Dependencies Between Random Variables Viewed as Entropy Areas

by Jeff Edmonds
York University

1 Entropy Areas

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.*

***X*, *Y*, and *Z*:** The random variables X, Y, and Z can be thought of in three different ways. Being able to comfortably switch between these adds deeper understanding of the concepts and more tools for being able to prove results.

Random Variables: X, Y, and Z are defined to be three random variables. What objects/events/values these takes on does not matter, only how the probability is distributed between them. For example, X could be an apple with probability $\frac{1}{3}$ and an orange with probability $\frac{2}{3}$.

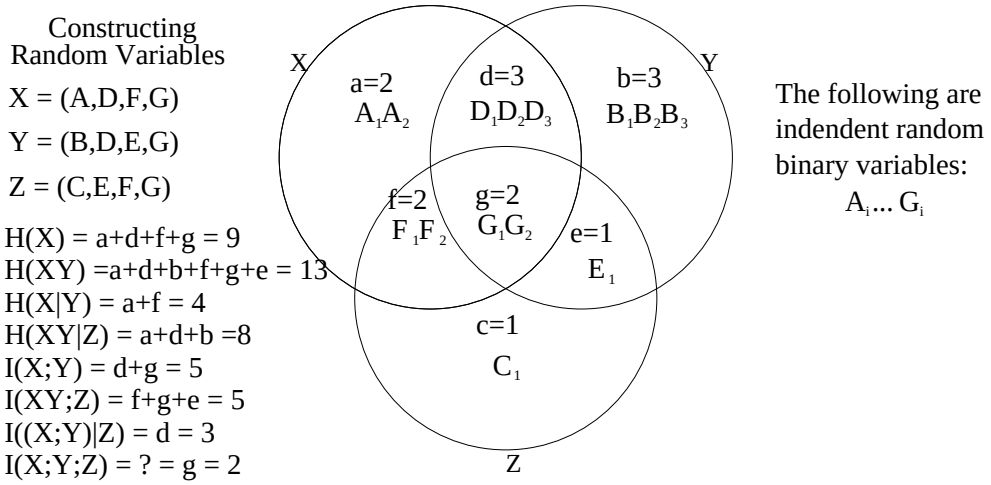
Information: A random variable adds a certain amount of uncertainty. Imagine that you are in the dark about which value X happens to take on after all the deciding coins have been flipped, however, your friend does. Your friend has more *information* than you do. An alternative way of viewing X is not as a random variable but as the information that your friend has that you do not have.

Entropy $H(X)$ and Pseudo Entropy: The *entropy of* X is denoted $H(X)$. It is the (expected) number of bits needed to specify the value of X. Suppose for example that $X \in \{0,1\}^H$ was a uniformly random H bit string. Clearly then the number of bits needed to specify X is $H(X) = H$. In this example, the probability that X takes on any particular string x is $\Pr[X = x] = 2^{-H}$. One solves for H as $-\log_2(\Pr[X = x])$. As such let us define *pseudo entropy* to be $H'(X) = -\log_2(\Pr[X = x]) = -\log_2(2^{-H}) = H$. This works great if as in this example $\Pr[X = x]$ is the same for each value x that X might take on. If this is not the case then, the formal definition of entropy is $H(X) = \text{Exp}_x - \log_2(\Pr[X = x])$.

One of the advantages of logarithms is that it turns multiplications into additions. If $X \in \{0,1\}^{H_1}$ and $Y \in \{0,1\}^{H_2}$ are uniformly independent random H_1 and H_2 bit strings, then $\Pr[X = x \ \& \ Y = y] = \Pr[X = x] \times \Pr[Y = y] = 2^{-H_1} \times 2^{-H_2} = 2^{-(H_1+H_2)}$. Correspondingly, the number of bits to communicate both X and Y is $H'(XY) = -\log_2(\Pr[X = x \ \& \ Y = y]) = -\log_2(2^{-(H_1+H_2)}) = H_1 + H_2$, which is the number of bits $H(X) = H_1$ needed to communicate X plus the number $H(Y) = H_2$ needed for Y.

Entropy Areas: Draw on the paper three intersecting circles and label them X, Y, and Z. Think of the circle X containing the information corresponding to knowing

what value X takes on. The area outside of the circle X correspond to information that has nothing to do with the information X. We say that this information is *independent* of X.



Constructing Random Variables from Entropy Areas:

Binary Variables: The simplest random variable are independent binary ones. Each takes on the value 0 or 1 with probability $\frac{1}{2}$. Let $A_i, B_i, \ldots, G_i$ denote such variables. Let $a = 2, b = 3, \ldots, g = 2$ denote the numbers of each of these types of variables.

Areas of Regions: Let the areas of the regions in the above figure be of these same values $a = 2, b = 3, \ldots, g = 2$.

Combining: These binary random variables can be combined. For example in the above figure, the random vector X is constructed as $X = \langle A_1, A_2, D_1, D_2, D_3, F_1, F_2, G_1, G_2 \rangle$, $Y = \langle B_1, B_2, B_3, D_1, D_2, D_3, E_1, G_1, G_2 \rangle$, and $Z = \langle C_1, E_1, F_1, F_2, G_1, G_2 \rangle$.

Joint and Conditional Entropy:

Entropy $H(X)$: The *entropy of* X is denoted $H(X)$. It is the (expected) number of bits needed to specify the value of X. In this example, this is clearly $a+d+f+g = 9$. This is represented by the area of the circle labeled X. Its pseudo entropy is defined as $H'(X) = -\log_2(\Pr[X = x])$.

Joint Entropy $H(XY)$: The number of bits needed to specify both X and Y is at most that needed to specify them separately if they share information. Here $H(XY) = a + d + b + f + g + e = 14$. This is represented by the area $X \cup Y$. The pseudo joint entropy is defined $H'(XY) = -\log_2(\Pr[X = x \ \& \ Y = y])$.

Conditional Entropy $H(X|Y)$: If you already know Y, then the number of bits needed to specify X may be less than that needed to specify X when you don't

know Y. Here $H(X|Y) = a+f = H(XY)-H(Y)$. When I know Y, I like to cover up the Y circle. Then $H(X|Y)$ is the remaining area of X, i.e. of $X-Y = X\cap\overline{Y}$. This is the area of $X \cup Y$ minus the area of Y.

Pseudo conditional entropy is defined $H'(X|Y) = -\log_2(\Pr[X = x \mid Y = y])$. Recall that the formal definition of *conditional probability* is $\Pr[X = x \mid Y = y]) = \Pr[X = x \ \& \ Y = y])/\Pr[Y = y]$. Note that this corresponds to our definition that $H(X|Y) = H(XY) - H(Y)$.

Mutual Information $I(X;Y)$: $I(X;Y)$ is the information that is common to both X and Y. Here $I(X;Y) = d+g = H(X) - H(X|Y) = H(X) + H(Y) - H(XY)$. This is because if you paint X and then paint Y, the left and right each get painted once and the middle twice. If you subtract off $H(XY)$, then the left and right are no longer painted and the middle is painted only once. This 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.

Recall that the formal definition of X and Y being *independent* is that $\Pr[X = x \ \& \ Y = y]) = \Pr[X = x] \times \Pr[Y = y]$. Equivalently $\Pr[X = x] \times \Pr[Y = y]/\Pr[X = x \ \& \ Y = y]) = 1$ or its negative logarithm is zero. This motivates defining the *pseudo mutual entropy* as $I'(X;Y) = -\log_2(\Pr[X = x \mid Y = y])$. Note that this corresponds to our definition that $I(X;Y) = H(X) + H(Y) - H(XY)$.

Conditional Mutual Information $I((X;Y)|Z)$: Similarly, $H(XY|Z) = a+d+b$ is the number of bits to learn X and Y if you already know Z. Cover up Z and then look at the remaining area of XY.

Joint Mutual Information $I(XY;Z)$: $I(XY;Z) = f+g+e$ is what you learn about X and Y from Z. It is the intersection of XY and Z.

SOURCE OF CONFUSION! $I(X;Y;Z)$: The area of $X\cap Y\cap Z$ does not really have a defined meaning using entropy, conditional entropy, and mutual entropy. We will denote this area by $I(X;Y;Z)$. It is the source of lots of confusion because its area can be negative!

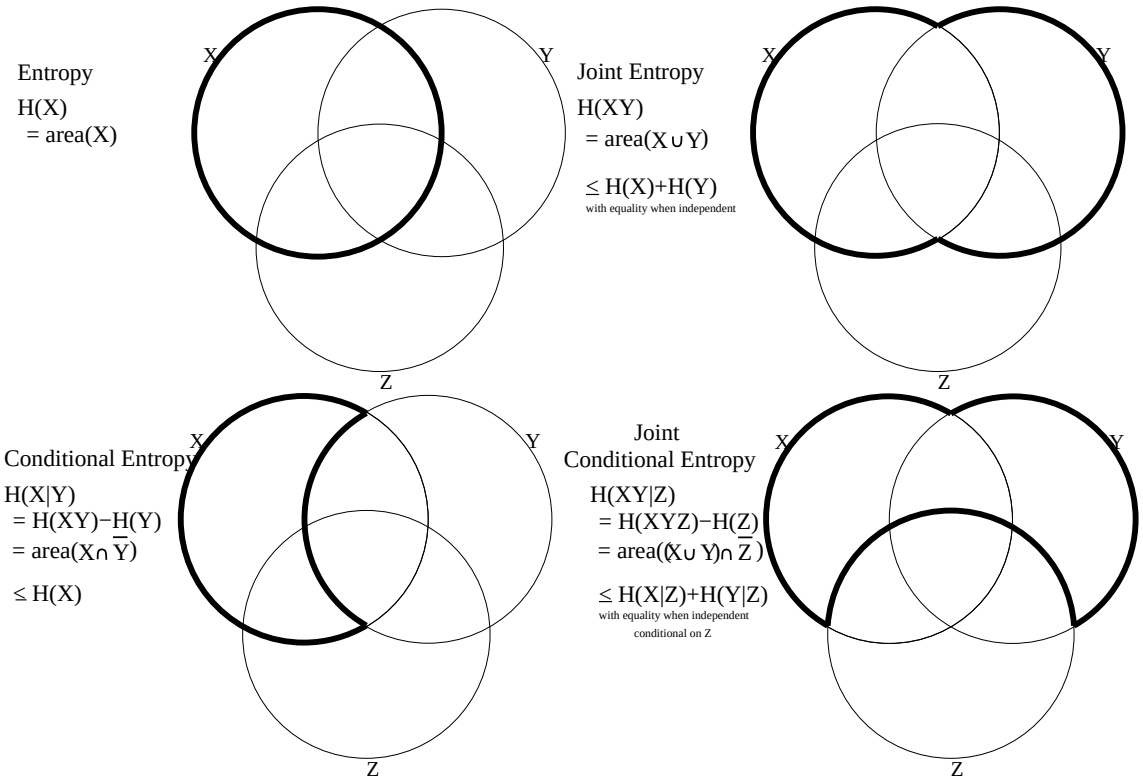
Solving for 2^n-1 Unknowns: If your goal is to construct random variables with some given entropy and dependence, then these entropy areas help a lot. One has to set up and solve equations for all the equality/inequities that you have. But what are the unknowns $c_1, c_2, c_3, \ldots$ that one is solving for. It turns out that with n random variables $X_1, \ldots, X_n$, there are 2^n-1 unknowns.

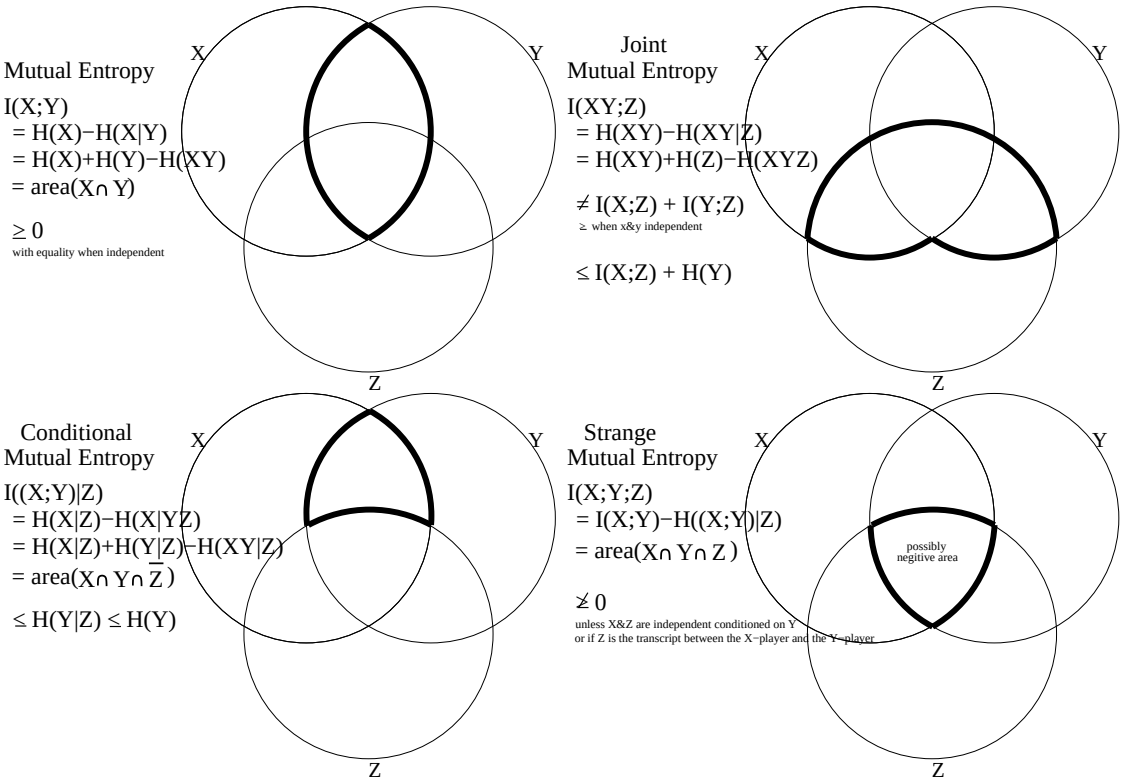
Primitives = Unions: One way to define the 2^n-1 unknowns is as follows. For each nonempty subset $S \subseteq [X_1, \ldots, X_n]$, define unknown c_S to be $H(S)$. This is the area of the *union* $\cup_{i\in S}X_i$. This value is straight forward to compute as $H'(S) = -\log_2(\Pr[\{X_i = x_i|\forall i \in S\}])$. Hence, we call it a *primitive* value. For example, $H(XY)$ represent the information consisting of knowing both the value of X and of Y. The problem is that these can be hard to work with.

Duals = Intersections: Now note that the three intersecting circles in the above figure partition the page into $2^3 = 8$ separate regions with areas denoted with $a = 2, b = 3, \ldots, g = 2$. The eighth is the area outside of the three circles. Lets call each of these the *dual primitives*. We specify a dual primitive, by specifying for each random variable whether we are considering the area inside or outside of its corresponding circle. For example $X \cap \overline{Y} \cap Z$, denotes the area that in X, in Z, and outside of Y. There are $2^n - 1$ of these because for each nonempty subset $S \subseteq [X_1, \ldots, X_n]$, define unknown c'_S to be the area of the *intersection* $\cup_i (X_i \text{ if } i \in S \text{ else } \overline{X_i})$. Given the values of each of these unknowns c'_S, one can easily compute all the entropy, joint entropies, and mutual entropies as done above. Similarly, if you know the area of each primitive, one can compute the area of each dual primitive, because there are $2^n - 1$ linear equations and $2^n - 1$ unknowns. We will prove that with the three variables X, Y, and Z, all the dual primitive areas $\langle a, b, \ldots, g \rangle$ are non-negative, except the intersection g of all three.

NOT Venn Diagrams: Be very careful not to confuse these circles with Venn diagrams. In a Venn diagram, X, Y, and Z must be random variables with taking on either *true* or *false*. Inside the circle X represent all results of the coin flips that leads to X being true and outside represent all those for which X is false. The area of the circle is the probability that X is true.

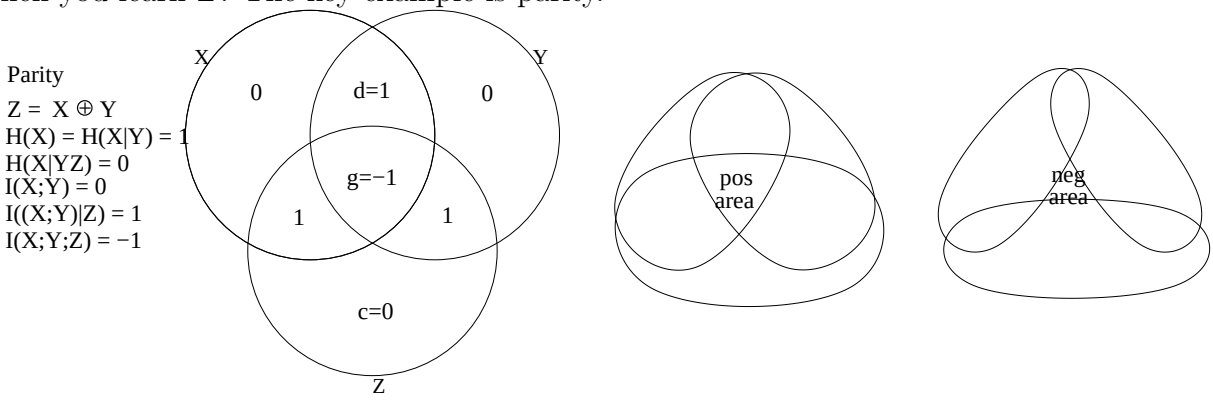
2 Definitions and Lemmas via Figures





3 Negative $I(X;Y;Z)$

Entropy is great for many applications. But there is one place where the intuition breaks. This occurs because the strange area $I(X;Y;Z) = Area(X \cap Y \cap Z)$ can be negative. It occurs when $I((X;Y)|Z) \geq I(X;Y)$, i.e. X and Y become more dependent on each other when you learn Z. The key example is parity.



$I(X;Y;Z) = Area(X \cap Y \cap Z) = g$**:** The dual primitive $X \cap Y \cap Z$ does not have a direct defined meaning using entropy, conditional entropy, and mutual entropy. We have seen that $I(X;Y) = Area(X \cap Y)$ is the information that is mutual between X and Y, hence for ease of notation, let us denote the area of $X \cap Y \cap Z$ by $I(X;Y;Z)$. For what ever it means this would be the information that is "mutual between all three."

$\boldsymbol{I(X;Y;Z) = I(X;Y) - I((X;Y)|Z) \not\geq 0}$**:** $I(X;Y)$ measures the mutual information between X and Y or how dependent they are on each other. It is how much you learn about one from the other. In the figure, $I(X;Y) = d + g$ is the area of the intersection $X \cap Y$.

Similarly, $I((X;Y)|Z) = d$ is the same after you learn Z. Recall it is the intersection $X \cap Y$ remaining after covering up Z.

It follows that $I(X;Y;Z) = g = (d+g) - d = I(X;Y) - I((X;Y)|Z)$. Clearly, it is negative when $I((X;Y)|Z) \geq I(X;Y)$, i.e. there X and Y become more dependent on each other when you learn Z.

Parity: The key example making this middle area negative is parity. This example often comes up as a counter example to things. Let X and Y be independent boolean random variables. Let $Z = X \oplus Y$. Note that when you don't know Z, then X and Y are independent, so $I(X;Y) = H(X) = H(Z) = 1$, but if you know it, then they are completely dependent, namely when $Z = 0$, then $X = Y$ and when $Z = 1$, $X = \overline{Y}$, giving $I((X;Y)|Z) = d = 1$. Hence, $I(X;Y) - I((X;Y) = 0 - 1 = -1$. Because you can rearrange the equation in all ways $X = Y \oplus Z$, it follows the situation between X, Y, and Z is symmetrical.

Lets examine this situation more. I tell you both X and Y, then Z is determined. We write this as $H(Z|XY) = 0$. Recall this says that if you cover up both X and Y then whats remaining of Z, namely $c = 0$.

We still need that $H(X) = 1$ because X is determined by one bit. According the picture $H(X) = 0 + 1 + 1 - 1 = 1$ as needed.

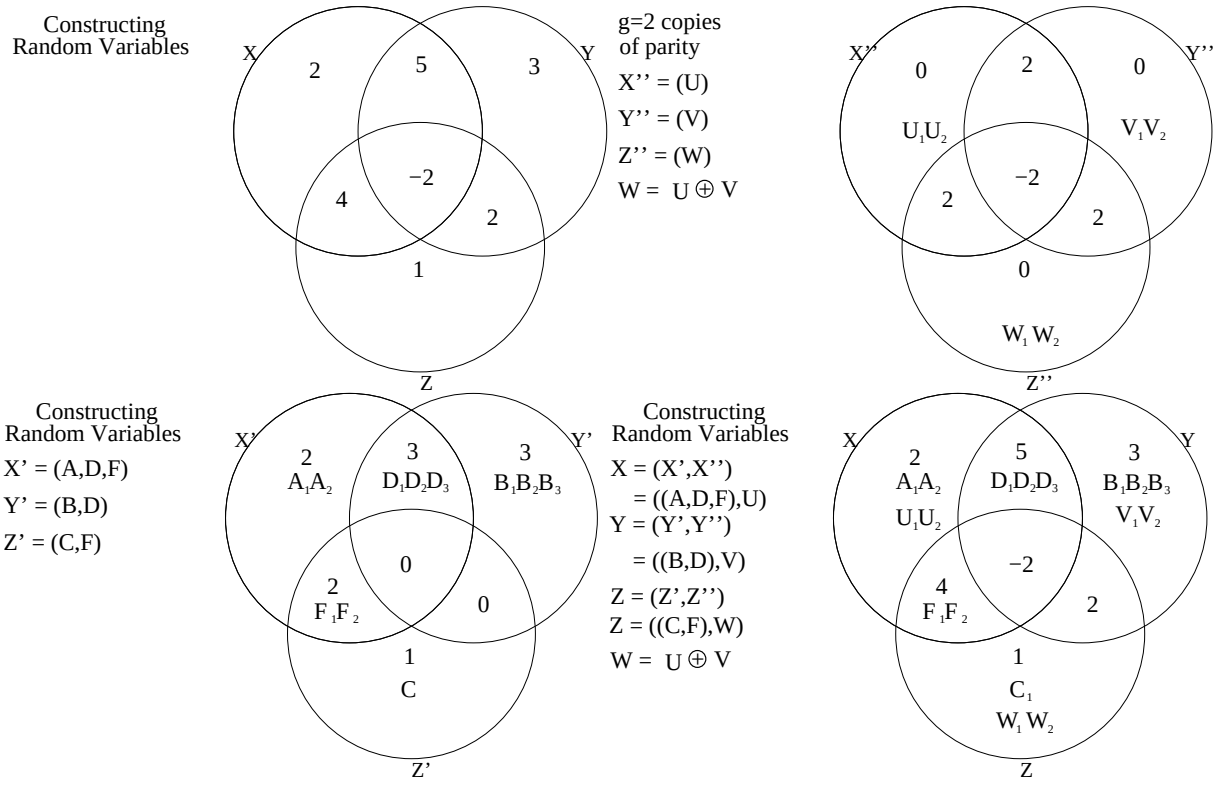
Marriage: Here is another set of random variables that have negative area and maybe are more practical. Say Z denotes whether or not X and Y are married. This is like half of parity. When you learn $Z = 1$, you learn that X and Y are completely dependent, namely $X = Y$, giving $I((X;Y)|Z = 1) > I(X;Y)$. On the other hand, when you learn $Z = 0$, you learn that X and Y are completely independent giving $I((X;Y)|Z = 0) < I(X;Y)$. It is not clear which of these effects wins out for a random Z.

Crunching the numbers gives $I(X;Y;Z) = I(X;Y) - I((X;Y)|Z) = 0.19 - 0.5 = -0.31$. In fact, we tried a number of scenarios where if $Z = 0$, then X and Y are a little more independent then when $Z = 1$. In all the cases we tried this area was negative.

Note that this set of random variables is the only one discussed in this paper which can't be decomposed into independent boolean variables.

Negative Area: The figure shows how one might interpret negative area. If three circles over lap, then their intersection should be positive. But what happens if they form a circle, with X overlapping with Y, Y with Z and Z with X, and with space in the middle of the circle. Might the area of the middle $X \cap Y \cap Z$ be considered negative?

Possible Random Variables and Areas of Primitives: In trying to prove lemmas about entropy, one needs to know the range of possible interactions between random variables. Recall that the n circles corresponding to n random variables partition the paper into $2^n - 1$ dual primitives like $X \cap \overline{Y} \cap Z$. Suppose you are told the areas $\langle a, b, \ldots, g \rangle$ of each of these dual primitives. One might ask, for which such values can one construct random variables such that the entropy, conditional entropy, and mutual entropy are all given by these areas. We have answered this question given only three random variables.



Requirements on Dual Primitive Areas: The long list of lemmas below will prove that the areas $\langle a, b, \ldots, g \rangle$ of each dual primitive must be nonnegative, with the exception of the area g of $X \cap Y \cap Z$ and that for each pair of random variables the mutual information $I(X;Y) = X \cap Y$ must also be nonnegative. These are the only requirement for constructing random variables X, Y, and Z with these values.

Proof: Above we showed how the construction $X = \langle A, D, F, G \rangle$, $Y = \langle B, D, E, G \rangle$, and $Z = \langle C, E, F, G \rangle$ solves the problem when all of the require areas $\langle a, b, \ldots, g \rangle$ are integers and $I(X;Y;Z) = Area(X \cap Y \cap Z) = g \geq 0$. To deal with arbitrary reals, have the binary variables be unfair. Now suppose that $I(X;Y;Z) = g$ for some negative integer g. See the first of the above four figures, in which $g = -2$. The only way we know to make $I(X;Y;Z)$ negative is with parity and if we copy parity $-g$ times, then $I(X'';Y'';Z'')$ becomes g. See the upper right figure above.

One thing to note is that the areas of all the entropy measure is additive when you concatenate independent variables. More formally, suppose the areas of the dual primitives for $\langle X', Y', Z' \rangle$ are $\langle a', b', \ldots, g' \rangle$ and that those for $\langle X'', Y'', Z'' \rangle$ are $\langle a'', b'', \ldots, g'' \rangle$, where these two lots are. Now define $X = \langle X', X'' \rangle$, $Y = \langle Y', Y'' \rangle$, $Z = \langle Z', Z'' \rangle$. It follows that the areas of the dual primitives for $\langle X, Y, Z \rangle$ are $\langle a, b, \ldots, g \rangle = \langle a' + a'', b' + b'', \ldots, g' + g'' \rangle$.

Here we know the $\langle a, b, \ldots, g \rangle$ that we want and we know the $\langle a'', b'', \ldots, g'' \rangle$ that we got with $-g$ copies of parity. We reverse engineer by calculating $\langle a', b', \ldots, g' \rangle = \langle a - a'', b - b'', \ldots, g - g'' \rangle$. These are the values in the lower left figure. Note that for $-g$ copies of parity $g'' = g$ and $d'' = -g$. Hence, $g' = g - g'' = 0$. Also we know that $I(X;Y) = d + g \geq 0$, giving that $d \geq -g$, which gives that $d' = d - d'' \geq 0$. Hence, all our new areas are positive. Hence we can construct $\langle X', Y', Z' \rangle$ as we did before. Again see the lower left figure.

We conclude by defining $X = \langle X', X'' \rangle$, $Y = \langle Y', Y'' \rangle$, and $Z = \langle Z', Z'' \rangle$ and we are done. See the lower right figure.

Boolean Case Unknown: One thing that the authors do not know how to do is to construct X, Y, and Z from the dual primitive areas $\langle a, b, \ldots, g \rangle$ when X, Y, and Z are required to each be boolean.