

Mutual Information Extraction

Master thesis

Ori Mosenzon

September 12, 2004

Contents

1	abstract	3
2	Preface	4
3	Encapsulating dependency	5
3.1	Introduction	5
3.2	Formalization	5
3.3	The need for approximation	8
3.4	The optimization problem	9
4	A framework for dealing with mutual relation of random variable	10
4.1	I-measure	12
4.2	Usage examples	16
4.2.1	The notion of $\Lambda(X; Y; T)$	16
4.2.2	Three variables Markov chain	17
4.3	re-formalization of the original problem	17
4.4	Special cases of the optimization problem	18
4.5	historical notes	20
5	Coping with the general problem	21
6	The deterministic variant	25
6.1	Decomposition of entropy and mutual information	25
6.2	A greedy agglomerative algorithm	26
6.3	Finding the best deterministic extractor is hard	28
7	Deterministic function of two variables	30
7.1	Motivation	30
7.2	A characterization of a two-variable-extractor	31
7.3	A Cartesian multiplication extractor	31
7.4	The partition lattice	32
7.5	The GCP	33
7.6	A search algorithm for a good partition	33

8 Discussion

37

A

39

Chapter 1

abstract

Consider two random variables, X and Y . The mutual relation between the variables can vary between complete independence to complete dependency, when one variable is a deterministic function of the other. The measure of mutual information $I(X;Y)$ quantifies the amount of dependency between X and Y , but states nothing about its nature.

In this work we try to capture this dependency by using a new random variable that we call 'an extractor'. A perfect extractor is a variable that contains all the information X gives on Y and no other information.

It turns out that in the general case, there exist no perfect extractor, so the best we can do is to look for a good approximation. We develop a general framework to deal with problems of information-relations among several variables. Using this framework, we approach the problem from several different directions.

Chapter 2

Preface

This work began as an attempt to understand dependency between random variables using concepts of Information Theory. It was very much motivated by the work of Tishby et al “The Information Bottleneck Method” (see [7]). Although the motivation for this research was similar to the motivation of the Bottleneck method, a different approach was taken.

After the research evolved, much effort was put, in order to build a general framework to deal with problems of this type. The main tool of this framework was a correspondence between sets with an additive function and information measures on random variables. After the framework was constructed, we learned about the work made by Yeung ([8],[9] and [5]). Since Yeung developed the same framework and took it much further than we did, our work could not be considered original.

Several different paths were taken in order to cope with the general problem that is presented in the first chapter. None of these paths reached a full maturity. The result is various attempts to solve variants of the general problem in addition to the general framework and the problem formalization.

Chapter 3

Encapsulating dependency

3.1 Introduction

The problem of understanding dependencies among empirical results is one of the most fundamental aspects of any experimental research. When trying to understand phenomena in nature, we frequently sense that there is a mutual influence between two different entities we observe, but it is often hard to point out what is the nature of this connection.

As an example we can think of a medical researcher that investigates heart faults. The researcher might notice that there is a correlation between heart attacks and the use of a certain medicine. It is much easier to detect the correlation than it is to understand why the two factors are correlated. Imagine that the researcher deduces that the reason that the medicine causes heart attacks is that it binds to a certain protein and deactivate it. By that deduction, the researcher made a big step forward, she now *understands* the correlation she observed before. This understanding can lead to practical solution for the problem and can help predicting observations in the future. The main question that we will ask, is whether it is possible to get a better understanding of a correlation, just by analyzing the statistics of various entities of the domain. Back to our example, we will ask whether it is possible to get some understanding of the correlation between heart attack and the medicine, just by looking at joint frequency of various factors of the heart and the medicine.

To formalize this question mathematically, we will think of the two correlated entities as random variables. We will look for another random variable that encapsulates the correlation or at least some of it.

3.2 Formalization

Let X and Y be two discrete random variables, and let $P(X, Y)$ be their joint probability function. The joint distribution can be represented as a $|Y| \times |X|$ matrix of nonnegative values that sums to 1.

The joint probability matrix contains all the statistical information about the two variables. By summing entries and normalizations one can compute the marginals and the conditional distributions. It is important to point out that the marginal probabilities $P(X)$, $P(Y)$ do not determine $P(X, Y)$ but rather set constraints and leave a certain degree of freedom. Fixing the marginal probabilities enable to construct joint probabilities that ranges from independence of the variable to a high correlation among them.

Information Theory supplies a measure of the quantity of this dependency. The mutual information, denoted as $I(X, Y)$, quantifies how much information (in bits) each variables tells about the other.

We can think of $I(X, Y)$ as the amount of uncertainty which is reduced from the uncertainty about Y by the knowing of X (and vice versa). The formal definition of $I(X; Y)$ is as follows:

$$I(X; Y) = \sum_{x,y} P(x,y) \log \left(\frac{P(x,y)}{P(x)P(y)} \right)$$

Mutual information has some nice properties that are consistence with our intuition of “quantity of relationship” between two lotteries:

- $I(X, X) = H(X)$.
- $I(X, Y) = I(Y, X)$.
- $I(X, Y) \geq 0$ with equality if and only if X is independent of Y (i.e. $P(X, Y) = P(X)P(Y)$)

Where the $H(X)$ is the entropy of X which is defined as:

$$H(X) = - \sum_x P(x) \log(P(x))$$

Although we have a good *quantifier* for the dependency between the variables, we don't yet have a *description* of that dependency. We would like to know *what* each variable tells about the other, not only *how much* it tells.

One can claim that by knowing the value of X , the information we obtain is the conditional distribution of Y , i.e., $P(Y|x)$. Figure 3.1 illustrates this claim. Although this claim is correct, we would like to capture the meaning of the mutual information in a different way. We will seek for a new random variable T , that encapsulates the information X and Y tell about each other. Encapsulating the dependency by a random variable will enable handling the mutual information in the same way we handle the information of the original correlated entities.

The demand that T will contain *all* the information X and Y tell about each other, has a natural formal interpretation in terms of conditional independence. We will say that T contains all the mutual information among X and Y if $I(X; Y|T) = 0$ or equivalently, if the Markovian condition $X - T - Y$, holds.

It is not hard to find such T . For example, choosing T to be X , Y or the joint variable - (X, Y) will do. In spite of the fact that these choices indeed

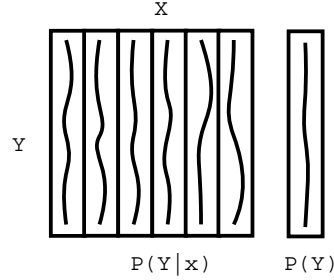


Figure 3.1: The mutual information can be seen as the average reduction of the entropy of Y which is caused by knowing the value of X . $I(X; Y) = H(P(Y)) - \sum_x P(x)H(P(Y|x))$

contain all the mutual information, we will not think of them as encapsulating the information. This is due to the fact that they contain additional information which is not relevant to the dependency between X and Y . If we choose $T = X$ we will have $I(X; Y|T) = 0$ but T contain more information: $H(T) = H(X) = H(X|Y) + I(X; Y) > I(X; Y)$ (assuming X is not a deterministic function of Y).

Thus, a random variable T that *exactly* encapsulates the mutual information between X and Y should satisfy:

- $I(X; Y|T) = 0$.
- $H(T) = I(X; Y)$.

It is easy to see (we will show that shortly) that in the general case, there is no T that encapsulates the mutual information. A perfect encapsulator exists only when the joint distribution $P(X, Y)$ is of a very special form. This leads us to look for an approximation: A random variable that contains as much relevant information as possible and as little irrelevant information, as possible.

Succeeding in finding a variable that encapsulates most of the dependency between two random variable can be used for compression and for understanding the domain. The idea of encapsulating mutual dependency by a new random variable was introduced by Tishby, Pereira and Bialek [7]. Their idea was to extract the mutual dependency by a new variable which is a stochastic function of X that preserves information about Y while losing information on X . The algorithm to find such variable is called “The information bottleneck method”. Although the motivation of this research is similar, we will try to handle it in a different way. We define a general optimization problem that represents the seek for a good extractor - T . The information bottleneck problem is a special case

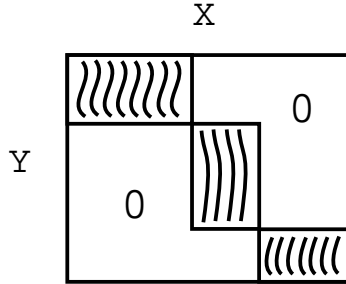


Figure 3.2: The joint distribution of two variables that have a perfect extractor of their mutual information. Each value of the extractor corresponds to a block of an independent distribution.

of that problem. We will try to deal with other special cases as well as with the general problem. Our hope is that this research will lead to new algorithms for extraction of mutual information and for a better understanding of the concept of dependency between random variables.

3.3 The need for approximation

As mentioned in the previous section, a variable T is said to encapsulate the mutual information of X and Y if: $I(X; Y|T) = 0$ and $H(T) = I(X; Y)$. The first condition represents the fact that T contains all the mutual information and the second condition represents the fact that it doesn't contain any additional information.

Suppose that T satisfies those two conditions. From the fact that T contains all the mutual information we have:

$$H(T) \geq I(X; T) \geq I(X; Y) \tag{3.1}$$

The first inequality is true for any two random variables T and X . The second is the data processing inequality (see-4.23). The demand that $H(T) = I(X; Y)$, implies that the inequality 3.1 is actually an equality, thus:

$$H(T) = I(T; X) = H(T) - H(T|X)$$

This means that $H(T|X) = 0$ and that T is a deterministic function of X . Using the same arguments on Y we obtain that T is also a deterministic function of Y . Thus, if T encapsulates the mutual information, there exist two functions g and h such that $T = g(X)$ and $T = h(Y)$. If we group the values of X according to the values of g and group the values of Y according to the values of h , the joint distribution $P(X, Y)$ must have the same structure as in Figure 3.2

Thus, a perfect extractor exists only when X and Y distribute in a very special way. In the general case, the best we can do is to look for an approximation of a perfect extractor.

3.4 The optimization problem

As we have seen, we must compromise on the original demands on T . Instead of insisting that T will contain all the mutual information, we wish to capture most of it. Similarly, instead of insisting that T will contain no irrelevant information, we will try to minimize it. This brings us to the following formalization of an optimization problem:

Choose the cardinality of T and the conditional probability values $P(t|x, y)$ in order to minimize the following expression:

$$\omega_1 I(X; Y|T) + \omega_2 H(T) \tag{3.2}$$

subject to the constrains:

$$\forall x, y : \sum_t P(t|x, y) = 1 \tag{3.3}$$

$$\forall x, y, t : P(t|x, y) \geq 0 \tag{3.4}$$

Where ω_1 and ω_2 represents the trade-off between the importance of capturing the mutual information and the importance of being compact. In the next chapter we will examine a genral framework to deal with mutual relations of random variables. This general tool will give some additional insights and will enable to refine and to generalize this optimization problem.

Chapter 4

A framework for dealing with mutual relation of random variable

Since our optimization problem deals with mutual relations among (three) random variables, we wish to gain more understanding about the domain of variables interrelations. We want to know what are the possible information measures that a set of variable can have, what are the implications from knowing a certain fact on the variables, and so forth. For example, the fact that two variables are independent conditioning on a third variable, implies that the entropy of the third variable is greater or equal than the mutual information of the two variables. We will try to construct a general framework in which a fact like this, will have a clear and intuitive meaning.

The main tool we are going to use is an interesting correspondence between mutual relations among random variables and mutual relations among sets. The main idea could be illustrated by an example. Let X_1 and X_2 be two arbitrary random variables. It is known that the following equations hold:

$$H(X_1, X_2) = H(X_1|X_2) + H(X_2) = H(X_2|X_1) + H(X_1) \quad (4.1)$$

$$I(X_1; X_2) = H(X_1) + H(X_2) - H(X_1, X_2) \quad (4.2)$$

$$= H(X_1) - H(X_1|X_2) \quad (4.3)$$

$$= H(X_2) - H(X_2|X_1) \quad (4.4)$$

These equations can be summarized in an intuitive way by a Venn-diagram like the one that is illustrated on the upper part of Figure 4.1. This illustration appears in [1] but without any additional formalization.

Now, assume that \widetilde{X}_1 and \widetilde{X}_2 are two sets and μ is an additive function on the σ -field generated by $\{\widetilde{X}_1, \widetilde{X}_2\}$. Assume also that $\mu(\widetilde{X}_1) = H(X_1)$, $\mu(\widetilde{X}_2) =$

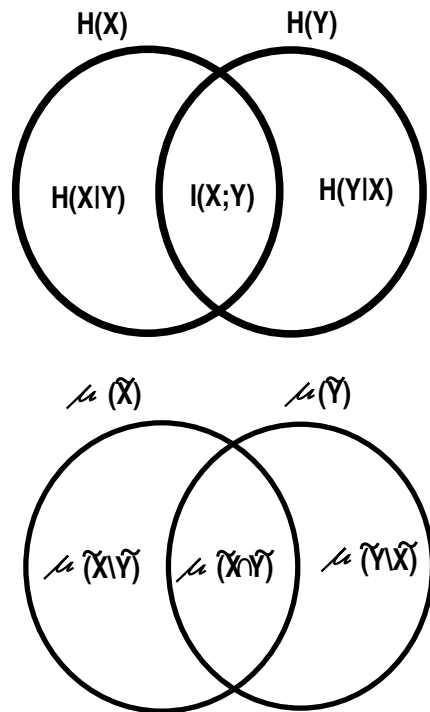


Figure 4.1: Venn diagrams that represents mutual relations between random variables (upper) and between sets (lower).

$H(X_2), \mu(\widetilde{X}_1 \cup \widetilde{X}_2) = H(X_1, X_2), \mu(\widetilde{X}_1 \setminus \widetilde{X}_2) = H(X_1|X_2), \mu(\widetilde{X}_2 \setminus \widetilde{X}_1) = H(X_2|X_1)$ and $\mu(\widetilde{X}_1 \cap \widetilde{X}_2) = I(X_1; X_2)$. At this point it is not clear that such sets indeed exist, but we will prove that shortly. The following corresponding equations are easily verified by examining lower part of Figure 4.1:

$$\mu(\widetilde{X}_1 \cup \widetilde{X}_2) = \mu(\widetilde{X}_1 \setminus \widetilde{X}_2) + \mu(\widetilde{X}_2) = \mu(\widetilde{X}_2 \setminus \widetilde{X}_1) + \mu(\widetilde{X}_1) \quad (4.5)$$

$$\mu(\widetilde{X}_1 \cap \widetilde{X}_2) = \mu(\widetilde{X}_1) + \mu(\widetilde{X}_2) - \mu(\widetilde{X}_1 \cup \widetilde{X}_2) \quad (4.6)$$

$$= \mu(\widetilde{X}_1) - \mu(\widetilde{X}_1 \setminus \widetilde{X}_2) \quad (4.7)$$

$$= \mu(\widetilde{X}_2) - \mu(\widetilde{X}_2 \setminus \widetilde{X}_1) \quad (4.8)$$

Note that the entropy-equations 4.1-4.4 correspond to the set-equations 4.5-4.8.

It turns out that this correspondence can be generalized to any natural number of variables/sets. The set-structure of the information quantities, gives insight and tools to deal with inference and optimization problems that consists of interrelated random variables. In the simple example of two variables, this set-structure does not give additional insights to the well known basic equalities. After we prove the general correspondence, we will give several examples of how this correspondence could be beneficial in the case of three variables.

4.1 I-measure

Theorem 1 (information-set-correspondence) *for any set of random variable X_1, \dots, X_n , there exists an analogical set of sets $\widetilde{X}_1, \dots, \widetilde{X}_n$ and a signed measure function μ on the σ -field generated by those sets, such that:*

1. $H(X_i) = \mu(\widetilde{X}_i)$
2. $H(X_i|X_j) = \mu(\widetilde{X}_i \setminus \widetilde{X}_j)$
3. $I(X_i; X_j) = \mu(\widetilde{X}_i \cap \widetilde{X}_j)$
4. $I(X_i; X_j|X_k) = \mu(\widetilde{X}_i \cap \widetilde{X}_j \setminus \widetilde{X}_k)$
5. *Any of the above equation will hold also for sets of variables. The correspondence sets will be unions of the analogical sets. For example, for any $\sigma_1, \sigma_2 \in 1, \dots, n$ equation 2 will take the form:*

$$H(\{X_i|i \in \sigma_1\}|\{X_i|i \in \sigma_2\}) = \mu \left((\cup_{i \in \sigma_1} \widetilde{X}_i) \setminus (\cup_{i \in \sigma_2} \widetilde{X}_i) \right)$$

proof: The formal proof contains some technical details that might obscure the global picture. Before we give this proof, we will try to illustrate the general idea in a more intuitive manner.

The proof is based on the fact that information quantities can be decomposed into sum (or subtraction) of joint entropies in the same manner that a value of an

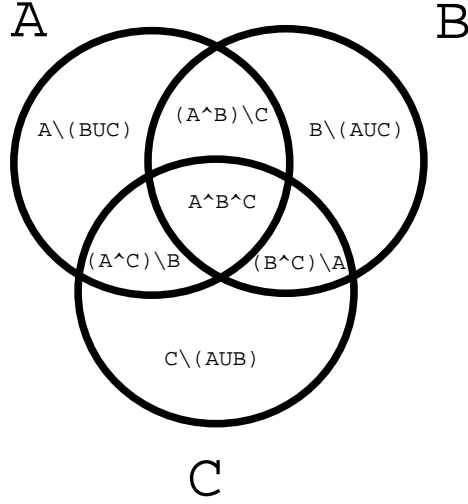


Figure 4.2: The atoms of three sets A, B and C .

additive function on set-expression can be decomposed into sum (or subtraction) of union expressions. For example, the information quantity $H(X|Y)$ can be decomposed into joint entropies in the following way:

$$H(X|Y) = H(X, Y) - H(Y)$$

The value $\mu(X \setminus Y)$ can be decomposed into union expressions, in a similar way:

$$\mu(X \setminus Y) = \mu(X \cup Y) - \mu(Y)$$

The proof describes how to construct the sets $\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_n$ and an additive function μ that meet the requirements of the theorem. The sets could be any sets that have a full algebra i.e. an algebra in which any expression of sets is not the empty set.

The additive set is defined in the following way:

$$\mu(\cup_{i \in \sigma} \tilde{X}_i) \equiv H(\{X_i | i \in \sigma\})$$

A priori, it is not clear why an additive function can take those values and why those values fully characterize it. In order to show why this is indeed true, we define an additive function by setting its values on the atoms. The atoms are $2^n - 1$ disjoint sets of which any set-expression can be describe as a union of them. Figure 4.2 illustrates the atoms of three sets A, B and C .

It is clear that an additive function can take any value on an atom and that the values on the atoms fully characterize the additive function.

The next step is to express the function values on union expression in terms of values on atom and vise versa. We show that there exist a regular linear

transformation from the atom values into the union values. This part of the proof contains the technical details that are somewhat hard to follow. The existence of the regular linear transformation implies that the union expressions can take any value and that they, like the values on atoms, fully characterize the additive function.

The similar decomposition and the similar values on the union expression are sufficient to show that any two correspondence expressions has the same value. For example:

$$H(X|Y) = H(X, Y) - H(Y) = \mu(X \cup Y) - \mu(Y) = \mu(X \setminus Y)$$

We will now give the formal proof.

Given an arbitrary set of random variable X_1, \dots, X_n it is sufficient to construct $\tilde{X}_1, \dots, \tilde{X}_n$ and a function μ such that for any $\sigma_1, \sigma_2, \sigma_3 \in \{1, \dots, n\}$:

$$\begin{aligned} I(\{X_i|i \in \sigma_1\}; \{X_i|i \in \sigma_2\}|\{X_i|i \in \sigma_3\}) &= \\ \mu\left(\left(\bigcup_{i \in \sigma_1} \tilde{X}_i\right) \cap \left(\bigcup_{i \in \sigma_2} \tilde{X}_i\right) \setminus \bigcup_{i \in \sigma_3} \tilde{X}_i\right) & \quad (4.9) \end{aligned}$$

If we choose $\sigma_1 = \sigma_2 = \{i\}$ and $\sigma_3 = \phi$, we obtain equation 1. If we choose $\sigma_1 = \sigma_2 = \{i\}$ and $\sigma_3 = \{j\}$, we obtain equation 2. If we choose $\sigma_1 = \{i\}$, $\sigma_2 = \{j\}$ and $\sigma_3 = \phi$, we obtain equation 3.

To obtain the rest of the equations, we can make similar choices of non singleton sets.

We will now describe the construction process. Let $\tilde{X}_1, \dots, \tilde{X}_n$ be sets such that for any $\phi \neq \sigma \subset \{1, \dots, n\}$:

$$\bigcap_{i \in \sigma} \tilde{X}_i \setminus \bigcup_{i \in \bar{\sigma}} \tilde{X}_i \neq \phi$$

Where $\bar{\sigma} = \{1, \dots, n\} \setminus \sigma$. One way to construct such sets is to define $\tilde{X}_i = \{\sigma | \sigma \subset \{1, \dots, n\}, i \in \sigma\}$. Now,

$$\bigcap_{i \in \sigma} \tilde{X}_i \setminus \bigcup_{i \in \bar{\sigma}} \tilde{X}_i = \{\sigma\} \neq \phi$$

We shall call the sets of the form $\bigcap_{i \in \sigma} \tilde{X}_i \setminus \bigcup_{i \in \bar{\sigma}} \tilde{X}_i$ - *atoms*, because any set-expression of $\tilde{X}_1, \dots, \tilde{X}_n$ can be described as a disjoint union of atoms.

Because of the fact that μ is an additive function, the values of μ on the atoms fully characterize it. Hence, we need to define $2^n - 1$ values, the values of $\mu(\bigcap_{i \in \sigma} \tilde{X}_i \setminus \bigcup_{i \in \bar{\sigma}} \tilde{X}_i)$ in order to construct μ .

We will now show that by defining alternative $2^n - 1$ values, the values of $\mu(\bigcup_{i \in \sigma} \tilde{X}_i)$, one can obtain an alternative characterization μ .

First we will describe $\mu(\bigcup_{i \in \sigma} \tilde{X}_i)$ as a linear combination of measures of atoms:

$$\begin{aligned} \mu(\bigcup_{i \in \sigma} \tilde{X}_i) &= \mu\left(\bigcup_{\psi \cap \sigma \neq \phi} (\bigcap_{i \in \psi} \tilde{X}_i \setminus \bigcup_{i \notin \psi} \tilde{X}_i)\right) \\ &= \sum_{\psi \cap \sigma \neq \phi} \mu(\bigcap_{i \in \psi} \tilde{X}_i \setminus \bigcup_{i \notin \psi} \tilde{X}_i) \end{aligned}$$

Second, we describe $\mu(\cap_{i \in \sigma} \tilde{X}_i \setminus \cup_{i \in \bar{\sigma}} \tilde{X}_i)$ as a linear combination of terms of the form $\mu(\cup_{i \in \sigma} \tilde{X}_i)$:

$$\begin{aligned} \mu(\cap_{i \in \sigma} \tilde{X}_i \setminus \cup_{i \in \bar{\sigma}} \tilde{X}_i) &= \\ \sum_{\phi \neq s \subset \sigma} (-1)^{|\phi|+1} \mu(\cup_{i \in \phi} \tilde{X}_i \setminus \cup_{i \in \bar{\sigma}} \tilde{X}_i) &= \end{aligned} \quad (4.10)$$

$$\sum_{\phi \neq s \subset \sigma} (-1)^{|\phi|+1} \left(\mu(\cup_{i \in \phi} \tilde{X}_i \cup_{i \in \bar{\sigma}} \tilde{X}_i) - \mu(\cup_{i \in \bar{\sigma}} \tilde{X}_i) \right) = \quad (4.11)$$

$$\begin{aligned} \left(\sum_{\phi \neq s \subset \sigma} (-1)^{|\phi|+1} \mu(\cup_{i \in \phi \cup \bar{\sigma}} \tilde{X}_i) \right) - \left(\mu(\cup_{i \in \bar{\sigma}} \tilde{X}_i) \sum_{\phi \neq s \subset \sigma} (-1)^{|\phi|+1} \right) &= \\ \sum_{\phi \neq s \subset \sigma} (-1)^{|\phi|+1} \mu(\cup_{i \in \phi \cup \bar{\sigma}} \tilde{X}_i) - \mu(\cup_{i \in \bar{\sigma}} \tilde{X}_i) & \end{aligned} \quad (4.12)$$

Equation 4.10 is a version of the inclusion/exclusion formula. See proposition 4 at the appendix for a proof. Equation 4.11 was obtained using the following equation:

$$\mu(A \cup B) - \mu(B) = \mu(A \setminus B)$$

which is true for any sets A and B (see proposition 6 at the appendix). Equation 4.12 is true because:

$$\sum_{\phi \neq s \subset \sigma} (-1)^{|\phi|+1} = 1$$

(See proposition 5 at the appendix for a proof)

We have seen that there exists a regular linear transformation that translates $2^n - 1$ terms of the form $\mu(\cup_{i \in \sigma} \tilde{X}_i)$ into $2^n - 1$ terms of the form $\mu(\cap_{i \in \sigma} \tilde{X}_i \setminus \cup_{i \in \bar{\sigma}} \tilde{X}_i)$. Thus, by assigning arbitrary values to the former terms, we determine a legitimate measure function on the sets.

Now, let us assign:

$$\mu(\cup_{i \in \sigma} \tilde{X}_i) \equiv H(\{X_i | i \in \sigma\})$$

which concludes the construction.

It is remained to show that under this construction, equations of the form

4.9 indeed hold.

$$\begin{aligned}
I(\{X_i|i \in \sigma_1\}; \{X_i|i \in \sigma_2\}|\{X_i|i \in \sigma_3\}) &= H(\{X_i|i \in \sigma_1\}|\{X_i|i \in \sigma_3\}) \\
&\quad - H(\{X_i|i \in \sigma_1\}|\{X_i|i \in \sigma_2 \cup \sigma_3\}) \\
&= H(\{X_i|i \in \sigma_1 \cup \sigma_3\}) - H(\{X_i|i \in \sigma_3\}) \\
&\quad - H(\{X_i|i \in \sigma_1 \cup \sigma_2 \cup \sigma_3\}) + H(\{X_i|i \in \sigma_2 \cup \sigma_3\}) \\
&= \mu(\cup_{i \in \sigma_1 \cup \sigma_3} \tilde{X}_i) - \mu(\cup_{i \in \sigma_3} \tilde{X}_i) \tag{4.13} \\
&\quad - \mu(\cup_{i \in \sigma_1 \cup \sigma_2 \cup \sigma_3} \tilde{X}_i) + \mu(\cup_{i \in \sigma_2 \cup \sigma_3} \tilde{X}_i)
\end{aligned}$$

$$= \mu(\cup_{i \in \sigma_1} \tilde{X}_i \setminus \cup_{i \in \sigma_3} \tilde{X}_i) \tag{4.14}$$

$$\begin{aligned}
&\quad - \mu(\cup_{i \in \sigma_1} \tilde{X}_i \setminus \cup_{i \in \sigma_2 \cup \sigma_3} \tilde{X}_i) \\
&= \mu((\cup_{i \in \sigma_1} \tilde{X}_i) \cap (\cup_{i \in \sigma_2} \tilde{X}_i) \setminus \cup_{i \in \sigma_3} \tilde{X}_i) \tag{4.15}
\end{aligned}$$

Equation 4.13 is a direct use of our definition of μ . Equation 4.14 is true due to the following equation which hold for any sets A and B :

$$\mu(A \cup B) - \mu(B) = \mu(A \setminus B)$$

(see proposition 6 at the appendix)

Equation 4.15 is a consequence of the following equation:

$$\mu(A \setminus C) - \mu(A \setminus (B \cup C)) = \mu(A \cap B \setminus C)$$

(see proposition 8 at the appendix)

4.2 Usage examples

4.2.1 The notion of $\Lambda(X; Y; T)$

We will now define a new information quantity of three random variable:

$$\Lambda(X; Y; T) \equiv I(X; Y) - I(X; Y|T) \tag{4.16}$$

By using Theorem 1 we construct three corresponding sets \tilde{X} , \tilde{Y} and \tilde{T} . Thus:

$$I(X; Y) - I(X; Y|T) = \mu(\tilde{X} \cap \tilde{Y}) - \mu(\tilde{X} \cap \tilde{Y} \setminus \tilde{T}) \tag{4.17}$$

$$= \mu(\tilde{X} \cap \tilde{Y} \cap \tilde{T}) \tag{4.18}$$

Equation 4.18 was obtained using proposition 7 in the appendix. The last term shows that the quantity $\Lambda(X; Y; T)$ is symmetric. Thus, $\Lambda(X; Y; T) = \Lambda(T; X; Y) = \Lambda(Y; T; X)$. Note that $\Lambda(X; Y; T)$ might be negative. A classic example is when X and Y are two independent binary random variable uniformly distributed and $T = X \oplus Y$. In this case, we have $I(X; Y) = 0$, $I(X; Y|T) = 1$ and $\Lambda(X; Y; T) = -1$.

4.2.2 Three variables Markov chain

Suppose $X - T - Y$ form a Markov chain. An equivalent representation of that fact is that $I(X; Y|T) = 0$. When we examine the definition of Λ in this case, we obtain:

$$\Lambda(X; Y; T) = I(X; Y) - I(X; Y|T) = I(X; Y)$$

Using Theorem 1 in the general case of three variables, we obtain:

$$H(T) = \Lambda(X; Y; T) + I(T; X|Y) + I(T; Y|X) + H(T|X, Y)$$

In the case of a Markov chain, the last equation becomes:

$$H(T) = I(X; Y) + I(T; X|Y) + I(T; Y|X) + H(T|X, Y) \quad (4.19)$$

$$\Rightarrow H(T) \geq I(X; Y) \quad (4.20)$$

Thus, a random variable that *seperates* two random variables, has an entropy which is larger or equal to the mutual information among the variables.

Another corollary regarding a Markov chain is the 'Data Processing Inequality'.

$$I(X; T) = \Lambda(X; Y; T) + I(X; T|Y) \quad (4.21)$$

$$= I(X; Y) + I(X; T|Y) \quad (4.22)$$

$$\Rightarrow I(X; T) \geq I(X; Y) \quad (4.23)$$

4.3 re-formalization of the original problem

Now, after gaining some new intuition about the inter-relations among three random variables, let us look again at the formalization of the original optimization problem. We now wish to obtain a better understanding about the meaning of the quantities we are trying to minimize.

In order to find a random variable T that captures the dependency between X and Y , we minimize the expression:

$$\omega_1 I(X; Y|T) + \omega_2 H(T)$$

The quantity $H(T)$ can be decomposed into four quantities that corresponds to atoms. We will use this decomposition to gain a finer view on the optimization expression. Substituting $H(T)$ with its decomposition, we obtain:

$$\begin{aligned} \omega_1 I(X; Y|T) + \omega_2 H(T) &= \\ \omega_1 I(X; Y|T) + \omega_2 (I(X; T|Y) + I(Y; T|X) + H(T|X, Y) + \Lambda(X; Y; T)) &= \\ \omega_1 I(X; Y|T) + \omega_2 (I(X; T|Y) + I(Y; T|X) + H(T|X, Y)) + \omega_2 I(X; Y) - \omega_2 I(X; Y|T) & \end{aligned} \quad (4.24)$$

Recall that $I(X; Y)$ is fixed, and so we can remove it from the optimization expression without effecting the solution.

The original goal, as recalled, was to find such T that will contain as much information about the dependency and as little irrelevant information. In terms of the decomposition expression, we wish to maximize $\Lambda(X; Y; T)$ and minimize $I(X; T|Y)$, $I(Y; T|X)$ and $H(T|X, Y)$. This leads us to try to minimize the following expression:

$$-\omega_1 \Lambda(X; Y; T) + \omega_2 (I(X; T|Y) + I(Y; T|X) + H(T|X, Y))$$

again, since $I(X; Y)$ is fixed, it is the same as minimizing:

$$\omega_1 I(X; Y|T) + \omega_2 (I(X; T|Y) + I(Y; T|X) + H(T|X, Y))$$

Notice that the last expression is equivalent to expression 4.24 because we can choose (without loss of generality) this ω_1 to be $\omega_1 - \omega_2$ of expression 4.24. Thus, the last expression is just another formalization of our original optimization problem. Nevertheless, we prefer this formalization because here, ω_1 expresses exactly the importance we give to the capture of dependency and ω_2 expresses the importance of not including irrelevant information.

The optimization expression above assumes that all the irrelevant informations have equal importance. We can generalize the problem and give each irrelevant information its own importance:

$$\omega_1 I(X; Y|T) + \omega_2 I(X; T|Y) + \omega_3 I(Y; T|X) + \omega_4 H(T|X, Y)$$

4.4 Special cases of the optimization problem

The general optimization problem consist of four weighted information quantities. Each quantity is non negative and the weight represents the importance that this quantity will be small. Thus, if the value of a quantity is not important to us, we will give it a weight that equals zero. If we insist that a certain quantity will be zero, we will give it an infinite weight.

We will now examine several special cases of the optimization problem.

Figures 4.3 illustrates the Venn diagrams that correspond to the various cases.

1. The case when $\omega_1 = \infty$ which means that $I(X; Y|T) = 0$. In this case we look for a compact variable that contains *all* the information of the dependency. The decomposition of $H(T)$ is:

$$H(T) = I(X; T|Y) + I(Y; T|X) + H(T|X, Y) + I(X; Y)$$

Thus, the information of T consist of the dependency information and three other types of information we try to minimize.

2. When $\omega_4 = \infty$ i.e $H(T|X, Y) = 0$. In this case we look for a deterministic function of (X, Y) that captures the dependency.

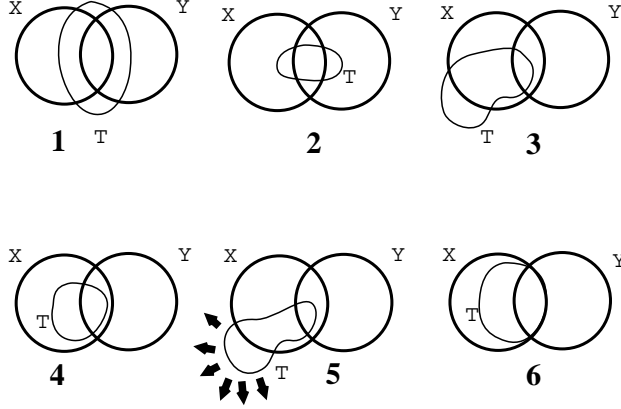


Figure 4.3: Venn diagrams of various special cases of the optimization problem.

3. When $\omega_3 = \infty$ i.e $I(Y; T|X) = 0$ or $P(T|X) = P(T|X, Y)$. Here, we try to find a compact extractor that is a stochastic function of X .
4. When $\omega_3 = \omega_4 = \infty$, we look for a deterministic function of X that captures the dependency:

$$H(T|X) = H(T|X, Y) + I(Y; T|X) = 0 + 0$$

This problem was considered by Tishby and Slonim at [6] where they introduced an agglomerative approximation algorithm for the problem.

5. When $\omega_3 = \infty$ and $\omega_4 = 0$, we look for a stochastic function of X that preserves information about Y while losing information about X . Since $I(Y; T|X) = 0$ and $\omega_4 H(T|X, Y) = 0$, we have:

$$I(T; Y) = \Lambda(X; Y; T) + I(Y; T|X) = \Lambda(X; Y; T)$$

$$\begin{aligned} & \omega_1 I(X; Y|T) + \omega_2 I(X; T|Y) + \omega_3 I(Y; T|X) + \omega_4 H(T|X, Y) \\ = & \omega_1 (I(X; Y) - \Lambda(X; Y; T)) + \omega_2 (I(T; X) - \Lambda(X; Y; T)) \\ = & \omega_1 (I(X; Y) - I(T; Y)) + \omega_2 (I(T; X) - I(T; Y)) \\ = & -(\omega_1 + \omega_2) I(T; Y) + \omega_2 I(T; X) + \omega_1 I(X; Y) \end{aligned}$$

Trying to minimize the last expression under the constrains is to try to loose information about X and keep information about Y . This problem is the Information bottleneck problem as defined by Tishby et al [7].

6. When $\omega_1 = \omega_3 = \omega_4 = \infty$, we look for a sufficient statistic that captures the dependency. We have already seen that $\omega_3 = \omega_4 = 0$ constrains T to be a deterministic function of X . The fact that $I(X; Y|T) = 0$ means that the Markovian condition $X - T - Y$, holds. This means that T is a sufficient statistic of X on Y .

4.5 historical notes

The original idea of the correspondence between information measures and sets is due to Hu ([3]). In his original paper he shows that any information-measures equation which is true for any random variables, has its correspondence equation that is true for any sets and additive function. Hu did not try to construct a particular additive function and sets that will have the same values of the correspondence information measures.

This idea of Hu was greatly developed by Yeung ([8], [9] and [5]) who introduced the concept of I-measure in a similar way as it was introduced in this chapter. Yeung showed how the concept of I-measure can be of a great value for dealing with problems of various dependent random variables. The work that was presented in this chapter was done without knowing about Yeung's work.

Chapter 5

Coping with the general problem

Before we deal with the various special cases, we will examine the general optimization problem and see what can be said about the solution. As described previously, the formalization of the general optimization problem is as follows:

Find a cardinality of T (denoted as $|T|$) and $|T| \cdot |X| \cdot |Y|$ nonnegative values that are denoted as $P(t|x, y)$ that minimize the expression:

$$\omega_1 I(X; Y|T) + \omega_2 I(X; T|Y) + \omega_3 I(Y; T|X) + \omega_4 H(T|X, Y) \quad (5.1)$$

s.t:

$$\forall x, y : \sum_t P(t|x, y) = 1$$

Substitution each information quantity with its definition, we obtain:

$$\begin{aligned} & \omega_1 E \left[\log \frac{P(X, Y|T)}{P(X|T)P(Y|T)} \right] + \omega_2 E \left[\log \frac{P(X, T|Y)}{P(X|Y)P(T|Y)} \right] + \\ & \omega_3 E \left[\log \frac{P(Y, T|X)}{P(Y|X)P(T|X)} \right] + \omega_4 E [\log P(T|X, Y)] \\ = & \omega_1 E \left[\log \frac{P(X, Y, T)P(T)}{P(X, T)P(Y, T)} \right] + \omega_2 E \left[\log \frac{P(X, T, Y)P(Y)}{P(X, Y)P(T, Y)} \right] + \\ & \omega_3 E \left[\log \frac{P(Y, T, X)P(X)}{P(Y, X)P(T, X)} \right] + \omega_4 E \left[\log \frac{P(T, X, Y)}{P(X, Y)} \right] + \\ = & (\omega_1 + \omega_2 + \omega_3 + \omega_4) E [\log P(X, Y, T)] - (\omega_1 + \omega_3) E [\log P(X, T)] - \\ & (\omega_1 + \omega_2) E [\log P(Y, T)] + \omega_1 E [\log P(T)] + \omega_3 E [\log P(X)] + \\ & \omega_2 E [\log P(Y)] + \omega_4 E [\log P(X, Y)] \end{aligned}$$

In order to find solutions for the constrained optimization problem, we will look

for fixed points of the following Lagrangian:

$$\begin{aligned}\mathcal{L}(P(t|x, y)) &= (\omega_1 + \omega_2 + \omega_3 + \omega_4)E[\log P(X, Y, T)] - \\ &\quad (\omega_1 + \omega_3)E[\log P(X, T)] - (\omega_1 + \omega_2)E[\log P(Y, T)] + \\ &\quad \omega_1 E[\log P(T)] + \omega_3 E[\log P(X)] + \omega_2 E[\log P(Y)] + \\ &\quad \omega_4 E[\log P(X, Y)] + \sum_{x, y} \lambda_{x, y} \left(1 - \sum_t P(t|x, y)\right)\end{aligned}$$

Now, we need to see when the partial derivatives equals zero:

$$\begin{aligned}\frac{\partial \mathcal{L}(P(t|x, y))}{\partial P(t_0|x_0, y_0)} &= (\omega_1 + \omega_2 + \omega_3 + \omega_4) \frac{\partial E[\log P(X, Y, T)]}{\partial P(t_0|x_0, y_0)} - (\omega_1 + \omega_3) \frac{\partial E[\log P(X, T)]}{\partial P(t_0|x_0, y_0)} \\ &\quad - (\omega_1 + \omega_2) \frac{\partial E[\log P(Y, T)]}{\partial P(t_0|x_0, y_0)} + \omega_1 \frac{\partial E[\log P(T)]}{\partial P(t_0|x_0, y_0)} + \lambda_{x_0, y_0}\end{aligned}\quad (5.2)$$

Hence, we need to find the derivative of expressions of the form $E(\log P(Z))$ where Z is a subset of $\{X, Y, T\}$. We denote $Z_{x, y, t}$ as the values subset of x, y, t which corresponds to the subset Z .

Lemma 1

$$\frac{\partial E[\log P(Z)]}{\partial P(t_0|x_0, y_0)} = p(x_0, y_0) [\log P(Z_{x_0, y_0, t_0}) + 1]$$

proof:

$$\begin{aligned}\frac{\partial E[\log P(Z)]}{\partial P(t_0|x_0, y_0)} &= \frac{\partial \sum_{x, y, t} P(x, y, t) \log P(Z_{x, y, t})}{\partial P(t_0|x_0, y_0)} \\ &= \sum_{x, y, t} \frac{\partial P(x, y, t)}{\partial P(t_0|x_0, y_0)} \log P(Z_{x, y, t}) + \sum_{x, y, t} P(x, y, t) \frac{\partial \log P(Z_{x, y, t})}{\partial P(t_0|x_0, y_0)} \\ &= \sum_{x, y, t} \frac{\partial P(x, y) P(t|x, y)}{\partial P(t_0|x_0, y_0)} \log P(Z_{x, y, t}) + \sum_{x, y, t} P(x, y, t) \frac{1}{P(Z_{x, y, t})} \frac{\partial P(Z_{x, y, t})}{\partial P(t_0|x_0, y_0)} \\ &= P(x_0, y_0) \log P(Z_{x_0, y_0, t_0}) + \sum_{x, y, t} \frac{P(x, y, t)}{P(Z_{x, y, t})} \frac{\partial P(Z_{x, y, t})}{\partial P(t_0|x_0, y_0)}\end{aligned}\quad (5.3)$$

Now,

$$\begin{aligned}\frac{\partial P(Z_{x, y, t})}{\partial P(t_0|x_0, y_0)} &= \sum_{x', y', t'} \frac{\partial P(x', y', t', Z_{x, y, t})}{\partial P(t_0|x_0, y_0)} \\ &= \sum_{x', y', t'} \frac{\partial P(x', y') P(t'|x', y') P(Z_{x, y, t}|x', y', t')}{\partial P(t_0|x_0, y_0)}\end{aligned}$$

The expression $P(Z_{x, y, t}|x', y', t')$ equals one whenever $Z_{x, y, t} = Z_{x', y', t'}$ and equals zero otherwise. Hence:

$$\frac{\partial P(Z_{x, y, t})}{\partial P(t_0|x_0, y_0)} = \begin{cases} P(x_0, y_0) & \text{when } Z_{x, y, t} = Z_{x_0, y_0, t_0} \\ 0 & \text{when } Z_{x, y, t} \neq Z_{x_0, y_0, t_0} \end{cases}$$

By using this result in 5.3. we obtain:

$$\begin{aligned}
\frac{\partial E[\log P(Z)]}{\partial P(t_0|x_0, y_0)} &= P(x_0, y_0) \log P(Z_{x_0, y_0, t_0}) + \sum_{\{x, y, t\} \setminus Z} \frac{P(Z_{x_0, y_0, t_0}, \{x, y, t\} \setminus Z)}{P(Z_{x_0, y_0, t_0})} P(x_0, y_0) \\
&= P(x_0, y_0) \log P(Z_{x_0, y_0, t_0}) + P(x_0, y_0) \frac{P(Z_{x_0, y_0, t_0})}{P(Z_{x_0, y_0, t_0})} \\
&= P(x_0, y_0) [\log P(Z_{x_0, y_0, t_0}) + 1]
\end{aligned}$$

□

Using Lemma 1 on equation 5.2 we obtain:

$$\begin{aligned}
\frac{1}{P(x_0, y_0)} \frac{\partial \mathcal{L}(P(t|x, y))}{\partial P(t_0|x_0, y_0)} &= (\omega_1 + \omega_2 + \omega_3 + \omega_4)(\log P(x_0, y_0, t_0) + 1) \\
&\quad - (\omega_1 + \omega_3)(\log P(x_0, t_0) + 1) \\
&\quad - (\omega_1 + \omega_2)(\log P(y_0, t_0) + 1) \\
&\quad + \omega_1(\log P(t_0) + 1) + \frac{\lambda_{x_0, y_0}}{P(x_0, y_0)}
\end{aligned}$$

Under the assumption that $P(x_0, y_0) \neq 0$.

We denote $\beta_1 = \omega_1 + \omega_2 + \omega_3 + \omega_4$, $\beta_2 = -(\omega_1 + \omega_3)$, $\beta_3 = -(\omega_1 + \omega_2)$ and $\beta_4 = \omega_1$. Thus, a necessary condition for an optimal point will be:

$$\begin{aligned}
0 &= \beta_1(\log P(t_0|x_0, y_0) + \log P(x_0, y_0) + 1) \\
&\quad + \beta_2(\log P(x_0, t_0) + 1) + \beta_3(\log P(y_0, t_0) + 1) \\
&\quad + \beta_4(\log P(t_0) + 1) + \frac{\lambda_{x_0, y_0}}{P(x_0, y_0)}
\end{aligned}$$

If $\beta_1 \neq 0$ we have:

$$\begin{aligned}
\log P(t_0|x_0, y_0) &= -\log P(x_0, y_0) - \frac{\beta_2}{\beta_1} \log P(x_0, t_0) - \frac{\beta_3}{\beta_1} \log P(y_0, t_0) \\
&\quad - \frac{\beta_4}{\beta_1} \log P(t_0) - \frac{\beta_1 + \beta_2 + \beta_3 + \beta_4}{\beta_1} - \frac{\lambda_{x_0, y_0}}{\beta_1 P(x_0, y_0)}
\end{aligned}$$

Or,

$$\begin{aligned}
P(t_0|x_0, y_0) &= P(x_0, y_0)^{-1} \cdot P(x_0, t_0)^{-\frac{\beta_2}{\beta_1}} \cdot P(y_0, t_0)^{-\frac{\beta_3}{\beta_1}} \cdot P(t_0)^{-\frac{\beta_4}{\beta_1}} \\
&\quad \cdot 2^{-\frac{\beta_1 + \beta_2 + \beta_3 + \beta_4}{\beta_1} - \frac{\lambda_{x_0, y_0}}{\beta_1 P(x_0, y_0)}} \\
&= P(x_0, t_0)^{\frac{\omega_1 + \omega_3}{\omega_1 + \omega_2 + \omega_3 + \omega_4}} \cdot P(y_0, t_0)^{\frac{\omega_1 + \omega_2}{\omega_1 + \omega_2 + \omega_3 + \omega_4}} \cdot P(t_0)^{-\frac{\omega_1}{\omega_1 + \omega_2 + \omega_3 + \omega_4}} \\
&\quad \cdot \left(P(x_0, y_0)^{-1} \cdot 2^{-\frac{P(x_0, y_0)\omega_4 + \lambda_{x_0, y_0}}{(\omega_1 + \omega_2 + \omega_3 + \omega_4)P(x_0, y_0)}} \right)
\end{aligned}$$

The last equation expresses a relation between one optimization variable and the other variables. This relation must hold for an optimal solution and can

serve for an iterative algorithm that converges to a set of variables that satisfies it. The last element (enclosed in parentheses) does not contain optimization variables. If we try to find a solution using an iterative algorithm over the variables, this element will be a constant.

Note that this equation expresses the desire to find a separator. Instead of the general equation that holds for any three variables:

$$P(t_0|x_0, y_0) = P(x_0, y_0)^{-1}P(t_0)P(x_0|t_0)P(y_0|x_0, t_0)$$

we have an equation of the form:

$$P(t_0|x_0, y_0) = P(x_0, y_0)^{-1}P(t_0)^{c_1}P(x_0|t_0)P(y_0|t_0)^{c_2}$$

where c_1 and c_2 are constants.

Recall that the random variable T is a separator if and only if $P(y_0|x_0, t_0) = P(y_0|t_0)$.

Chapter 6

The deterministic variant

In this chapter we will try to extract the mutual information by a deterministic function of one of the variables. The correspondence I-diagram for this approximation variant appears as case 4 in 4.3. As in the original problem, We are given a joint distribution of two random variables $P(X, Y)$ but this time, we try to encapsulate the mutual dependency by a variable T that satisfies: $T = f(X)$.

6.1 Decomposition of entropy and mutual information

First, we will show that a function of X , induces an interesting decomposition of the mutual information $I(X; Y)$. The fact that $T = f(X)$ is equivalent to the fact that $H(T|X) = 0$. Now, according to Theorem 1 (the information-set-correspondence), we have:

$$H(T|X) = H(T|X, Y) + I(T; Y|X)$$

Since $H(T|X)$ and $H(T|X, Y)$ equal zero, also $I(T; Y|X)$ equal zero and thus the Markovian condition $T - X - Y$, holds.

Proposition 1 (mutual information decomposition) :

if

$$I(T; Y|X) = 0$$

then :

$$I(X; Y) = I(T; Y) + \sum_t P(t)I(X; Y|t) \quad (6.1)$$

proof:

Assume $I(T; Y|X) = 0$. Now,

$$I(X; Y) = \Lambda(X; Y; T) + I(X; Y|T) \quad (6.2)$$

$$= I(T; Y) - I(T; Y|X) + I(X; Y|T) \quad (6.3)$$

$$= I(T; Y) + I(X; Y|T) \quad (6.4)$$

$$= I(T; Y) + \sum_t P(t)I(X; Y|t) \quad (6.5)$$

Where 6.2 and 6.3 use two different expressions for $\Lambda(X; Y; T)$, 6.4 is true since $I(T; Y|X) = 0$ and 6.5 is just a usage of the definition of conditional mutual information. \square

Let us now think about the meaning of proposition 1 in the case that $T = f(X)$. In this case, for each $t_0 \in \mathcal{T}$ and $y_0 \in \mathcal{Y}$ we have:

$$P(t_0, y_0) = \sum_{x \in \{x|f(x)=t_0\}} P(x, y_0)$$

Which means that any column of $P(T, Y)$ consists of sums of columns of $P(X, Y)$. Figure 6.1 illustrates this fact. Looking at 6.1 in that light, reveals that the mutual information $I(X; Y)$ is decomposed into the mutual information of the coarse matrix $I(T; Y)$ and a convex combination of refined mutual informations that were merged.

In the case that T only merges two columns, the information lost is the weighted information of the those columns matrix. More formally, if

$$T = f(X) = \begin{cases} t_1 & \text{when } X = x_1 \text{ or } X = x_2 \\ X & \text{when } X = x_i, i > 2 \end{cases}$$

Then,

$$I(X; Y) - I(T; Y) = \sum_t P(t)I(X; Y|t) = P(t_1)I(X; Y|t_1)$$

6.2 A greedy agglomerative algorithm

The last result suggests a greedy agglomerative algorithm for extracting the mutual information by a deterministic variable. The algorithm receives as inputs $P(X, Y)$ and k , the desired cardinality of T - k . The output is a deterministic extractor T .

1. begin with $T = X$
2. construct $\hat{T} = f(T)$ such that f merges the two values t_1 and t_2 that minimizes the expression:

$$(P(t_1) + P(t_2)) I(T; Y|T \in \{t_1, t_2\})$$

3. make $T = \hat{T}$

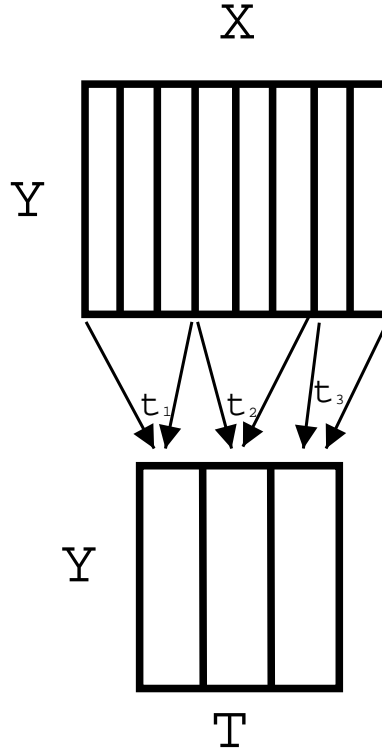


Figure 6.1:

4. if $|T| = k$ stop, else goto 2

The algorithm does not guarantee an optimal extractor of cardinality k . There might be more compact extractors (smaller $H(T)$) that reach a better information extraction (greater $I(T; Y)$).

In each step the algorithm merges two columns that reduces the minimal amount of mutual information. If there exist a merge that does not reduce the mutual information at all, such a merge is chosen first. Thus, the algorithm finds the minimal sufficient statistic first, and only then, look for a variable that loose some information.

Note that the quantity we minimize in each step is the weighted Jentsen-Shannon distance between the two columns. The Jentsen-Shannon distance between two distributions P_1, P_2 is defined as follows:

$$JS_{[\pi, 1-\pi]}(P_1, P_2) = H(\pi P_1 + (1 - \pi)P_2) - (\pi H(P_1) + (1 - \pi)H(P_2))$$

when π and $1 - \pi$ are the weights.

And thus,

$$\begin{aligned}
I(T; Y|T \in \{t_1, t_2\}) &= H(Y|T \in \{t_1, t_2\}) - H(Y|T, T \in \{t_1, t_2\}) \\
&= H(P(Y|T = t_1)P(T = t_1|T \in \{t_1, t_2\}) + P(Y|T = t_2)P(T = t_2|T \in \{t_1, t_2\})) \\
&\quad - (P(T = t_1|T \in \{t_1, t_2\})H(P(Y|T = t_1)) + P(T = t_2|T \in \{t_1, t_2\})H(P(Y|T = t_2))) \\
&= JS_{[P(T=t_1|T \in \{t_1, t_2\}), P(T=t_2|T \in \{t_1, t_2\})]}(P(Y|T = t_1), P(Y|T = t_2))
\end{aligned}$$

Thus, in each step, the algorithm seeks for two conditional distributions that minimize the Jentsen-Shannon distance. If there exist identical conditional probabilities, they will be picked first, of course. Because of that fact, the algorithm finds the minimal sufficient statistic, before it seeks for an extractor that loses information.

The agglomerative algorithm that was described here is due to Tishby et al ([6]) although the approach taken there, differs from ours.

6.3 Finding the best deterministic extractor is hard

The greedy algorithm does not guarantee to generate the best possible extractor. We will now show, that even the problem of finding the best deterministic extractor with a cardinality equals 2, is NP-hard.

We will describe a polynomial reduction to the Subset-Sum problem. The Subset-Sum problem can be described as follows: given a finite set $S \subset \mathbb{N}$ and number a $k \in \mathbb{N}$, decide whether there exist a subset $S' \subseteq S$ such that $\sum_{s' \in S'} s' = k$. This decision problem is known to be in the NP-Complete set.

Suppose we know how to find the best deterministic extractor of X on Y . As a special case, we know how to find the best extractor of cardinality 2 when X is a function of Y .

Since, $I(X; Y) = H(X)$ and since $|X| = 2$, the problem is to find a partition on the elements of X into two sets, with probabilities of p and $1 - p$ such that $H_0(p, 1 - p)$ is maximal. $H_0(p, 1 - p)$ is maximal when $|p - 1/2|$ is minimal.

Thus, the solution for the best extractor of cardinality 2 gives a solution for the following problem: given a finite set $H \subset \mathbb{Q}$ such that $\sum_{h \in H} h = 1$, find a set $H' \subset H$ such that $|\sum_{h' \in H'} h' - 1/2|$ is minimal. Or, stated as a decision problem: determine whether there exist H' such that $\sum_{h' \in H'} h' = 1/2$.

The last decision problem can be stated using natural numbers in the following way: Given a finite set $S \subset \mathbb{N}$, decide whether there exist a subset $S' \subset S$ such that

$$\sum_{s' \in S'} s' = \frac{\sum_{s \in S} s}{2}$$

In order to translate this alternative representation to the previous representation, we need to divide each element by $\sum_{s \in S} s$.

Now, consider again the problem of Subset-Sum. We denote $K = \sum_{s \in S} s$ and k as the desired sum of the problem. In the case that $k = K/2$, we have already showed how to make the decision.

In the case that $K > k > K/2$, we define a new set $\bar{S} = S \cup \{a\}$ when $a = 2k - K$. We know how to decide whether there is $\bar{S}' \subset \bar{S}$ which sums to k . If there is such \bar{S}' , there is a subset of \bar{S} that sums to k and doesn't include a and therefor it is a subset of S . This subset might be \bar{S}' or $\bar{S} \setminus \bar{S}'$.

Deciding whether there exist such a \bar{S}' , is equal in this case to deciding whether there exist S' that sums to k .

In the last case, when $0 < k < K/2$, we will decide whether there is a subset that sums to $K - k$ which is greater than $K/2$. If there exist such a subset S' , $S \setminus S'$ will sums to k . \square

Chapter 7

Deterministic function of two variables

7.1 Motivation

A natural question that arises from the discussion of the last chapter is whether it is possible to gain a better approximation of the mutual information by using a function of the *two* variables rather than a function of just one variable.

We already know how to find the best function of one variable that preserves *all* the information about the other, this is the minimal sufficient statistic. Can we find a better variable that preserves all the information by allowing it to be a function of the two variables? More formally, we are looking for a new variable $T = f(X, Y)$ which is a separator, and has minimal entropy.

First, we will give an example that will serve as a motivation for this question. The following joint probability matrix will allow a good two variables function extractor but no good one-variable-extractor.

Example 1

Let $P(X, Y)$ be a joint probability function of two *independent* random variable of cardinality N and M , respectively. We will assume that M and N are both even numbers. Let $\{p_1, \dots, p_N\}$ be N different values in the interval $(0, 1)$. for each $j \in \{1, \dots, N\}$ and $i \in \{1, \dots, M/2\}$ we will set $P'(i, j) = p_i P(i, j)$. For each $j \in \{1, \dots, N\}$ and $i \in \{M/2 + 1, \dots, M\}$ we will set $P'(i, j) = (1 - p_i)P(i, j)$.

Similarly, let $\{q_1, \dots, q_M\}$ be different values of the interval $(0, 1)$. We construct P'' as follows: for each $i \in \{1, \dots, M\}$ for each $j \in \{1, \dots, N/2\}$ we'll set $P''(i, j) = q_j P'(i, j)$ and for each $j \in \{N/2 + 1, \dots, N\}$ we'll set $P''(i, j) = (1 - q_j)P'(i, j)$. As a final step, we will normalize P'' . figure 7.1 illustrates the matrix of P'' .

One can verify that P'' has the following properties:

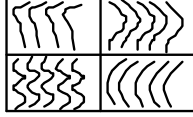


Figure 7.1: An example of a joint probability matrix that enables a good two variable extractor but no good one-variable-extractor. The curves represent similar conditional probabilities inside blocks.

- The matrix can be divided into 4 blocks, each of rank that equals one.
- No two columns or two rows are similar (i.e. are dependent as vectors).

Those properties suggest that there exists a random variable that is a function of the two variables and that extracts all the mutual information among them. The cardinality of this extractor will be only four. On the other hand, no one-variable-extractor exists, other than the original variables.

In our example, each value of the extractor, corresponded to a combinatorial rectangle of the original matrix. We will now show that this was not a coincidence but the general case.

7.2 A characterization of a two-variable-extractor

The fact that T is a function of the two variables suggests that each value of T corresponds to a set of entries of the matrix of $P(X, Y)$, the values that satisfy $P(x, y|t_0) > 0$.

Proposition 2 *Each value of T corresponds to a combinatorial rectangle in the matrix $P(X, Y)$.*

proof:

Let t be a value of T . Let S_x be the set of all values of X that correspond to t_0 i.e.

$$S_x = \{x|P(x|t) > 0, x \in \mathcal{X}\}$$

In the same manner we define

$$S_y = \{y|P(y|t) > 0, y \in \mathcal{Y}\}$$

Since T is a separator, for each $x, y \in \mathcal{X} \times \mathcal{Y}$ we have $P(x, y|t) = P(x|t)P(y|t)$. Thus, $P(x, y|t) > 0 \Leftrightarrow x, y \in S_x \times S_y$. \square

7.3 A Cartesian multiplication extractor

We will now restrict ourselves to a more specific question: given a joint probability matrix $P(X, Y)$, find a function of X , named \hat{X} and a function of Y , named

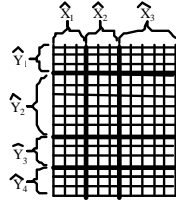


Figure 7.2: An example of a coarse partition on a matrix which is a Cartesian multiplication of a partition on the rows and a partition on the columns

\hat{Y} such that $X - (\hat{X}, \hat{Y}) - Y$ form a Markov chain and $H(\hat{X}, \hat{Y})$ is minimal. We are now looking for a partition of the original matrix into a coarser grid. This problem is different from the general problem because it is not an arbitrary partition into combinatorial rectangles but a partition which is a Cartesian multiplication of a partition of the rows and a partition of the columns. Figure 7.2 illustrates the situation.

7.4 The partition lattice

Since we are going to deal with various partitions of the rows and the columns of the matrix, let us examine the structure of the partitions set and agree on certain graphical notations. The partitions binary relation “finer (or equal) than” is a reflexive, antisymmetric and transitive relation, and hence is a partial order on the partitions set. If one partition is finer than the other, we will say that it is “smaller”. If one partition is coarser than another partition, we will say that it is “bigger”. In order to simplify the writing, we will sometimes use “smaller” and “bigger” instead of “smaller or equal” and “bigger or equal” . The exact meaning should be clear from the context.

For any subset of partitions, there is a unique partition that is smaller than all the partitions of the subset and bigger than all the other partitions that are smaller than the partitions of the subset. This partition is called the “Infimum” of the partitions subset. Similarly, the “Suprimum” is a unique partition that is bigger than all the partitions of the subset and smaller than all the other partitions that are bigger than the partitions of the subset.

The fact that any partitions subset has an infimum and a suprimum, implies that the partitions set with the specified partial order, forms a lattice.

To illustrate the partitions lattice, we draw various vertexes that represent partitions and directed edges between them that represent the order. The direction of the edges is always from the smaller to the bigger. The general layout of the vertexes is of a diamond form such that the bottom vertex represents the infimum of all the partitions and the upper vertex represents the suprimum of all the partitions. Figure 7.3 is an example of such illustration.

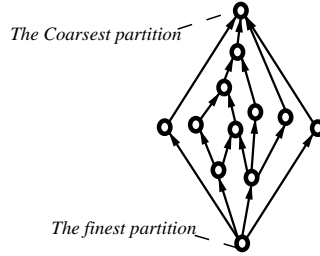


Figure 7.3: An example of a partitions lattice. Each vertex represents a partition and each directed edge represents the fact that one partition is finer than the other

7.5 The GCP

In attempt to answer the last question we will define a new operation that we will call: “*Greatest Common Partition (GCP)*”. The GCP will get as a parameter a subset of columns and its value will be a partition on the rows of the matrix. The GCP’s rows partition will be the most coarse partition that will divide all the columns into parts of similar distributions. Figure 7.4 illustrates the operation of the GCP. The GCP operation will be used by an algorithm that finds partitions of rows and partition of columns. These partitions will define the new random variable T and will guarantee that it will be a separator.

Let M be a matrix of the joint probability $P(X, Y)$ and Let $S_c = \{c_1, \dots, c_m\}$ be the set of all the columns of M . For each $S \subset S_c$, $GCP(S)$ will be a partition of the rows of M which is constructed in the following way: Let M_s be the $n \times |S|$ matrix formed by ordering all the columns of S in some order. The rows partition will be the partition that group similar rows of M_s .

Proposition 3 *GCP(S) is coarser (or equal) than any other partition of rows into similar parts of the columns in S*

proof: Let Prt be a partition of the rows into similar parts of all the columns. Prt will divide M_s into combinatorial blocks of rank equals to one. Each class of Prt will contain similar rows of M_s . Thus, the partition that groups all the similar rows into the same class must be coarser (or equal) than Prt . \square

7.6 A search algorithm for a good partition

In this section, we will describe an algorithm that searches for coarse mutual partitions. The algorithm will exploit the special structure of the domain in order to restrict the search space.

We start with the following observation: Given a column partition, $\{s_1, s_2, \dots, s_k\}$, where each s_i is a set of columns, The coarsest rows partition that corresponds

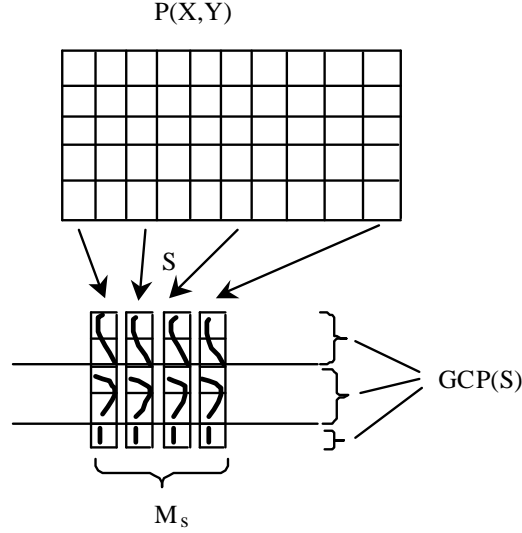


Figure 7.4: Illustration of the GCP function

to it, is:

$$\text{Inf}\{GCP\{s_1\}, GCP\{s_2\}, \dots, GCP\{s_k\}\} \quad (7.1)$$

Clearly, any rows partition that corresponds to the columns partition, is \leq any GCP of a columns set. The reverse statement is also true: any rows partition that is \leq all the GCPs, corresponds with the columns partition. Thus, the infimum of all the GCPs is a rows partition that is \geq than all other corresponding rows partition. We will call this partition “The implied” rows partition of the given columns partition.

It is also clear that if we unite two classes of the columns partition such that the union GCP is \geq the implied partition, the union action does not change the implied partition. Figure 7.5 illustrates this case.

In order to find coarse mutual partitions, we use the following algorithm:

1. Start by using the finest (smallest) columns partition $\{\{c_1\}, \{c_2\}, \dots, \{c_n\}\}$
The implied partition of this partition is the coarsest (greatest) rows partition: $\{\{r_1, r_2, \dots, r_n\}\}$.
2. Unify any two column classes that does not change the implied partition.
A unification does not change the implied partition if and only if, the GCP of the unified class is \geq the current implied partition.
3. Output the current columns partition and its implied partition. If the columns partition is the coarsest, exit.
4. Choose two column classes and unify them.

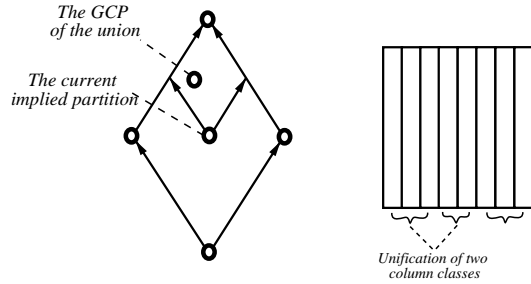


Figure 7.5: A union of two classes that has a GCP that is bigger than the current one, does not change the implied rows partition

5. go back to step 2.

Figure 7.6 illustrates this algorithm.

Note that we do not specify how we choose the two classes to be unified (step 4). Naturally, we will try to choose a couple that reduces the implied partition as little as possible. If one optional couple induces a greater implied partition than another couple, obviously, we will prefer it. If however, two possible couples induce incomparable implied partitions, we will have to try both. Thus, the described algorithm is indeed a search algorithm on the columns partition lattice.

Let us examine the operation of the above algorithm on the joint probability of example 1. The first unification step (2) will not unify any columns. At the choosing step (4), if two column of the proposed coarse partition are chosen, the next unification step will reach the proposed partition. If the two chosen columns belongs to different classes of the proposed partition, the implied partition will probably be much smaller than the proposed partition. Thus, on this example the search algorithm finds the good solution using only few iterations.

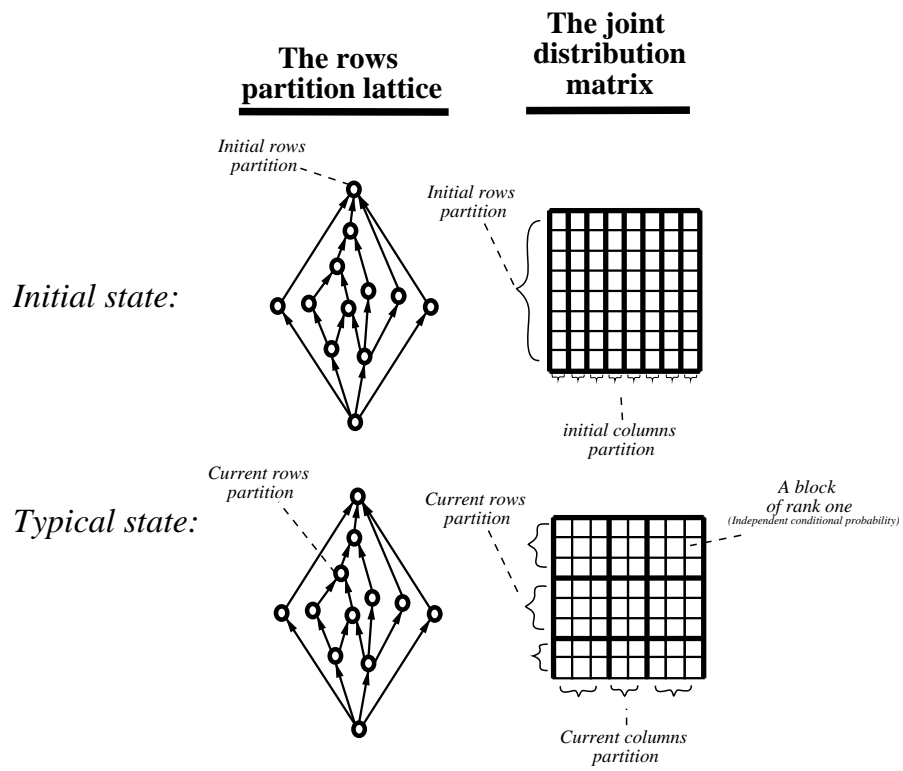


Figure 7.6: The partition algorithm...

Chapter 8

Discussion

This work is an attempt to deal with a fundamental question about the meaning of dependency among random variables. Although there is a natural way to quantify the amount of the dependency, there is no formal way to describe its nature. Our research tries to capture the dependency by a new random variable and to claim that this variable represents the dependency. It is soon becomes clear, that in the general case, there is no random variable that exactly captures the dependency. Our goal becomes finding a good approximation for a perfect representation of the dependency.

As a first stage, this question was formalized as an optimization problem of Information-Theory quantities. Then, a general framework to deal and to visualize relationship among random variables, was developed. This framework was used to present variants of the original question.

The next chapter was devoted to an attempt to handle the most general case using regular optimization tools. This direction did not reach full maturity. The following chapter tries to deal with a specific variant of the original question, capturing dependency using a deterministic function of one of the variables. We have shown that if we look for the best extractor of a given cardinality, the problem is NP-hard. Instead of looking for the best extractor we have presented an agglomerative greedy algorithm that finds an approximation. The difference between its solution and the optimal solution is still an open question. The last chapter deals with another variant of the general question: how to capture the dependency by a deterministic function of the two variables. First we show why a function of two variables can achieve better results than a function of one variable. Then, we introduce an iterative search algorithm for this problem. The algorithm exploit the structure of the solutions space and greatly reduce the size of the search. There is still much experimental work to be done on this algorithm. The result of those experiments should be compared with the result of the previous algorithm on practical data, results that can be found at [6].

This work is composed of various research path, each of which could have been developed much further. One of the main reasons for this is that the work was supposed to be a Ph.d thesis and was eventually converted to a master

thesis.

We think that the main importance of this work is not the answers that it gives to certain question, but rather the questions it rises, the general framework to deal with them and the insights it gives to the subject of interrelation among random variables. We feel that this framework could be used much further. A demonstration of such usage can be found in the work of Yeung [8], [9] and [5]. Another example of the advantage of this way of thinking about random variables can be found at [4].

Appendix A

This is a version of the inclusion/exclusion formula. It expresses a measure on intersection of sets under subtraction, in terms of unions of sets under subtraction.

Proposition 4 (inclusion/exclusion in context)

$$\mu(\cap_{i \in \sigma} \tilde{X}_i \setminus \cup_{i \in \bar{\sigma}} \tilde{X}_i) = \sum_{\phi \neq s \subset \sigma} -1^{|\phi|+1} \mu\left(\cup_{i \in s} \tilde{X}_i \setminus \cup_{i \in \bar{\sigma}} \tilde{X}_i\right)$$

for example, for any arbitrary sets $\tilde{X}_1, \tilde{X}_2, \tilde{X}_3$ and an arbitrary additive function μ , the following equation holds:

$$\mu(\tilde{X}_1 \cap \tilde{X}_2 \setminus \tilde{X}_3) = \mu(\tilde{X}_1 \setminus \tilde{X}_3) + \mu(\tilde{X}_2 \setminus \tilde{X}_3) - \mu(\tilde{X}_1 \cup \tilde{X}_2 \setminus \tilde{X}_3)$$

here σ is $\{1, 2\}$.

proof:

$$\mu(\cap_{i \in \sigma} \tilde{X}_i \setminus \cup_{i \in \bar{\sigma}} \tilde{X}_i) = \mu\left(\cap_{i \in \sigma} (\tilde{X}_i \setminus \cup_{j \in \bar{\sigma}} \tilde{X}_j)\right) \quad (\text{A.1})$$

$$= \sum_{\phi \neq s \subset \sigma} -1^{|\phi|+1} \mu\left(\cup_{i \in s} (\tilde{X}_i \setminus \cup_{j \in \bar{\sigma}} \tilde{X}_j)\right) \quad (\text{A.2})$$

$$= \sum_{\phi \neq s \subset \sigma} -1^{|\phi|+1} \mu\left(\cup_{i \in s} \tilde{X}_i \setminus \cup_{i \in \bar{\sigma}} \tilde{X}_i\right) \quad (\text{A.3})$$

Where A.1 is true because of the following equation that holds for any sets A, B and C :

$$(A \cap B) \setminus C = (A \setminus C) \cap (B \setminus C)$$

Equation A.2 is the regular inclusion/exclusion formula. Equation A.3 is true because of the following equation:

$$(A \setminus C) \cup (B \setminus C) = (A \cup B) \setminus C$$

Proposition 5

$$\sum_{\phi \neq s \subset \sigma} (-1)^{|\phi|+1} = 1$$

proof:

$$\sum_{\phi \neq s \subset \sigma} (-1)^{|s|+1} = \sum_{i=1}^{|\sigma|} \binom{|\sigma|}{i} (-1)^{i+1} \quad (\text{A.4})$$

$$0 = (1-1)^{|\sigma|} \quad (\text{A.5})$$

$$= \sum_{i=0}^{|\sigma|} \binom{|\sigma|}{i} (-1)^i \quad (\text{A.6})$$

$$= 1 - \sum_{i=1}^{|\sigma|} \binom{|\sigma|}{i} (-1)^{i+1} \quad (\text{A.7})$$

$$= 1 - \sum_{\phi \neq s \subset \sigma} (-1)^{|s|+1} \quad (\text{A.8})$$

Proposition 6 *let A and B be arbitrary sets, and let μ be an arbitrary additive function, the following equation holds:*

$$\mu(A \cup B) - \mu(B) = \mu(A \setminus B)$$

proof:

$$\begin{aligned} \mu(A \cup B) &= \mu(B \uplus (A \setminus B)) \\ &= \mu(B) + \mu(A \setminus B) \end{aligned}$$

Proposition 7

$$\mu(A \cap B) = \mu(A) - \mu(A \setminus B)$$

proof:

$$\mu(A) = \mu((A \setminus B) \uplus (A \cap B)) = \mu(A \setminus B) + \mu(A \cap B)$$

Proposition 8 *let A, B and C be arbitrary sets, and let μ be an arbitrary additive function, the following equation holds:*

$$\mu(A \setminus C) - \mu(A \setminus (B \cup C)) = \mu(A \cap B \setminus C)$$

proof:

$$\begin{aligned} \mu(A \setminus C) &= \mu((A \setminus B) \uplus ((A \cap B) \setminus C)) \\ &= \mu(((A \setminus B) \setminus C) \uplus ((A \cap B) \setminus C)) \\ &= \mu(A \setminus (B \cup C)) + \mu((A \cap B) \setminus C) \end{aligned}$$

[1] [3] [4] [2]

Bibliography

- [1] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991.
- [2] I. Csiszar and Korner. *Information Theory: Coding Theorem for Discrete Memoryless Systems*. Academic Press, and Budapest: Akademiai Kiado, New York, 1981.
- [3] Guo ding Hu. On the amount of information (in russian). *Teor. Veroyatnost. i Primenen.*, 4:447–455, 1962.
- [4] N. Slonim N. Friedman, O. Mosenzon and N. Tishby. Multivariate information bottleneck. *UAI 2001*. See www.cs.huji.ac.il/~nir/publications.html.
- [5] T. T. Lee R. W. Yeung and Z. Ye. Information theoretic characterization of conditional mutual independence and markov random fields. *IEEE Transactions on information Theory*, 48(7), July 2002.
- [6] N. Slonim and N. Tishby. Agglomerative information bottleneck. In *Neural Information Processing Systems (NIPS-99)*, pages 617–623. 1999.
- [7] N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In *Proc. 37th Allerton Conference on Communication and Computation*. 1999.
- [8] R. W. Yeung. A framework for linear information inequalities. *IEEE Trans. Inform. Theory*, 43(6):1924–1934, 1997.
- [9] Z. Zhang and R. W. Yeung. On characterization of entropy functions via information inequalities. *IEEE Trans. Inform. Theory*, 44:1440–1452, 1998.