# Estimating the Unique Information of Continuous Variables in Recurrent Networks

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

The integration and transfer of information from multiple sources to multiple targets is a core motive of neural systems. The emerging field of partial information decomposition (PID) provides a novel information-theoretic lens into these mechanisms by identifying synergistic, redundant, and unique contributions to the mutual information between one and several variables. While many works have studied aspects of PID for Gaussian and discrete distributions, the case of general continuous distributions is still uncharted territory. In this work we present a method for estimating the unique information in continuous distributions, for the case of one versus two variables. Our method solves the associated optimization problem over the space of distributions with fixed bivariate marginals by combining copula decompositions and techniques developed to optimize variational autoencoders. We obtain excellent agreement with known analytic results for Gaussians, and illustrate the power of our new approach in several brain-inspired neural models. Our method is capable of recovering the effective connectivity of a chaotic network of rate neurons, and uncovers a complex trade-off between redundancy, synergy and unique information in recurrent networks trained to solve a generalized XOR task.

## 1 Introduction and background

In neural systems, often multiple neurons are driven by one external event or stimulus; conversely multiple neural inputs can converge onto a single neuron. A natural question in both cases is how multiple variables hold information about the singleton variable. In their seminal work [1], Williams and Beer proposed an axiomatic extension of classic information theory to decompose the mutual information between multiple source variables and a single target variable in a meaningful way. For the case of two sources $X_1, X_2$, their partial information decomposition (PID) amounts to expressing the mutual information of $X_1, X_2$ with a target $Y$ as a sum of unique ($U_1$, $U_2$), redundant ($R$) and synergistic ($S$) non-negative contributions:

$$I(Y:(X_1,X_2)) = U(Y:X_1\backslash X_2) + U(Y:X_2\backslash X_1) + R(Y:(X_1,X_2)) + S(Y:(X_1,X_2))\,, \tag{1.1}$$

which also obey the consistency equations

$$I(Y:X_1) = R(Y:(X_1,X_2)) + U(Y:X_1\backslash X_2)\,, \tag{1.2}$$
$$I(Y:X_2) = R(Y:(X_1,X_2)) + U(Y:X_2\backslash X_1)\,. \tag{1.3}$$

The PID has proved useful in understanding information processing by distributed systems in a diverse array of fields including machine learning [2, 3], earth science [4] and cellular automata [5], and particularly in neuroscience [6–9], where notions of synergy and redundancy, traditionally

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

considered mutually exclusive and distinguished by the sign of

$$\begin{aligned}\Delta &= I(Y:(X_1,X_2)) - I(Y:X_1) - I(Y:X_2)\,,\\ &= S(Y:(X_1,X_2)) - R(Y:(X_1,X_2))\,,\end{aligned} \tag{1.4}$$

have long played a central role in the quest to understand how neural circuits integrate information from multiple sources [10–13]. The novelty of the PID framework here is in separating the measures of synergy and redundancy in (1.4).

The above abstract formulation of PID provides three equations for four unknowns, and only becomes operational once one of $U_1$, $U_2$, $R$, or $S$ is defined. This has been done in [14] via a definition of the unique information:

**Definition 1** (BROJA [14])**.** *Given three random variables $(Y, X_1, X_2)$ with joint probability density $p(y, x_1, x_2)$, the unique information $U_1$ of $X_1$ with respect to $Y$ is*

$$U(Y:X_1\backslash X_2) = \min_{q\in Q} I_q(Y:X_1|X_2)\,, \tag{1.5}$$

$$= \min_{q\in Q} \int dy dx_1 dx_2\, q(y,x_1,x_2) \log\left(\frac{q(y,x_1|x_2)}{q(y|x_2)q(x_1|x_2)}\right)\,, \tag{1.6}$$

*where*

$$Q = \{q(y,x_1,x_2)\,|\,q(y,x_i) = p(y,x_i), i=1,2\}\,. \tag{1.7}$$

In words, we minimize the conditional mutual information $I(Y:X_1|X_2)$ over the space of density functions that preserve the marginal densities $p(y, x_1)$ and $p(y, x_2)$. The above definition implies, along with (1.2)-(1.3), that the unique and redundant information only depend on the marginals $p(y, x_1), p(y, x_2)$, and that the synergy can only be estimated from the full $p(y, x_1, x_2)$.

The original definition in [14] was limited to discrete random variables. Here, we show that the extension to continuous variables is well-defined and can be practically estimated.

**Motivation from decision theory** [14]. Consider for simplicity discrete variables. A decision maker $\mathrm{DM}_1$ can choose an action $a$ from a finite set $\mathcal{A}$, and receives a reward $u(a, y)$ based on the selected action and the state $y$, which occurs with probability $p(y)$. Notably, $\mathrm{DM}_1$ has no knowledge of $y$, but observes instead a random signal $x_1$ sampled from $p(x_1|y)$. Choosing the action maximizing the expected reward for each $x_1$, his maximal expected reward is

$$R_1 = \sum_{x_1} p(x_1) \max_{a|x_1} \sum_y p(y|x_1)u(a,y)\,. \tag{1.8}$$

$\mathrm{DM}_1$ is said to have no unique information about $y$ w.r.t. another decision maker $\mathrm{DM}_2$ that observes $x_2 \sim p(x_2|y)$ – if $R_2 \geq R_1$ for any set $\mathcal{A}$, any distribution $p(y)$, and any reward function $u(a, y)$. A celebrated theorem by Blackwell [15, 16] states that such a generic advantage by $\mathrm{DM}_2$ occurs iff there exist a stochastic matrix $q(x_1|x_2)$ which satisfies

$$p(x_1|y) = \sum_{x_2} p(x_2|y)q(x_1|x_2)\,. \tag{1.9}$$

But this occurs precisely when the unique information (1.5) vanishes, since then there exists a joint distribution $q(y, x_1, x_2)$ in $Q$ for which $y \perp x_1|x_2$, which implies $q(x_1|x_2, y) = q(x_1|x_2)$, and thus (1.9) holds. Similar results exist for continuous variables [17, 18]. Thus the unique information from Definition 1 quantifies a departure from Blackwell's relation (1.9).

In this work we present a definition and a method to estimate the BROJA unique information for generic continuous probability densities. Our approach is based on the observation that the constraints (1.7) can be satisfied with an appropriate copula parametrization, and makes use of techniques developed to optimize variational autoencoders. We only consider one-dimensional $Y, X_1, X_2$ for simplicity, but the method can be naturally extended to higher dimensional cases. In Section 2 we review related works, in Section 3 we present our method and Section 4 contains several illustrative examples.

## 2 Related works

Partial information decomposition offers a solution to a repeated question that was not addressed by 'classical' information theory regarding the relations between two sources and a target [1]. From a mathematical perspective a 'functional definition' has to be made, meaning that such a definition should align with our intuitive notions. Yet, as shown in [19], not all intuitively desirable properties of a PID can be realized simultaneously. Thus, different desirable properties are chosen for distinct application scenarios. Thus, various proposals for decomposition measures are not seen as conflicting but as having different operational interpretations: For example, the BROJA approach used here builds on desiderata from decision theory, while other approaches appeal to game theory [20] or the framework of Kelly gambling [21]. Yet other approaches use arguments from information geometry [22]. Other approaches assume agents receiving potentially conflicting or incomplete information about the source variables for the purpose of inference or decryption (see e.g. [23, 24]). In [25] the authors separate the specific operational interpretations of PID measures from the general structure of information decomposition.

The actual computation of the BROJA unique information is non-trivial, even for discrete variables. Optimization methods exits for the latter case [26–28], and analytic solutions are only known when all the variables are univariate binary [29]. For continuous probability densities, an earlier definition aligned with the BROJA measure was made by Barret [30], but only applies to Gaussian variables. For Barret's measure, an analytic solution is known when $p(y, x_1, x_2)$ is a three-dimensional Gaussian density [30], but does not generalize to higher dimensional Gaussians [31].

PID has been previously applied to *in vitro* neural data to elucidate how single neuron computational properties depend on network topology [7].

## 3 Bounding and estimating the unique information

**Parametrizing the optimization space with copulas.** To characterize the optimization space $Q$ in (1.5)-(1.7), it is convenient to recall that according to Sklar's theorem [32], any $n$-variate probability density can be expressed as

$$p(x_1 \dots x_n) = p(x_1) \dots p(x_n) c(u_1 \dots u_n)\,, \tag{3.1}$$

where $p(x_i)$ is the marginal and $u_i = F(x_i)$ is the CDF of each variable. The dependency structure among the variables is encoded in the function $c{:}[0,1]^n \to [0,1]$. This is a *copula* density, a probability density on the unit hypercube with uniform marginals [33],

$$\int\limits_{[0,1]^{n-1}} \prod_{j=1, j\neq i}^{n} du_j\, c(u_1 \dots u_n) = 1 \quad \forall i\,. \tag{3.2}$$

Note that under univariate reparametrizations $z_i' = g(z_i)$, the $u_i$'s and the copula $c$ remain invariant, and that the copula of any marginal distribution is the marginal of the copula. For an overview of copulas in machine learning, see [34].

**Proposition 1.** *Under the BROJA Definition 1 of unique information, all the terms of the partial information decomposition in (1.1)-(1.3) are independent of the univariate marginals $p(x_1), p(x_2), p(y)$, and only depend on the copula $c(u_y, u_1, u_2)$.*

*Proof.* Expressing $q(y, x_1, x_2), q(x_1, x_2), q(y, x_2)$ via copula decompositions (3.1), and changing variables as $du_y = q(y)dy$, etc., the objective function in (1.6) becomes

$$I_q(Y : X_1|X_2) \quad = \quad \int\limits_{[0,1]^3} du_y du_1 du_2\, c(u_y, u_1, u_2) \log\left(\frac{c(u_y, u_1, u_2)}{c(u_y, u_2)c(u_1, u_2)}\right)\,. \tag{3.3}$$

Thus the optimization objective and the unique information are independent of the univariate marginals. A similar result holds for the mutual information terms in the l.h.s. of (1.1)-(1.3).[1] It follows that none of the PID terms in (1.1)-(1.3) depend on the univariate marginals, and therefore all the PID terms are invariant under univariate reparametrizations of $(y, x_1, x_2)$. □

To find the copula degrees of freedom that parametrize the optimization space $Q$ in (1.7), consider the factorization

$$p(y, x_1, x_2) = p(x_1)p(y|x_1)p(x_2|y, x_1)\,. \tag{3.4}$$

Using the copula decomposition (3.1) for $n = 2$, the last two factors in (3.4) can be expressed as

$$\begin{aligned} p(y|x_1) &= c(u_y, u_1)p(y)\,, && (3.5)\\ p(x_2|y, x_1) &= \frac{p(x_1, x_2|y)}{p(x_1|y)}\,, && (3.6)\\ &= c_{1,2|y}(u_{1|y}, u_{2|y})p(x_2|y)\,, && (3.7)\\ &= c_{1,2|y}(u_{1|y}, u_{2|y})c(u_y, x_2)p(x_2)\,, && (3.8) \end{aligned}$$

where we defined the conditional CDFs,

$$u_{i|y} = F(u_i|u_y) = \frac{\partial C(u_y, u_i)}{\partial u_y} \qquad i = 1, 2 \tag{3.9}$$

and $C(u_y, u_i)$ is the CDF of $c(u_y, u_i)$. Note that the function $c_{1,2|y}(u_{1|y}, u_{2|y})$ in (3.7) is not the conditional copula $c(u_1, u_2|u_y)$, but rather the copula of the conditional $p(x_1, x_2|y)$. Using expressions (3.5) and (3.8), the full density (3.4) becomes

$$p(y, x_1, x_2) = p(y)p(x_1)p(x_2)c(u_y, u_1, u_2)\,, \tag{3.10}$$

where

$$c(u_y, u_1, u_2) = c(u_y, u_1)\, c(u_y, u_2)c_{1,2|y}(u_{1|y}, u_{2|y})\,. \tag{3.11}$$

This is a simple case of the pair-copula construction of multivariate distributions [37–39], which allows to expand any $n$-variate copula as a product of (conditional) bivariate copulas.

**Proposition 2.** *The copula of the conditional, $c_{1,2|y}(\cdot, \cdot)$, parametrizes the space $Q$ in (1.7).*

*Proof.* Since $q(y, x_i) = p(y, x_i)$ $(i = 1, 2)$, the copula factors in

$$p(y, x_i) = p(y)p(x_i)\, c(u_y, u_i)\,, \qquad i = 1, 2 \tag{3.12}$$

are fixed in $Q$. Therefore, parametrizing by $\theta$ the elements $q(y, x_1, x_2) \in Q$, a copula decomposition similar to (3.10)-(3.11) necessarily has the form

$$c_\theta(u_y, u_1, u_2) = c(u_y, u_1)\, c(u_y, u_2)\, c_{1,2|\theta(y)}(u_{1|y}, u_{2|y})\,. \tag{3.13}$$

This differs from (3.11) only in $c_{1,2|\theta(y)}(u_{1|y}, u_{2|y})$, which depends on the function $\theta(y)$ whose parameters are the optimization variables.[2] □

Inserting now the expression (3.13) into the objective function (3.3) we get

$$I_q[\theta] = \mathbb{E}_{c_\theta(u_y,u_1,u_2)} \log \left[c(u_y, u_1)c_{1,2|\theta(y)}(u_{1|y}, u_{2|y})\right] - \mathbb{E}_{c_\theta(u_1,u_2)} \log c_\theta(u_1, u_2)\,, \tag{3.14}$$

which is our objective function and satisfies the marginal constraints (1.7). Note that apart from the optimization parameters $\theta$, it depends on the bivariate copulas $c(u_y, u_1)$ and $c(u_y, u_2)$ which should be estimated from the observed data. Given $D$ observations $\{y^{(i)}, x_1^{(i)}, x_2^{(i)}\}_{i=1}^D$, we map each value to $[0, 1]$ via the empirical CDFs of each coordinate $(y, x_1, x_2)$. This yields a data set

[1] The connection between mutual information and copulas was discussed in [35, 36].

[2] We note that in multivariate pair-copula expansions it is common to assume constant conditioning parameters $\theta$ [40], but we do not make such a simplifying assumption.

$\{u_y^{(i)}, u_1^{(i)}, u_2^{(i)})\}_{i=1}^D$, with $O(D \log D)$ cost. The latter set is used to estimate copula densities $c(u_y, u_1)$ and $c(u_y, u_2)$ by fitting several parametric and non-parametric copula models [41], and choosing the best pair of models using the AIC criterion.[3] From the learned copulas we also get the conditional CDF functions $u_{i|y} = F(u_i|u_y)$ that appear in the arguments of the first term in (3.14).

**A variational upper bound.** Minimizing (3.14) directly w.r.t. $\theta$ is challenging because the second term depends on the copula marginal $c_\theta(u_1, u_2)$ which has no closed form, as it requires integrating (3.13) w.r.t. $u_y$. We introduce instead an inference distribution $r_\phi(u_y|u_1, u_2)$, with parameters $\phi$, that approximates the conditional copula $c_\theta(u_y|u_1, u_2)$, and consider the bound

$$\log c_\theta(u_1, u_2) = \log \int du'_y \, c_\theta(u'_y, u_1, u_2) \geq \int du'_y \, r_\phi(u'_y|u_1, u_2) \, \log \frac{c_\theta(u'_y, u_1, u_2)}{r_\phi(u'_y|u_1, u_2)} \,, \quad (3.15)$$

which follows from Jensen's inequality and is tight when $r_\phi(u'_y|u_1, u_2) = c_\theta(u'_y|u_1, u_2)$. This expression gives an upper bound on $I_q[\theta]$, which can be minimized jointly w.r.t. $(\theta, \phi)$.

A disadvantage of the bound (3.15) is that its tightness depends strongly on the expressiveness of the inference distribution $r_\phi(u'_y|u_1, u_2)$. This situation can be improved by considering a multiple-sample generalization proposed by [43],

$$\log c_\theta(u_1, u_2) \geq D_{K,\theta,\phi}(u_1, u_2) \equiv \mathbb{E}_{p(u_y^{(1)} \ldots u_y^{(K)})} \log \left[ \frac{1}{K} \sum_{k=1}^{K} \frac{c_\theta(u_y^{(k)}, u_1, u_2)}{r_\phi(u_y^{(k)}|u_1, u_2)} \right] \,, \quad (3.16)$$

where the expectation is w.r.t. $K$ independent samples of $r_\phi(u'_y|u_1, u_2)$. $D_{K,\theta,\phi}(u_1, u_2)$ coincides with the lower bound in (3.15) for $K = 1$ and satisfies [43]

$$D_{K+1,\theta,\phi}(u_1, u_2) \geq D_{K,\theta,\phi}(u_1, u_2), \quad (3.17)$$

$$\lim_{K\to\infty} D_{K,\theta,\phi}(u_1, u_2) = \log c_\theta(u_1, u_2) \,. \quad (3.18)$$

Thus, even when $r_\phi(u'_y|u_1, u_2) \neq c_\theta(u'_y|u_1, u_2)$, the bound can be made arbitrarily tight for large enough $K$. Inserting (3.16) in (3.14), we get finally

$$I_q[\theta] \leq B_1[\theta] + B_2[\theta, \phi] \,, \quad (3.19)$$

where

$$B_1[\theta] = \mathbb{E}_{c_\theta(u_y, u_1, u_2)} \log \left[ c(u_y, u_1) c_{1,2|\theta(y)}(u_{1|y}, u_{2|y}) \right] \,, \quad (3.20)$$

$$B_2[\theta, \phi] = -\mathbb{E}_{c_\theta(u_1, u_2)} D_{K,\theta,\phi}(u_1, u_2) \,, \quad (3.21)$$

and we minimize the r.h.s. of (3.19) w.r.t. $(\theta, \phi)$. Low-variance estimates of the gradients to perform the minimization can be obtained with the reparametrization trick [44, 45], as discussed in detail in Appendix A. In our examples below we use for $c_{1,2|\theta(y)}$ a bivariate Gaussian copula (reviewed in Appendix B). Such a copula has just one parameter $\theta \in [-1, +1]$, and thus the optimization is done over the space of functions $\theta(y) = \theta(u_y)$:$[0, 1] \to [-1, +1]$, which we parametrize with a two-layer neural network. Similarly, we parametrize $r_\phi(u_y|u_1, u_2)$ with a two-layer neural network. Details of these networks are in Appendix D.

While the term $B_2$ in our bound is similar to the negative of the ELBO bound in importance weighted autoencoders (IWAEs) [43], there are some differences between the two settings, the most important being that we are interested in the precise value of the bound at the minimum, rather than the learned functions $c_\theta, r_\phi$. Note also that our latent variables $u_y^{(k)}$ are one-dimensional, as opposed to the usual higher dimensional latent distributions of variational autoencoders, and that the empirical expectation over data observations in IWAEs is replaced in $B_2$ by the expectation over $c_\theta(u_1, u_2)$, whose parameters are also optimized.

**Estimating the other PID terms** In the following we adopt the minimal value taken by the upper bound (3.19) as our estimate of $U_1$. The other terms in the partial information decomposition

[3] For this fitting/model selection step, we used the `pyvinecopulib` python package [42].

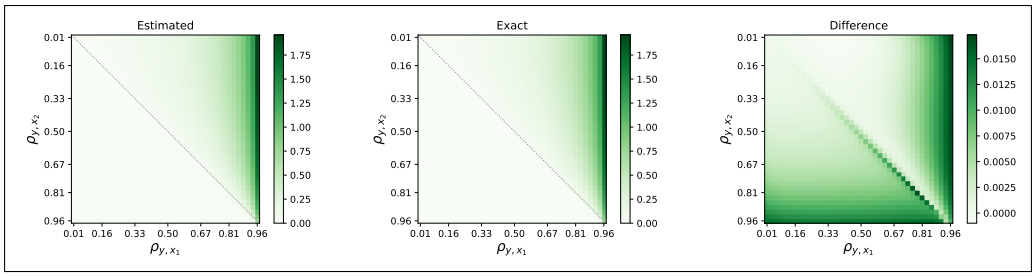


Figure 1: **Estimated vs. exact values of unique information for Gaussians.** For a three-dimensional Gaussian, we show estimates of $U(Y : X_1 \backslash X_2)$ as a function of the correlations $\rho_{y,x_i} (i = 1, 2)$, compared with the exact results from [30]. Only for Gaussian distributions are exact results known for continuous variables.

are obtained from the consistency relations (1.1)-(1.3), after estimating the mutual informations $I(Y : (X_1, X_2)), I(Y : X_1), I(Y : X_2)$. There are several methods for the latter. In our examples, we use the observed data to fit additional copulas $c(u_1, u_2)$ and $c_{12|\theta(y)}$ and estimate $I(Y : X_1) \simeq \frac{1}{D} \sum_{i=1}^{D} \log c(u_y^{(i)}, u_1^{(i)})$ and similarly for the other terms. Note that all our estimates have sources of potential bias. Firstly, the estimation of the parametric copulas is subject to model or parameter misspecification. Secondly, the optimized bound might not saturate, biasing the $U_1$ estimate upwards.

# 4 Examples

**Comparison with exact results for Gaussians** Consider a three dimensional Gaussian with correlations $\rho_{y,x_i}$ between $y, x_i$ for $i = 1, 2$. As shown in [30], the solution to (1.5) in this case is

$$U(Y : X_1 \backslash X_2) = \frac{1}{2} \log \left( \frac{1 - \rho_{y,x_2}^2}{1 - \rho_{y,x_1}^2} \right) \mathbb{1} \left[ \rho_{y,x_2} < \rho_{y,x_1} \right]. \tag{4.1}$$

Figure 1 compares the above expression with estimates from our method. Here we know that $c_{y,1}$ and $c_{y,2}$ are Gaussian copulas, with parameters $\rho_{y,x_1}, \rho_{y,x_2}$, and we assumed a Gaussian copula for $c_{1,2|y,\theta}(u_{1|y}, u_{2|y})$ as well. For each pair of values $\rho_{y,x_1}, \rho_{y,x_2}$. In this and the rest of the experiments, we optimized the parameters $(\theta, \phi)$ using the ADAM algorithm [46] with a fixed learning rate $10^{-2}$ during 1200 iterations. In each iteration we used $M = 3000, K = 50$ samples. The results reported correspond to the mean of the bound in the last 100 iterations. The comparison in Figure 1 shows excellent agreement.

**Model systems of three neurons.** The nature of information processing of neural systems is a prominent area of application of the PID framework, since synergy has been proposed as natural measure of information modification [7, 47]. We consider two models:

$$\begin{array}{cc} \textbf{M1} & \textbf{M2} \\ (X_1, X_2) \sim \mathcal{N}(0, \rho_{12}^2), & (X_1, X_2) \sim \mathcal{N}(0, \rho_{12}^2), \\ Y = \tanh(w_1 X_1 + w_2 X_2). & Y = X_1^2 / \left(0.1 + w_1 X_1^2 + w_2 X_2^2\right). \end{array} \tag{4.2}$$

Both models are parameterized by the correlation $\rho_{12}$ and weights $w_1, w_2$. Model 1 is a particularly simple neural network. The $\tanh$ activation does not affect its copula, and even for a linear activation function the variables are not jointly Gaussian since $Y$ is deterministic on $(X_1, X_2)$. Model 2 is inspired by a normalization operation widely believed to be canonical in neural systems [48] and plays a role in common learned image compression methods [49]. The results, presented in Figure 2. are obtained from 3000 samples from each model

**Computational aspects of connectivity in recurrent neural circuits.** We apply our continuous variable PID to understand computational aspects of the information transfer between recurrently

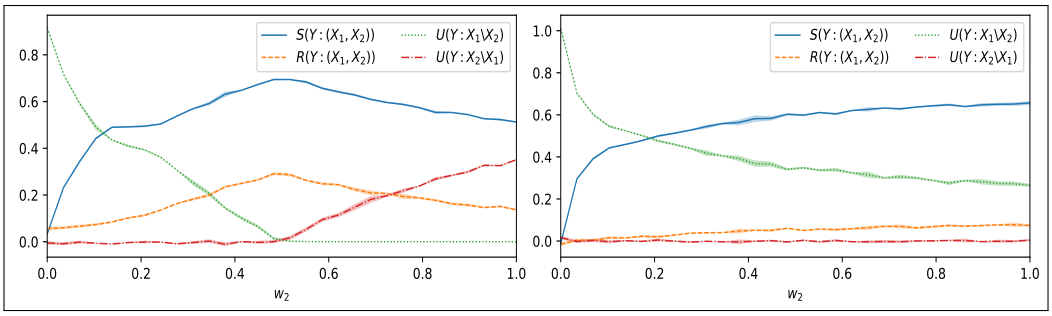


Figure 2: **Partial information decomposition for two neural network models.** In both models (4.2) we fixed $w_1 = 0.5, \rho_{12} = 0.3$, and show the PID terms as a function of the synaptic strength $w_2$, normalized by $I(Y:(X_1, X_2))$. We show mean (lines) and standard deviations (shaded area around each line) from 3 runs. *Left:* Model 1: The input of greatest weight conveys all the unique information, and synergy and redundancy both peak as $w_1 = w_2$. *Right:* Model 2: The second input $X_2$ has negligible unique information contribution, but its synaptic strength $w_2$ modulates the synergistic term, associated to the modification of information the neuron performs [47].

coupled neurons (Fig. 3). A large amount of work has been devoted to applying information theoretic measures for quantifying directed pairwise information transfer between nodes in dynamic networks and neural circuits [50]. Transfer entropy (TE) [51] is the agreed upon measure to estimate the directed transfer of information between pairs of neurons [52, 53], and is sometimes approximated by linear Granger causality. Intuitively, TE between a process $X$ and a process $Y$ measures how much the past of $X$, $X^-$, can help predicting the future of $Y$, $Y^+$, accounting for its past $Y^-$. Although TE quantifies how much information is transferred between neurons, it does not shed light on the computation emerging from the interaction of $X^-$ and $Y^-$. Simply put, the information transferred from $X^-$ could enter $Y^+$, independently of the past state $Y^-$, or it could be fused in a non-trivial way with the information in the state in $Y^-$[54, 55]. This is reflected in the following decomposition: $TE(X \to Y) = I(Y^+{:}X^-|Y^-) = U(Y^+{:}X^- \setminus Y^-) + S(Y^+{:}X^-, Y^-)$. Here, we show that PID uncovers the fusion (or modification [47]) of information implied by $S(Y^+{:}X^-, Y^-)$, beyond merely detecting the directed functional interactions estimated by traditional TE methods (Appendix F, for details). To investigate this, we simulated chaotic networks of rate neurons with an a-priori causal structure consisting of two sub-networks $X$ and $Y$ (Fig. 3a, see [56] for more details on causal analyses of the network model). The sub-network $X$ is a Rossler attractor of three neurons obeying the dynamical equations:

$$\begin{cases} \dot{X}_1 = -X_2 - X_3 \\ \dot{X}_2 = X_1 + \alpha X_2 \\ \dot{X}_3 = \beta + X_3(X_1 - \gamma) \end{cases}$$

where $\{\alpha, \beta, \gamma\} = \{0.2, 0.2, 5.7\}$. Neurons within the sub-network $Y$ obey the dynamical equations

$$\dot{Y} = -\lambda Y + 10 \tanh(J_{YX} X + J_{YY} Y)$$

where $J_{YX}$ is the weight matrix of sparse projections from $X$ to $Y$, sampled as independent Bernoulli variables, and $J_{YY}$ is the recurrent weight matrix of the $Y$ subnetwork, sampled as independent Gaussian variables with standard deviation $g = 1.5$. No projections exist from $Y$ to $X$. We refer to $X$ and $Y$ as the upstream and downstream sub-networks. We simulated time series from this network (exhibiting chaotic dynamics Fig. 3a) and estimated the PID as unique, redundant, and synergistic contribution of neuron $i$ and neuron $j$ at time $t$ in shaping the future of neuron $j$ at time $t+1$. For each pair of neurons $Z_i, Z_j \in \{X_1, X_2, Y_1, Y_2\}$ we treated $(Z_i^t, Z_j^t, Z_j^{t+1})_{t=1}^T$ as iid samples and ran PID on these triplets ($i, j$ represent rows and columns in Fig. 3b-d). The PID uncovered the functional architecture of the network and further revealed non-trivial interactions between neurons belonging to the different subnetworks, encoded in the four matrices $S, R, U_1, U_2$ (details in Fig. 3d). The sum of the relevant unique information terms and the synergies was found to be consistent with the TE (Fig. 3c, TE equal to $S + U_1$, up to estimation bias). The TE itself captured the network effective connectivity, consistent with previous results [53, 56].

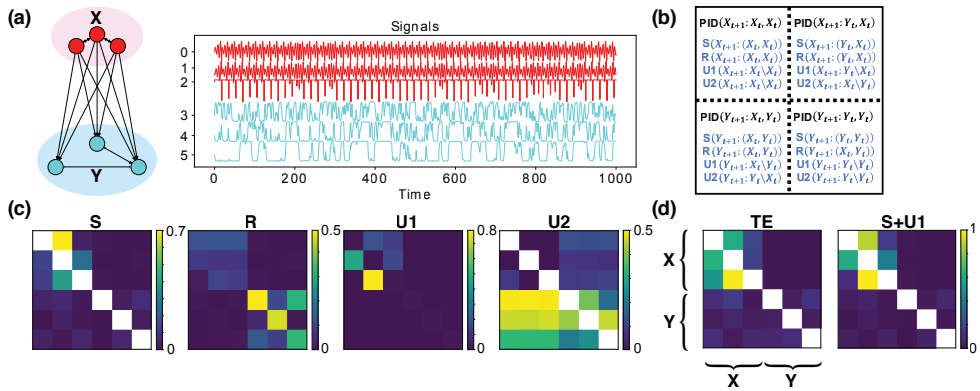


Figure 3: **PID uncovers the effective connectivity and the algorithmic interactions in a chaotic network of rate neurons. a**: Schematics of recurrent network architecture (left) and representative activity (right). **b**: Schematic of the PID decomposition of network nodes. **c:** PID decomposition into $S, R, U_1, U_2$ for the rate network. The future activity of $Y$ neurons (bottom left block) depends on a synergistic interaction ($S$) of $X$ and $Y$ past activities, plus the unique information in their past ($U_1$ and $U_2$, respectively). The future of $X$ neurons only depends on unique information in the past of $X$ neurons and their synergistic interactions. The interactions between the $X$ and $Y$ subnetworks only contain synergistic information regarding the future of $Y$ but no redundant information; the latter is only present in the interactions confined within each subnetwork. **d**: The transfer entropy (TE), estimated via IDTxl [57], recovers the sum $S + U_1$.

**Towards algorithm identification in trained recurrent neural networks.** A fundamental goal in neuroscience is to understand the computational mechanisms emerging from the collective interactions of recurrent neural circuits leading to cognitive function and behavior. Here, we show that PID opens a new window for assessing how specific computations arise from recurrent neural interactions. Unlike MI or TE, the PID quantifies the alternative ways in which a neuron determines the information in its output from its inputs, and thus can be a sensitive marker of different computational strategies. We here trained RNNs as models of cortical circuits [58] and used the PID to elucidate how the computations emerging from recurrent neural interactions contribute to task performance. We trained RNNs to solve a generalized version of the classic XOR classification problem with target labels corresponding to odd vs. even mixture components (Fig. 4a). Stimuli were presented for one time step ($t = 0$) and the network was trained to report the decision at $t = 10$. By tracking the temporal trajectories of the hidden layer activity we found that the network recurrent dynamics (represented as unfolded in time in Fig. 4b) progressively pulls the two input classes in opposite directions along the output weights (see Appendix). We used PID to dissect how the computation emerged from the recurrent interaction between hidden neurons, and to elucidate whether all successfully trained networks had a similar profile in terms of the PID terms. If so, this hints at a single computational strategy across these networks. If not, it is safe to assume that task performance is reached via different mechanisms, despite identical network architecture and training algorithm.

We found that on average across multiple networks S, R, and U rose with task difficulty (Fig. 4c), yet at all difficulties, individual networks differed strongly with respect to the ratio $S/U$, i.e. there were networks with larger average synergy across neuron pairs compared to the average unique information, and vice versa. For simple networks like the ones used here, one can inspect receptive fields to understand the reason for this differential behaviour (Fig. 4d-e). Indeed, networks with high average unique information displayed 'grandmother-cell'-like neurons, that would alone classify a large parts of the sample space, while in networks with higher average synergy such cells were absent (Fig. 4d). The emergence of these 'grandmother-cell'-like receptive fields is due to the recurrent dynamics. While in a feedforward architecture ($W_{XX} = 0$) hidden layer receptive fields are captured by hyperplanes in input space, in the RNN the receptive fields are time dependent, where later times are interpreted as deeper layers (Fig. 4b) and thus can capture highly non-linear features in input space. The advantage of PID versus a manual inspection of receptive fields is

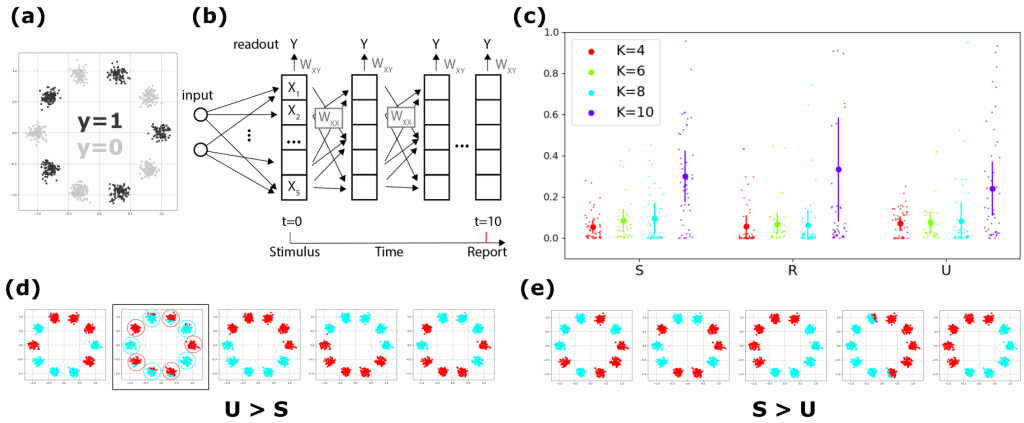


Figure 4: **PID of RNNs trained to solve generalized XOR problem. a**: Input data drawn from a 2D Gaussian Mixture Model with $K$ mixture components $X \sim \sum_{k=1}^{K} \frac{1}{K}\mathcal{N}(X|\mu_k, \sigma I)$ (the means $\mu_{1:K}$ lie at the vertices of a regular K-gon centered on the origin, $\mu_k = (\sin\frac{2\pi k}{K}, \cos\frac{2\pi k}{K})$; grey and black dots represent the two class labels). **b**: Two layer network with 2D input layer, 5 recurrently connected hidden neurons $X$ and one readout neuron $Y$; RNN activity unfolds in time (horizontal axis). The input is presented at time $t = 0$, then withdrawn, and the RNN is trained with BPTT to report the decision at $t = 10$. In this representation, layers correspond to time-steps and weights $W_{XX}$ are shared between layers. **c**: PID between output $Y(t)$ and pairs of hidden neurons $X_i(t-1), X_j(t-1)$ for $t = 10$ yielding $S, R, U_1, U_2$ (distribution over 1000 input samples for each task $K$; 20 networks per task). Harder tasks led to an increase in PID measures. **c**: Example receptive fields for a network with $U > S$ shows emergence of grand-mother cells in the hidden layer (red and blue colors represent hidden neurons outputs; grandmother cell, second from left). **d**: Example receptive fields for a network with $S > U$, relying on higher synergy between neurons to solve the task.

twofold: First, the PID framework abstracts an generalizes descriptions of receptive fields as being e.g. 'grandmother-cell'-like; thus the concept of unique information stays relevant even in scenarios where the concept of a receptive field becomes meaningless, or inaccessible. Second, the quantitative outcomes of a PID rest only on information theory, not specfic assumptions about neural coding or computational strategies, and can be obtained for large numbers of neurons.

## 5 Conclusions

We present a partial information decomposition measure for continuous variables with arbitrary probability densities, thereby extending the popular BROJA PID measure for discrete variables. Extending PID measures to continuous variables drastically broadens the possible applications of the PID framework. This is important as the PID framework itself provides key insights into the way a complex system represents and modifies information in a computation – via asking which variables carry information about a target uniquely (such that it can only be obtained from that variable), redundantly, or only synergistically with other variables. Answering these questions is pivotal to understanding distributed computation in complex systems in general, and neural coding in particular. We believe that the methods presented here will allow PIDs to be extended efficiently in neuroscience for multiple continuous sources with potentially complex dependency structures, as would be common in cellular imaging data or activation properties of brain modules or areas in functional imaging. More generally, the approach we presented here would be relevant for other application domains such as machine learning, biomedical science, finance, and the physical sciences.

# References


[1] Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

[2] Tycho Tax, Pedro AM Mediano, and Murray Shanahan. The partial information decomposition of generative neural network models. *Entropy*, 19(9):474, 2017.

[3] Patricia Wollstadt, Sebastian Schmitt, and Michael Wibral. A rigorous information-theoretic definition of redundancy and relevancy in feature selection based on (partial) information decomposition. *arXiv preprint arXiv:2105.04187*, 2021.

[4] Allison E Goodwell et al. Debates—does information theory provide a new paradigm for earth science? *Water Resources Research*, 56(2):e2019WR024940, 2020.

[5] Benjamin Flecker, Wesley Alford, John M Beggs, Paul L Williams, and Randall D Beer. Partial information decomposition as a spatiotemporal filter. *Chaos*, 21(3):037104, 2011.

[6] Michael Wibral, Joseph T Lizier, and Viola Priesemann. Bits from brains for biologically inspired computing. *Frontiers in Robotics and AI*, 2:5, 2015.

[7] Nicholas M Timme, Shinya Ito, Maxym Myroshnychenko, Sunny Nigam, Masanori Shimono, Fang-Chin Yeh, Pawel Hottowy, Alan M Litke, and John M Beggs. High-degree neurons feed cortical computations. *PLoS computational biology*, 12(5):e1004858, 2016.

[8] Michael Wibral, Viola Priesemann, Jim W Kay, Joseph T Lizier, and William A Phillips. Partial information decomposition as a unified approach to the specification of neural goal functions. *Brain and cognition*, 112:25–38, 2017.

[9] Jim W Kay, WA Phillips, Jaan Aru, Bruce P Graham, and Matthew E Larkum. A Bayesian decomposition of BAC firing as a mechanism for apical amplification in neocortical pyramidal neurons. *bioRxiv*, page 604066, 2019.

[10] Itay Gat and Naftali Tishby. Synergy and redundancy among brain cells of behaving monkeys. In *Advances in neural information processing systems*, pages 111–117, 1999.

[11] Naama Brenner, Steven P Strong, Roland Koberle, William Bialek, and Rob R de Ruyter van Steveninck. Synergy in a neural code. *Neural computation*, 12(7):1531–1552, 2000.

[12] Elad Schneidman, William Bialek, and Michael J Berry. Synergy, redundancy, and independence in population codes. *Journal of Neuroscience*, 23(37), 2003.

[13] Rodrigo Quian Quiroga and Stefano Panzeri. Extracting information from neuronal populations: information theory and decoding approaches. *Nature Reviews Neuroscience*, 10(3):173–185, 2009.

[14] Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014.

[15] D Blackwell. Comparison of experiments. Proc. 2nd Berkeley Symp. Math. Stats. and Probability, 1951.

[16] Moshe Leshno and Yishay Spector. An elementary proof of Blackwell's theorem. *Mathematical Social Sciences*, 25(1):95–98, 1992.

[17] Erik Torgersen. *Comparison of statistical experiments*. Cambridge University Press, 1991.

[18] Lucien Le Cam. Comparison of experiments: A short review. *Lecture Notes-Monograph Series*, pages 127–138, 1996.

[19] Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, and Jürgen Jost. Shared information—new insights and problems in decomposing information in complex systems. In Thomas Gilbert, Markus Kirkilionis, and Gregoire Nicolis, editors, *Proceedings of the European Conference on Complex Systems 2012*, pages 251–269, Cham, 2013. Springer International Publishing.

[20] Robin AA Ince. Measuring multivariate redundant information with pointwise common change in surprisal. *Entropy*, 19(7):318, 2017.

[21] Conor Finn and Joseph T Lizier. Pointwise partial information decomposition using the specificity and ambiguity lattices. *Entropy*, 20(4):297, 2018.

[22] Malte Harder, Christoph Salge, and Daniel Polani. Bivariate measure of redundant information. *Physical Review E*, 87(1):012130, 2013.

[23] Johannes Rauh. Secret sharing and shared information. *Entropy*, 19(11):601, 2017.

[24] Abdullah Makkeh, Aaron J Gutknecht, and Michael Wibral. Introducing a differentiable measure of pointwise shared information. *Physical Review E*, 103(3):032149, 2021.

[25] Aaron J Gutknecht, Michael Wibral, and Abdullah Makkeh. Bits and pieces: Understanding information decomposition from part-whole relationships and formal logic. *arXiv preprint arXiv:2008.09535*, 2020.

[26] Pradeep Kr Banerjee, Johannes Rauh, and Guido Montúfar. Computing the unique information. In *ISIT*, 2018.

[27] Abdullah Makkeh, Dirk Oliver Theis, and Raul Vicente. Bivariate partial information decomposition: The optimization perspective. *Entropy*, 19(10):530, 2017.

[28] Abdullah Makkeh, Dirk Oliver Theis, and Raul Vicente. Broja-2pid: A robust estimator for bivariate partial information decomposition. *Entropy*, 20(4):271, 2018.

[29] Johannes Rauh, Maik Schünemann, and Jüergen Jost. Properties of unique information. *arXiv preprint arXiv:1912.12505*, 2019.

[30] Adam B Barrett. Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems. *Physical Review E*, 91(5), 2015.

[31] Gabriel Schamberg and Praveen Venkatesh. Partial Information Decomposition via Deficiency for Multivariate Gaussians. *arXiv preprint arXiv:2105.00769*, 2021.

[32] M Sklar. Fonctions de repartition a n dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8:229–231, 1959.

[33] Harry Joe. *Multivariate Models and Multivariate Dependence Concepts*. CRC Press, May 1997.

[34] Gal Elidan. Copulas in machine learning. In *Copulae in mathematical and quantitative finance*, pages 39–60. Springer, 2013.

[35] Rafael S Calsaverini and Renato Vicente. An information-theoretic approach to statistical dependence: Copula information. *EPL (Europhysics Letters)*, 88(6):68003, 2009.

[36] Jian Ma and Zengqi Sun. Mutual information is copula entropy. *Tsinghua Science & Technology*, 16(1):51–54, 2011.

[37] Tim Bedford and Roger M Cooke. Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial intelligence*, 32(1-4):245–268, 2001.

[38] Kjersti Aas, Claudia Czado, Arnoldo Frigessi, and Henrik Bakken. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and economics*, 44(2):182–198, 2009.

[39] Claudia Czado. Pair-copula constructions of multivariate copulas. In *Copula theory and its applications*, pages 93–109. Springer, 2010.

[40] Thomas Nagler and Claudia Czado. Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. *Journal of Multivariate Analysis*, 151:69–89, 2016.

[41] Roger B Nelsen. *An Introduction to Copulas*. Springer Science & Business Media, June 2007.

[42] Thomas Nagler and Thibault Vatter. pyvinecopulib, November 2020.

[43] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *ICLR 2016*, 2016.

[44] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *NIPS 2015*.

[45] George Tucker, Dieterich Lawson, Shixiang Gu, and Chris J Maddison. Doubly reparameterized gradient estimators for Monte Carlo objectives. *ICLR*, 2019.

[46] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR 2015*.

[47] Joseph T Lizier, Benjamin Flecker, and Paul L Williams. Towards a synergy-based approach to measuring information modification. In *2013 IEEE Symposium on Artificial Life (ALIFE)*, pages 43–51. IEEE, 2013.

[48] Matteo Carandini and David J Heeger. Normalization as a canonical neural computation. *Nat. Rev. Neurosci.*, 13(1):51–62, November 2011.

[49] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. *ICLR*, 2017.

[50] Andrew T Reid, Drew B Headley, Ravi D Mill, Ruben Sanchez-Romero, Lucina Q Uddin, Daniele Marinazzo, Daniel J Lurie, Pedro A Valdés-Sosa, Stephen José Hanson, Bharat B Biswal, et al. Advancing functional connectivity research from association to causation. *Nature neuroscience*, 22(11):1751–1760, 2019.

[51] Thomas Schreiber. Measuring information transfer. *Physical review letters*, 85(2):461, 2000.

[52] Raul Vicente, Michael Wibral, Michael Lindner, and Gordon Pipa. Transfer entropy—a model-free measure of effective connectivity for the neurosciences. *Journal of computational neuroscience*, 30(1):45–67, 2011.

[53] Leonardo Novelli and Joseph T Lizier. Inferring network properties from time series using transfer entropy and mutual information: validation of multivariate versus bivariate approaches. *Network Neuroscience*, 5(2):373–404, 2021.

[54] Paul L Williams and Randall D Beer. Generalized measures of information transfer. *arXiv preprint arXiv:1102.1507*, 2011.

[55] Michael Wibral, Conor Finn, Patricia Wollstadt, Joseph T Lizier, and Viola Priesemann. Quantifying information modification in developing neural networks via partial information decomposition. *Entropy*, 19(9):494, 2017.

[56] Amin Nejatbakhsh, Francesco Fumarola, Saleh Esteki, Taro Toyoizumi, Roozbeh Kiani, and Luca Mazzucato. Predicting perturbation effects from resting state activity using functional causal flow. *bioRxiv*, 2020.

[57] Patricia Wollstadt, Joseph T Lizier, Raul Vicente, Conor Finn, Mario Martinez-Zarzuela, Pedro Mediano, Leonardo Novelli, and Michael Wibral. Idtxl: The information dynamics toolkit xl: a python package for the efficient analysis of multivariate information dynamics in networks. *Journal of Open Source Software*, 4(34):1081, 2019.

[58] Valerio Mante, David Sussillo, Krishna V Shenoy, and William T Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *nature*, 503(7474):78–84, 2013.

[59] Tom Rainforth, Adam Kosiorek, Tuan Anh Le, Chris Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh. Tighter variational bounds are not necessarily better. In *ICML*, 2018.

[60] Murray Rosenblatt. Remarks on a multivariate transformation. *The annals of mathematical statistics*, 23(3):470–472, 1952.

[61] Michael Wibral, Nicolae Pampu, Viola Priesemann, Felix Siebenhühner, Hannes Seiwert, Michael Lindner, Joseph T Lizier, and Raul Vicente. Measuring information-transfer delays. *PloS one*, 8(2):e55809, 2013.

[62] Pedro AM Mediano, Fernando Rosas, Robin L Carhart-Harris, Anil K Seth, and Adam B Barrett. Beyond integrated information: A taxonomy of information dynamics phenomena. *arXiv preprint arXiv:1909.02297*, 2019.

[63] Heysem Kaya, Pmar Tüfekci, and Fikret S Gürgen. Local and global learning methods for predicting power of a combined gas & steam turbine. In *Proceedings of the international conference on emerging trends in computer and electronics engineering icetcee*, pages 13–18, 2012.

[64] Pınar Tüfekci. Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. *International Journal of Electrical Power & Energy Systems*, 60:126–140, 2014.

# Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] Only for the code.

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Its clear form the text
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We reported the full distributions.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] We used a publicly available dataset in the appendix.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]