Max-Planck-Institut für Mathematik in den Naturwissenschaften Leipzig

Computing the Unique Information

by

Pradeep Kumar Banerjee, Johannes Rauh,and Guido Montúfar

Preprint no.: 73 2017

Computing the Unique Information

Pradeep Kr. Banerjee[1], Johannes Rauh[1], and Guido Montúfar[1,2]

[1]Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany
[2]Departments of Mathematics and Statistics, UCLA, USA

September 25, 2017

Abstract

Given a set of predictor variables and a response variable, how much information do the predictors have about the response, and how is this information distributed between unique, complementary, and shared components? Recent work has proposed to quantify the unique component of the decomposition as the minimum value of the conditional mutual information over a constrained set of information channels. We present an efficient iterative divergence minimization algorithm to solve this optimization problem with convergence guarantees, and we evaluate its performance against other techniques.

Keywords: Positive information decomposition, mutual information, alternating divergence minimization

1 Introduction

When Shannon proposed to use entropy in order to quantify information, he had in mind a very specific setting of communication over a noisy channel. Since then, the use of entropic quantities has been greatly expanded, with successful applications in statistical physics, complex systems, neural networks, and machine learning. In particular, transfer entropy is used as a tool to study causality in dynamical systems (Schreiber 2000), and mutual information as a criterion in feature selection (Vergara and Estévez 2014). In many applications, the effort of estimating entropic quantities, which may be considerable, is out-weighed by the performance gain.

Despite the success of information theory, there are still many open questions about the nature of information. In particular, since information is not a conservation quantity, it is difficult to describe how information is distributed over composite systems. Clearly, different subsystems may have exclusive (or unique) information, or they may have redundant information. Moreover, synergy effects complicate the analysis: It may happen that some information is not known to any subsystem but can only be recovered from knowledge of the entire system. An example is the checksum of several digits, which can only be computed when all digits are known. Such synergy effects abound in cryptography, where the goal is that the encrypted message alone contains no information about the original message without knowledge of the key. It is also suspected that synergy plays a major role in the neural code (Latham and Nirenberg 2005).

In spite of their conceptual importance, so far there is no consensus on how to measure or extract the unique, shared, and synergistic portions of joint information, even though there have been several proposals (e.g., McGill 1954, Bell 2003). Williams and Beer (2010) proposed a principled approach to decomposing the total mutual information of a system into positive components corresponding to a lattice of subsystems. This was followed up by the axiomatic approach from Bertschinger et al. (2014), quantifying the unique, shared, and synergistic information based on ideas from decision theory. We focus on their definitions, detailed in Section 2.

Although theoretically promising, these definitions involve an optimization problem that complicates experimentation and applications. Indeed, Bertschinger et al. (2014) note that the optimization problem, although convex, can be very ill conditioned, and difficulties have been reported, with out-of-the-box methods or other implementations either failing to produce the correct results, or taking extremely long to converge. In Section 3, we derive an alternating divergence minimization algorithm for solving the union information optimization problem. Our algorithm is conceptually similar to the Blahut-Arimoto algorithm (Blahut 1972), which is used for computing the capacity of an information channel as the maximum of the mutual information over a constrained set of joint probability distributions. However, there are significant differences, especially in relation to the nature of the constraints.

In Section 4, we run numerical experiments comparing our algorithm with other computation approaches. Our algorithm consistently returns accurate solutions, while still requiring far less computation time than other methods. We wrap up and give a brief outlook in Section 5. Relevant notations are included in Appendix B.

2 Quantifying the unique information

While much research has focused on finding an information measure for a single aspect (like synergy), the seminal paper by Williams and Beer (2010) introduced an approach to find a complete decomposition of the total mutual information $I(S; Y_1, \ldots, Y_k)$ about a signal S that is distributed among a family of random variables $Y_1, \ldots, Y_k$. Here, the total mutual information is expressed as a sum of non-negative terms with a well-defined interpretation corresponding to the different ways in which information can have aspects of redundant, unique, or synergistic information. For example, in the case $k = 2$, writing $Y_1 \equiv Y$ and $Y_2 \equiv Z$, the decomposition is of the form

$$I(S;Y,Z) = \underbrace{SI(S;Y,Z)}_{\text{shared (redundant)}} + \underbrace{CI(S;Y,Z)}_{\text{complementary (synergistic)}} + \underbrace{UI(S;Y\backslash Z)}_{\text{unique } Y \text{ wrt } Z} + \underbrace{UI(S;Z\backslash Y)}_{\text{unique } Z \text{ wrt } Y}, \quad (1)$$

where $SI(S;Y,Z)$, $CI(S;Y,Z)$, $UI(S;Y\backslash Z)$, and $UI(S;Z\backslash Y)$ are nonnegative functions that depend continuously on the joint distribution of (S,Y,Z). Furthermore, these functions are required to satisfy the intuitive equations

$$\begin{aligned} I(S;Y) &= SI(S;Y,Z) + UI(S;Y\backslash Z), \\ I(S;Z) &= SI(S;Y,Z) + UI(S;Z\backslash Y). \end{aligned} \quad (2)$$

Combining these equations, it follows that the co-information can be written as the difference of redundant and synergistic information, which agrees with the general interpretation of co-information:

$$CoI(S;Y;Z) := I(S;Y) - I(S;Y|Z) = SI(S;Y,Z) - CI(S;Y,Z). \quad (3)$$

Similarly, the conditional mutual information satisfies

$$I(S;Y|Z) = I(S;Y,Z) - I(S;Z) = CI(S;Y,Z) + UI(S;Y\backslash Z). \tag{4}$$

The decomposition is illustrated in Figure 1a.

Although the above framework is very appealing, there is no general agreement on how to define the corresponding functions for shared, unique, and synergistic information. When Williams and Beer (2010) presented their information decomposition framework, they also proposed specific measures. However, their functions have been criticized as overestimating redundant and synergistic information, while underestimating unique information (Griffith and Koch 2014)[1]. Another proposal of information measures for the bivariate case ($k = 2$) that involves information-geometric ideas is presented in Harder et al. (2013).

Here we follow the approach from Bertschinger et al. (2014) and use the functions SI, UI and CI defined there, since it is the most principled approach, based on ideas from decision theory and having an axiomatic characterization. This approach covers only $k = 2$, but situations with larger k can be analyzed by grouping the variables. The decomposition is based on the idea that unique and shared information, $UI(S;Y\backslash Z)$ and $SI(S;Y,Z)$, should depend only on the marginal distributions of the pairs (S,Y) and (S,Z). It gives similar values as the functions defined in Harder et al. (2013). Incidentally, the bivariate synergy measure derived from this approach agrees with the synergy measure defined by Griffith and Koch (2014) for arbitrary k.

For some finite state spaces $\mathcal{Y}, \mathcal{Z}, \mathcal{S}$, let $\mathbb{P}_{\mathcal{S}\times\mathcal{Y}\times\mathcal{Z}}$ be the set of all joint distributions of (S,Y,Z). Given $P \in \mathbb{P}_{\mathcal{S}\times\mathcal{Y}\times\mathcal{Z}}$, let

$$\Delta_P := \big\{Q \in \mathbb{P}_{\mathcal{S}\times\mathcal{Y}\times\mathcal{Z}} \colon Q_{SY}(s,y) = P_{SY}(s,y) \text{ and } Q_{SZ}(s,z) = P_{SZ}(s,z)\big\} \tag{5}$$

denote the set of joint distributions of (Y,Z,S), that have the same marginals on (S,Y) and (S,Z) as P. Bertschinger et al. (2014) define the unique information that Y conveys about S with respect to Z as

$$UI(S;Y\backslash Z) := \min_{Q\in\Delta_P} I_Q(S;Y|Z). \tag{6}$$

By (1) and (2), specifying (6) fixes the other three functions in (1), which are then

$$UI(S;Z\backslash Y) := \min_{Q\in\Delta_P} I_Q(S;Z|Y), \tag{7}$$

$$SI(S;Y,Z) := I(S;Y) - \min_{Q\in\Delta_P} I_Q(S;Y|Z) = \max_{Q\in\Delta_P} CoI_Q(S;Y;Z), \tag{8}$$

$$CI(S;Y,Z) := SI(S;Y,Z) - CoI(S;Y;Z) = I(S;Y,Z) - \min_{Q\in\Delta_P} I_Q(S;Y,Z). \tag{9}$$

Since Δ_P is compact and the mutual information is a continuous function, these maxima and minima are all well-defined.

Example 1 (Census data)**.** We illustrate the mutual information decomposition by a brief analysis of the US 1994 census income data set (Lichman 2013). Here the task is to relate a list of predictor variables with a binary response variable. The predictors include: age (continuous variable divided into 4 categories: < 24, 24–35, 36–50, > 50), sex (binary: Male,

[1]For example, in the case of two independent variables Y, Z and a copy $S = (Y,Z)$, the measure of Williams and Beer assigns 0 bit of unique information to Y and Z, and all available information is interpreted as either redundant or synergistic.

Female), race (5 values: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black), education level (4 values: Basic-schooling, Attended-HS, Bachelors-and-above, Vocational), occupation (14 values: Tech-support, Craft-repair, Other-service, etc.), and hours-per-week (continuous variable grouped into 2 categories:$\leq$ 40, $>$ 40). The response is the yearly income, with values $>$ 50K and $\leq$ 50K.

Figure 1 shows the evaluation of the information decomposition (6)–(9) on this data set, computed using the algorithm that we will present in Section 3. We see, for instance, that most of the information that race and occupation have for predicting income, is uniquely in the occupation. On the other hand, education and sex have about equally large shared and complementary information. Age has a large unique information about income with respect to sex, as does occupation with respect to hours-per-week. These results appear quite reasonable. They illustrate how the information decomposition allows us to obtain a fine grained quantitative analysis of the relationships between predictors and responses.

3 Computing the information decomposition

We only need to solve one of the optimization problems (6)–(9) in order to obtain all the terms in the information decomposition. Actually we can solve another equivalent optimization problem, namely for a function called the *union information*, defined as

$$I_\cup(S;Y,Z) := I(S;Y,Z) - CI(S;Y,Z) = \min_{Q\in\Delta_P} I_Q(S;Y,Z). \tag{10}$$

We first note that this is a convex minimization problem:

Proposition 1. *The optimization problems* (6) *and* (10) *are convex. Moreover,* (6)–(10) *are equivalent in that a distribution Q solves one of them if and only if it solves all of them.*

Proof. The equivalence of the optimization problems is clear by construction. Moreover,

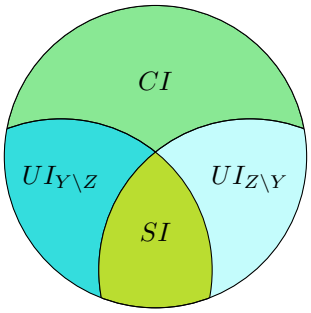
$$\min_{Q\in\Delta_P} I_Q(S;Y,Z) = H(S) - \max_{Q\in\Delta_P} H_Q(S|Y,Z),$$

since $H(S)$ is constant on Δ_P. Thus, convexity of the optimization problems follows from the fact that $H_Q(S|Y,Z)$ is concave with respect to the joint distribution (Cover and Thomas 1991). □

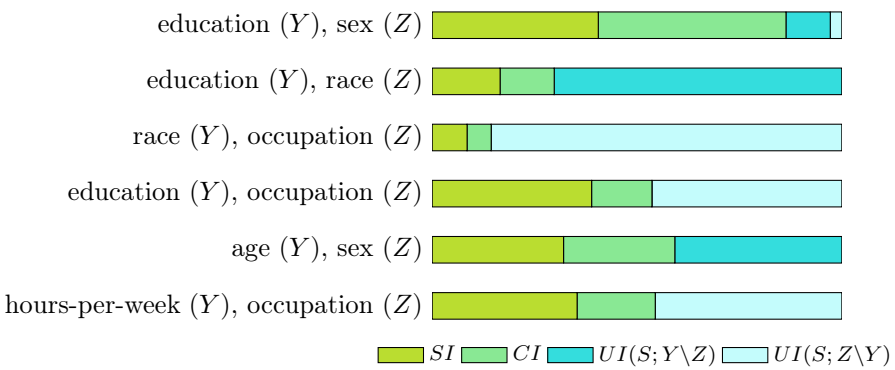
The target function $I_Q(S;Y,Z)$ is convex, but not strictly convex. For certain P, the solution is not unique, and close by, the conditioning number of the optimization problem becomes very bad (Bertschinger et al. 2014) Even though the optimizing $Q \in \Delta_P$ may not be unique, convexity guarantees that any local optimizer is a global optimizer and that the optimum value is unique.

Double minimization formulation. The mutual information can be written in the form $I_P(S;Y,Z) = \min_{R_{YZ}\in\mathbb{P}_{\mathcal{Y}\times\mathcal{Z}}} D(P\|P_S R_{YZ})$, with the minimum attained at $R^*_{YZ} = P_{YZ}$ (see, e.g., Csiszár and Körner 2011; eq. (8.7)). With this expression, we can rewrite (10) as a double minimization problem:

$$I_\cup(S;Y,Z) = \min_{Q\in\Delta_P} \min_{R_{YZ}\in\mathbb{P}_{\mathcal{Y}\times\mathcal{Z}}} D(Q\|Q_S R_{YZ}). \tag{11}$$



(a)



(b)

Figure 1: (a) Illustration of the decomposition (1) of the mutual information of a pair (Y, Z) and S into the complementary (synergistic) information CI, the unique information UI of X with respect to Y and conversely, and the shared (redundant) information SI. (b) Information decomposition evaluated on the US census data set: The attributes Y and Z predict the income level (S), i.e., whether a person makes over USD 50K per year. In this figure, each bar is normalized by the total mutual information $I(S; Y, Z)$, to highlight the relative values of SI, CI and UI.

Conditional probability formulation. The minimization problem (11) can also be studied and solved over a set of conditional probabilities, instead of the set Δ_P that consists of joint probability distributions. In fact, Δ_P is in bijection with $\Delta_{P,S} := \bigtimes_{s\in\mathcal{S}} \Delta_{P,s}$, where

$$\Delta_{P,s} := \big\{Q_{YZ} \in \mathbb{P}_{\mathcal{Y}\times\mathcal{Z}} \colon Q_Y(y) = P_{Y|S}(y|s) \text{ and } Q_Z(z) = P_{Z|S}(z|s)\big\}, \quad s \in \mathcal{S}. \tag{12}$$

The set $\Delta_{P,s}$ is the linear family of probability distributions of (Y, Z) defined by fixing the marginal distributions of Y and Z to be those of $P_{YZ|s}$. Any joint distribution $Q \in \Delta_P$ has the form $Q = P_S Q_{YZ|S}$ with $Q_{YZ|S} \in \Delta_{P,S}$. In turn, the optimization problem (11) can be written as

$$\begin{aligned} I_\cup(S;Y,Z) &= \min_{Q_{YZ|S}\in\Delta_{P,S}} \min_{R_{YZ}\in\mathbb{P}_{\mathcal{Y}\times\mathcal{Z}}} D(P_S Q_{YZ|S} \| P_S R_{YZ}) \\ &= \min_{R_{YZ}\in\mathbb{P}_{\mathcal{Y}\times\mathcal{Z}}} \sum_s P_S(s) \min_{Q_{YZ|s}\in\Delta_{P,s}} D(Q_{YZ|s}\|R_{YZ}). \end{aligned} \tag{13}$$

Alternating divergence minimization. An alternating algorithm iteratively fixes one of the two free variables and optimizes over the other. Starting with some $R_{YZ}^{(0)} \in \mathbb{P}_{\mathcal{Y}\times\mathcal{Z}}$, recursively define

$$Q_{YZ|s}^{(i+1)} = \operatorname*{arg\,min}_{Q_{YZ|s}\in\Delta_{P,s}} D(Q_{YZ|s}\|R_{YZ}^{(i)}) \quad \text{for each } s \in \mathcal{S}, \tag{14a}$$

$$R_{YZ}^{(i+1)} = \operatorname*{arg\,min}_{R_{YZ}\in\mathbb{P}_{\mathcal{Y}\times\mathcal{Z}}} D(P_S Q_{YZ|S}^{(i+1)} \| P_S R_{YZ}). \tag{14b}$$

With suitable initialization, this iteration converges to a pair attaining the global optimum:

Theorem 1. *Given $P \in \mathbb{P}_{\mathcal{S}\times\mathcal{Y}\times\mathcal{Z}}$ and an initial value $R_{YZ}^{(0)} \in \mathbb{P}_{\mathcal{Y}\times\mathcal{Z}}$ of full support, the iteration* (14) *converges. The limit $\lim_{i\to\infty} P_S Q_{YZ|S}^{(i)}$ is a global optimum of the minimization problem* (13)*.*

Proof. For any P, the subsets Δ_P and $\{P_S R_{YZ} \colon R_{YZ} \in \mathbb{P}_{\mathcal{Y}\times\mathcal{Z}}\}$ of $\mathbb{P}_{\mathcal{S}\times\mathcal{Y}\times\mathcal{Z}}$ are compact and convex. The statement then follows from (Csiszár and Shields 2004; Corollary 5.1). □

The optimization problem is convex but not strictly convex and it may have several global optimizers. Still the sequence of divergence values converges to the minimum (13).

Implementation. Pseudocode for the alternating divergence minimization algorithm for computing the union information (admUI) is in Algorithm 1. We next discuss the two steps separately.

Step 1 Using the product structure of $\Delta_{P,S}$, we can break the computation of $Q_{YZ|S}^{(i+1)}$ into smaller problems as given in (14a). This kind of minimization problem can be solved using generalized iterative scaling (pseudocode in Algorithm 2):

Theorem 2. *The nonnegative functions b_n on $\mathcal{Y}\times\mathcal{Z}$ defined recursively by*

$$b_0(y,z) = R_{YZ}(y,z), \quad b_{n+1}(y,z) = b_n(y,z) \left[\frac{P_{Y|S}(y|s)}{\sum_z b_n(y,z)}\right]^{1/2} \left[\frac{P_{Z|S}(z|s)}{\sum_y b_n(y,z)}\right]^{1/2}, \tag{15}$$

converge to $\arg\min_{Q_{YZ|s}\in\Delta_{P,s}} D(Q_{YZ|s}\|R_{YZ})$, that is, the I-projection of R_{YZ} to $\Delta_{P,s}$.

Proof. The set $\Delta_{P,s}$ is a linear family with statistics δ_y, $y \in \mathcal{Y}$ and δ_z, $z \in \mathcal{Z}$. The claim follows from (Csiszár and Shields 2004; Theorem 5.2). □

Step 2 Using the variational representation for $I_{P_S Q_{YZ|S}}(S; Y, Z)$ discussed prior (11), we can write the minimizer of (14b) in closed form as

$$R_{YZ}^{(i+1)}(y, z) = \sum_{s \in \mathcal{S}} P_S(s) Q_{YZ|S}^{(i+1)}(y, z|s). \tag{16}$$

Stopping criterion The iteration (14) can be stopped when

$$\max_{y \in \mathcal{Y}, z \in \mathcal{Z}} \log \frac{Q_{YZ|S}^{(i+1)}(y,z|s)}{Q_{YZ|S}^{(i)}(y,z|s)} \leq \epsilon, \quad \text{for all } s \in \mathcal{S}, \tag{17}$$

for some prescribed $\epsilon > 0$. When this condition is satisfied, ϵ is an upper bound on the difference of the current value of the divergence and the minimum (see Csiszár and Shields 2004; Corollary 5.1).

ϵ is a parameter of the algorithm. In our experiments, we chose $\epsilon = 10^{-6}$.

For the I-projection, we propose a heuristic stopping criterion. The iteration (15) can be stopped when the squared distance between subsequent distributions is less than the square of some prespecified ϵ_1. We found $\epsilon_1 = 10^{-2}\epsilon$ a good standard value.

In general, the distributions returned in Step 1 are not exact, and this needs to be accounted for in the stopping criterion. In Appendix A we show that it is possible to guarantee ϵ optimality of the overall optimization, if the outer loop is interrupted when $\max_{y,z,s} \log \tilde{Q}_{YZ|S}^{(i+1)}(y, z|s)/\tilde{Q}_{YZ|S}^{(i)}(y, z|s) \leq \frac{\epsilon}{3}$, and iteration (15) is interrupted when $\|\tilde{\eta}^{(i)} - \eta\|_1 \leq \tilde{Q}_{YZ|S}^{(i)}(y, z|s)\frac{\epsilon}{12}$. Here $\tilde{\eta}^{(i)}$ and η denote the expectation parameters of the current iterate and of the target, respectively.

Modifications. There are various obvious modifications of our algorithms that could contribute to an improved performance. The stopping criterion does not need to be evaluated in every iteration. Evaluating it once every 20 iterations can save about 10% of the total computation time, as we found in numerical experiments. The optimization problems in Step 1 are equivalent to maximum likelihood estimation on an exponential family. They can be solved using negative log likelihood gradient descent, L-BFGS, and others. For large systems, the optimization can be run in parallel for blocks of s values. The stopping criterion discussed previously will work regardless of the iterative optimization method used on Step 1.

Algorithm 1 Alternating divergence minimization for the union information (admUI)

Input: Marginals P_{SY} and P_{SZ}
Output: $Q^* = \arg\min_{Q \in \Delta_P} I_Q(S;Y,Z)$
Initialization: Some $R_{YZ}^{(0)}$ from the interior of $\mathbb{P}_{\mathcal{Y}\times\mathcal{Z}}$. Set $i = 0$.
while not converged **do**
    for all $s \in \operatorname{supp}(P_S)$ **do in parallel**
        $Q_{YZ|s}^{(i+1)} \leftarrow \arg\min_{Q_{YZ|s} \in \Delta_{P,s}} D(Q_{YZ|s} \| R_{YZ|s}^{(i)})$ (Algorithm 2) ▷ Step 1
    end for
    $R_{YZ}^{(i+1)}(y,z) \leftarrow \sum_{s\in\mathcal{S}} P_S(s) Q_{YZ|S}^{(i+1)}(y,z|s)$ ▷ Step 2
    $i \leftarrow i+1$
end while
return $P_S Q_{YZ|S}^{(i)}$

Algorithm 2 The I-projection of R_{YZ} to $\Delta_{P,s}$

Input: Marginals P_{SY} and P_{SZ}, some $s \in \mathcal{S}$, and target distribution R_{YZ}
Output: $Q_{YZ|s}^* = \arg\min_{Q_{YZ|s} \in \Delta_{P,s}} D(Q_{YZ|s} \| R_{YZ})$
Initialization: $b_0(y,z) \leftarrow R_{YZ}(y,z)$. Set $n = 0$.
while not converged **do**
    $b_{n+1}(y,z) \leftarrow b_n(y,z) \left[\frac{P_{Y|S}(y|s)}{\sum_z b_n(y,z)}\right]^{1/2} \left[\frac{P_{Z|S}(z|s)}{\sum_y b_n(y,z)}\right]^{1/2}$
    $n \leftarrow n+1$
end while
return b_n

4 Experiments

Comparison with other methods. We compare the performance of our alternating divergence minimization algorithm admUI against two other optimization methods. We implemented the admUI algorithm in Matlab R2017a as a Matlab executable (MEX). Our first baseline is the general purpose optimizer `fmincon` from the Matlab optimization package. We also compare it against `fmincon` including the derivative and the Hessian. Figure 2 shows the mean of the values of the unique information computed on 250 joint distributions of (S, Y, Z) sampled uniformly at random from the probability simplex. We are interested in the accuracy of the computations and the required computation time as the state spaces increase in size. In terms of accuracy, all methods perform similarly (lower values reflect more accurate outcomes of the minimization). However, our algorithm allows for significant savings in terms of computation time. In fact, the black-box `fmincon` and `fmincon` with only the gradient included failed to give any answer beyond $|\mathcal{Y}| = 12$ in a reasonable amount of time (see last row in Figure 2). We note that for admUI we did not parallelize the computations in Step 1, which we expect will provide additional savings, especially for systems with large $\mathcal{S}$ (last row in the Figure 2).

Accuracy and stopping criterion. To test the accuracy and efficiency of the admUI algorithm for high-dimensional systems, we consider the COPY distribution: Y and Z are independent uniformly distributed random variables and $S = (Y, Z)$. The unique information of Y with respect to S for the COPY distribution is just the mutual information $I(S; Y) = H(Y)$ (in the language of Harder et al. (2013), UI satisfies the *identity property*, as shown by Bertschinger et al. (2014)). Hence we can use this example to test the accuracy of the solutions produced by different optimizers. Table 1 compares the admUI algorithm and `fmincon` (with gradient and Hessian included) in terms of the error and computation times for different cardinalities of $\mathcal{Y}$. We chose $\mathcal{Z} = \mathcal{Y}$ and $\mathcal{S} = \mathcal{Y} \times \mathcal{Y}$ so that overall size of the system scales as $|\mathcal{Y}|^4$. Compared to the admUI, the computation time and error grow at a much faster rate for `fmincon`.

For admUI, we consider the two stopping criteria discussed in Section 3, with several choices of the accuracy parameter ϵ. Stop 1 is the heuristic and Stop 2 is the rigorous method. The stopping criterion was evaluated in every iteration. As can be seen from the table, both criteria allow us to control the error. The heuristic has a lower computational overhead compared to the rigorous stopping criterion. On the other hand, the error bound of the rigorous criterion appears to be somewhat pessimistic, and seems to perform well even with a much larger ϵ.

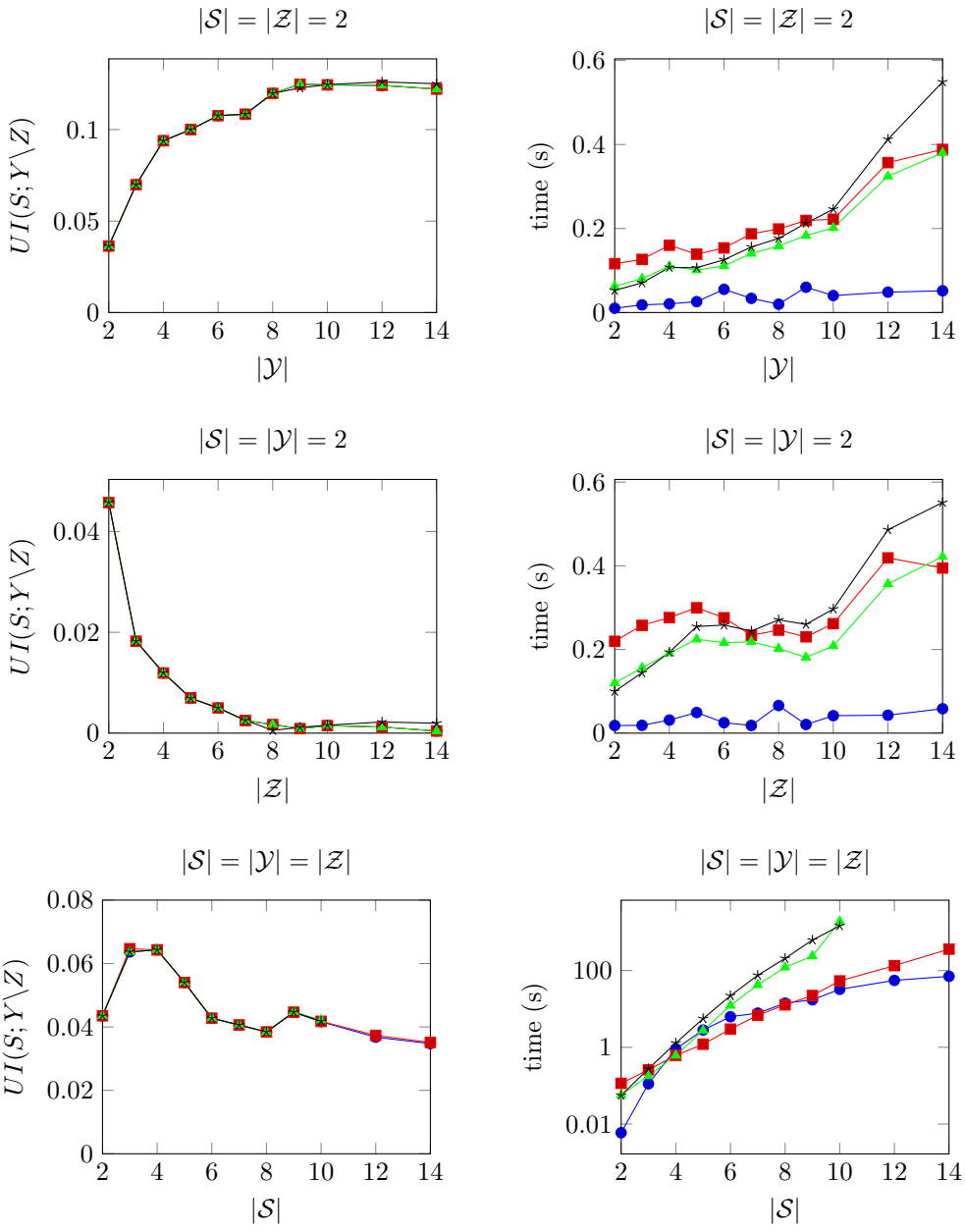


Figure 2: Comparison of the admUI algorithm (—●—) with the general purpose optimizer `fmincon` when including the gradient and Hessian (—■—), only the gradient (—▲—), and when including none (—⋆—). The left panel shows the average values of the computed unique information, $UI(S;Y\backslash Z)$ for 250 distributions sampled uniformly at random from the probability simplex. The right panel shows the average computation (wall-clock) time on an Intel 2.60GHz CPU. Note that the last row corresponds to much larger systems.

Table 1: Comparison of admUI and `fmincon` on the COPY example.

Size	ϵ	admUI				`fmincon`[1]	
		Stop 1		Stop 2			
		Error	Time (ms)	Error	Time (ms)	Error	Time (ms)
2^4	10^{-8}	$1.94 \cdot 10^{-9}$	4.38	$9.16 \cdot 10^{-10}$	$9.03 \cdot 10^{1}$	$9.52 \cdot 10^{-5}$	$2.38 \cdot 10^{2}$
	10^{-5}	$1.97 \cdot 10^{-6}$	5.36	$6.67 \cdot 10^{-7}$	$6.45 \cdot 10^{1}$		
	10^{-3}	$1.09 \cdot 10^{-4}$	4.19	$5.01 \cdot 10^{-5}$	$1.03 \cdot 10^{1}$		
4^4	10^{-8}	$1.63 \cdot 10^{-9}$	11.09	$7.24 \cdot 10^{-10}$	$2.27 \cdot 10^{2}$	$1.50 \cdot 10^{-4}$	$4.17 \cdot 10^{2}$
	10^{-5}	$1.84 \cdot 10^{-6}$	5.77	$5.38 \cdot 10^{-7}$	$2.67 \cdot 10^{2}$		
	10^{-3}	$1.03 \cdot 10^{-4}$	5.06	$4.13 \cdot 10^{-5}$	$2.59 \cdot 10^{2}$		
7^4	10^{-8}	$3.15 \cdot 10^{-9}$	6.23	$4.93 \cdot 10^{-10}$	$2.42 \cdot 10^{3}$	$2.32 \cdot 10^{-4}$	$8.61 \cdot 10^{3}$
	10^{-5}	$1.43 \cdot 10^{-6}$	4.49	$3.71 \cdot 10^{-7}$	$2.41 \cdot 10^{3}$		
	10^{-3}	$0.81 \cdot 10^{-4}$	7.68	$2.89 \cdot 10^{-5}$	$1.97 \cdot 10^{3}$		
10^4	10^{-8}	$2.60 \cdot 10^{-9}$	14.67	$3.71 \cdot 10^{-10}$	$9.38 \cdot 10^{3}$	$3.51 \cdot 10^{-4}$	$4.86 \cdot 10^{5}$
	10^{-5}	$1.18 \cdot 10^{-6}$	12.11	$2.82 \cdot 10^{-7}$	$9.20 \cdot 10^{3}$		
	10^{-3}	$0.66 \cdot 10^{-4}$	11.90	$2.22 \cdot 10^{-5}$	$8.73 \cdot 10^{3}$		

[1] `fmincon` with gradient, Hessian, and options: Algorithm = interior-point, MaxIterations = 10^4, MaxFunctionEvaluations = 10^5, OptimalityTolerance = 10^{-6}, ConstraintTolerance = 10^{-8}.

5 Discussion

We developed an efficient algorithm to compute a decomposition of information in composite systems that was proposed by Bertschinger et al. (2014), but for which the computation had remained a challenge so far. Our algorithm comes with convergence guarantees and a rigorous stopping criterion ensuring ϵ optimality of the solution. We tested the computation time and accuracy of our algorithm against other general purpose constrained convex optimizers. In various experiments, our algorithm showed a very good performance both in terms of computation time and accuracy.

One may ask whether the computational complexity of the function UI prohibits its use in applications, given that already computing or estimating a mutual information is challenging. One major problem when estimating the mutual information is the difficulty in estimating the joint distribution of many variables. In this respect, UI compares well, since $UI(S; Y \setminus Z)$ does not depend on the joint distribution of all variables, but only on the marginal distributions of pairs (S, Y) and (S, Z). In those applications where the main problem is the estimation of the joint distribution given the data at hand, UI is easier to treat than the mutual information.

We hope that our algorithm will contribute means to test the information decomposition on larger systems than was possible so far, and also to use it in settings such as feature selection, robotics (Ghazi-Zahedi and Rauh 2015, Ghazi-Zahedi et al. 2017), and the analysis of multivariate systems, in particular stochastic neural networks, which so far has been pursued only with simpler types of measures Tax et al. (2017).

Appendix

A Stopping criterion

Outer loop with errors. The stopping criterion (17) for the outer loop of Algorithm 1 tests $\max_{s,y,z} \log \frac{Q^{(i+1)}(y,z|s)}{Q^{(i)}(y,z|s)} \leq \epsilon$, which ensures that the objective function has reached a value within ϵ of optimal. We need to describe the behavior of this test when using approximations $\tilde{Q}^{(i)}$ and $\tilde{Q}^{(i+1)}$ instead of the exact distributions $Q^{(i)}$ and $Q^{(i+1)}$. Consider any s, y, z and abbreviate $q^{(i)} \equiv Q^{(i)}(y,z|s)$ and $\tilde{q}^{(i)} \equiv \tilde{Q}^{(i)}(y,z|s)$.

Proposition 2. *Let $\epsilon > 0$. If*

$$|\tilde{q}^{(i)} - q^{(i)}| \leq \tilde{q}^{(i)} \frac{\epsilon}{12}, \quad |\tilde{q}^{(i+1)} - q^{(i+1)}| \leq \tilde{q}^{(i+1)} \frac{\epsilon}{12}, \quad \textit{and} \quad \log \frac{\tilde{q}^{(i+1)}}{\tilde{q}^{(i)}} \leq \frac{\epsilon}{3},$$

then

$$\log \frac{q^{(i+1)}}{q^{(i)}} \leq \epsilon.$$

Proof. By direct evaluation. □

In turn, testing the stopping criterion with $\epsilon_1 \leq \frac{\epsilon}{3}$ allows us to conclude ϵ optimality, if the approximate distributions plugged in are within $\epsilon_0 \leq \min\{\tilde{q}^{(i)}, \tilde{q}^{(i+1)}\}\frac{\epsilon}{12}$ of the actual distributions, in each entry.

Inner loop. Now we want to find a criterion to interrupt the iteration from Algorithm 2 with the guarantee that $|\tilde{q} - q| \leq \epsilon_0$ for some prespecified ϵ_0.

Note that the optimization in Theorem 2 takes place over the set of distributions of the form $\frac{1}{Z(R,q_Y,q_Z)} R(y,z) q_Y(y) q_Z(z)$, where q_Y and q_Z are arbitrary probability distributions over Y and Z respectively, R is the distribution that we want to approximate with a distribution from the linear family $\Delta_{P,s}$, and $Z(R, q_Y, q_Z)$ is the normalizing partition function. This is an exponential family with sufficient statistics $\mathbb{1}_{y'}$, $y' \in Y$, $\mathbb{1}_{z'}$, $z' \in Z$, computing the marginal distributions on Y and Z. (This is similar to an independence model, but with a non uniform reference measure.) The solution to this optimization problem is the unique distribution Q_{YZ} within the exponential family, that is also contained in $\Delta_{P,s}$, meaning that its marginal distributions (which correspond to the expectation parameters) satisfy $Q_Y(y) = \eta_y = P_{Y|S}(y|s)$ and $Q_Z(z) = \eta_z = P_{Z|S}(z|s)$. We want to bound the error $|\tilde{q} - q|$ in terms of the error $|\tilde{\eta} - \eta|$ of the expectation parameters.

Conjecture 1. $\|\tilde{q} - q\|_\infty \leq \|\tilde{\eta} - \eta\|_1$.

Extensive computer experiments seem to confirm that Conjecture 1 is true. Assuming this, the stopping criterion is

$$\begin{aligned}\|\tilde{\eta} - \eta\|_1 = &\sum_{z \in Z \setminus \{1\}} |(\sum_y \frac{1}{Z} R(y,z) q_Y(y) q_Z(z)) - P_{Z|S}(z|s)| \\ &+ \sum_{y \in Y \setminus \{1\}} |(\sum_z \frac{1}{Z} R(y,z) q_Y(y) q_Z(z)) - P_{Y|S}(z|s)| \\ \leq &\epsilon_0.\end{aligned}$$

Summarizing, we can guarantee ϵ optimality of the overall optimization, if the outer loop is interrupted when $\log \frac{\tilde{q}^{(i+1)}}{\tilde{q}^{(i)}} \leq \frac{\epsilon}{3}$, and the inner loop is interrupted when $\frac{\|\tilde{\eta}^{(i)} - \eta\|_1}{\min \tilde{q}^{(i)}} \leq \frac{\epsilon}{12}$.

B Notation

We use capital letters to denote random variables and script for the corresponding finite alphabets. We write P_S for the probability distribution of S, which is a vector with entries $P_S(s)$, $s \in \mathcal{S}$. The support of P_S is the set $\mathrm{supp}(P_S) = \{s \in \mathcal{S} : P_S(s) \neq 0\}$. The set of all probability measures on $\mathcal{S}$ is denoted $\mathbb{P}_{\mathcal{S}}$. A Markov kernel from $\mathcal{S}$ to $\mathcal{Y}$ is a measurable function $P_{Y|S} : \mathcal{S} \to \mathbb{P}_{\mathcal{Y}}$, represented by a matrix with rows $P_{Y|S=s} = P_{Y|s} \in \mathbb{P}_{\mathcal{Y}}$, $s \in \mathcal{S}$.

We use the following quantities:

- The entropy $H(P_S)$ of a distribution $P_S \in \mathbb{P}_{\mathcal{S}}$ is $H(P_S) := -\sum_{s \in \mathcal{S}} P_S(s) \log P_S(s)$.
- The Kullback-Leibler divergence from P_S to Q_S is

$$D(P_S \| Q_S) := \sum_{s \in \mathcal{S}} P_S(s) \log \tfrac{P_S(s)}{Q_S(s)}.$$

- The conditional divergence is

$$D(P_{Y|S} \| Q_{Y|S} | P_S) := \mathbb{E}_{s \sim P_S}[D(P_{Y|S=s} \| Q_{Y|S=s})].$$

- The mutual information of two random variables S and Y is $I(S;Y) := D(P_{SY} \| P_S P_Y)$. Equivalently, $I(S;Y) = D(P_{Y|S} \| P_Y | P_S) = D(P_{S|Y} \| P_S | P_Y)$. We use a subscript to specify the underlying distribution, e.g., $I_Q(S;Y|Z)$, under $Q = Q_{SYZ}$.
- The conditional mutual information of S and Y given Z is

$$I_Q(S;Y|Z) = \sum_z Q_Z(z) \sum_{s,y} Q_{SY|Z}(s,y|z) \log \tfrac{Q_{SY|Z}(s,y|z)}{Q_{S|Z}(s|z) Q_{Y|Z}(y|z)}.$$

References

A. J. Bell. The co-information lattice. In *Proceedings of the Fourth International Symposium on Independent Component Analysis and Blind Signal Separation (ICA 03)*, 2003.

N. Bertschinger, J. Rauh, E. Olbrich, J. Jost, and N. Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014.

R. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, 18(4):460–473, 1972.

T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, first edition, 1991.

I. Csiszár and J. Körner. *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

I. Csiszár and P. C. Shields. Information theory and statistics: A tutorial. *Foundations and Trends® in Communications and Information Theory*, 1(4):417–528, 2004.

K. Ghazi-Zahedi and J. Rauh. Quantifying morphological computation based on an information decomposition of the sensorimotor loop. In *Proceedings of the 13th European Conference on Artificial Life (ECAL 2015)*, pages 70–77, July 2015.

K. Ghazi-Zahedi, C. Langer, and N. Ay. Morphological computation: Synergy of body and brain. *Entropy*, 19(9), 2017. ISSN 1099-4300. doi: 10.3390/e19090456. URL `http://www.mdpi.com/1099-4300/19/9/456`.

V. Griffith and C. Koch. Quantifying synergistic mutual information. In M. Prokopenko, editor, *Guided Self-Organization: Inception*, volume 9 of *Emergence, Complexity and Computation*, pages 159–190. Springer Berlin Heidelberg, 2014.

M. Harder, C. Salge, and D. Polani. A bivariate measure of redundant information. *Physical Review E*, 87:012130, Jan 2013.

P. E. Latham and S. Nirenberg. Synergy, redundancy, and independence in population codes, revisited. *Journal of Neuroscience*, 25(21):5195–5206, 2005.

M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

W. McGill. Multivariate information transmission. *IRE Transactions on Information Theory*, 4(4):93–111, 1954.

T. Schreiber. Measuring information transfer. *Physical Review Letters*, 85(2):461, 2000.

T. M. Tax, P. A. Mediano, and M. Shanahan. The partial information decomposition of generative neural network models. *Entropy*, 19(9), 2017. ISSN 1099-4300. doi: 10.3390/e19090474. URL `http://www.mdpi.com/1099-4300/19/9/474`.

J. Vergara and P. Estévez. A review of feature selection methods based on mutual information. *Neural Computing & Applications*, 24:175–186, 2014.

P. Williams and R. Beer. Nonnegative decomposition of multivariate information. *arXiv:1004.2515v1*, 2010.