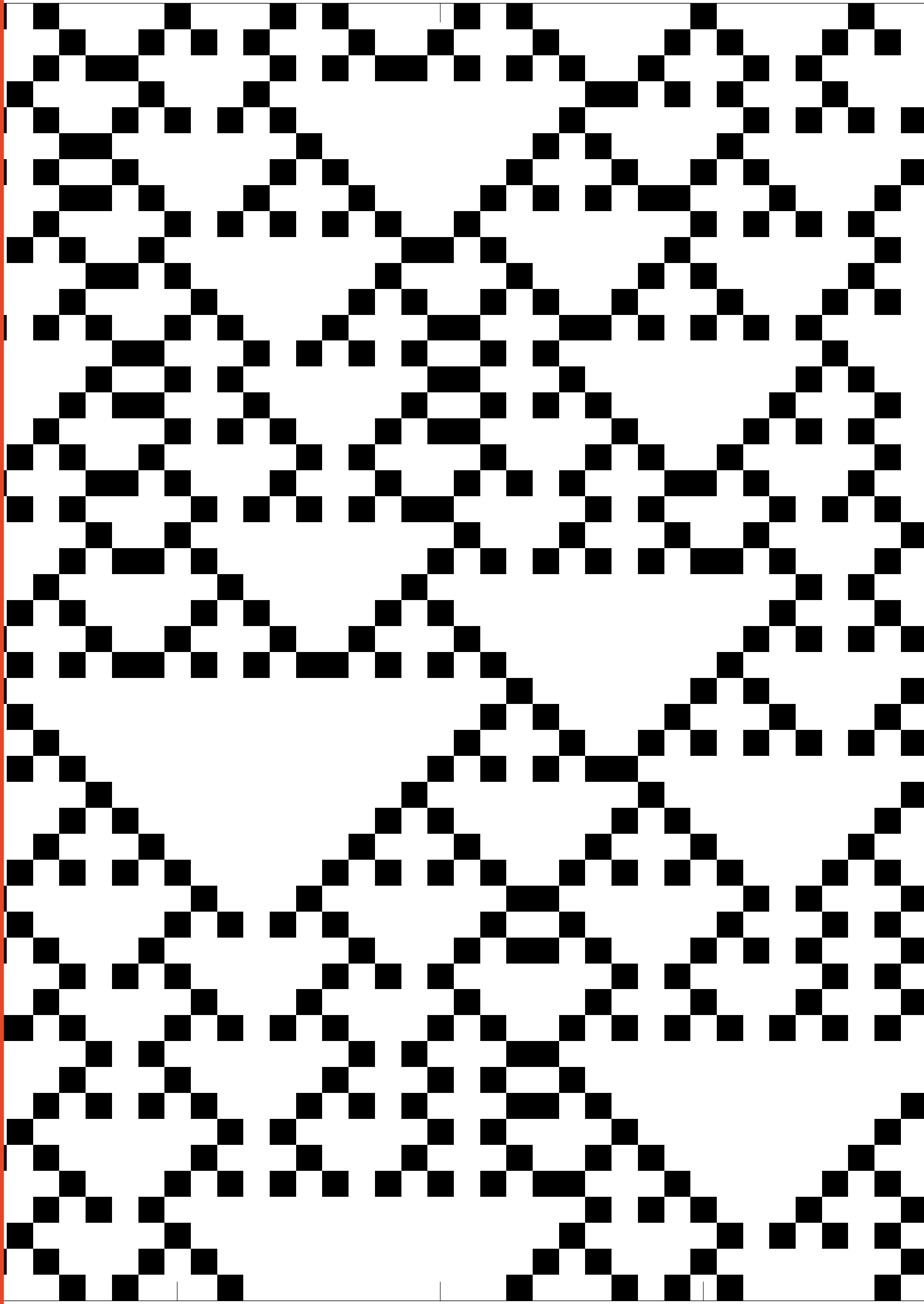
What is information?



Dr. Joseph Lizier

THE UNIVERSITY OF SYDNEY

What is Information: session outcomes

- Ability to express ideas about information and relate that to uncertainty.
- Understand further fundamental measures of information theory including: mutual information, conditional mutual information.
- Ability to partially construct Matlab code to compute such measures, and apply that code to examples.

- Primary references:
  - Cover and Thomas, "Elements of Information Theory", Hoboken, New Jersey: John Wiley and Sons, Inc., 2006 (2nd ed.); section 2.2-2.5, 2.6, 2.8
  - Mackay, "Information Theory, Inference, and Learning Algorithms", Cambridge: Cambridge University Press, 2003; sections 2.6, chapter 8.
  - Bossomaier, Barnett, Harré, Lizier, "An Introduction to Transfer Entropy: Information Flow in Complex Systems", Springer, Cham, 2016; section 3.2.1-3.2.4.
  - Lizier, "JIDT: An information-theoretic toolkit for studying the dynamics of complex systems", Frontiers in Robotics and AI, 1:11, 2014; Appendix A.1 and A.3

Cross entropy and Kullback-Leibler divergence

- Cross-entropy:

$$\mathrm{G}(p||q) = \sum_{x \in A_x} p(x) \log_2 \frac{1}{q(x)}$$

  - Average code length if using the PDF *q(x)* to optimally encode x, which has actual PDF *p(x)*

- Kullback-Leibler (KL) divergence:

$$\mathrm{D}(p||q) = \sum_{x \in A_x} p(x) \log_2 \frac{p(x)}{q(x)}$$

  - Average coding penalty from using the PDF *q(x)* to optimally encode x, which has actual PDF *p(x)*
  - $\mathrm{D}(p||q) \geq 0$ with equality iff *q*=*p*
    - You always incur a cost for using incorrect PDF!

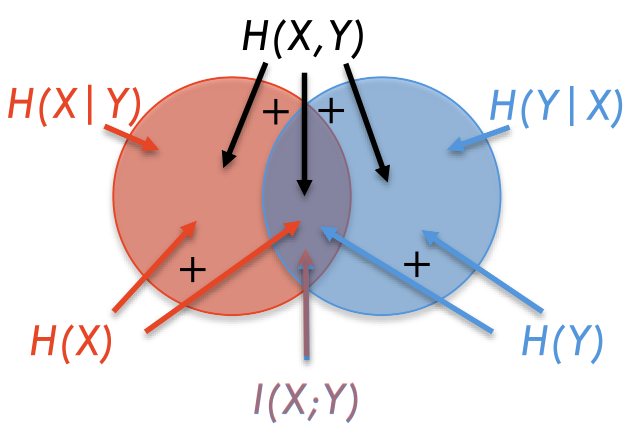
T. M. Cover and J. A. Thomas. Elements of Information Theory. Wiley-Interscience, New York, 1991. Section 2.3

Bossomaier, Barnett, Harré, Lizier, "An Introduction to Transfer Entropy: Information Flow in Complex Systems", Springer, Cham, 2016; section 3.2.4

Mutual information (MI)

– Mutual information *I(X;Y)* is the reduction in uncertainty or surprise about one variable *X* that we obtain from another variable *Y*

– Interpretation 1: from Venn diagrams –



$$\mathrm{I}(X;Y) = \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y)$$
$$\mathrm{I}(X;Y) = \mathrm{H}(X) - \mathrm{H}(X|Y)$$
$$\mathrm{I}(X;Y) = \mathrm{H}(Y) - \mathrm{H}(Y|X)$$
$$\mathrm{I}(X;Y) = \mathrm{I}(Y;X)$$

Properties:

– $0 \leq \mathrm{I}(X;Y) \leq \min(\mathrm{H}(X), \mathrm{H}(Y))$

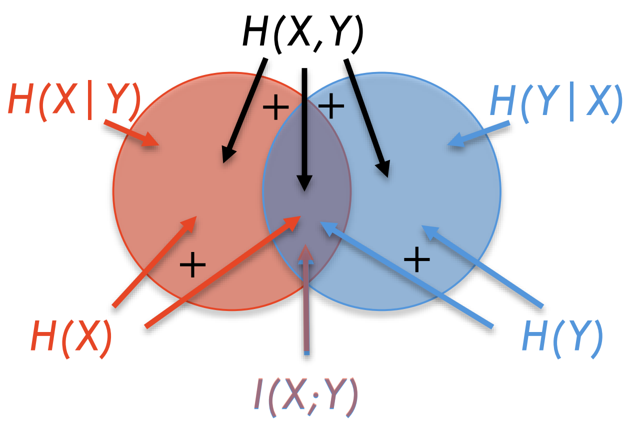
– Is symmetric in *X* and *Y*

– $\mathrm{I}(X;Y) = \mathrm{H}(X) \rightarrow \mathrm{H}(X|Y) = 0$

T. M. Cover and J. A. Thomas. Elements of Information Theory. Wiley-Interscience, New York, 1991. Section 2.4

Mutual information (MI)

- Mutual information *I(X;Y)* is the reduction in uncertainty or surprise about one variable *X* that we obtain from another variable *Y*
- Interpretation 2: KL divergence / Bayesian view –



$$\mathrm{I}(X;Y) = \sum_{x\in A_x, y\in A_y} p(x,y)\log_2 \frac{p(x,y)}{p(x)p(y)}$$

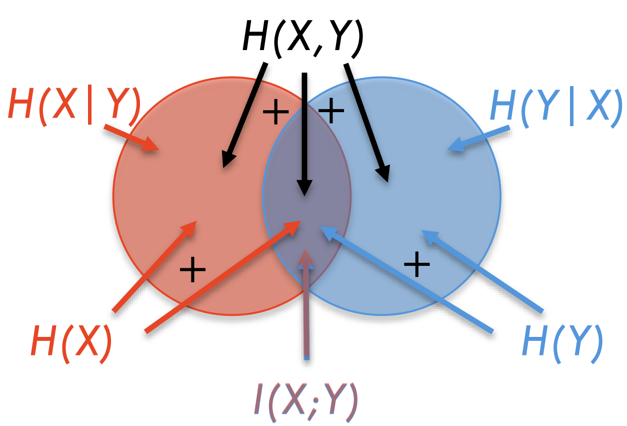
$$\mathrm{I}(X;Y) = \sum_{x\in A_x, y\in A_y} p(x,y)\log_2 \frac{p(x|y)}{p(x)}$$

Properties:

- $\mathrm{I}(X;Y) = \mathrm{D}(p(x,y)||p(x)p(y))$
- MI is code length penalty for coding {*x*,*y*} assuming *x* and *y* are independent, or for coding *x* without using knowledge of *y*.

Mutual information (MI)

- Mutual information *I(X;Y)* is the reduction in uncertainty or surprise about one variable *X* that we obtain from another variable *Y*
- Interpretation 3: statistical view –



$$\mathrm{I}(X;Y) = \sum_{x \in A_x, y \in A_y} p(x,y) \log_2 \frac{p(x|y)}{p(x)}$$

Properties:

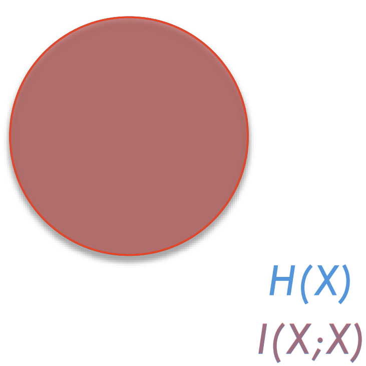
- $\mathrm{I}(X;Y) = 0 \leftrightarrow$ *X* is independent of *Y*
- MI is a non-linear form of correlation

T. M. Cover and J. A. Thomas. Elements of Information Theory. Wiley-Interscience, New York, 1991. Section 2.3
Bossomaier, Barnett, Harré, Lizier, "An Introduction to Transfer Entropy: Information Flow in Complex Systems", Springer, Cham, 2016; section 3.2.2

Mutual information (MI)

- Mutual information *I(X;Y)* is the reduction in uncertainty or surprise about one variable *X* that we obtain from another variable *Y*
- Interpretation 4: self-information and uncertainty –

$$\mathrm{I}(X;X) = \mathrm{H}(X) + \mathrm{H}(X) - \mathrm{H}(X,X)$$
$$\mathrm{I}(X;X) = \mathrm{H}(X)$$



Properties:

- Entropy *H(X)* (uncertainty) is equivalent to the self-information *I(X;X)* (uncertainty reduction) obtained from that variable about itself.
- Entropy and information are complementary quantities!

T. M. Cover and J. A. Thomas. Elements of Information Theory. Wiley-Interscience, New York, 1991. Section 2.4

Mutual information (MI)

- Mutual information *I(X;Y)* is the reduction in uncertainty or surprise about one variable *X* that we obtain from another variable *Y*
- Interpretation 5: Kelly Gambling –

$$\mathrm{I}(X;Y) = \sum_{x \in A_x, y \in A_y} p(x,y) \log_2 \frac{p(x|y)}{p(x)}$$

  - If we gamble on *x*, with fair odds (payout is *1/p(x)* for each winner *x*), and invest all our capital on each race, over repeated races:
    - Best strategy is to spread investments as per *p(x)*
    - If we have side information *y*, best to invest as per *p(x|y)*
    - *I(X;Y)* as the average growth rate of investment return when investing as per *p(x|y)*, with respect to gambling as per *p(x)*.

T. M. Cover and J. A. Thomas. Elements of Information Theory. Wiley-Interscience, New York, 1991. Chapter 6
C. Finn and J.T. Lizier, "Pointwise Partial Information Decomposition Using the Specificity and Ambiguity Lattices", Entropy, 20, 297, 2018.

Pointwise or local Mutual information

- Mutual information *i(x;y)* is the reduction in uncertainty or surprise about one sample *x* of variable *X* that we obtain from one sample *y* of another variable *Y*

$$
\begin{aligned}
\mathrm{i}(x;y) &= \mathrm{h}(x) + \mathrm{h}(y) - \mathrm{h}(x,y) \\
\mathrm{i}(x;y) &= \mathrm{h}(x) - \mathrm{h}(x|y) \\
\mathrm{i}(x;y) &= \mathrm{h}(y) - \mathrm{h}(y|x) \\
\mathrm{i}(x;y) &= \log_2 \frac{p(x|y)}{p(x)} \\
\mathrm{I}(X;Y) &= \langle \mathrm{i}(x;y) \rangle
\end{aligned}
$$

- *i(x;y)* > 0 means *p(x|y)* > *p(x)*, so *y* increased our expectation that *x* would occur, positively informing us.
- *i(x;y)* < 0 means *p(x|y)* < *p(x)*, so *y* reduced our expectation that *x* would occur, misinforming us.
  - e.g. when the weather report says 'sunshine' but it actually rains, we may have *p*(`rain|sunny_forecast`) = 0.05 whilst *p*(`rain`)=0.2.
  - But: *on average* over all samples *Y* provides *I(X;Y)* ≥ *0.*

Bossomaier, Barnett, Harré, Lizier, "An Introduction to Transfer Entropy: Information Flow in Complex Systems", Springer, Cham, 2016; section 3.2.2/3.2.2.1

Pointwise or local Mutual information

- Mutual information *i(x;y)* is the reduction in uncertainty or surprise about one sample *x* of variable *X* that we obtain from one sample *y* of another variable *Y*
- Interpretation 6: information comes from effect of *exclusions* –

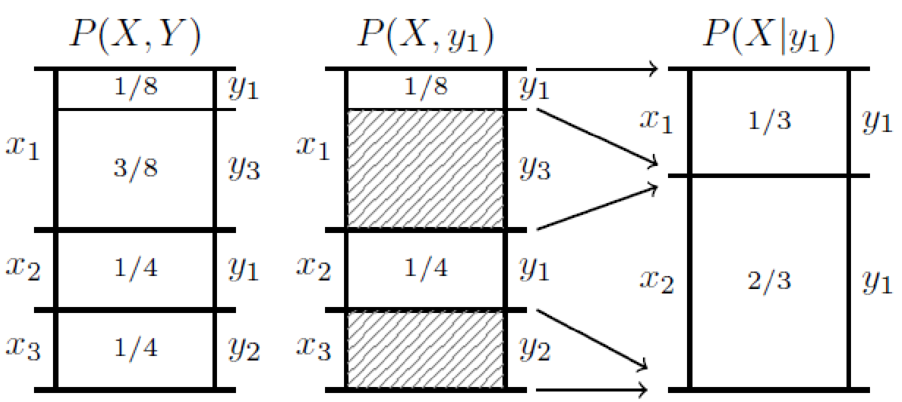


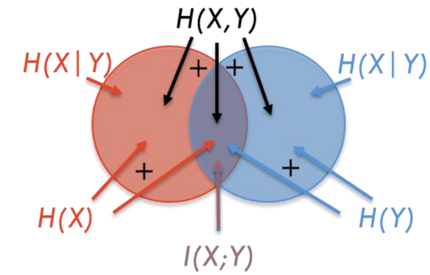
Fig. 1. Probability mass diagrams, which use length to represent the probability mass of each joint event (x, y) (summing to 1 over all (x, y) of course). These illustrate: a. (left) starting from the joint distribution $P(X, Y)$; b. (middle) the occurrence of the event $Y = y_1$ leads to exclusions of $\bar{y}_1 = \{\mathcal{Y} \setminus y_1\} = \{y_2, y_3\}$ to leave $P(X, y_1)$; c. (right) and the remaining space is then normalised into $P(X|y_1)$.

1. Learning the value $Y=y_1$ leads to exclusions in the joint space $P(X,y_1)$. The potential "value" of the exclusions is $h(y_1)$.
2. Renormalise the probability space to get $P(X|y_1)$.
3. Compare $P(x_1)$ and $P(x_1|y_1)$ for the event x_1 which occurred.

- Consider how exclusions in Guess Who provide information in this way …

C. Finn & J.T. Lizier, "Decomposing Local Information into Directed Positive and Negative Components", arXiv:1801.09223, 2018.

Mutual information (MI) – code

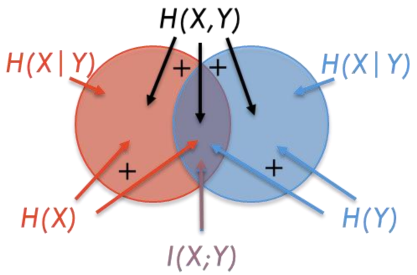


- The reduction in uncertainty or surprise about one variable *X* that we obtain from another variable *Y*

$$\mathrm{i}(x;y) = \mathrm{h}(x) + \mathrm{h}(y) - \mathrm{h}(x,y)$$
$$\mathrm{I}(X;Y) = \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y)$$

- Exercise: Let's code it!
  1. Edit the Matlab function `mutualinformation(p)` to return the MI between *X* and *Y* for the joint probability *p*.
     a. You can assume *p* is 2D (`p(x,y)`); this is the input.
     b. Trick: can we use our existing `entropy()` and `jointentropy()`?
     c. Test: `mutualinformation([0.5,0;0,0.5])`=1
     d. Test: `mutualinformation([0.25,0.25;0.25,0.25])` = 0
     e. Guess Who? I(`sex;earings`)? Construct p(`sex,earings`) first. Why is there MI here?

Mutual information (MI) – code



– The reduction in uncertainty or surprise about one variable X that we obtain from another variable Y

$$\mathrm{i}(x;y) = \mathrm{h}(x) + \mathrm{h}(y) - \mathrm{h}(x,y)$$
$$\mathrm{I}(X;Y) = \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y)$$

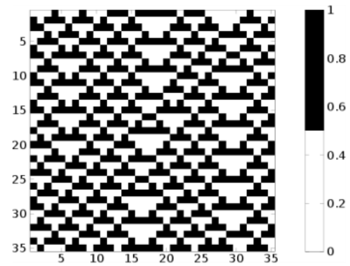
– Exercise: Let's code it!

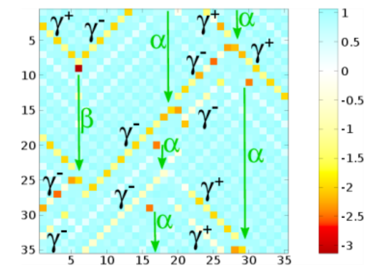
1. Edit the Matlab function `mutualinformationempirical(xn,yn)` to return the MI between X *and* Y from empirical samples x_n, y_n:
   a. Input is samples x_n, y_n.
   b. Trick: can we use our existing `jointentropyempirical()`?
   c. Test: `mutualinformationempirical([0,0,1,1], [0,1,0,1])` =0
   d. Test: `mutualinformationempirical([0,0,1,1], [0,0,1,1])` =1

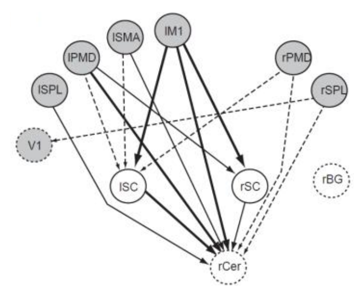
Mutual information (MI)

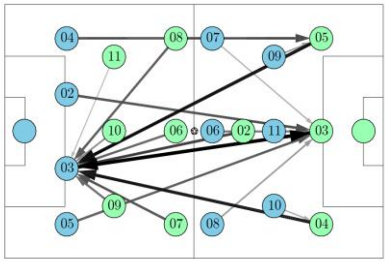
- Is a great model-free tool to:
  - detect relationships between variables;
  - reveal patterns;
  - show how such relationships and patterns fluctuate in time.

- Example uses:
  - Feature selection in machine learning
  - Space-time characterisation of information processing in complex systems – see later!
  - Inferring relationships (i.e. networks) in multivariate time-series data (e.g. brain imaging) – see later!









J. T. Lizier. "Measuring the dynamics of information processing on a local scale in time and space". In M. Wibral, R. Vicente, and J. T. Lizier, editors, "Directed Information Measures in Neuroscience", Springer, Berlin/Heidelberg, 2014; pp. 161–193.
J. T. Lizier, J. Heinzle, A. Horstmann, J.-D. Haynes, & M. Prokopenko. "Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fMRI connectivity". J. Computational Neuroscience, 30 (1):85–107, 2011.
O.M. Cliff, J.T. Lizier, P. Wang, X.R. Wang, O. Obst, M. Prokopenko, "Quantifying Long-Range Interactions and Coherent Structure in Multi-Agent Dynamics", Artificial Life, vol. 23, no. 1, pp. 34-57, 2017.

Conditional mutual information (CMI)

- *I(X;Y|Z)* is the reduction in uncertainty or surprise about one variable *X* that we obtain from another variable *Y*, given the value of another variable Z.
- Interpretation 1: in the context of Z –

$$\mathrm{I}(X;Y|Z) = \mathrm{H}(X|Z) + \mathrm{H}(Y|Z) - \mathrm{H}(X,Y|Z)$$
$$\mathrm{I}(X;Y|Z) = \mathrm{H}(X|Z) - \mathrm{H}(X|Y,Z)$$
$$\mathrm{I}(X;Y|Z) = \mathrm{H}(Y|Z) - \mathrm{H}(Y|X,Z)$$
$$\mathrm{I}(X;Y|Z) = \mathrm{I}(Y;X|Z)$$
$$\mathrm{I}(X;Y|Z) = \mathrm{I}(X;Y,Z) - \mathrm{I}(X;Z)$$

All maths changes to be conditional on Z

Ok this one is new!

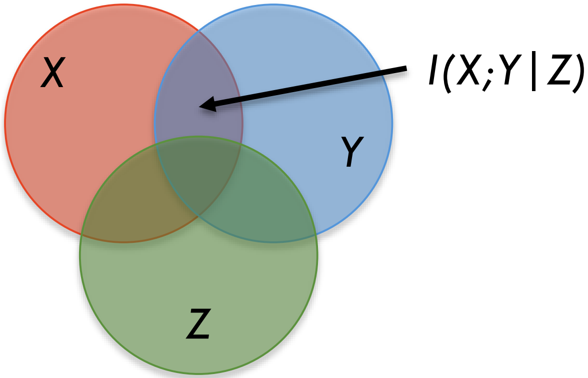
Properties:

- $0 \leq \mathrm{I}(X;Y|Z) \leq \min\big(\mathrm{H}(X|Z), \mathrm{H}(Y|Z)\big)$
  - e.g. if *Z* explains *X* (*H(X|Z)*=0), then $\mathrm{I}(X;Y|Z)$=0
- Is symmetric in *X* and *Y*
- $\mathrm{I}(X;Y|Z) = \mathrm{H}(X|Z) \rightarrow \mathrm{H}(X|Y,Z) = 0$

T. M. Cover and J. A. Thomas. Elements of Information Theory. Wiley-Interscience, New York, 1991. Section 2.5

Conditional mutual information (CMI)

- $I(X;Y|Z)$ is the reduction in uncertainty or surprise about one variable X that we obtain from another variable Y, given the value of another variable Z.
- Be warned against using Venn diagrams to interpret 3-term entropies!
  - Areas in the diagram add up correctly **but** the diagram gives the misleading impression that all areas are positive! (They aren't!)



  - Mackay emphasises that there are no other well-defined "3-term entropies"

T. M. Cover and J. A. Thomas. Elements of Information Theory. Wiley-Interscience, New York, 1991. Problem 2.25

D. J. C. MacKay. Information Theory, Inference, and Learning Algorithms. Cambridge University Press, Cambridge, 2003. Exercise 8.8 (sol'n p. 143) & p. 139

Conditional mutual information (CMI)

- *I(X;Y|Z)* is the reduction in uncertainty or surprise about one variable *X* that we obtain from another variable *Y*, given the value of another variable Z.
- Interpretation 2: KL divergence / Bayesian view –

$$\mathrm{I}(X;Y|Z) = \sum_{x \in A_x, y \in A_y, z \in A_z} p(x,y,z) \log_2 \frac{p(x,y|z)}{p(x|z)p(y|z)}$$

$$\mathrm{I}(X;Y|Z) = \sum_{x \in A_x, y \in A_y, z \in A_z} p(x,y,z) \log_2 \frac{p(x|y,z)}{p(x|z)}$$

Properties:

- $\mathrm{I}(X;Y|Z) = \mathrm{D}(p(x,y|z)||p(x|z)p(y|z))$
- CMI is code length penalty for coding {*x*,*y*} assuming *x* and *y* are conditionally independent (on z), *or* for coding *x* without using knowledge of *y in addition* to knowledge of *z*.

Conditional mutual information (CMI)

- $I(X;Y|Z)$ is the reduction in uncertainty or surprise about one variable X that we obtain from another variable Y, given the value of another variable Z.
- Interpretation 3: statistical view –

$$\mathrm{I}(X;Y|Z) = \sum_{x \in A_x, y \in A_y, z \in A_z} p(x,y,z) \log_2 \frac{p(x|y,z)}{p(x|z)}$$

Properties:

- $\mathrm{I}(X;Y|Z) = 0 \leftrightarrow$ X, conditional on Z, is independent of Y
- CMI is a non-linear form of partial correlation

T. M. Cover and J. A. Thomas. Elements of Information Theory. Wiley-Interscience, New York, 1991. Section 2.5
Bossomaier, Barnett, Harré, Lizier, "An Introduction to Transfer Entropy: Information Flow in Complex Systems", Springer, Cham, 2016; section 3.2.3

Local or pointwise Conditional Mutual information

- $i(x;y|z)$ is the reduction in uncertainty or surprise about one sample x of variable X that we obtain from one sample y another variable Y, given the sample z of another variable Z.

$$\mathrm{i}(x;y|z) = \mathrm{h}(x|z) + \mathrm{h}(y|z) - \mathrm{h}(x,y|z)$$
$$\mathrm{i}(x;y|z) = \mathrm{h}(x|z) - \mathrm{h}(x|y,z)$$
$$\mathrm{i}(x;y|z) = \mathrm{h}(y|z) - \mathrm{h}(y|x,z)$$
$$\mathrm{i}(x;y|z) = \log_2 \frac{p(x|y,z)}{p(x|z)}$$
$$\mathrm{I}(X;Y|Z) = \langle \mathrm{i}(x;y|z) \rangle$$

- $i(x;y|z)$ may be positive or negative (as per $i(x;y)$)

Bossomaier, Barnett, Harré, Lizier, "An Introduction to Transfer Entropy: Information Flow in Complex Systems", Springer, Cham, 2016; section 3.2.3

Conditional Mutual information (CMI) – code

- $I(X;Y|Z)$ is the reduction in uncertainty or surprise about one variable X that we obtain from another variable Y, given the value of another variable Z.

$$\mathrm{i}(x;y|z) = \mathrm{h}(x|z) + \mathrm{h}(y|z) - \mathrm{h}(x,y|z)$$
$$\mathrm{I}(X;Y|Z) = \mathrm{H}(X|Z) + \mathrm{H}(Y|Z) - \mathrm{H}(X,Y|Z)$$

- Exercise: Let's code it! (empirical only)
  1. Edit the Matlab function `conditionalmutualinformationempirical(xn,yn,zn)` to return the CMI between X *and* Y *given* Z from empirical samples x_n,y_n,z_n:
     a. Input is samples x_n,y_n,z_n.
     b. Trick: can we use our existing `conditionalentropyempirical()`?
     c. Test: `CMI([0,0,1,1], [0,1,0,1], [0,1,0,1])` =0
     d. Test: `CMI([0,0,1,1], [0,0,1,1], [0,1,1,0])` =1
     e. *Challenge*: compute using $\mathrm{I}(X;Y|Z) = \mathrm{I}(X;Y,Z) - \mathrm{I}(X;Z)$
     f. *Challenge*: write `conditionalmutualinformation (p)` (p is a 3D matrix!)

Conditional and unconditional mutual information

Conditioning on Z in $I(X;Y|Z)$, as compared to $I(X;Y)$ can:

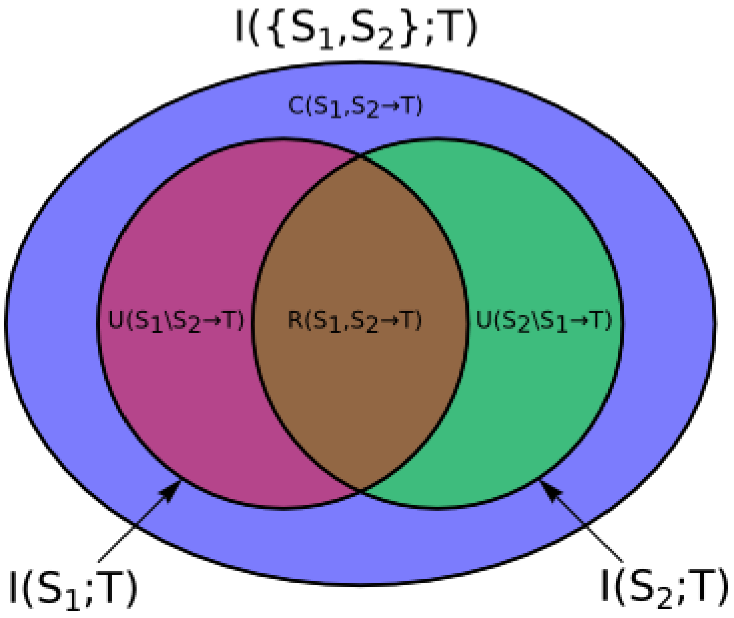
- Have no effect (if all variables are independent)
- Serve to decrease $I(X;Y|Z)$ compared to $I(X;Y)$
  - Y and Z carry redundant information about X.
  - Z explained away some of what could be detected by Y
  - e.g. If X=Y=Z are iid random bits, $I(X;Y|Z)=0$ although $I(X;Y)=1$
- Serve to increase $I(X;Y|Z)$ compared to $I(X;Y)$
  - Y and Z together provide synergistic information about X, which cannot be detected by examining either alone.
  - e.g. If X=Y `xor` Z, iid random bits, $I(X;Y|Z)=1$ although $I(X;Y)=0$.

- $I(X;Y|Z) - I(X;Y)$ being positive implies presence of synergy, or being negative implies presence of redundancy.
- But you can have both redundancy and synergy at once!

Williams and Beer, “Nonnegative decomposition of multivariate information”. arXiv:1004.2515, 2010.
Bossomaier, Barnett, Harré, Lizier, "An Introduction to Transfer Entropy: Information Flow in Complex Systems", Springer, Cham, 2016; section 3.2.3.1

Synergy and redundancy: Information decomposition

– Cannot measure redundancy and synergy with traditional info theory …
  – Need a new measure for redundancy (*out of scope*)



Williams and Beer, "Nonnegative decomposition of multivariate information". arXiv:1004.2515, 2010.

J.T. Lizier, N. Bertschinger, J. Jost, M. Wibral, "Information Decomposition of Target Effects from Multi-Source Interactions: Perspectives on Previous, Current and Future Work", Entropy, 20(4), 307, 2018

C. Finn and J.T. Lizier, "Pointwise Information Decomposition Using the Specificity and Ambiguity Lattices", Entropy, 20(4), 297, 2018

Chain rule for mutual information

- Chain rule for information:
  - $\mathrm{I}(X;Y,Z) = \mathrm{I}(X;Y) + \mathrm{I}(X;Z|Y)$
  - $\mathrm{I}(X;Y,Z) = \mathrm{I}(X;Z) + \mathrm{I}(X;Y|Z)$

$$I(X_1, X_2, \dots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_1 .. X_{i-1})$$

  - This is an information regression!

  - Same applies for $i(x;y)$, $I(X;Y|Z)$ and $i(x;y|z)$.

T. M. Cover and J. A. Thomas. Elements of Information Theory. Wiley-Interscience, New York, 1991. Section 2.5

Aside: Mutual information – derivation

– Pointwise mutual information between X and Y:

$$\mathrm{i}(x;y) = \log_2 \frac{p(x|y)}{p(x)}$$

and by implication $I(X;Y) = \langle i(x;y) \rangle$

– Is a unique form that satisfies four axioms:
  – Once-differentiability w.r.t. $p(x)$ and $p(x|y)$
  – Conditional form $i(x;y|z)$ matches $i(x;y)$ but with all PDFs conditioned on z
  – Additivity – $i(x;y,z) = i(x;z) + i(x;y|z)$
  – Separation for independent ensembles:
    - $p(x,y,u,v) = p(x,y)p(u,v) \rightarrow i(x,u;y,v) = i(x;y) + i(u,v)$

Fano, R.M.: Transmission of information: a statistical theory of communications. M.I.T. Press, Cambridge, MA, USA (1961)

What is information: summary

- We've been introduced to the ideas of uncertainty and surprise.
- Understand the meaning of information as uncertainty reduction
- Know how to calculate fundamental measures of information theory, from PDFs and empirically from data.

- *Coming up*: Move onto using a more advanced toolkit, and dealing with continuous-valued variables using a number of different estimators.

Questions

THE UNIVERSITY OF SYDNEY