Investigation of Alternative Measures for Mutual Information

Kuskonmaz, Bulut; Gundersen, Jaron Skovsted; Wisniewski, Rafal

Published in:
IFAC-PapersOnLine

DOI (link to publication from Publisher):
10.1016/j.ifacol.2022.09.016

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 55-16 (2022) 154–159

Investigation of Alternative Measures for Mutual Information*

Bulut Kuskonmaz * **Jaron S. Gundersen** * **Rafal Wisniewski** *

* *Department of Electronical Systems, Aalborg University, Aalborg, Denmark (e-mails: {bku,jaron,raf}@es.aau.dk).*

Abstract: Mutual information $I(X;Y)$ is a useful definition in information theory to estimate how much information the random variable Y holds about the random variable X. One way to define the mutual information is by comparing the joint distribution of X and Y with the product of the marginals through the Kullback-Leibler (KL) divergence. If the two distributions are close to each other there will be almost no leakage of X from Y since the two variables are close to being independent. In the discrete setting the mutual information has the nice interpretation of how many bits Y reveals about X. However, in the continuous case we do not have the same reasoning. This fact enables us to try different metrics or divergences to define the mutual information. In this paper, we are evaluating different metrics and divergences to form alternatives to the mutual information in the continuous case. We deploy different methods to estimate or bound these metrics and divergences and evaluate their performances.

1. INTRODUCTION

Mutual information has been used as a measure of privacy leakage in several contexts. It dates back to the introduction of information theory in Shannon (1948). Afterwards, it has been used to measure leakage in several contexts such as multiparty computation (MPC), differential privacy, and machine learning Cristiani et al. (2020); Farokhi and Kaafar (2020); Sankar et al. (2013); Urrutia (2018); Cuff and Yu (2016); Li et al. (2021). In this paper we focus on the use of mutual information in privacy-preserving distributed computations, e.g. MPC. Even though you want to compute on real numbers, an MPC protocol is usually converted into finite field operations, and hence the mutual information is between two discrete random variables. However, Tjell and Wisniewski (2021) suggest secret sharing scheme over the real numbers. Using this the MPC can be carried out directly in the real numbers which gives the advantage that some real number computations are easier to carry out. The drawback is that a share might leak a small amount of information, but as described in the paper the amount can be very limited. In any case, this gives rise to study and investigate different ways to measure the leakage in the continuous case.

The interpretation of the mutual information $I(X;Y)$ in the discrete case is how many bits Y on average reveals about X, so if we want to keep X private we want this information to be small if someone learns Y. However, this bit interpretation goes out of the window when the random variables are continuous random variables. Therefore, in the continuous case there might be other alternatives which are just as good as mutual information to measure information leakage. The mutual information compare the joint distribution of X and Y to the product of the marginals through the KL-divergence. Intuitively, this makes sense since if X and Y are independent the product of the marginals and the joint coincide and hence mutual information is equal to zero. A large difference means dependency and hence Y tells more about X.

There could be other alternatives which can be just as good as the KL-divergence. For instance several other measures of distances between probability distributions exist such as Jensen-Shannon divergence, Wasserstein distance, or Total variation distance. In this paper, we define these alternatives for mutual information (Section 2), consider different ways to estimate them (Section 3), and evaluate the performance of the estimators (Section 4).

2. PRELIMINARIES

Let $\mathcal{X}$ be a probability space with σ-algebra $\mathcal{E}$ and measure P. We consider two random variables X and Y on this space, i.e. $X, Y \colon \mathcal{X} \to \mathbb{R}^d$. We equip the measurable space $\mathbb{R}^d$ with the Borel σ-algebra $\mathcal{B}$ and Lebesgue measure λ. The measure on $\mathbb{R}^d$ induced by X given by $P(X^{-1}(B))$ for all $B \in \mathcal{B}$ is called the distribution P_X (similarly for Y) and the probability density function (pdf) $p(x)$ if it exists is defined to be the function satisfying

$$P(X^{-1}(B)) = \int_B p(x) d\lambda(x) \tag{1}$$

for all $B \in \mathcal{B}$, also known as the Radon-Nikodym derivative $p(x) = \frac{dP_X}{d\lambda}$. Hence, we remark that we will use the following two equivalent notations for integrals

$$\int_{\mathcal{X}} dP_X = \int_{\mathbb{R}^d} p(x) d\lambda(x). \tag{2}$$

2.1 Definition of Mutual Information and Alternatives

Now, we present the definitions of the different divergences/metrics and we start with the original definition of mutual information from the KL-divergence. [1]

* This work was supported by Poul Due Jensens Foundation

[1] We remark that the mutual information can also be defined using the differential entropy but this is equivalent to the KL-definition.

10.1016/j.ifacol.2022.09.016

Definition 1. Let P and Q be two probability distributions, where P is absolutely continuous with respect to Q. The Kullback-Leibler (KL) divergence between them is defined as

$$D_{KL}(P \parallel Q) = \int_{\mathcal{X}} \log\left(\frac{dP}{dQ}\right) dP = \int_{\mathbb{R}^d} p(x) \log\left(\frac{p(x)}{q(x)}\right) d\lambda(x), \tag{3}$$

where $\frac{dP}{dQ}$ is the Radon-Nikodym derivative and the last equality holds if the pdf's $p(x)$ and $q(x)$ exists.

Definition 2. Let $X\colon \mathcal{X} \to \mathbb{R}^d$ and $Y\colon \mathcal{Y} \to \mathbb{R}^m$ be two continuous random variables with distributions P_X and P_Y. If $P_{(X,Y)}$ is the joint distribution of X and Y, then the mutual information between X and Y is given by

$$I_{KL}(X;Y) = D_{KL}(P_{(X,Y)} \parallel P_X \otimes P_Y). \tag{4}$$

As we can see, the mutual information is a comparison (using the KL-divergence) of the joint distribution with the product of the marginals. However, as described in the introduction we can also make this comparison using other divergences/metrics on probability measures. We will introduce some of these below. We first consider f-divergences Rényi (1961), which is a class of divergences the KL-divergence also belongs to. An f-divergence between two probability distributions P and Q, where P is absolutely continuous with respect to Q, is defined as

$$D_f(P \parallel Q) = \int_{\mathcal{X}} f\left(\frac{dP}{dQ}\right) dQ, \tag{5}$$

and if both P and Q are absolutely continuous with respect to the Lebesgue measure λ we have the densities satisfying $dP = p(x)d\lambda(x)$ and $dQ = q(x)d\lambda(x)$ implying that

$$D_f(P \parallel Q) = \int_{\mathbb{R}^d} f\left(\frac{p(x)}{q(x)}\right) q(x) d\lambda(x). \tag{6}$$

With this notation the KL-divergence is an f-divergence with $f(t) = t\log(t)$.

Definition 3. Let P and Q be two probability distributions. The Jensen-Shannon (JS) divergence of P and Q is an f-divergence with $f(t) = \frac{1}{2}\left((t+1)\log(\frac{2}{t+1}) + t\log(t)\right)$.

From this, it can be deduced that [2]

$$\begin{aligned} D_{JS}(P \parallel Q) &= \frac{1}{2} D_{KL}(P \parallel \frac{P+Q}{2}) + \frac{1}{2} D_{KL}(Q \parallel \frac{P+Q}{2}) \\ &= \frac{1}{2} \int_{\mathbb{R}} p(x) \log\left(\frac{2p(x)}{p(x)+q(x)}\right) + q(x) \log\left(\frac{2q(x)}{p(x)+q(x)}\right) d\lambda(x), \end{aligned} \tag{7}$$

Definition 4. Let P and Q be two probability distributions. The Total variation (TV) divergence between them is an f-divergence with $f(t) = \frac{1}{2}|t-1|$.

Again, this shows that

$$D_{TV}(P \parallel Q) = \frac{1}{2} \int_{\mathcal{X}} |dP - dQ| = \frac{1}{2} \int_{\mathbb{R}^d} |p(x) - q(x)| d\lambda(x). \tag{8}$$

Intuitively, it makes sense to compare how much the two densities differ from each other throughout the whole domain. But on the other hand, the total variation can sometimes be a too strong metric.

As an alternative to the f-divergences we also consider the Wasserstein distance which is a metric.

[2] The JS-divergence is a symmetric version of the KL-divergence.

Definition 5. Let P and Q be two probability measures on a metric space and let $d(x,y)$ be a metric on this space. Then the k'th Wasserstein distance is

$$W_k(P,Q) = \inf_{\gamma \in \Gamma(P,Q)} \left(\int_{\mathcal{X}\times\mathcal{X}} d(x,y)^k d\gamma(x,y) \right)^{\frac{1}{k}} \tag{9}$$

where $\Gamma(P,Q)$ is the set of all couplings of P and Q, i.e. the set of measures having P and Q as the marginals.

One can see γ as a transportation plan for transforming P into Q. Therefore, $W_1(P,Q)$ is also known as the earth mover distance since it measures the cost of "moving the mass" from P to Q. Since W_1 has the earth mover interpretation this will be the one we are focusing on the most in this paper.

In fact, both the W_1 and the TV distance can be described as an integral probability metric (IPM) Müller (1997).

Definition 6. For two probability measures P and Q on a measurable space $\mathcal{X}$, IPM is given by

$$\gamma(\mathcal{F}, P, Q) = \sup_{f \in \mathcal{F}} \int f(x) dP(x) - \int f(y) dQ(y) \tag{10}$$

where $\mathcal{F}$ is a space for measurable functions on $\mathcal{X}$.

For different $\mathcal{F}$ we obtain different metrics. In fact, due to the Rubenstein-Kantorovich duality setting the class $\mathcal{F} = \mathcal{L}_M$ where, $\mathcal{L}_M$ is the set of all functions that are 1-Lipschitz functions Villani (2009) we obtain the Wasserstein-1 distance as an IPM. Similarly, we can capture the TV by setting $\mathcal{F}_{TV} = \left\{ f \mid \|f\|_\infty \leq \frac{1}{2} \right\}$ [3] where $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$.

To measure information leakage we compare the joint distribution with the product of the marginals of two random variables X and Y, like in (4). Since the other divergences/metrics can be used as alternatives for measuring the information leakage we will use a mutual information-like notation. Hence, we write $I_{JS}(X;Y)$, $I_{TV}(X;Y)$, and $I_{W_k}(X;Y)$ when we do the same comparison as in (4) but with another divergence/metric.

2.2 Relation Between Different Measures

There are a lot of relations between the divergences. In this version we just state a few of them and refer the reader to the full version for more details.

$$0 \leq D_{JS}(P \parallel Q) \leq D_{TV}(P \parallel Q) \leq 1, \tag{11}$$

$$D_{TV}(P \parallel Q) \leq \sqrt{\frac{1}{2} D_{KL}(P \parallel Q)}, \tag{12}$$

$$D_{TV}(P \parallel Q) \leq \sqrt{1 - e^{-D_{KL}(P\|Q)}}. \tag{13}$$

We also mention, W_k is increasing with k, i.e. $W_{k_1}(P,Q) \leq W_{k_2}(P,Q)$ when $k_1 \leq k_2$ Villani (2009).

3. MUTUAL INFORMATION AS LEAKAGE MEASURE IN MPC

In MPC n parties would like to compute $f(x_1, \ldots, x_n)$ where x_i is held by the i'th party. The computation needs to be secure even in the presence of an adversary corrupting a number of the parties. This means for instance

[3] We choose $\frac{1}{2}$ as our bound to be consistent with definition 4 referring to $(\beta - \alpha = 1)$ in Theorem 2 in Sriperumbudur et al. (2009)

that the adversary is not allowed to learn more than it will learn from the input of the corrupted parties and the output $f(x_1, \dots, x_n)$. While the security requires a simulation proof, the privacy requirement can be stated via mutual information in the following way

$$I(X_i; \text{View}_{\mathcal{A}}) = I(X_i; f(X_1, \dots, X_n), \{X_j\}_{j\in\mathcal{A}}) \quad (14)$$

where $\text{View}_{\mathcal{A}}$ is everything the adversary $\mathcal{A}$ sees through the algorithm. Some privacy-preserving algorithms do not guarantee equality in (14) but ensure that $I(X_i; \text{View}_{\mathcal{A}})$ is not much higher than $I(X_i; f(X_1, \dots, X_n), \{X_j\}_{j\in\mathcal{A}})$. Hence, it is interesting to be able to compute $I(X_i; \text{View}_{\mathcal{A}})$ but the density of $\text{View}_{\mathcal{A}}$ is not always known which makes it difficult for computation. Hence, approximating this mutual information is interesting from an MPC perspective.

Thus, we present different ways to approximate the different metrics/divergences defined in Section 2.1 on probability distributions P and Q in the following sections. The approximations are based on samples $\mathbf{X} = \{x_i\}_{i=1}^N$ from a random variable X having distribution P and density $p(x)$ and samples $\mathbf{Y} = \{y_i\}_{i=1}^N$ from a random variable Y having distribution Q and density $q(y)$. Often in practice p and q are not known so we approximate them.

3.1 Leakage Estimation via Histograms

In this section we give a general way to approximate the pdf's using histograms. We build up histograms by splitting the domain into K bins and count the number of instances in each bins. I.e. the domain equals $\bigcup_{i=1}^K B_i$ where $B_i \cap B_j = \emptyset$ when $i \neq j$. This gives rise to an approximate pdf of p, where we let $n_{p,i}$ be the number of instances in bin B_i from the samples of X;

$$\hat{p}(x) = \sum_{i=1}^{K} \mathbb{1}_{x\in B_i} \frac{n_{p,i}}{N \cdot \lambda(B_i)}, \quad (15)$$

where $\mathbb{1}_{x\in B_i}$ is the indicator function for x in B_i. We remark that the B_i's are disjoint and hence for each x exactly one of the terms in the sum is nonzero. Hence, $\hat{p}(x) = \frac{n_{p,i}}{N\cdot\lambda(B_i)}$ if $x \in B_i$. Since all the divergences we look at are f-divergences, they can be described as an integral with respect to the Lebesgue measure of some function of the pdf's, i.e. $f(p(x), q(x))$. Hence, we can split the interval up in a sum of intervals where we integrate over a constant. It means that

$$\begin{aligned}\int_{\mathbb{R}^d} f(p(x), q(x))d\lambda(x) &= \sum_{i=1}^{K} \int_{B_i} f(p(x), q(x))d\lambda(x) \\ \approx \sum_{i=1}^{K} \int_{B_i} f(\hat{p}(x), \hat{q}(x))d\lambda(x) &= \sum_{i=1}^{K} \lambda(B_i)c_i,\end{aligned} \quad (16)$$

where the last equality follows from the fact that $\hat{p}(x)$ and $\hat{q}(x)$ is constant inside bin B_i and hence $f(\hat{p}(x), \hat{q}(x)) = c_i$ is constant inside this bin. This implies the following approximations of the different divergences.

$$\begin{aligned}D_{KL,hist}(P \parallel Q) &= \sum_{i=1}^{K} \frac{n_{p,i}}{N} \log\left(\frac{n_{p,i}}{n_{q,i}}\right) \\ D_{TV,hist}(P \parallel Q) &= \frac{1}{2}\sum_{i=1}^{K} \left|\frac{n_{p,i}}{N} - \frac{n_{q,i}}{N}\right|\end{aligned} \quad (17)$$

$$\begin{aligned}D_{JS,hist}(P \parallel Q) =& \frac{1}{2}\left(\sum_{i=1}^{K} \frac{n_{p,i}}{N} \log\left(\frac{2n_{p,i}}{n_{q,i}+n_{p,i}}\right)\right. \\ &\left. + \frac{n_{q,i}}{N} \log\left(\frac{2n_{q,i}}{n_{q,i}+n_{p,i}}\right)\right).\end{aligned} \quad (18)$$

3.2 Wasserstein Distance Via Optimal Transport

Optimal transport is often formulated in a discrete setting, so we start by considering the optimal transport between two discrete distributions. We consider $p(x)$ and $q(y)$ and they can be described by two vectors $\mathbf{p}$ and $\mathbf{q}$ where the entries are the probabilities for the different outcomes and hence the entries in $\mathbf{p}$ (and $\mathbf{q}$) sum to 1. Optimal transport describes the cost (how far the mass needs to be moved and how much mass) of transporting $p(x)$ to $q(y)$.

The solution to the optimal transportation problem between $\mathbf{p}$ and $\mathbf{q}$ is nothing but a matrix $[M_{ij}]_{i,j=1}^N \in \mathbb{R}_+^{N\times N}$ where the element M_{ij} is the amount of mass transported from p_i to q_j. In order to find the optimal transportation plan M using the cost function C, consider

$$\begin{aligned}d(P, Q) =& \min_{M\geq 0} \langle C, M\rangle \\ &\text{subject to } M\mathbf{1} = \mathbf{p}, \\ &\qquad M^T\mathbf{1} = \mathbf{q},\end{aligned} \quad (19)$$

where $\mathbf{1}$ is the vector of all ones. The optimal transportation plan between $\mathbf{p}$ and $\mathbf{q}$ is obtained after solving the linear program (LP) problem in (19) Haasler et al. (2021).

Now we look at the continuous case. We still have $p(x)$ and $q(y)$ but they are now continuous functions. The equivalence of minimizing $\langle C, M\rangle$ with respect to M in the discrete setting is to take the infimum with respect to all possible $m(x, y)$ of $\int_{\mathbb{R}^d\times\mathbb{R}^d} c(x, y)m(x, y)d\lambda(x, y)$ satisfying $\int_{\mathbb{R}^d} m(x, y)\,\mathrm{d}\lambda(x) = q(y)$ and $\int_{\mathbb{R}^d} m(x, y)\,\mathrm{d}\lambda(y) = p(x)$. But with $c(x, y)$ being a metric (and in our case the Euclidean distance) this is nothing else than the W_1 distance from definition 5. Approximating $p(x)$ and $q(x)$ by histograms from samples we can approximate the Wasserstein distance by solving (19) setting the entries in $\mathbf{p}$ equal $\frac{n_{p,i}}{N}$ and similarly for $\mathbf{q}$. In this case we need to define the distance matrix C but a natural way to do so, is to compute the distance between the centers of the bins.

The optimal transportation is very useful when we have a relatively small number of bins. However, when we have a large amount of bins, solving the LP problem can be computationally heavy. To address this issue, it is suggested to apply the Sinkhorn distance. The optimization problem can be converted into a Sinkhorn distance between two probability vectors $\mathbf{p}$ and $\mathbf{q}$ by introducing a Lagrange multiplier for the entropy constraint as

$$\begin{aligned}d_\lambda(P, Q) =& \min_{M\geq 0} \langle C, M\rangle - \frac{1}{\lambda}h(M) \\ &\text{subject to } M\mathbf{1} = \mathbf{p}, \\ &\qquad M^T\mathbf{1} = \mathbf{q},\end{aligned} \quad (20)$$

where $\lambda \geq 0$ is a tuning parameter that scales the entropy constraint $h(M) = -\sum_{i,j=1}^N M_{ij}\log(M_{ij})$ (Cuturi (2013), equation (2)). By optimizing d_λ, implies an upper bound on W_1 distance which computationally should be easier

to compute. Furthermore, we mention that if λ is large, the Sinkhorn distance would be a good approximation of the Wasserstein distance. We used Algorithm (1) in Cuturi (2013) and please refer to here for a detailed description of the Sinkhorn distance.

3.3 A KL-estimator From Samples

In this section we describe a KL-divergence estimator given a set of samples from two distributions. The estimator is presented in Perez-Cruz (2008). The estimator is computed by approximating $p(x)$ and $q(y)$ around x_i by looking at the k-th nearest neighbor to x_i. They show that even though the approximations of $p(x)$ and $q(y)$ do not necessarily converge to $p(x)$ and $q(y)$ the estimator will converge to the true KL-divergence when increasing the sample size. The estimator is given by

$$D_{KL,k-nn}(P \parallel Q) = \frac{d}{n}\sum_{i=1}^{N}\log\left(\frac{s_k(x_i)}{r_k(x_i)}\right) + \log\left(\frac{N}{N-1}\right), \quad (21)$$

where $r_k(x_i)$ is the Euclidean distance from x_i to the k-the nearest neighbor in $\mathbf{X}\setminus\{x_i\}$, and $s_k(x_i)$ is the Euclidean distance to the k-the nearest neighbor in $\mathbf{Y}$.

3.4 Leakage Estimation via Kernel Mean Embedding

The use of kernel functions range widely in classical machine learning topics for deploying inner product $\langle x\,, x^{'}\rangle$ of two data instances $x, x^{'} \in \mathcal{X}$ which measures the distance between those instances. However, linear functions applied to the inner product sometimes fail when generalizing the distance measure.

In order to overcome this issue, one can apply a "kernel trick" and make the distance measure accurate enough by replacing the inner product with a possible non-linear mapping. Kernel methods rely on kernel functions and can be defined as an inner product of a mapping function which transforms data instances into a higher dimensional feature space as $k(x, x^{'}) = \langle\phi(x)\,, \phi(x^{'})\rangle_{\mathscr{H}}$ where $k(\cdot,\cdot)$ is the kernel function, $\phi(\cdot)$ is the mapping function for data instances as $\{\phi : \mathcal{X} \to \mathscr{H}, x \to \phi(x)\}$, and $\langle\phi(x)\,, \phi(x^{'})\rangle_{\mathscr{H}}$ is the inner product in the reproducing kernel Hilbert space (RKHS) $\mathscr{H}$. By introducing this trick, which depends on substituting the $\langle x\,, x^{'}\rangle$ with $\langle\phi(x)\,, \phi(x^{'})\rangle_{\mathscr{H}}$, it is possible to apply the inner product in a higher dimension, and measure the similarities between x and $x^{'}$ Muandet et al. (2016). In this way, we do not need to explicitly construct $\phi(x)$ and not need to know $\mathscr{H}$ specifically. It will be sufficient to use positive definite kernels in $\mathscr{H}$ for the benefits. As a such kernel, we used the Gaussian kernel $k\,(x, x') = \exp(-\frac{\|x-x'\|_2^2}{2\sigma^2})$ with $\sigma = \sqrt{1/2}$ for this paper.

We define yet another metric from IPM, namely the maximum mean discrepancy (MMD) (with the help of kernel function eventually) using $\mathcal{F} := \{f \mid \|f\|_{\mathscr{H}} \leq 1\}$ to get the functions from a unit ball in $\mathscr{H}$. Let $X, X', Y,$ and Y' be independent representations as $X, X' \sim P$ and $Y, Y' \sim Q$. We show in the full version that the MMD can be computed as

$$\begin{aligned}\mathrm{MMD}^2[\mathscr{H}, P, Q] = \mathbb{E}_{P,P}[k(X, X')] + \mathbb{E}_{Q,Q}[k(Y, Y')] \\ - 2\mathbb{E}_{P,Q}[k(X, Y)]\end{aligned} \quad (22)$$

which leads to the unbiased estimator of the MMD Borgwardt et al. (2006) below

$$\begin{aligned}\widehat{\mathrm{MMD}}^2_u[\mathscr{H}, \mathbf{X}, \mathbf{Y}] = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}^{N} k(x_i, x_j) \\ + \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}^{N} k(y_i, y_j) - \frac{2}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k(x_i, y_j).\end{aligned} \quad (23)$$

4. EXPERIMENT AND RESULTS

Since the motivation is especially leakage of information in a distributed computation such as a MPC protocol we evaluate the metrics in such setups. We take our inspiration from Tjell and Wisniewski (2021) where secret sharing is defined over the real numbers. In contrast to the traditional finite field MPC a share might reveal some (but limited) information. Hence, we will evaluate the leakage of a share and a small MPC using this concept.

4.1 A Secret and Its Share Scenario

First, we consider the leakage of information of a normal distributed $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ from an obfuscated version of X, namely $X - R$ where $R \sim \mathcal{N}(\mu_r, \sigma_r^2)$. This can also be seen as a share of X in the real number secret sharing scheme from Tjell and Wisniewski (2021). The joint distribution of $(X, X - R)$ is

$$\mathcal{N}\left(\begin{bmatrix}\mu_x \\ \mu_x - \mu_r\end{bmatrix}, \begin{bmatrix}\sigma_x^2 & \sigma_x^2 \\ \sigma_x^2 & \sigma_x^2 + \sigma_r^2\end{bmatrix}\right), \quad (24)$$

and the product of the marginals is

$$\mathcal{N}\left(\begin{bmatrix}\mu_x \\ \mu_x - \mu_r\end{bmatrix}, \begin{bmatrix}\sigma_x^2 & 0 \\ 0 & \sigma_x^2 + \sigma_r^2\end{bmatrix}\right), \quad (25)$$

and hence we can actually determine the KL-divergence explicitly in this case. In our experiments, we assume that both X and R has 0 mean so for simplicity we do the same here. In this situation we have

$$I_{KL}(X; X - R) = \frac{1}{2}\ln\left(1 + \frac{\sigma_x^2}{\sigma_r^2}\right). \quad (26)$$

In the experiments we set $\sigma_x^2 = 1$ and $\sigma_r^2 = 10$ implying that

$$I_{KL}(X; X - R) = \frac{1}{2}\ln(1.1) = 0.048. \quad (27)$$

The upper bound on TV in (12) is less than the bound in (13) in this case, and it implies that

$$I_{JS}(X; X - R) \leq I_{TV}(X; X - R) \leq 0.154. \quad (28)$$

However, using that the JS-divergence is a symmetrized version of the KL-divergence, a result from Durrieu et al. (2012) implies that

$$I_{JS}(X; X - R) \leq 0.0356. \quad (29)$$

For details about this we refer to the full version of our paper, where we also argue that

$$I_{W_1}(X; X - R) \leq W_2(P_{(X,X-R)}, P_X \otimes P_{(X-R)}) = 0.292 \quad (30)$$

when we use the covariance matrices from (24) and (25).

4.2 Multiplication Using Three Parties Scenario

In this scenario we consider a situation with three parties where two of them having a value s and t respectively and

the parties wants to learn the product st. We treat s and t as outcomes of random variables S and T and we use the Shamir secret sharing scheme from Tjell and Wisniewski (2021) with privacy threshold 1 and evaluation points $p_1 = -1$, $p_2 = 1$, and $p_3 = 2$. In this situation the shares can be constructed by evaluating $f_s(x) = s + (r_s - s)x$ at p_i and hence the shares of s is $(2s - r_s, r_s, -s + 2r_s)$. Notice that having two shares leaks everything about s. The algorithm goes on like this. The party having s, secret shares s by sending $f_s(p_i)$ to the i'th party. Similarly, is t shared by sending $f_t(p_i)$ to the i'th party. Now, the i'th party computes $f_s(p_i)f_t(p_i)$ and sends this value to the other two parties. Since this is a evaluation of a degree-2 polynomial and each party has three evaluations they can determine the polynomial having constant term st. We assume that the first party is not having an input and we want to evaluate how much he learns about s from this distributed algorithm. I.e. we want to evaluate

$$I(S; 2S - R_s, 2T - R_t, R_t R_s, (-S + 2R_s)(-T + 2R_t)). \quad (31)$$

In the experiments we assume that S and T are following a $\mathcal{N}(0, 1)$ distribution and R_s and R_t are normal as well with 0 mean and variance $\sigma_r^2 = 10$.

4.3 Experimental Setup

For both scenarios explained in Section 4.1 and Section 4.2, we consider the convergences for all divergences and metrics with respect to the number of samples and the number of bins used. We report and comment on the run time for each approximation and computation. However, we were unable to compute the Wasserstein and Sinkhorn distance due to memory problems for the scenario in Section 4.2 which indicates that the two methods are impractical for measuring information leakage in higher dimensions. Furthermore, the k-nearest neighbor approach does not seem to converge for the amount of samples we were able to evaluate. Hence, we only show the convergence for the histogram-based divergences in this case.

Due to definition 1 on the KL-divergence errors occur if we have an empty bin for $q(x)$ (since we cannot divide by 0). To overcome this issue, we replace these 0's with a small number (10^{-8}). In our experiments we use $\lambda = 700$ for the Sinhkhorn distance.

4.4 Results

In Figure 1, we observe for the scenario described in Section 4.1 that we have a nice convergence for the histogram based estimations from Section 3.1 of the KL-, TV- and JS-divergences when sufficient amount of samples are used. After 10^5 samples, divergences started to converge smoothly and their standard deviation narrows down. Furthermore, the KL-divergence seems to converge close to its right value in (27). On the other hand, the histogram based Wasserstein distance started to converge when we used a 10^7 number of samples while the Sinkhorn estimation converged using 10^5 samples. The knn estimators for different k oscillates for the samples smaller than 10000 in the scenario 4.1. However, it started to converge when 2000 samples were used. MMD approximation also seems converged as well with the sufficient amount of samples.

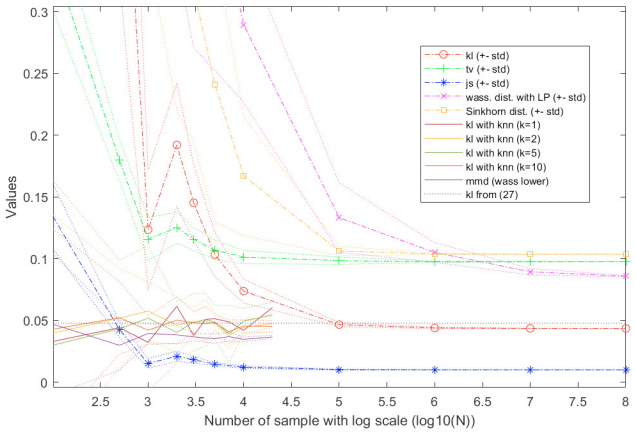

Fig. 1. Average values of metrics for the scenario in 4.1 with respect to logarithm of the number of samples. Dotted lines represent the standard deviation. Number of bins are 24.

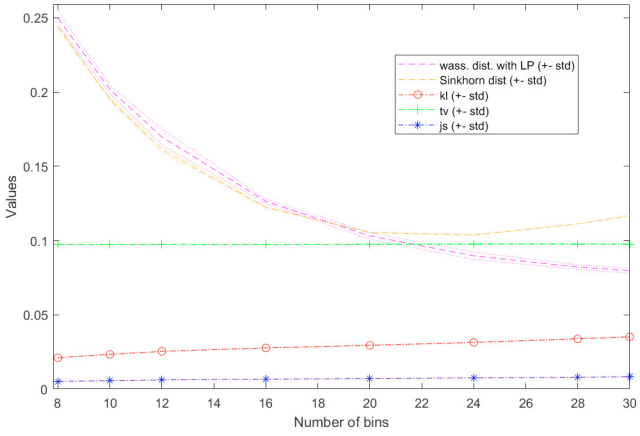

Fig. 2. Average values of metrics for the scenario in 4.1 with respect to number of bins. Dotted lines represent the standard deviation. Number of samples are $N = 10^7$

In Figure 2, we evaluate for the scenario 4.1 that histogram based W_1 via LP started to converge when using 30 bins in both dimensions for the histograms (in total 900 bins). Also Sinkhorn behaves like an upper bound for W_1 at the convergence level. The change in the number of bins does not seem to effect the behavior of TV- and JS-divergence. Table 1 shows the advantage of Sinkhorn over LP calculation for W_1 in terms of the speed. For divergences, it seems that the choice of the number of bins has a minor effect on run time.

	8	10	12	16	20	24	28	30
kl with histograms	0.23	0.19	0.19	1.1	0.23	0.24	0.35	6.9
tv with histograms	0.12	0.09	0.90	0.14	0.10	0.11	0.12	1.0
js with histograms	0.16	0.21	0.12	0.14	0.14	0.17	0.28	1.2
wass. dist. with LP	14	31	52	206	737	1993	4859	12073
Sinkhorn dist.	0.88	1.2	1.3	2.6	7.8	15.1	35.7	99.3

Table 1. Histogram based: Average run times in milliseconds for scenario 4.1 compared to number of bins.

In Table 2, it is obvious that the run time increases gradually for MMD approximation as the number of samples in use increases in the both scenarios 4.1 and 4.2. When using knn estimator, there is insignificant increase in runtime as we use more samples for scenario 4.1 unlike MMD. Hence, knn estimator works poorly for the scenario 4.2 since it did not converge to a specific value as we increase the number of samples for the estimation.

	100	500	10^3	2×10^3	3×10^3	4×10^3	5×10^3	7×10^3	10^4	2×10^4
knn (for all k) for scenario 4.1	5	12	20	35	48	67	80	104	142	312
mmd with Gaussian for scenario 4.1	9	102	421	2066	4002	7351	11601	22822	47644	193679
mmd with Gaussian for scenario 4.2	6	101	412	2207	4778	8430	15471	25546	50493	198473

Table 2. Average run times of the methods in 3.3 and 3.4 in milliseconds for different sample sizes.

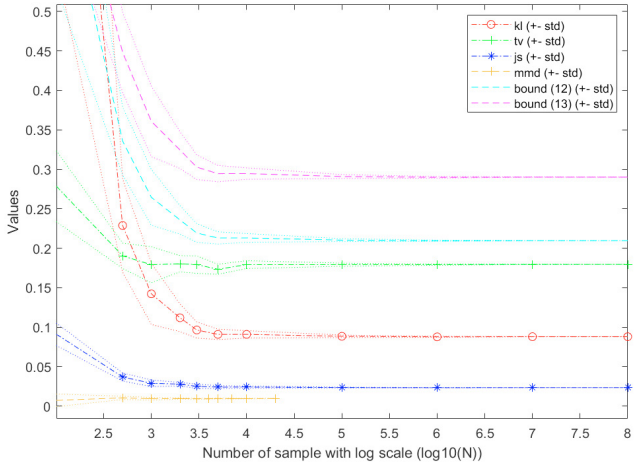

Fig. 3. Average values of metrics for scenario 4.2 with respect to logarithm of the number of samples. Dotted lines represent the standard deviation. Number of bins are 24.

In Figure 3, we observe for the scenario described in Section 4.2 that the convergence of approximations started to emerge when we use 10^5 number of samples for histogram based estimations of the KL-, TV- and JS-divergences. Standard deviation is quite tight after such a number of samples used. Since we do not know the explicit value of KL, we plug in the estimate into the bounds in (12) and (13), which are shown in the figure as well. MMD approximation could only be possible to run with 20000 samples maximum but it is sufficient for its convergence.

5. CONCLUSION AND FUTURE WORK

To sum up, we evaluate the possible divergences and metrics to measure the mutual information $I(X;Y)$ for specific scenarios using MPC in this paper. Results show that the histogram-based estimators of the divergences are strong for approximating the mutual information in terms of the number of samples used and the run time of the approximation. The MMD metric is also a useful measure for its convergence but it can be computationally heavy for applications requiring a high number of samples. Wasserstein distance is a quite informative metric as well. On the other hand, calculating it using LP or estimating it with Sinkhorn's algorithm becomes useless for the samples in the 5 dimension.

REFERENCES

Borgwardt, K.M., Gretton, A., Rasch, M.J., Kriegel, H.P., Schölkopf, B., and Smola, A.J. (2006). Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14), e49–e57.

Cristiani, V., Lecomte, M., and Maurine, P. (2020). Leakage assessment through neural estimation of the mutual information. In *Applied Cryptography and Network Security Workshops*, 144–162. Springer International Publishing, Cham.

Cuff, P. and Yu, L. (2016). Differential privacy as a mutual information constraint. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, 43–54. Association for Computing Machinery, New York, NY, USA. doi: 10.1145/2976749.2978308.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2292–2300.

Durrieu, J.L., Thiran, J.P., and Kelly, F. (2012). Lower and upper bounds for approximation of the kullback-leibler divergence between gaussian mixture models. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4833–4836. doi:10.1109/ICASSP.2012.6289001.

Farokhi, F. and Kaafar, M.A. (2020). Modelling and quantifying membership information leakage in machine learning.

Haasler, I., Karlsson, J., and Ringh, A. (2021). Control and estimation of ensembles via structured optimal transport. *IEEE Control Systems Magazine*, 41(4), 50–69.

Li, Q., Gundersen, J.S., Heusdens, R., and Christensen, M.G. (2021). Privacy-preserving distributed processing: Metrics, bounds and algorithms. *IEEE Transactions on Information Forensics and Security*, 16, 2090–2103. doi: 10.1109/TIFS.2021.3050064.

Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. (2016). Kernel mean embedding of distributions: A review and beyond. *arXiv preprint arXiv:1605.09522*.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2), 429–443.

Perez-Cruz, F. (2008). Kullback-leibler divergence estimation of continuous distributions. In *2008 IEEE International Symposium on Information Theory*, 1666–1670. doi:10.1109/ISIT.2008.4595271.

Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, 547–561. University of California Press.

Sankar, L., Rajagopalan, S.R., and Poor, H.V. (2013). Utility-privacy tradeoffs in databases: An information-theoretic approach. *IEEE Transactions on Information Forensics and Security*, 8(6), 838–852. doi: 10.1109/TIFS.2013.2253320.

Shannon, C.E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x.

Sriperumbudur, B.K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G.R. (2009). On integral probability metrics,\phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*.

Tjell, K. and Wisniewski, R. (2021). Privacy in distributed computations based on real number secret sharing.

Urrutia, F. (2018). Information theory for multi-party peer-to-peer communication protocols. (théorie de l'information pour protocoles de communication peer-to-peer).

Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer.