Optimizing Bivariate Partial Information Decomposition

Abdullah Makkeh, Dirk Oliver Theis*

Institute of Computer Science of the University of Tartu
Ulikooli 17, 51014 Tartu, Estonia
{makkeh,dotheis}@ut.ee

September 23, 2018

Abstract

None of the BROJA information decomposition measures $\text{SI}, \text{CI}, \text{UI}_y, \text{UI}_z$ are convex or concave over the probability simplex. In this paper, we provide formulas for the sub-gradient and super-gradients of any of the information decomposition measures. Then we apply these results to obtain an optimum of some of these information decomposition measures when optimized over a constrained set of probability distributions.

1 Introduction

Terminology and notation

We use the common shorthand $[n] := \{1, \ldots, n\}$. For vectors, we use the following summation convention: Replacing an index by an asterisk $*$ has the effect summing over all the possible values, e.g., for $p \in \mathbb{R}^{A \times B \times C}$, the term $p_{a,*c}$ stands for $(\sum_{b \in B} p_{a,b,c})$, e.g.,

$$p_{a,*c} = (\sum_{b \in B} p_{a,b,c})(\sum_{a \in A} p_{a,b,c})$$

All random variables considered in this paper have finite range (unless explicitly stated otherwise). Denote by $R_X$ the range\(^1\) of the (finite-range) random variable $X$.

For a (finite) set $X$, we denote the probability simplex by

$$\Delta^X := \{p \in \mathbb{R}_+^X \mid p_* = 1\}$$

For us, a probability distribution on a set $X$, is a vector in $\Delta^X$.

---

*Supported by the Estonian Research Council, ETAG (Eesti Teadusagentuur), through PUT Exploratory Grant #620, and by the European Regional Development Fund through the Estonian Center of Excellence in Computer Science, EXCS.

\(^1\)The range is a set with the property $P(X = x) > 0$ for all $x \in R_X$, and $P(X = x) = 0$ for all $x \not\in R_X$. If a range exists it is unique; if the range exists and is finite, we say that the random variable has “finite range”.

1
2 Main Theorem: Derivatives of PID-Quantities

\[ M(p) := \max_q h(q) \]  
over \( q \in \mathbb{R}^{S \times Y \times Z} \)  
subject to \( q_{s,y,z} = p_{s,y,z} \) for all \((s, y) \in S \times Y\);  
(1c) \( q_{s,z} = p_{s,z} \) for all \((s, z) \in S \times Z\);  
(1d) \( q_{s,y,z} \geq 0 \) for all \((s, y, z) \in S \times Y \times Z\).  
(1e)

Proposition 1 (Corollary 3 in [3]). A feasible point \( q \) is an optimal solution to (1), if and only if there exist \( \lambda \in \mathbb{R}^{S \times Y} \) and \( \mu \in \mathbb{R}^{S \times Z} \) satisfying the following:

(a) For all \((y, z) \in Y \times Z\) with \( q_{*,y,z} > 0 \):
\[ \lambda_{s,y} + \mu_{s,z} = \ln \left( \frac{q_{s,y,z}}{q_{*,y,z}} \right) \quad \text{holds for all } s \in S; \]

(b) For all \((y, z) \in Y \times Z\) with \( q_{*,y,z} = 0 \), there is a probability distribution \( \varrho \) with support \( S \) such that
\[ \lambda_{s,y} + \mu_{s,z} \leq \ln(\varrho_{y,z}^s) \quad \text{holds for all } s \in S. \]

If \( q, \lambda, \mu \) are as in the proposition, then we say that \( \lambda, \mu \) are Lagrange multipliers certifying optimality.

Lemma 2. Suppose \( p \) has full support. Let \( q \) be an optimal solution of (1), and let \( \lambda, \mu \) be Lagrange multipliers certifying optimality.

(a) If \( q_{*,y,z} > 0 \) for all \((s, y, z) \in S \times Y \times Z\), then \( M \) is differentiable in \( p \), and we have
\[ \frac{\partial M(p)}{\partial s_{y,z}} = -\lambda_{s,y} - \mu_{s,z} \]
(2)

(b) In any case, the vector defined by
\[ g(p)_{s,y,z} := -\lambda_{s,y} - \mu_{s,z} \]
(3)
is a super-gradient on \( M \) in the point \( p \).

Proof. If \( q \) is the optimal solution of (1), then
\[ M(p) = \max_q h(q) = -\min_q -h(q) = h(q) \]
where \( h(q) = H(S \mid Y, Z) \). So, the gradient of \( M \) in \( p \) is
\[ \nabla M(p) = -\ln \left( \frac{q_{s,y,z}}{q_{*,y,z}} \right)_{s,y,z} \]
(4)

If \( q_{s,y,z} > 0 \) for all \((s, y, z) \in S \times Y \times Z\), then \( q_{*,y,z} > 0 \) for all \((y, s) \in Y \times Z\) and so \( M \) is differentiable in \( p \). Moreover, Equation (2) follows from the fact that \( q, \lambda, \mu \) are as in Proposition 1 and the gradient defined in (4).

From [3, Proposition 2] and Proposition 1, we have \( \lambda_{s,y} + \mu_{s,z} \) is a sub-gradient to \( M'(p) := \min_q -h(q) \) subject to the constraints (1c), (1d), and (1e) in the point \( p \). Hence \(-\lambda_{s,y} - \mu_{s,z}\) is a super-gradient on \( M \) in the point \( p \).

We would like to emphasize that, in this lemma as well as in the following results, the condition that \( p \) has full support is only there to simplify notation, and can be readily abandoned.
Lemma 3 ([5], Lemma 2.73). Let \( f : \mathbb{R}^n \to \mathbb{R} \) be a convex function and \( p \in \text{dom}(f) \). A vector \( g \) is a subgradient of \( f \) in the point \( p \) iff

\[
f'(p, d) \geq g^T d \quad \text{for all } d \in \mathbb{R}^n.
\]

Theorem 4. Suppose \( p \) has full support. Let \( q \) be an optimal solution of (1), and let \( \lambda, \mu \) be Lagrange multipliers certifying optimality.

(a) If \( q_{s,y,z} > 0 \) for all \( (s, y, z) \in S \times Y \times Z \), then CI, SI, UIy, UIz are all differentiable in \( p \), and we have

\[
\begin{align*}
\partial_{s,y,z} \text{CI}(p) &= \ln \left( \frac{p_{s,y,z}}{p_{s,y,z}^*} \right) - \lambda_{s,y} - \mu_{s,z} \quad (5a) \\
\partial_{s,y,z} \text{SI}(p) &= -1 + \ln \left( \frac{p_{s,y,z}^*}{p_{s,y,z}} \right) - \lambda_{s,y} - \mu_{s,z} \quad (5b) \\
\partial_{s,y,z} \text{UIy}(p) &= \ln \left( \frac{p_{s,y,z}^*}{p_{s,y,z}} \right) + \lambda_{s,y} + \mu_{s,z} \quad (5c) \\
\partial_{s,y,z} \text{UIz}(p) &= \ln \left( \frac{p_{s,y,z}^*}{p_{s,y,z}} \right) + \lambda_{s,y} + \mu_{s,z} \quad (5d)
\end{align*}
\]

(b) In any case, the vectors defined by

\[
\begin{align*}
g_{\text{CI}}(p)_{s,y,z} &= \ln \left( \frac{p_{s,y,z}}{p_{s,y,z}^*} \right) - \lambda_{s,y} - \mu_{s,z} \quad (6a) \\
g_{\text{SI}}(p)_{s,y,z} &= -1 + \ln \left( \frac{p_{s,y,z}^*}{p_{s,y,z}} \right) - \lambda_{s,y} - \mu_{s,z} \quad (6b)
\end{align*}
\]

are local super-gradivents of CI and SI respectively and the vectors defined by

\[
\begin{align*}
g_{\text{UIy}}(p)_{s,y,z} &= \ln \left( \frac{p_{s,y,z}^*}{p_{s,y,z}} \right) + \lambda_{s,y} + \mu_{s,z} \quad (7a) \\
g_{\text{UIz}}(p)_{s,y,z} &= \ln \left( \frac{p_{s,y,z}^*}{p_{s,y,z}} \right) + \lambda_{s,y} + \mu_{s,z} \quad (7b)
\end{align*}
\]

are local subgradients of UIy and UIz in the point \( p \) respectively.

Proof. For (a), Bertschinger et al. in [1] defined the partial information decomposition as follows:

\[
\begin{align*}
\text{CI}(p) &= \text{MI}(S; Y, Z) - \min_q \text{MI}(S; Y, Z) \\
\text{SI}(p) &= \max_q \text{CoI}(S; Y; Z) \\
\text{UIy}(p) &= \min_q \text{MI}(S; Y | Z) \\
\text{UIz}(p) &= \min_q \text{MI}(S; Z | Y)
\end{align*}
\]

where the optimization is subject to the constraints (1c), (1d), and (1e). Using the definition of \( \text{MI}(S; Y, Z) \) and the chain rule, we get

\[
\begin{align*}
\text{CI}(p) &= M(p) - H(S | Y, Z) \\
\text{SI}(p) &= M(p) + \text{MI}(S; Y) - H(S | Z) \\
\text{UIy}(p) &= \text{MI}(S; Z) + H(S) - M(p) \\
\text{UIz}(p) &= \text{MI}(S; Y) + H(S) - M(p)
\end{align*}
\]

where \( H(S | Y, Z), H(S | Z), \text{MI}(S; Y), \) and \( \text{MI}(S; Y) \) are functions of \( p \). By direct computations the equations in (a) follow using the fact \( \partial_{s,y,z} M(p) = -\lambda_{s,y} - \mu_{s,z} \).

3
For (b), let
\[
\begin{align*}
g_{CI}(p) &= H(S \mid Y, Z) \\
g_{SI}(p) &= MI(S; Y) - H(S \mid Z) \\
g_{UI_y}(p) &= MI(S; Z) + H(S) \\
g_{UI_z}(p) &= MI(S; Y) + H(S).
\end{align*}
\] (8)

Since \( p \) has a full support then all the functions in (8) are differentiable and
\[
\begin{align*}
g'_{CI}(p, d) &= \sum_{s,y,z} \ln \left( \frac{p_{s,y,z}}{p_{s,y,z}} \right) d_{s,y,z} \\
g'_{SI}(p, d) &= \sum_{s,y,z} \left( \ln \left( \frac{p_{s,y,z}p_{s,y,z}}{p_{s,y,z}p_{s,y,z}} \right) - 1 \right) d_{s,y,z} \\
g'_{UI_y}(p, d) &= \sum_{s,y,z} \ln \left( \frac{p_{s,y,z}}{p_{s,y,z}} \right) d_{s,y,z} \\
g'_{UI_z}(p, d) &= \sum_{s,y,z} \ln \left( \frac{p_{s,y,z}}{p_{s,y,z}} \right) d_{s,y,z}.
\end{align*}
\] (9)

From Lemma 2 and Lemma 3, \( g(p) \) is a super-gradient of \( M \) at \( p \) and for any \( d \in \mathbb{R}^{S \times Y \times Z} \), we have \(-M'(p, d) \geq -g^Td\). Hence, the vectors defined by (6a) and (6b) are super-gradients of CI and SI respectively and the vectors defined by (7a) and (7b) are local subgradients of UI\(_y\) and UI\(_z\) in the point \( p \) respectively.

**Corollary 5.** Let \( I \) be any of CI, SI, UI\(_y\), UI\(_z\). At the points where \( I \) is not smooth it is
(a) concave, in the case of \( I = CI, SI \);
(b) convex, in the case of \( I = UI_y, UI_z \).

**Proof.** Using Theorem (a), the vectors \( g_{CI}(p) \) and \( g_{SI}(p) \) are local super-gradients of CI and SI and the vectors \( g_{UI_y}(p) \) and \( g_{UI_z}(p) \) are local sub-gradients of UI\(_y\) and UI\(_z\) in the point \( p \). From this, the statements in this Corollary follow.

### 3 Application I: Extractable Shared Information

Let \( S, Y, Z \) are random variables with joint probability distribution \( p \), and denote by \( S, Y, Z \) the ranges, respectively, of \( S, Y, Z \).

For a set \( R \) and a \( m \in \mathbb{N} \), a **stochastic** \( ([m] \times R) \)-matrix is a matrix \( \Pi \) with \( m \) rows (indexed \( 1, \ldots, m \) as usual) and columns indexed by the elements of \( R \), whose entries are nonnegative reals such that \( \Pi_{s,z} = 1 \). Let \( p \) be a probability distribution on \( S \times Y \times Z \), and \( \Pi \) be a stochastic \( ([m] \times S) \)-matrix. Then we define the probability distribution \( \Pi(p) \) as follows:
\[
\Pi_t(s,y,z) := \sum_{s \in R_{S_i}(p)} \Pi_{t,s}p_{s,y,z}, \text{ for all } t \in [m] \text{ and } (y,z) \in Y \times Z.
\]

Rauh et al. [4] define two “extractable” versions of shared information. Let \( S, Y, Z \) be random variables with distribution \( p \in \mathbb{P}^{S \times Y \times Z} \). The **extractable shared information** of \( S, Y, Z \) is defined as
\[
SI_{\text{ext}}(p) := \sup_f SI(f(S); Y, Z)
\] (10)

where the supremum is taken over all functions \( f: S \to T \), where \( S \) is the range of \( S \) and \( T \) is an arbitrary finite set. The **probabilistically extractable shared information** is defined as
\[
SI_{\text{prext}}(p) := \sup_T SI(T; Y, Z)
\] (11)
where the supremum is taken over all random variables \( T \) (with finite range) which are conditionally independent of \( Y, Z \) given \( S \).

It is straightforward that the extractable shared information of \( p \) is the value of the following optimization problem:

\[
\text{SI}_{\text{ext}}(p) := \max_{\Pi} \text{SI}(\Pi(p))
\]

over \( \Pi \in \mathbb{R}^{[m] \times S} \)

subject to

\[
\Pi_{t,s} = 1 \quad \text{for all } s \in S
\]

\[
\Pi_{t,s} \geq 0 \quad \text{for all } (t, s) \in [m] \times S
\]

\[
\Pi_{t,s} \in \mathbb{Z} \quad \text{for all } (t, s) \in [m] \times S.
\]

To see why this is the same as the definition \((10)\), given in \([4]\), let us take random variables \( S, Y, Z \) with distribution \( p \). The integrality constraints \((12d)\) — together with the nonnegativity inequalities \((12c)\) and the equation — have precisely the effect of ensuring that for every \( s \) in the range of \( S \) there exists a unique \( t \in [m] \) such that \( \Pi_{t,s} = 1 \). In other words, \( \Pi \) defines a mapping from \( Rg_S \) to \([m]\). Since \( m \) is the size of the range of \( S \), the optimization problem \((12)\) simply optimizes over all functions defined on the range of \( S \), which is exactly \((10)\).

Similarly, the probabilistically extractable shared information is the value of the following optimization problem:

\[
\sup_{m \geq |S|} \text{SI}(\Pi(p))
\]

over \( m \geq |S| \)

subject to

\[
\Pi_{t,s} = 1 \quad \text{for all } s \in S
\]

\[
\Pi_{t,s} \geq 0 \quad \text{for all } (t, s) \in [m] \times S.
\]

To see why this is equivalent to the definition \((11)\), given in \([4]\), consider the relation

\[
\Pi_{t,s} = \mathbb{P}(T = t \mid S = s).
\]

Given \( \Pi \), it defines a random variable \( T \) which is conditionally independent of \( X_1, \ldots, X_k \) given \( S \), such that \( \Pi(p) \) is the distribution of \( (T, X_1, \ldots, X_k) \). On the other hand, given a random variable \( T \) conditionally independent of \( X_1, \ldots, X_k \) given \( S \), setting \( m := \max Rg_S \), relation \((14)\) defines a \( \Pi \) such that \( \Pi(p) \) is the distribution of \( (T, X_1, \ldots, X_k) \). We invite the reader to check these claims — or read the detailed proof in \([2, \text{Lemma 5.2.1}]\).

There are two significant differences between the \((12)\) and \((13)\). Firstly, it lacks the integrality constraints, making it a continuous optimization problem. Secondly, the dimension, \( m \), is a variable, making the optimization problem infinite dimensional (as observed in \([4]\)), and thus basically intractable from an algorithmic point of view. (The lower bound \( m \geq S \) is redundant, see Lemma \(6 \) below).

The following optimization problem, however, is a standard continuous optimization problem to which we can apply our results: For a fixed value of \( m \in \mathbb{N} \), let us define

\[
\text{SI}_{\text{m}}(p) := \max_{\Pi} \text{SI}(\Pi(p))
\]

over \( \Pi \in \mathbb{R}^{m \times S} \)

subject to

\[
\Pi_{t,s} = 1 \quad \text{for all } s \in S
\]

\[
\Pi_{t,s} \geq 0 \quad \text{for all } (t, s) \in [m] \times S.
\]

\(^2\)Approximation through is thinkable.
The following lemma is quite obvious (see \cite[Lemma 5.2.2]{makkeh} for a detailed proof).

**Lemma 6.** The sequence \( m \mapsto \text{SI}^\bullet(m) \) is non-decreasing and for every fixed \( m_0 \geq |S| \),

\[
\text{SI}^\text{ext}(p) \leq \text{SI}^\bullet_{m_0}(p) \leq \sup_{m \geq 0} \text{SI}^\bullet_m(p) = \sup_{m \geq 0} \text{SI}^\text{prext}(p).
\]

**Acknowledgements**

This research was supported by the Estonian Research Council, ETAG (*Eesti Teadusagentuur*), through PUT Exploratory Grant #620. We also gratefully acknowledge funding by the European Regional Development Fund through the Estonian Center of Excellence in Computer Science, EXCS.

**References**

[1] Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J.; Ay, N. Quantifying unique information. *Entropy* 2014, 16, 2161–2183.

[2] Makkeh, A. Applications of Optimization in Some Complex Systems. Ph.D. Thesis, University of Tartu, Tartu, Estonia, forthcoming.

[3] Makkeh, A.; Theis, D. O.; Vicente, R. Bivariate Partial Information Decomposition: The Optimization Perspective. *Entropy* 2017, 19(10) 530.

[4] Rauh, J.; Banerjee, P. Kr.; Olbrich, E.; Jost, J.; Bertschinger, N. On Extractable Shared Information *Entropy* 2017, 19 (10) 328.

[5] Ruszczyński, A. P. Nonlinear optimization *Princeton university press* 2006, 13.