Computing the Unique Information

Guido Montúfar
montufar@math.ucla.edu

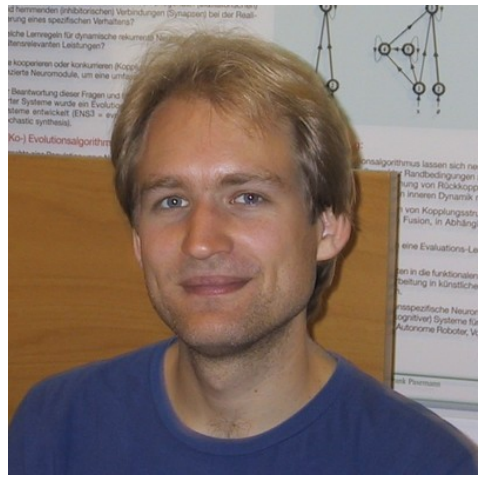
1st Workshop on Semantic Information, CVPR, June 2019

UCLA

MAX-PLANCK-GESELLSCHAFT

erc
European Research Council
Established by the European Commission

Pradeep Kr. Banerjee



Johannes Rauh

Computing the Unique Information, Proc. IEEE ISIT, [BRM18]

1. Disentangling Synergy and Redundancy
2. The Unique Information
3. Computing the Unique Information
4. Applications

Disentangling Synergy and Redundancy

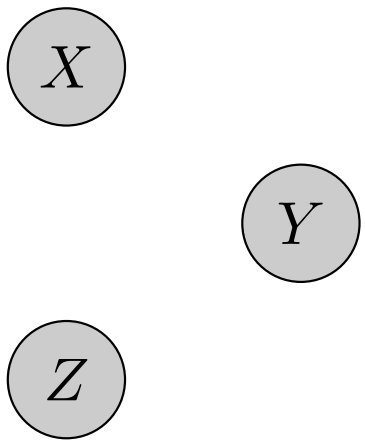
Information Decompositions

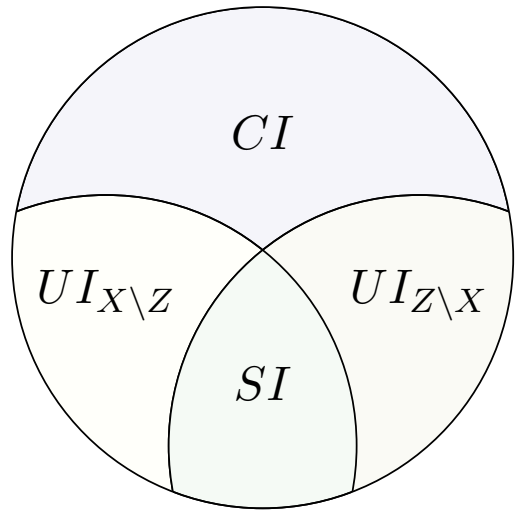
- Consider variables Y, X, Z, where Y is of interest.
  Want to predict Y based on X or Z.
- We can consider the mutual information $I(Y;X)$ or $I(Y;Z)$.
  This measures redundant + unique information of either.
  It does not tell us the unique component.
- We can consider $I(Y;X,Z)$.
  This includes synergistic information that X, Z convey about Y.
  It does not tell synergy and redundancy apart.

Information Decompositions

$$
\begin{cases}
I(Y;X) & = \underbrace{UI(Y;X\backslash Z)}_{\text{unique } X \text{ wrt } Z} + \underbrace{SI(Y;X,Z)}_{\text{shared (redundant)}} \\
I(Y;X|Z) & = \underbrace{UI(Y;X\backslash Z)}_{\text{unique } X \text{ wrt } Z} + \underbrace{CI(Y;X,Z)}_{\text{complementary (synergistic)}} \\
I(Y;X,Z) & = UI(Y;X\backslash Z) + SI(Y;X,Z) \\
& \quad + UI(Y;Z\backslash X) + CI(Y;X,Z)
\end{cases}
$$

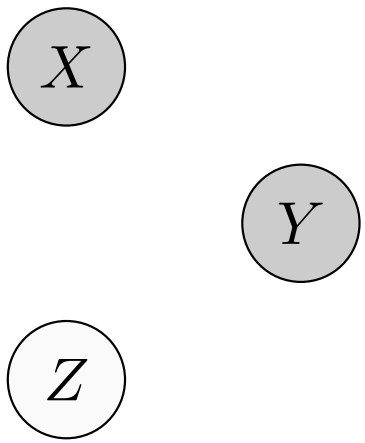
$I(Y; X, Z)$

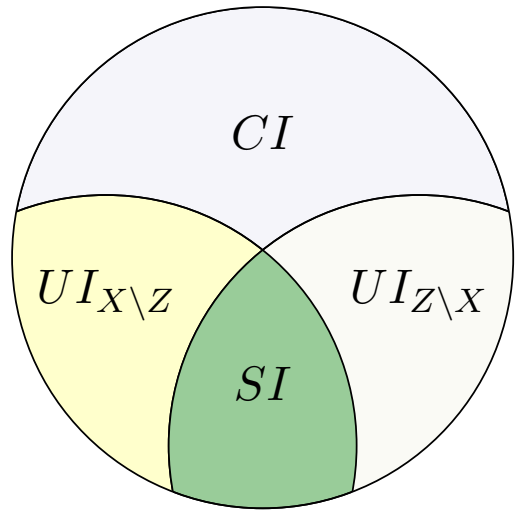




$I(Y; X, Z)$

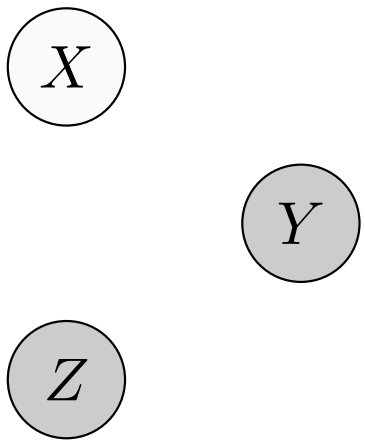
$I(Y;X)$

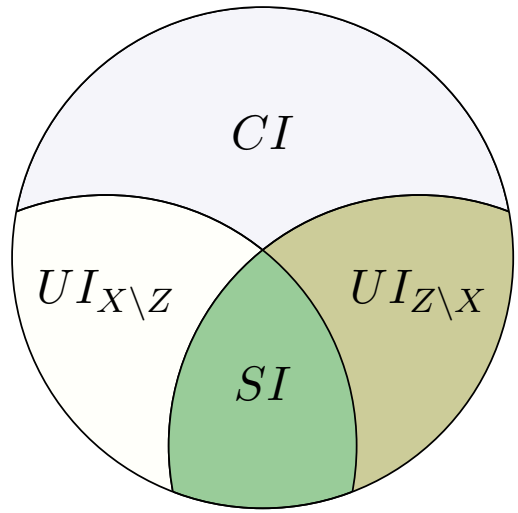




$I(Y;X,Z)$

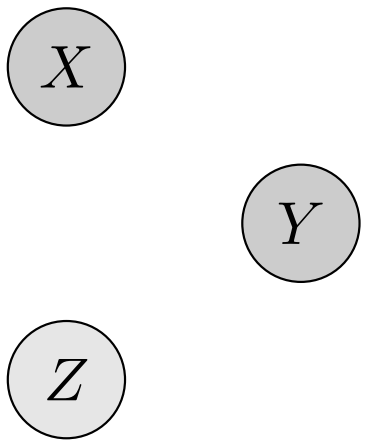
$I(Y;Z)$

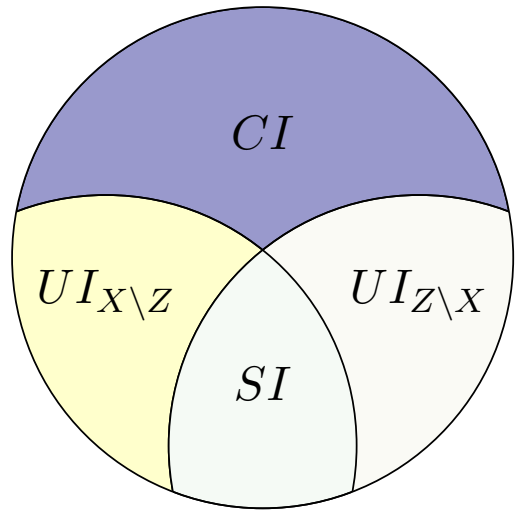




$I(Y;X,Z)$

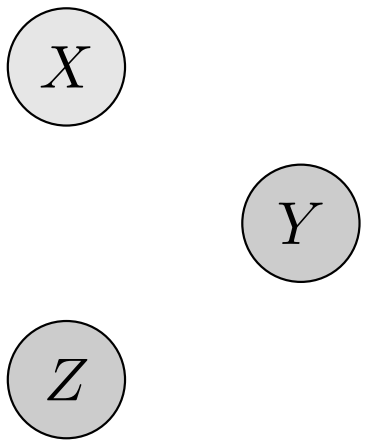
$$I(Y;X|Z)$$

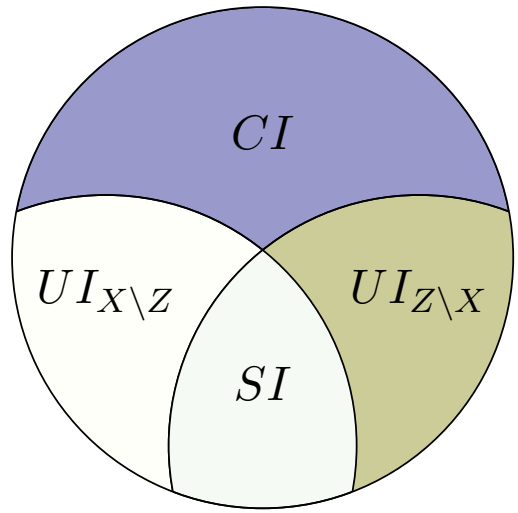




$$I(Y;X,Z)$$

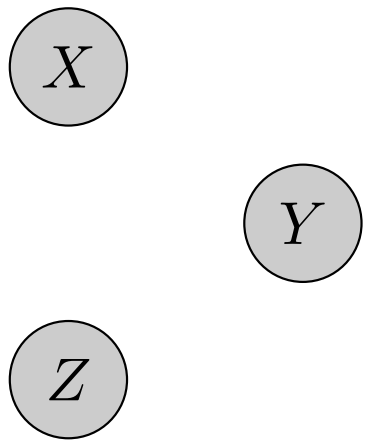
$$I(Y;Z|X)$$

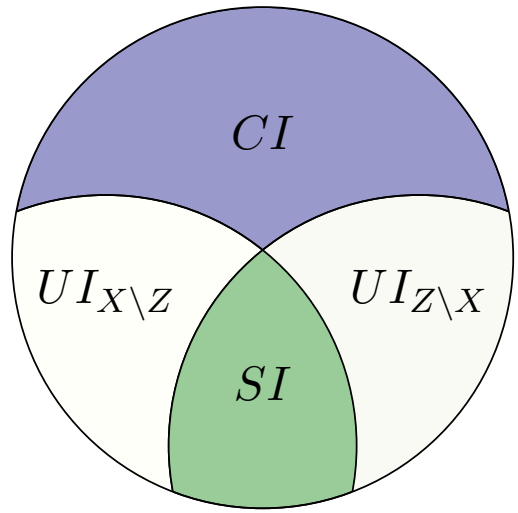




$$I(Y;X,Z)$$

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





$I(Y;X,Z)$

$CoI = SI - CI$ can be negative.

Information is not a conserved quantity

Conditioning on an additional random variable can decrease or increase the mutual information.
$CoI(Y;X;Z) = I(Y;X) - I(Y;X|Z)$ can be positive or negative.

1. If $Y = X = Z$ uniform binary, then

$$0 = I(Y;X|Z) \leq I(Y;X) = 1.$$

X and Z convey the same redundant information about Y.

2. If Y and X independent binary, and $Z = Y \oplus_2 X$, then

$$1 = I(Y;X|Z) \geq I(Y;X) = 0.$$

Neither X nor Z individually convey any information about Y, but jointly they convey synergistic information about Y.

Information is not a conserved quantity

Conditioning on an additional random variable can decrease or increase the mutual information.
$CoI(Y;X;Z) = I(Y;X) - I(Y;X|Z)$ can be positive or negative.

1. If $Y = X = Z$ uniform binary, then

$$0 = I(Y;X|Z) \leq I(Y;X) = 1.$$

X and Z convey the same redundant information about Y.

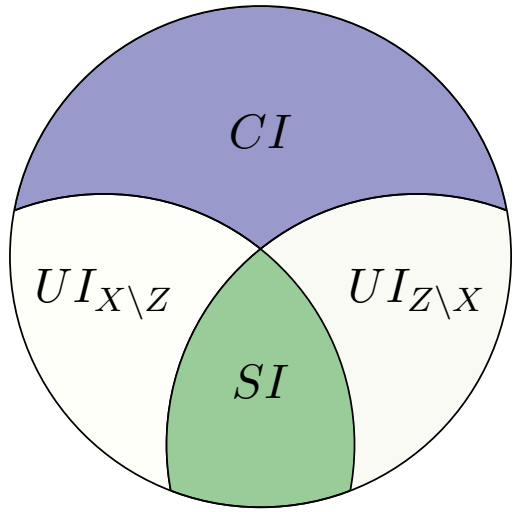
2. If Y and X independent binary, and $Z = Y \oplus_2 X$, then

$$1 = I(Y;X|Z) \geq I(Y;X) = 0.$$

Neither X nor Z individually convey any information about Y, but jointly they convey synergistic information about Y.

Explaining away: Knowledge of a common effect might render positive or negative dependence between the causes

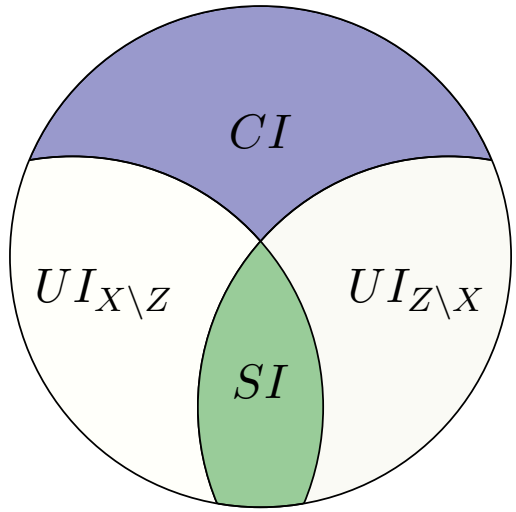
Goal: disentangle synergy and redundancy



$I(Y;X,Z)$

Need to fix a degree of freedom

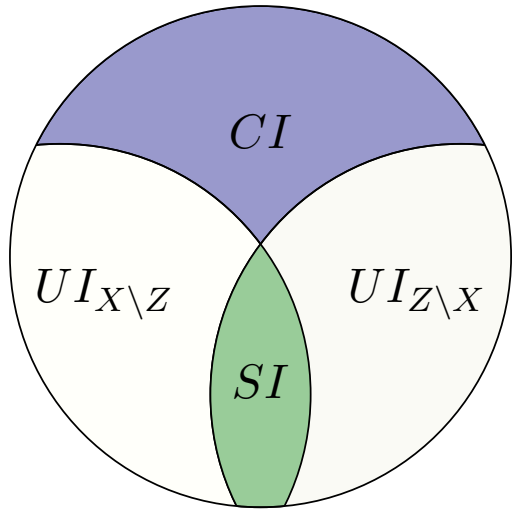
Goal: disentangle synergy and redundancy



$I(Y;X,Z)$

Need to fix a degree of freedom

Goal: disentangle synergy and redundancy



$I(Y; X, Z)$

Need to fix a degree of freedom

How do we measure synergy and redundancy?

- Synergy: the whole is more than the sum of parts (sat Aristotle)
- McGill proposed coinformation (CoI) in [McG54] as a generalization of mutual information. CoI can be negative
- In neuroscience, negative values of CoI are interpreted as synergy and positive values as redundancy
- Williams and Beer (2010) proposed a framework to decompose the mutual information $I(Y; X_1, X_2, \ldots, X_n)$ into nonnegative components
- For the bivariate case, two approaches have gained traction: Harder et al. (2013) and Bertschinger et al. (2014)

Information Axiomatic vs Operational

Following the axiomatic characterization of entropy, Shannon (1948) states:

"The real justification of these definitions resides in their implications"

The Unique Information

Decision problems

- Idea: If X has unique information about Y w.r.t. Z, then there must be a situation in which X "performs better" than Z.

Decision problems

- Idea: If X has unique information about Y w.r.t. Z, then there must be a situation in which X "performs better" than Z.
- Consider a decision problem:

  An agent chooses an action $a \in \mathcal{A}$, with $\mathcal{A}$ finite.

  She receives a reward $u(a, y)$ depending on the outcome y of Y.

  Should she prefer to observe X or Z?

Decision problems

- Idea: If X has unique information about Y w.r.t. Z, then there must be a situation in which X "performs better" than Z.
- Consider a decision problem:

  An agent chooses an action $a \in \mathcal{A}$, with $\mathcal{A}$ finite.

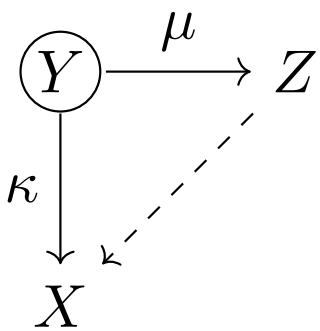
  She receives a reward $u(a, y)$ depending on the outcome y of Y.

  Should she prefer to observe X or Z?
- X has no unique information if it is better to choose Z for any $\mathcal{A}, u$.

  This idea only specifies the zero set of $UI(Y; X \backslash Z)$.

The Unique Information (UI)

Bertschinger, Rauh, Olbrich, Jost, Ay (2014)



$$UI_P(Y;X\backslash Z) := \min_{Q\in\Delta_P} I_Q(Y;X|Z)$$

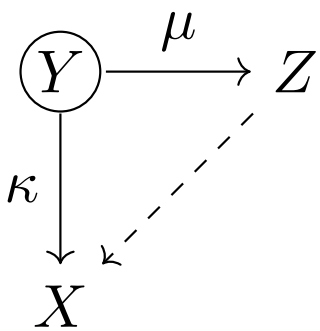
$$\Delta_P = \{Q_{YXZ} \colon Q_{YX} = P_{YX},\ Q_{YZ} = P_{YZ}\}$$

When the diagram commutes, $UI(Y;X\backslash Z)$ vanishes.

When this vanishes, we can always discard X in favor of Z.

The Unique Information (UI)

Bertschinger, Rauh, Olbrich, Jost, Ay (2014)



$$UI_P(Y;X\backslash Z) := \min_{Q\in\Delta_P} I_Q(Y;X|Z)$$

$$\Delta_P = \{Q_{YXZ} \colon Q_{YX} = P_{YX},\ Q_{YZ} = P_{YZ}\}$$

When the diagram commutes, $UI(Y;X\backslash Z)$ vanishes.

When this vanishes, we can always discard X in favor of Z.

Convex program over a polytope of dimension $|\mathcal{Y}|(|\mathcal{X}|-1)(|\mathcal{Z}|-1)$

Property $(*)$

- Choosing between X and Z only depends on the marginal distributions of the pairs (Y, X) and (Y, Z).
- Plausible property for information decompositions:

  $UI(Y; X \backslash Z)$ only depends on the margins (Y, X) and (Y, Z).
- In this case, also $SI(Y; X, Z)$ and $UI(Y; Z \backslash X)$ only depend on the margins (Y, X) and (Y, Z).

  Hence only the synergy involves "interactions" between X and Z.

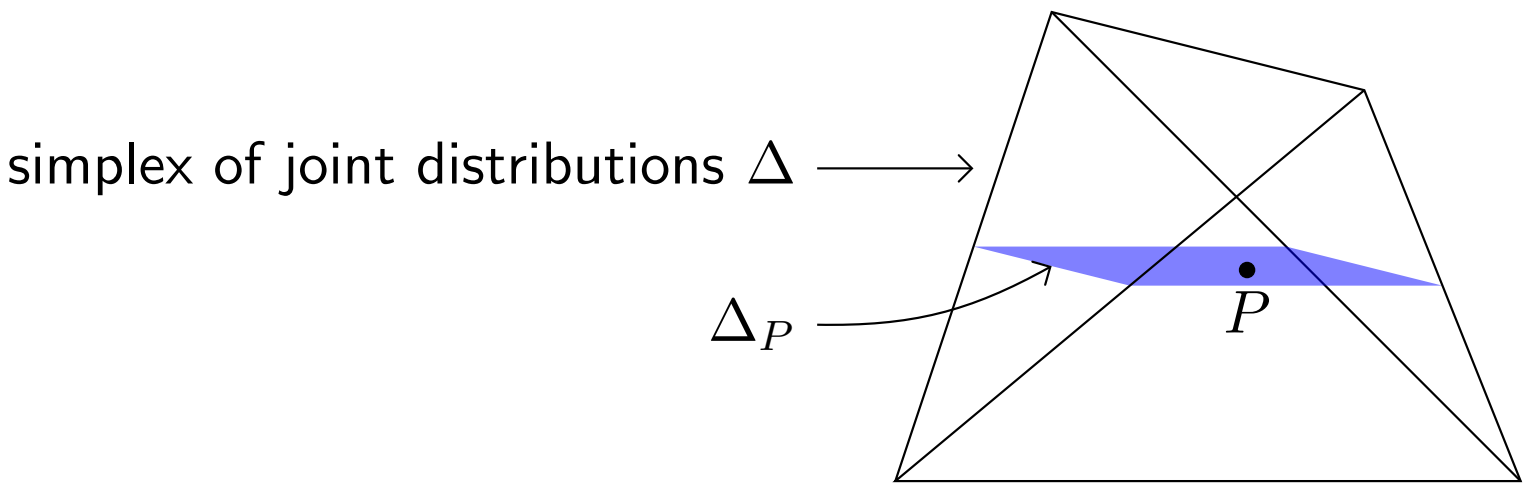
Property $(*)$

Property $(*)$ holds if and only if the functions

$$Q \mapsto UI_Q(Y;X\backslash Z),\ Q \mapsto UI_Q(Y;Z\backslash X),\ Q \mapsto SI_Q(Y;X,Z)$$

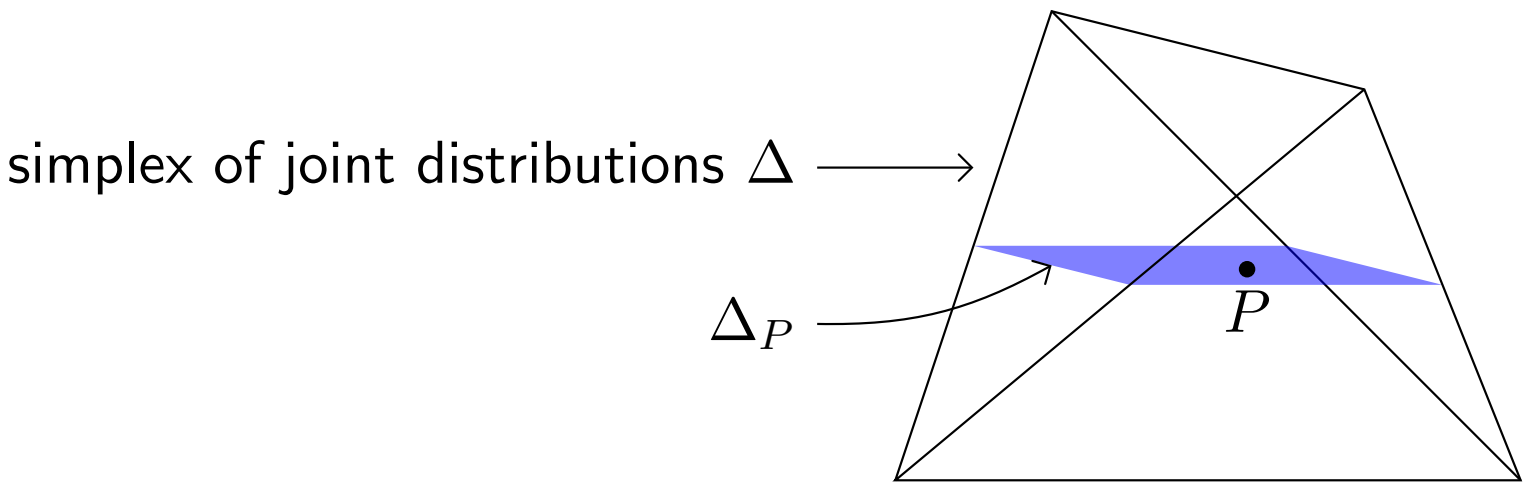
are constant on

$$\Delta_P := \Big\{ Q \in \Delta \ :\ Q(Y,X) = P(Y,X), \quad Q(Y,Z) = P(Y,Z) \Big\}.$$



Property $(**)$

- For each $P \in \Delta$ there is $Q \in \Delta_P$ with $CI_Q(Y; X, Z) = 0$.
- This says that for any choice of the margins, there is a joint distribution with these margins which has zero synergy.



Theorem 1 (BROJA decomposition)

The only information decomposition that satisfies $(*)$ *and* $(**)$ *is:*

$$UI(Y;X\backslash Z) := \min_{Q\in\Delta_P} I_Q(Y;X|Z),$$
$$UI(Y;Z\backslash X) := \min_{Q\in\Delta_P} I_Q(Y;Z|X),$$
$$SI(Y;X,Z) := \max_{Q\in\Delta_P} CoI_Q(Y;X;Z),$$
$$CI(Y;X,Z) := I(Y;X,Z) - \min_{Q\in\Delta_P} I_Q(Y;X,Z).$$

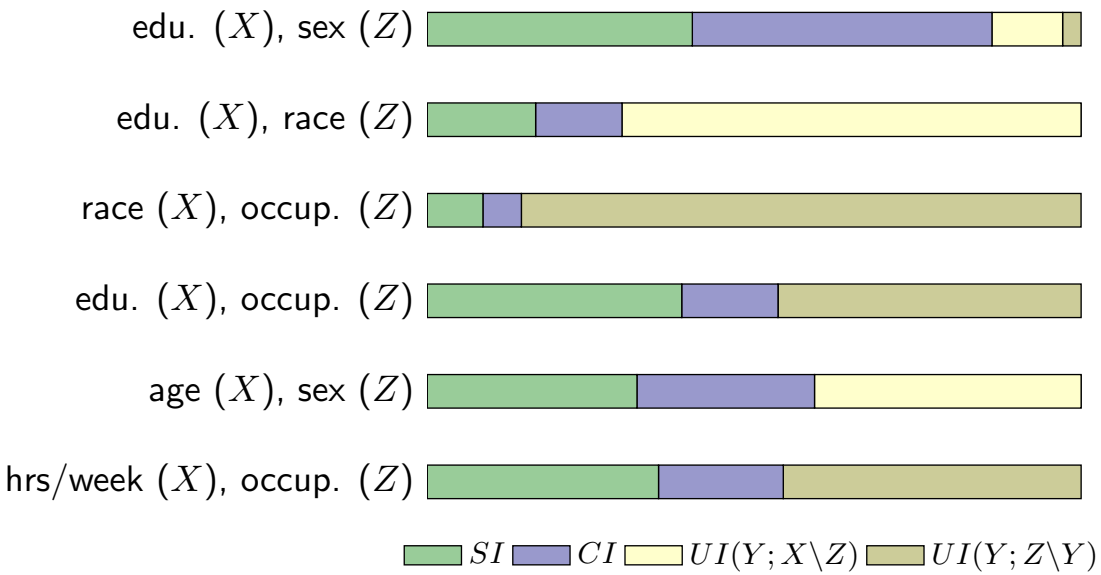
All these problems have the same solution.
The optimization problem is convex.

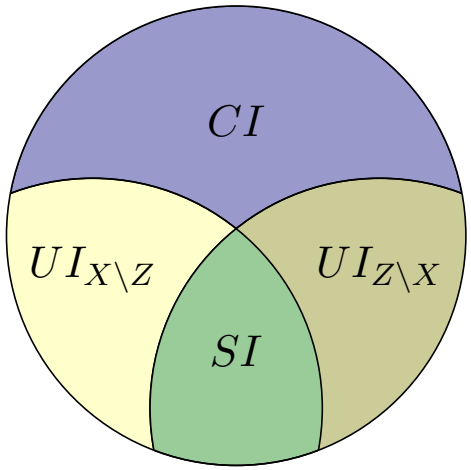
Example: The US 1994 census data set

[M. Lichman, UCI ML Repo, 2013, http://archive.ics.uci.edu/ml]

- Sample of the US population from 1994 with 48,842 entries
- Contains features such age, race, education, gender, income level, etc.
- Prediction task: Binary income level threshold USD 50,000.

Example: The US 1994 census data set





$I(Y;X,Z)$

- *Education* and *sex* convey about equally large *shared* and *complementary* information about income.
- Most of the information that *race* and *occupation* convey about income, is *uniquely* in the *occupation*.

  Classical approach: Test if $Y - Z - X$, i.e., $I(Y;X|Z) = 0$.

  New insight: Test if $I(Y;X|Z) = CI + UI_{X\backslash Z}$ is *purely synergistic*.

Absolute values 0.0853 0.0506 0.0988 0.1095 0.1159 0.1148

Slide taken from a recent talk

The BROJA[1] decomposition is:

- the theoretically best studied bivariate information decomposition to date;
- the only bivariate information decomposition with a complete axiomatic characterization;
- motivated by decision theory (Blackwell setting).

Small print:

- It is non-trivial to compute

[1] Bertschinger, Rauh, Olbrich, Jost, Ay 2014 (Entropy)

Computing the Unique Information

Computing the UI

- Challenge: Although convex, the problem can be very ill-conditioned
- Existing libraries implement Frank-Wolfe algo to compute the UI
  - Python package `dit` https://github.com/dit/dit
  - Custom implementation by Makkeh et al (2017) using the Python interior-point solver `CVXOPT` https://cvxopt.org/
- We develop an efficient double minimization algorithm, related to the classical Blahut-Arimoto algorithm for computing the channel capacity
  Library: https://github.com/infodeco/computeUI

Computing the UI

A convex program

$$UI(Y;X\backslash Z) := \min_{Q\in\Delta_P} I_Q(Y;X|Z)$$

Equivalent problem

$$\begin{aligned} I_\cup(Y;X,Z) :=& I(Y;Z) + UI(Y;X\backslash Z) \\ =& \min_{Q\in\Delta_P} I_Q(Y;X,Z) \end{aligned}$$

Double minimization formulation

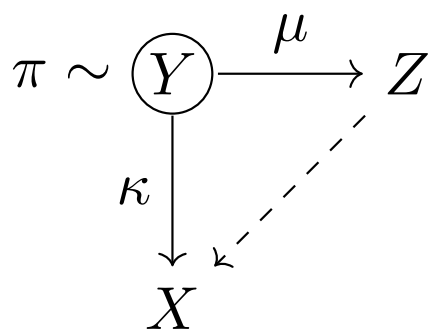
$$I_\cup(Y;X,Z) = \min_{Q\in\Delta_P} I_Q(Y;XZ) = \min_{Q\in\Delta_P} \min_{R_{XZ}\in\mathbb{P}_{\mathcal{X}\times\mathcal{Z}}} D(Q\|R_{XZ}Q_Y).$$

Conditional probability formulation

$$I_\cup(Y;X,Z) = \min_{R_{XZ}\in\mathbb{P}_{\mathcal{X}\times\mathcal{Z}}} \sum_y \pi(y) \min_{Q_{XZ|y}\in\Delta_{P,y}} D(Q_{XZ|y}\|R_{XZ}).$$

(for parallelization: we can compute each y separately)

Computing the UI



$$\Delta_P = \left\{Q \in \mathbb{P}_{\mathcal{Y}\times\mathcal{X}\times\mathcal{Z}} \colon Q_{YX}(y,x) = \pi(Y)\kappa_y(x),\ Q_{YZ}(y,z) = \pi(y)\mu_y(z)\right\}$$

$$\Delta_{P,Y} = \times_{y\in\mathcal{Y}}\Delta_{P,y}$$

$$\Delta_{P,y} = \left\{Q_{XZ} \in \mathbb{P}_{\mathcal{X}\times\mathcal{Z}} \colon Q_X(x) = \kappa_y(x) \text{ and } Q_Z(z) = \mu_y(z)\right\}, \mathsf{y} \in \mathcal{Y}$$

Alternating divergence minimization (admUI)

Algorithm: Initialize $R^{(0)}_{XZ}$ with full support. Recursively define

Step 1: $Q^{(i+1)}_{XZ|y} = \text{argmin}_{Q_{XZ|y} \in \Delta_{P,y}} D(Q_{XZ|y} \| R^{(i)}_{XZ})$ for each $y \in \mathcal{Y}$

Step 2: $R^{(i+1)}_{XZ} = \text{argmin}_{R_{XZ} \in \mathbb{P}_{\mathcal{X} \times \mathcal{Z}}} D(Q^{(i+1)}_{XZ|Y} \| R_{XZ} | \pi)$

Step 1 can be obtained, e.g., via iterative scaling, in parallel for all y:

Theorem 2 (I-projection)

The nonnegative functions b_n on $\mathcal{X} \times \mathcal{Z}$ defined recursively by

$$b_0(x,z) = R_{XZ}(x,z),$$

$$b_{n+1}(x,z) = b_n(x,z) \left[\frac{\kappa_y(x)}{\sum_z b_n(x,z)}\right]^{1/2} \left[\frac{\mu_y(z)}{\sum_y b_n(x,z)}\right]^{1/2},$$

converge to $\operatorname{argmin}_{Q_{XZ|y} \in \Delta_{P,y}} D(Q_{XZ|y} \| R_{XZ})$.

Step 2 admits a closed-form solution:

$$R_{XZ}^{(i+1)}(x,z) = \sum_{y \in \mathcal{Y}} \pi(y) Q_{XZ|Y}^{(i+1)}(x,z|y).$$

admUI algorithm: Convergence properties

Theorem 3 (admUI convergence)

Given π, κ, μ and an initial value $R^{(0)}_{XZ} \in \mathbb{P}_{\mathcal{X}\times\mathcal{Z}}$ of full support, the admUI iteration converges. The limit $\lim_{i\to\infty} \pi Q^{(i)}_{XZ|Y}$ is a global optimum.

admUI algorithm: Stopping criteria and Time complexity

Stopping criterion (main loop) The admUI iteration can be stopped when

$$\max_{x\in\mathcal{X},z\in\mathcal{Z}} \log \frac{Q^{(i+1)}_{XZ|Y}(x,z|y)}{Q^{(i)}_{XZ|Y}(x,z|y)} \leq \epsilon, \quad \text{for all } y \in \mathcal{Y},$$

for some prescribed accuracy $\epsilon > 0$.

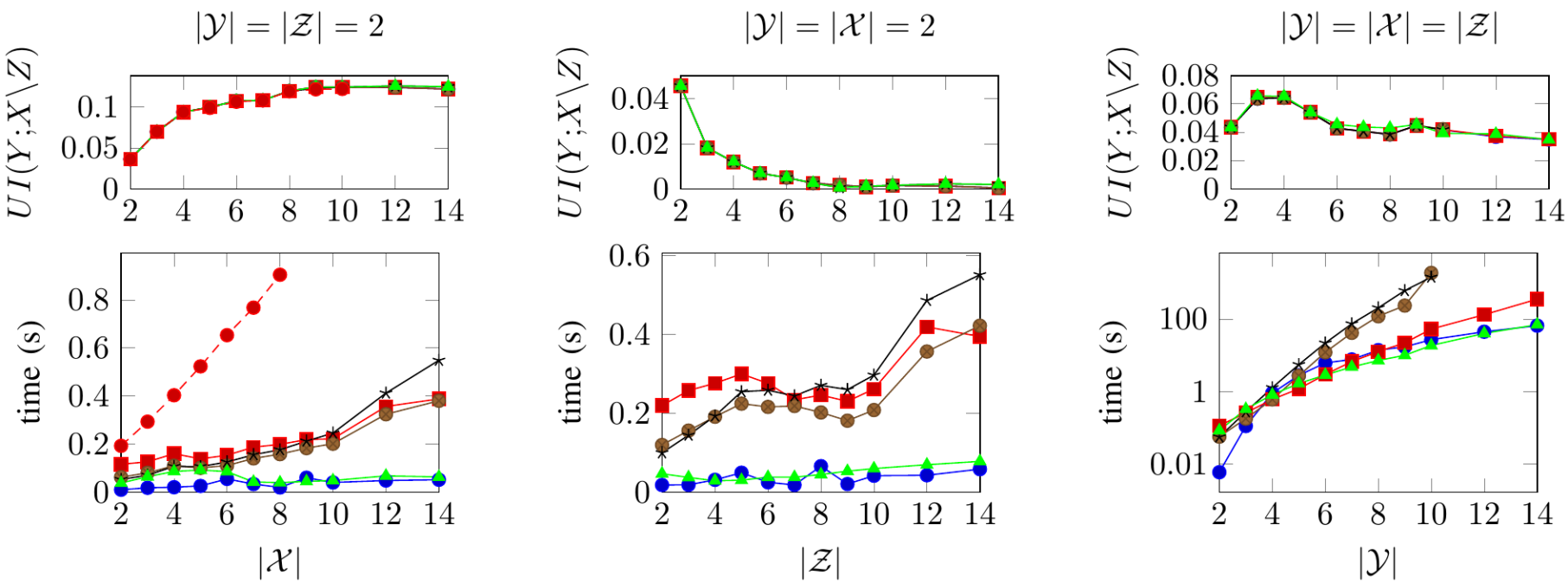
Stopping criterion (I-projection) The error in approximating the true solution scales inversely with the number of iterations. In practice, choosing an accuracy parameter $\epsilon' = 10^{-2}\epsilon$ yields good convergence.

Time complexity Complexity of one iteration of the admUI algorithm is dominated by Step 1 which costs about $\mathcal{O}(\frac{|\mathcal{Y}||\mathcal{X}||\mathcal{Z}|\log(|\mathcal{X}||\mathcal{Z}|)}{\epsilon'})$ to find a solution within ϵ' of the true solution.

Experiments

Code: `https://github.com/infodeco/computeUI`

Averages over 100 random joint distributions of computed UI value and computation time (wall-clock) on an Intel 2.60 GHz CPU.



Experiments

Code: `https://github.com/infodeco/computeUI`

Comparison of `admUI` and `fmincon` for the COPY example $Y = (X, Z)$

Size	ϵ	`admUI`		`fmincon`	
		Error	Time (ms)	Error	Time (ms)
2^4	10^{-8}	$9.16 \cdot 10^{-10}$	$9.03 \cdot 10^{1}$	$9.52 \cdot 10^{-5}$	$2.38 \cdot 10^{2}$
	10^{-5}	$6.67 \cdot 10^{-7}$	$6.45 \cdot 10^{1}$		
	10^{-3}	$5.01 \cdot 10^{-5}$	$1.03 \cdot 10^{1}$		
4^4	10^{-8}	$7.24 \cdot 10^{-10}$	$2.27 \cdot 10^{2}$	$1.50 \cdot 10^{-4}$	$4.17 \cdot 10^{2}$
	10^{-5}	$5.38 \cdot 10^{-7}$	$2.67 \cdot 10^{2}$		
	10^{-3}	$4.13 \cdot 10^{-5}$	$2.59 \cdot 10^{2}$		
7^4	10^{-8}	$4.93 \cdot 10^{-10}$	$2.42 \cdot 10^{3}$	$2.32 \cdot 10^{-4}$	$8.61 \cdot 10^{3}$
	10^{-5}	$3.71 \cdot 10^{-7}$	$2.41 \cdot 10^{3}$		
	10^{-3}	$2.89 \cdot 10^{-5}$	$1.97 \cdot 10^{3}$		
10^4	10^{-8}	$3.71 \cdot 10^{-10}$	$9.38 \cdot 10^{3}$	$3.51 \cdot 10^{-4}$	$4.86 \cdot 10^{5}$
	10^{-5}	$2.82 \cdot 10^{-7}$	$9.20 \cdot 10^{3}$		
	10^{-3}	$2.22 \cdot 10^{-5}$	$8.73 \cdot 10^{3}$		

`fmincon` (with gradient and Hessian included) settings: Algorithm = interior-point, MaxIterations = 10^4, MaxFunctionEvaluations = 10^5, OptimalityTolerance = 10^{-6}, ConstraintTolerance = 10^{-8}.

Reduce error by 5+ and time by 1+ orders of magnitude

Applications

- Refined statistical analysis (e.g. census data)
- Representation learning and Information Bottlenecks (e.g. IB is zero synergy bottleneck, deficiency bottleneck)
- Analysis of regularizers (e.g. dropout to avoid conspiracies)
- Feature selection (e.g. robustness and decision theoretic advantages)
- RL regularizers / morphological computation
- Secret key

Examples of synergy in the sciences

- Neurosciences
  - Synergy in neural code (coInformation) – a pair of spikes closely spaced in time can jointly convey more than twice the information carried by a single spike [BSK$^+$00]. can be negative
  - Synergy, redundancy, and independence (correlational importance) – nonnegative measure in neural coding [LN05]. can exceed mutual information
  - Hierarchical decomposition (multi-information) – difference of entropies of information projections [SBB03]. no data processing inequality / operational interpretation
- Cryptography - secret sharing
  - Integer addition modulo k is an important building block in many secret sharing schemes
  - Checksum of several digits can only be computed when all digits are known

IB as zero-synergy bottleneck

- IB maximize $I(Y;Z) - \beta I(X;Z)$ subject to $Y - X - Z$.

Lemma 4

If $Y - X - Z$ is a Markov chain, then

$$UI(Y;X\backslash Z) = I(Y;X|Z) = I(Y;X) - I(Y;Z).$$

This implies that the synergy vanishes, $CI(Y;X,Z) = 0$.

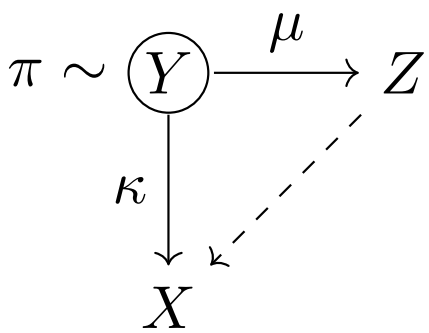
- IB can be interpreted as a "zero-synergy bottleneck". Equivalent formulation of IB: Minimize over $e = P(Z|X)$

$$I(Y;X|Z) + \beta I(X;Z) = UI(Y;X\backslash Z) + \beta I(X;Z)$$

The *sufficiency* term depends on the (Y,X) and (Y,Z)-marginals.

The *minimality* term depends on the (X,Z)-marginal.

UI and deficiencies



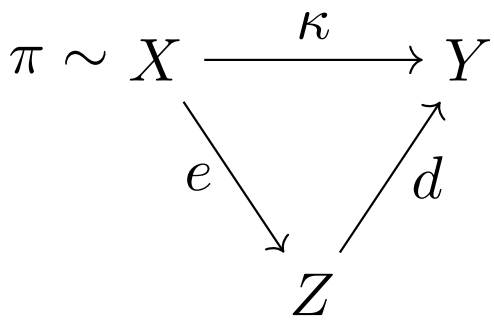
- The UI is similar in spirit to a generalized version of the Le Cam deficiency [LC64, Rag11]
- Deficiencies measure the cost of approximating one channel from another via Markov kernels.
- The input *deficiency of* μ *w.r.t.* κ is

$$\delta^\pi(\mu,\kappa) := \min_{\lambda \in \mathsf{M}(\mathcal{Z};\mathcal{X})} D(\kappa \| \lambda \circ \mu | \pi)$$

- The deficiency is upper bounded by the UI: $\delta^\pi(\mu,\kappa) \leq UI(Y;X\backslash Z)$
- $\delta^\pi(\mu,\kappa)$ satisfies the following property:

$$\delta^\pi(\mu,\kappa) = 0 \text{ if and only if } UI(Y;X\backslash Z) = 0$$

UI the deficiency bottleneck



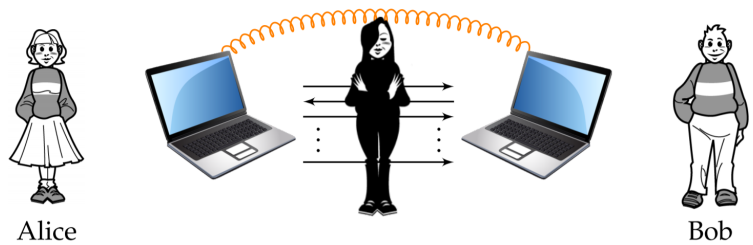
- The output *deficiency of* d *w.r.t.* κ is

$$\delta_o^\pi(d,\kappa) := \min_{e \in \mathsf{M}(\mathcal{X};\mathcal{Z})} D(\kappa \| d \circ e | \pi)$$

- By convexity of the KL divergence, $\delta_o^\pi(d,\kappa) \leq I(Y;X|Z)$
- The *deficiency bottleneck* [BM18] minimizes

$$\delta_o^\pi(d,\kappa) + \beta I(Z;X)$$

Operational interpretation



Secret key agreement task

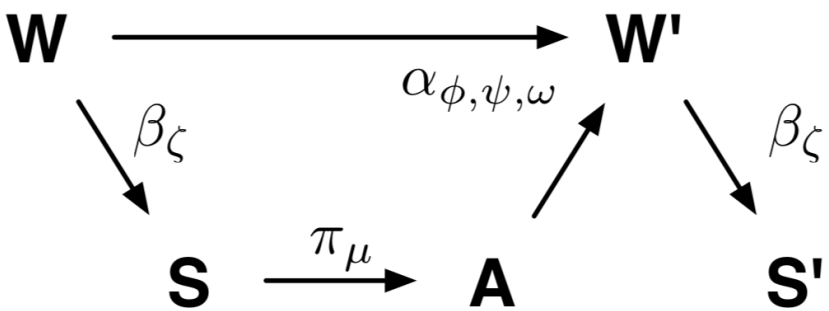
- Alice, Bob and an adversary Eve observe resp. i.i.d. copies of Y, X, Z, where $(Y, X, Z) \sim P$
- Alice and Bob can perform local operations on their subsystems. In addition, Alice can send a message to Bob over a public channel transparent to Eve
- Alice and Bob wish to convert P into a secret key that is uncorrelated with Eve. The maximum (asymptotic) rate at which Alice and Bob can compute a key (length of the key) is called the one-way secret key rate [AC93]

Theorem 5 ([RBOJ19])

$UI(Y; X \backslash Z)$ is an upper bound on the one-way secret key rate.

Quantifying morphological computation

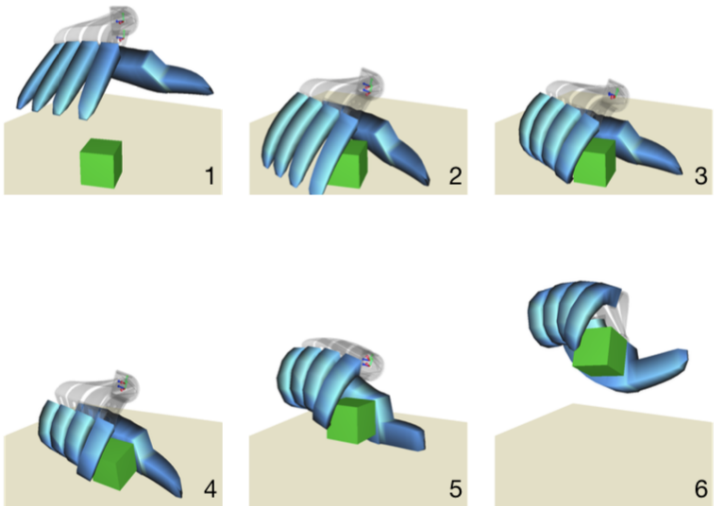
MC: contribution of the embodiment to a behavior

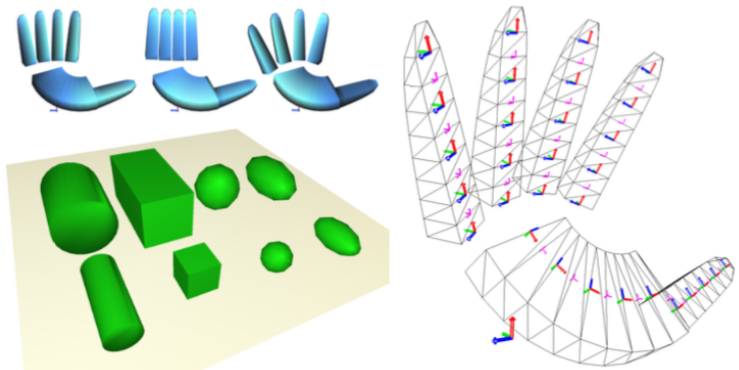


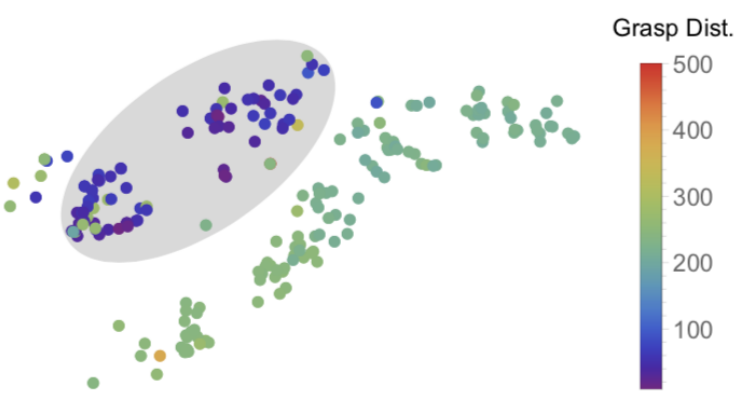
UI has been proposed as a measure of MC [GZR15]

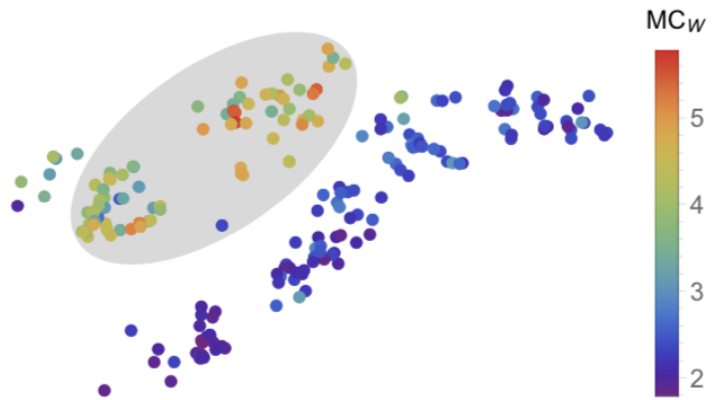
$$MC = UI(W'; W \backslash A)$$

Quantifying morphological computation









MC: the good the bad and the ugly [ZDM$^+$17]

Discussion

- Positive information decompositions allow us to disentangle synergy and redundancy, which is important in multiple applications
- Defining such decompositions has been a long quest
- The BROJA is an option with axiomatic and operational appeal, but involves a non trivial optimization problem
- We formulated a scalable algorithm for this

References I

Rudolf Ahlswede and Imre Csiszár.

Common randomness in information theory and cryptography. I. Secret sharing.

IEEE Transactions on Information Theory, 39(4):1121–1132, 1993.

David Blackwell.

Equivalent comparisons of experiments.

The Annals of Mathematical Statistics, 24(2):265–272, 1953.

Pradeep Kr Banerjee and Guido Montúfar.

The variational deficiency bottleneck.

arXiv preprint arXiv:1810.11677, 2018.

Nils Bertschinger and Johannes Rauh.

The Blackwell relation defines no lattice.

In *Proc. IEEE ISIT*, pages 2479–2483. IEEE, 2014.

References II

Pradeep Kr. Banerjee, Johannes Rauh, and Guido Montúfar.

Computing the unique information.

In *Proc. IEEE ISIT*, pages 141–145. IEEE, 2018.

Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay.

Quantifying unique information.

Entropy, 16(4):2161–2183, 2014.

Naama Brenner, Steven P Strong, Roland Koberle, William Bialek, and Rob R de Ruyter van Steveninck.

Synergy in a neural code.

Neural computation, 12(7):1531–1552, 2000.

References III

Keyan Ghazi-Zahedi and Johannes Rauh.

Quantifying morphological computation based on an information decomposition of the sensorimotor loop.

The 2018 Conference on Artificial Life: A Hybrid of the European Conference on Artificial Life (ECAL) and the International Conference on the Synthesis and Simulation of Living Systems (ALIFE), (27):70–77, 2015.

Lucien Le Cam.

Sufficiency and approximate sufficiency.

The Annals of Mathematical Statistics, pages 1419–1455, 1964.

Peter E. Latham and Sheila Nirenberg.

Synergy, redundancy, and independence in population codes, revisited.

Journal of Neuroscience, 25(21):5195–5206, 2005.

References IV

Ueli M Maurer.

Secret key agreement by public discussion from common information.

IEEE Transactions on Information Theory, 39(3):733–742, 1993.

W. McGill.

Multivariate information transmission.

IRE Transactions on Information Theory, 4(4):93–111, 1954.

Ueli M. Maurer and Stefan Wolf.

Unconditionally secure key agreement and the intrinsic conditional information.

IEEE Transactions on Information Theory, 45(2):499–514, 1999.

Maxim Raginsky.

Shannon meets Blackwell and Le Cam: Channels, codes, and statistical experiments.

In *Proc. IEEE ISIT*, pages 1220–1224. IEEE, 2011.

References V

Johannes Rauh, Pradeep Kr. Banerjee, Eckehard Olbrich, and Jürgen Jost.

Unique information and secret key decompositions.

In *Proc. IEEE ISIT (to appear)*. IEEE, 2019.

Elad Schneidman, William Bialek, and Michael J Berry.

Synergy, redundancy, and independence in population codes.

Journal of Neuroscience, 23(37):11539–11553, 2003.

Paul Williams and Randall Beer.

Nonnegative decomposition of multivariate information.

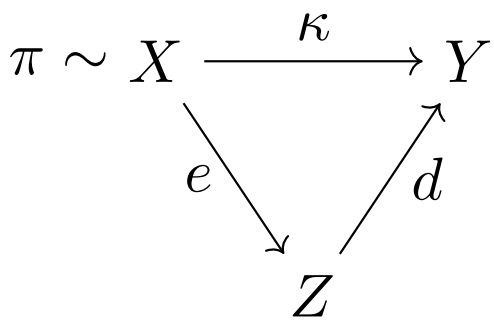
arXiv:1004.2515v1, 2010.

K. Zahedi, R. Deimel, G. Montufar, V. Wall, and O. Brock.

Morphological computation: The good, the bad, and the ugly.

In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 464–469, Sep. 2017.

The Information Bottleneck (IB) problem



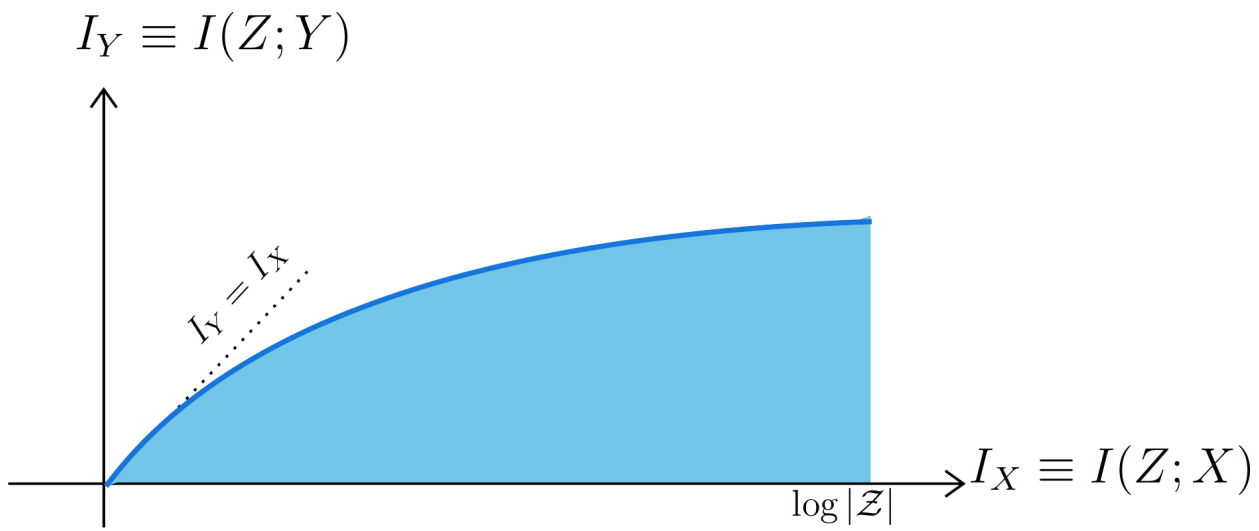
- X is an observation or *input* variable and Y a correlated *output* variable of interest.
- The channel κ gives the true relation between the input and output. In general, it is unknown and only accessible through examples
- Want to learn a more "structured" version of κ.
- Find a pair of maps (e, d) so that Z preserves as much *relevant information* as possible about the output (sufficiency) while maximally compressing the input (minimality)

The IB curve

- IB maximizes

$$I(Y;Z) - \beta I(X;Z),$$

  over all $e \in \mathsf{M}(\mathcal{X};\mathcal{Z})$, where $\beta \in [0,1]$ is a regularization parameter
- The IB curve traces $I(Z;Y)$ (sufficiency) vs. $I(Z;X)$ (minimality) for different values of β
- The IB curve is a concave and monotonically non-decreasing function



The Variational Deficiency Bottleneck (VDB)

- The *deficiency bottleneck* minimizes

$$\delta^{\pi}(d, \kappa) + \beta I(Z; X)$$

  over all $e \in \mathsf{M}(\mathcal{X}; \mathcal{Z})$, $d \in \mathsf{M}(\mathcal{Z}; \mathcal{Y})$.

- The *rate term* admits a simple *variational upper bound*:

$$I(Z; X) \leq \int p(x, z) \log \frac{e(z|x)}{r(z)} \, dx \, dz, \quad \text{for any } r.$$

- The Variational Deficiency Bottleneck (VDB) objective:

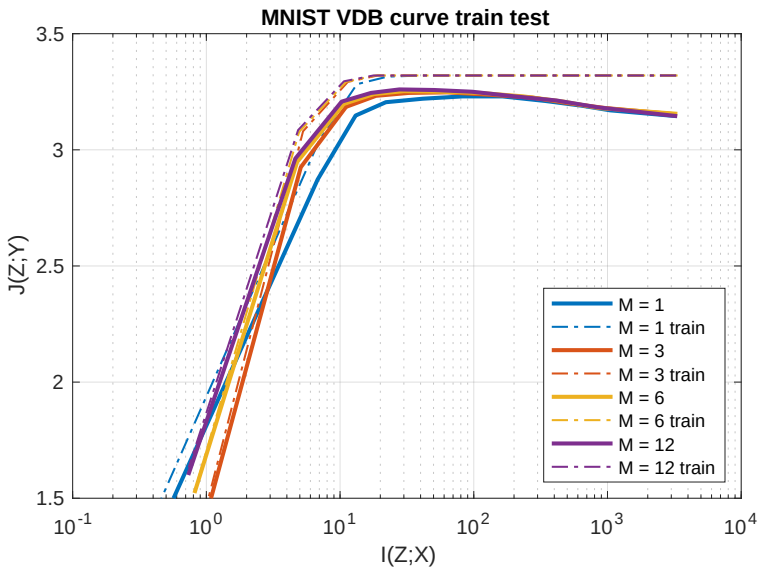
$$\mathcal{L}_{VDB}(e, d) := \mathbb{E}_{(x,y)\sim p_{\mathcal{D}}} \left[-\log(\int d(y|z) e(z|x) \, dz) + \beta D(e(Z|x) \| r(Z)) \right]$$

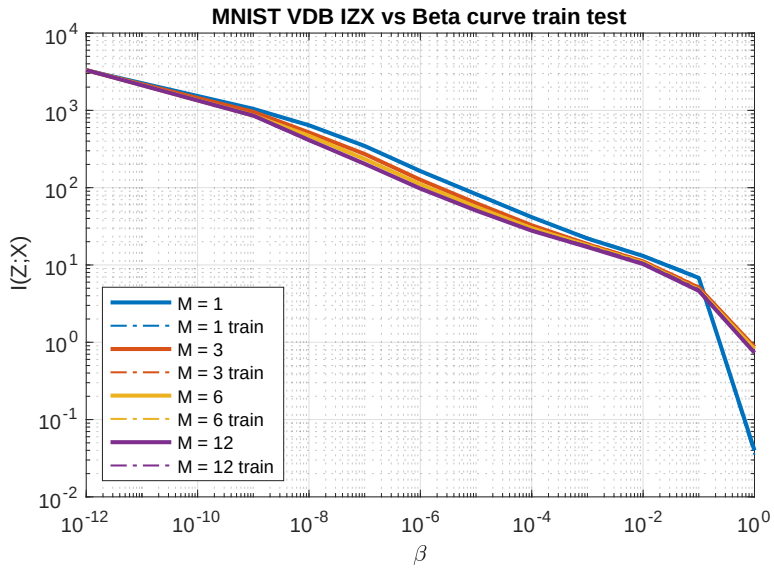
- Relation of the deficiency to the cross-entropy loss:

$$\mathbb{E}_{(x,y)\sim p_{\mathcal{D}}} \left[-\log(\int d(y|z) e(z|x) \, dz) \right] \leq \mathbb{E}_{(x,y)\sim p_{\mathcal{D}}} \left[\int - e(z|x) \log d(y|z) dz \right]$$

The VDB curve for the MNIST dataset

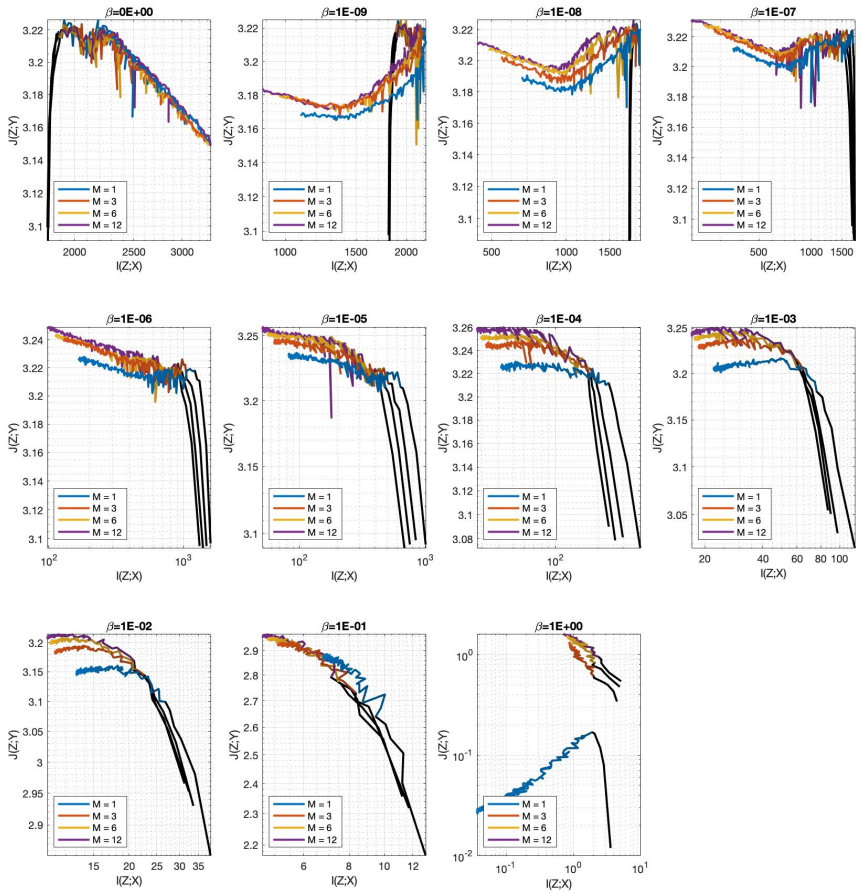
- In the VDB, "more sufficient" means "less deficient"
- The term corresponding to sufficiency is
  $J(Z;Y) := H(Y) - \mathbb{E}_{(x,y)\sim p_{\mathcal{D}}(x)} [-\log(\int d(y|z)e(z|x)\,dz)]$





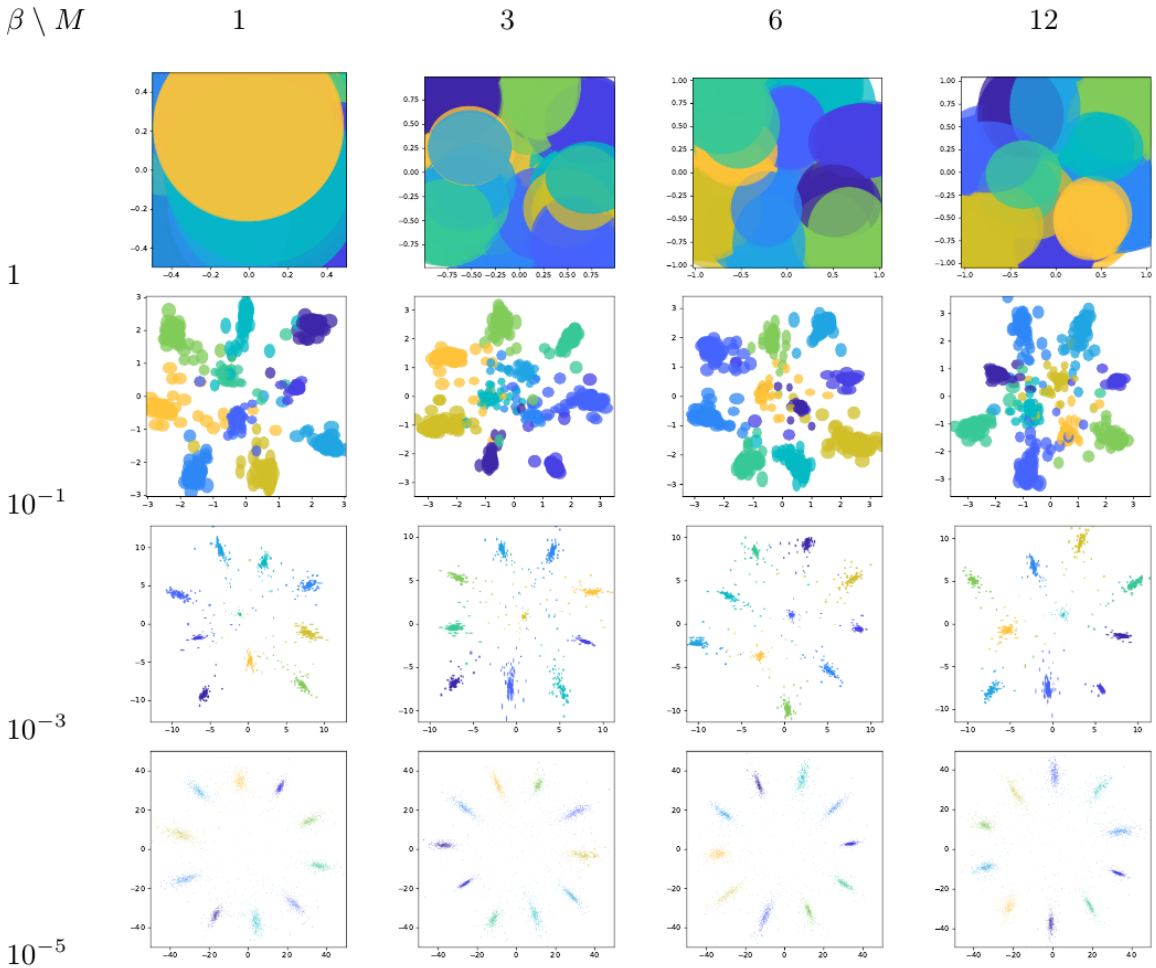
- M is the number of encoder output samples.
- $M = 1$ corresponds to the Variational Information Bottleneck (VIB).
- Encoder: 784 inputs–1024 ReLU–1024 ReLU–512 linear output units.
- Decoder: A softmax layer.

IB-plane learning curves on MNIST



- Evolution of $J(Z;Y)$ vs. $I(Z;X)$ over 200 training epochs (dark to light color) with 256D representation.
- For good β values, early epochs are mainly fitting ($J(Z;Y)\ \nearrow$), while later epochs are mainly discarding information about the input ($I(Z;X)\ \searrow$).
- At higher values of M (our method), the representation captures more information about the output while discarding more information about the input.

2D representations of MNIST



- Posterior distributions of 1000 input images after training.
- Color corresponds to the class label.
- For $\beta = 10^{-5}$, the representations of different classes are well separated.

Blackwell property of the UI

UI satisfies the following key property which we call the *Blackwell property* of the UI:

Lemma 6 (Vanishing UI [BRO+14, Lemma 6])

For a given joint distribution P_{YXZ}, $UI(Y;X\backslash Z)$ vanishes, i.e., $\min_{Q\in\Delta_P} I_Q(Y;X|Z)=0$, if and only if there exists a random variable X' such that $Y-Z-X'$ is a Markov chain and $P_{YX'}=P_{YX}$.

Blackwell's theorem [Bla53, BR14] implies:

Corollary 7

A vanishing $UI(Y;X\backslash Z)$ is equivalent to the fact that any decision problem in which the task is to predict Y can be solved just as well with the knowledge of Z as with the knowledge of X.

Secret key rate definition

Definition 8 ([Mau93])

The *two-way secret key rate* denoted $S_{\leftrightarrow}(Y;X|Z)$, is the maximum rate R such that for every $\epsilon > 0$ and sufficiently large n, there exists a two-way public communication protocol that outputs keys K and K' (ranging over some common set $\mathcal{K}$) satisfying

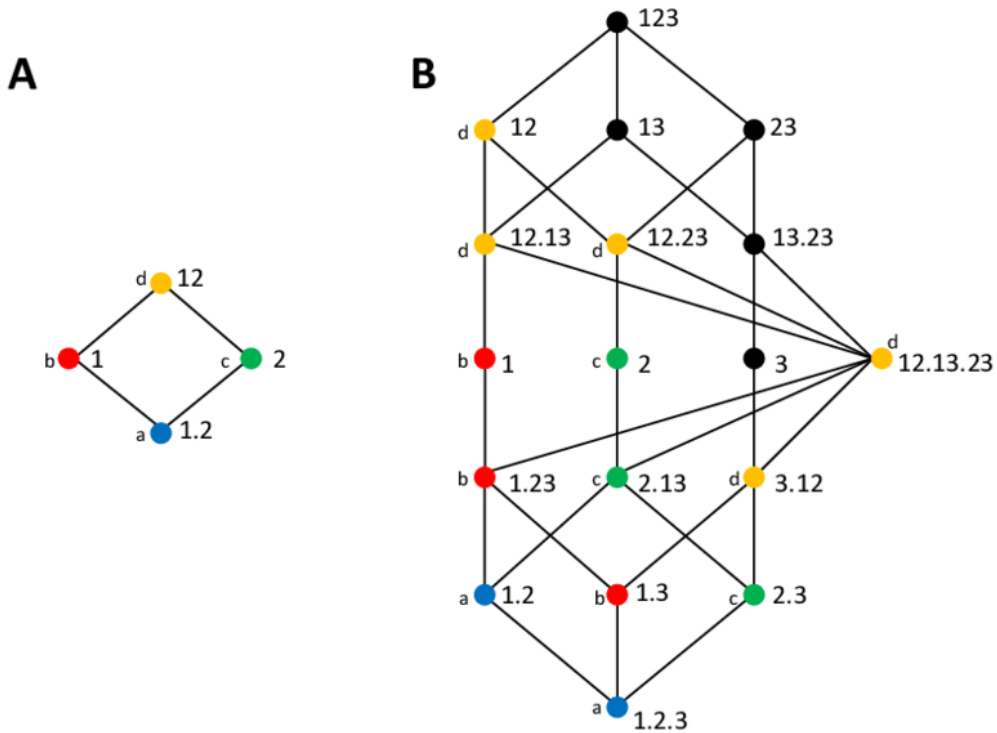
$$\begin{aligned} \Pr[K = K'] &\geq 1 - \epsilon && \textit{(reliability)}, \\ \tfrac{1}{n} H(K) &> \tfrac{1}{n} \log |\mathcal{K}| - \epsilon && \textit{(uniformity)}, \\ \tfrac{1}{n} I(K; C, Z^n) &\leq \epsilon && \textit{(weak secrecy)}, \end{aligned}$$

and achieving $\frac{1}{n} H(K) \geq R - \epsilon$, where C is the amount of public communication consumed in the protocol. We say that the protocol is *one-way* if Alice is allowed to send only one message and Bob none. The corresponding key rate is called the *one-way secret key rate* $S_{\rightarrow}(Y;X|Z)$.

The first and second condition ensure, resp., that the keys are equal to each other with high probability and that they are almost uniformly distributed. The third condition ensures that the *rate* at which Eve learns information about the keys is negligibly small. MODEL: Maurer proposed the following interactive model called the *source model* for secret key agreement [Mau93, MW99]. Alice, Bob and Eve observe n i.i.d. copies of random variables Y, X and Z resp., where (Y, X, Z) is distributed according to some joint distribution known to all parties, called the *source*. Alice and Bob wish to agree on a common secret key by communicating interactively over a public channel transparent to Eve. A *two-way* public communication protocol proceeds in rounds, where Alice and Bob exchange messages in alternating order, with Alice sending messages in the odd rounds and Bob in the even rounds. Each message is a function of the sender's observation and all the messages exchanged so far. At the end of the protocol, Alice (resp., Bob) computes a key K (resp., K') as a function of Y^n (resp., X^n) and C, the set of all exchanged messages.

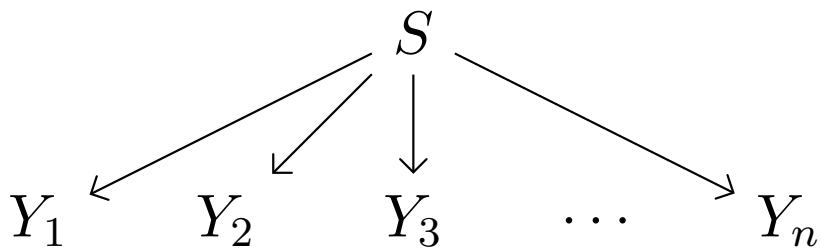
More than three variables

There is a decomposition scheme proposed by Williams and Beer [WB10].



The idea of Williams and Beer

What happens for more random variables?



Partial information lattice (Williams, Beer 2010):
Decomposition framework based on a measure of redundancy

$$I_\cap(S : A_1; \ldots; A_k), \qquad A_1, \ldots, A_k \subseteq \{Y_1, \ldots, Y_n\}.$$

Basic idea: Classify information according to "who knows what".

Due to synergistic effects, "who" not only refers to random variables, but to subsets of random variables.

Axioms for Shared Information

Williams and Beer propose that $I_\cap$ should satisfy:

- $I_\cap(S : A_1; \ldots; A_k)$ is symmetric in $A_1, \ldots, A_k$.
- $I_\cap(S : A_1) = MI(S : A_1)$.
- $I_\cap(S : A_1; \ldots; A_k; A_{k+1}) \leq I_\cap(S : A_1; \ldots; A_k)$, with equality if $A_i \subseteq A_{k+1}$ for some $i \leq k$.

Notes:

1. Intuition of the last axiom: Information on the left is a subset of information on the right.
2. The axioms imply that it is enough to know the values $I_\cap(S : A_1; \ldots; A_k)$, where $(A_1; \ldots; A_k)$ is an antichain; i.e. $A_i \not\subseteq A_j$ for all $i \neq j$.

Local positivity

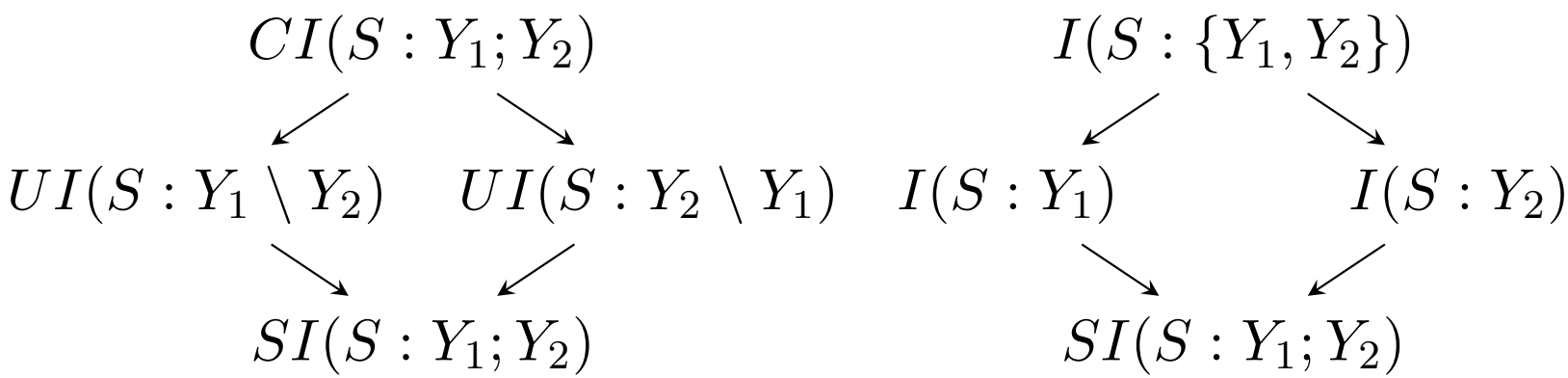
To decompose the information, we need a function I_∂ with

$$I_\cap(S : A_1; \ldots; A_k) = \sum_{(B_1,\ldots,B_l)\preceq(A_1,\ldots,A_k)} I_\partial(S : B_1; \ldots; B_l).$$

I_∂ can be computed by Moebius inversion.
If $I_\partial \geq 0$, then we call $I_\cap$ locally positive.

Example

For $n = 2$ we obtain:



Status of the Williams-Beer program

Open problem:

How to define the functions $I_\cap$ and I_∂.

Current status:

1. There is no convincing proposal for a general information decomposition along this framework.
2. To the contrary, there are some impossibility results.
3. There is a number of bivariate $(n = 2)$ information decompositions.

Questions:

1. Is the PI lattice correct? Is the WB program doable?
2. How to distinguish different (bivariate) information decompositions?