Multivariate Information Decomposition

DARE Centre

Conor Finn

March 2021

THE UNIVERSITY OF SYDNEY

Overview

Information Theory

Information content

How much information h do I get from knowing the outcome x of a variable X?

Information content

How much information h do I get from knowing the outcome x of a variable X?

- If the outcome was certain, then we should get no information.

Information content

How much information h do I get from knowing the outcome x of a variable X?

- If the outcome was certain, then we should get no information.
- The more surprising the outcome, the more information we should get.

Information content

How much information h do I get from knowing the outcome x of a variable X?

- If the outcome was certain, then we should get no information.
- The more surprising the outcome, the more information we should get.
- If two independent outcomes occur, then the total information that we get should be equal to the sum of the information that we get from each outcome.

Information content

How much information h do I get from knowing the outcome x of a variable X?

- If the outcome was certain, then we should get no information.
- The more surprising the outcome, the more information we should get.
- If two independent outcomes occur, then the total information that we get should be equal to the sum of the information that we get from each outcome.

The unique function that satisfies the criteria is called the **information content**,

$$h(x) = \log \frac{1}{p(x)} = -\log p(x) \geq 0.$$

Joint and conditional information content

With a second random variable Y, we can consider the joint information content,

$$h(x, y) = -\log p(x, y) \geq 0.$$

- Since $p(x, y) \leq p(x), p(y)$, we know that $h(x, y) \geq h(x), h(y)$.

Joint and conditional information content

With a second random variable Y, we can consider the joint information content,

$$h(x, y) = -\log p(x, y) \geq 0.$$

- Since $p(x, y) \leq p(x), p(y)$, we know that $h(x, y) \geq h(x), h(y)$.

We can also consider the conditional information content,

$$\begin{aligned} h(x|y) &= -\log p(x|y) = -\log p(x, y) + \log p(y) \\ &= h(x, y) - h(y) \geq 0. \end{aligned}$$

Entropy, joint entropy and conditional entropy

The expected information content of a random variable is called the **entropy**,

$$H(X) = \mathrm{E}_X[h(x)] = -\sum_{x \in X} p(x) \log p(x) \geq 0,$$

- Non-negativity of the entropy follows directly from that of the information content.

Entropy, joint entropy and conditional entropy

The expected information content of a random variable is called the **entropy**,

$$H(X) = \mathrm{E}_X[h(x)] = -\sum_{x \in X} p(x) \log p(x) \geq 0,$$

- Non-negativity of the entropy follows directly from that of the information content.

Similarly, we have the joint and conditional entropy, which are also non-negative,

$$H(X,Y) = \mathrm{E}_{XY}[h(x,y)] = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x,y) \geq 0,$$

$$H(X|Y) = \mathrm{E}_{XY}[h(x|y)] = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x|y) \geq 0.$$

Mutual information and pointwise mutual information

The **pointwise mutual information** is defined as follows,

$$i(x;y) = h(x) + h(y) - h(x,y).$$

Mutual information and pointwise mutual information

The **pointwise mutual information** is defined as follows,

$$i(x; y) = h(x) + h(y) - h(x, y).$$

- In contrast to the various information contents, this function is not non-negative.

Mutual information and pointwise mutual information

The **pointwise mutual information** is defined as follows,

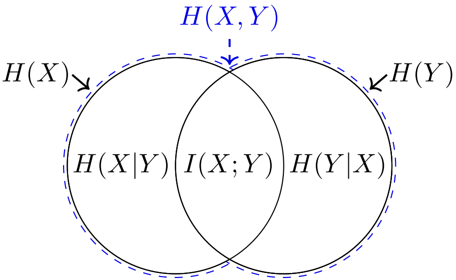
$$i(x;y) = h(x) + h(y) - h(x,y).$$

- In contrast to the various information contents, this function is not non-negative.

Nevertheless, the **mutual information** is non-negative,

$$\begin{aligned} I(X;Y) &= \mathrm{E}_{XY}[i(x,y)] \\ &= H(X) + H(Y) - H(X,Y) \\ &= \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= D_{\mathsf{KL}}(P_{XY} || P_X \otimes P_Y) \geq 0. \end{aligned}$$

Summarising the basic functions



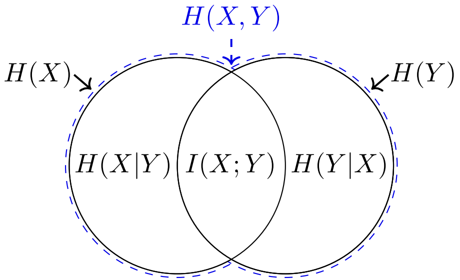
$$H(X) + H(Y) \geq H(X,Y) \geq H(X),\ H(Y) \geq 0$$

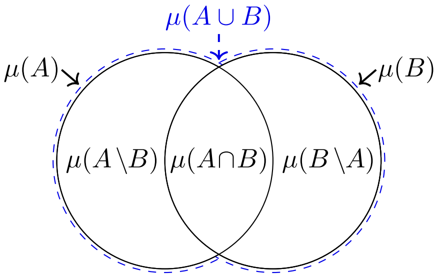
$$H(X|Y) = H(X,Y) - H(Y) \geq 0$$

$$H(Y|X) = H(X,Y) - H(X) \geq 0$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \geq 0$$

Summarising the basic functions

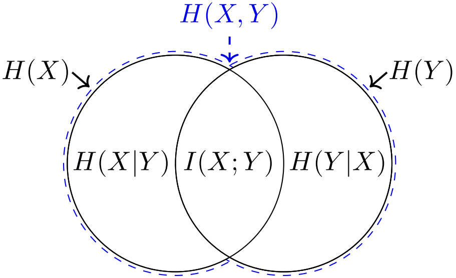


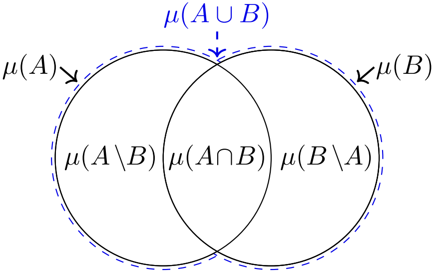


$$H(X) + H(Y) \geq H(X,Y) \geq H(X),\ H(Y) \geq 0$$

$$H(X|Y) = H(X,Y) - H(Y) \geq 0$$
$$H(Y|X) = H(X,Y) - H(X) \geq 0$$
$$I(X;Y) = H(X) + H(Y) - H(X,Y) \geq 0$$

Summarising the basic functions





$$H(X) + H(Y) \geq H(X,Y) \geq H(X),\ H(Y) \geq 0$$

$$\mu(A) + \mu(B) \geq \mu(A \cup B) \geq \mu(A),\ \mu(B) \geq 0$$

$$H(X|Y) = H(X,Y) - H(Y) \geq 0$$
$$H(Y|X) = H(X,Y) - H(X) \geq 0$$
$$I(X;Y) = H(X) + H(Y) - H(X,Y) \geq 0$$

$$\mu(A \setminus B) = \mu(A \cup B) - \mu(B) \geq 0$$
$$\mu(B \setminus A) = \mu(A \cup B) - \mu(A) \geq 0$$
$$\mu(A \cap B) = \mu(A) + \mu(B) - \mu(A \cup B) \geq 0$$

Multivariate mutual information

McGill generalised the MI by defining the **multivariate mutual information**,

$$\begin{aligned} I(X;Y;Z) = H(X) + H(Y) + H(Z) \\ - H(X,Y) - H(X,Z) - H(Y,Z) \\ + H(Z,Y,Z) \end{aligned}$$

$$= I(X;Y) + I(X;Z) - I\big(X;(Y,Z)\big)$$

$$= \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y)p(x,z)p(y,z)}{p(x)p(y)p(z)p(x,y,z)}$$

Multivariate mutual information

McGill generalised the MI by defining the **multivariate mutual information**,

$$
\begin{aligned}
I(X;Y;Z) = {} & H(X) + H(Y) + H(Z) \\
& - H(X,Y) - H(X,Z) - H(Y,Z) \\
& + H(Z,Y,Z) \\
= {} & I(X;Y) + I(X;Z) - I\big(X;(Y,Z)\big) \\
= {} & \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y)p(x,z)p(y,z)}{p(x)p(y)p(z)p(x,y,z)}
\end{aligned}
$$

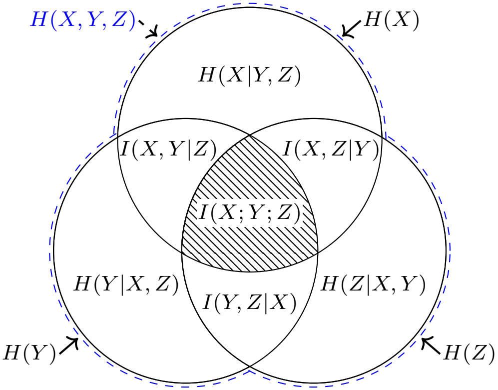
- MMI is not non-negative.
- MMI has “no intuitive meaning”.

Multivariate mutual information

McGill generalised the MI by defining the **multivariate mutual information**,

$$
\begin{aligned}
I(X;Y;Z) = {} & H(X) + H(Y) + H(Z) \\
& - H(X,Y) - H(X,Z) - H(Y,Z) \\
& + H(Z,Y,Z) \\
= {} & I(X;Y) + I(X;Z) - I\big(X;(Y,Z)\big) \\
= {} & \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y)p(x,z)p(y,z)}{p(x)p(y)p(z)p(x,y,z)}
\end{aligned}
$$

- MMI is not non-negative.
- MMI has "no intuitive meaning".



Partial Information Decomposition

Unique, redundant and synergistic information

Say S_1 and S_2 provide information about T

- Several types of information

Unique, redundant and synergistic information

Say S_1 and S_2 provide information about T

- Several types of information
  - **Unique information** $U(S_1 \setminus S_2; T)$

Unq

p	s_1	s_2	t
1/4	0	0	0
1/4	0	1	0
1/4	1	0	1
1/4	1	1	1

Unique, redundant and synergistic information

Say S_1 and S_2 provide information about T

- Several types of information
  - **Unique information** $U(S_1 \setminus S_2; T)$
  - **Redundant information** $R(S_1, S_2; T)$

UNQ

p	s_1	s_2	t
1/4	0	0	0
1/4	0	1	0
1/4	1	0	1
1/4	1	1	1

RDN

p	s_1	s_2	t
1/2	0	0	0
1/2	1	1	1

Unique, redundant and synergistic information

Say S_1 and S_2 provide information about T

- Several types of information
  - **Unique information** $U(S_1 \setminus S_2; T)$
  - **Redundant information** $R(S_1, S_2; T)$
  - **Synergistic information** $C(S_1, S_2; T)$

Unq

p	s_1	s_2	t
1/4	0	0	0
1/4	0	1	0
1/4	1	0	1
1/4	1	1	1

Rdn

p	s_1	s_2	t
1/2	0	0	0
1/2	1	1	1

Xor

p	s_1	s_2	t
1/4	0	0	0
1/4	0	1	1
1/4	1	0	1
1/4	1	1	0

Unique, redundant and synergistic information

Say S_1 and S_2 provide information about T

- Several types of information
  - **Unique information** $U(S_1 \setminus S_2; T)$
  - **Redundant information** $R(S_1, S_2; T)$
  - **Synergistic information** $C(S_1, S_2; T)$

- Mutual information captures

$$I(T; S_1) = U(S_1 \setminus S_2; T) + R(S_1, S_2; T)$$
$$I(T; S_2) = U(S_2 \setminus S_1; T) + R(S_1, S_2; T)$$

Unq

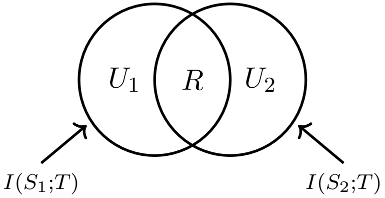
p	s_1	s_2	t
1/4	0	0	0
1/4	0	1	0
1/4	1	0	1
1/4	1	1	1

Rdn

p	s_1	s_2	t
1/2	0	0	0
1/2	1	1	1

Xor

p	s_1	s_2	t
1/4	0	0	0
1/4	0	1	1
1/4	1	0	1
1/4	1	1	0



Unique, redundant and synergistic information

Say S_1 and S_2 provide information about T

- Several types of information
  - **Unique information** $U(S_1 \setminus S_2; T)$
  - **Redundant information** $R(S_1, S_2; T)$
  - **Synergistic information** $C(S_1, S_2; T)$

- Mutual information captures

$$I(T; S_1) = U(S_1 \setminus S_2; T) + R(S_1, S_2; T)$$
$$I(T; S_2) = U(S_2 \setminus S_1; T) + R(S_1, S_2; T)$$

- Joint mutual information captures

$$I\big((S_1, S_2); T\big) = U(S_1 \setminus S_2; T) + U(S_2 \setminus S_1; T) + R(S_1, S_2; T) + C(S_1, S_2; T)$$

UNQ

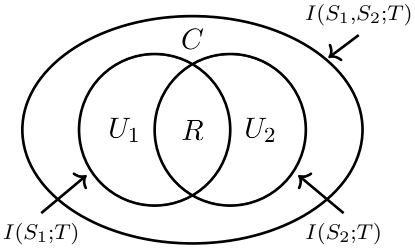
p	s_1	s_2	t
1/4	0	0	0
1/4	0	1	0
1/4	1	0	1
1/4	1	1	1

RDN

p	s_1	s_2	t
1/2	0	0	0
1/2	1	1	1

XOR

p	s_1	s_2	t
1/4	0	0	0
1/4	0	1	1
1/4	1	0	1
1/4	1	1	0



Information decomposition

Information decomposition for two source variable is an algebraic problem,

$$
\begin{aligned}
I(T;S_1) &= U(S_1 \setminus S_2;T) + R(S_1,S_2;T),\\
I(T;S_2) &= U(S_2 \setminus S_1;T) + R(S_1,S_2;T),\\
I\big((S_1,S_2);T\big) &= U(S_1 \setminus S_2;T) + U(S_2 \setminus S_1;T) + R(S_1,S_2;T) + C(S_1,S_2;T).
\end{aligned}
$$

Information decomposition

Information decomposition for two source variable is an algebraic problem,

$$\begin{aligned} I(T;S_1) &= U(S_1 \setminus S_2;T) + R(S_1,S_2;T), \\ I(T;S_2) &= U(S_2 \setminus S_1;T) + R(S_1,S_2;T), \\ I\big((S_1,S_2);T\big) &= U(S_1 \setminus S_2;T) + U(S_2 \setminus S_1;T) + R(S_1,S_2;T) + C(S_1,S_2;T). \end{aligned}$$

- Explains why the mutual information is not non-negative,

$$\begin{aligned} I(X;Y;Z) &= I(X;Y) + I(X;Z) - I\big(X;(Y,Z)\big) \\ &= R(S_1,S_2;T) - C(S_1,S_2;T). \end{aligned}$$

Information decomposition

Information decomposition for two source variable is an algebraic problem,

$$
\begin{aligned}
I(T;S_1) &= U(S_1 \setminus S_2;T) + R(S_1,S_2;T),\\
I(T;S_2) &= U(S_2 \setminus S_1;T) + R(S_1,S_2;T),\\
I\big((S_1,S_2);T\big) &= U(S_1 \setminus S_2;T) + U(S_2 \setminus S_1;T) + R(S_1,S_2;T) + C(S_1,S_2;T).
\end{aligned}
$$

- Explains why the mutual information is not non-negative,

$$
\begin{aligned}
I(X;Y;Z) &= I(X;Y) + I(X;Z) - I\big(X;(Y,Z)\big)\\
&= R(S_1,S_2;T) - C(S_1,S_2;T).
\end{aligned}
$$

- Can we define one of the quantities to solve the system?
- Can we generalise this to more than two sources?

Partial information decomposition (PID)

Axiomatic framework for multivariate information decomposition (Williams and Beer, 2010)

- Derived from axioms a measure of redundancy $I_\cap$ must satisfy.

Partial information decomposition (PID)

Axiomatic framework for multivariate information decomposition (Williams and Beer, 2010)

- Derived from axioms a measure of redundancy $I_\cap$ must satisfy.

Axioms

(1) *Symmetry*: $I_\cap(S_1, \ldots, S_n; T)$ is invariant under permutations of the S_i's

(2) *Self-redundancy*: $I_\cap(S_i : T) = I(S_i; T)$

(3) *Monotonicity*: $I_\cap(S_1; \ldots; S_n; T) \leq I_\cap(S_1; \ldots; S_{n-1}; T)$

Partial information decomposition (PID)

Axiomatic framework for multivariate information decomposition (Williams and Beer, 2010)

- Derived from axioms a measure of redundancy $I_\cap$ must satisfy.

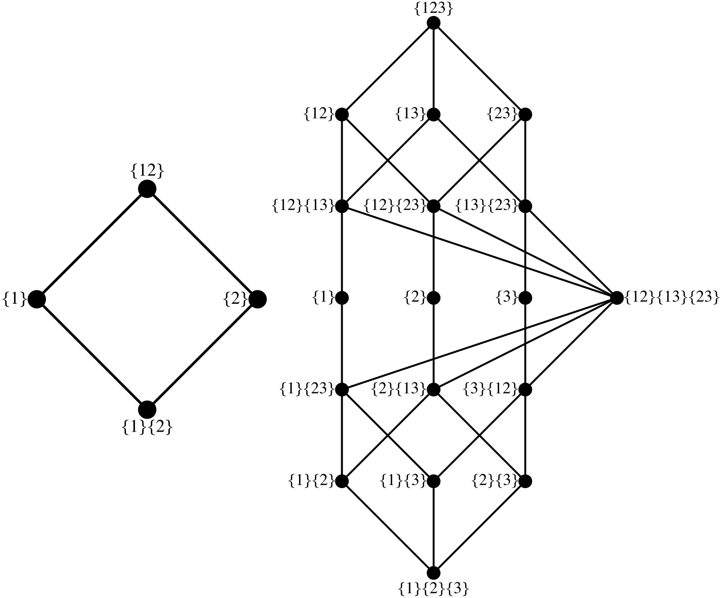
Axioms

(1) *Symmetry*: $I_\cap(S_1, \ldots, S_n; T)$ is invariant under permutations of the S_i's

(2) *Self-redundancy*: $I_\cap(S_i : T) = I(S_i; T)$

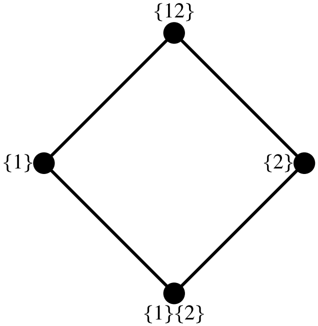
(3) *Monotonicity*: $I_\cap(S_1; \ldots; S_n; T) \leq I_\cap(S_1; \ldots; S_{n-1}; T)$

- Works for an arbitrary number of variables $S_1, \ldots, S_n$.
- Based on the intuitive notions from set theory.
- Williams and Beer used these axioms to derive the redundancy lattice.


{12}
{1}
{2}
{1}{2}
{123}
{12}
{13}
{23}
{12}{13}
{12}{23}
{13}{23}
{1}
{2}
{3}
{12}{13}{23}
{1}{23}
{2}{13}
{3}{12}
{1}{2}
{1}{3}
{2}{3}
{1}{2}{3}

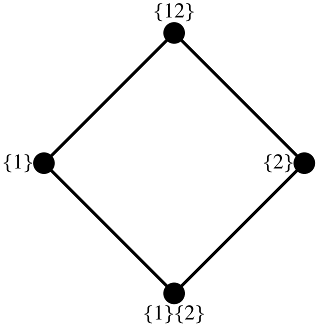
Möbius inversion

A Möbius inversion over the lattice yields partial information atoms and equations:



Möbius inversion

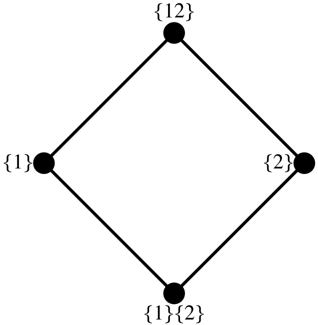
A Möbius inversion over the lattice yields partial information atoms and equations:



$$R(S_1, S_2; T) \quad = I_\cap(S_1, S_2; T)$$

Möbius inversion

A Möbius inversion over the lattice yields partial information atoms and equations:

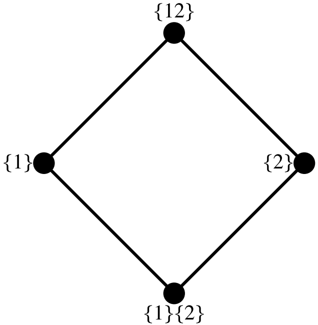


$$U(S_1 \setminus S_2; T) = I(S_1; T) - R(S_1, S_2; T)$$
$$U(S_2 \setminus S_1; T) = I(S_2; T) - R(S_1, S_2; T)$$

$$R(S_1, S_2; T) \quad = I_{\cap}(S_1, S_2; T)$$

Möbius inversion

A Möbius inversion over the lattice yields partial information atoms and equations:



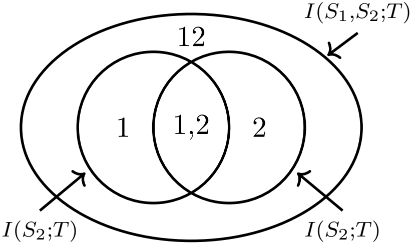
$$\begin{aligned} C(S_1, S_2; T) &= I\big((S_1, S_2); T\big) - R(S_1, S_2; T) \\ &\quad - U(S_1 \setminus S_2; T) - U(S_2 \setminus U_1; T) \end{aligned}$$

$$U(S_1 \setminus S_2; T) = I(S_1; T) - R(S_1, S_2; T)$$
$$U(S_2 \setminus S_1; T) = I(S_2; T) - R(S_1, S_2; T)$$

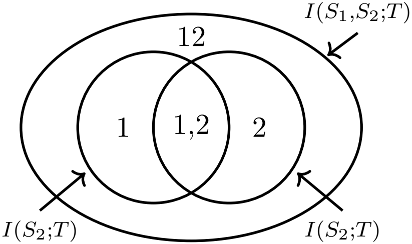
$$R(S_1, S_2; T) = I_\cap(S_1, S_2; T)$$

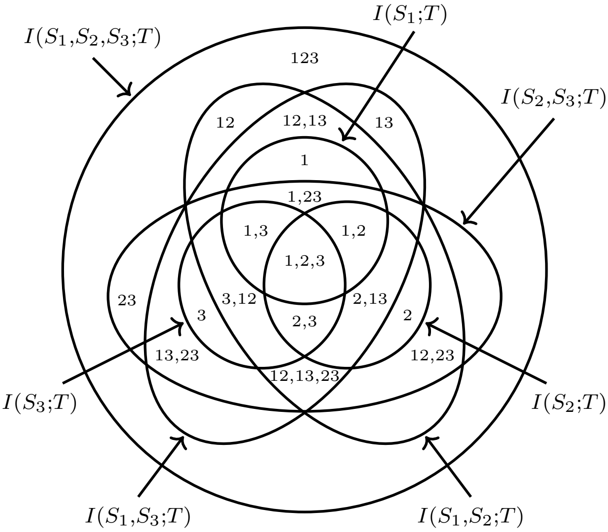
Atoms of partial information

Atoms of partial information



Atoms of partial information





But, we’re not yet complete...

Can we generalise this to more than two sources?

Can we define one of the quantities to solve the system?

But, we're not yet complete...

Can we generalise this to more than two sources?

- Yes $\leftarrow$ redundancy lattice.

Can we define one of the quantities to solve the system?

But, we're not yet complete...

Can we generalise this to more than two sources?

- Yes $\leftarrow$ redundancy lattice.

Can we define one of the quantities to solve the system?

- Not yet.

But, we're not yet complete...

Can we generalise this to more than two sources?

- Yes $\leftarrow$ redundancy lattice.

Can we define one of the quantities to solve the system?

- Not yet.
- There are many proposals (Lizier et al., 2018).

Redundancy measures: $I_{\min}$

Original measure of redundancy introduced by Williams and Beer (2010),

$$I_{\min}(S_1, \ldots, S_n) = \sum_t p(t) \min_{S_i} i(t; S_i),$$

where $i(t; S_i)$ is the specific information, which satisfies $I(S_i; T) = \mathrm{E}_T[i(t; S_i)]$.

Redundancy measures: $I_{\min}$

Original measure of redundancy introduced by Williams and Beer (2010),

$$I_{\min}(S_1, \ldots, S_n) = \sum_t p(t) \min_{S_i} i(t; S_i),$$

where $i(t; S_i)$ is the specific information, which satisfies $I(S_i; T) = \mathrm{E}_T[i(t; S_i)]$.

- Widely criticised after its introduction: two bit copy and pointwise unique problem.

Redundancy measures: $I_{\min}$

Original measure of redundancy introduced by Williams and Beer (2010),

$$I_{\min}(S_1, \ldots, S_n) = \sum_t p(t) \min_{S_i} i(t; S_i),$$

where $i(t; S_i)$ is the specific information, which satisfies $I(S_i; T) = \mathrm{E}_T[i(t; S_i)]$.

- Widely criticised after its introduction: two bit copy and pointwise unique problem.

Two bit copy

p	s_1	s_2	t
1/4	0	0	00
1/4	0	1	01
1/4	1	0	10
1/4	1	1	01

Redundancy measures: $I_{\min}$

Original measure of redundancy introduced by Williams and Beer (2010),

$$I_{\min}(S_1, \dots, S_n) = \sum_t p(t) \min_{S_i} i(t; S_i),$$

where $i(t; S_i)$ is the specific information, which satisfies $I(S_i; T) = \mathrm{E}_T[i(t; S_i)]$.

- Widely criticised after its introduction: two bit copy and pointwise unique problem.

Two bit copy

p	s_1	s_2	t
1/4	0	0	00
1/4	0	1	01
1/4	1	0	10
1/4	1	1	01

PW unique

p	s_1	s_2	t
1/4	0	1	1
1/4	1	0	1
1/4	0	2	2
1/4	2	0	2

Redundancy measures: $I_{\min}$

Original measure of redundancy introduced by Williams and Beer (2010),

$$I_{\min}(S_1, \ldots, S_n) = \sum_t p(t) \min_{S_i} i(t; S_i),$$

where $i(t; S_i)$ is the specific information, which satisfies $I(S_i; T) = \mathrm{E}_T[i(t; S_i)]$.

- Widely criticised after its introduction: two bit copy and pointwise unique problem.

Two bit copy

p	s_1	s_2	t
1/4	0	0	00
1/4	0	1	01
1/4	1	0	10
1/4	1	1	01

PW unique

p	s_1	s_2	t
1/4	0	1	1
1/4	1	0	1
1/4	0	2	2
1/4	2	0	2

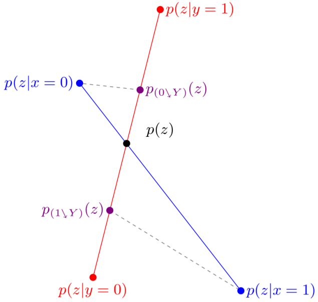
- "The problem is $I_{\min}$ does not distinguish whether sources carry the same information or just the same amount of information"

Redundancy measures: I_{red}

Based on information geometric methods from Harder et al. (2013),

$$I_{\text{red}}(S_1, S_2; T) = \min\left\{I_T^\pi(S_1 \searrow S_2)\,,\; I_T^\pi(S_2 \searrow S_1)\right\},$$

where $I_Z^\pi(X \searrow Y)$ is the MI between T and S_1 expressed in terms of T and S_2.

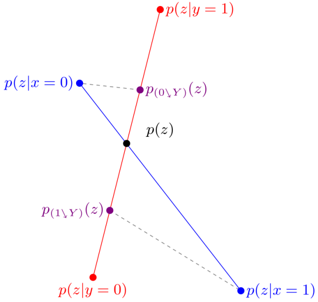


Redundancy measures: I_{red}

Based on information geometric methods from Harder et al. (2013),

$$I_{\text{red}}(S_1, S_2; T) = \min\left\{ I_T^\pi(S_1 \searrow S_2)\,,\ I_T^\pi(S_2 \searrow S_1) \right\},$$

where $I_Z^\pi(X \searrow Y)$ is the MI between T and S_1 expressed in terms of T and S_2.



- Only works for two source variables (Rauh et al., 2014).
- Not clear why this captures the redundant information.
- No meaningful pointwise interpretation.

Unique information measure: $\widetilde{UI}$

Based on a decision-theoretic motivation from Bertschinger et al. (2014),

$$\widetilde{UI}(S_1 \setminus S_2; T) = \min_{Q \in \Delta_P} I_Q(T; S_1 | S_2),$$

where Δ_P is the set of all joint distributions of the triple (S_1, S_2, T) that have the same marginal distributions of the pairs (S_1, T) and (S_2, T).

Unique information measure: $\widetilde{UI}$

Based on a decision-theoretic motivation from Bertschinger et al. (2014),

$$\widetilde{UI}(S_1 \setminus S_2; T) = \min_{Q \in \Delta_P} I_Q(T; S_1 | S_2),$$

where Δ_P is the set of all joint distributions of the triple (S_1, S_2, T) that have the same marginal distributions of the pairs (S_1, T) and (S_2, T).

- Again, only works for two source variables (Rauh et al., 2014).

Unique information measure: $\widetilde{UI}$

Based on a decision-theoretic motivation from Bertschinger et al. (2014),

$$\widetilde{UI}(S_1 \setminus S_2; T) = \min_{Q \in \Delta_P} I_Q(T; S_1 | S_2),$$

where Δ_P is the set of all joint distributions of the triple (S_1, S_2, T) that have the same marginal distributions of the pairs (S_1, T) and (S_2, T).

- Again, only works for two source variables (Rauh et al., 2014).
- Is equivalent to an approach proposed by Griffith and Koch (2014).

Unique information measure: $\widetilde{UI}$

Based on a decision-theoretic motivation from Bertschinger et al. (2014),

$$\widetilde{UI}(S_1 \setminus S_2; T) = \min_{Q \in \Delta_P} I_Q(T; S_1 | S_2),$$

where Δ_P is the set of all joint distributions of the triple (S_1, S_2, T) that have the same marginal distributions of the pairs (S_1, T) and (S_2, T).

- Again, only works for two source variables (Rauh et al., 2014).
- Is equivalent to an approach proposed by Griffith and Koch (2014).

PW UNIQUE

p	s_1	s_2	t
1/4	0	1	1
1/4	1	0	1
1/4	0	2	2
1/4	2	0	2

Unique information measure: $\widetilde{UI}$

Based on a decision-theoretic motivation from Bertschinger et al. (2014),

$$\widetilde{UI}(S_1 \setminus S_2; T) = \min_{Q \in \Delta_P} I_Q(T; S_1 | S_2),$$

where Δ_P is the set of all joint distributions of the triple (S_1, S_2, T) that have the same marginal distributions of the pairs (S_1, T) and (S_2, T).

- Again, only works for two source variables (Rauh et al., 2014).
- Is equivalent to an approach proposed by Griffith and Koch (2014).

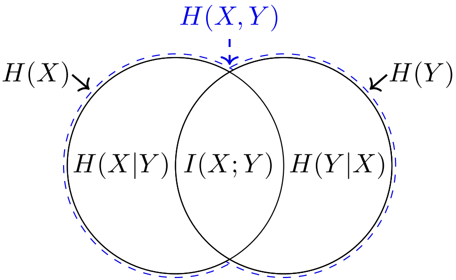
PW UNIQUE

p	s_1	s_2	t
1/4	0	1	1
1/4	1	0	1
1/4	0	2	2
1/4	2	0	2

$\widetilde{UI}(X : Y) = \widetilde{UI}(X : Y) = 0$ bit

Shared Marginal Information

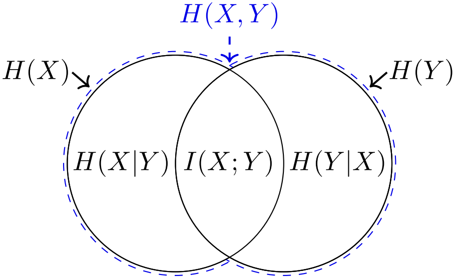
Why can we use Venn diagrams?

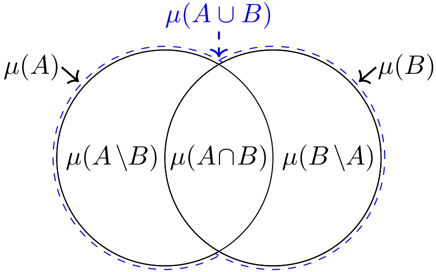


$$H(X) + H(Y) \geq H(X,Y) \geq H(X),\ H(Y) \geq 0$$

$$H(X|Y) = H(X,Y) - H(Y) \geq 0$$
$$H(Y|X) = H(X,Y) - H(X) \geq 0$$
$$I(X;Y) = H(X) + H(Y) - H(X,Y) \geq 0$$

Why can we use Venn diagrams?





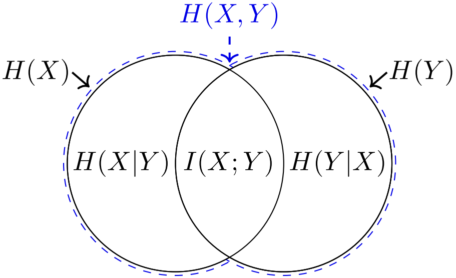
$$H(X) + H(Y) \geq H(X,Y) \geq H(X),\ H(Y) \geq 0$$

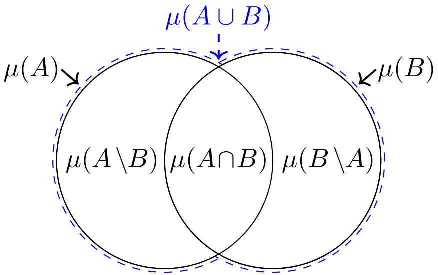
$$H(X|Y) = H(X,Y) - H(Y) \geq 0$$

$$H(Y|X) = H(X,Y) - H(X) \geq 0$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \geq 0$$

Why can we use Venn diagrams?





$$H(X) + H(Y) \geq H(X,Y) \geq H(X),\ H(Y) \geq 0$$

$$\mu(A) + \mu(B) \geq \mu(A \cup B) \geq \mu(A),\ \mu(B) \geq 0$$

$$H(X|Y) = H(X,Y) - H(Y) \geq 0$$
$$H(Y|X) = H(X,Y) - H(X) \geq 0$$
$$I(X;Y) = H(X) + H(Y) - H(X,Y) \geq 0$$

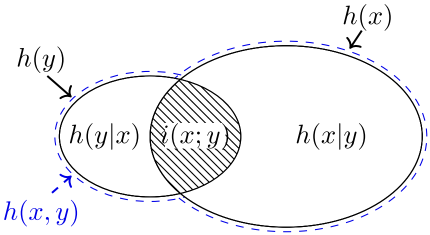
$$\mu(A \setminus B) = \mu(A \cup B) - \mu(B) \geq 0$$
$$\mu(B \setminus A) = \mu(A \cup B) - \mu(A) \geq 0$$
$$\mu(A \cap B) = \mu(A) + \mu(B) - \mu(A \cup B) \geq 0$$

Venn diagrams and information content

- The information content satisfies the following inequalities:

$$h(x, y) \geq h(x),\; h(y) \geq 0,$$

$$h(x|y) = h(x, y) - h(y) \geq 0,$$
$$h(y|x) = h(x, y) - h(x) \geq 0.$$



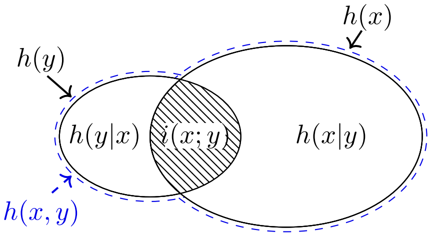
- In contrast to the joint and marginal entropy, the joint information content is not upper bounded by the sum of the marginal information contents.

Venn diagrams and information content

- The information content satisfies the following inequalities:

$$h(x,y) \geq h(x),\, h(y) \geq 0,$$

$$h(x|y) = h(x,y) - h(y) \geq 0,$$
$$h(y|x) = h(x,y) - h(x) \geq 0.$$



- In contrast to the joint and marginal entropy, the joint information content is not upper bounded by the sum of the marginal information contents.
- Thus, the pointwise mutual information is not non-negative,

$$i(x;y) = h(x) + h(y) - h(x,y).$$

Joint and independent information

	Johnny	Alice	Bob
			
Observations:	(X, Y)	X	Y
Knows:	$P(X, Y)$	$P(X)$	$P(Y)$

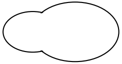
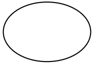
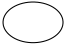
Joint and independent information

	Johnny	Alice	Bob
			
Observations:	(X, Y)	X	Y
Knows:	$P(X, Y)$	$P(X)$	$P(Y)$
Realisation:	(x, y)	x	y

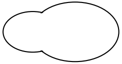
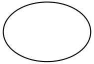
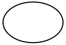
Joint and independent information

	Johnny	Alice	Bob
			
Observations:	(X, Y)	X	Y
Knows:	$P(X, Y)$	$P(X)$	$P(Y)$
Realisation:	(x, y)	x	y
Information:	$h(x, y)$	$h(x)$	$h(y)$

Joint and independent information

	Johnny	Alice	Bob
			
Observations:	(X, Y)	X	Y
Knows:	$P(X, Y)$	$P(X)$	$P(Y)$
Realisation:	(x, y)	x	y
Information:	$h(x, y)$	$h(x)$	$h(y)$
Venn diagram:			

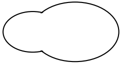
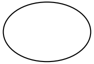
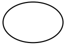
Joint and independent information

	Johnny	Alice	Bob
			
Observations:	(X, Y)	X	Y
Knows:	$P(X, Y)$	$P(X)$	$P(Y)$
Realisation:	(x, y)	x	y
Information:	$h(x, y)$	$h(x)$	$h(y)$
Venn diagram:			

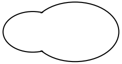
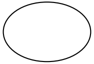
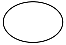
- Johnny knows more than either Alice or Bob.
- This is supported by the inequalities

$$h(x, y) \geq h(x),\ h(y) \geq 0$$
$$h(x|y) = h(x, y) - h(y) \geq 0$$
$$h(y|x) = h(x, y) - h(x) \geq 0$$

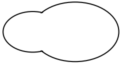
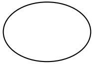
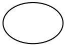
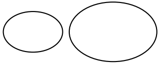
Joint and independent information

	Johnny	Alice	Bob	Indy
				
Observations:	(X, Y)	X	Y	-
Knows:	$P(X, Y)$	$P(X)$	$P(Y)$	
Realisation:	(x, y)	x	y	
Information:	$h(x, y)$	$h(x)$	$h(y)$	
Venn diagram:				

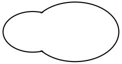
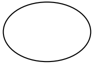
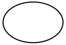
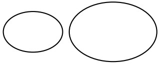
Joint and independent information

	Johnny	Alice	Bob	Indy
				
Observations:	(X, Y)	X	Y	-
Knows:	$P(X, Y)$	$P(X)$	$P(Y)$	$P(X)$ & $P(Y)$
Realisation:	(x, y)	x	y	(x, y)
Information:	$h(x, y)$	$h(x)$	$h(y)$	
Venn diagram:				

Joint and independent information

	Johnny	Alice	Bob	Indy
				
Observations:	(X, Y)	X	Y	-
Knows:	$P(X, Y)$	$P(X)$	$P(Y)$	$P(X)$ & $P(Y)$
Realisation:	(x, y)	x	y	(x, y)
Information:	$h(x, y)$	$h(x)$	$h(y)$	$h(x) + h(y)$
Venn diagram:				

Joint and independent information

	Johnny	Alice	Bob	Indy
				
Observations:	(X, Y)	X	Y	-
Knows:	$P(X, Y)$	$P(X)$	$P(Y)$	$P(X)$ & $P(Y)$
Realisation:	(x, y)	x	y	(x, y)
Information:	$h(x, y)$	$h(x)$	$h(y)$	$h(x) + h(y)$
Venn diagram:				

- Indy can have more or less information than Johnny (no inequality to say otherwise).

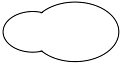
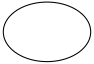
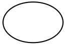
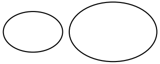
Joint and independent information

	Johnny	Alice	Bob	Indy
Observations:	(X, Y)	X	Y	-
Knows:	$P(X, Y)$	$P(X)$	$P(Y)$	$P(X)$ & $P(Y)$
Realisation:	(x, y)	x	y	(x, y)
Information:	$h(x, y)$	$h(x)$	$h(y)$	$h(x) + h(y)$

Venn diagram:

- Indy can have more or less information than Johnny (no inequality to say otherwise).
- Sometimes Indy thinks he has more information than Johnny despite knowing less.

Joint and independent information

	Johnny	Alice	Bob	Indy
				
Observations:	(X, Y)	X	Y	-
Knows:	$P(X, Y)$	$P(X)$	$P(Y)$	$P(X)$ & $P(Y)$
Realisation:	(x, y)	x	y	(x, y)
Information:	$h(x, y)$	$h(x)$	$h(y)$	$h(x) + h(y)$
Venn diagram:				

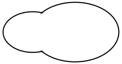
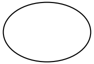
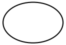
- Indy can have more or less information than Johnny (no inequality to say otherwise).
- Sometimes Indy thinks he has more information than Johnny despite knowing less.
- This occurs because Indy assumes that the marginal realisations are independent.

Marginal information sharing

Idea: replace Indy with Eve who makes no assumptions about the information.

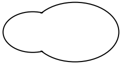
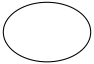
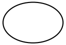
Marginal information sharing

Idea: replace Indy with Eve who makes no assumptions about the information.

	Johnny	Alice	Bob
			
Observations:	(X, Y)	X	Y
Knows:	$P(X, Y)$	$P(X)$	$P(Y)$
Realisation:	(x, y)	x	y
Information:	$h(x, y)$	$h(x)$	$h(y)$
Venn diagram:			

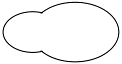
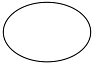
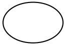
Marginal information sharing

Idea: replace Indy with Eve who makes no assumptions about the information.

	Johnny	Alice	Bob	Eve
				
Observations:	(X, Y)	X	Y	-
Knows:	$P(X, Y)$	$P(X)$	$P(Y)$	
Realisation:	(x, y)	x	y	
Information:	$h(x, y)$	$h(x)$	$h(y)$	
Venn diagram:				

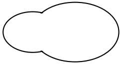
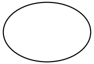
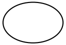
Marginal information sharing

Idea: replace Indy with Eve who makes no assumptions about the information.

	Johnny	Alice	Bob	Eve
				
Observations:	(X, Y)	X	Y	-
Knows:	$P(X, Y)$	$P(X)$	$P(Y)$	$P(X)$ & $P(Y)$
Realisation:	(x, y)	x	y	(x, y)
Information:	$h(x, y)$	$h(x)$	$h(y)$	
Venn diagram:				

Marginal information sharing

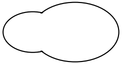
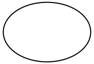
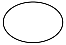
Idea: replace Indy with Eve who makes no assumptions about the information.

	Johnny	Alice	Bob	Eve
				
Observations:	(X, Y)	X	Y	-
Knows:	$P(X, Y)$	$P(X)$	$P(Y)$	$P(X)$ & $P(Y)$
Realisation:	(x, y)	x	y	(x, y)
Information:	$h(x, y)$	$h(x)$	$h(y)$	
Venn diagram:				

- Eve has at least as much information as Alice and Bob, but no more than Johnny.

Marginal information sharing

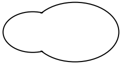
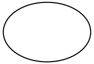
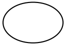
Idea: replace Indy with Eve who makes no assumptions about the information.

	Johnny	Alice	Bob	Eve
				
Observations:	(X, Y)	X	Y	-
Knows:	$P(X, Y)$	$P(X)$	$P(Y)$	$P(X)$ & $P(Y)$
Realisation:	(x, y)	x	y	(x, y)
Information:	$h(x, y)$	$h(x)$	$h(y)$	
Venn diagram:				

- Eve has at least as much information as Alice and Bob, but no more than Johnny.
- Easy to show that Eve's information is given by $\max\big(h(x), h(y)\big) =: h(x \sqcup y)$.

Marginal information sharing

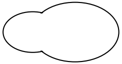
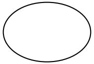
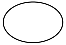
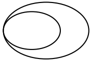
Idea: replace Indy with Eve who makes no assumptions about the information.

	Johnny	Alice	Bob	Eve
				
Observations:	(X, Y)	X	Y	-
Knows:	$P(X, Y)$	$P(X)$	$P(Y)$	$P(X)$ & $P(Y)$
Realisation:	(x, y)	x	y	(x, y)
Information:	$h(x, y)$	$h(x)$	$h(y)$	$\max\big(h(x), h(y)\big)$
Venn diagram:				

- Eve has at least as much information as Alice and Bob, but no more than Johnny.
- Easy to show that Eve's information is given by $\max\big(h(x), h(y)\big) =: h(x \sqcup y)$.

Marginal information sharing

Idea: replace Indy with Eve who makes no assumptions about the information.

	Johnny	Alice	Bob	Eve
				
Observations:	(X, Y)	X	Y	-
Knows:	$P(X, Y)$	$P(X)$	$P(Y)$	$P(X)$ & $P(Y)$
Realisation:	(x, y)	x	y	(x, y)
Information:	$h(x, y)$	$h(x)$	$h(y)$	$\max\big(h(x), h(y)\big)$
Venn diagram:				

- Eve has at least as much information as Alice and Bob, but no more than Johnny.
- Easy to show that Eve's information is given by $\max\big(h(x), h(y)\big) =: h(x \sqcup y)$.

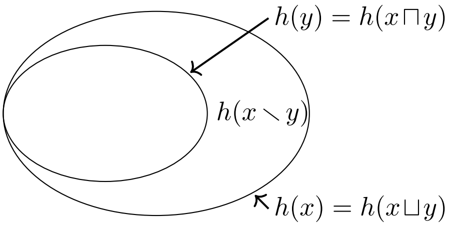
Union and intersection information content

- Union information content

$$h(x \sqcup y) = \max\big(h(x), h(y)\big).$$

- The union information content satisfies

$$h(x)+h(y) \geq h(x \sqcup y) \geq h(x),\ h(y) \geq 0$$



Union and intersection information content

- Union information content

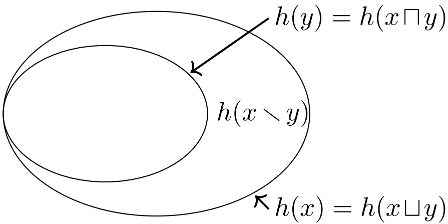
$$h(x \sqcup y) = \max\big(h(x), h(y)\big).$$

- The union information content satisfies

$$h(x) + h(y) \geq h(x \sqcup y) \geq h(x),\ h(y) \geq 0$$

- Unique information content

$$\begin{aligned} h(x \setminus y) &= h(x \sqcup y) - h(y) \\ &= \max\big(h(x) - h(y), 0\big) \geq 0 \\ h(y \setminus x) &= h(x \sqcup y) - h(x) \\ &= \max\big(0, h(y) - h(x)\big) \geq 0 \end{aligned}$$



Union and intersection information content

- Union information content

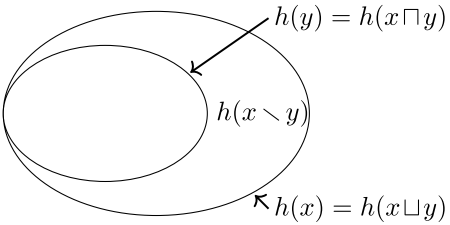
$$h(x \sqcup y) = \max\big(h(x), h(y)\big).$$

- The union information content satisfies

$$h(x) + h(y) \geq h(x \sqcup y) \geq h(x),\ h(y) \geq 0$$

- Unique information content

$$\begin{aligned} h(x \setminus y) &= h(x \sqcup y) - h(y) \\ &= \max\big(h(x) - h(y), 0\big) \geq 0 \\ h(y \setminus x) &= h(x \sqcup y) - h(x) \\ &= \max\big(0, h(y) - h(x)\big) \geq 0 \end{aligned}$$



- Intersection information content

$$\begin{aligned} h(x \sqcap y) &= h(x) + h(y) - h(x \sqcup y) \\ &= \min\big(h(x), h(y)\big) \geq 0. \end{aligned}$$

Union and intersection information content

- Union information content

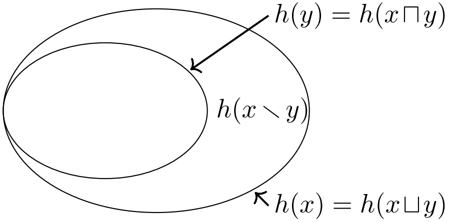
$$h(x \sqcup y) = \max\big(h(x), h(y)\big).$$

- The union information content satisfies

$$h(x)+h(y) \geq h(x \sqcup y) \geq h(x),\ h(y) \geq 0$$

- Unique information content

$$\begin{aligned} h(x \setminus y) &= h(x \sqcup y) - h(y) \\ &= \max\big(h(x) - h(y), 0\big) \geq 0 \\ h(y \setminus x) &= h(x \sqcup y) - h(x) \\ &= \max\big(0, h(y) - h(x)\big) \geq 0 \end{aligned}$$



- Intersection information content

$$\begin{aligned} h(x \sqcap y) &= h(x) + h(y) - h(x \sqcup y) \\ &= \min\big(h(x), h(y)\big) \geq 0. \end{aligned}$$

- Decomposition

$$h(x \sqcup y) = h(x \sqcap y) + h(x \setminus y) + h(y \setminus x)$$

Union and intersection entropy

- Union entropy

$$H(X \sqcup Y) = \mathrm{E}_{XY}\left[h(x \sqcup y)\right]$$

- The union entropy satisfies

$$H(X) + H(Y) \geq H(X \sqcup Y) \geq H(X),\ H(Y) \geq 0$$

- Unique information content

$$H(X \setminus Y) = \mathrm{E}_{XY}\left[h(x \smallsetminus y)\right]$$
$$H(Y \setminus X) = \mathrm{E}_{XY}\left[h(y \smallsetminus x)\right]$$

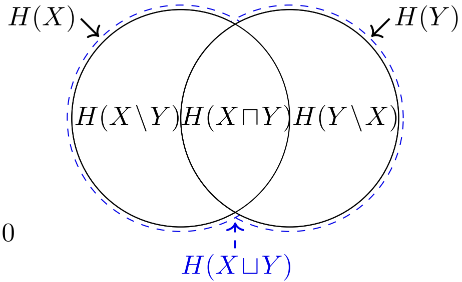
- Intersection information content

$$\begin{aligned} H(X \sqcap Y) &= H(X) + H(Y) - H(X \sqcup Y) \\ &= \mathrm{E}_{XY}\left[h(x \sqcap y)\right] \end{aligned}$$

- Decomposition

$$H(X \sqcup Y) = H(X \sqcap Y) + H(X \setminus Y) + H(Y \setminus X)$$

Union and intersection entropy



- Union entropy

$$H(X \sqcup Y) = \mathrm{E}_{XY}\left[h(x \sqcup y)\right]$$

- The union entropy satisfies

$$H(X)+H(Y) \geq H(X\sqcup Y) \geq H(X),\ H(Y) \geq 0$$

- Unique information content

$$H(X \setminus Y) = \mathrm{E}_{XY}\left[h(x \searrow y)\right]$$
$$H(Y \setminus X) = \mathrm{E}_{XY}\left[h(y \searrow x)\right]$$

- Intersection information content

$$H(X \sqcap Y) = H(X) + H(Y) - H(X \sqcup Y)$$
$$= \mathrm{E}_{XY}\left[h(x \sqcap y)\right]$$

- Decomposition

$$H(X \sqcup Y) = H(X \sqcap Y) + H(X \setminus Y) + H(Y \setminus X)$$

Generalised marginal information sharing

- Thus far, we have only considered two marginal observers, Alice and Bob.
- With a third observer, Charlie, we could consider three-way information sharing $h(x \sqcup y \sqcup z)$.
- We could also consider sharing information through intermediaries. For example:

Generalised marginal information sharing

- Thus far, we have only considered two marginal observers, Alice and Bob.
- With a third observer, Charlie, we could consider three-way information sharing $h(x \sqcup y \sqcup z)$.
- We could also consider sharing information through intermediaries. For example:
  - If Alice and Bob share their information with Dan, then his information will be given by the union information $h(x \sqcup y)$.

Generalised marginal information sharing

- Thus far, we have only considered two marginal observers, Alice and Bob.
- With a third observer, Charlie, we could consider three-way information sharing $h(x \sqcup y \sqcup z)$.
- We could also consider sharing information through intermediaries. For example:
  - If Alice and Bob share their information with Dan, then his information will be given by the union information $h(x \sqcup y)$.
  - If Charlie and Dan subsequently share their information with Eve, then her information will be given by $h\big((x \sqcup y) \sqcup z\big)$.

Generalised marginal information sharing

- Thus far, we have only considered two marginal observers, Alice and Bob.
- With a third observer, Charlie, we could consider three-way information sharing $h(x \sqcup y \sqcup z)$.
- We could also consider sharing information through intermediaries. For example:
  - If Alice and Bob share their information with Dan, then his information will be given by the union information $h(x \sqcup y)$.
  - If Charlie and Dan subsequently share their information with Eve, then her information will be given by $h\big((x \sqcup y) \sqcup z\big)$.
  - Since Eve ultimately ends up with the same marginal information, we would expect that
$$h(x \sqcup y \sqcup z) = h\big((x \sqcup y) \sqcup z\big).$$

Generalised marginal information sharing

- Thus far, we have only considered two marginal observers, Alice and Bob.
- With a third observer, Charlie, we could consider three-way information sharing $h(x \sqcup y \sqcup z)$.
- We could also consider sharing information through intermediaries. For example:
  - If Alice and Bob share their information with Dan, then his information will be given by the union information $h(x \sqcup y)$.
  - If Charlie and Dan subsequently share their information with Eve, then her information will be given by $h\big((x \sqcup y) \sqcup z\big)$.
  - Since Eve ultimately ends up with the same marginal information, we would expect that
$$h(x \sqcup y \sqcup z) = h\big((x \sqcup y) \sqcup z\big).$$
- To understand the distinct ways of sharing marginal information we must understand the algebraic properties of the union and intersection information content.

Algebraic properties of share marginal information content

- Idempotent

$$h(x \sqcup x) = h(x)$$
$$h(x \sqcap x) = h(x)$$

Algebraic properties of share marginal information content

- Idempotent

$$h(x \sqcup x) = h(x)$$
$$h(x \sqcap x) = h(x)$$

- Commutative

$$h(x \sqcup y) = h(y \sqcup x)$$
$$h(x \sqcap y) = h(y \sqcap x)$$

Algebraic properties of share marginal information content

- Idempotent

$$h(x \sqcup x) = h(x)$$
$$h(x \sqcap x) = h(x)$$

- Commutative

$$h(x \sqcup y) = h(y \sqcup x)$$
$$h(x \sqcap y) = h(y \sqcap x)$$

- Associative

$$\begin{aligned} h(x \sqcup y \sqcup z) &= h\big((x \sqcup y) \sqcup z\big) \\ &= h\big(x \sqcup (y \sqcup z)\big) \\ h(x \sqcap y \sqcap z) &= h\big((x \sqcap y) \sqcap z\big) \\ &= h\big(x \sqcap (y \sqcap z)\big) \end{aligned}$$

Algebraic properties of share marginal information content

- Idempotent

$$h(x \sqcup x) = h(x)$$
$$h(x \sqcap x) = h(x)$$

- Commutative

$$h(x \sqcup y) = h(y \sqcup x)$$
$$h(x \sqcap y) = h(y \sqcap x)$$

- Associative

$$\begin{aligned} h(x \sqcup y \sqcup z) &= h\big((x \sqcup y) \sqcup z\big) \\ &= h\big(x \sqcup (y \sqcup z)\big) \\ h(x \sqcap y \sqcap z) &= h\big((x \sqcap y) \sqcap z\big) \\ &= h\big(x \sqcap (y \sqcap z)\big) \end{aligned}$$

- Absorption

$$h\big(x \sqcup (x \sqcap y)\big) = h(x)$$
$$h\big(x \sqcap (x \sqcup y)\big) = h(x)$$

Algebraic properties of share marginal information content

- **Idempotent**

$$h(x \sqcup x) = h(x)$$
$$h(x \sqcap x) = h(x)$$

- **Commutative**

$$h(x \sqcup y) = h(y \sqcup x)$$
$$h(x \sqcap y) = h(y \sqcap x)$$

- **Associative**

$$\begin{aligned} h(x \sqcup y \sqcup z) &= h\big((x \sqcup y) \sqcup z\big) \\ &= h\big(x \sqcup (y \sqcup z)\big) \\ h(x \sqcap y \sqcap z) &= h\big((x \sqcap y) \sqcap z\big) \\ &= h\big(x \sqcap (y \sqcap z)\big) \end{aligned}$$

- **Absorption**

$$h\big(x \sqcup (x \sqcap y)\big) = h(x)$$
$$h\big(x \sqcap (x \sqcup y)\big) = h(x)$$

- **Distributive**

$$h\big(x \sqcup (y \sqcap z)\big) = h\big((x \sqcup y) \sqcap (x \sqcup z)\big)$$
$$h\big(x \sqcap (y \sqcup z)\big) = h\big((x \sqcap y) \sqcup (x \sqcap z)\big)$$

Algebraic properties of share marginal information content

- Idempotent

$$h(x \sqcup x) = h(x)$$
$$h(x \sqcap x) = h(x)$$

- Commutative

$$h(x \sqcup y) = h(y \sqcup x)$$
$$h(x \sqcap y) = h(y \sqcap x)$$

- Associative

$$\begin{aligned} h(x \sqcup y \sqcup z) &= h\big((x \sqcup y) \sqcup z\big) \\ &= h\big(x \sqcup (y \sqcup z)\big) \\ h(x \sqcap y \sqcap z) &= h\big((x \sqcap y) \sqcap z\big) \\ &= h\big(x \sqcap (y \sqcap z)\big) \end{aligned}$$

- Absorption

$$h\big(x \sqcup (x \sqcap y)\big) = h(x)$$
$$h\big(x \sqcap (x \sqcup y)\big) = h(x)$$

- Distributive

$$h\big(x \sqcup (y \sqcap z)\big) = h\big((x \sqcup y) \sqcap (x \sqcup z)\big)$$
$$h\big(x \sqcap (y \sqcup z)\big) = h\big((x \sqcap y) \sqcup (x \sqcap z)\big)$$

- Connexity, i.e. either

$$h(x \sqcup y) = h(x) \text{ and } h(x \sqcap y) = h(y)$$

or

$$h(x \sqcup y) = h(y) \text{ and } h(x \sqcap y) = h(x)$$

Algebraic structure of shared marginal information content

- Idempotent, commutative, associative and connected by absorbtion implies a lattice.

Algebraic structure of shared marginal information content

- Idempotent, commutative, associative and connected by absorbtion implies a lattice.
- If the operators are also distributive then its a distributive lattice.

Algebraic structure of shared marginal information content

- Idempotent, commutative, associative and connected by absorbtion implies a lattice.
- If the operators are also distributive then its a distributive lattice.
- If they are also connex, then the algebraic structure is a total order.

Algebraic structure of shared marginal information content

- Idempotent, commutative, associative and connected by absorbtion implies a lattice.
- If the operators are also distributive then its a distributive lattice.
- If they are also connex, then the algebraic structure is a total order.

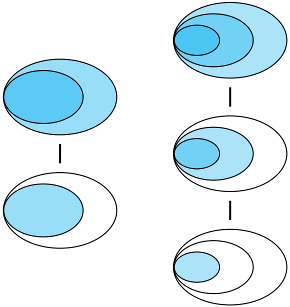
$$h(x \sqcup y) = h(x)$$

$$h(x \sqcap y) = h(y)$$

$$h(x \sqcup y \sqcup z) = h(x) = h(x \sqcup y) = h(x \sqcup z)$$

$$h(x \sqcap y) = h(y \sqcup z) = h(y)$$

$$h(x \sqcap y \sqcap z) = h(z) = h(x \sqcap z) = h(y \sqcap z)$$



Algebraic properties of share marginal entropy

- Idempotent

$$H(X \sqcup X) = H(X)$$
$$H(X \sqcap X) = H(X)$$

- Commutative

$$H(X \sqcup Y) = H(Y \sqcup X)$$
$$H(X \sqcap Y) = H(Y \sqcap X)$$

- Associative

$$\begin{aligned} H(X \sqcup Y \sqcup Z) &= H\big((X \sqcup Y) \sqcup Z\big) \\ &= H\big(X \sqcup (Y \sqcup Z)\big) \\ H(X \sqcap Y \sqcap Z) &= H\big((X \sqcap Y) \sqcap Z\big) \\ &= H\big(X \sqcap (Y \sqcap Z)\big) \end{aligned}$$

- Absorption

$$H\big(X \sqcup (X \sqcap Y)\big) = H(X)$$
$$H\big(X \sqcap (X \sqcup Y)\big) = H(X)$$

- Distributive

$$H\big(X \sqcup (Y \sqcap Z)\big) = H\big((X \sqcup Y) \sqcap (X \sqcup Z)\big)$$
$$H\big(X \sqcap (Y \sqcup Z)\big) = H\big((X \sqcap Y) \sqcup (X \sqcap Z)\big)$$

Algebraic properties of share marginal entropy

- Idempotent

$$H(X \sqcup X) = H(X)$$
$$H(X \sqcap X) = H(X)$$

- Commutative

$$H(X \sqcup Y) = H(Y \sqcup X)$$
$$H(X \sqcap Y) = H(Y \sqcap X)$$

- Associative

$$H(X \sqcup Y \sqcup Z) = H\big((X \sqcup Y) \sqcup Z\big) = H\big(X \sqcup (Y \sqcup Z)\big)$$
$$H(X \sqcap Y \sqcap Z) = H\big((X \sqcap Y) \sqcap Z\big) = H\big(X \sqcap (Y \sqcap Z)\big)$$

- Absorption

$$H\big(X \sqcup (X \sqcap Y)\big) = H(X)$$
$$H\big(X \sqcap (X \sqcup Y)\big) = H(X)$$

- Distributive

$$H\big(X \sqcup (Y \sqcap Z)\big) = H\big((X \sqcup Y) \sqcap (X \sqcup Z)\big)$$
$$H\big(X \sqcap (Y \sqcup Z)\big) = H\big((X \sqcap Y) \sqcup (X \sqcap Z)\big)$$

- Connexity is the only property that does not hold for the entropy.

Algebraic properties of share marginal entropy

- Idempotent

$$H(X \sqcup X) = H(X)$$
$$H(X \sqcap X) = H(X)$$

- Commutative

$$H(X \sqcup Y) = H(Y \sqcup X)$$
$$H(X \sqcap Y) = H(Y \sqcap X)$$

- Associative

$$\begin{aligned} H(X \sqcup Y \sqcup Z) &= H\big((X \sqcup Y) \sqcup Z\big) \\ &= H\big(X \sqcup (Y \sqcup Z)\big) \\ H(X \sqcap Y \sqcap Z) &= H\big((X \sqcap Y) \sqcap Z\big) \\ &= H\big(X \sqcap (Y \sqcap Z)\big) \end{aligned}$$

- Absorption

$$H\big(X \sqcup (X \sqcap Y)\big) = H(X)$$
$$H\big(X \sqcap (X \sqcup Y)\big) = H(X)$$

- Distributive

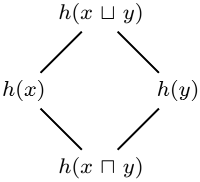
$$H\big(X \sqcup (Y \sqcap Z)\big) = H\big((X \sqcup Y) \sqcap (X \sqcup Z)\big)$$
$$H\big(X \sqcap (Y \sqcup Z)\big) = H\big((X \sqcap Y) \sqcup (X \sqcap Z)\big)$$

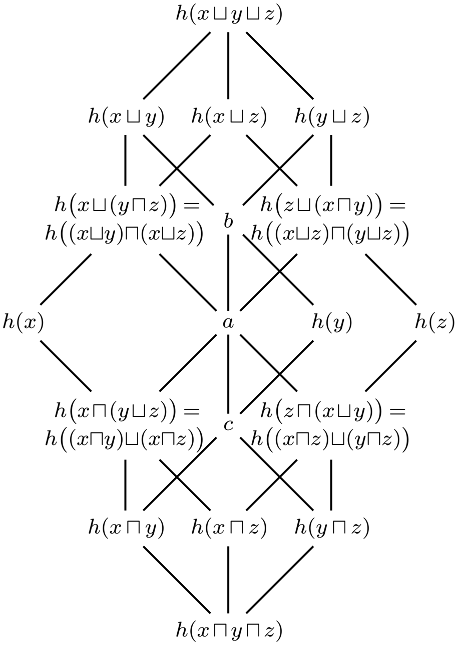
- Connexity is the only property that does not hold for the entropy.
- Therefore, the shared marginal entropy forms a distributive lattice.

$$
\begin{aligned}
a &= h\big((x \sqcup y) \sqcap (x \sqcup z) \sqcap (y \sqcup z)\big)\\
&= h\Big(\big(x \sqcup (y \sqcap z)\big) \sqcap \big(y \sqcup (x \sqcap z)\big)\Big)\\
&= h\Big(\big(x \sqcup (y \sqcap z)\big) \sqcap \big(z \sqcup (x \sqcap y)\big)\Big)\\
&= h\Big(\big(y \sqcup (x \sqcap z)\big) \sqcap \big(z \sqcup (x \sqcap y)\big)\Big)\\
&= h\Big(\big(y \sqcap (x \sqcup z)\big) \sqcup \big(z \sqcap (x \sqcup y)\big)\Big)\\
&= h\Big(\big(x \sqcap (y \sqcup z)\big) \sqcup \big(z \sqcap (x \sqcup y)\big)\Big)\\
&= h\Big(\big(x \sqcap (y \sqcup z)\big) \sqcup \big(y \sqcap (x \sqcup z)\big)\Big)\\
&= h\big((x \sqcap y) \sqcup (x \sqcap z) \sqcup (y \sqcap z)\big)
\end{aligned}
$$

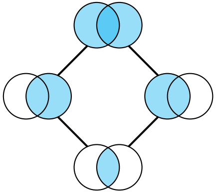
$$b = h\big(y \sqcup (x \sqcap z)\big) = h\big((x \sqcup y) \sqcap (y \sqcup z)\big)$$

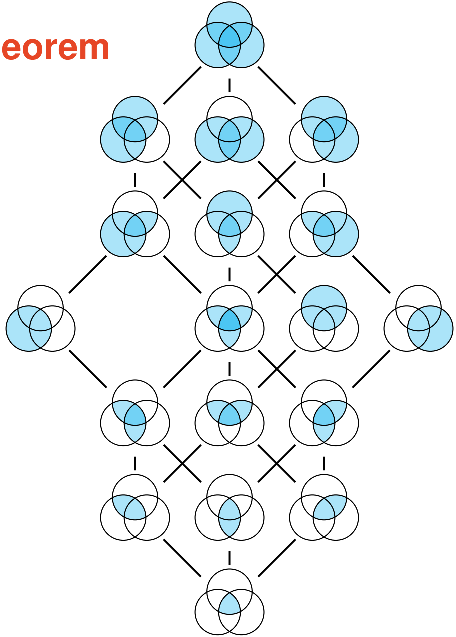
$$c = h\big(y \sqcap (x \sqcup z)\big) = h\big((x \sqcap y) \sqcup (y \sqcap z)\big)$$



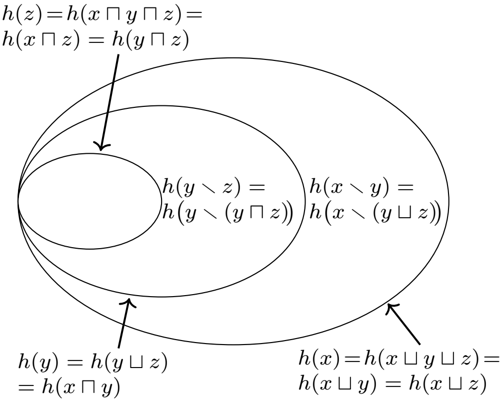


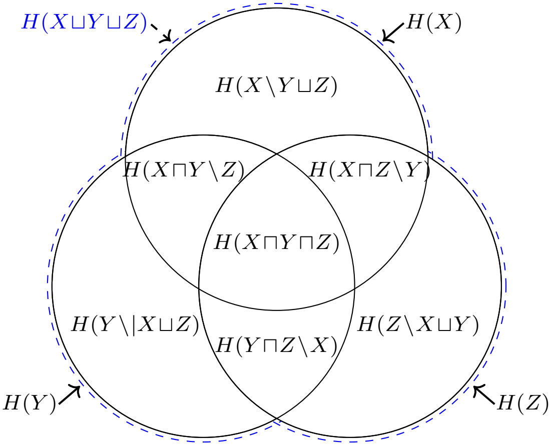
Birkhoff’s representation theorem





Shared marginal information content and entropy



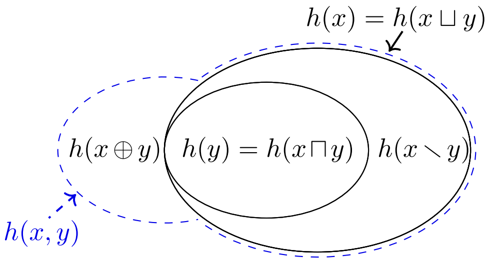


Relating PID to Shared Marginal Information

Synergistic information content

- Eve has no more information than Johnny

$$h(x, y) \geq h(x \sqcup y)$$



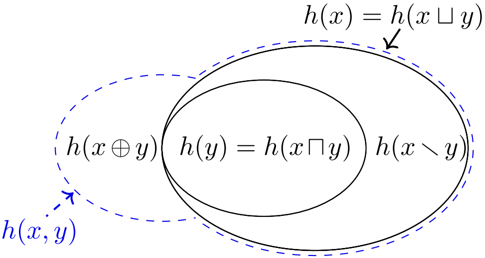
Synergistic information content

- Eve has no more information than Johnny

$$h(x, y) \geq h(x \sqcup y)$$

- Synergistic information content

$$h(x \oplus y) = h(x, y) - h(x \sqcup y)$$
$$= \min\big(h(y|x), h(x|y)\big) \geq 0$$



Synergistic information content

- Eve has no more information than Johnny

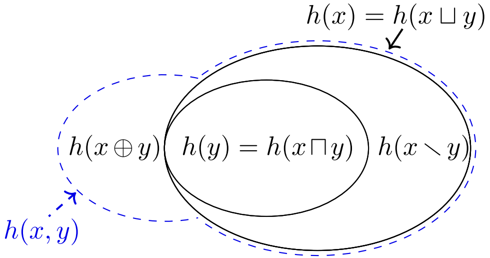
$$h(x, y) \geq h(x \sqcup y)$$

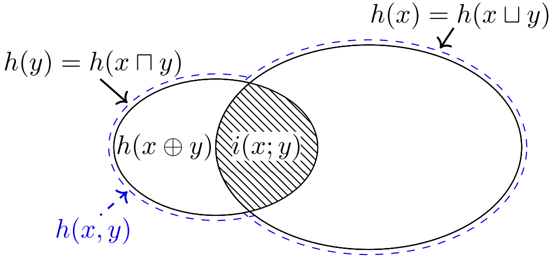
- Synergistic information content

$$h(x \oplus y) = h(x, y) - h(x \sqcup y) = \min\big(h(y|x), h(x|y)\big) \geq 0$$

- Mutual information content

$$i(x; y) = h(x \sqcup y) - h(x \oplus y)$$





Synergistic information content

- Eve has no more information than Johnny

$$h(x, y) \geq h(x \sqcup y)$$

- Synergistic information content

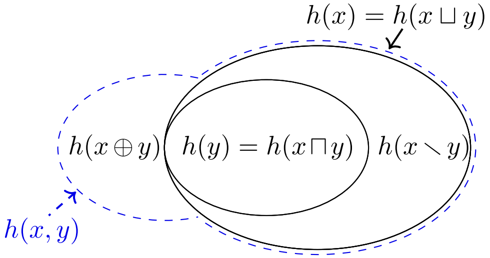
$$h(x \oplus y) = h(x, y) - h(x \sqcup y)$$
$$= \min\big(h(y|x), h(x|y)\big) \geq 0$$

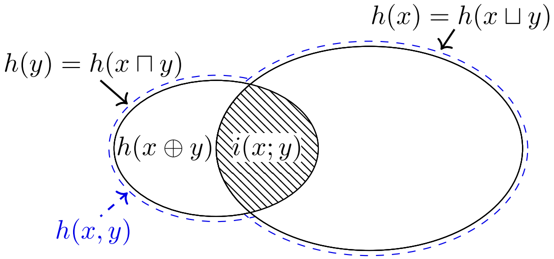
- Mutual information content

$$i(x; y) = h(x \sqcup y) - h(x \oplus y)$$

- Decomposition

$$h(x, y) = h(x \setminus y) + h(y \setminus x) + h(x \sqcap y) + h(x \oplus y)$$

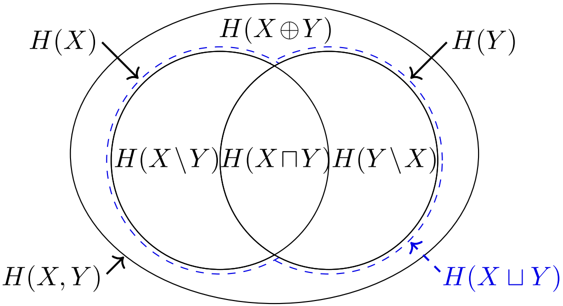




Synergistic entropy

- Synergistic entropy

$$H(X \oplus Y) = H(X,Y) - H(X \sqcup Y)$$
$$= \mathrm{E}_{XY}\left[h(x \oplus y))\right] \geq 0$$



- Mutual information

$$I(X;Y) = H(X \sqcup Y) - H(X \oplus Y)$$

- Decomposition

$$H(X,Y) = H(X \setminus Y) + H(Y \setminus X) + H(X \sqcap Y) + H(X \oplus Y)$$

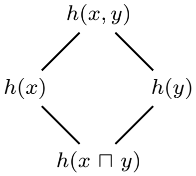
- We can also generalise this argument to any number of joint sources.
  - The redundancy lattice from PID then appears as a by-product (a sub-algebra).

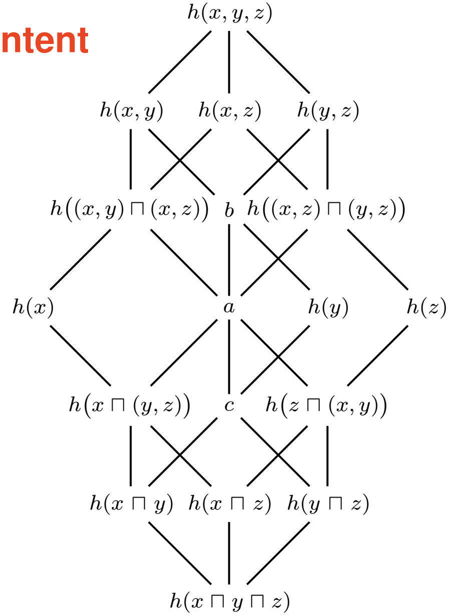
Redundancy lattice for info content

$$a = h\big((x,y) \sqcap (x,z) \sqcap (y,z)\big)$$

$$b = h\big((x,y) \sqcap (y,z)\big)$$

$$c = h\big(y \sqcap (x,z)\big)$$





Recovering the redundancy lattice

We can consider conditional variants of the shared marginal information contents, e.g.

$$h(x \sqcap y|z) = h(x|z) + h(y|z) - h(x \sqcup y|z) = \min\big(h(x|z), h(y|z)\big).$$

We can evaluate the equivalent pointwise mutual information term, e.g.

$$i(s_1 \sqcap s_2; t) = h(s_1 \sqcap s_2) - h(s_1 \sqcap s_2|t).$$

- Complete argument for this is provided in (Finn and Lizier, 2018b).
- This yields a pointwise partial information decomposition (Finn and Lizier, 2018a).
- Take the expectation to recover partial information decomposition.

Conclusions

Takeaway points

Information decomposition is an interesting and active area of information theory.

- Theory is not yet completely understood.
- The are a ton of potential applications.

Takeaway points

Information decomposition is an interesting and active area of information theory.

- Theory is not yet completely understood.
- The are a ton of potential applications.

Regarding my research:

- The union and intersection information content are fundamental quantities.
- Birkoff's representation theorem rigorously connects them to the algebra of sets.
- The redundancy lattice appears as a by product when considering joint variables.
- Pointwise PID is reasonably well developed (more so that most other approaches).
  - Works for the information content as well as the mutual information.
  - One of the only approaches that works for an arbitrary number of sources.

Potential applications

- Neuroscience:
  - Information theory can measure neural information storage and transfer
  - Quantifying information modification requires multivariate information theory

Potential applications

- Neuroscience:
  - Information theory can measure neural information storage and transfer
  - Quantifying information modification requires multivariate information theory
- Feature selection in machine learning:
  - Consider a data set with known heart disease risk factors:
  - Smoker or non-smoker might contribute a large amount of unique information;
  - Obesity and diabetes might be largely redundant;
  - Genetic risks or age might be important synergistically with other features.

Potential applications

- Neuroscience:
  - Information theory can measure neural information storage and transfer
  - Quantifying information modification requires multivariate information theory
- Feature selection in machine learning:
  - Consider a data set with known heart disease risk factors:
  - Smoker or non-smoker might contribute a large amount of unique information;
  - Obesity and diabetes might be largely redundant;
  - Genetic risks or age might be important synergistically with other features.
- Network coding:
  - High-dimensional redundancies need to be removed.
  - Shannon's theory is not a very useful for network coding.

Future work

- Further understand the algebraic structure of multivariate information.
- Relating the existing approaches.
- Continuous information decomposition

References

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

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

Conor Finn and Joseph T. Lizier. Probability mass exclusions and the directed components of mutual information. *Entropy*, 20(11):826, 2018b.

Virgil Griffith and Christof Koch. Quantifying synergistic mutual information. In *Guided Self-Organization: Inception*, pages 159–190. Springer, 2014.

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

Joseph Lizier, Nils Bertschinger, Juergen Jost, and Michael Wibral. Information decomposition of target effects from multi-source interactions: Perspectives on previous, current and future work. *Entropy*, 20(4):307, Apr 2018. ISSN 1099-4300. doi: 10.3390/e20040307. URL `http://dx.doi.org/10.3390/e20040307`.

Johannes Rauh, Nils Bertschinger, Eckehard Olbrich, and Jürgen Jost. Reconsidering unique information: Towards a multivariate information decomposition. In *2014 IEEE International Symposium on Information Theory*, pages 2232–2236. IEEE, 2014.

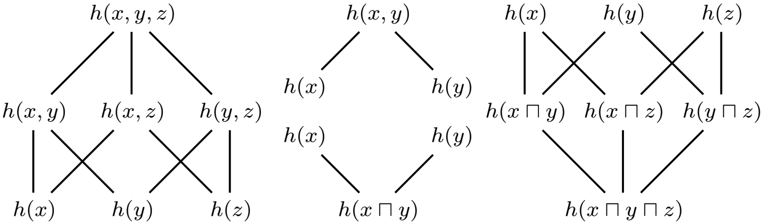
Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

Generalising the synergistic information

- The algebraic structure of joint information is also a lattice.
- Need to explore the relationship with shared marginal information.

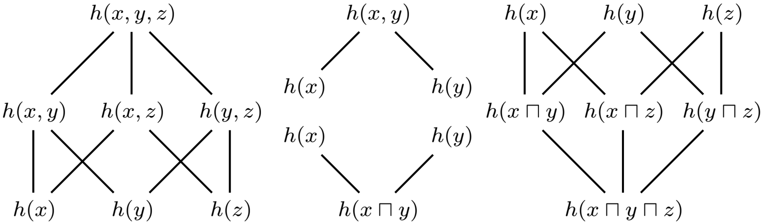
Generalising the synergistic information

- The algebraic structure of joint information is also a lattice.
- Need to explore the relationship with shared marginal information.
- Consider the respective semilattices generated by joint and intersection information.



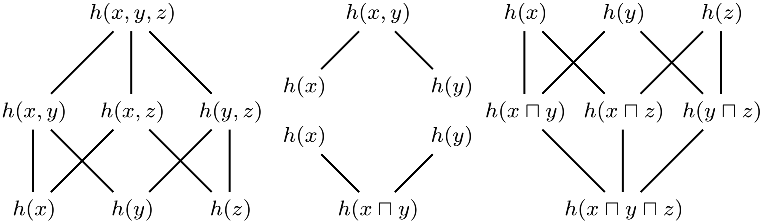
Generalising the synergistic information

- The algebraic structure of joint information is also a lattice.
- Need to explore the relationship with shared marginal information.
- Consider the respective semilattices generated by joint and intersection information.

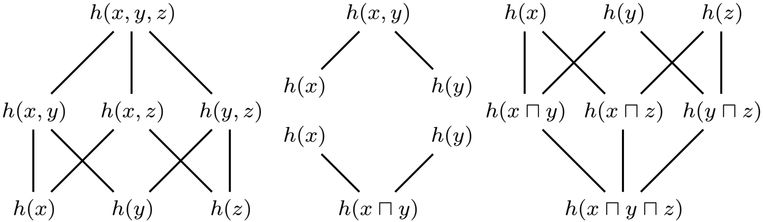


- Are these semilattices be connected by absorption?

Connecting the joint and intersection information

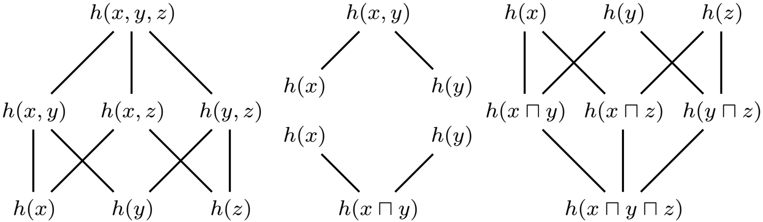


Connecting the joint and intersection information



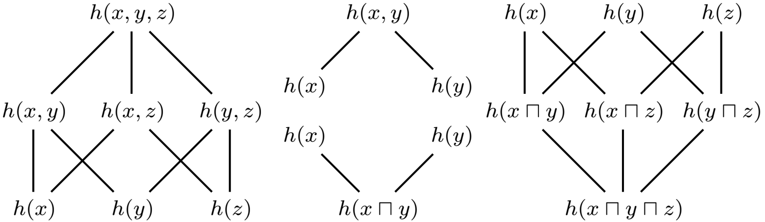
- Intersection information content absorbs the joint information content $h(x \sqcap (x,y))$.

Connecting the joint and intersection information



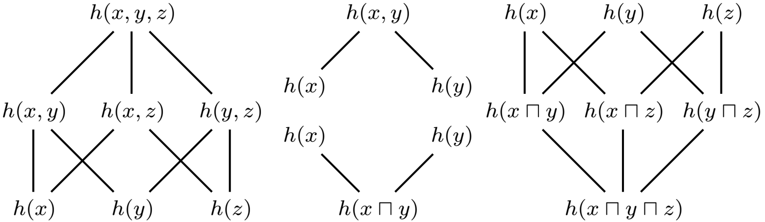
- Intersection information content absorbs the joint information content $h(x \sqcap (x,y))$.
- However, the joint information content does not absorb the intersection information content since $h(x,(x \sqcap y))$ is equal to $h(x,y)$ for $h(x) \geq h(y)$, i.e. is not equal to $h(x)$.

Connecting the joint and intersection information



- Intersection information content absorbs the joint information content $h(x \sqcap (x, y))$.
- However, the joint information content does not absorb the intersection information content since $h(x, (x \sqcap y))$ is equal to $h(x, y)$ for $h(x) \geq h(y)$, i.e. is not equal to $h(x)$.
- Nevertheless, this means that we do get a lattice if we consider the intersection information content of the various joint information contents (but not vice versa).

Connecting the joint and intersection information



- Intersection information content absorbs the joint information content $h(x \sqcap (x,y))$.
- However, the joint information content does not absorb the intersection information content since $h(x, (x \sqcap y))$ is equal to $h(x,y)$ for $h(x) \geq h(y)$, i.e. is not equal to $h(x)$.
- Nevertheless, this means that we do get a lattice if we consider the intersection information content of the various joint information contents (but not vice versa).
- This substructure is the redundancy lattice from partial information decomposition!