol{y}}$, and $P_{\boldsymbol{x}, \boldsymbol{y}}$ or their copulas in order to accurately learn the mutual information $I(\boldsymbol{y}; \boldsymbol{x})$. The error made by (A-MIND) in estimating the joint copula entropy $h(\boldsymbol{u}_y, \boldsymbol{u}_x)$ can offset the errors made estimating the individual copula entropies $h(\boldsymbol{u}_y)$ and $h(\boldsymbol{u}_x)$, so that we may perfectly estimate the mutual information with a finite dimensional statistics function ϕ_m, without accurately learning the copula distributions.

3.3 Iterative MIND

When either (MIND) or (A-MIND) are sparse in the sense that each coordinate of ϕ_m^k is a function of some but not all coordinates of $\boldsymbol{u}$, they can be broken down into smaller, cheaper, more robust and cacheable problems.

Theorem 3.3. *Let us assume that $\psi_m(\boldsymbol{u})$ takes the form $\psi_m(\boldsymbol{u}) = (\eta_1(w, \boldsymbol{v}_1), \ldots, \eta_q(w, \boldsymbol{v}_q))$ where w is a coordinate of $\boldsymbol{u}$ and the vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_q$ are made of coordinates of $\boldsymbol{u}$ but share no common coordinate and do not include w. If the constraint $E_P[\psi_m(\boldsymbol{u})] = \boldsymbol{\beta}_m := (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q)$ is feasible, then we have*

$$p_M\left(\boldsymbol{u}; \phi_m^k, \boldsymbol{\alpha}_m^k\right) = \prod_{i=1}^q p_M(w, \boldsymbol{v}_i; \eta_i, \boldsymbol{\beta}_i), \tag{9}$$

$$h_M\left(\boldsymbol{u}; \phi_m^k, \boldsymbol{\alpha}_m^k\right) = \sum_{i=1}^q h_M(w, \boldsymbol{v}_i; \eta_i, \boldsymbol{\beta}_i). \tag{10}$$

The proof is provided in Appendix B.4. Theorem 3.3 is useful in single-output problems when blocks of explanatory variables are known or assumed to be independent conditional on the output.

More generally, when each coordinate of ϕ_m^k depends on some but not all coordinates of $\boldsymbol{u}$, the problem (A-MIND) can be broken down into smaller problems that can be solved iteratively.

Theorem 3.4. *Let $\boldsymbol{u} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_q)$ and let us assume that $\psi_m(\boldsymbol{u})$ takes the form $\psi_m(\boldsymbol{u}) = (\eta_1(\boldsymbol{v}_1), \ldots, \eta_q(\boldsymbol{v}_q), \gamma(\boldsymbol{u}))$ where all η_i and γ satisfy (P1). If the constraint $E_P[\psi_m(\boldsymbol{u})] = \boldsymbol{\beta}_m := (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q, \bar{\boldsymbol{\beta}})$ is feasible, then the maximizer of (A-MIND) has density of the form*

$$p_{AM}\left(\boldsymbol{u}; \phi_m^k, \boldsymbol{\beta}_m\right) = e^{\boldsymbol{\theta}^T \phi_m^k(\boldsymbol{u})} \prod_{i=1}^q p_{AM}(\boldsymbol{v}_i; \eta_i, \boldsymbol{\beta}_i) \tag{11}$$

where $\boldsymbol{\theta}$ is the only constant such that $E_{P_{AM}}\left[\phi_m^k(\boldsymbol{u})\right] = \boldsymbol{\alpha}_m^k$.

Moreover, the solution reads

$$h_{AM}\left(\boldsymbol{u}; \phi_m^k, \boldsymbol{\beta}_m\right) = -\boldsymbol{\theta}^T \boldsymbol{\alpha}_m^k + \sum_{i=1}^q h_{AM}(\boldsymbol{v}_i; \eta_i, \boldsymbol{\beta}_i)$$

$$- \sum_{i=1}^q KL\left[p_{AM}\left(\boldsymbol{v}_i; \phi_m^k, \boldsymbol{\beta}_m\right)||p_{AM}(\boldsymbol{v}_i; \eta_i, \boldsymbol{\beta}_i)\right],$$

where $p_{AM}\left(\boldsymbol{v}_i; \phi_m^k, \boldsymbol{\beta}_m\right)$ is the marginal of the maximizer of the full (A-MIND) problem and $p_{AM}(\boldsymbol{v}_i; \eta_i, \boldsymbol{\beta}_i)$ the maximizer of the (A-MIND) problem pertaining to $\boldsymbol{v}_i$. Furthermore,

$$-\boldsymbol{\theta}^T \boldsymbol{\alpha}_m^k \leq \sum_{i=1}^q KL\left[p_{AM}\left(\boldsymbol{v}_i; \phi_m^k, \boldsymbol{\beta}_m\right)||p_{AM}(\boldsymbol{v}_i; \eta_i, \boldsymbol{\beta}_i)\right],$$

and the equality holds if and only if $\boldsymbol{\theta} = 0$, condition satisfied if $\forall \boldsymbol{u}, \ \gamma(\boldsymbol{u}) = 1$.

The proof is provided in Appendix B.5. Essentially, to solve a full (A-MIND) problem, we may partition $\boldsymbol{u}$ into q blocks, solve the q (A-MIND) problems in parallel using within-block constraints, determine whether $\prod_{i=1}^{q} p_{\mathrm{AM}}(\boldsymbol{v}_i; \eta_i, \boldsymbol{\beta}_i)$ satisfies the between-blocks constraints to an acceptable tolerance, and if not solve the full (A-MIND) problem using the parameters of $\prod_{i=1}^{q} p_{\mathrm{AM}}(\boldsymbol{v}_i; \eta_i, \boldsymbol{\beta}_i)$ as initial parameters.

4 Estimation

We now turn to estimating the parameters of the solutions of the problems (A-MIND) and (MIND). With ϕ_m^k satisfying (P5), let us consider the problem:

$$\min_{\boldsymbol{\theta}} \quad -\boldsymbol{\theta}^T \boldsymbol{\alpha}_m^k + \int_{[0,1]^d} e^{\boldsymbol{\theta}^T \phi_m^k(\boldsymbol{u})} d\boldsymbol{u}. \quad \text{(CVX-MIND)}$$

4.1 Convex Estimation

Lemma 4.1. *The optimization problem (CVX-MIND) is strictly convex.*

Lemma 4.2. *The minimizer of problem (CVX-MIND) is the maximizer of problem (A-MIND).*

Theorem 4.1. *Let ϕ_m^k satisfy (P3) and (P5). If $\hat{\boldsymbol{\beta}}_{m,n}$ is a consistent estimator of $\boldsymbol{\beta}_m := E_{P_{\boldsymbol{u}_z}}[\psi_m(\boldsymbol{u})]$, then for every $m > 0$*

$$h_{AM}\left(\boldsymbol{u}; \phi_m^k, \hat{\boldsymbol{\beta}}_{m,n}\right) \underset{k,n \to \infty}{\longrightarrow} h_M(\boldsymbol{u}; \psi_m, \boldsymbol{\beta}_m), \quad \text{(A)}$$

and for every $k > 0$

$$h_{AM}\left(\boldsymbol{u}; \phi_m^k, \hat{\boldsymbol{\beta}}_{m,n}\right) \underset{m,n \to \infty}{\longrightarrow} h(\boldsymbol{u}_z). \quad \text{(B)}$$

See Appendix B.6 for the proof of Lemma 4.1, Appendix B.7 for the proof of Lemma 4.2, and Appendix B.8 for the proof of Theorem 4.1.

4.2 Choice of k and ψ_m

It follows from Theorem 4.1 that k controls how close to uniform marginals of p_{AM} are, while m controls how close $h_{\mathrm{AM}}(\boldsymbol{u}_z)$ is to $h(\boldsymbol{u}_z)$. When the object of study is to learn the copula itself not just its entropy, k should be as large as necessary. Note however that the maximum-entropy problem (A-MIND) inherently favors distributions that are as close to uniform as allowed by empirical evidence, so that a large k might not be needed in practice.

An example family $(\psi_m)_m$ that satisfies (P3) are polynomials of degree m, thanks to the Stone-Weierstrass

theorem (Rudin (1973)). With this choice of ψ_m, our approach can be regarded as a maximum-entropy Taylor expansion of the true log copula density $\log p(\boldsymbol{u}_z)$. Recalling that $\rho_{ij} = 12E[u_i u_j] - 3$ is the population version of the Spearman rank correlation between associated primal variables (Nelsen (2007)), it follows that with m as small as 2, we can capture all smooth 1-to-1 associations[2] between any two coordinates of $\boldsymbol{z}$.

Monomials with degree $m > 2$ grow combinatorially in number, but do not provide as much insights per term as ψ_2. Thus, before considering higher degree polynomials, we suggest leveraging Proposition 2.3 in combination with ψ_2 to incorporate specific types of smooth but non-1-to-1 associations. For instance, using $f_{\boldsymbol{\mu}}(\boldsymbol{x}) := |\boldsymbol{x} - \boldsymbol{\mu}|$, where the absolute value is coordinatewise, and solving the problem (MIND) with $m = 2$ to estimate $I(\boldsymbol{y}; \boldsymbol{x})$ through $I(\boldsymbol{y}; \boldsymbol{x}, f_{\boldsymbol{\mu}}(\boldsymbol{x}))$, allows us to reveal any possible associations of the type 'a coordinate of $\boldsymbol{y}$ tend(s) to be monotonically related to the departure of coordinates of $\boldsymbol{x}$ from some baseline values', as well as all smooth 1-to-1 associations between coordinates of $\boldsymbol{x}$ and $\boldsymbol{y}$. Good examples for $\boldsymbol{\mu}$ are the sample median or mean of $\boldsymbol{x}$. To capture departures from a standard range of values rather than a single one, $f_{\boldsymbol{\mu}}$ can be passed through an ϵ-insensitive loss function. When output y_j is a quasi-periodic function of x_i (e.g. x_i is time and output y_j is seasonal), $f_\pi(x_i) = x_i - \lfloor x_i/\pi_i \rfloor \pi_i$ allows us to capture seasonality-adjusted effects.

4.3 Handling Categorical Data

Categorical and non-ordinal variables should be ordinarily encoded as customary, and ordinal data should be treated as continuous variables. The only practical requirements for the validity of this approach are i) to use a ranking function that assigns different ranks to all inputs including ties (e.g. scipy's 'rankdata' function with method 'ordinal'), and ii) to avoid encoding methods that may result in linearly dependent coordinates (e.g. one-hot-encoding on a binary non-ordinal categorical variable). When a suitable ranking function is not available a small random jitter may be added to ordinal variables to remove ties.

This approach is mathematically valid thanks to the quantization characterization of the mutual information (Cover (1999), Definition 8.54). See Appendix A.2 for more details.

---

[2] A continuous function of one variable is 1-to-1 if and only if it is either decreasing or increasing, both of which are captured by Spearman's rank correlation.

4.4   Properties

We summarize some key properties of our mutual information estimator I_{AM}.

Non-negativity: I_{AM}, is always non-negative thanks to requirement (P4) and Theorem 3.4.

Unboundedness for every n: Because I_{AM} only depends on estimated expected statistics, not i.i.d. samples themselves, and because said estimated expected statistics may take extreme values for any n, I_{AM} cannot be upper-bounded by a function of n.

Marginal-Invariance: Our entire approach does not depend on marginal distributions. Additionally, our copula entropy estimator is invariant by any increasing univariate feature transformation (it depends on the data solely through ranks), and any smooth 1-to-1 univariate feature transformation for k large enough enough (Proposition 2.1 and Theorem 4.1-A).

Low Variance: Our approach depends on the data generating distribution solely through the expected statistics constraints, which are estimated using Equation (2) with $O(1/n)$ MSE rate (Ruschendorf (1976)). By the delta method, both the associated natural parameters $\boldsymbol{\theta}$ and the corresponding copula entropies $-\boldsymbol{\theta}^T \hat{\boldsymbol{\alpha}}_{m,n}$, and therefore I_{AM}, have MSE rate $O(1/n)$.

Consistency: The fact that I_{AM} is a consistent estimator of the true mutual information is a direct consequence of the consistency of the rank estimator Equation (2) and of the individual copula entropy estimators as a result (see Theorem 4.1-B).

Low Complexity: With our choice of ϕ_m, pre-optimization complexity is dominated by the computation of ranks, which scales in $\mathcal{O}(d^2 n \log n)$, while optimization can scale in $\mathcal{O}(d^2)$ using gradient descent, where d is the number of inputs and output(s). Calculating the integral over $[0,1]^d$ is only required while solving (CVX-MIND) to compute the gradient and possibly the Hessian; it is not needed to calculate the optimal entropy itself. Thus, a crude approximation using naive Monte Carlo at every learning step is good enough; the resulting algorithm, which can be regarded as mini-batch stochastic gradient descent on a convex objective (Bottou (2010)), will converge to the right solution even for large d.

5   Applications

We begin by applying our approach to clarifying a common misconception.

5.1   The Multivariate Gaussian is Highly Structured

The use of multivariate Gaussian variables is often justified by the fact that they are maximum-entropy (or the least informative of all distributions supported on $\mathbb{R}^d$) under Pearson correlation constraints. Such a choice is equivalent to assuming that marginals are Gaussian and the copula is the Gaussian copula. It might surprise the reader to know that the Gaussian copula is in fact highly structured/informative. For
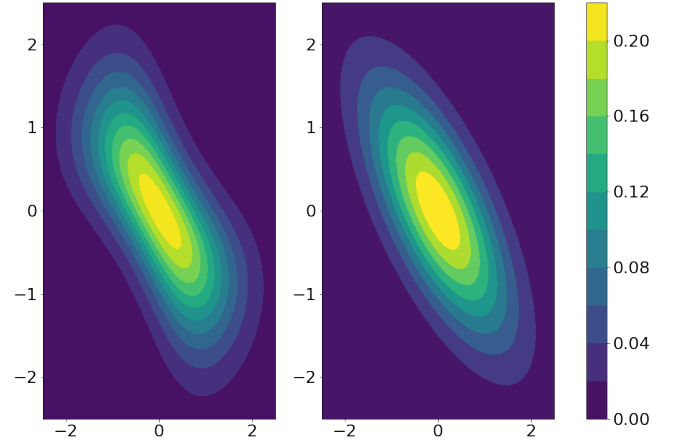

Figure 1: Two bivariate pdfs with standard normal marginals, and the same Spearman rank correlation -0.7. The copula is Gaussian on the right, and maximum-entropy under the Spearman rank correlation constraint on the left.

instance, it can be seen in Figure 1 that the Gaussian copula posits that tails are much more tightly coupled than the corresponding[3] maximum-entropy copula.

Additionally, as illustrated in Figure 4 in the Appendix, the bivariate Gaussian pdf is on average about 10% off, and up to 50% off from the pdf with the same marginals and copula the least informative copula with the same Spearman correlation.[4] When applying the maximum-entropy principle in the primal space, constraints should always be broken down into constraints that solely apply to the copula and constraints that solely apply to marginals (if any). If this is not done, marginal entropies will tend to dominate the copula entropy in the entropy decomposition of Proposition 2.2, and the copula will tend to be low entropy. When constraints are so separable, Equation (1) allows us to break down the optimization problem into two, one maximum-entropy problem about the

---

[3]With the same Spearman correlation structure.

[4]The Spearman rank correlation ρ_s of a bivariate Gaussian with Pearson correlation ρ reads $\rho_s = \frac{6}{\pi} \arcsin\left(\frac{\rho}{2}\right)$ (Kruskal (1958)).

<p>Inductive Mutual Information Estimation: A Convex Maximum-Entropy Copula Approach</p>

copula, which our approach allows solving, and one about marginals. This is not possible using covariance matrices as maximum-entropy constraints given that Pearson's correlation is not a functional of the copula: it *does* depend on marginals. As expected, marginal entropies in this case do dominate the copula entropy in the maximum-entropy problem, which explains why univariate Gaussians are indeed high entropy but the Gaussian copula is low entropy/highly structured.
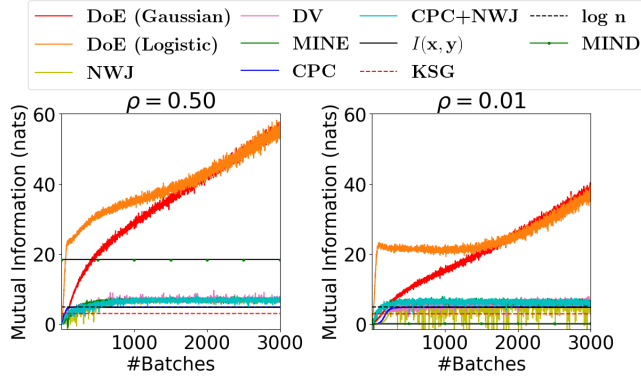

Figure 2: Estimation of the mutual information between two 128-dimensional vectors $\boldsymbol{x} = (x_1, \ldots, x_d)$ and $\boldsymbol{y} = (y_1, \ldots, y_d)$ from a draw of 1000 i.i.d. samples. (x_i, y_i) are i.i.d. Gaussians with mean zero, unit marginal variance, and correlation ρ. For models using deep neural networks, we run the experiment 10 times and report for each batch number the estimate that is the closest to the ground truth. The ground truth is the black solid line labeled $I(\boldsymbol{x}; \boldsymbol{y})$.

5.2 MIND is Far More Data-Efficient Than Competing Approaches

Next, we illustrate that our approach is far more data-efficient than all alternatives, in both low and high mutual information settings.

We repeat the experiment of McAllester and Stratos (2020), and estimate the mutual information between two d-dimensional vectors $\boldsymbol{x} = (x_1, \ldots, x_d)$ and $\boldsymbol{y} = (y_1, \ldots, y_d)$, where (x_i, y_i) are i.i.d. Gaussians with mean zero, unit marginal variance, and correlation ρ. The true mutual information in this case is $I(\boldsymbol{y}; \boldsymbol{x}) = -\frac{d}{2} \log \left(1 - \rho^2\right)$. We reuse the exact same settings as McAllester and Stratos (2020), except for one simple change. Rather than drawing a fresh mini-batch from the true data generating distribution, which is equivalent to using 384000 i.i.d. samples in total, we generate 1000 i.i.d. samples used by all experiments and from which mini-batches are sampled. We run the experiment in a high ($\rho = 0.5$) and a low ($\rho = 0.01$) mutual information setting. We use the code provided

$I(\boldsymbol{y}; \boldsymbol{x})$	MIND	DoE	MINE	NWJ	KSG
18.41	**18.36**	53.88	6.92	7.22	2.98
0.01	**0.08**	37.49	5.83	1.67	2.89

Table 1: Estimates of the mutual information between two 128-dimensional vectors from 1000 i.i.d. samples in low and high mutual information settings using various models. The ground truth is the column $I(\boldsymbol{y}; \boldsymbol{x})$ and the closest model to it is in bold.

by the authors of McAllester and Stratos (2020) at `https://github.com/karlstratos/doe` for all models but KSG (Kraskov et al. (2004)) and MIND. For MIND, we use second order polynomials as ϕ_m. As it can be seen in Figure 2 and in Table 1, MIND is the only model able to come anywhere close to the ground truth in high or low mutual information settings. DoE models clearly overshoot in both settings. If we refer to Figure 2 of McAllester and Stratos (2020), we may conclude that DoE models need about 2000×128 i.i.d. samples to converge to the ground truth in this experiment, which is 256 times more than what MIND requires. Variational models in the primal space overshoot in low mutual information settings and seem to be upper-bounded by $O(\log n)$ in high mutual information settings. As for the nonparametric KSG estimator, it struggles with large input dimensions.

5.3 Copula Entropy Regularized Generative Adversarial Networks (CER-GANs)

Finally, we illustrate that our approach may be used to prevent mode collapse in GANs.

We recall that GANs (Goodfellow et al. (2014)) are very effective generative models made of two modules: a generator whose state is represented by a function $G : \mathcal{Z} \subset \mathbb{R}^q \to \mathcal{X}$, and a discriminator whose state is represented by a function $D : \mathcal{X} \to [0, 1]$. The aim is to learn a function G that maps a simple noise or code distribution $P_{\boldsymbol{z}}$ supported on $\mathcal{Z}$ to a true data generating distribution of interest $P_{\boldsymbol{x}}$ supported on $\mathcal{X}$, so that we may draw samples from the true distribution $P_{\boldsymbol{x}}$ (e.g. realistic images) simply as $G(\boldsymbol{z})$ with $\boldsymbol{z} \sim P_{\boldsymbol{z}}$. To do so, a GAN alternates between two steps. The first step consists of learning a discriminating function D that, as the predictive probability of a binary classifier, is effective at telling draws from the true distribution $P_{\boldsymbol{x}}$ apart from draws of the form $G(\boldsymbol{z})$ (i.e. that are fake). This is done by maximizing the likelihood-like objective

$$\mathcal{L}_D := E_{P_{\boldsymbol{x}}} \left[\log D\left(\boldsymbol{x}\right)\right] + E_{P_{\boldsymbol{z}}} \left[\log \left(1 - D\left(G(\boldsymbol{z})\right)\right)\right]$$

over functions D induced by a deep neural network, for a given G. The second step consists of updating the state G of the generator so as to trick the discriminator
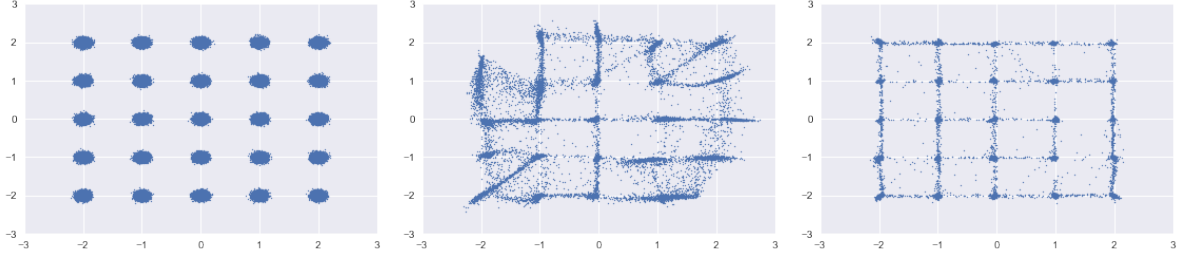
Figure 3: Illustration of GANs trained on the 25 Gaussians dataset (left) without regularization (middle), and with copula entropy regularization (right). Both models use the same discriminator and generator architectures and were stopped after 500 epochs.

into thinking that fake samples are real. This is done by minimizing the objective

$$\mathcal{L}_G := E_{P_z} \left[\log \left(1 - D \left(G(z) \right) \right) \right]$$

over functions G induced by another deep neural network. When P_x is multi-modal, as is often the case in real-life applications, if the generator becomes good at sampling from a mode of P_x, then it will keep generating samples near the same mode, as no term in the objectives $\mathcal{L}_D$ and $\mathcal{L}_G$ incentivizes the generator to keep exploring beyond a mode. This pathology of GANs is known as mode collapse (Che et al. (2016)). Belghazi et al. (2018) identified as possible solution regularizing $\mathcal{L}_G$ with the entropy of $G(z)$ so as to foster exploration, but the authors considered this solution intractable. Instead they followed Chen et al. (2016) and focused on cases where, in addition to z, the generator uses meta-data c that implicitly identify modes of the true distribution P_x (e.g. the digit in the case of MNIST). The authors realized that a low mutual information $I\left(G(z, c); c \right)$ between fake samples and the associated codes reveals mode collapse, and consequently proposed regularizing $\mathcal{L}_G$ with the negative of the foregoing mutual information term:

$$\tilde{\mathcal{L}}_G = \mathcal{L}_G - \beta I \left(G(z, c); c \right), \ \beta > 0.$$

We propose an alternative that does not rely on meta-data.

It follows from Sklar's theorem (Theorem 2.1) that to prevent the generator from collapsing to a mode, it suffices to prevent the copula of $G(z)$ from collapsing to a mode, which can be done by regularizing the generator's loss function with the entropy of the copula of $G(z)$:

$$\bar{\mathcal{L}}_G = \mathcal{L}_G - \beta h \left(u_{G(z)} \right), \ \beta > 0.$$

Using the problem (CVX-MIND), we may write the regularized batch training step of the generator as

$$\min_G \ \frac{1}{b} \sum_{i=1}^{b} \log \left[1 - D \left(G(z_i) \right) \right] + \beta \left[\theta^T \phi \left(\hat{u}_i \right) - e^{\theta^T \phi(u_i)} \right]$$

where $\beta > 0$, ϕ is the constraint function, $u_i \sim \mathcal{U} \left([0,1]^d \right)$ are standard uniforms, and $\hat{u}_i = \frac{\text{rg}(G(z_i))}{b+1}$, the rank being understood as within-batch. The generator step is now followed by the batch copula step

$$\min_\theta \ \frac{1}{b} \sum_{i=1}^{b} -\theta^T \phi \left(\hat{u}_i \right) + e^{\theta^T \phi(u_i)}.$$

We call this model Copula Entropy Regularized GANs (CER-GANs). Figure 3 illustrates the efficacy of CER-GANs on the 25-Gaussians dataset (Belghazi et al. (2018)).

6   Conclusion

We propose a novel approach for solving maximum-entropy copula problems under flexible linear constraints as a convex optimization problem, and we apply our finding to estimating the mutual information between two random vectors. Our approach is inductive in that it relies on the data generating distribution solely through some association-revealing nonparametric properties; it does not assume we have enough data to fully characterize the underlying true joint distribution. This allows the resulting estimator, which we denote MIND, to be considerably more data efficient than all competing models. We show that MIND can accurately estimate the mutual information even when the sample size n is not large enough to fully characterize the true data generating distribution. For large n settings, we show that MIND is a consistent estimator of the true mutual information and has MSE rate $O(1/n)$. Beyond mutual information estimation, we show that mode collapse in GANs can be mitigated by adding a regularizing term that maximizes the copula entropy of the generator using MIND.

An implementation of MIND can be accessed from the Python package 'kxy' available from Pypi (by running 'pip install kxy') or GitHub (https://github.com/kxytechnologies/kxy-python).