Dependencies Between Random Variables Viewed as Entropy Areas

(Stuff cut out of short version)

by Jeff Edmonds
York University

Entropy is a hugely useful concept. We discuss it here in terms of the thermo dynamics, the expected number of bits needed to generate/specify a random object, and in compressing text.

1 Entropy

Thermo Dynamics: The second law of thermo dynamics says that the Entropy of a closed system always increases. In physics, Entropy is a measure of how much usable energy there is. This amounts to how much disorder there is in a system. It is measured as some constant times the log of the number of *micro* states the system might be in given one knows the *macro* state. 100 years later, Shannon related Entropy to information theory. Because it takes $\log_2 N$ bits to specify one of N states, the Entropy of a system can be viewed as (a constant times) the number of bits of information needed to reveal the micro state. For example, if the macro state consist of a nicely ordered crystal, then there are few possible positions that the atoms may be in and it would take very few bits to reveal where they all are. On the other hand, if the macro state consist of a hot gas, each atom has some unknown location and velocity. It would then take a lot of bits to reveal all of this information.

Compressing Text: Suppose you had text consisting of a sequence of objects each from the set $\{Obj_1, \ldots, Obj_N\}$. Your task is to compress this text by allocating to each object Obj_i a code consisting of a short bit string. If all n of the objects O_i are equally likely to appear in the text, then it makes sense to allocated each of them a code of length $\log n$. However, if some objects appear much more frequently then they should be allocated much shorter codes. One challenge with stringing together codes of different lengths is being able to uniquely decode what the original sequence of objects was. For example, you can not allocate Obj_1 the string 10, Obj_2 the string 11, and Obj_3 the string 1011, because then we would not know whether to decode 1011 as $Obj_1 Obj_2$ or as Obj_3. It is sufficient to require that no code is the prefix of another code. This is best viewed a putting the objects Obj_i on the leaves of a binary tree. Label each left edge zero and each right edge one. The code for Obj_i will be the string of labels in the path from the root to the leaf it is on. One decodes 1011... by starting at the root and heading left or right down the tree as indicated by the bits. When one reaches a leaf, the object O_i at this leaf is outputted and one starts back at the root in order to decode the next object.

Code Length: The next task is to decide the optimal length I_i for each code. Focus for a moment only on the i^{th} object. We will argue that if it appears with probability p_i

then optimally it should be allocated a code of length $I(p_i) = \log_1(\frac{1}{p_p})$. (Of course if this number is not an integer, then we might have to round it up a bit.) Here are two arguments for this. We have not considered all the other objects, but suppose they all had this same probability p_i of occurring. Then there would be $\frac{1}{p_i}$ objects and it would require $I(p_i) = \log_2(\frac{1}{p_i})$ bits to specify this object. The second argument is that if we allocate I_i bits to object Obj_i, then this object will be placed on a leaf of the binary tree at level I_i. In a full binary tree, there are 2^{I_i} nodes at this level. Hence, it is reasonable to say that that the object Obj_i has "used up" $\frac{1}{2^{I_i}}$ of the tree. Given that it appears with probability p_i it should only "use up" a p_i fraction of the tree. This motivates setting I_i so that $\frac{1}{2^{I_i}} = p_i$. Solving gives that $I = \log_2(\frac{1}{p_i})$.

Building the Tree: Having decided to allocated code of length $I(p_i) = \log_2(\frac{1}{p_i})$ to object Obj_i, the next task is to allocated the codes themselves. It turns out, that there is always a way to build a binary tree with the objects Obj_i on its leaves so that objects Obj_i is at depth $\lceil \log_2(\frac{1}{p_i}) \rceil$.

Expected Code Length: Given that we allocated a code of length $I(p_i) = \log_2(\frac{1}{p_i})$ to object Obj_i, we can now compute the expected length of the code. Recall how expectation is computed.

$$H(\{p_i\}) = \text{Exp}_{i\in\{p_i\}} I(p_i) = \sum_i p_i I(p_i) = \sum_i p_i \log_2(\frac{1}{p_i})$$

It turns out that, given the probabilities p_i, this is the optimal expected length of the code. This then becomes a lower bound on how much the text can be compressed or a measure of the "information content" of the text. It is referred to as the *entropy* H of the probability distribution.

Entropy of a Probability Distribution: A probability distribution $\mathcal{D}$ on a set of objects/states $\{Obj_1, \ldots, Obj_N\}$ is defined by specifying for each state/object obj_i, the probability p_i of being in that state or of choosing that object. Entropy $H(\mathcal{D})$ measures the amount of randomness in a probability distribution. As seen above, it is the expected number bits that need to be communicated in order to specify which object was chosen. Within one or two, it is also the expected number of fair coins that need to be flipped to generate an object according to this distribution (See homework question).

2 Summery of Lemmas

Lemma 1 *The relationships between the entropies, joint entropies, conditional entropies, and the mutual information between three random variables X, Y, Z is equivalent to the relationships between the areas of three over lapping circles X, Y, Z.*

1. **Entropy:** $H(X) = \sum_x p(x)L(x)$, where $p(x)$ is short for $Pr(X = x)$ and $L(x)$ is short for $\log(1/p(x))$. $H(X) = area(X)$.

Intuitively, $H(X)$ can be thought of as the expected length of shortest message to tell someone in the optimum way the value X happens to take on, where $L(x)$ is the length measured in bits to say $X = x$.

2. $0 \leq H(X) \leq \log_2(\#\text{ of different values})$.

3. $H(F(X)) \leq H(X)$ with equality when 1-1

4. **Joint Entropy:** $H(XY) = \sum_x \sum_y p(xy)L(xy)$, where $p(xy)$ is short for $Pr(X = x \ \&\ Y = y)$. $H(XY) = area(X \cup Y)$.

   Intuitively, It is the expected length of message to say both X & Y.

5. $H(XY) \leq H(X) + H(Y)$ with equality iff independent.

6. **Conditional Entropy:** $H(X|Y) = H(XY) - H(Y) = area(X \cap \overline{Y})$.

   Intuitively, $H(X|Y)$ is the expected length of message to tell you X after I have already told you Y.

7. $H(X|Y) \geq 0$ with equality iff Y determines X.

8. $H(X|Y) \leq H(X)$ with equality when independent.

9. $H(XY|Z) \leq H(X|Z) + H(Y|Z)$ with equality iff independent conditional on Z.

10. $H(F(X,Y)) \neq \sum_y H(F(X,y))$.

11. **Mutual Information:** $I(X;Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(XY) = area(X \cap Y)$.

    Intuitively, $I(X;Y)$ is the information that is common to both X and Y. It is the amount about X you learn from me telling you Y. Perhaps surprisingly, this is equal to the amount about Y you learn from me telling you X.

12. $I(X;Y) \geq 0$ with equality iff independent.

13. $I(X;Y) \leq H(X)$ with equality iff $X = Y$.

14. **Joint Mutual Information:** $I(XY;Z) = area((X \cup Y) \cap Z)$.

15. $I(XY;Z) \not\approx I(X;Z) + I(Y;Z)$ with $\geq$ if X and Y are independent.

16. $I(XY;Z) \leq I(X;Z) + H(Y)$.

17. $I(U;AR) \leq I(UR;A)$ when U and R are independent.

18. **Conditional Mutual Information:** $I((X;Y)|Z) = H(X|Z) - H(X|YZ) = area(X \cap Y \cap \overline{Z})$.

    Intuitively, it is the information common to X and Y after you have already told me Z. Or after you have already told me Z, it is the amount of additional information I learn about X from you telling me Y.

19. $0 \leq I((X;Y)|Z) \leq H(Y|Z) \leq H(Y)$.

20. $I((F(X,Y);Z)|Y) \leq I(X;Z)$.

21. **Strange Area:** $I(X;Y;Z) = Area(X \cap Y \cap Z) = I(X;Y) - I((X;Y)|Z) \not\geq 0$.

22. $I((X;Y;Z)|W) \geq 0$ when X and Z are conditionally independent given Y and W.

23. **Group Learning:** $\sum_j I((X_j;Z)|Y_j) \leq I((\langle X_1, X_2, \ldots, X_n \rangle;Z)|\langle Y_1, Y_2, \ldots, Y_n \rangle) \leq H(Z)$ assuming for all disjoint subsets J and $J' \subseteq [n]$, X_J and $X_{J'}$ are independent conditional on $\langle Y_J, Y_{J'} \rangle$, and X_J is independent of $Y_{J'}$ conditional on Y_J.

    Intuitively, the sum of the amounts people in your class who came in with the knowledge X_i learned individually about their question Y_i from your lecture Z is at most that learned collectively.

24. **Possible Random Variables and Areas of Primitives:** Areas of circles X, Y, and Z correspond to random variables X, Y, and Z if and only if the area of each *dual primitive* must be positive, with the exception of $X \cap Y \cap Z$, which could be negative and for each pair of random variables the mutual information $I(X;Y) = X \cap Y$ must also be positive.

25. **Communication Complexity:** Let Π denote the transcript of a communication between two players with inputs X and Y and with private random bits. Consider a third player, Alice. Alice knowing both X and Y sends a message A consisting of m bits to the Y player (or to both of them.) Then the X and Y players have the conversation with transcript Π.

26. **Negative Area:**

    (a) $I(X;Y;\Pi) \geq 0$

    (b) $I(X;Y;A\Pi) \geq -m$.

    (c) $\sum_{k \in [K], j \in [n]} I(X_{\langle k,j \rangle};Y_j;A\Pi) \geq -m$.
    This requires that conditioned on Y, the rows of X are independent.
    It also requires that conditioned on Y, the bits $X_{\langle k,j \rangle}$ and $X_{\langle k,j' \rangle}$ are independent.
    It also requires that conditioned on Y_{-j}, the bits $X_{\langle k,j \rangle}$ and Y_j are independent.
    This is confusing. Just make all the bits independent.

    (d) $\sum_{k \in [K], j \in [n]} I((X_{\langle k,j \rangle};Y_j;A\Pi)|\langle X_{\langle *,-j \rangle}, Y_{-j} \rangle) \geq -nm$.
    This requires that conditioned on Y, the rows of X are independent.

    Where X is a matrix with $k \in [K], j \in [n], X_{\langle k,j \rangle}$ and Y_j a row.

27. **What People Learn:** Suppose there is an eves dropper, Eve, who learns the conversation Π but knows neither X nor Y. By definition $I(XY;\Pi)$ is what she learns about the inputs $\langle X, Y \rangle$ from their conversation Π, $I(X;\Pi)$ is what she learns about the X-player's inputs X, and $I(Y;\Pi)$ about the Y-player's inputs Y. The X-player already knows X, and hence from Π the amount about Y that he learns is denoted $I((Y;\Pi)|X)$. Similarly, the Y-player learns $I((X;\Pi)|Y)$ about X.

- $I((Y;Z)|X) + I((X;Z)|Y) \geq I(XY;Z) \geq I(X;Z) + I(Y;Z)$ when X and Y are independent.
- $I((Y;\Pi)|X) + I((X;\Pi)|Y) \leq I(XY;\Pi) \leq I(X;\Pi) + I(Y;\Pi)$.
- Same with equality when X and Y are independent.
- $I((Y;A\Pi)|X) + I((X;A\Pi)|Y) \leq I(XY;A\Pi) + m$.

3 Proofs of the Lemmas

Entropy $H(X)$ (Proof of Lm1.1): Just as there are three ways of understanding the random variable X, there are corresponding ways of understanding the *entropy* $H(X)$ of X.

Random Variables: $H(X)$ is said to be the *entropy of* X.

Computationally: It is formally defined as $H(X) = \sum_x p(x)L(x)$, where $p(x)$ is short for $Pr(X = x)$ and $L(x)$ is short for $\log(1/p(x))$.

Information: $H(X)$ can be thought of as the expected length of shortest message to tell someone in the optimum way the value X happens to take on, where $L(x)$ is the length measured in bits to say $X = x$.

Generating X: If you were to write a program that flips as few coins as possible in order to determine which value X should take, then $H(X)$ would be the expected number of coins that would need be flipped, where $L(x)$ is the number flipped when the program decides that $X = x$.

Circles: $H(X)$ can also be viewed as the area of the circle representing X.

$0 \leq H(X) \leq \log_2(\#$ of different values) (Proof of Lm1.2): One can identify each n different objects, using a binary label containing $\log_2(n)$ bits. Given this is one way to communicate X, the optimal way takes at most this many bits.

$H(F(X)) \leq H(X)$ with Equality when 1-1 (Proof of Lm1.3): Here F is a function that maps each objects/events/values that X takes to some other object/event/value. For each of these new objects f, there is a probability that $F(X,Y) = f$. Hence, F becomes a random variable in its own right. Remember that the which objects X takes on does effect the entropy $H(X)$ of X, only how the probability is distributed between them. Hence, if X takes on the objects apple, orange, and pair, and f maps these to 1, 2, and 3, then the entropy does not change. However, if f collapse apple and orange to the same value 1, then the entropy goes down.

Proof: The key observation is that for each x, $\Pr(F(X)=F(x)) \geq \Pr(X=x)$, because when ever $X = x$, we have that $F(X) = F(x)$, but it may be that $F(X) = F(x') = F(x)$ when $X = x' \neq x$. From this we bound the entropy. $H(F(X)) = \sum_f \Pr(F(X,Y) = f)\log(1/\Pr(F(X) = f) = \sum_f [\sum_x p_x(\text{whether } F(x)=f)]\log(1/\Pr(F(X)=f) = \sum_x p_x \log(1/\Pr(F(X)= F(x)) \leq \sum_x p_x \log(1/\Pr(X=x)$.

Joint Entropy $H(XY)$ (Proof of Lm1.4): (Often written $H(X,Y)$ or $H(\langle X,Y\rangle)$).

- **Random Variables:** If X and Y are random variables then $\langle X,Y\rangle$ is the *joint* random variable telling you both the value of X and of Y. The joint entropy, $H(XY)$, is to defined to be the entropy of $\langle X,Y\rangle$.
- **Computationally:** It is formally defined as $H(XY) = \sum_x \sum_y p(xy)L(xy)$, where $p(xy)$ is short for $Pr(X = x \;\&\; Y = y)$.
- **Information:** It follows that $H(XY)$ is the expected length of message to say both X & Y.
- **Circles:** $H(XY)$ is the area of $X \cup Y$.
- **$H(XY) \leq H(X) + H(Y)$ with Equality iff Independent (Proof of Lm1.5):**
  - **Intuition:** Intuitively, this is true because one could say both X and Y by saying X and then saying Y, but perhaps one could save time by saying them together if some of the information overlaps. But if X & Y are independent, then no information overlaps and $H(XY) = H(X) + H(Y)$.
  - **Proof:** With independence the proof is as follows.

    $L(xy) = \log(1/p(xy))$, which by independence is $\log(1/(p(x)p(y))) = L(x) + L(y)$. This gives $H(XY) = \sum_x \sum_y p(xy)L(xy) = \sum_x \sum_y p(xy)L(x) + \sum_x \sum_y p(xy)L(y) = \sum_x \left[\sum_y p(xy)\right] L(x) + \sum_y \left[\sum_x p(xy)\right] L(y) = \sum_x p(x)L(x) + \sum_y p(y)L(y) = H(x) + H(y)$.

    The proof that without independence $H(XY) \leq H(X) + H(Y)$ is harder. Note that the terms for which $L(xy) > L(x) + L(y)$ have their weight $p(xy)$ larger because $p(xy) > p(x)p(y)$. Similarly the terms for which $L(xy) < L(x) + L(y)$ have their weight $p(xy)$ smaller because $p(xy) < p(x)p(y)$.

Conditional Entropy $H(X|Y)$ (Proof of Lm1.6):

- **Random Variables:** $H(X|Y)$ is defined to be the entropy of X *conditional on* Y.
- **Computationally:** It is formally defined as $H(X|Y) = \sum_y p(y)H(X|y)$, but this is not so intuitive.
- **Information:** Intuitively, $H(X|Y)$ is the expected length of message to tell you X after I have already told you Y.
- **Circles:** Pictorially, $H(X|Y)$ is the area of $X - Y = X \cap \overline{Y}$. This is the area of $X \cup Y$ minus the area of Y.
- **$H(X|Y) = H(XY) - H(Y)$:** This seems to be a more intuitive and a more useful definition of mutual information.
  - **Intuition:** Thinking of mutual information as areas of area of $X \cup Y$ minus the area of Y leads us to understand that $H(X|Y) = H(XY) - H(Y)$ is true.
  - **Primitives:** I like this definition because it is wrt the "primitives" $H(X)$, $H(Y)$ and $H(XY)$.

Conditional Probabilities: I also like it because it looks like $p(x|y) = \frac{p(xy)}{p(y)}$ when you take the log of both sides.

Proof: The proof of $H(X|Y) = H(XY) - H(Y)$ is as follows.
$H(X|Y) = \sum_y p(y) H(X|y) = \sum_y p(y) \left[\sum_x p(x|y) L(x|y)\right]$
$= \sum_y p(y) \left[\sum_x \frac{p(xy)}{p(y)} \log(\frac{p(y)}{p(xy)})\right] = \sum_x \sum_y p(xy) \left[L(xy) - L(y)\right]$
$= \left[\sum_x \sum_y p(xy) L(xy)\right] - \left[\sum_x \sum_y p(xy) L(y)\right] = H(XY) - \sum_y \left[\sum_x p(xy)\right] L(y)$
$= H(XY) - \sum_y p(y) L(y) = H(XY) - \sum_y p(y) L(y) = H(XY) - H(Y)$.

$H(X|Y) \geq 0$ with Equality iff Y Determines X (Proof of Lm1.7):

Intuition: After telling you Y, either you know X in which case $H(X|Y) = 0$ or I need to tell you more.

Proof: The proof is easy using our new definition, $H(X|Y) = H(XY) - H(Y)$ which is positive because surely it take more to tell you X and Y than to simply tell you Y.

$H(X|Y) \leq H(X)$ with Equality when Independent (Proof of Lm1.8):

Intuition: The intuition is that it can only be easier to tell you X after I have only told you Y.

Proof: The proof is easy using our new definition, $H(X|Y) = H(XY) - H(Y) \leq [H(X) + H(Y)] - H(Y) = H(X)$.

$H(XY|Z) \leq H(X|Z) + H(Y|Z)$ with Equality iff Independent Conditional on Z (P

Intuition: After I have told you Z, it is no harder to tell you X and Y together than to tell you each separately.

Proof: The proof of this is hard like that for $H(XY) \leq H(X) + H(Y)$.

$H(F(X,Y)) \neq \sum_y H(F(X,y))$ (Proof of Lm1.10): It is natural to assume that because entropy is a weighted sum that one can decompose it into a weighted sum. However, it is not this easy. $\sum_y H(F(X,y)) = \sum_y H(F(X,Y)|Y = y) = H(F(X,Y)|Y) \neq H(F(X,Y))$.

Mutual Information $I(X;Y)$ (Proof of Lm1.11):

Random Variables: $I(X;Y)$ is said to be the *mutual information* between X and Y.

Information: $I(X;Y)$ is the information that is common to both X and Y.
It is the amount about X you learn from me telling you Y.
Perhaps surprisingly, this is equal to the amount about Y you learn from me telling you X.

Computationally: It is formally defined as $I(X;Y) = H(X) - H(X|Y)$,
namely how much less do I have to tell you to tell you X after I have told you Y.

Primitives: Decomposed it into primitives gives $I(X;Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(XY)$.

- **Circles:** $I(X;Y)$ is the area of $X \cap Y$. This is the area of X plus that of Y minus that of $X \cup Y$.
- **$\langle X;Y\rangle$ is Not a Random Variable:** When seeing $H(X;Y)$, it is tempting to think of $\langle X;Y\rangle$ as a random variable and/or information to be be communicated. This is not the case and can lead to faulty intuition at times. The information $\langle X;Y\rangle$ common between X and Y is something that may be hard to communicate separate from communicating X and Y. It is not a random variable to communication in its own right. Maybe this is why they use the letter I in $I(X;Y)$ instead of an H as in $H(X;Y)$.
- **$I(X;Y) \geq 0$ with Equality iff Independent (Proof of Lm1.12):**
  - **Intuition:** $I(X;Y) \geq 0$ is true because the amount of common information cant be negative. Clearly if X and Y are independent then $I(X;Y) = 0$, because they contain no information about each other.
  - **Proof:** $I(X;Y) \geq 0$ follows from our new definition $I(X;Y) = H(X) + H(Y) - H(XY)$ and that $H(XY) \leq H(X) + H(Y)$ with equality iff independent.
- **$I(X;Y) \leq H(X)$ with Equality iff $X = Y$ (Proof of Lm1.13):**
  - **Intuition:** $I(X;Y) \geq 0$ is true because the amount of common information cant be negative. Clearly if X and Y are independent then $I(X;Y) = 0$, because they contain no information about each other.
  - **Proof:** $H(X) - I(X;Y) = H(X) - [H(X) - H(X|Y)] = H(X|Y) \geq 0$.

Joint Mutual Information $I(XY;Z)$ (Proof of Lm1.14): When I tell you Z, what does this tell you about the joint entropy of X and Y?

- **$I(XY;Z) \not\approx I(X;Z) + I(Y;Z)$ (Proof of Lm1.15):** When understanding a subject, it is important to understand not only what is true but also what intuitively one might think is true but is not. Despite ones intuition, no direct comparison can be made between $I(XY;Z)$ and $I(X;Z) + I(Y;Z)$.
  - **$I(XY;Z) \not\leq I(X;Z) + I(Y;Z)$:**
    - **False Intuition:** One might first guess that this would be similar to $H(XY) \leq H(X) + H(Y)$, namely that the amount you learn about XY from Z is at most the you learn about X plus the amount you learn about Y.
    - **Example:** Consider the example in which $X = Y$ and $I(X;Z) > 0$. Then $I(XY;Z) = I(X;Z) = I(Y;Z)$. It follows that for this example $I(XY;Z) < 2I(XY;Z) = I(X;Z) + I(Y;Z)$.
  - **$I(XY;Z) \not\geq I(X;Z) + I(Y;Z)$:**
    - **False Intuition:** On the other hand, one could argue the opposite. You learn at least as much about Z from X and Y, than you learn about Z from each of them separately.
    - **Example:** Consider the example $Z = X \oplus Y$, where X and Y are independent boolean variables. Here knowing Z tells you nothing about X and

similarly nothing about Y. However, knowing Z tells you a full bit of information about $\langle X, Y\rangle$, namely their parity. It follows that for this example $1 = I(XY; Z) > I(X; Z) + I(Y; Z) = 0$.

$I(XY; Z) \geq I(X; Z) + I(Y; Z)$ if X and Y are independent: See the intuition and example above.

Proof: The proof is as follows. $I(XY; Z) = H(XY) - H(XY|Z)$. Because of independence, we have that $H(XY) = H(X) + H(Y)$. However, we can only be sure that $H(XY|Z) \leq H(X|Z) + H(Y|Z)$ unless, X and Y are independent conditional on Z. It follows that $I(XY; Z) \geq [H(X) - H(X|Z)] + [H(Y) - H(Y|Z)] = I(X; Z) + I(Y; Z)$.

Key: This is key in proving lower bound in communication complexity.

$I(XY; Z) \leq I(X; Z) + H(Y)$ (Proof of Lm1.16): This is similar in nature to the previous comparison, but it is true.

Intuition: The intuition is as follows. Suppose someone knows X and from this can deduce $I(X; Z)$ about Z. Then suppose someone tells him Y. Now he knows X and Z. The amount that he can now deduce about Z is denoted $I(XY; Z)$. Surely, this is at most the about $I(X; Z)$ that he could deduce before plus the number of bits $H(Y)$ needed to communicate Y.

Proof: This can be proved as follows. Look at the 2^3 dual primitive formed from the intersections of the three circles for X, Y, and Z. In each separate dual primitive area put a +1 when within the area $X \cap Z$ for $I(X; Z)$ and another +1 when within the area Z. Similarly, put a -1 when within each dual primitive area within $XY \cap Z$ for $I(XY; Z)$. Summing these gives a zero in every dual primitive area except for three, which are the union of $X \cap Y$ and $Y - (X \cup Z)$. These two areas are equal to $I(X; Y)$ and $H(Y|XZ)$. All this proves that $RHS - LHS = I(X; Z) + H(Y) - I(XY; Z) = I(X; Y) + H(Y|XZ)$. One could prove this more formally by expanding each into the primitives and making sure they all cancel. We also know that both $I(X; Y)$ and $H(Y|XZ)$ are positive. The statement follows.

$I(U; AR) \leq I(UR; A)$ when U and R are independent (Proof of Lm1.17): Proof: $I(U; AR) = I(U; A|R) + I(U; R) = I(U; A|R) + 0 \leq I(U; A|R) + I(A; R) = I(UR; A)$. ■

Conditional Mutual Information $I((X; Y)|Z)$ (Proof of Lm1.18): It is often written as $I(X; Y|Z)$ but its is parsed as $I((X; Y)|Z)$.

Information: $I((X; Y)|Z)$ is the information common to X and Y after you have already told me Z. Or after you have already told me Z, $I((X; Y)|Z)$ is the amount of additional information I learn about X from you telling me Y.

Computationally: $I((X; Y)|Z) = H(X|Z) - H(X|YZ)$
$= H(X|Z) + H(Y|Z) + H(XY|Z)$
$= H(XZ) + H(YZ) - H(Z) - H(XYZ)$.

Circles: $I((X;Y)|Z)$ is the area of $(X \cap Y) - Z$ which is the single dual primitive $X \cap Y \cap \overline{Z}$.

$0 \leq I((X;Y)|Z) \leq H(Y|Z) \leq H(Y)$ (Proof of Lm1.19): All of these are reasonable and useful things that are true but need to be proved.

Proof of $I((X;Y)|Z) \geq 0$: By definition $I((X;Y)|Z) = H(X|Z) + H(Y|Z) + H(XY|Z)$. But we have already seen that $H(XY|Z) \leq H(X|Z) + H(Y|Z)$.

Proof of $I((X;Y)|Z) \leq H(Y|Z)$: $H(Y|Z) - I((X;Y)|Z) = H(Y|Z) - [H(Y|Z) - H(Y|XZ)] = H(Y|XZ) \geq 0$.

Proof of $H(Y|Z) \leq H(Y)$: $H(Y) - H(Y|Z) = I(Y;Z) \geq 0$.

$I((F(X,Y);Z)|Y) \leq I(X;Z)$ (Proof of Lm1.20): The intuition is that if I tell you the value of Y, then $F(X,Y)$ simply becomes a function $F_y(X) = F(X,y)$ dependent only on X. As we have seen above $H(F_y(X)) \leq H(X)$. Similarly, $I((F_y(X);Z) \leq I(X;Z)$.

Proof: Could write one ????

Strange Area $I(X;Y;Z) = Area(X \cap Y \cap Z) \not\geq 0$ (Proof of Lm1.21): See Section **??** to see how parity can give a negative area in the middle.

$I(X;Y;Z) = I(X;Y) - I((X;Y)|Z) \not\geq 0$: The question $I(X;Y)$ vs $I((X;Y)|Z)$ is whether or not there can be more common information between X and Y after I tell you Z.

False Intuition: We already have seen $I(W;Z) = H(W) - H(W|Z) \geq 0$. Hence, it would be natural to generalize to $I(X;Y;Z) = I((X;Y);Z) = H((X;Y)) - H((X;Y)|Z) \geq 0$. However, we will see that this is not true. It is not directly true because $\langle X;Y \rangle$ is not a random variable in its own right.

Circles: Lets try to expand $I(X;Y) - I((X;Y)|Z) = [H(X) + H(Y) - H(XY)] - [H(XZ) + H(YZ) - H(Z) - H(XYZ)]$. Look at the three intersecting circles. In each dual primitive area put a +1 when the area is added and a -1 when it is subtracted. Summing these gives a zero every where except for a one in the intersection dual primitive area $X \cap Y \cap Z$. This concurs with the intuition that this is the information that is common between all three of X, Y and Z, namely $I(X;Y;Z) = I(X;Y) - I((X;Y)|Z)$.

$I((X;Y;Z)|W) \geq 0$ when X and Z are conditionally independent given Y and W (

Positive Area: $I(X;Y;Z)$ is the area of the dual primitive $X \cap Y \cap Z$. It is counter intuitive for this area to be negative. We claim, in fact, that most counter intuitive things about mutual entropy arise from the fact that this area can be negative. Hence, it is interesting to look at conditions in which it is not. Conditioning it on another variable W, just makes the result more general. Assuming W is a constant removes mention of it.

$I((X;Y)|Z) \leq I(X;Y)$ when X and Z are conditionally independent given Y: This is a restatement of the above because by definition $I(X;Y;Z) = I(X;Y) - I((X;Y)|Z)$. This was listed in the [YJLS]. I am not sure how they used, it but we will find it very useful.

Proof: By assumption, X and Z are conditionally independent given Y and W. Formally this means $I((X;Z)|YW) = 0$. We have not considered the intersections of four random variables. But we can think of $U = YW$ as corresponding to $U = Y \cup W$. Above we defined $I((X;Z)|U) = area(X \cap Z \cap \overline{U}) = area(X \cap Z \cap \overline{Y \cup W}) + area(X \cap Z \cap \overline{Y} \cap \overline{W}) = 0$.

What is always true is that mutual information $I((X;Z)|W) = area(X \cap Z \cap \overline{W}) \geq 0$.

It follows that $I((X;Y;Z)|W) = area(X \cap Z \cap Y \cap \overline{W}) = area(X \cap Z \cap \overline{W}) - area(X \cap Z \cap Y \cap \overline{W}) = I((X;Z)|W) - 0 \geq 0$.

Group Learning (Proof of Lm1.23): $\sum_j I((X_j;Z)|Y_j) \leq I((\langle X_1, X_2, \ldots, X_n\rangle;Z)|\langle Y_1, Y_2, \ldots, Y_n\rangle) \leq H(Z)$ assuming for all disjoint subsets J and $J' \subseteq [n]$, X_J and $X_{J'}$ are independent conditional on $\langle Y_J, Y_{J'}\rangle$, and X_J is independent of $Y_{J'}$ conditional on Y_J.

Intuition: Suppose you are teaching a class to n people. For each $j \in [n]$, let Y_j denote the information that the j^{th} person knows before the lecture. Let X_j denote the information that he personally wants to learn, hopefully from the lecture. Let Z denote the information taught at the lecture. Then $I((X_j;Z)|Y_j)$ denotes how much the j^{th} person learns during the lecture about his personal question and $I((\langle X_1, X_2, \ldots, X_n\rangle;Z)|\langle Y_1, Y_2, \ldots, Y_n\rangle)$ denotes how much the class collectively learns about their collective questions. Requiring that X_J and $X_{J'}$ are independent conditional on $\langle Y_J, Y_{J'}\rangle$ asserts that for any disjoint subsets of people $J \neq J' \in [n]$, given their combined knowledge, their questions are independent of each other. Requiring that X_J is independent of $Y_{J'}$ conditional on Y_J asserts that given what the J of people know, the other people's previous knowledge will not help him with his personal question. The conclusion is that the sum of the amounts learned individually is at most that learned collectively.

Proof: It is only necessary to prove it for two people, then the result for n people follows by induction.

(X_1 and X_2 are independent conditional on $\langle Y_1, Y_2\rangle$) thought of as areas of circles translates into $area(X_1 \cap X_2 \cap \overline{Y}_1 \cap \overline{Y}_2) = 0$.

($I((X_1;X_2)|Y_1Y_2Z) \geq 0$) translates into $area(X_1 \cap X_2 \cap \overline{Y}_1 \cap \overline{Y}_2 \cap \overline{Z}) \geq 0$.

Combining the last two statements gives that $area(X_1 \cap X_2 \cap \overline{Y}_1 \cap \overline{Y}_2 \cap Z) = area(X_1 \cap X_2 \cap \overline{Y}_1 \cap \overline{Y}_2) - area(X_1 \cap X_2 \cap \overline{Y}_1 \cap \overline{Y}_2 \cap \overline{Z}) \leq 0$. Call this statement (1).

(X_2 is independent of Y_1 given Y_2) translates into $area(X_2 \cap Y_1 \cap \overline{Y}_2) = 0$.

($I((X_2;Y_1)|Y_2Z) \geq 0$) translates into $area(X_2 \cap Y_1 \cap \overline{Y}_2 \cap \overline{Z}) \geq 0$.

Combining the last two statements gives that $area(X_2 \cap Y_1 \cap \overline{Y}_2 \cap Z) = area(X_2 \cap Y_1 \cap \overline{Y}_2) - area(X_2 \cap Y_1 \cap \overline{Y}_2 \cap \overline{Z}) \leq 0$.

We use this result to bound the following. $I((X_2; Z)|Y_2) = area(X_2 \cap \overline{Y}_2 \cap Z) = area(X_2 \cap Y_1 \cap \overline{Y}_2 \cap Z) + area(X_2 \cap \overline{Y}_1 \cap \overline{Y}_2 \cap Z) \leq area(X_2 \cap \overline{Y}_1 \cap \overline{Y}_2 \cap Z) = area(X_1 \cap X_2 \cap \overline{Y}_1 \cap \overline{Y}_2 \cap Z) + area(\overline{X}_1 \cap X_2 \cap \overline{Y}_1 \cap \overline{Y}_2 \cap Z)$. Call this statement (2).

By symmetry of statement (2), get the following statement (3) that $I((X_1; Z)|Y_1) \leq area(X_1 \cap X_2 \cap \overline{Y}_1 \cap \overline{Y}_2 \cap Z) + area(X_1 \cap \overline{X}_2 \cap \overline{Y}_1 \cap \overline{Y}_2 \cap Z)$.

$I((\langle X_1, X_2\rangle ; Z)| \langle Y_1, Y_2\rangle) = area((X_1 \cup X_2) \cap \overline{Y}_1 \cap \overline{Y}_2 \cap Z) = area(X_1 \cap X_2 \cap \overline{Y}_1 \cap \overline{Y}_2 \cap Z) + area(\overline{X}_1 \cap X_2 \cap \overline{Y}_1 \cap \overline{Y}_2 \cap Z) + area(X_1 \cap \overline{X}_2 \cap \overline{Y}_1 \cap \overline{Y}_2 \cap Z)$. Call this statement (4).

Combining statements (2,3,4) gives that $I((X_1; Z)|Y_1) + I((X_2; Z)|Y_2) - I((\langle X_1, X_2\rangle ; Z)| \langle Y_1, Y_2\rangle) \leq area(X_1 \cap X_2 \cap \overline{Y}_1 \cap \overline{Y}_2 \cap Z)$. Then by statement (1), this less than or equal to zero. ■

4 Predictability

In this section we introduce Russell Impagliazzo's idea of *Predictability.*

- J. Edmonds, R. Impagliazzo, S. Rudich, and J. Sgall, "Communication Complexity Towards Lower Bounds on Circuit Depth," *Journal of Computational Complexity,* 10: pp 210-246, 2001. Previously in *FOCS, Symp. Foundations of Computer Science*, pp. 249-257, 1991.

For example, suppose over a sequence of communication bits, the A-Player tells the B-Player, "I am not telling you any of my values, but I will tell you that if my coin flip is heads, then all k bits in my vector are zero. On the other hand, if this flip is tails, then I reveal no information about this vector." The question now is whether the B-Player is considered to know k or zero bits about this vector. A useful information measure for many applications is Entropy. Because half the time k bits about the vector are revealed and half the time 0 bits are revealed, Entropy measures the number of bits revealed as the average $(k+0)/2 = k/2$. Our adversary, however, wants to be more cautious by assuming that the B-Player knows more than this. We define a measure of "predictability" to be the probability of guessing the value. If the B-Player completely knows the vector then the probability of guessing it is 1 and if he knows nothing about it, then the probability is 2^{-k}. The measure is the average of these, $(1 + 2^{-k})/2 \approx 1/2$. A predictability of 1/2 is then interpreted to mean that the B-Player knows everything but "1 bit" about this vector. The next section formally defines this measure.

Def^n 4.1 *Let $v \in [k]^\pi$ a vector be a vector indexed by entries in π and values in $[1..k]$ and let S be a set of such vectors. For every subset of indices $\rho \subseteq \pi$, let $Proj(v, \rho) \in [k]^\rho$ be the sub-vector of v indexed by the indices in ρ and let $Proj(S, \rho) = \{Proj(v, \rho) \mid v \in S\}$ be the* **projection** *of S onto ρ. A function R from $Proj(S, \rho)$ to S is called an* **extension function** *if for all $w \in Proj(S, \rho)$, $Proj(R(w), \rho) = w$.*

Suppose an input v specifies for each index $i \in \pi$, a value $v_i = Proj(v,i) \in [k]$. Then there are $k^{|\pi|}$ possible inputs v. Suppose someone restricts this set of possible inputs to a set of size $|S| \geq \left(\frac{r}{k}\right)^{\ell} \times k^{|\pi|}$. We will interpret this as them revealing some $t = \ell \log(k/r)$ bits about the input. An interesting question is, for $i \in \pi$, how many of these bits were communicated "about the i^{th} element" conditioned on knowing the other elements? Since the actual bits communicated could depend on all the elements, this is not a clear-cut issue. Our measure is computed as follows. Choose a random vector $w \in Proj(S, \pi - i)$ giving values of all the elements other than i. The set $\{v_i \in [k] \mid \langle w, v_i \rangle \in S\} \equiv \{v \in s | \ Proj(v, \pi - i) = w\}$ is the set of values for the i^{th} element (or full vectors) consistent with our chosen values w for the other elements. We define the unpredictability of v_i to be the expected number of such choices.

Def^n 4.2 *The* **unpredictability** *of the i^{th} element in S is $UnPred_i(S) = \frac{|S|}{|Proj(S, \pi - i)|}$.* [1]

If the i^{th} element v_i of $v \in S$ is fixed as a function of the other elements, then $UnPred_i(S) = 1$. If this element is completely undetermined, then $UnPred_i(S) = k$. If $UnPred_i(S) = r$, we can think of $t = \log(k/r)$ as the "number of bits known about element i", since if S is the set of inputs consistent with this number of independent bits communicated about v_i, then $UnPred_i(S) = \frac{k}{2^t} = r$.

Suppose $t = \ell \log(k/r)$ bits have been communicated about the vectors v in S, i.e., $|S| = 2^{-t} \cdot k^{|\pi|}$. A natural property to want is that at most ℓ elements of v can have more than $\frac{t}{\ell} = \log(k/r)$ bits "revealed about it". The exemplary counter example is the following. Suppose that the sum of the elements over the field $[k]$ was revealed. This requires only $\log(k)$ bits to be communicated. On one hand, it feel like nothing has been revealed about any one element because each still can uniformly take on any value. On the other hand, each element has been completely revealed, conditioned on knowing the other elements. This gives that for each $i \in \pi$, $UnPred_i(S) = 1$, implying that $\log(k)$ bits have been communicated "about each of the elements". We will get around this problem as follows. When an element $i \in \pi$ becomes highly predictable in S, we start ignoring it by considering only the possible settings of the other elements $Proj(S, \pi - i)$ in place of S. The value of element i is fixed as a function of the other elements using an extension function $R(w)$ as defined in Definition 4.1. In our previous example, if any one of element v_i is fixed to be $v_i = sum - \sum_{j \neq i} v_j$, then no information at all is known about the remaining elements. The following lemma then gives us what we wanted, that if at most $t = \ell \log(k/r)$ bits have been revealed about S, then there exists a set of at most ℓ elements such that, if we fixed them in this way, then no more than $\frac{t}{\ell} = \log(k/r)$ bits have been "revealed about" any of the other elements and hence there are still $\frac{k}{2^{t/\ell}} = r$ values left for each.

Lemma 2 *[Lemma 4.6 in [EIRS]] Let $|S| \geq \left(\frac{r}{k}\right)^{\ell} \times k^{|\pi|}$ and $|\pi| \geq \ell > 0$. Then there exists a subset $\sigma \subseteq \pi$ of at most ℓ of elements such that if we* reveal *these, then each of the unrevealed elements is still highly unpredictable, namely $\forall i \in \pi - \sigma$, $UnPred_i(Proj(S, \pi - \sigma)) > r$.*

[1] This definition of *unpredictability* is one over the the definition of *predictability* defined in [EIRS]. They also give a second equivalent definition. Also our $Proj(v, \rho)$ is their $Proj(v, \pi - \rho)$.

Proof:

Initially, let $\sigma = \emptyset$. We will keep adding indices to σ, maintaining the property that

$$\frac{|Proj(S, \pi - \sigma)|}{k^{|\pi - \sigma|}} \geq \left(\frac{r}{k}\right)^{\ell - |\sigma|}.$$

Clearly this is true for $\sigma = \emptyset$, because $Proj(S, \pi) = S$. Now assume for $\sigma \subseteq \pi$ the property holds and that there is an index $i \in \pi - \sigma$ for which

$$\frac{|Proj(S, \pi - \sigma)|}{|Proj(S, \pi - \sigma \cup \{i\})|} = UnPred_i(Proj(S, \pi - \sigma)) \leq r.$$

It follows that

$$\frac{|Proj(S, \pi - \sigma \cup \{i\})|}{k^{|\pi - \sigma - \{i\}|}} \geq \frac{1}{r} \cdot \frac{|Proj(S, \pi - \sigma)|}{k^{|\pi - \sigma - \{i\}|}} = \frac{k}{r} \cdot \frac{|Proj(S, \pi - \sigma)|}{k^{|\pi - \sigma|}} \geq \frac{k}{r} \cdot \left(\frac{r}{k}\right)^{\ell - |\sigma|} = \left(\frac{r}{k}\right)^{\ell - |\sigma \cup \{i\}|}.$$

Thus the property holds for $\sigma \cup \{i\}$. Eventually, for all $i \in \pi - \sigma$, $UnPred_i(Proj(S, \pi - \sigma)) > r$. Since $Proj(S, \pi - \sigma) \subseteq [k]^{\pi - \sigma}$ and $r < k$, it follows that

$$1 \geq \frac{|Proj(S, \pi - \sigma)|}{k^{|\pi - \sigma|}} \geq \left(\frac{r}{k}\right)^{\ell - |\sigma|}$$

and thus $|\sigma| \leq l$. ■