Discrete Entropy

Joshua Erde

Department of Mathematics,
Universität Hamburg.

Contents

1 Probability Theory

This section is intended as a short introduction to the very basics of probability theory, covering only the basic facts about finite probability spaces that we will need to use in this course.

Definition. A *probability space* is a triple $(\Omega, \Sigma, \mathbb{P})$, where Ω is a set, $\Sigma \subseteq 2^\Omega$ is a σ-algebra, and $\mathbb{P}$ is a measure on Σ with $\mathbb{P}(\Omega) = 1$. To recall, Σ is a σ-algebra means:

- $\emptyset \in \Sigma$;
- If $A \in \Sigma$ then $A^c = \Omega \setminus A \in \Sigma$;
- If $(A_i : i \in \mathbb{N})$ are in Σ then $\bigcup_{i=1}^\infty A_i \in \Sigma$.

Note that, by the second and third condition, Σ is also closed under countable intersections. $\mathbb{P} : \Sigma \to \mathbb{R}$ is a measure means :

- $\mathbb{P}$ is non-negative;
- $\mathbb{P}(\emptyset) = 0$;
- For all countable families of disjoint sets $(A_i : i \in \mathbb{N})$ in Σ,

$$\mathbb{P}\left(\bigcup_{i=1}^\infty A_i\right) = \sum_{i=1}^\infty \mathbb{P}(A_i).$$

The elements of Σ are called *events* and the elements of Ω are called *elementary events*. For an event A, $\mathbb{P}(A)$ is called the *probability of* A.

During this course we will mostly consider *finite probability spaces*, those where Ω is finite and $\Sigma = 2^\Omega$. In this case the probability measure $\mathbb{P}$ is determined by the value it takes on elementary events. That is, given any function $p : \Omega \to [0, 1]$ that satisfies $\sum_{\omega \in \Omega} p(\omega) = 1$, the function on Σ given by $\mathbb{P}(A) = \sum_{\omega \in A} p(\omega)$ is a probability measure.

In a finite probability space, the most basic example of a probability measure is the *uniform distribution* on Ω, where

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} \text{ for all } A \subseteq \Omega.$$

One elementary fact that we will use often is the following, often referred to as the union bound:

Lemma 1.1 (Union bound)**.** *For any countable family of events* $(A_i : i \in \mathbb{N})$ *in* Σ*,*

$$\mathbb{P}\left(\bigcup_{i=1}^\infty A_i\right) \leq \sum_{i=1}^\infty \mathbb{P}(A_i)$$

Proof. For each $i \in \mathbb{N}$ let us define

$$B_i = A_i \setminus (\bigcup_{j=1}^{i-1} A_j).$$

Then $B_i \subseteq A_i$, and so $\mathbb{P}(B_i) \leq \mathbb{P}(A_i)$, and also $\bigcup_{i=1}^{\infty} B_i = \bigcup_{i=1}^{\infty} A_i$. Therefore, since the events $B_1, B_2, \ldots, B_n$ are disjoint, by the countable additivity of $\mathbb{P}$

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(B_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

□

Definition. Two events $A, B \in \Sigma$ are *independent* if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

More generally, a set of events $\{A_1, A_2, \ldots, A_n\}$ is *mutually independent* if, for any subset of indices $I \subseteq [n]$,

$$\mathbb{P}\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \mathbb{P}(A_i).$$

It is important to note that the notion of mutual independence is stronger than simply having pairwise independence of all the pairs A_i, A_j. Intuitively, the property of independence of two events, A and B, should mean that knowledge about whether or not A occurs should not influence the likelihood of B occurring. This intuition is made formal with the idea of *conditional probability.*

Definition. Given two events $A, B \in \Sigma$ such that $\mathbb{P}(B) \neq 0$, we define the *conditional probability of A, given that B occurs*, as

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Note that, as expected, A and B are independent if and only if $\mathbb{P}(A|B) = \mathbb{P}(A)$.

Definition. A *random variable* on a probability space $(\Omega, \Sigma, \mathbb{P})$ is a $\mathbb{P}$-measurable function $X : \Omega \to E$ to some measurable space E. That is, E is a set together with a σ-algebra Σ_E on E such that for any measurable $A \in \Sigma_E$

$$\{\omega \in \Omega \colon X(\omega) \in A\} \in \Sigma.$$

Given a measurable set $A \subseteq E$ the probability that the value X takes lies in A is $\mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\})$ which we will write as $\mathbb{P}(X \in A)$. It will sometimes be convenient to think about random variables not as functions from some probability space to a measurable space, but just in terms of the *distributions* on the measurable space they determine.

What do we mean by a distribution? Well, for every measurable set $A \in \Sigma_E$ we can assign it a measure $\hat{\mathbb{P}}(A) = \mathbb{P}(X \in A)$. It is not hard to check that the triple $(E, \Sigma_E, \hat{\mathbb{P}})$ is then a probability space. So, in fact, this is just another word for a notion we already have, that of a probability measure on Σ_E, and indeed for every probability space $(\Omega, \Sigma, \mathbb{P})$ the function $\mathrm{id} : \Omega \to \Omega$ is a random variable whose distribution agrees with the measure $\mathbb{P}$.

However, we tend to make a philosophical distinction between probability spaces $(\Omega, \Sigma, \mathbb{P})$ and the distribution of a random variable X. The former we tend to treat as merely sources of randomness, so that we don't care about the elements of the set Ω, whereas we quite often care about the specific values that a random variable takes, and the probability that it takes those values.

Given two random variables X and Y we write $X \sim Y$ if X and Y have the same distribution, that is, if $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$ for every $A \in \Sigma_E$.

A particularly common case is a *real random variable* when $E = \mathbb{R}$ and Σ_E is the borel σ-algebra on $\mathbb{R}$. However in this course we will mostly be interested in *discrete random variables*, that is random variables where the range of X is finite. Note that, in particular, this will be true whenever $(\Omega, \Sigma, \mathbb{P})$ is a finite probability space.

Definition. The *expectation* of a real random variable X is

$$\mathbb{E}(X) = \int_\Omega X(\omega)\, d\mathbb{P}(\omega).$$

In the case of a finite probability space this can be expressed more clearly as

$$\mathbb{E}(X) = \sum_{\omega \in \Omega} p(\omega) X(\omega).$$

The set of random variables forms an algebra over $\mathbb{R}$ with addition and multiplication defined pointwise. For example the random variable $X + Y$ is the function from Ω to $\mathbb{R}$ defined by $(X + Y)(\omega) = X(\omega) + Y(\omega)$.

Lemma 1.2 (Linearity of expectation)**.** *For any two random variables X and Y*

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$

Proof.

$$\begin{aligned}\mathbb{E}(X + Y) &= \int_\Omega (X + Y)(\omega)\, d\mathbb{P}(\omega) = \int_\Omega X(\omega) + Y(\omega)\, d\mathbb{P}(\omega) \\ &= \int_\Omega X(\omega)\, d\mathbb{P}(\omega) + \int_\Omega Y(\omega)\, d\mathbb{P}(\omega) = \mathbb{E}(X) + \mathbb{E}(Y).\end{aligned}$$

□

So expectation is linear, however in general it is not multiplicative. Indeed $\mathbb{E}(XY)$ can be quite different to $\mathbb{E}(X)\mathbb{E}(Y)$, however if the two random variable are independent the two will coincide.

Definition. Two random variables X, Y are *independent* if, for any two measurable sets $A, B \subseteq \mathbb{R}$ we have

$$\mathbb{P}(X \in A \text{ and } Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).$$

More generally, a set of random variables $\{X_1, X_2, \ldots, X_n\}$ is *mutually independent* if, for any subset of indices $I \subseteq [n]$ and any set of measurable sets $\{A_i \subseteq \mathbb{R} : i \in I\}$ we have

$$\mathbb{P}\left(X_i \in A_i \text{ for all } i \in I\right) = \prod_{i \in I} \mathbb{P}(X_i \in A_i).$$

We note the following useful fact, although we will not need it in the course.

Lemma 1.3. *For any two independent random variables, X and Y,*

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$$

Suppose that X and Y are both random variables on the same probability space $(\Omega, \Sigma, \mathbb{P})$, perhaps with different codomains E_X and E_Y. We say that X *determines* Y if Y is a function of X. That is, if there exists some function $f : E_X \to E_Y$ such that for every $\omega \in \Omega$

$$Y(\omega) = f(X(\omega)).$$

We will also want to use at various times Jensen's inequality, which can be neatly phrased as a probabilistic statement.

Lemma 1.4. *[Jensen's inequality] Let X be a real random variable and let $g : \mathbb{R} \to \mathbb{R}$ be convex. Then*

$$\mathbb{E}(g(X)) \geq g(\mathbb{E}(X)).$$

Proof. Let $L(x) = a + bx$ be the line tangent to g at the point $E(X)$. Since g is convex, g lies above the line L and hence $g(x) \geq L(x)$ for all $x \in \mathbb{R}$. Hence

$$\mathbb{E}(g(X)) \geq \mathbb{E}(L(X)) = \mathbb{E}(a + bX) = a + b\mathbb{E}(X) = L\mathbb{E}(X) = g(\mathbb{E}(X)).$$

□

Remark 1.5. *Note that, by considering $-g$ in the above theorem, we can conclude that if instead g is concave, then*

$$\mathbb{E}(g(X)) \leq g(\mathbb{E}(X)).$$

We will also use throughout the notes the following notation for comparing growth rates of functions, which it will be useful to be familiar with. Given two functions $f, g : \mathbb{N} \to \mathbb{R}$ we say that:

- $f = O(g)$ if there exists $C > 0$ such that for all sufficiently large n, $f(n) \leq Cg(n)$;
- $f = \Omega(g)$ if there exists $C > 0$ such that for all sufficiently large n, $f(n) \geq Cg(n)$;
- $f = o(g)$ if for sufficiently large n, $f(n) \leq Cg(n)$, for any fixed $C > 0$;
- $f = \omega(g)$ if for sufficiently large n, $f(n) \geq Cg(n)$, for any fixed $C > 0$;

2 Entropy

The idea of entropy originated in statistical mechanics. Broadly, given a thermodynamic system, such as a gas or a liquid if we know some global properties of the system, e.g temperature, volume, energy, there are many different *microstates*, that is configurations of the individual particles within the system, which are consistent with these measurements.

As an example imagine flipping 1000 coins. We have a global measurement, the number of heads, but for each particular value for this, there are many different configurations of the specific states each of the 1000 coins landed in which achieve this number of heads.

Under a broad assumption that each of these microstates are equally likely, Boltzmann defined entropy of the system to be $k_B \log(\#$ of microstates) where k_B is some constant. Gibbs generalized this to microstates with unequal probabilities and gave the formula

$$S = -k_B \sum p_i \log(p_i),$$

where S is the entropy, p_i is the probability of the ith microstates, and the sum ranges over all the microstates. This reduces to Boltzmann's formula when the p_i are equal.

The second law of thermodynamics states that the entropy of an isolated system never decreases, and so such systems naturally 'tend' towards the state with maximum entropy, known as thermodynamic equilibrium. This was an attempt to formalise the idea that there is a natural 'direction' to natural processes, for example to explain why heat is transferred from hotter objects to cooler objects, rather than the other way round (which would not by itself contradict the conservation of energy in a process).

In the 40s Shannon discovered that a similar function arose quite naturally in the study of information theory, and at the suggestion of Von Neumann, called it entropy.

"You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage." - John von Neumann

More recently the notion of entropy has found many unexpected applications in mathematics, in particular in combinatorics, but also in other fields such as geometry and ergodic theory.

Given a discrete random variable X let us write $p(x)$ for the probability $\mathbb{P}(X = x)$ for each x in the range of X. We define the *entropy* of the random variable X to be

$$\mathbb{H}(X) := \sum_x p(x) \log\left(\frac{1}{p(x)}\right).$$

Note that this quantity is always non-negative.

It might be helpful to think of entropy, at least heuristically, as a measure of the expected amount of 'surprise' we have upon discovering the value of X. We then have the following heuristic argument for why $\mathbb{H}(X)$ should be defined as above.

If we have an event A, such as the event that $X = x$ for some x, the amount of 'surprise' we have at the event A happening should just be some function $f(p)$ of $p := \mathbb{P}(A)$. There are a number of reasonable conditions we should expect f to satisfy:

- $f(1) = 0$, since a certain event is no surprise;
- f should be decreasing, since rarer events are more surprising;
- f is continuous;
- $f(pq) = f(p) + f(q)$, which can be motivated by considering independent events happening with probability p and q;
- finally, for normalisation we may as well assume $f(1/2) = 1$.

It turns out that $f(p) = \log \frac{1}{p}$ is the unique function satisfying these constraints. Then, $\mathbb{H}(X)$ is the expected value, taken over the range of X, of the surprise of the event that X takes each value, and so $\mathbb{H}(X)$ is the only 'reasonable' function representing the idea following these heuristics.

Broadly, a key idea in the course will be to be consider a random variable X as a having a *product structure*, in the sense that X is really some vector of random variables $(X_1, \dots, X_n)$. Whilst X might represent some 'global' information, these coordinates X_i might be much simpler, and represent some 'local' information.

For example, suppose we have a graph G and we consider a random variable X which chooses a perfect matching from G uniformly at random. Formally, X takes values in $2^{E(G)}$, that is subsets of the edge set of G, which we may identify with the space $\{0,1\}^{E(G)}$, where an edge set is mapped to its characteristic vector. X then has a very natural product structure: for each $e \in E(G)$ we can consider the random variable X_e, which is given by the coordinate of X corresponding to e. Each X_e has a very simple structure, it takes the value 0 with some fixed probability and the value 1 with some fixed probability, but the relationship between different coordinates can be very complex.

A motivating idea for the next section is that we will try to develop tools that allow us to relate $\mathbb{H}(X)$ to the quantities $\mathbb{H}(X_i)$ when $X = (X_1, \dots, X_n)$.

However, note in the above example there is not a unique 'product structure' that we can impose on this random variable X. Indeed instead we could consider the following: For each $v \in V(G)$ let us denote by Y_v the random variable which is given by the neighbour of v in X. Then the vector $Y = (Y_v \colon v \in V(G))$ also in some way represents the same information as X. For our purposes, we will care about the entropy of the random variables, and the following lemma will show that, two random variables which both determine each other will have the same entropy

Lemma 2.1. *Let X and Y be discrete random variables such that X determines Y. Then $\mathbb{H}(Y) \leq \mathbb{H}(X)$.*

Proof. Suppose that X takes values in $\mathcal{X}$ and Y takes values in $\mathcal{Y}$. By definition there is some function $f : \mathcal{X} \to \mathcal{Y}$ such that $Y = f(X)$. Then, for every $y \in \mathcal{Y}$ we have that $p(y) =$

$\sum_{x:\ f(x)=y} p(x)$ and $p(x) \leq p(f(x))$. Hence

$$\begin{aligned}\mathbb{H}(Y) &= \sum_{y\in\mathcal{Y}} p(y)\log\left(\frac{1}{p(y)}\right)\\ &= \sum_{y\in\mathcal{Y}}\sum_{x:\ f(x)=y} p(x)\log\left(\frac{1}{p(y)}\right)\\ &= \sum_{x\in\mathcal{X}} p(x)\log\left(\frac{1}{p(f(x))}\right)\\ &\leq \sum_{x\in\mathcal{X}} p(x)\log\left(\frac{1}{p(x)}\right) = \mathbb{H}(X).\end{aligned}$$

□

Hence if X determines Y and X determines Y then $\mathbb{H}(X) = \mathbb{H}(Y)$. So, for example, since every matching determines, and is determined by, the list of neighbours of each vertex in G, we know that $\mathbb{H}(X) = \mathbb{H}(Y)$ in the above example.

A fundamental example of a random vector is the *Bernoulli random variable* X, which takes two values, 0 and 1 with probability p and $1-p$ respectively, then

$$\mathbb{H}(X) = p\log\left(\frac{1}{p}\right) + (1-p)\log\left(\frac{1}{1-p}\right),$$

and so as $p \to 1$ or 0, $\mathbb{H}(X) \to 0$. Since this value will come up later in the course, we will write

$$h(p) := p\log\left(\frac{1}{p}\right) + (1-p)\log\left(\frac{1}{1-p}\right).$$

It is not hard to see that the entropy of this particular X is maximised when $p = 1/2$, when $\mathbb{H}(X) = 1$, and in fact in general we have that:

Lemma 2.2. *Let X be a discrete random variable and let R be the range of X.*

$$\mathbb{H}(X) \leq \log(|R|).$$

with equality if X is uniformly distributed.

Proof. We will use the following form of Jensen's inequality. Let f be concave on $[a,b]$, $\lambda_i \geq 0$ such that $\sum_{i=1}^n \lambda_i = 1$ and let $x_1, \ldots x_n \in [a,b]$. Then if we consider a real random variable Y taking the values x_i with probability λ_i, we have by Lemma 1.4

$$\sum_{i=1}^n \lambda_i f(x_i) = \mathbb{E}(f(Y)) \leq f(\mathbb{E}(Y)) = f\left(\sum_{i=1}^n \lambda_i x_i\right).$$

We note that $f(x) = \log(x)$ is a concave function on $(0,\infty)$, which can be seen since its derivative $\frac{1}{x}$ is decreasing on $(0,\infty)$, and so

$$\mathbb{H}(X) = \sum_{x\in R} p(x)\log\left(\frac{1}{p(x)}\right) \leq \log\left(\sum_{x\in R}\frac{p(x)}{p(x)}\right) = \log(|R|).$$

Finally it is easy to see that if X is uniformly distributed then $p(x) = \frac{1}{|R|}$ for each $x \in R$ and so $\mathbb{H}(X) = \log(|R|)$. □

This gives a useful connection between entropy and counting. We are going to define a whole host of generalisations of the entropy function, and in order to try and give you some intuition for such things, and give some working examples of calculating entropy, we'll keep a motivating example in mind as we go through these definitions.

Consider the probability space Ω given by a sequence of N fair coin flips for N very large, and the random variable $X : \Omega \to \{0,1\}^{[N]}$ where $X_i = 1$ if the ith coin flip was heads and 0 if it was tails. For every subset $A \subset [N]$ we can consider the random variable X_A given by the restriction of X to just the coordinates in A. In this way we have a correspondence between random variables and subsets.

Since X_A is uniformly distributed on $\{0,1\}^A$, Lemma 2.2 tells us that $\mathbb{H}(X) = \log|\{0,1\}^A| = \log 2^{|A|} = |A|$. So, in this setting there is a correspondence between the entropy of X_A and the cardinality of the set A.

Given two discrete random variables, X and Y, we define the *joint entropy* (X,Y) to be

$$\mathbb{H}(X,Y) = \sum_x \sum_y p(x,y) \log\left(\frac{1}{p(x,y)}\right),$$

where, as before, $p(x,y) := \mathbb{P}(X = x, Y = y)$. Note that, if X and Y are independent then, by definition $p(x,y) = p(x)p(y)$ for all $x \in X$ and $y \in Y$, and so

$$\begin{aligned}
\mathbb{H}(X,Y) &= \sum_x \sum_y p(x,y) \log\left(\frac{1}{p(x,y)}\right) \\
&= \sum_x \sum_y p(x)p(y) \log\left(\frac{1}{p(x)p(y)}\right) \\
&= \sum_x \sum_y p(x)p(y) \left(\log\left(\frac{1}{p(x)}\right) + \log\left(\frac{1}{p(y)}\right)\right) \\
&= \sum_x p(x) \log\left(\frac{1}{p(x)}\right) \sum_y p(y) + \sum_y p(y) \log\left(\frac{1}{p(y)}\right) \sum_x p(x) \\
&= \sum_x p(x) \log\left(\frac{1}{p(x)}\right) + \sum_y p(y) \log\left(\frac{1}{p(y)}\right) \\
&= \mathbb{H}(X) + \mathbb{H}(Y)
\end{aligned}$$

However, in general that will not be the case.

So, in our example if we have two subsets A and B, what will the joint entropy of X_A and X_B be? Well X_A takes values in $\{0,1\}^A$ and X_B takes values in $\{0,1\}^B$, but given $x \in \{0,1\}^A$ and $y \in \{0,1\}^B$ it's not necessarily true that $p(x,y) = p(x)p(y)$, that is, the random variables X_A and X_B are not necessarily independent. Indeed, since X_A and X_B are restrictions of the same random variable X, for every $i \in A \cap B$ we have $(X_A)_i = (X_B)_i$.

So, what will the term $p(x,y)$ look like? Well, for a fixed $x \in \{0,1\}^A$, if y disagrees with x in a coordinate $i \in A \cap B$, then $p(x,y)$ is clearly 0. Otherwise, since A and B were both uniformly distributed over their range, $p(x,y) = 2^{|A\cap B|-|A|-|B|}$ and there are exactly $2^{|B|-|A\cap B|}$

such $y \in \{0,1\}^B$ which agree with x on $\{0,1\}^{A\cap B}$. Hence we can calculate

$$\begin{aligned}\mathbb{H}(X_A, X_B) &= \sum_x \sum_y p(x,y) \log\left(\frac{1}{p(x,y)}\right) \\ &= \sum_x 2^{|B|-|A\cap B|} \cdot 2^{|A\cap B|-|A|-|B|} \log 2^{|A|+|B|-|A\cap B|} \\ &= 2^{|A|+|B|-|A\cap B|} \cdot 2^{|A\cap B|-|A|-|B|} \log 2^{|A|+|B|-|A\cap B|} \\ &= |A| + |B| - |A\cap B| = |A \cup B|.\end{aligned}$$

So, in this context the joint entropy corresponds to the cardinality of the union $A \cup B$.

We also define the *conditional entropy* of Y given X in the following way. Let us write, as another shorthand, $p(y|x) := \mathbb{P}(Y = y|X = x)$, and similarly $p(x|y)$. We define

$$\begin{aligned}\mathbb{H}(Y|X) &:= \sum_x p(x) \sum_y p(y|x) \log\left(\frac{1}{p(y|x)}\right) \\ &= \sum_x p(x)\mathbb{H}(Y|X = x) \\ &= \mathbb{E}_x\big(\mathbb{H}(Y|X = x)\big).\end{aligned}$$

Where the first equation is a definition, and the other equalities are merely different ways to rewrite this quantity. Note the difference between $\mathbb{H}(Y|X = x)$, which is the entropy of the random variable $(Y|X = x)$, and $\mathbb{H}(Y|X)$, which is the expected value of the latter over all possible values of x. In particular, $(Y|X)$ is not a random variable.

Back to our example, given subsets A and B and considering $\mathbb{H}(X_B|X_A)$, what will $p(y|x)$ be? Well, as before, given a fixed x, this term is 0 unless x and y agree on $\{0,1\}^{A\cap B}$, and if they do agree on $A \cap B$ then it is clear that $p(y|x) = 2^{-|B\setminus A|}$. Also, for each x, there are exactly $2^{|B|-|A\cap B|} = 2^{|B\setminus A|}$ such y which agree with x on $\{0,1\}^{A\cap B}$. Hence we can calculate

$$\begin{aligned}\mathbb{H}(X_B|X_A) &:= \sum_x p(x) \sum_y p(y|x) \log\left(\frac{1}{p(y|x)}\right) \\ &= \sum_x p(x) 2^{|B\setminus A|} 2^{-|B\setminus A|} \log\left(2^{|B\setminus A|}\right) \\ &= 2^{|A|} 2^{-|A|} \log\left(2^{|B\setminus A|}\right) \\ &= |B \setminus A|.\end{aligned}$$

So, in this context the conditional entropy corresponds to the cardinality of the set difference $B \setminus A$.

We can think of the conditional entropy as being the expected surprise in learning the value of Y, given that the value of X is known. We might expect, heuristically, that having extra knowledge should only decrease how surprised we are, and indeed that turns out to be the case:

Lemma 2.3 (Dropping conditioning)**.** *Let X, Y and Z be discrete random variables. Then*

$$\mathbb{H}(Y|X) \leq \mathbb{H}(Y).$$

Furthermore

$$\mathbb{H}(Y|X, Z) \leq \mathbb{H}(Y|X)$$

with equality if X determines Z.

Proof. We will just prove the first part of the lemma, the second part will be left as an exercise. Noting that $p(y)p(x|y) = p(x)p(y|x) = p(x,y)$, we see that

$$\begin{aligned}
\mathbb{H}(Y|X) &= \sum_x p(x) \sum_y p(y|x) \log\left(\frac{1}{p(y|x)}\right) \\
&= \sum_y p(y) \sum_x p(x|y) \log\left(\frac{1}{p(y|x)}\right) \\
&\leq \sum_y p(y) \log\left(\sum_x \frac{p(x|y)}{p(y|x)}\right) \\
&= \sum_y p(y) \log\left(\sum_x \frac{p(x)}{p(y)}\right) \\
&= \sum_y p(y) \log\left(\frac{1}{p(y)}\right) \\
&= \mathbb{H}(Y).
\end{aligned}$$

Where in the above we make repeated use of the fact that, if we sum the probabilities that a random variable takes a specific value over its entire range,1 then the result is 1, and Jensen's inequality (See Lemma 2.2) in the third line. □

Using our correspondence between the set world and the random variable world, we can now use Lemma 2.3 to say something about sets. Indeed, we have that

$$|B \setminus A| = \mathbb{H}(X_B|X_A) \leq \mathbb{H}(X_B) = |B|.$$

In a similar fashion, any identity or inequality about entropy will specialise to a combinatorial identity or inequality about finite sets. The converse is not true, and we shall see some examples of this later, but sometimes it can give intuition about what identities may hold. For example, we know that $|A \cup B| = |A| + |B \setminus A|$. Translating this back into the language of entropy would give the statement $\mathbb{H}(X_A, X_B) = \mathbb{H}(X_A) + \mathbb{H}(X_B|X_A)$, which we will see in fact holds for all pairs of random variables.

Lemma 2.4 (Chain rule)**.** *Let X and Y be discrete random variables. Then*

$$\mathbb{H}(X,Y) = \mathbb{H}(X) + \mathbb{H}(Y|X).$$

Proof.

$$\begin{aligned}
\mathbb{H}(X,Y) &= \sum_x \sum_y p(x,y) \log\left(\frac{1}{p(x,y)}\right) \\
&= \sum_x \sum_y p(x,y) \log\left(\frac{1}{p(x)p(y|x)}\right) \\
&= \sum_x \sum_y p(x,y) \log\left(\frac{1}{p(x)}\right) + \sum_x \sum_y p(x,y) \log\left(\frac{1}{p(y|x)}\right) \\
&= \sum_x p(x) \log\left(\frac{1}{p(x)}\right) + \sum_x \sum_y p(x)p(y|x) \log\left(\frac{1}{p(y|x)}\right) \\
&= \mathbb{H}(X) + \sum_x p(x) \sum_y p(y|x) \log\left(\frac{1}{p(y|x)}\right) \\
&= \mathbb{H}(X) + \mathbb{H}(Y|X).
\end{aligned}$$

□

One can also define the joint entropy of a sequence of discrete random variables $X_1, X_2, \ldots, X_n$ in a similar way and by induction it follows that

$$\mathbb{H}(X_1, X_2, \ldots, X_n) = \mathbb{H}(X_1) + \mathbb{H}(X_2|X_1) + \ldots \mathbb{H}(X_n|X_1, X_2, \ldots, X_{n-1}).$$

We shall sometimes also refer to this as the *chain rule*. Note that, by Lemma 2.3 and Lemma 2.4 we have that

$$\mathbb{H}(X_1, X_2, \ldots, X_n) \leq \sum_i \mathbb{H}(X_i). \tag{2.1}$$

This seemingly quite simple statement is really quite useful, since it allows us to reduce the calculation of the entropy of a single random variable, to the calculation of many, hopefully simpler, random variables. Often, using this we can turn quite 'global' calculations into 'local' ones which are much easier to deal with.

So far we have an analogue of set union and set difference, so a natural idea would be consider the entropic function corresponding to intersection. Since $|A \cap B| = |A| + |B| - |A \cup B|$ this quantity should be represented by $\mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y)$. We call this the *mutual information of X and Y* and it is denoted by $\mathbb{I}(X;Y)$. Note that, by Lemma 2.4

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

As the name suggests, we can think of this quantity of measuring the amount of information that X and Y share, and indeed this should be the amount of information 'left' from $H(X)$ after we get rid of the information remaining in X once we know Y, $\mathbb{H}(X|Y)$. From Lemma 2.3 if follows that $\mathbb{I}(X;Y) \geq 0$, and in fact by analysing when we get equality in Jensen's inequality one can show that $I(X;Y) = 0$ if and only if X and Y are independent. Hence, the mutual information is in some way a measure of the dependence of the random variables X and Y.

The final definition we will make is the *conditional mutual information of X and Y given Z*, which we write as $\mathbb{I}(X;Y|Z)$. The definition of this quantity is perhaps obvious given the name

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

Naively, Lemma 2.3 and Lemma 2.4 again suggest that this quantity should be non-negative, however strictly to deduce this we will need to prove a conditional version of Lemma 2.4, whose proof we will leave as an exercise.

Lemma 2.5. *Let X, Y and Z be discrete random variables. Then*

$$\mathbb{H}(X, Y|Z) = \mathbb{H}(X|Z) + \mathbb{H}(Y|X, Z).$$

Given Lemma 2.5 and Lemma 2.3 it follows that

$$\mathbb{I}(X; Y|Z) := \mathbb{H}(X|Z) + \mathbb{H}(Y|Z) - \mathbb{H}(X, Y|Z) = \mathbb{H}(Y|Z) - \mathbb{H}(Y|X, Z) \geq 0.$$

3 Information Theory

3.1 Secure Encryption

Suppose we have a set of *messages* $\mathcal{M}$ that we might wish to encrypt and a set of *keys* $\mathcal{K}$ that we can use to encrypt these messages. That is, every pair $m \in \mathcal{M}$ and $k \in \mathcal{K}$ of a message and a key can be used to generate some encrypted text $c \in \mathcal{C}$, or *ciphertext*.

Normally we have some (pseudo)-random method of generating keys $k \in \mathcal{K}$, which determines some random variable K on $\mathcal{K}$. If we also think of the messages as coming from some, independent, random variable M, we can think of an *encryption scheme* for M in a very broad sense as being a pair of random variables K and C, representing the key and the encrypted text such that K and C together determine M. This last condition is just saying that we can decrypt the message given the key and the ciphertext.

A classical encryption scheme would consist of some deterministic function $e : \mathcal{M} \times \mathcal{K} \to \mathcal{C}$ such that for each $k \in \mathcal{K}$ the function $e(\cdot, k) \to \mathcal{C}$ is injective, and then taking $C = e(M, K)$. Moreover, normally K is chosen to be uniform over $\mathcal{K}$.

Our definition above is slightly more general, as it allows for additional randomness in the generation of the ciphertext, and doesn't require the key to be chosen uniformly.

What does it mean for an encryption scheme to be *secure*? We want that someone who doesn't know the key cannot infer any information about the message from the ciphertext. To put this in terms of entropy, we want that there is no mutual information between C and M. Recall that this is equivalent to saying that C and M are independent.

Definition (Perfectly secure encryption scheme)**.** An encryption scheme K, C for M is *perfectly secure* if $I(M; C) = 0$.

There is an obvious example of a perfectly secure encryption scheme which is known as a *one-time pad*. We may assume without loss of generality that the messages are long binary strings, that is $\mathcal{M} = \{0, 1\}^n$. If we take the set of keys to also be $\mathcal{K} = \{0, 1\}^n$, and let K be uniformly distributed independently of M, then we can consider the function $e(m, k) = m + k$ where addition is taken in $\mathbb{Z}_2^n$.

Theorem 3.1. *The one time pad is perfectly secure.*

Proof. It will be sufficient to show that M and $C = M + K$ are independent. Explicitly, we need to check that for any $x, y \in \{0, 1\}^n$

$$\mathbb{P}(M = x \text{ and } C = y) = \mathbb{P}(M = x)\mathbb{P}(C = y).$$

First we note that C is uniformly distributed on $\{0,1\}^n$, since for any $x \in \{0,1\}^n$

$$\begin{aligned}\mathbb{P}(C = x) &= \sum_{(y,z)\colon y+z=x} \mathbb{P}(M = y \text{ and } K = z) \\ &= \sum_y \mathbb{P}(M = y)\mathbb{P}(K = x - y) \\ &= \sum_y \mathbb{P}(M = y)2^{-n} = 2^{-n},\end{aligned}$$

where the second line follows since M and K are independent.

However, then

$$\begin{aligned}\mathbb{P}(M = x \text{ and } C = y) &= \mathbb{P}(M = x)\mathbb{P}(C = y|M = x) \\ &= \mathbb{P}(M = x)\mathbb{P}(K = y - x) \\ &= \mathbb{P}(M = x)2^{-n} \\ &= \mathbb{P}(M = x)\mathbb{P}(C = y).\end{aligned}$$

It follows that $I(M;C) = 0$. □

However this clearly isn't a very efficient method of encryption, since it requires the two parties to share a key which is as large as the message itself. However Shannon showed that this is essentially necessary for a secure encryption scheme, in the sense that, in an perfectly secure encryption scheme the set of keys must contain as least as much information as the messages.

Theorem 3.2. *If K, C is a perfectly secure encryption scheme for M then $\mathbb{H}(K) \geq \mathbb{H}(M)$.*

Proof. Since K, C is a perfectly secure encryption scheme for M, by assumption $I(M;C) = 0$. Furthermore, since K and C together determine M it follows that $\mathbb{H}(M|K,C) = 0$. This can be seen for example from the chain rule since $\mathbb{H}(K,C,M) = \mathbb{H}(K,C)$ (by the comment after Lemma 2.1), and so by the chain rule we have

$$\mathbb{H}(M|K,C) = \mathbb{H}(K,C,M) - \mathbb{H}(K,C) = 0.$$

Hence,

$$\begin{aligned}\mathbb{H}(M) &= I(M;C) - \mathbb{H}(C) + \mathbb{H}(M,C) \\ &= \mathbb{H}(M,C) - \mathbb{H}(C) \\ &\leq \mathbb{H}(M,C,K) - \mathbb{H}(C) \\ &= \mathbb{H}(M,K|C) \\ &= \mathbb{H}(K|C) + \mathbb{H}(M|K,C) \\ &= \mathbb{H}(K|C) \leq \mathbb{H}(K).\end{aligned}$$

□

As a more concrete example, if both M and K are uniformly distributed then Theorem 3.2 says that

$$\log|\mathcal{K}| = \mathbb{H}(K) \geq \mathbb{H}(M) = \log|\mathcal{M}|.$$

That is, $|\mathcal{K}| \geq |\mathcal{M}|$ and so we need at least as many different keys as we have messages.

3.2 Data Compression

3.2.1 Uniquely Decodable and Prefix-free Codes

Suppose we have a set of elements $\mathcal{X}$, for example the alphabet of some language, and we wish to store or transmit an element $x \in \mathcal{X}$ using a binary string. It is clear that we can represent each element of $x \in \mathcal{X}$ by a unique binary string of length $n = \lceil \log |\mathcal{X}| \rceil$ and so we can transmit an arbitrary element by sending at most n bits of information.

However if there is some distribution, given by a random variable X, on $\mathcal{X}$ in which some elements are more likely to appear than others, then it might be that we can exploit this to find an encoding whose length is shorter on average. That is, we might hope to encode more likely values by shorter strings.

Given a set S let us write S^* for the set of finite sequences of elements of S and given $(s_1, s_2, \ldots, s_m) = C \in S^*$ let us write $||C|| = m$ for the length of the sequence. Formally, given a random variable X on $\mathcal{X}$, which we call a *source*, we wish to find an injective function $C : \mathcal{X} \to \{0,1\}^*$ which will minimize

$$\mathbb{E}(||C(X)||).$$

We call such a function an *encoding* of $\mathcal{X}$ (or X). This isn't a particularly interesting problem, since it is relatively clear that greedily assigning the most likely elements of $\mathcal{X}$ to the shortest strings will minimise the expected length.

However, more generally, we may want to transmit messages consisting of strings of elements from $\mathcal{X}$. It is clear that we can extend an encoding $C : \mathcal{X} \to \{0,1\}^*$ to an encoding $C^* : \mathcal{X}^* \to \{0,1\}^*$ in the obvious way by concatenation, however this might cause ambiguities when there are distinct sequences of elements $(x_1, x_2, \ldots, x_m)$ and $(x'_1, x'_2, \ldots, x'_m) \in \mathcal{X}^*$ such that

$$C(x_1)|C(x_2)|\ldots|C(x_m) = C(x'_1)|C(x'_2)|\ldots|C(x'_m).$$

Of course, it is possible to 'solve' this problem by adding an extra character to our encoding that indicates separation between elements, but this requires the use of an extra character, and also increases the length of the encoding of each message. There is however a different way to approach this problem, we say an encoding $C : \mathcal{X} \to \{0,1\}^*$ is *uniquely decodable* if the extension C^* of C via concatenation is injective. Again, a natural question to ask for a random variable X supported on $\mathcal{X}$ is how small can $\mathbb{E}(||C(X)||)$ be for a uniquely decodable C.

There is perhaps an obvious condition which implies that C is uniquely decodable, which is that C is *prefix-free*, that is, there is no $x, x' \in \mathcal{X}$ such that $C(x)$ is a prefix of $C(x')$. If C if prefix free then given a message $(x_1, x_2, \ldots, x_m) \in \mathcal{X}^*$ we can recover this sequence from

$$C^*(x_1, x_2, \ldots, x_m) = C(x_1)|C(x_2)|\ldots|C(x_m)$$

by reading the string from left to right. Prefix-freeness guarantees that the first $C \in \{0,1\}^*$ that is a prefix of this word is indeed $C(x_1)$, and so in a recursive manner we can recover the message $(x_1, x_2, \ldots, x_m)$.

This also highlights another benefit of prefix-free encodings, you can decode them sequentially in this manner, and so given only a prefix of the encoded message you can still decode a prefix of

the message. For a general encoding C you may have to know the entire coded message before being able to recover any of the original message.

Since being prefix-free is a considerably stronger condition than being uniquely decodable, we might expect that the 'best' prefix-free encoding for X is on average longer than the 'best' uniquely decodable encoding, and a natural question to ask would be how much worse? However, it will in fact turn out that the two coincide, and are essentially given by the entropy of X.

Given a source X and an encoding C let us write

$$\ell_C(X) := \mathbb{E}(||C(X)||) = \sum_{x\in\mathcal{X}} \mathbb{P}(X = x)||C(x)||,$$

and let $\mathcal{C} = \{C(x)\colon x \in \mathcal{X}\} \subseteq \{0,1\}^*$ be the range of C.

As useful tool will be Kraft-McMillan inequality.

Lemma 3.3. *[Kraft-McMillan inequality] If C is a uniquely decodable code with range $\mathcal{C}$ then*

$$\sum_{c\in\mathcal{C}} \frac{1}{2^{||c||}} \leq 1.$$

Furthermore if $\{\ell_x\colon x \in \mathcal{X}\}$ is a (multi)-set of numbers such that

$$\sum_{x\in\mathcal{X}} \frac{1}{2^{\ell_x}} \leq 1$$

then there exists a prefix-free code C such that $||C(x)|| = \ell_x$ for all $x \in \mathcal{X}$.

Remark 3.4. *Note that we proved this Theorem for prefix-free codes on the first example sheet. The stronger result tells us that we if we can find a prefix-free encoding to a set of strings with specified lengths, we can in fact find a uniquely decodable encoding to a set of strings of the same length.*

Proof. In fact, we showed on the example sheet that if

$$\sum_{x\in\mathcal{X}} \frac{1}{2^{\ell_x}} \leq 1,$$

then there exists even a prefix-free code C such that $||C(x)|| = \ell_x$ for all $x \in \mathcal{X}$. The proof is essentially to show that a greedy choice always works. Hence, it remains to show the forward implication.

Suppose C is uniquely decodable and let us write

$$S := \sum_{c\in\mathcal{C}} \frac{1}{2^{||c||}}.$$

The idea behind this proof is a variant of something sometimes known as the 'tensor product trick'. Let N be the length of the longest string in $\mathcal{C}$, then, since there are at most 2^ℓ strings of length ℓ, it is clear that for every $\ell \in \mathbb{N}$ the strings of length ℓ contribute at most 1 to the sum S. Hence clearly $S \leq N$, but we wish to show that $S \leq 1$. What we will show is that the

unique decodability of C implies that we can in fact get an inequality of the form $S^k \leq Nk$ for every k, and by taking the limit as $k \to \infty$ we will recover the result.

So, why should this inequality hold? We can write

$$S^k = \left(\sum_{c \in \mathcal{C}} \frac{1}{2^{|c|}}\right)^k = \sum_{c_1, c_2, \ldots, c_k \in \mathcal{C}} \frac{1}{2^{\sum_{i=1}^k ||c_i||}}.$$

Let us think about $\mathcal{C}^k$, the set of all sequences $(c_1, c_2, \ldots, c_k)$ of length k from $\mathcal{C}$. Given such a sequence we can consider the concatenation $c_1|c_2|\ldots|c_k \in \{0,1\}^*$, and since C is uniquely decodable the mapping $(c_1, c_2, \ldots, c_k) \mapsto c_1|c_2|\ldots|c_k$ is injective. Hence, since $||c_1|c_2|\ldots|c_k|| = \sum_{i=1}^k ||c_i||$,

$$\begin{aligned} S^k &= \sum_{c_1, c_2, \ldots, c_k \in \mathcal{C}} \frac{1}{2^{\sum_{i=1}^k ||c_i||}} \\ &= \sum_{(c_1, c_2, \ldots, c_k) \in \mathcal{C}^k} \frac{1}{2^{||c_1|c_2|\ldots|c_k||}} \\ &\leq kN \end{aligned}$$

where the last inequality holds again because the set of strings of length ℓ for any fixed ℓ give a contribution of at most 1 to the sum, since the mapping $(c_1, c_2, \ldots, c_k) \mapsto c_1|c_2|\ldots|c_k$ is injective, and furthermore all strings have length at most kN. □

Using this inequality it is relatively simple to show the following bound on the average length of an encoding, originally proved by Shannon.

Theorem 3.5. *[Shannon] For any source X and uniquely decodable encoding C*

$$\mathbb{H}(X) \leq \ell_C(X)$$

and furthermore for any source X there exists a prefix-free C such that

$$\ell_C(X) \leq H(X) + 1.$$

Proof. By definition we have that

$$\begin{aligned} \mathbb{H}(X) - \ell_C(X) &= -\sum_{x \in \mathcal{X}} p(x) \log p(x) - \mathbb{E}(||C(X)||) \\ &= -\sum_{x \in \mathcal{X}} p(x) \left(\log p(x) + ||C(x)||\right) \\ &= \sum_{x \in \mathcal{X}} p(x) \log \left(\frac{1}{2^{||C(x)||} p(x)}\right) \end{aligned}$$

By Jensen's inequality we can put the sum on the inside

$$\mathbb{H}(X) - \ell_C(X) \leq \log \left(\sum_{x \in \mathcal{X}} p(x) \frac{1}{2^{||C(x)||} p(x)}\right) = \log \left(\sum_{x \in \mathcal{X}} \frac{1}{2^{||C(x)||}}\right) = \log \left(\sum_{c \in \mathcal{C}} \frac{1}{2^{||c||}}\right)$$

and finally by the Lemma 3.3 we can conclude that

$$\mathbb{H}(X) - \ell_C(X) \leq \log 1 = 0.$$

For the upper bound let us define, for every $x \in \mathcal{X}$

$$\ell_x = \left\lceil \log\left(\frac{1}{p(x)}\right) \right\rceil.$$

It follows that

$$\sum_{x\in\mathcal{X}} \frac{1}{2^{\ell_x}} \leq \sum_{x\in\mathcal{X}} p(x) = 1.$$

Hence by Lemma 3.3 there exists a prefix-free code C such that $||C(x)|| = \ell_x$ for all $x \in \mathcal{X}$. However this code now satisfies

$$\begin{aligned}
\ell_C(X) &= \sum_{x\in\mathcal{X}} p(x)||C(x)|| \\
&= \sum_{x\in\mathcal{X}} p(x)\ell_x \\
&\leq \sum_{x\in\mathcal{X}} p(x)\left(\log\left(\frac{1}{p(x)}\right) + 1\right) \\
&= \mathbb{H}(X) + \sum_{x\in\mathcal{X}} p(x) \\
&= \mathbb{H}(X) + 1.
\end{aligned}$$

□

3.2.2 Huffman Codes

Given a source X Lemma 3.3 tell us that

$$\min_{C \text{ prefix-free}} \ell_C(X) = \min_{C \text{ uniquely decodable}} \ell_C(X)$$

and Theorem 3.5 gives us a way to construct a prefix-free code C with

$$\ell_C(X) \leq \mathbb{H}(X) + 1 \leq \min_{C \text{ prefix-free}} \ell_C(X) + 1.$$

However these codes in general will not be optimal. In turns out one can give an explicit description of an optimal code, which is called a *Huffman code* after David Huffman who discovered them.

One way to view the construction is as follows: We will build a subgraph of the binary tree with leaves corresponding to $\mathcal{C}$. We start by taking $|\mathcal{X}|$ independent vertices, labelled with the value $p(x)$. In the first stage we take the two vertices with the smallest labels $p(x)$ and $p(x')$ and join them both to a new vertex which we label with $p(x) + p(x)'$ and consider it as their parent. In a general step we consider all vertices in the forest with no parents, we choose the two with the smallest labels ℓ_1 and ℓ_2 and we join them both to a new parent vertex with label $\ell_1 + \ell_2$. This continues until the vertex set is connected.

The resulting graph is clearly a subgraph of the binary tree with $|\mathcal{X}|$ leaves, and by arbitrarily choosing a $\{0,1\}$ labelling of the pair of edges from each vertex to its children we can assign to each leaf a string $C(x) \in \{0,1\}^*$. Note that, by construction the code C is prefix-free.

3.3 Guessing From Partial Information

Suppose there is some random variable X whose value we wish to know, but we are only given some partial information in the form of a random variable Y and we have to make a 'guess' of the value of X based on Y. That is, we have some guessing function g which gives us a random variable $g(Y) = \hat{X}$, and we are interested in 'how accurate' our guess is.

For example the random variable X might be a message sent to us across a 'noisy channel', where the message we receive is Y. We are interested in how likely our guess is to be correct, that is, $\mathbb{P}(X \neq \hat{X}) := p_e$.

It turns out that the entropy function gives us a lower bound on how good our guessing function can be in terms of $\mathbb{H}(X|Y)$.

Theorem 3.6. *[Fano's inequality] Let X and Y be discrete random variables taking values in $\mathcal{X}$ and $\mathcal{Y}$ respectively,and let $g : \mathcal{Y} \to \mathcal{X}$ be any function. If we define $\hat{X} = g(Y)$ and $p_e = \mathbb{P}(X \neq \hat{X})$ as above then*

$$h(p_e) + p_e \log(|\mathcal{X}| - 1) \geq \mathbb{H}(X|Y).$$

In particular, since $h(p_e) \leq 1$

$$p_e \geq \frac{\mathbb{H}(X|Y) - 1}{\log(|\mathcal{X}| - 1)}.$$

Proof. Let us define a random variable S which takes the value 1 if $X \neq \hat{X}$ and 0 if $X = \hat{X}$. By assumption $\mathbb{P}(S = 1) = p_e$ and $\mathbb{P}(S = 0) = 1 - p_e$ and so $\mathbb{H}(S) = h(p_e)$. Now $\mathbb{H}(X|Y) = \mathbb{H}(E, X|Y) - \mathbb{H}(E|X, Y)$ but E is determined by X and Y, since $\hat{X}$ is a function of Y. Hence $\mathbb{H}(X|Y) = \mathbb{H}(E, X|Y) - 0 = \mathbb{H}(E, X|Y)$.

Now $\mathbb{H}(E, X|Y) = \mathbb{H}(E|Y) + \mathbb{H}(X|E, Y) \leq \mathbb{H}(E) + \mathbb{H}(X|E, Y)$. However by the definition of conditional entropy we can split $\mathbb{H}(X|E, Y)$ into two parts

$$\begin{aligned} H(X|E, Y) &= \mathbb{P}(E = 0)H(X|Y, E = 0) + \mathbb{P}(E = 1)H(X|Y, E = 1) \\ &= (1 - p_e)H(X|Y, E = 0) + p_e H(X|Y, E = 1). \end{aligned}$$

However, if $E = 0$ then $X = \hat{X}$ and so, conditioned on $E = 0$, X is determined by Y. Hence the first term is 0, and since

$$\mathbb{H}(X|Y, E = 1) = \mathbb{E}_y \mathbb{H}(X|Y = y, E = 1)$$

we can bound the second term by $p_e \log(|\mathcal{X}| - 1)$ since if we fix $y \in \mathcal{Y}$ then $(X|Y = y, E = 1)$ can take any value in $\mathcal{X}$ except $g(y)$. It follows that

$$\mathbb{H}(X|Y) \leq \mathbb{H}(E) + \mathbb{H}(X|E, Y) \leq h(p_e) + p_e \log(|\mathcal{X}| - 1).$$

□

3.4 Shannon's Channel Coding Theorem

A *channel* is a way to model the transmission of some message. We have some set $\mathcal{X}$ of messages which are to be transmitted and a set $\mathcal{Y}$ of possible outputs and the channel is modelled by a conditional probability distribution $p_{Y|X}$. That is, for each $x \in \mathcal{X}$ we have a distribution $p_{Y|X}(\cdot|x)$ on $\mathcal{Y}$ which is the distribution of the output of the channel when x is the input.

Two simple examples are the binary symmetric channel, where $\mathcal{X} = \mathcal{Y} = \{0,1\}$ and each message has a ε chance of resulting in the wrong output, or the binary erasure channel where $\mathcal{X} = \{0,1\}$ and $\mathcal{Y} = \{0,1,\perp\}$ and each message has $(1-\varepsilon)$ chance of being transmitted correctly and a ε chance of being 'lost', and outputting $\perp$.

Given a particular channel $p_{Y|X}$ we are interested in how much information we can reliably transmit across the channel. In order to do this we have a *source* which we wish to transmit. which we will imagine comes as a sequence $(U_1, \dots, U_k)$ of independent, uniform binary random variable. In order to transmit this source over the channel we first have to encode it using some encoding function $e : \{0,1\}^k \to \mathcal{X}^n$ for some n. Note that we are allowed to take $n \geq k$, and in this way we might hope to increase the accuracy of our message by sending a sequence with some error-correcting properties, however this comes at the cost of sending more message over the channel.

The encoded sequence $(X_1, \dots, X_n) = e(U_1, \dots, U_k)$ is transmitted over the channel, with output $(Y_1, \dots, Y_n)$ and we then decode this sequence using some decoding function $d : \mathcal{Y}^n \to \{0,1\}^k$ to get a sequence $(\hat{U}_1, \dots, \hat{U}_k)$. An *encoding scheme* for a source $(U_1, \dots, U_k)$ and a channel $p_{Y|X}$ is a pair (e, d) of an encoding and decoding function. We define the *rate* of an encoding scheme to be $R = \frac{k}{n}$, the number of bits communicated per message sent over the channel.

Finally we would like a way to measure how close our decoded sequence $(\hat{U}_1, \dots, \hat{U}_k)$ is to the source $(U_1, \dots, U_k)$. For this purpose we define the *error probability* as

$$p_e := \mathbb{P}(\text{There exists } i \text{ with } \hat{U}_i \neq U_i).$$

Given channel we say a rate R is *achievable* if for each source there exists an encoding scheme with rate R such that $p_e \to 0$ as the length of the source tends to ∞. The *capactity* of a channel is the supremum over all achievable rates.

Theorem 3.7 (Shannon's Channel Coding Theorem)**.** *For any channel $p_{Y|X}$ the capacity of the channel is equal to $\max_{p_X} I(X;Y)$.*

Proof. Let $C := \max_{p_X} I(X;Y)$, we need to show two statements. Firstly if we have an arbitrary $\varepsilon > 0$ then we can find an encoding scheme for each source with rate $C - \varepsilon$ such that $p_e \to 0$ and secondly, again for an arbitrary $\varepsilon > 0$, the rate $C + \varepsilon$ is not achievable.

Let us begin by showing the second. Suppose that we have an encoding scheme with rate R and consider a source of length k. If the error probability is low, then we should expect that $\hat{U}_i$ is distributed very similarly to U_i for each i, and since $\hat{U}_i$ is determined by $Y_{[n]}$, this would imply that we can almost recreate $U_{[k]}$ from $Y_{[n]}$, and so their mutual information must be high. We can make this assertion precise with Fano's inequality.

However, we have a Markov chain $U_{[k]} \to X_{[n]} \to Y_{[n]}$, and so as we will see on the example sheet, the mutual information between $U_{[k]}$ and $Y_{[n]}$ is at most the mutual information between $X_{[n]}$ and $Y_{[n]}$, which by assumption is at most nC. Our aim will be to show that if n is small compared to k, this will contradict the bound given by Fano's inequality.

So, let's try to make this precise. By the chain rule

$$\mathbb{H}(X_{[n]}, Y_{[n]}, U_{[k]}) = \mathbb{H}(X_{[n-1]}, Y_{[n-1]}, U_{[k]}) + \mathbb{H}(X_n | X_{[n-1]}, Y_{[n-1]}, U_{[k]}) + \mathbb{H}(Y_n | X_{[n]}, Y_{[n-1]}, U_{[k]}).$$

However, X_n is determined by $U_{[k]}$, and hence the middle term is 0, and Y_n is independent of $(X_{[n-1]}, Y_{[n-1]}, U_{[k]})$ conditioned on X_n and hence

$$\mathbb{H}(X_{[n]}, Y_{[n]}, U_{[k]}) = \mathbb{H}(X_{[n-1]}, Y_{[n-1]}, U_{[k]}) + \mathbb{H}(Y_n | X_n).$$

It follows by induction that $\mathbb{H}(X_{[n]}, Y_{[n]}, U_{[k]}) = \mathbb{H}(U_{[k]}) + \sum_{i=1}^n \mathbb{H}(Y_i|X_i)$. However, again since $X_{[n]}$ is determined by $U_{[k]}$, $\mathbb{H}(X_{[n]}, Y_{[n]}, U_{[k]}) = \mathbb{H}(Y_{[n]}, U_{[k]})$ and so

$$\mathbb{H}(Y_{[n]}, U_{[k]}) = \mathbb{H}(U_{[k]}) + \sum_{i=1}^n \mathbb{H}(Y_i|X_i).$$

Therefore

$$I(U_{[k]}; Y_{[n]}) = \mathbb{H}(Y_{[n]}) - \sum_{i=1}^n \mathbb{H}(Y_i|X_i) \leq \sum_{i=1}^n (\mathbb{H}(Y_i) - (Y_i|X_i)) = \sum_{i=1}^n I(X_i; Y_i) \leq n.C$$

Hence, since the U_is are independent and uniform,

$$\mathbb{H}(U_{[k]}|Y_{[n]}) = \mathbb{H}(U_{[k]}) - I(U_{[k]}; Y_{[n]}) \geq k - nC$$

However, if we consider Fano's inequality (Theorem 3.6), applied to the random variables $U_{[k]}, Y_{[n]}$ and the decoding function $d : \mathcal{Y} \to \{0,1\}^k$ then we see that

$$p_e \geq \frac{\mathbb{H}(U_{[k]}|Y_{[n]})}{k-1} \geq \frac{k - nC}{k} = 1 - \frac{C}{R}$$

Hence, if $R > C + \varepsilon$, then it follows that p_e is bounded away from 0.

For the converse, suppose we are given an arbitrary $\varepsilon > 0$, we wish to find an encoding scheme with rate $R = C - \varepsilon$. We will show using the probabilistic method that such an encoding scheme must exist.

The key idea here is that of *ε-typical sequences*. Suppose we have a random variable X taking values in $\mathcal{X}$. Given $n \in \mathbb{N}$ and $(x_1, \ldots, x_n) \in \mathcal{X}^n$ let us write

$$p(x_1, \ldots, x_n) = \prod_{i=1}^n \mathbb{P}(X_i = x_i)$$

where $X_1, \ldots, X_n$ are independent and identically distributed as X. Suppose we sample points in $\mathcal{X}^n$ according to the product distribution, what does a typical point look like?

Well, for each $x \in \mathcal{X}$ we expect there to be $np(x)$ many $x_i = x$. This means that a typical point $(x_1, \ldots, x_n) \in \mathcal{X}^n$ should have probability about

$$p(x_1, \ldots, x_n) = \prod_{x \in \mathcal{X}} p(x)^{np(x)}$$

and so the information theoretic content of this point, that is the contribution to the entropy $\mathbb{H}(X_{[n]})$ will be

$$-\log\left(\frac{1}{p(x_1, \ldots, x_n)}\right) = -\log\left(\frac{1}{\prod_{x \in \mathcal{X}} p(x)^{np(x)}}\right) = n \sum_{x \in \mathcal{X}} p(x) \log\left(\frac{1}{p(x)}\right) = n\mathbb{H}(X).$$

Note that $\mathbb{H}(X_{[n]}) = n\mathbb{H}(X)$, and so, since the entropy is just the average over points y in the range of $\log \frac{1}{p(y)}$, these points contribute approximately the correct amount to this average. Rearranging the above, we could equivalently say

$$p(x_1, \ldots, x_n) = 2^{-n\mathbb{H}(X)}.$$

We say a sequence $(x_1, \ldots, x_n) \in \mathcal{X}^n$ is *ε-typical* if

$$2^{-n(\mathbb{H}(X)+\varepsilon)} \leq p(x_1, \ldots, x_n) \leq 2^{-n(\mathbb{H}(X)-\varepsilon)}.$$

Let us denote the set of ε-typical sequences in $\mathcal{X}^n$ by $A_\varepsilon^n(X)$.

Let us note some properties of $A_\varepsilon^n(X)$

- If $(x_1, \ldots, x_n) \in A_\varepsilon^n(X)$ then $\mathbb{H}(X) - \varepsilon \leq -\frac{1}{n}\log(p(x_1, \ldots, x_n)) \leq \mathbb{H}(X) + \varepsilon$;
- $p(A_\varepsilon^n(X)) > 1 - o(n)$;
- $|A_\varepsilon^n(X)| \leq 2^{n\mathbb{H}(X)+\varepsilon}$;
- $|A_\varepsilon^n(X)| \geq (1 - o(n))2^{n\mathbb{H}(X)-\varepsilon}$.

The first is just a restatement of the definition and the second follows from the law of large numbers. More precisely the (weak) law of large numbers says that if X is a real random variable with mean μ and $X_1, X_2, \ldots, X_n$ are independently distributed as X then $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \to \mu$ in probability. That is, $\mathbb{P}(|\overline{X}_n - \mu| > \varepsilon) \to 0$ as $n \to \infty$ for any fixed $\varepsilon > 0$. To see why this implies the second consider the random variable given by $Y = \frac{1}{p(X)}$. We have that $\mathbb{E}(Y) = \mathbb{H}(X)$ by definition and if $X_1, \ldots, X_n$ are independently distributed as X then $Y_i = \frac{1}{p(X_i)}$ are independently distributed as Y and

$$\begin{aligned}
\mathbb{P}((X_1, \ldots, X_n) \notin A_\varepsilon^n(X)) &= \mathbb{P}\left(|-\frac{1}{n}\log(p(X_1, \ldots, X_n)) - \mathbb{H}(X)| > \varepsilon\right) \\
&= \mathbb{P}\left(|-\frac{1}{n}\sum_{i=1}^n \log(p(X_i)) - \mathbb{H}(X)| > \varepsilon\right) \\
&= \mathbb{P}\left(|\frac{1}{n}\sum_{i=1}^n Y_i - \mathbb{E}(Y)| > \varepsilon\right) \to 0.
\end{aligned}$$

For the third we note that

$$1 \geq p(A_\varepsilon^n(X)) = \sum_{(x_1,\ldots,x_n)\in A_\varepsilon^n(X)} p(x_1,\ldots,x_n) \geq |A_\varepsilon^n(X)|2^{-n(\mathbb{H}(X)+\varepsilon)},$$

from which the claim follows. Similarly again by the law of large numbers we may take n large enough that

$$1-\varepsilon \leq p(A_\varepsilon^n(X)) = \sum_{(x_1,\ldots,x_n)\in A_\varepsilon^n(X)} p(x_1,\ldots,x_n) \leq |A_\varepsilon^n(X)|2^{-n(\mathbb{H}(X)-\varepsilon)}.$$

Note that typical sequences are not in general the most likely ones! For example if X is a biased binary variable, say with success probability $2/3$ then the most likely sequence is $(1,1,\ldots,1)$, but this is far from typical. In fact a typical sequence will have roughly 2/3rds of it's entries equal to 1.

Similarly given X as above and Y taking values in $\mathcal{Y}$ we can define *ε-jointly-typical sequences* for X and Y is a sequence $(x_1,\ldots,x_n,y_1,\ldots,y_n)$ such that

- $2^{-n(\mathbb{H}(X)+\varepsilon)} \leq p(x_1,\ldots,x_n) \leq 2^{-n(\mathbb{H}(X)-\varepsilon)}$;
- $2^{-n(\mathbb{H}(Y)+\varepsilon)} \leq p(y_1,\ldots,y_n) \leq 2^{-n(\mathbb{H}(Y)-\varepsilon)}$;
- $2^{-n(\mathbb{H}(X,Y)+\varepsilon)} \leq p(x_1,\ldots,x_n,y_1,\ldots,y_n) \leq 2^{-n(\mathbb{H}(X,Y)-\varepsilon)}$;

where $p(x_1,\ldots,x_n,y_1,\ldots,y_n) = \prod_{i=1}^n \mathbb{P}((X_i,Y_i)=(x_i,y_i))$ where each (X_i,Y_i) are independent and distributed as the joint distribution of (X,Y). We denote the set of ε-jointly-typical sequences in $(\mathcal{X}^n,\mathcal{Y}^n)$ as $A_\varepsilon^n(X,Y)$.

Let us note some properties of $A_\varepsilon^n(X,Y)$

- If $(x_1,\ldots,x_n,y_1,\ldots,y_n)\in A_\varepsilon^n(X,Y)$ then
  - $\mathbb{H}(X)-\varepsilon \leq -\frac{1}{n}\log(p(x_1,\ldots,x_n)) \leq \mathbb{H}(X)+\varepsilon$;
  - $\mathbb{H}(Y)-\varepsilon \leq -\frac{1}{n}\log(p(y_1,\ldots,y_n)) \leq \mathbb{H}(Y)+\varepsilon$
  - $\mathbb{H}(X,Y)-\varepsilon \leq -\frac{1}{n}\log(p(x_1,\ldots,x_n,y_1,\ldots,y_n)) \leq \mathbb{H}(X,Y)+\varepsilon$
- $p(A_\varepsilon^n(X,Y)) > 1-o(n)$ for n sufficiently large;
- $|A_\varepsilon^n(X,Y)| \leq 2^{n\mathbb{H}(X,Y)+\varepsilon}$;
- $|A_\varepsilon^n(X,Y)| \geq (1-o(n))2^{n\mathbb{H}(X,Y)-\varepsilon}$ for n sufficiently large.

Finally, we will need the following lemma, whose proof we sketch

Lemma 3.8. *Let $(\hat{X}_i,\hat{Y}_i)$ be independent and distributed with $p_{\hat{X}_i,\hat{Y}_i} = p_X\cdot p_Y$, then for all $\varepsilon>0$ there is an n_0 such that for all $n>n_0$*

$$(1-o(n))2^{-n(I(X;Y)+3\varepsilon)} \leq \mathbb{P}((\hat{X}_1,\ldots,\hat{X}_n,\hat{Y}_1,\ldots,\hat{Y}_n)\in A_\varepsilon^n(X,Y)) \leq 2^{-n(I(X;Y)-3\varepsilon)}$$

Proof. Well $\mathbb{P}((\hat{X}_1,\dots,\hat{X}_n,\hat{Y}_1,\dots,\hat{Y}_n)\in A_\varepsilon^n(X,Y))$ is just the sum over all ε-jointly-typical sequences $(x_1,\dots,x_n,y_1,\dots,y_n)$ of the probability that $(\hat{X}_1,\dots,\hat{X}_n,\hat{Y}_1,\dots,\hat{Y}_n)=(x_1,\dots,x_n,y_1,\dots,y_n)$, and there are approximately $2^{n\mathbb{H}(X,Y)}$ many such sequences.

However, since the $\hat{X}_i$ and the $\hat{Y}_i$ are independent, this is just the product of the probabilities that $(\hat{X}_1,\dots,\hat{X}_n)=(x_1,\dots,x_n)$ and $(\hat{Y}_1,\dots,\hat{Y}_n))=(y_1,\dots,y_n)$. However, since $\hat{X}_i\sim X$ and $\hat{Y}_i\sim Y$ it follows from the fact that $(x_1,\dots,x_n,y_1,\dots,y_n)$ is ε-jointly-typical that these probabilities are approximately $2^{-n\mathbb{H}(X)}$ and $2^{-n\mathbb{H}(Y)}$ respectively. That is

$$\begin{aligned}\mathbb{P}((\hat{X}_1,\dots,\hat{X}_n,\hat{Y}_1,\dots,\hat{Y}_n) &\approx \sum_{(x_1,\dots,x_n,y_1,\dots,y_n)\in A_\varepsilon^n(X,Y)} 2^{-n\mathbb{H}(X)}2^{-n\mathbb{H}(Y)}\\ &\approx 2^{n\mathbb{H}(X,Y)}2^{-n\mathbb{H}(X)}2^{-n\mathbb{H}(Y)}\\ &= 2^{-nI(X;Y)}.\end{aligned}$$

□

So, given a channel, let $p=p_X$ be the distribution on X optimising $I(X;Y)$, recall that we wish to show that the rate $I(X;Y)-4\varepsilon$ is achievable, so let us pick sufficiently large k and n with $I(X;Y)-4\varepsilon\leq R=\frac{k}{n}<I(X;Y)-3\varepsilon$.

We pick our encoding function $e:\{0,1\}^k\to\mathcal{X}^n$ by choosing the image of each $s\in\{0,1\}^k$ at random from $\mathcal{X}^n$ with distribution given by the product distribution of p. That is, for every s individually and every $(x_1,\dots,x_n)$ the probability that $e(s)=(x_1,\dots,x_n)$ is $\prod_i p(x_i)$.

We then define a decoding function as follow. Given output $(y_1,\dots,y_n)$ we look at all the sequences $(x_1,\dots,x_n,y_1,\dots,y_n)$ where $(x_1,\dots,x_n)=e(s)$ for some $s\in\{0,1\}^k$. Now, if $e(s)=(x_1,\dots,x_n)$ were the input that resulted in the output $(y_1,\dots,y_k)$, since each pair (x_i,y_i) was chosen according to the distribution (X,Y) and so we should expect this sequence to be ε-jointly-typical with very high probability, so a sensible decoding function would map $(y_1,\dots,y_k)$ to s. Of course, there might be no suitable s, or there might be more than one of them, but we shall see that this happens so rarely for it not to be a problem.

So our decoding function d be

$$d(y_1,\dots,y_n)=\begin{cases}s \text{ if } (e(s),y_1,\dots,y_n)\in A_\varepsilon^n(X,Y) \text{ and } (e(s'),y_1,\dots,y_n)\notin A_\varepsilon^n(X,Y) \text{ for all } s\neq s'\in\{0,1\}^k\\ \text{error} \qquad \text{otherwise}\end{cases}$$

So, we have picked a random encoding function e and that determines our decoding function d. We would like to say that, with high probability, (e,d) show that the desired rate is achievable. That is, since the probability of error p_e is now a function of the encoding function e, we would like to say that it is likely that for a randomly chosen e the error probability is small. So, let us try and estimate the expected probability of error $\mathbb{E}(p_e(e))$. If we can show this tends to 0 with n then by the first moment method we know there is **some** function e for which this holds.

Since the source is uniform, for a given e we just need to calculate the expected probability that we transmit a particular string s correctly. Furthermore, since we chose the values of e

at random according to p independently for each $s \in \{0,1\}^k$ we want to estimate the following quantity:

Given a family of random variables $(X_1(s'), \ldots, X_n(s'))$ for all $s' \in \{0,1\}^k$ each independently distributed as $X_{[n]}$ and a random variable $(Y_1(s), \ldots, Y_n(s))$ where each $Y_i(s)$ is distributed as $(Y|X = X_i(s))$ (and independently of the other $X_i(s')$) we wish to bound from above the probability that one of our two bad events occur. That is either,

- $(X_1(s), \ldots, X_n(s), Y_1(s), \ldots, Y_n(s)) \notin A_\varepsilon^n(X,Y)$, or
- $(X_1(s'), \ldots, X_n(s'), Y_1(s), \ldots, Y_n(s)) \in A_\varepsilon^n(X,Y)$ for all $s \neq s' \in \{0,1\}^k$.

By the union bound this is at most the sum of the probabilities. Hence

$$\begin{aligned}\mathbb{E}(p_e(e)) \leq{}& \mathbb{P}\left((X_1(s), \ldots, X_n(s), Y_1(s), \ldots, Y_n(s)) \notin A_\varepsilon^n(X,Y)\right) \\ &+ \sum_{s' \neq s} \mathbb{P}\left((X_1(s'), \ldots, X_n(s'), Y_1(s), \ldots, Y_n(s)) \in A_\varepsilon^n(X,Y)\right)\end{aligned}$$

The first term can clearly be bounded above by the definition of ε-jointly-typical sequences. However, since the values of e were chosen independently, $(Y_1(s), \ldots, Y_n(s))$ is independent of $(X_1(s'), \ldots, X_n(s'))$ for all $s \neq s'$. Hence Lemma 3.8 gives an upper bound for the probability of the terms in the sum. Putting this all together we get

$$\begin{aligned}\mathbb{E}(p_e(e)) &\leq o(n) + \sum_{s' \neq s} 2^{-n(I(X;Y)-3\varepsilon)} \\ &= o(n) + 2^{k-1} 2^{-n(I(X;Y)-3\varepsilon)} \\ &\leq o(n) + 2^{nR} 2^{-n(I(X;Y)-3\varepsilon)} \\ &= o(n) + 2^{-n(I(X;Y)-R-3\varepsilon)}\end{aligned}$$

Where we used that $k - 1 \leq k = nR$. However since $R < I(X;Y) - 3\varepsilon$ it follows that the second term also tends to 0 with n.

□

4 Combinatorial Applications

4.1 Brégman's Theorem

The *permanent* of an $n \times n$ matrix A is

$$\text{perm}(A) = \sum_{\sigma \in S_n} \prod_{i=1}^{n} a_{i\sigma(i)}$$

where S_n is the set of permutations of $[n]$. Note that this is very close to the definition of $\det(A)$, only with the factor of $(-1)^{\text{sgn}(\sigma)}$ removed. Given a 0/1 matrix A we should expect that we can bound the permanent in terms of the number of non-zero entries of A in some way. In 1963 Minc gave a very natural conjecture for a bound given the row sums.

Conjecture 4.1 (Minc's Conjecture)**.** *Let A be an $n \times n$ 0/1 matrix such that the sum of the entries of the ith row is r_i. Then*

$$perm(A) \leq \prod_{i=1}^{n} (r_i!)^{\frac{1}{r_i}}.$$

It turns out this conjecture can be very easily transformed into an equivalent conjecture about graphs. There is a natural correspondence between $n \times n$ 0/1 matrices and bipartite graphs with partition classes of size n. Given such a matrix A we can consider a graph G on vertex set (V, W) where $V = \{v_1, \ldots, v_n\}$ and $W = \{w_1, \ldots, w_n\}$ with an edge between v_i and w_j if and only if $a_{ij} = 1$.

Now, a permutation σ gives a non-zero contribution to $\text{perm}(A)$ if and only if $a_{i\sigma(i)} = 1$ for all $i \in [n]$, that is, if and only if $(v_i, w_{\sigma(i)})$ is an edge for every $i \in [n]$. However, since σ is injective, $\{(v_i, w_{\sigma(i)}) \colon i \in [n]\}$ gives a perfect matching of G. Conversely, any perfect matching M of G determines a permutation σ of $[n]$ given by $\sigma(i) = j$ such that $(v_i, w_j) \in M$, and the contribution of this permutation to $\text{perm}(A)$ is non-zero. Putting this together we see that if we write $\Phi(G)$ for the set of perfect matchings of G and $\phi(G) = |\Phi(G)|$ then

$$\text{perm}(A) = \phi(G).$$

Since the row sums of A are precisely the degrees of vertices in V, an upper bound on the permament of A in terms of the row sums is equivalent to an upper bound on the number of perfect matchings of G in terms of the degrees of vertices in one partition class. Minc's conjecture was proved by Brégman's, and so is now known as Brégman's Theorem, but we will give a proof using entropy methods due to Radhakrishnan.

Theorem 4.2 (Brégman's Theorem)**.** *Let G be a bipartite graph on vertex classes A and B such that $|A| = |B| = n$. Then*

$$\phi(G) \leq \prod_{v \in A} (d(v)!)^{\frac{1}{d(v)}}.$$

Proof. Let M be a perfect matching of G chosen uniformly at random from $\Phi(G)$. For convenience we will associate A with the set $[n]$ in the natural way, and denote by d_i the degree of the vertex i. For each $i \in [n]$ let X_i be the neighbour of i in M and we identify

M with $X = (X_1, X_2, \ldots, X_n)$. More precisely, since M determines and is determined by $(X_1, X_2, \ldots, X_n)$ it follows that $\mathbb{H}(M) = \mathbb{H}(X)$.

Since M is unifromly distributed over $\phi(G)$ possibilities we have that $\mathbb{H}(M) = \mathbb{H}(X) = \log(\phi(G))$. Hence if we can bound $\mathbb{H}(X)$ from above, we can also bound $\phi(G)$. Note that to get the stated bound we would need to show that

$$\mathbb{H}(X) \leq \sum_{i=1}^{n} \frac{\log(d_i!)}{d_i}.$$

A naive first approach might be the use the sub-additivity of entropy to say

$$\mathbb{H}(X) \leq \sum_{i=1}^{n} H(X_i),$$

and since there are at most d_i possibilities for the random variable X_i we have that

$$\mathbb{H}(X) \leq \sum_{i=1}^{n} \mathbb{H}(X_i) \leq \sum_{v \in A} \log(d_i).$$

However, by Stirling's approximation, $\log(d_i!)/d_i \sim \log(d_i/e)$, and so this bound is not enough. However perhaps we can improve this bound by using the chain rule, since we have

$$\mathbb{H}(X) = \sum_{i=1}^{n} \mathbb{H}(X_i | X_1, X_2, \ldots X_{i-1}).$$

We can think of this as revealing the matching one edge at a time, and working out the remaining entropy at each step given what we know. Now instead of just using the naive bound for each X_i we can hopefully take into account the fact that, if we already know $X_1, , X_2, \ldots X_{i-1}$ this may reduce the number of possibilities for X_i, since some of the vertices $1, 2, \ldots, i-1$ may be matched to neighbours of i in M, reducing the range of X_i.

However, since the matching M was random and the ordering of A were arbitrary, we don't know how many neighbours of i have already been used in M by vertices $j < i$. However, given any permutation σ of $[n]$ we can apply the chain rule with respect to the ordering given by σ to see

$$\mathbb{H}(X) = \sum_{i=1}^{n} \mathbb{H}(X_{\sigma(i)} | X_{\sigma(1)}, X_{\sigma(2)}, \ldots X_{\sigma(i-1)}).$$

For each matching M there are some orderings that will give a significant improvement on the bound above, so if we average over all possible choices of σ

$$\mathbb{H}(X) \leq \frac{1}{n!} \sum_{\sigma} \sum_{i=1}^{n} \mathbb{H}(X_{\sigma(i)} | X_{\sigma(1)}, X_{\sigma(2)}, \ldots X_{\sigma(i-1)}),$$

we might hope to get a reasonable improvement in our bound.

For each $i \in [n]$ and permutation σ let us write $J_{\sigma,i} = \{k \colon \sigma(k) < \sigma(i)\} \subseteq [n] \setminus \{i\}$. Each term in the sum above is of the form $\mathbb{H}(X_i | X_{J_{\sigma,i}})$. So we can re-write the sum as

$$\begin{aligned}\mathbb{H}(X) &\leq \frac{1}{n!}\sum_{\sigma}\sum_{i=1}^{n}\mathbb{H}(X_{\sigma(i)}|X_{\sigma(1)}, X_{\sigma(2)}, \dots X_{\sigma(i-1)}) \\ &= \frac{1}{n!}\sum_{i=1}^{n}\sum_{\sigma}\mathbb{H}(X_i|X_{J_{\sigma,i}})\end{aligned}$$

For each of these terms, if we think about calculating the sequentially, we've reduced the range of X_i by how many of the previously exposed X_j lie in $N(i)$, the neighbourhood of i. For each fixed value of $X_{J_{\sigma,i}}$, which corresponds to some sequence of $|J_{\sigma,i}|$ many neighbours of i, say C, the entropy of X_i conditioned on $X_{J_{\sigma,i}} = C$ can be bounded above by $\log(|N(i) \setminus C|)$. So, let us denote by $N_\sigma(i) = N(i) \setminus \{X_j : j \in J_{\sigma,i}\}$ the vertices in the neighbourhood of i without those already chosen by some X_j.

It follows that, for any fixed σ and i we can calculate as follows, where C ranges over sequences of vertices in $N(i)$ of length $|J_{\sigma,i}|$

$$\begin{aligned}\mathbb{H}(X_i|X_{J_{\sigma,i}}) &= \sum_{C}\mathbb{P}(X_{J_{\sigma,i}} = C)\mathbb{H}(X_i|X_{J_{\sigma,i}} = C) \\ &\leq \sum_{j=1}^{d_i}\mathbb{P}(|N_\sigma(i)| = j)\log j\end{aligned}$$

Where we used the definition of conditional entropy, and then Lemma 2.2. However, since we're picking a random matching, it doesn't seem like we have any control over this improvement, since we don't know how much this will reduce the range of X_i.

However, for any fixed matching M, if we pick a random permutation σ, we claim that the size of $|N_\sigma(i)|$ is in fact uniformly distributed between 1 and d_i. Indeed, for a given matching we only care about the order in which we pick i and the vertices matched in M to the neighbours of i. Since i is equally likely to be chosen in any position in this list, the claim follows. In other words, for a fixed matching M, the proportion of σ such that $|N_\sigma(i)| = k$ is $\frac{1}{d_i}$ for each $1 \leq k \leq d_i$.

Since this is true separately for each particular matching, then it is also true when we pick a random matching. So, even though we can't bound any of the terms $\mathbb{P}(|N_\sigma(i)| = j)$ for a fixed σ, we can bound their average.

That is to say, if we pick M and σ both uniformly at random then

$$\mathbb{P}_{\sigma,M}(|N_\sigma(i)| = j) = 1/d_i$$

and hence, by definition

$$\frac{1}{n!}\sum_{\sigma}\mathbb{P}(|N_\sigma(i)| = j) = \frac{1}{d_i}$$

Hence,

$$\begin{aligned}
\mathbb{H}(X) &= \frac{1}{n!}\sum_{\sigma}\sum_{i=1}^{n}\mathbb{H}(X_{\sigma(i)}|X_{\sigma(1)}, X_{\sigma(2)}, \dots X_{\sigma(i-1)}) \\
&= \sum_{i=1}^{n}\frac{1}{n!}\sum_{\sigma}\mathbb{H}(X_i|X_{J_{\sigma,i}}) \\
&\leq \sum_{i=1}^{n}\frac{1}{n!}\sum_{\sigma}\sum_{j=1}^{d_i}\mathbb{P}(|N_\sigma(i)| = j)\log(j) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{d_i}\left(\sum_{\sigma}\frac{1}{n!}\mathbb{P}(|N_\sigma(i)| = j)\right)\log(j) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{d_i}\frac{\log j}{d_i} = \sum_{i=1}^{n}\frac{\log(d_i!)}{d_i}
\end{aligned}$$

giving the bound as claimed. □

Note that this bound is tight. If we take G to be $\frac{n}{d}$ copies of $K_{d,d}$ then we have that $d(v) = d$ for all $v \in A$ and every matching consists of picking one from the $d!$ possible matchings on each $K_{d,d}$. Therefore.

$$\phi(G) = \prod_{i=1}^{\frac{n}{d}} d! = \prod_{v\in A}(d(v)!)^{\frac{1}{d(v)}}.$$

A natural question to ask is what happens for a non-bipartite G? It turns out a similar bound can be given, and as we will see in the examples sheet, it can actually be derived in a clever way from Brégman's Theorem.

Theorem 4.3. *Let $G = (V, E)$ be a graph with $|V| = 2n$. Then*

$$\phi(G) \leq \prod_{v\in V}(d(v)!)^{\frac{1}{2d(v)}}.$$

4.2 Sidorenko's Conjecture

4.2.1 Coupling

One commonly studied random variable, that we will be interested in in this section, is that of a *random graph*.

Definition. The random graph $G(n, p)$ is a random variable taking values in the set

$$\{G\colon G \text{ a graph with } |G| = n\},$$

where the probability of each graph G with m edges is

$$p(G) = p^m(1-p)^{\binom{n}{2}-m}$$

We note that, this corresponds to the usual notion of picking a random graph by including every potential edge independently with probability p. If we take $p = 1/2$ it is not hard to see that this random variable is uniformly distributed on all graphs with n vertices, and so statements about $G(n, 1/2)$ can be though of as statements about the 'average' graph.

For the next section we will need to use a tool from probability theory that allows us to relate two unrelated random variables by considering them both as marginal distributions of a pair of random variables living in the same space. As a simple example, consider two random variables $X \sim G(n, p)$ and $Y \sim G(n, q)$ where $p < q$. It seems obvious that for increasing properties P of graphs, such as being connected, $\mathbb{P}(X \text{ has property P}) < \mathbb{P}(Y \text{ has property P})$, but showing this explicitly is cumbersome.

However we could consider a different probability space, $\Omega = [0, 1]^{\binom{n}{2}}$ together with the obvious probability measure. Let's define two random variables X' and Y' as follows:

Let's order the edges of K_n as $\{e_1, e_2, \ldots, e_{\binom{n}{2}}\}$ and let $X'(\omega)$ be the graph where $e_i \in E(X'(\omega))$ if and only if $\omega_i \leq p$. Similarly let $Y'(\omega)$ be the graph where $e_i \in E(Y'(\omega))$ if and only if $\omega_i \leq q$.

Now, it's relatively easy to show that $X' \sim X$, which is to say that X' takes the same values with the same probabilities as X, and also $Y' \sim Y$, however, the two random variables X' and Y' are now closely related. Indeed, for any $\omega \in \Omega$ it is easy to see that $X'(\omega) \subseteq Y'(\omega)$. In particular, for any increasing property P of graphs if $X'(\omega)$ has property P than so does $Y'(\omega)$. Hence

$$\mathbb{P}(X \text{ has property P}) = \mathbb{P}(X' \text{ has property P}) < \mathbb{P}(Y' \text{ has property P}) = \mathbb{P}(Y \text{ has property P}).$$

For our purposes we will want to consider a slightly more general situation. The intuition will be that there is a certain part of X and Y which 'looks the same' and we will want a coupling which agrees on this part.

Suppose we have two discrete random variables X_1 and X_2 taking values in $\mathcal{X}_1$ and $\mathcal{X}_2$ and a third random variable X_3 taking values in $\mathcal{X}_3$ together with maps $\psi_i : \mathcal{X}_i \to \mathcal{X}_3$ for $i = 1, 2$ such that $\psi_i(X_i) = X_3$ for $i = 1, 2$.

Let $\mathcal{Y} = \{(x_1, x_2) \in \mathcal{X}_1 \times \mathcal{X}_2 : \psi_1(x_1) = \psi_2(x_2)\}$. A random variable $Y = (Y_1, Y_2)$ taking values in $\mathcal{Y}$ is a *coupling of* X_1 *and* X_2 *over* X_3 if it's marginal distributions in the first and second coordinate are X_1 and X_2 respectively. In this situation there is an 'obvious' choice for Y, namely

$$\mathbb{P}(Y = (x_1, x_2)) = \frac{\mathbb{P}(X_1 = x_1)\mathbb{P}(X_2 = x_2)}{\mathbb{P}(X_3 = \psi_1(x_1) = \psi_2(x_2))}.$$

We call this the *conditionally independent coupling of* X_1 *and* X_2 *over* X_3. Note that the

marginal distribution of the first coordinate here is indeed X_1 since for any $x_1 \in \mathcal{X}_1$

$$\begin{aligned}
\mathbb{P}(Y_1 = x_1) &= \sum_{x_2 \colon (x_1,x_2)\in\mathcal{Y}} \mathbb{P}(Y = (x_1, x_2)) \\
&= \sum_{x_2 \colon (x_1,x_2)\in\mathcal{Y}} \frac{\mathbb{P}(X_1 = x_1)\mathbb{P}(X_2 = x_2)}{\mathbb{P}(X_3 = \psi_1(x_1) = \psi_2(x_2))} \\
&= \frac{\mathbb{P}(X_1 = x_1)}{\mathbb{P}(X_3 = \psi_1(x_1) = \psi_2(x_2))} \sum_{x_2 \colon (x_1,x_2)\in\mathcal{Y}} \mathbb{P}(X_2 = x_2) \\
&= \frac{\mathbb{P}(X_1 = x_1)}{\mathbb{P}(X_3 = \psi_1(x_1) = \psi_2(x_2))} \sum_{x_2 \colon \psi_2(x_2)=\psi_1(x_1)} \mathbb{P}(X_2 = x_2) \\
&= \frac{\mathbb{P}(X_1 = x_1)}{\mathbb{P}(X_3 = \psi_1(x_1) = \psi_2(x_2))} \mathbb{P}(X_3 = \psi_2(x_2)) \\
&= \mathbb{P}(X_1 = x_1).
\end{aligned}$$

A similar calculation shows that the marginal distribution of the second coordinate is X_2.

Intuitively, $\mathcal{Y}$ is the set of all possible values of the pair (X_1, X_2) which agree on the parts of X_1 and X_2 corresponding to X_3. The conditionally independent coupling is then in some sense the 'most independent' coupling. Conditioned on a particular value x_3 for X_3, Y can only take values in $\{(x_1, x_2) \colon \psi_1(x_1) = \psi_2(x_2) = x_3\}$ and in the conditionally independent coupling the value of the first and second coordinate will be independent, conditioned on the value of X_3.

One particular useful feature of the conditionally independent coupling is that it maximises the entropy. That is, if X_1, X_2 and X_3 are as above, Y is the conditionally independent coupling of X_1 and X_2 over X_3, and Z is any other coupling then

$$H(Z) \leq H(Y). \tag{4.1}$$

To see this we will need the following small lemma:

Lemma 4.4. *Let U, V and W be discrete random variables and suppose that U and V are conditionally independent given W. Then*

$$\mathbb{H}(U|V, W) = \mathbb{H}(U|W).$$

Proof. Since U and V are conditionally independent given W it follows that $\mathbb{H}(U, V|W) = \mathbb{H}(U|W) + \mathbb{H}(V|W)$. However by Lemma 2.5

$$\mathbb{H}(U|W) + \mathbb{H}(V|W) = \mathbb{H}(U, V|W) = \mathbb{H}(U|V, W) + \mathbb{H}(V|W)$$

from which the claim follows. □

Then, given X_1, X_2, X_3, Y and Z as above let us note the following facts:

- $Y_i \sim Z_i \sim X_i$ for $i = 1, 2$;
- $\psi_1(Y_1) = \psi_2(Y_2) \sim \psi_1(Z_1) = \psi_2(Z_2) \sim X_3$;

- Y_1 and Y_2 are mutually independent given $\psi_1(Y_1)$.

So we can calculate

$$\begin{aligned}
\mathbb{H}(Z) &= \mathbb{H}(Z_1, Z_2, \psi_1(Z_1)) \\
&= \mathbb{H}(Z_1, Z_2|\psi_1(Z_1)) + \mathbb{H}(\psi_1(Z_1)) \\
&= \mathbb{H}(Z_1|Z_2, \psi_1(Z_1)) + \mathbb{H}(Z_2|\psi_1(Z_1)) + \mathbb{H}(X_3) \\
&\leq \mathbb{H}(Z_1|\psi_1(Z_1)) + \mathbb{H}(Z_2|\psi_1(Z_1)) + \mathbb{H}(X_3) \\
&= \mathbb{H}(Z_1|\psi_1(Z_1)) + \mathbb{H}(Z_2|\psi_2(Z_2)) + \mathbb{H}(X_3) \\
&= \mathbb{H}(Y_1|\psi_1(Y_1)) + \mathbb{H}(Y_2|\psi_2(Y_2)) + \mathbb{H}(\psi_1(Y_1)) \\
&= \mathbb{H}(Y_1|Y_2, \psi_1(Y_1)) + \mathbb{H}(Y_2|\psi_1(Y_1)) + \mathbb{H}(\psi_1(Y_1)) \\
&= \mathbb{H}(Y_1, Y_2, \psi_1(Y_1)) = \mathbb{H}(Y).
\end{aligned}$$

4.2.2 Sidorenko's Conjecture

A *graph homomorphism* $f : H \to G$ is a map which preserves adjacency. Given two graphs G and H we can define the homomorphism density $t(H, G)$ as the proportion of all maps from $V(H)$ to $V(G)$ which are homomorphisms, or equivalently, the probability that a random such map will be a homomorphism. Explicitly if we write hom(H,G) for the set of homomorphisms from H to G then

$$t(H, G) = \frac{|\hom(H, G)|}{v(G)^{v(H)}}.$$

For example, if we take $H = K_2$ to be a single edge then

$$t(K_2, G) = \frac{2e(G)}{v(G)^2}.$$

Conjecture 4.5 (Sidorenko's Conjecture)**.** *For every bipartite graph H and every graph G*

$$t(H, G) \geq t(K_2, G)^{e(H)}.$$

Note that, if we take G to be a random graph $G(n, p)$ with $p = t(K_2, G)$ then $t(K_2, G)^{e(H)} = p^{e(H)}$ is just the probability that a particular copy of H is a subgraph of $G(n, p)$. Taking the sum over all possible copies of H (and noting that almost all maps from $v(H)$ to $v(G)$ are injective as long as n is large enough), it follows that

$$\mathbb{E}(t(H, G)) = \frac{\mathbb{E}(|\hom(H, G)|)}{v(G)^{v(H)}} = \frac{(1 + o(1))v(G)^{v(H)}p^{e(H)}}{v(G)^{v(H)}} = (1 + o(1))p^{e(H)}.$$

and so by the first moment method, this bound is asymptotically tight.

This conjecture is around 25 years old, and there has been much partial progress towards it, showing that it holds for many classes of bipartite graphs. Quite recently some progress was made independently by Colon, Kim, Lee and Lee, and Szegedy using entropy methods. Let's consider how we might used entropy to attack this problem.

Suppose we take some random variable X which takes values in $\hom(H, G)$. By Lemma 2.2 we know that $\log |\hom(H, G)| \geq \mathbb{H}(X)$ for any such X and so

$$\log t(H, G) \geq \mathbb{H}(X) - v(H) \log v(G).$$

Hence, to show that $\log(t(H,G)) \geq \log(t(K_2,G)^{e(H)})$, it would be sufficient to find an X such that

$$\mathbb{H}(X) \geq e(H)\log t(K_2,G) + v(H)\log v(G) = e(H)\log 2e(G) + (v(H) - 2e(H))\log v(G).$$

We can re-write this in a slightly nicer form by letting V be a vertex chosen uniformly at random from $V(G)$ and E be an *oriented* edge chosen uniformly at random from $\overrightarrow{E}(G)$, or in other words V is uniform on $\hom(K_1,G)$ and E is uniform of $\hom(K_2,G)$. Then $\mathbb{H}(V) = \log v(G)$ and $\mathbb{H}(E) = \log 2e(G)$ and so we want to find an X such that

$$\mathbb{H}(X) \geq e(H)\mathbb{H}(E) + (v(H) - 2e(H))\mathbb{H}(V)$$

Since X gives a distribution on $\hom(H,G)$, for every subgraph $H' \subseteq H$ we can consider the marginal distribution of H' in X, that is, the induced distribution coming from the projection $\hom(H,G) \to \hom(H',G)$. In particular, every oriented edge $(u,v) \in \overrightarrow{E}(H)$ has a marginal distribution in X (taking values in $\hom(K_2,G)$), which tells us for each oriented edge $(x,y) \in \overrightarrow{E}(G)$ how likely it is that in a random homomorphism chosen by X, (u,v) is mapped to (x,y). A natural property to ask of X is that each of these marginal distributions are uniform on $\hom(K_2,G)$.

Definition. A *witness variable* for a bipartite graph H is a family of random variables $(X(G) : G \text{ a graph})$ such that for every G:

1. $X(G)$ is a random variable taking values in $\hom(H,G)$;
2. For every edge $(u,v) \in \overrightarrow{E}(H)$ the marginal distribution of (u,v) in X is uniform.
3. $\mathbb{H}(X(G)) \geq e(H)\mathbb{H}(E(G)) + (v(H) - 2e(H))\mathbb{H}(V(G))$.

Where, in a slight abuse of notation, we have written $V(G)$ and $E(G)$ for the uniform random variables on $V(G)$ and $\overrightarrow{E}(G)$. In what follows we'll normally be talking about a fixed G, and so we'll just write X, V and E for these random variables. The discussion above shows that the existence of a witness variable for H is a sufficient condition for H to satisfy Sidorenko's conejcture.

Theorem 4.6. *If H has a witness variable, then H satisfies Sidorenko's conjecture.*

So far perhaps we haven't really done very much, in fact we've made things slightly harder for ourselves by asking that X satisfy this second condition. However, the useful thing about this reformulation is that we will be able to build witness variables for graphs H by combining witness variables for smaller graphs. This will allow us to inductively show that classes of graphs built using certain graph operations will satisfy Sidorenko's conjecture. However to get started we will need a base case, for which we can use a single edge.

Lemma 4.7. *K_2 has a witness variable.*

Proof. Property 2 tells us that if K_2 has a witness variable, it must be the uniform distribution on $\hom(K_2,G)$, which we are writing as E. So, we just have to check that Property 3 holds for $X = E$. Indeed

$$e(K_2)\mathbb{H}(E) + (v(K_2) - 2e(K_2))\mathbb{H}(V) = 1 \cdot \mathbb{H}(E) + (2-2) \cdot \mathbb{H}(V) = \mathbb{H}(E) = \mathbb{H}(X).$$

□

So, we have a base case. What operations might the set of graphs with a witness variable be closed under?

Definition. Given two graphs H_1 and H_2 and subsets $S_1 \subseteq V(H_1)$ and $S_2 \subseteq V(H_2)$ together with a bijection $f : S_1 \to S_2$ we define the *glued graph* $H = H_1 \oplus_f H_2$ in the obvious way. That is, the vertex set $V(H) = (V(H_1) \cup V(H_2))/(s \sim f(s) \colon s \in S_1)$ and $E(H)$ is the image of $E(H_1) \cup E(H_2)$ under the quotient map (except we delete parallel edges). We will denote by S the image of S_1 (and S_2) in H.

Lemma 4.8. *Let H_1 and H_2 be graphs with witness variables X_1 and X_2 and let $S_1 \subseteq V(H_1)$ and $S_2 \subseteq V(H_2)$ be independent sets. Suppose there is a bijection $g : S_1 \to S_2$ such that the marginal distribution of S_1 in X_1 is the same as the marginal distribution of $g(S_1)$ in X_2, which we will denote by X_S. Let $H = H_1 \oplus_g H_2$ and let Y be the conditionally independent coupling of X_1 and X_2 over X_S. Then Y is a witness variable for H.*

Proof. To formally define Y we need to understand the maps ψ_i from X_i to X_S. Informally, this is pretty simple. X_1 and X_2 take values in $\hom(H_i, G)$, that is, functions $f : V(H_i) \to V(G)$. For each such function the restriction of f to S_i is a function in $\hom(S_i, G)$ and so ψ_i is just the map which restricts the image of X_i to S_i. With these maps the range of Y is then

$$\mathcal{Y} := \{(f_1, f_2) \colon f_i \in \hom(H_i, G) \text{ and } f_1\big|_{S_1} = f_2\big|_{g(S_1)}\}.$$

However there is an obvious bijection from $\hom(H, G) \to \mathcal{Y}$ given by $f \mapsto (f\big|_{V(H_1)}, f\big|_{V(H_2)})$. Hence Y satisfies Property 1.

For Property 2 we need to check the marginal distribution of an edge $(x, y) \in E(H)$ in Y. Since S is independent, every edge in H is contained in H_1 or H_2, say without loss of generality $(x, y) \in E(H_1)$. The marginal distribution of H_1 in Y is X_1, since Y is a coupling of X_1 and X_2, and the marginal distribution of (x, y) in Y is determined by the marginal distribution of H_1 in Y. Hence, since X_1 was a witness variable for H_1 it follows that the marginal distribution of (x, y) in X_1, and hence in Y, is E.

Finally, for Property 3 we have to estimate the entropy of Y. We know that $Y_1 \sim X_1$ and $Y_2 \sim X_2$, and furthermore we can consider the random variable Y_S given by the marginal distribution of S_1 in Y, noting that $Y_S \sim X_S$.

Since Y determines and is determined by the triple (Y_1, Y_2, Y_S) we have that $\mathbb{H}(Y) = \mathbb{H}(Y_1, Y_2, Y_S)$. Hence by Lemma 2.4

$$\mathbb{H}(Y) = \mathbb{H}(Y_1, Y_2, Y_S) = \mathbb{H}(Y_S) + \mathbb{H}(Y_1|Y_S) + \mathbb{H}(Y_2|Y_1, Y_S).$$

Since, by construction, Y_1 and Y_2 are conditionally independent given Y_S, by Lemma 4.4

$$\mathbb{H}(Y) = \mathbb{H}(Y_S) + \mathbb{H}(Y_1|X_S) + \mathbb{H}(Y_2|Y_S).$$

On the other hand, since each Y_i determines Y_S, again by Lemma 2.4 we can write

$$\mathbb{H}(Y_i) = \mathbb{H}(Y_i, Y_S) = \mathbb{H}(Y_S) + H(Y_i|Y_S).$$

Combining these equalities we see

$$\mathbb{H}(Y) = \mathbb{H}(Y_1) + \mathbb{H}(Y_2) - \mathbb{H}(Y_S) = \mathbb{H}(X_1) + \mathbb{H}(X_2) - \mathbb{H}(X_S).$$

Note that submodularity gives $\mathbb{H}(Y) \leq \mathbb{H}(Y_1)+\mathbb{H}(Y_2)-\mathbb{H}(Y_S)$, however for our purposes we wish to prove a lower bound of $\mathbb{H}(Y)$, so it is useful that we get an equality here for the conditionally independent coupling.

Then, by our assumption that X_1 and X_2 are witness variables we can bound them from below as follows

$$\begin{aligned}
\mathbb{H}(Y) &= \mathbb{H}(X_1) + \mathbb{H}(X_2) - \mathbb{H}(X_S) \\
&\geq e(H_1)\mathbb{H}(E) + (v(H_1) - 2e(H_1))\mathbb{H}(V) + e(H_2)\mathbb{H}(E) + (v(H_2) - 2e(H_2))\mathbb{H}(V) - \mathbb{H}(X_S) \\
&= \big(e(H_1) + e(H_2)\big)\mathbb{H}(E) + \big((v(H_1) + v(H_2)) - 2(e(H_1) + e(H_2))\big)\mathbb{H}(V) - \mathbb{H}(X_S) \\
&= e(H)\mathbb{H}(E) + \big(v(H) + |S| - 2e(H)\big)\mathbb{H}(V) - \mathbb{H}(X_S) \\
&= e(H)\mathbb{H}(E) + \big(v(H) - 2e(H)\big)\mathbb{H}(V) + \big(|S|\mathbb{H}(V) - \mathbb{H}(X_S)\big).
\end{aligned}$$

Now, X_S is a random variable on $V(G)^S$, and hence

$$H(X_S) \leq \log |V(G)|^{|S|} = |S| \log |V(G)| = |S|\mathbb{H}(V).$$

It follows that

$$\mathbb{H}(Y) \geq e(H)\mathbb{H}(E) + \big(v(H) - 2e(H)\big)\mathbb{H}(V),$$

and so Y is indeed a witness variable for H. □

So, we have a procedure for building new graphs with witness variables from other graphs with witness variables. However, we can only do so when we have two independent sets which have the same marginal distribution in the witness variables.

In fact, we already have some control over the marginal distribution of certain vertex sets in witness variables. Indeed, since the marginal distribution of any edge of H is uniform, this actually determines the marginal distribution of each vertex as well.

Lemma 4.9. *Suppose H has no isolated vertices. Let D be a random variable taking values in $V(G)$ where for each $v \in V(G)$*

$$\mathbb{P}(D = v) = \frac{d(v)}{2e(G)}.$$

Then if X is a witness variable for H, then the marginal distribution of any vertex of H in X is precisely D.

Proof. Note that, in the uniformly distributed random variable E, the marginal distribution of a vertex in this edge is D. Indeed, if we let $E = (U, V)$ then for every $x \in V(G)$

$$\mathbb{P}(U = X) = \sum_{(x,y)\in E(G)} \mathbb{P}(E = (x, y)) = \frac{d(x)}{2e(G)}.$$

However, since H has no isolated vertices, every vertex $v \in V(H)$ is incident to some edge e. Then, since the marginal distribution of e in X is E, it follows that the marginal distribution of v in X is the marginal distribution of v in E, which is precisely D. □

In particular, if $|S_1| = |S_2| = 1$ in Lemma 4.8 then their marginal distributions in X_1 and X_2 will agree. Therefore we can conclude

Lemma 4.10. *If H_1 and H_2 are connected and have witness variables and H is formed by gluiing H_1 and H_2 along a single vertex, then H has a witness variable.*

A simple corollary of this is that trees satisfy Sidorenko's conjecture.

Corollary 4.11. *Every tree has a witness variable, and hence satisfies Sidorenko's conjecture.*

Proof. We prove this by induction on the number of vertices. The base case if $T = K_2$, which follows from Lemma 4.7. Suppose the Corollary holds for all trees with $\leq n$ vertices. Given a tree T on $n+1$ vertices we pick a leaf e and consider $T - e$. $T - e$ has a witness variable by assumption, and T is formed by gluing $T - e$ and K_2 along a single vertex. Hence by Lemmas 4.7 and 4.10 it follows that T has a witness variable. □

Corollary 4.12. *Let T be a tree, $S \subseteq V(T)$ be independent and let $g : S \to S$ be the identity map. Then $T \oplus_g T$ has a witness variable.*

In particular, every even cycle has a witness variable and hence satisfies Sidorenko's conjecture.

Proof. By Corollary 4.11 T has a witness variable and since g is the identity map it is trivially true that the conditions of Lemma 4.8 hold. Hence $T \oplus_g T$ has a witness variable. □

Both of these corollaries were already shown to hold by Sidorenko in his original paper using the Cauchy-Schwartz and Hölder inequalities, however with relatively little extra work we can use these ideas to prove that a class of graphs which are called *tree-arrangeable* satisfy Sidorenko's conjecture, which was the best known result before the use of entropy.

Definition. Suppose H is a bipartite graph on (A, B, E). Let us define two operations for extending H.

- We may add a single vertex v to A and connect it to a single vertex $b \in B$, or
- We may add a single vertex v to B and connect it to a subset of $N(b)$ for some vertex $b \in B$.

A graph is called *tree-arrangeable* if it can be built from K_2 via sequence of these operations. For example, trees can be seen to be tree-arrangeable. Such graphs were defined by Kim, Lee and Lee, who gave an equivalent description of such graphs and proved that they satisfy Sidorenko's conjecture. Using the entropy tools of Szegedy we can give a simple alternative proof.

We will need a slight strengthening of the concept of a witness variable. Note that, if we fix a vertex $v \in V(H)$ then $N_H(v)$ is an independent set and the marginal distribution of any one of these neighbours is D, but these distributions are not independent. However it would be nice if these marginal distributions were 'as independent as possible', that is, if we let X_v be the marginal distribution of a fixed vertex $v \in V(H)$, we would like that once we fix a value for

X_v, say $X_v = y \in V(G)$ that the random variables $\{(X_w|X_v = y)\colon w \in N_H(v)\}$ are mutually independent and identically distributed. Note that, since the marginal distribution of every edge is uniform, these random variables are always identically distributed since $\mathbb{P}(X_w = x|X_v = y) = \frac{1}{d(y)}$ for every $x \in N_G(y)$.

Suppose H is a bipartite graph with bipartition classes A and B. We say that X is a *balanced witness variable* for H if it is a witness variable for H and for every $v \in B$ and $y \in V(G)$ the random variables $\{(X_w|X_v = y)\colon w \in N_H(v)\}$ are mutually independent. In other words, the distribution of $N_H(v)$ can be obtained by choosing a sample from D and then picking $d(v)$ neighbours of this vertex independently and uniformly.

Theorem 4.13. *Let H be a tree-arrangeable graph. Then H has a balanced witness variable, and hence satisfies Sidorenko's conjecture.*

Proof. Our proof will proceed via induction on $v(H)$, and this induction will need to use the stronger condition of having a balanced witness variable, which is why we introduced the notion.

The base case is again $H = K_2$. We know K_2 has a witness variable, the uniform distribution on the edges of G, and for each vertex the neighbourhood is a single vertex, which trivially satisfies the stronger condition.

Suppose then that the inductive hypothesis holds for all $v(H) \leq n$ and let H be a tree-arrangeable graph on $n + 1$ vertices. By definition there is some $v \in V(H)$ such that H can be obtained from $H - v$ from one of the two operations defined above.

Let us first assume that $v \in A$ and is connected to a single $b \in B$, that is, $H = H - v \oplus_g K_2$ where g maps a vertex of K_2 to b. As in Lemma 4.8 let X_1 and X_2 be witness variables for $H - v$ and K_2 respectively and let Y be the conditionally independent coupling of X_1 and X_2 over X_b, the marginal distribution of b in X_1. Note that $X_2 \sim E$ and $X_b \sim D$. By Lemma 4.8 Y is a witness variable for H, so it remains to check that Y is a balanced witness variable. Let us define Y_w to be the marginal distribution of w in Y for each $w \in V(H)$ and let X_w be the marginal distribution of w in X_1 for $w \in V(H - v)$.

For every $b \neq b' \in B$ the neighbourhood of b' in H is the same as the neighbourhood as b' in $H - v$. Since the marginal distribution of $H - v$ in Y is just X_1, it follows that the random variables $\{Y_w\colon w \in N_H(b')\} \cup \{Y_{b'}\}$ have the same joint distribution as the random variables $\{X_w\colon w \in N_H(b')\} \cup \{X_{b'}\}$. In particular, for any $y \in V(G)$ the set $\{(Y_w|Y_{b'} = y)\colon w \in N_H(b')\}$ has the same distribution as the set $\{(X_w|X_{b'} = y)\colon w \in N_H(b')\}$ and hence, since X_1 is a balanced witness variable, are mutually independent.

Now, $N_H(b) = N_{H-v}(b) + v$. However, since $N_{H-v}(b) \subseteq H - v$ and Y is the conditionally independent coupling of X_1 and X_2 over X_b it follows that, conditioned on the value of Y_b, the marginal distribution of $N_{H-v}(b)$ in Y is independent of the marginal distribution of v in Y. In particular, conditioned on the value of Y_b, $\{Y_w\colon w \in N_{H-v}(b)\}$ is independent of Y_v. Then, since X_1 and X_2 were both balanced witness variables it follows that for any $y \in V(G)$ the set $\{(Y_w|Y_b = y)\colon w \in N_H(b)\}$ is mutually independent.

So, let us assume there is a $v \in B$ which is connected to a subset of the neighbourhood of some other $b \in B$. Hence H can be obtained by $(H-v)\oplus_g K_{1,m}$ where g is a map from the leaves of $K_{1,m}$ to $N_H(v)$. Note that, since $K_{1,m}$ can be built from K_2 using only the first operation,

we may assume by induction that both $H - v$ and $K_{1,m}$ have balanced witness variables X_1 and X_2.

We claim that the set of leaves of $K_{1,m}$ has the same marginal distribution in X_2 as the independent set $N_H(v)$ in X_1. Indeed, this follows from the assumption that both X_1 and X_2 are balanced witness variables, since the leaves of $K_{1,m}$ are distributed uniformly and independently as neighbours of the centre, whose distribution is D, and similarly the vertices of $N_H(b)$ are distributed uniformly and independently as neighbours of b, whose distribution is D, and hence so is the subset $N_H(v)$.

Hence we may apply Lemma 4.8 to see that the conditionally independent coupling of X_1 and X_2 over $X_{N_H(v)}$ is a witness variable for H. It remains to check that this witness variable is balanced.

However this claim follows for every $v \neq b \in B$ from that fact that X_1 is a strong witness variable, and for v from the fact that X_2 is a strong witness variable. □

4.3 Shearer's lemma and projection inequalities

4.3.1 Shearer's Lemma

Given a sequence of discrete random variables random variables $X_1, X_2, \ldots, X_n$ and some subset $A \subseteq [n]$ let define $X_A := (X_i : i \in A)$.

Lemma 4.14 (Shearer's inequality)**.** *Let $X_1, X_2, \ldots, X_n$ be discrete random variables and $\mathcal{A}$ a collection (not necessarily distinct) of subsets of $[n]$, such that each $i \in [n]$ is in at least m members of $\mathcal{A}$. Then*

$$H(X_1, X_2, \ldots, X_n) \leq \frac{1}{m} \sum_{A \in \mathcal{A}} H(X_A).$$

Proof. Let $A = \{a_1, a_2, \ldots, a_k\}$ with $a_1 < a_2 < \ldots < a_k$. We have that

$$\begin{aligned} H(X_A) &= H(X_{a_1}) + H(X_{a_2}|X_{a_1}) + \ldots + H(X_{a_k}|X_{a_1}, X_{a_2}, \ldots, X_{a_{k-1}}) \\ &\geq H(X_{a_1}|X_{<a_1}) + H(X_{a_2}|X_{<a_2}) + \ldots + H(X_{a_k}|X_{<a_k}), \end{aligned}$$

where $X_{<i} = (X_1, X_2, \ldots X_{i-1})$. This follows from repeated applications of the chain rule, and the fact that entropy only decreases if we condition on more variables. Therefore

$$\begin{aligned} \sum_{A \in \mathcal{A}} H(X_A) &\geq m. \sum_{i \in [n]} H(X_i|X_{<i}) \\ &= m.H(X_1, X_2, \ldots, X_n) \end{aligned}$$

□

4.3.2 The Bollobás-Thomason Box Theorem

Shearer's Lemma is closely related to notions of isoperimetry, the relation between the volume of a shape and it's 'perimeter' in the following way. If we think about a shape $S \subseteq \mathbb{R}^n$ with area

$|S|$ then we can think about the process of picking a random point inside of S. This determines a vector $X = (X_1, \ldots, X_n)$ where the X_i are dependent on each other, depending on what the shape S is.

Suppose we take a very fine grid approximating $\mathbb{R}^n$, we can then think of S as being a discrete subset of this grid, whose number of points is proportional to $|S|$. Since this vector $X = (X_1, \ldots, X_n)$ now has some finite range, we can relate the volume of S directly to the entropy of X. That is

$$H(X) = \log |S|.$$

How can we interpret the random variable X_A for $A \subset [n]$? Well in this case, this is relatively clear, these correspond to the projections on the shape S onto the subspace spanned by the coordinates in A. That is, if we let S_A be the projection of S onto the subspace

$$\{(x_1, \ldots, x_n) : x_i = 0 \text{ for all } i \in A\}$$

Then the range of X_A is the 'volume' (in the $n - |A|$-dimensional sense) of S_A. We will write S_j for $S_{\{j\}}$.

In this way, Shearer's inequality gives us a way to relate the volume of a shape to it's lower dimensional projections. For example, if we just consider the 1-dimensional projections, we have the famous Loomis Whitney inequality:

Theorem 4.15 (The Loomis-Whitney inequality)**.** *Let $S \subset \mathbb{Z}^n$ then,*

$$|S|^{n-1} \leq \prod_{i=1}^{n} |S_{[n]\setminus\{i\}}|$$

For example, in two dimensions this simply says that the area of a shape can be bounded above by the product of its one-dimensional projections, a relatively trivial fact. But even in three-dimensions it is not clear what the relationship should be between the volume of a shape and its projections onto two dimensional subspaces.

Notice that, this theorem is tight when $|S|$ is a 'box', that is, a set of the form $[1, m_1] \times [1, m_2] \times \ldots \times [1, m_n]$. Indeed, the volume of $|S|$ is $\prod_{i=1}^{n} m_i$ and the volume of the projection of S onto the hyperplane where $x_i = 0$ is just $\prod_{j \neq i} m_i$. This is perhaps not surprising, as a box represents the case where the X_is are independent, where we get equality in the argument for Shearer's inequality.

In fact, we will show a more general theorem, and deduce the Loomis-Whitney theorem as a corollary. We say a collection of sets $\mathcal{C} = \{C_1, \ldots, C_m\} \subset 2^{[n]}$ is a *k-uniform cover* if each $i \in [n]$ belongs to exactly k many of the C_j.

Theorem 4.16 (Uniform covers theorem)**.** *Let $S \subset \mathbb{Z}^n$ and let $\mathcal{C} \subset 2^{[n]}$ be a k-uniform cover, then*

$$|S|^k \leq \prod_{C \in \mathcal{C}} |S_C|$$

Remark 4.17. *Note that $\mathcal{C} = \{[n] \setminus \{i\} : i \in [n]\}$ is an $(n-1)$-uniform cover of $[n]$, and so Theorem 4.15 follows from Theorem 4.16.*

Proof. Let us choose a points $X = (X_1, \ldots, X_n)$ uniformly at random from S. Then, $H(X) = \log |S|$. By Lemma 4.14 it follows that

$$H(X) \leq \frac{1}{k} \sum_{C \in \mathcal{C}} H(X_C).$$

However, the range of X_C is $|S_C|$ and so it follows that

$$H(X) \leq \frac{1}{k} \sum_{C \in \mathcal{C}} \log |S_C|.$$

Combining the two equatons we see that

$$\log |S| \leq \frac{1}{k} \sum_{C \in \mathcal{C}} \log |S_C|$$

and so

$$|S|^k \leq \prod_{C \in \mathcal{C}} |S_C|,$$

as claimed. □

As before, if we consider the 1-uniform cover $\{\{i\} : i \in [n]\}$, Theorem 4.16 tells us the elementary fact the volume of a shape can be bounded by the product of its one-dimensional projections. We also note that, whilst Theorem 4.16 is a simple consequence of Shearer's Lemma, it is not hard to show that Shearer's Lemma is a consequence of Theorem 4.16.

By taking limits of finer and finer grids it is possible to show that Theorem 4.16 also holds for subsets of $\mathbb{R}^n$ with the lebesgue measure. In fact a rather amazing strengthening of Theorem 4.16 can be shown to hold, which is known as the Bollobás-Thomason Box Theorem. In what follows we will write $|S|$ for the Lebesgue measure of a set $S \subseteq \mathbb{R}^n$.

Theorem 4.18 (Bollobás-Thomason Box Theorem)**.** *Let $S \subset \mathbb{R}^n$ be compact. Then there is a box $A \subset \mathbb{R}^n$ such that $|A| = |S|$ and $|A_I| \leq |S_I|$ for all $I \subseteq [n]$.*

That is, for any shape we can find a box of the same volume such that *every* lower dimensional projection of this box has smaller volume than the corresponding projection of S. This immediately tells us that for any upper bound we might want to prove for the volume of a set in terms of the volumes of its projection, we only have to check that it holds for boxes.

Indeed, if we know that for every box A, $|A| \leq f(A_I : I \subset [n])$ for some function f which is increasing in each coordinate, then for any S we have that $|S| = |A| \leq f(A_I : I \subset [n]) \leq f(S_I : I \subset [n])$.

It is possible to prove this theorem via a continuous version of Theorem 4.16 (which can be proven analytically using H older's inequality) and a careful inductive argument, but we can also do so using an entropy argument, however to do so we will need to define a notion of entropy for continuous random variables. Suppose X is a continuous random variable taking values in $\mathbb{R}^n$ with probability density function f, then a natural guess for the *entropy* of X is the following:

$$\mathbb{H}(X) = -\int f(x) \log f(x) dx$$

where integration is with respect to the Lebesgue measure. As we saw on the example sheets this does not inherit every property of the discrete entropy, for example it can take negative values. However, many of useful properties of $\mathbb{H}$ are still true in this setting, and we will assume without proof that the following are true:

- If $\mathbb{P}(X \in S) = 1$ then $\mathbb{H}(X) \leq |S|$ with equality if (and only if) X is uniform on S;
- For any X and Y, $\mathbb{H}(X|Y) \leq \mathbb{H}(X)$.
- If we write $X = (X_1, \ldots, X_n)$ and $X_I = (X_i \colon i \in I)$ then

$$\mathbb{H}(X) = \sum_{i=1}^{n} \mathbb{H}(X_i | X_{[i-1]}).$$

Proof of Theorem 4.18. Let X be a random variable uniformly distributed on S, then $\mathbb{H}(X) = \log(|S|)$. Let us define $a_i = 2^{\mathbb{H}(X_i|X_{[i-1]})}$ and let $A = [0, a_1] \times [0, a_2] \times \ldots \times [0, a_n]$ be a box in $\mathbb{R}^n$.

Now, for any $I \subseteq [n]$, X_I takes values in S_I, and hence $\mathbb{H}(X_I) \leq \log |S_I|$. On the other hand, using the chain rule we see that, if $I = \{i_1 < i_2 < \ldots < i_k\}$

$$\begin{aligned}
\mathbb{H}(X_I) &= \mathbb{H}(X_{i_1}) + \mathbb{H}(X_{i_2}|X_{i_1}) + \ldots + \mathbb{H}(X_{i_k}|X_{i_1}, X_{i_2}, \ldots, X_{i_{k-1}}) \\
&\geq \sum_{j \in I} \mathbb{H}(X_j | X_{[j-1]}) \\
&= \sum_{j \in I} \log a_j \\
&= \log \left(\prod_{j \in I} a_j \right) \\
&= \log |A_I|.
\end{aligned}$$

Hence, $\log |S_I| \leq \log |A_I|$ and so $|S_I| \leq |A_I|$. □

Since, as mentioned above, Theorem 4.16 is in fact equivalent to Shearer's lemma we might expect there to be an entropy equivalent to the Box Theorem, and we shall show on the example sheet that this is the case.

Theorem 4.19. *Let $X = (X_1, \ldots, X_n)$ be a discrete random variable. Then there are non-negative constants $h_1, \ldots, h_n$ such that $\mathbb{H}(X) = \sum_i h_i$ and*

$$\sum_{i \in I} h_i \leq \mathbb{H}(X_I)$$

for every $I \subseteq [n]$.

4.3.3 Isoperimetry

Given a space with a notion of volume and boundary, the isoperimetric problem asks how small the boundary of a shape of fixed volume can be, and perhaps even what can be said about the structure of shapes which achieve this minimum.

The Loomis-Whitney theorem can be thought of as a type of isoperimetric inequality, bounding the volume of a shape in terms of it's lower dimensional projections. We can in fact relate Shearer's Lemma directly to some isopermetric theorems in graphs.

In a graph, there is a natural notion of 'volume' given by the cardinality of a set of vertices and there are a few natural notions of boundary one can consider. In terms of vertices we could consider either the *inner or outer vertex boundary* of a subset $S \subseteq V$, which is

$$\partial_v^+(S) := \{w \in V(G) \setminus S\colon \text{ there exists } v \in S \text{ with } (v,w) \in E(G)\},$$

or

$$\partial_v^-(S) := \{w \in S\colon \text{ there exists } v \in V(G) \setminus S \text{ with } (v,w) \in E(G)\}.$$

Or, in terms of edges we could consider the *edge-boundary*, which is

$$\partial(S) := \{(v,w) \in E(G) \setminus S\colon v \in S \text{ and } w \in V(G) \setminus S\}.$$

For any of these types of boundaries we can consider the isoperimetric problem in a graph G. These problems have been well studied in particularly structured, lattice-like graphs.

As a simple example, suppose $G = Q_d$ is the d-dimensional hypercube. The isoperimetric problem, for both edge and vertex boundary, in Q_d has been well studied. In fact, the problem of finding a subset $S \subseteq V(Q_d)$ of fixed size which minimises $\partial(S)$ was first considered due to an application to error-correcting codes.

Suppose we wish to encode the integers $\{1, \dots, 2^d\}$ into $\{0,1\}^d$ so that we can send them across a channel, but we want to try and make sure that we don't lose too much information if some bit is incorrectly transmitted. To put it another way, for every edge $(u,v) \in Q_d$ there is some error we get by transmitting u instead of v, and we wish to minimise the sum of these errors over the whole edge set.

Explicitly, we wish to find a bijection $f : \{0,1\}^d \to [2^d]$ such that the sum

$$\sum_{(u,v) \in E(Q_d)} |f(u) - f(v)|$$

is minimised. Harper gave an explicit way to construct all functions achieving the minimum using the following observation: If we let S_ℓ be $f^{-1}([\ell])$ then we can express

$$\sum_{(u,v) \in E(Q_d)} |f(u) - f(v)| = \sum_{\ell \in [2^d]} \partial(S_\ell).$$

Indeed, an edge between $f^{-1}(i)$ and $f^{-1}(j)$, with say $i < j$, contributes $j - i$ to the left hand side, but the edge $(f^{-1}(i), f^{-1}(j))$ is in the boundary of $S_i, S_{i+1}, \dots, S_{j-1}$ and so contributes $j - i$ to the right hand side.

Harper showed how to construct functions which in fact minimised $\partial(S_\ell)$ for each ℓ individually, which then also clearly give a minimum for the sum.

In other words, this gives an ordering of the vertices of Q_d such that for each ℓ the edge boundary of the first ℓ vertices is the minimum size of an edge boundary of *any* set of ℓ vertices in Q_d.

A particular consequence of this is that the edge boundary for sets of size 2^k is minimised by a subcube of size k, that is a set of the form

$$S = \{(x_1, x_2, \dots, x_d) \colon x_i = y_i \text{ for all } i \in I\}$$

for some $I \subseteq [d]$ of size $d-k$ and a fixed vector $y \in \{0,1\}^I$. In this case we have that $|S| = 2^k$ and every point in S has $d-k$ neighbours outside of S, and so $|\partial(S)| = (d-k)2^k$. Using Shearer's lemma we can give a short proof of an isoperimetric inequality for Q_d which is tight for subcubes.

Lemma 4.20. *Let $S \subseteq V(Q_d)$ then*

$$|\partial(S)| \geq \log\left(\frac{2^d}{|S|}\right)|S|.$$

Proof. Let X be a random variable uniformly distributed over S and let $X = (X_1, \dots, X_d)$ be it's marginal distributions on the coordinates of Q_d. Since X is uniform, $\mathbb{H}(X) = \log|S|$. Let $X^{(i)} = (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_d)$ for each $i \in [d]$. By Shearer's lemma

$$\mathbb{H}(X) \leq \frac{1}{d-1}\sum_{i=1}^{d}\mathbb{H}(X^{(i)})$$

or equivalently

$$\mathbb{H}(X) \geq \sum_{i=1}^{d}(\mathbb{H}(X) - \mathbb{H}(X^{(i)})) = \sum_{i=1}^{d}\mathbb{H}(X_i|X^{(i)}).$$

If we let

$$S^{(i)} = \left\{(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_d) \colon \begin{cases}(x_1, \dots, x_{i-1}, 0, x_{i+1}, \dots, x_d) \in S, \text{ or} \\ (x_1, \dots, x_{i-1}, 1, x_{i+1}, \dots, x_d) \in S\end{cases}\right\}, \tag{4.2}$$

be the range of $X^{(i)}$, then by definition of conditional entropy

$$\mathbb{H}(X_i|X^{(i)}) = \sum_{x^{(i)} \in S^{(i)}}\mathbb{P}(X^{(i)} = x^{(i)})\mathbb{H}(X_i|X^{(i)} = x^{(i)}).$$

However, we can split into two cases. Given $(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_d) \in S^{(i)}$ either one or both of $(x_1, \dots, x_{i-1}, 0, x_{i+1}, \dots, x_d)$ and $(x_1, \dots, x_{i-1}, 1, x_{i+1}, \dots, x_d)$ are in S.

In the first case, $\mathbb{P}(X^{(i)} = x^{(i)}) = \frac{1}{|S|}$, since X was uniformly distributed, and $\mathbb{H}(X_i|X^{(i)} = x^{(i)})$ takes a single value, and hence has 0 entropy. In the second case, $\mathbb{P}(X^{(i)} = x^{(i)}) = \frac{2}{|S|}$ and $\mathbb{H}(X_i|X^{(i)} = x^{(i)})$ is uniformly distributed over 2 values, and hence has entropy one.

However, there is a clear bijection between points of the first type and edges from S to S^c in 'direction' i. Indeed, $x \in S$ has a neighbour outside s in direction i if and only if $x^{(i)}$ is of the first type. So, since each $x^{(i)}$ of the second type corresponds to two $x \in S$, Hence

$$\mathbb{H}(X_i|X^{(i)}) = (|S| - |\partial_i(S)|)\frac{1}{|S|}$$

where $\partial_i(S)$ is the edge boundary in direction i. It follows that

$$\log|S| = \mathbb{H}(X) \geq \sum_{i=1}^{d} \mathbb{H}(X_i|X^{(i)}) \geq \sum_{i=1}^{d} (|S| - |\partial_i(S)|)\frac{1}{|S|} \geq d - \frac{|\partial S|}{|S|}.$$

Re-arranging we see that

$$|\partial(S)| \geq |S|(d - \log|S|) = |S| \log\left(\frac{2^d}{|S|}\right)$$

as claimed. □

We can also consider other graphs. For example, consider the l_1 grid on $\mathbb{Z}^d$, that is, the graph where two points $x, y \in \mathbb{Z}^d$ are joined by an edge if $||x - y||_1 = 1$. Both the vertex and edge-isoperimetric problem have been fully solved in these graphs, in the strong sense that there is some ordering of the vertices such that initial segments of size i have the smallest boundary over all sets of size i for each i, as in Harper's theorem. For the vertex boundary the optimal sets grow like balls of fixed radius around the origin, whereas for the edge boundary the sets grow like cubes. If we write ∂_i as before for the edge boundary in direction i we see that $\partial(S) = \sum_i \partial_i(S)$. Given a cube of side length n we have $|S| = n^d$ and $\partial(S) = \sum_i \partial_i(S) = 2dn^{d-1} = 2d|S|^{\frac{(d-1)}{d}}$.

Again, we can use Shearer's Lemma, or in this case the Loomis-Whitney inequality directly, to give an edge-isoperimetric inequality which is tight for cubes.

Theorem 4.21. *Let $S \subseteq \mathbb{Z}^d$ be finite. Then*

$$|\partial(S)| \geq 2d|S|^{\frac{(d-1)}{d}}.$$

Proof. We can think of each point in the projected set $S_{[d]\setminus\{i\}}$ as corresponding to multiple points in $S \cap L$, where L is an infinite line in $\mathbb{Z}^d$ formed by fixing all other co-ordinates and letting the ith co-ordinate vary. For each point in the projected set there are at least two edges in the boundary ∂_i in the ith direction, coming from the points in $S \cap L$ with largest and smallest ith co-ordinate respectively.

Hence, $\partial(S) = \sum_i \partial_i(S) \geq \sum_i 2|S_{[d]\setminus\{i\}}|$. However by the arithmetic geometric mean inequality

$$\sum_i 2|S_{[d]\setminus\{i\}}| \geq 2d\left(\prod_i |S_{[d]\setminus\{i\}}|\right)^{\frac{1}{d}}.$$

Hence, by the Loomis-Whitney inequality

$$\partial(S) \geq 2d\left(\prod_i |S_{[d]\setminus\{i\}}|\right)^{\frac{1}{d}} \geq 2d|S|^{\frac{(d-1)}{d}}.$$

□

This approach also allows one to prove a 'stability' type result for edge-isoperimetry in $\mathbb{Z}^d$. Recall from the exercise classes that we showed that given a random variable X taking values

on $\mathcal{X}$ if U is the uniform random variable on $\mathcal{X}$ then we can bound $\mathbb{H}(X)$ from above with some function of $||X-U||_1$, this was Pinsker's inequality. Using Pinsker's inequality one can show the following 'stability' version of the uniform covers theorem:

Theorem 4.22. *For every $d \geq 2$ there is some constant $c(d) > 0$ such that the following holds. Let $S \subset \mathbb{Z}^d$ and let $\mathcal{C} \subset 2^{[d]}$ be a k-uniform cover such that for every $i \neq j \in [n]$ there are at least $\alpha > 0$ many $C \in \mathcal{C}$ containing i but not j. If*

$$|S| \geq (1-\varepsilon)\left(\prod_{C\in\mathcal{C}} |S_C|\right)^{\frac{1}{k}}$$

then there exists a box $B \subset \mathbb{Z}^d$ such that

$$|S \triangle B| \leq c\frac{k}{\alpha}\varepsilon|S|.$$

Using the above one can follow the proof of Theorem 4.21 to show that any almost optimal shape is close in symmetric difference to some box. With some careful combinatorial arguments, one can deduce that this box in fact has to be very close to a cube, leading to the following stability version of Theorem 4.21,

Theorem 4.23. *Let $S \subseteq \mathbb{Z}^d$ be finite, such that*

$$|\partial(S)| \leq (1+\varepsilon)2d|S|^{\frac{(d-1)}{d}},$$

then there exists a cube $C \subseteq \mathbb{Z}^d$ such that

$$|S \triangle C| \leq 72d^{\frac{5}{2}}\sqrt{\varepsilon}|S|.$$

4.3.4 Counting Matroids

A matroid is very general combinatorial object which models a notion of independence, in the sense of vectors in a finite dimensional vector space. One classic example of this comes from the *cycle matroid* of a graph, where sets of edges are considered independent if they form a forest.

We will use the following definition of a matroid, mostly since it is the most compact, but there are many equivalent axiom schemes defining matroids in terms of their independent sets/cycles/rank functions etc. For us, a *matroid* will be a pair $(E, \mathcal{B})$ where E is a finite set and $\mathcal{B}$ is a non-empty collection of subsets of E, which we call *bases*, which satisfy the following base exchange axiom

$$\text{For all } B, B' \in \mathcal{B} \text{ and all } e \in B \setminus B', \text{ there exists an } f \in B' \setminus B \text{ such that } B - e + f \in \mathcal{B} \quad (4.3)$$

This axiom implies that every $B \in \mathcal{B}$ has the same cardinality, which we call the *rank* of the matroid. Let us write $m_{n,r}$ for the number of matroids of rank r on $E = [n]$ and m_n for the number of matroids on $E = [n]$.

Clearly $m_n \leq 2^{2^n}$, which is equivalent to $\log\log m_n \leq n$, and Piff gave an improved bound of $\log\log m_n \leq n - \log n + O(1)$. On the other hand, this isn't too far from the true value, as Knuth showed that

$$m_n \geq 2^{\frac{1}{n}\binom{n}{n/2}}$$

which means that $\log\log m_n \geq n - \frac{3}{2}\log n + O(1)$. Both of these bounds were shown in the 70s, and since then Piff's bound has been improved to almost match Knuth's lower bound. In this section we will present a very short argument that gives a slightly weaker bound than that of Knuth, whose proof was quite involved.

Given a set E and $r \leq |E|$ let us consider the set of collections of r-sets of E which define bases of a matroid, together with the empty set. That is

$$\mathcal{M}_{E,r} = \{\mathcal{B} \subseteq E^{(r)} \colon \mathcal{B} \text{ satisfies (4.3)}\}.$$

Note that, since we allow $\emptyset \in \mathcal{M}_{E,r}$ we have that $|\mathcal{M}_{E,r}| = m_{|E|,r} + 1$. It will be useful to identify the elements $\mathcal{B}$ of $\mathcal{M}_{E,r}$ with their characteristic vectors in the space $\mathbb{Z}_2^{E^{(r)}}$, where each coordinate corresponds to an r-set of E.

Given a matroid $(M, \mathcal{B})$ and a subset $T \subseteq E$ which is contained in some basis of M, then *contracting* T gives rise to a new matroid $M/T := (E \setminus T, \mathcal{B}/T)$ where

$$\mathcal{B}/T := \{B \setminus T \colon B \in \mathcal{B}, T \subseteq B\}$$

It follows from (4.3) that M/T is in fact a matroid, and so $\mathcal{B}/T \in \mathcal{M}_{E\setminus T, r-t}$. However, even if T is not contained in some basis of M, then the set $\mathcal{B}/T = \emptyset$ and so, even though $(E \setminus T, \mathcal{B}/T)$ is no longer a matroid, $\mathcal{B}/T \in \mathcal{M}_{E\setminus T, r-t}$.

Note also that we can view the contraction operation as a projection in $\mathbb{Z}_2^{E^{(r)}}$, where contracting T is equivalent to projecting onto the $\binom{|E|-t}{r-t}$ coordinates corresponding to r-sets of E containing T. This observation will allow us to use Shearer's Lemma to bound the number of matroids of a fixed rank.

Lemma 4.24. *Let $0 \leq t \leq r \leq n$. Then*

$$\frac{\log(m_{n,r}+1)}{\binom{n}{r}} \leq \frac{\log(m_{n-t,r-t}+1)}{\binom{n-t}{r-t}}$$

Proof. Let $n = |E|$ and let us consider a random variables X which is uniformly distributed over $\mathcal{M}_{E,r}$, where we will think of X as taking values in $\{0,1\}^{E^{(r)}}$. As always, since X is uniformly distributed, we have that

$$\mathbb{H}(X) = \log|\mathcal{M}_{E,r}| = \log(m_{n,r}+1).$$

Given $T \in E^{(t)}$ let us denote by X^T the projection of X onto the $\binom{|E|-t}{r-t}$ coordinates corresponding to r-sets of E containing T. Since X^T takes values in $\mathcal{M}_{E\setminus T, r-t}$ it follows that

$$\mathbb{H}(X^T) \leq \log|\mathcal{M}_{E\setminus T,r-t}| = \log(m_{n-t,r-t}+1).$$

If we let $A(T) = \{S \in E^{(r)} \colon T \subseteq S\}$ the we see that $\{A(T) \colon T \in E^{(t)}\}$ covers each element of $E^{(r)}$ precisely $\binom{r}{t}$ times, since each $S \in E^{(r)}$ is counted once for each of its subsets T of size t. Hence, by Lemma 4.14

$$\log(m_{n,r}+1) = \mathbb{H}(X) \leq \frac{1}{\binom{r}{t}} \sum_{T \in E^{(t)}} \mathbb{H}(X^T) \leq \frac{\binom{n}{t}}{\binom{r}{t}} \log(m_{n-t,r-t}+1).$$

However it is easy to verify that

$$\frac{\binom{n}{t}}{\binom{r}{t}} = \frac{\binom{n}{r}}{\binom{n-t}{r-t}}$$

from which the result follows. □

Lemma 4.24 allows us to use bounds for the number of lower rank matroids to bound the number of higher rank matroids. For example, since we know that there is exactly one matroid of rank 0, namely $\mathcal{B} = \emptyset$, we could use Lemma 4.24 with $t = r$ to see that

$$\log(m_{n,r} + 1) \leq \binom{n}{r} \log(m_{n-r,0} + 1) = \binom{n}{r}$$

which isn't an especially good bound, it is essentially the fact that $\emptyset \neq \mathcal{B} \subseteq [n]^{(r)}$. However, if we use matroids of rank 1 we can see that $m_{n,1} = 2^n - 1$, since any collection of singletons, apart from the empty set, satisfies (4.3). Hence, taking $t = r - 1$

$$\log(m_{n,r} + 1) =\leq \binom{n}{r} \frac{\log(m_{n-t,1} + 1)}{\binom{n-r}{1}} = \binom{n}{r} \frac{n}{n-r},$$

a slightly better bound. So, to get a good bound on $m_{n,r}$ we just need to find a good bound on $m_{n,k}$ for some small k, and it turns out $k = 2$ is actually sufficient for our purposes.

Lemma 4.25.

$$\log(m_{n,2} + 1) \leq (n + 1) \log(n + 1).$$

Proof. We first note that to every matroid of rank 2 we can associate a set $E_0 \subseteq [n]$ and a partition $\{E_1, \ldots, E_k\}$ of $[n] \setminus E_0$ such that

$$\mathcal{B} = \{\{e_1, e_2\} \colon e_1 \in E_i, e_2 \in E_j, 0 < i < j \leq k\}.$$

Indeed, let $E_0 := \{e \in E \colon e \notin B \text{ for all } B \in \mathcal{B}\}$. Let us define a relation on $E \setminus E_0$ by $e \sim f$ if $\{e, f\} \notin \mathcal{B}$. Note that $\sim$ is transitive. Indeed, suppose that $e \sim f$ and $f \sim g$ but $e \not\sim g$. By definition $\{e, g\} \in \mathcal{B}$ and since $f \in E \setminus E_0$ there exists $h \in E \setminus E_0$ such that $\{f, h\} \in \mathcal{B}$. By applying (4.3) to $h \in \{f, h\}$ and $\{e, g\}$ we see that $\{e, f\}$ or $\{f, g\} \in \mathcal{B}$, contradicting our assumption that $e \sim f$ and $f \sim g$.

Hence $\sim$ is an equivalence relation, and if we let $E_1, \ldots, E_k$ be the equivalence classes of $\sim$ on $[n] \setminus E_0$ we see that the claim about $\mathcal{B}$ holds. However, clearly different matroids determine different pairs $(E_0, \{E_1, \ldots, E_k\})$ and so we can bound $m_{n,2}$ by the number of such pairs. In fact, since no matroid determines the pair $([n], \emptyset)$, we can even bound $|\mathcal{M}_{[n],2}| = m_{n,2} + 1$ by this amount.

Clearly the function mapping the pair $(E_0, \{E_1, \ldots, E_k\})$ to the partition $\{E_0 \cup \{n+1\}, E_1, \ldots, E_k\}$ of $[n + 1]$ is injective, and hence this number is at most the number of partitions of $[n + 1]$. A very crude bound of $(n + 1)^{n+1}$ for this number will be sufficient for our purposes, although better bounds are known for the *Bell numbers* as they are known.

Hence

$$m_{n,2} + 1 = |\mathcal{M}_{[n],2}| \leq (n + 1)^{(n+1)}$$

from which the result follows. □

Using this we can give the following bound on m_n.

Theorem 4.26.

$$\log\log m_n \leq n - \frac{3}{2}\log n + \log\log n + O(1).$$

Proof. Applying Lemma 4.24 with $t = r-2$ we see that for each $r \leq n$

$$\begin{aligned}\log m_{n,r} \leq \log(m_{n,r}+1) &\leq \binom{n}{r}\frac{\log(m_{n-r+2,2}+1)}{\binom{n-r+2}{2}} \\ &\leq \binom{n}{r}\frac{(n+1)\log(n+1)}{\binom{n-r+2}{2}} \\ &= \binom{n+2}{r}\frac{2\log(n+1)}{n+2}.\end{aligned}$$

Since $m_n = \sum_{r\leq n} m_{n,r}$, it follows that $m_n \leq (n+1)\max_r m_{n,r}$ and so

$$\log m_n \leq \log(n+1) + \max_r \log(m_{n,r}).$$

We note that $\binom{n+2}{r}$ is maximised at $r = \lfloor\frac{(n+2)}{2}\rfloor$, and from the inequality

$$\binom{2m}{m} \leq \frac{2^{2m}}{\sqrt{2m}}$$

it follows that

$$\binom{n+2}{\lfloor\frac{(n+2)}{2}\rfloor} = O(2^n n^{-\frac{1}{2}}).$$

Hence, we can conclude

$$\log m_n \leq \log(n+1) + \binom{n+2}{\lfloor\frac{(n+2)}{2}\rfloor}\frac{2\log(n+1)}{n+2} = O(\log n 2^n n^{-\frac{3}{2}}).$$

Hence

$$\log\log m_n \leq n - \frac{3}{2}\log n + \log\log n + O(1).$$

□

It is tempting to hope that if we can get a better estimate for $m_{n,3}$ or $m_{n,4}$ we could even close the gap between this bound and Knuth's lower bound. However, unfortunately there are good lower bounds for the quantity $m_{n,k}$ for fixed k as $n \to \infty$ which imply that this approach cannot get rid of this $\log\log n$ term.

In fact, Lemma 4.24 holds in a slightly more general context. If we look at the proof, we never really used the fact that we were considering the set of *all* matroids of rank r, just that when we contracted we stayed within the class of matroids that we care about. Indeed the exact same proof shows that for any contraction-closed class of matroids $\mathcal{F}$ Lemma 4.24 still holds if we let $m_{n,r}$ be the set of matroids of rank r on a ground set $[n]$ in $\mathcal{F}$. This can allow us to prove interesting results about the relative density of certain contraction-closed classes of matroids in the class of all matroids.

A natural example of a contraction-closed class of matroids is the set of matroids not containing a fixed matroid as a minor. For example, suppose we let $U_{2,k}$ be the uniform matroid of rank 2 on k elements and let $Ex(U_{2,k}) = \{M \colon N \not\leq M\}$ be the set of matroids not containing $U_{2,k}$ as a minor. It is relatively easy to get a good bound on the number of such matroids of rank 3, specifically that

$$m'(n,3) \leq c_k(k^2)^n. \tag{4.4}$$

Indeed, suppose M is a simple (no loops or parallel edges) matroid of rank 3 without $U_{2,k}$ as a minor and let e be an edge of M. A quick detour into some matroid theory is necessary. Given a matroid M with a rank function r we can define the closure of a set X to be $\{f \colon r(X+f) = r(X)\}$. When M is rank 3 matroid the closure of a rank 2 set is called a *line*.

Consider the set of lines in which e is contained in. If there are at least k different lines, then when if we contract e, taking an edge from each of these lines will give a $U_{2,k}$ minor. Similarly, every line contains at most $k-1$ points, since if we restrict to a line of size k' we get a $U_{2,k'}$ minor. However, M is simple, every other edge f lies in a line with e and so $|E| \leq 1 + (k-1)(k-2)$.

Hence there is a glocal upper bound c_k of the number of simple matroids of rank 3 witjout a $U_{2,k}$ minor. Since every matroid is determined by its simplification and the assignment of its non-loop edges to parallel classes, (4.4) follows.

By the same methods as before, (4.4) together with Lemma 4.24 implies that

$$\log(m'_{n,r} + 1) \leq \binom{n}{r} \frac{12\log(k)}{n^2}(1 + o(1)),$$

which is sufficiently strong to show that almost all matroids contain $U_{2,k}$ as a minors.

4.3.5 Inequalities

Shearer's Lemma, in various guises, seems to crop up in many different contexts. In fact, many well known inequalities can be considered as specific cases of Shearer's Lemma, in a broad setting.

Let us give an example, before proving a generalisation of Shearer's Lemma that will prove a whole host of well-known inequalities. Suppose we have positive integers $a_1, \ldots, a_n$ and $b_1 \ldots, b_n$. Take pairwise disjoint subsets $A_1, A_2, \ldots, A_n$ of $\mathbb{Z}$ with $|A_i| = a_i$, and pairwise disjoint subsets $B_1, B_2, \ldots, B_n$ with $|B_i| = b_i$. Let $S_i = A_i \times B_i \subset \mathbb{Z}^2$ and $S = \bigcup_i S_i$. Note that, S is a set of $\sum_k a_k b_k$ many points in $\mathbb{Z}^2$.

Suppose we have pairwise disjoint subsets $A_1, A_2, \ldots, A_n$ of $\mathbb{Z}$ with $|A_i| = a_i$, and pairwise disjoint subsets $B_1, B_2, \ldots, B_n$ with $|B_i| = b_i$. Let $S_i = A_i \times B_i \subset \mathbb{Z}^2$ and $S = \bigcup_i S_i$. Note that, S is a set of $\sum_k a_k b_k$ many points in $\mathbb{Z}^2$.

We want to choose two points in S uniformly at random, but with the restriction that they both lie in the same S_i. Formally to do this let's consider a random variable I which chooses an index from $[1, n]$ with

$$\mathbb{P}(I = i) = \frac{a_i b_i}{\sum_k a_k b_k}.$$

Then, let $X = ((X_1, Y_1), (X_2, Y_2))$ be a pair of points chose uniformly at random from R_I. Note that both (X_1, Y_1) and (X_2, Y_2) are uniformly distributed amongst all the $\sum_k a_k b_k$ points in R, but they are not independent.

Now, since (X_1, Y_1) and (X_2, Y_2) are uniformly distributed we know that

$$\mathbb{H}(X_1, Y_1) + \mathbb{H}(X_2, Y_2) = \log\left(\sum_k a_k b_k\right) + \log\left(\sum_k a_k b_k\right) = 2\log\left(\sum_k a_k b_k\right).$$

However let us instead consider the distribution of the pairs (X_1, X_2) and (Y_1, Y_2). The first, whilst not necessarily uniform, is distributed over $\sum_k a_k^2$ values, and the second over $\sum_k b_k^2$ many values and so

$$\mathbb{H}(X_1, X_2) + \mathbb{H}(Y_1, Y_2) \leq \log\left(\sum_k a_k^2\right) + \log\left(\sum_k b_k^2\right).$$

However, we can see that $\mathbb{H}(X_1, Y_1) + \mathbb{H}(X_2, Y_2) = \mathbb{H}(X_1, X_2) + \mathbb{H}(Y_1, Y_2)$. Indeed, we note that (X_i, Y_i) determines I, but also X_i is independent of Y_i conditioned on the value of I, and so

$$\begin{aligned}
\mathbb{H}(X_1, Y_1) + \mathbb{H}(X_2, Y_2) &= \mathbb{H}(X_1, Y_1, I) + \mathbb{H}(X_2, Y_2, I) \\
&= \mathbb{H}(I) + \mathbb{H}(X_1|I) + \mathbb{H}(Y_1|X_1, I) + \mathbb{H}(I) + \mathbb{H}(X_2|I) + \mathbb{H}(Y_2|X_2, I) \\
&= \mathbb{H}(I) + \mathbb{H}(X_1|I) + \mathbb{H}(Y_1|I) + \mathbb{H}(I) + \mathbb{H}(X_2|I) + \mathbb{H}(Y_2|I)
\end{aligned}$$

however, X_1 is independent of X_2 given I, and similarly for Y_1 and Y_2 and so we can argue in reverse that

$$\begin{aligned}
&= \mathbb{H}(I) + \mathbb{H}(X_1|I) + \mathbb{H}(Y_1|I) + \mathbb{H}(I) + \mathbb{H}(X_2|I) + \mathbb{H}(Y_2|I) \\
&= \mathbb{H}(I) + \mathbb{H}(X_1|I) + \mathbb{H}(Y_1|I) + \mathbb{H}(I) + \mathbb{H}(X_2|X_1, I) + \mathbb{H}(Y_2|Y_1 I) \\
&= \mathbb{H}(X_1, X_2, I) + \mathbb{H}(Y_1, Y_2, I) = \mathbb{H}(X_1, X_2) + \mathbb{H}(Y_1, Y_2).
\end{aligned}$$

So we can conclude that

$$2\log\left(\sum_k a_k b_k\right) = \mathbb{H}(X_1, Y_1) + \mathbb{H}(X_2, Y_2) = \mathbb{H}(X_1, X_2) + \mathbb{H}(Y_1, Y_2) \leq \log\left(\sum_k a_k^2\right) + \log\left(\sum_k b_k^2\right).$$

By taking powers of both sides we get

$$\left(\sum_k a_k b_k\right)^2 \leq \left(\sum_k a_k^2\right)\left(\sum_k b_k^2\right)$$

which you should recognise as the Cauchy-Schwartz inequality.

So, how can we generalise this idea? As we saw on the example sheet, we can think of Shearer's Lemma as the following result about set systems (and in fact this is what Shearer originally proved).

Theorem 4.27. *Let $t \in \mathbb{N}$, $H = (V, E)$ be a hypergraph and $F_1, \ldots, F_r$ be subsets of V such that every vertex in V belongs to at least t of the sets F_i. Let $H_i = (V, E_i)$ be the projection hypergraphs, where $E_i = \{e \cap F_i \colon e \in E\}$. Then*

$$|E|^t \leq \prod_i |E_i|.$$

We will prove a weighted version of the above lemma, whose proof will closely follow our example, and then show how it can be used to deduce some other inequalities.

Lemma 4.28 (Weighted Shearer's Lemma)**.** *Let H, E, V, t and $F_1, \dots, F_r$ be as in Lemma 4.27 and let $w_i : E_i \to \mathbb{R}^+$ be a non-negative weight function for each E_i. Then*

$$\left(\sum_{e\in E}\prod_{i=1}^{r} w_i(e_i)\right)^t \leq \prod_{i=1}^{r}\sum_{e_i\in E_i} w_i(e_i)^t$$

Remark 4.29. *Firstly let us note that setting $w_i \equiv 1$ for all i gives us Lemma 4.27. Secondly, we can recover our example by taking the complete 1-uniform hypergraph $V = E = [n]$,taking $F_1 = F_2 = V$, so that $t = 2$, and letting $w_1(k) = a_k$ and $w_2(k) = b_k$ for all $k \in [n]$. Then the lemma allows us to conclude*

$$\left(\sum_{k\in[n]} a_k b_k\right)^2 \leq \left(\sum_{k\in[n]} a_k^2\right)\left(\sum_{k\in[n]} b_k^2\right).$$

Proof. For ease of presentation we will assume that $V = [n]$. Also, it will be sufficient to prove the result when all weights are positive integers. Let us create a multi-hypergraph from H by taking each edge $e \in E$ with multiplicity $\prod_i w(e_i)$, calling them $e^{(c_1,\dots,c_r)}$ where $1 \leq c_i \leq w_i(e_i)$, we call this graph $H' = (V, E')$. Similarly we will create from each H_i a multi-hypergraph $H_i' = (V, E_i')$ by taking each edge with multiplicity $w_i(e_i)$, calling them e_i^c where $1 \leq c \leq w_i(e_i)$.

Consider a random variable Y which is uniformly distributed over E', and consider the following random variable:

- $X = (X_1, \dots, X_n)$ is the characteristic vector of the edge Y (i.e $X_k = 1$ if and only if $k \in Y$);
- $C = (C_1, C_2, \dots, C_r)$ is the index of Y.

Note that Y is determined by and determines (X, C).

Since Y is uniform on E' we have that

$$\mathbb{H}(Y) = \log\left(\sum_{e\in E}\prod_{i=1}^{r} w_i(e_i)\right).$$

In a similar fashion to the example above, we would like to say that, if we pick t copies of Y independently, we can think of each copy of Y as being a 'vector' given by the projection of Y onto F_i (both in terms of the edge, and the index C_i). By regrouping these terms, we might hope to compare the entropy of these t copies of Y with the entropy of the collection $Y^1, \dots, Y^r$, where each Y^i is the t independent projections of Y onto F_i.

So, let us also define some random variables $Y^1, \dots, Y^r$ where Y^i is distributed on $(E_i')^t$ as follows: We choose an edge Z uniformly at random from E' and pick independently and with replacement t 'copies' of the edge $Z \cap F_i$ in E_i'. That is we have

- $X^i = (X^i_1, \dots, X^i_n)$ is the characteristic vector of the edge $Z \cap F_i$;
- $C^i = (C^i_1, C^i_2, \dots, C^i_t)$ are the indices of the t 'copies' of $Z \cap F_i$ we choose;

and Y^i is determined by and determines (X^i, C^i).

Let us note a few properties of these random variables. Firstly, the joint distribution of $(X^i_k \colon k \in F_i)$ is the same as the joint distribution of $(X_k \colon k \in F_i)$. Also, conditioned on a fixed value of X^i, $\{C^i_1, C^i_2, \dots, C^i_t)\}$ is mutually independent, and each is uniformly distributed between 1 and $w_i(X^i)$. Similarly, conditioned on a fixed value of X, C_i is distributed uniformly between 1 and $w_i(X \cap F_i)$.

Now, each Y^i can take at most $\sum_{e_i \in E_i} w_i(e_i)^t$ different values and so

$$\mathbb{H}(Y^i) \leq \log \left(\sum_{e_i \in E_i} w_i(e_i)^t \right).$$

Hence it will be sufficient to show that

$$t\mathbb{H}(Y) \leq \sum_{i=1}^{r} \mathbb{H}(Y^i).$$

However, by the chain rule

$$\mathbb{H}(Y) = \mathbb{H}(X, C) = \sum_{m=1}^{n} \mathbb{H}(X_m | X_{[m-1]}) + \sum_{i=1}^{r} \mathbb{H}(C_i | X, C_{[i-1]}) = \sum_{m=1}^{n} \mathbb{H}(X_m | X_{[m-1]}) + \sum_{i=1}^{r} \mathbb{H}(C_i | X),$$

since the C_i are mutually independent given X.

Similarly we can express

$$\begin{aligned} \mathbb{H}(Y^i) = \mathbb{H}(X^i, C^i) &= \sum_{m \in F_i} \mathbb{H}(X^i_m | X^i_\ell \colon \ell < m, \ell \in F_i) + \sum_{k=1}^{t} \mathbb{H}(C^i_k | X^i, C^i_{[k-1]}) \\ &= \sum_{m \in F_i} \mathbb{H}(X^i_m | X^i_\ell \colon \ell < m, \ell \in F_i) + \sum_{k=1}^{t} \mathbb{H}(C^i_k | X^i) \end{aligned}$$

where again we have use the fact that the C^i_k are independent given X^i. However, as we noted $(X^i_m \colon m \in F_i)$ has the same distribution as $(X_m \colon m \in F_i)$, and also $(C^i_k | X^i = e_i)$ has the same distribution as $(C_i | X = e)$ for all $e \in E$. Hence, since C^i_k is only depends on $(X^i_m \colon m \in F_i)$, it follows that

$$\sum_{m \in F_i} \mathbb{H}(X^i_m | X^i_\ell \colon \ell < m, \ell \in F_i) + \sum_{k=1}^{t} \mathbb{H}(C^i_k | X^i) = \sum_{m \in F_i} \mathbb{H}(X_m | X_\ell \colon \ell < m, \ell \in F_i) + t\mathbb{H}(C_i | X)$$

Hence we can write

$$\begin{aligned}\sum_{i=1}^{r} h(Y^i) &= \sum_{i=1}^{r}\left(\sum_{m\in F_i}\mathbb{H}(X_m|X_\ell\colon \ell<m, \ell\in F_i) + t\mathbb{H}(C_i|X)\right)\\ &\geq \sum_{i=1}^{r}\left(\sum_{m\in F_i}\mathbb{H}(X_m|X_{[m-1]})\right) + t\sum_{i=1}^{r}\mathbb{H}(C_i|X)\\ &\geq t\sum_{m=1}^{n}\mathbb{H}(X_m|X_{[m-1]}) + t\sum_{i=1}^{r}\mathbb{H}(C_i|X)\\ &= \mathbb{H}(Y).\end{aligned}$$

□

In fact it will be useful to consider an even more general weighted version, where we also allow the integer t to be 'fractionally covered'

Lemma 4.30. *Let H, E, V, t, F_i and w_i be as in Lemma 4.28. Let $\alpha = (\alpha_1, \dots, \alpha_r)$ be a vector of non-negative weights such that for each $v \in V$*

$$\sum_{i\colon v\in F_i}\alpha_i \geq 1$$

(i.e α is a 'fractional cover' of the hypergraph whose edges are the F_i). Then,

$$\sum_{e\in E}\prod_{i=1}^{r} w_i(e_i) \leq \prod_{i=1}^{r}\left(\sum_{e_i\in E_i} w_i(e_i)^{\frac{1}{\alpha_i}}\right)^{\alpha_i}.$$

This lemma can can deduce from Lemma 4.28 by constructing an appropriate multi-hypergraph for which the number of copies of each edge is determined by its weight. As we will see, many classical inequalities can be deduced from Lemma 4.30 by choosing the right hypergraph and fractional covering. We note also, as we will see on the example sheet, one can give a necessary condition for equality to hold in these generalised Lemma's, which also give information about the equality cases of these inequalities.

Theorem 4.31 (Hölder's Inequality)**.** *Let $a_1, \dots, a_k, b_1, \dots, b_k \in \mathbb{R}^+$ and $\lambda \in (0,1)$. Then,*

$$\sum_k a_k b_k \leq \left(\sum_k a_k^{\frac{1}{\lambda}}\right)^{\lambda}\left(\sum_k b_k^{\frac{1}{(1-\lambda)}}\right)^{(1-\lambda)}$$

Proof. We choose, as in the case of Cauchy-Schwartz, H to be the complete 1-uniform hypergraph on $[n]$, $F_1 = F_2 = [n]$, $w_1(k) = a_k$, $w_2(k) = b_k$) and let $\alpha = (\lambda, 1-\lambda)$. Note that the conditions of Lemma 4.30 are satisfied, and hence we can conclude that the above inequality holds. □

Theorem 4.32 (Generalised Hölder's Inequality)**.** *Let $\gamma_1, \dots \gamma_k \in \mathbb{R}^+$ be such that $\sum \frac{1}{\gamma_i} = 1$ and let $a_{ij} \in \mathbb{R}^+$ for $1 \leq i \leq n$ and $1 \leq j \leq r$. Then,*

$$\sum_{i=1}^{n}\prod_{j=1}^{r} a_{ij} \leq \prod_{j=1}^{r}\left(\sum_{i=1}^{n} a_{ij}^{\gamma_j}\right)^{\frac{1}{\gamma_j}}$$

Proof. As before let us choose H to be the complete 1-uniform hypergraph on $[n]$, $F_1 = F_2 = \ldots = F_r = [n]$ and $w_j(i) = a_{ij}$ for each i, j. Finally let us take the fractional cover $\alpha = (\frac{1}{\gamma_1}, \ldots, \frac{1}{\gamma_r})$. As before it is easy to see that the conditions of Lemma 4.30 are satisfied and hence we can conclude that the above inequality holds. which is to say □

Theorem 4.33 (Monotonicty of ℓ_p norms)**.** *Let $a_1, \ldots a_n \in \mathbb{R}^+$ and let $t < r \in \mathbb{N}$. Then*

$$\left(\sum_{k=1}^{n} a_k^r\right)^{\frac{1}{r}} \leq \left(\sum_{k=1}^{n} a_k^t\right)^{\frac{1}{t}}.$$

Proof. Let $V = [r]\times[n]$ and $E = \{e^{(k)} = [r]\times\{k\} \colon k \in [n]\}$. Let us take $F_i = \{i, i+1, \ldots, i+t-1\}$ mod $r\times[n]$ for $i = 1, \ldots, r$, $\alpha = (\frac{1}{t}, \frac{1}{t}, \ldots, \frac{1}{t})$ and let $w_i(e_i^{(k)}) = a_k$ for each i, k. Then, by Lemma 4.30

$$\sum_{k=1}^{n} a_k^r = \sum_{e\in E}\prod_{i=1}^{r} w_i(e_i) \leq \prod_{i=1}^{r}\left(\sum_{e_i\in E_i} w_i(e_i)^{\frac{1}{\alpha_i}}\right)^{\alpha_i} = \prod_{i=1}^{r}\left(\sum_{k=1}^{n} a_k^t\right)^{t},$$

which re-arranges to the desired inequality. □

Note that, from the above inequality it is easy to prove that this holds for arbitrary $q < p \in \mathbb{Q}^+$, and hence by a limiting argument for arbitrary $q < p \in \mathbb{R}^+$.

Theorem 4.34. *Let A, B, C be a $i \times j$, $j \times k$ and $k \times i$ matrices respectively. Then*

$$Tr(ABC) \leq \sqrt{Tr(AA^t)\,Tr(BB^t)\,Tr(CC^t)}.$$

Proof. Let H be a complete tripartite hypergraph graph on vertex sets I, J, K with $|I| = i, |J| = j, |K| = k$. Let $F_1 = I \cup J$, $F_2 = J \cup K$ and $F_3 = K \cup I$ and let $\alpha = (1/2, 1/2, 1/2)$. Finally for an edge $e = \{r, s, t\}$ let $w_1(e_1) = a_{rs}$, $w_2(e_2) = b_{st}$ and $w_3(e_3) = c_{tr}$. Then, by Lemma 4.30

$$\sum_{r,s,t} a_{rs}b_{st}c_{tr} = \sum_{e\in E}\prod_{i=1}^{r} w_i(e_i) \leq \prod_{i=1}^{r}\left(\sum_{e_i\in E_i} w_i(e_i)^{\frac{1}{\alpha_i}}\right)^{\alpha_i} = \left(\sum_{r,s} a_{rs}^2\right)^{\frac{1}{2}}\left(\sum_{s,t} b_{st}^2\right)^{\frac{1}{2}}\left(\sum_{t,r} c_{tr}^2\right)^{\frac{1}{2}},$$

which can be seen to be equivalent to the claimed inequality. □

Lemma 4.30 has an obvious continuous analogue, that in fact was shown by Finner well before the discrete version here was considered by Freidgut, from which many interesting functional inequalities can be deduced in a similar fashion. Lemma 4.28 was first considered by Friedgut and Rödl, who used it to give an entropy based proof of a hypercontractive inequality, a result that has been extremely useful in the study of boolean functions, by considering an appropriate hypergraph.

4.4 Embedding Graphs

Suppose we have a fixed graph H and a graph G with a fixed number ℓ of edges, how many 'copies' of H can there be in G? Of course, to answer this question we have to choose what we mean by 'copy', and in this section we will consider *embeddings* of graphs. An *embedding* of H

into G is an injective function $f : V(H) \to V(G)$ which preserves adjacency, in other words an injective homomorphism. Given a pair of graphs H and G we will write $\text{embed}(H, G)$ for the number of embeddings of H into G and

$$\text{embed}(H, \ell) := \max_{e(G)=\ell} \text{embed}(H, G).$$

The problem we will consider in this section is giving a good upper bound for $\text{embed}(H, \ell)$. As a simple example, let us consider the embedding number of a triangle $\text{embed}(K_3, \ell)$.

For any graph G with ℓ edges and $v \in V(G)$, v can be the 'top vertex' in at most 2ℓ triangles, by considering where the 'bottom' edge of the triangle is, but also can be the top vertex of at most $d(v)(d(v)-1) \leq d(v)^2$ many triangles, by considering where the edges adjacent to v are mapped. Since $\min\{d(v)^2, 2\ell\} \leq d(v)\sqrt{2\ell}$ it follows that

$$\text{embed}(K_3, G) \leq \sum_{v \in V(G)} d(v)\sqrt{2\ell} = 2\sqrt{2}\ell^{\frac{3}{2}},$$

and hence $\text{embed}(K_3, \ell) \leq 2\sqrt{2}\ell^{\frac{3}{2}}$. However this bound can be seen to be the correct order of magnitude since the complete graph on $\sqrt{2\ell}$ vertices contains at least $\sqrt{2\ell}(\sqrt{2\ell}-1)(\sqrt{2\ell}-2) \approx 2\sqrt{2}\ell^{\frac{3}{2}}$ embeddings of K_3, and has approximately ℓ edges.

There is an 'obvious' way to use entropy to try and bound $\text{embed}(H, G)$ from above. We let X be a embedding of H into G chosen uniformly at random, and then any bound on $\mathbb{H}(X)$ can be used to bound $\text{embed}(H, G)$. So, how might we bound $\mathbb{H}(X)$?

Well, let us assume that $e(G) = \ell$ and $V(H) = \{v_1, \dots, v_n\}$. If we let X_i be the image of v_i under the embedding X for each i then clearly X is determined by and determines $(X_i \colon i \in [n])$ and so we can instead estimate $\mathbb{H}(X_i \colon i \in [n])$. Now, we'd like to use one of our results on entropy to split this quantity up further, into more local random variables that we can estimate. We don't have too much control over where individual vertices are mapped, but we know that the range of an edge (X_i, X_j) is at most 2ℓ, and hence $\mathbb{H}(X_i, X_j) \leq \log(2\ell)$ for every $(v_i, v_j) \in E(H)$.

So, given a multi-set of edges $\mathcal{F}$ which covers each vertex m many times we can use Shearer's lemma to say

$$\mathbb{H}(X) \leq \frac{1}{m} \sum_{(i,j) \in \mathcal{F}} \mathbb{H}(X_i, X_j) \leq \frac{1}{m} |\mathcal{F}| \log(2\ell).$$

Clearly for any fixed H there is some family $\mathcal{F}$ minimising the quantity $\frac{|\mathcal{F}|}{m}$ and choosing such a family gives us a bound of the type $\text{embed}(H, G) \leq c\ell^{\frac{|\mathcal{F}|}{m}}$.

If we let $m = 1$ then we'd like to find the smallest family of edges $\mathcal{F}$ such that every vertex is in at least one of these edges. Such a family $\mathcal{F}$ is a *vertex cover* of H and the size of a smallest such family is the *vertex cover number* of H, commonly denoted by $\rho(H)$. However, it will turn out that by varying m we can often get a better bound. To this end let us define a fractional version of the above. It will be convenient to move from the discrete setting to a continuous version, which we do as follows.

Note that we could equivalently have defined a vertex cover to be some function $\varphi : E(H) \to \{0, 1\}$ such that for every $v \in V(H)$

$$\sum_{e \in E(G) \colon v \in e} \varphi(e) \geq 1,$$

and the vertex cover number to be the minimum of $\sum_{e\in E(H)}\varphi(e)$ over all vertex covers. The natural generalisation of this is then a *fractional cover* which is a function $\varphi : E(H) \to [0,1]$ such that for every $v \in V(H)$

$$\sum_{e\in E(G)\colon v\in e} \varphi(e) \geq 1,$$

and the *fractional cover number*, which we denote by $\rho^*(H)$, is the minimum of $\sum_{e\in E(H)}\varphi(e)$ over all fractional covers. Note that, since every vertex cover is a fractional cover, it follows that $\rho^*(H) \leq \rho(H)$. Clearly now the hope is that we can show some bound of the form

$$\mathrm{embed}(H,\ell) \leq c\ell^{\rho^*(H)}.$$

In fact, not only does this bound hold, but a corresponding lower bound of the form $c'c\ell^{\rho^*(H)}$ can also be given. The following was originally a theorem of Alon, but we give a proof due to Friedgut and Kahn, who actually proved a slightly more general result about embedding hypergraphs.

Theorem 4.35. *For every graph H there are constants $c_1, c_2 < 0$ such that for every ℓ*

$$c_1\ell^{\rho^*(H)} \leq \mathit{embed}(H,\ell) \leq c_2\ell^{\rho^*(H)}.$$

Proof. Let us first show that the upper bound holds. Let G be a graph with $e(G) = \ell$ and $V(H) = \{v_1,\dots,v_n\}$. Let X be an embedding of H into G chosen uniformly at random from all embeddings. As always we have that $\mathbb{H}(X) = \log(\mathrm{embed}(H,G))$ and so we wish to bound $\mathbb{H}(X)$ as above. Let X_i be the image of v_i under the embedding X, again as before $\mathbb{H}(X) = \mathbb{H}(X_i\colon i\in[n])$.

Given $\varepsilon > 0$, let φ be a fractional vertex cover taking rational values such that $\sum_{e\in E(H)}\varphi(e)$ is at most $\rho^*(H)+\varepsilon$. Then there is some integer $C \in \mathbb{N}$ such that $C\varphi(e) \in \mathbb{N}$ for all $e \in E(H)$. Let us take $\mathcal{F}$ to be a family of subsets of $[n]$ consisting of $C\varphi(e)$ many copies of each pair (i,j) such that $(v_i,v_j) \in E(H)$. Each $i \in [n]$ appears in at least

$$\sum_{e\in E(H)\colon v_i\in e} C\varphi(e) \geq C$$

many members of $\mathcal{F}$, since $\varphi(e)$ was a fractional cover of H. Hence by Shearer's Lemma

$$\mathbb{H}(X) \leq \frac{1}{C}\sum_{(i,j)\in\mathcal{F})} \mathbb{H}(X_i,X_j) = \frac{1}{C}\sum_{e=(v_i,v_j)\in E(H)} C\varphi(e)\mathbb{H}(X_i,X_j).$$

However, since $e(G) = \ell$, the range of (X_i,X_j) is at most 2ℓ and hence

$$\mathbb{H}(X) \leq \sum_{e=(v_i,v_j)\in E(H)} \varphi(e)\log(2\ell) \leq (\rho^*(H)+\varepsilon)\log(2\ell).$$

Hence, for all $\varepsilon > 0$ it follows that

$$\mathrm{embed}(H,G) \leq 2^{\rho^*(H)+\varepsilon}\ell^{\rho^*(H)+\varepsilon}$$

and so, by letting $\varepsilon \to 0$, the upper bound holds with $c_2 = 2^{\rho^*(H)}$.

In order to show the lower bound we will have to make use of the following consequence of the duality of linear programming. We can think of an independent set in a graph as a set of

vertices W such that each edge touches at most one vertex in W. Equivalent, we can think of the characteristic function of W, $\psi : V(H) \to \{0,1\}$, where we require that for each edge $e \in E(H)$

$$\sum_{v \in V(H):\, v \in e} \psi(v) \leq 1.$$

Then the independence number $\alpha(H)$ is the minimum of $\sum_{v \in V(H)} \psi(v)$ over all such functions.

In the same way we can define a *fractional independent set* to be a function $\psi : V(H) \to [0,1]$ such that for every edge $e \in E(H)$

$$\sum_{v \in V(H):\, v \in e} \psi(v) \leq 1,$$

and the *fractional independence number* of H, which we denote by $\alpha^*(H)$, to be the minimum of $\sum_{v \in V(H)} \psi(v)$ over all fractional independent sets.

It is easy to verify that for the integer versions $\alpha(H) \leq \rho(H)$, since every vertex cover needs at least one edge per vertex in an independent set, but in many cases this inequality is in fact strict, that is $\alpha(H) < \rho(H)$. However, the fundamental theorem of linear programming duality implies that the fractional versions are in fact equal, that is, $\alpha^*(H) = \rho^*(H)$. Using this we can construct a graph witnessing the lower bound in Theorem 4.35.

By the above, it will be sufficient to exhibit, for each ℓ, a graph with $\leq \ell$ edges such that $\text{embed}(H, G) \geq c_1 \ell^{\alpha^*(H)}$. Let us take an optimal fractional independent set ψ such that $\sum_{v \in V(H)} \psi(v) = \alpha^*(H)$, and let us assume in what follows, for ease of presentation that the quantity

$$\left(\frac{\ell}{e(H)}\right)^{\psi(v)}$$

is integral for each $v \in V$. We define a graph G as follows:

For each $v \in V(H)$ let $V(v)$ be a set of size $\left(\frac{\ell}{e(H)}\right)^{\psi(v)}$ and let $V(G) = \bigcup_{v \in V(H)} V(v)$. As the edge set of G we take all edges between $V(v)$ and $V(u)$ for every $(u,v) \in E(H)$. hence the number of edges in G is

$$\sum_{(u,v) \in E(H)} \left(\frac{\ell}{e(H)}\right)^{\psi(u)+\psi(v)}$$

However, since ψ is a fractional independent set and $(u,v) \in E(H)$, $\psi(u) + \psi(v) \leq 1$ and hence

$$E(G) \leq \sum_{(u,v) \in E(H)} \frac{\ell}{e(H)} = \ell.$$

Now, any function $f : V(H) \to V(G)$ such that $f(v) \in V(v)$ for all $v \in V(H)$ is an embedding of H into G and so

$$\text{embed}(H, G) \geq \prod_{v \in V(H)} |V(v)| = \left(\frac{\ell}{e(H)}\right)^{\sum_{v \in V(H)} \psi(v)} = \left(\frac{\ell}{e(H)}\right)^{\alpha^*(H)}.$$

Hence the lower bound follows with $c_1 = \left(\frac{1}{e(H)}\right)^{\rho^*(H)}$. □

4.5 Independent Sets in a Regular Bipartite Graph

Let G be a d-regular bipartite graph on $2n$ vertices with vertex classes A and B, and let $\mathcal{I}(G)$ be the class of independent subsets of $V(G)$. We would like to bound this number from above. As in the case of Breégman's Theorem, letting G be a disjoint union of $K_{d,d}$'s seems a natural guess for a best possible graph. Indeed in G it is clear that any independent set in G consists of an arbitrary subset taken from one side of each $K_{d,d}$. Therefore we have that

$$|\mathcal{I}(G)| = (2^{d+1} - 1)^{\frac{n}{d}}.$$

The following proof of a corresponding upper bound on $|\mathcal{I}(G)|$ using entropy methods is due to Kahn.

Theorem 4.36. *Let G be a d-regular bipartite graph on $2n$ vertices with vertex classes A and B, and let $\mathcal{I}(G)$ be the set of independent subsets of $V(G)$. Then*

$$|\mathcal{I}(G)| \leq (2^{d+1} - 1)^{\frac{n}{d}}$$

Proof. The basic idea of the proof is the same as in Theorem 4.2, we pick a random independent set I and estimate the entropy $H(I)$. As before we have that $H(I) = \log(|\mathcal{I}|)$.

We identify I with it's characteristic vector $(X_v : v \in A \cup B)$, note that I is determined by (X_A, X_B). The idea is that, rather than splitting X into X_v for each v, we can use the neighbourhoods of each $v \in A$ as a d-uniform cover of the vertices of B, and so use Shearer's Lemma to express X_B in terms of $X_{N(v)}$.

For each $v \in A$ let $N(v)$ be the neighbourhood of v in B. Each $w \in B$ is in exactly d of the sets $N(v)$ and so we have

$$\begin{aligned} H(I) &= H(X_A|X_B) + H(X_B) \\ &\leq \sum_{v \in A} H(X_v|X_B) + \frac{1}{d}\sum_{v \in A} H(X_{N(v)}) \\ &\leq \sum_{v \in A} \big(H(X_v|X_{N(v)}) + \frac{1}{d}H(X_{N(v)}\big), \end{aligned}$$

where the second line follows from Shearer's inequality, and the third since $N(v) \subset B$.

Fix some $v \in A$. Let χ_v be the indicator random variable of the event that $I \cap N(v) \neq \emptyset$, and let $p := \mathbb{P}(\chi_v = 0)$, that is the probability that $I \cap N(v) = \emptyset$. The nice thing about this random variable is that it contains all the information about $X_{N(v)}$ that we need to determine $H(X_v|X_{N(v)})$.

Hence ,

$$\begin{aligned} H(X_v|X_{N(v)}) &\leq H(X_v|\chi_v) \\ &= \mathbb{P}(\chi_v = 0)H(X_v|\chi_v = 0) + \mathbb{P}(\chi_v = 1)H(X_v|\chi_v = 1) \\ &= \mathbb{P}(\chi_v = 0)H(X_v|\chi_v = 0) \leq p, \end{aligned}$$

since the event $\chi_v = 1$ determines that $X_v = 0$, and since $H(X_v) \leq \log(|\text{range}(X_v)|) = 1$.

Also,

$$\begin{aligned} H(X_{N(v)}) &= H(X_{N(v)}, \chi_v) \\ &= H(\chi_v) + H(X_{N(v)}|\chi_v) \\ &\leq H(p) + (1-p)\log(2^d - 1), \end{aligned}$$

where $H(p) = p\log(1/p) + (1-p)\log(1/(1-p))$. Putting these inequalities together gives us

$$H(I) \leq \sum_{v \in A} (p + \frac{1}{d}\left(H(p) + (1-p)\log\left(2^d - 1\right)\right).$$

All that remains is to maximise the quantity on the right hand side according to p. It is a simple exercise to check that the function is convex, and to calculate its derivative, giving that the maximum is attained at $p = 2^d/(2^{d+1} - 1)$, and so $(1-p) = 2^d - 1/(2^{d+1} - 1)$ giving that:

$$\begin{aligned} H(I) &\leq \sum_{v \in A} (p + \frac{1}{d}\left(H(p) + (1-p)\log\left(2^d - 1\right)\right) \\ &= n\left(p + \frac{1}{d}\left(p\log(1/p) + (1-p)\log(1/(1-p)) + (1-p)\log\left(2^d - 1\right)\right)\right) \\ &= n\left(p + \frac{1}{d}\left(p\log\left(\frac{2^{d+1}-1}{2^d}\right) + (1-p)\log\left(\frac{2^{d+1}-1}{2^d - 1}\right) + (1-p)\log\left(2^d - 1\right)\right)\right) \\ &= n\left(p + \frac{1}{d}\left(p\log\left(2^{d+1} - 1\right) - pd + (1-p)\log\left(2^{d+1} - 1\right)\right)\right) \\ &= n\left(p - p + \frac{1}{d}\left((p + (1-p))\log\left(2^{d+1} - 1\right)\right)\right) \\ &= n\frac{1}{d}\log\left(2^{d+1} - 1\right) \end{aligned}$$

$$\log(|\mathcal{I}|) = H(I) \leq n.\frac{1}{d}\log\left(2^{d+1} - 1\right),$$

from which the result follows. □

5 Entropy Inequalities

Given a collection of discrete random variable $(X_i \colon i \in [n])$ there are $2^n - 1$ different joint distributions X_I whose entropy we can consider. If we write $h(I) := H(X_I)$, then many of the basic results about entropy can be expressed as linear inequalities between the $h(I)$.

For example, by Lemmas 2.3 and 2.4

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

and so $h(i,j) - h(i) - h(j) \leq 0$ for all $i, j \in [n]$. Alternatively, this is a consequence of the inequality $I(X;Y) \geq 0$.

Let $k = 2^n - 1$ and let us label the coordinates of $\mathbb{R}^k$ by the non-empty subsets of $[n]$. We say a vector $x \in \mathbb{R}^k$ is *entropic* if there exists some collection of discrete random variables $(X_i \colon i \in [n])$ such that $x_I = h(I)$ for every non-empty subset $I \subseteq [n]$. Note that, since $h(I) \geq 0$ for all $I \subseteq [n]$ not every vector is entropic. Let us define the region

$$\Gamma_n^* = \{x \in \mathbb{R}^k \colon x \text{ is entropic}\}.$$

Now, obviously Γ_n^* is restricted by all the inequalities we get by taking our known entropy inequalities and expressing them in terms of the $h(I)$. Let us call an inequality a *Shannon inequality* if it can be derived from an inequality of the form

$$I(X_U; X_V | X_W) \geq 0$$

where $U, V, W \subseteq [n]$. Let us first show that this makes sense, that is, every inequality of the above form is equivalent to some linear inequality in the $h(I)$.

Lemma 5.1. *Let $U, V, W \subseteq [n]$. Then there exist $(\lambda_I \in \mathbb{R} \colon I \subseteq [n])$ such that the inequality $I(X_U; X_V | X_W) \geq 0$ is equivalent to the inequality*

$$\sum_{I \subseteq [n]} \lambda_I h(I) \geq 0$$

Proof.

$$\begin{aligned} I(X_U; X_V | X_W) &= H(X_U|X_W) + H(X_V|X_W) - H(X_U, X_V|X_W) \\ &= H(X_U, X_W) - H(X_W) + H(X_V, X_W) - H(X_W) - H(X_U, X_V, X_W) + H(X_W) \\ &= H(X_U, X_W) + H(X_V, X_W) - H(X_W) - H(X_U, X_V, X_W) \\ &= h(U \cup W) + h(V \cup W) - h(W) - h(U \cup V \cup W). \end{aligned}$$

□

We know that every entropic x satisfies every Shannon inequality. A natural question to ask is, does this determine whether a vector x is entropic? To that end let us define Γ_n to be the set of points in $\mathbb{R}^k$ which satisfy every Shannon-inequality. Clearly $\Gamma_n^* \subseteq \Gamma_n$.

It can be shown that $\Gamma_2^* = \Gamma_2$ and, whilst Zhang and Yeung showed in 1997 that $\Gamma_3^* \neq \Gamma_3$, this is only false 'on the boundary', in that the the closure $\overline{\Gamma}_3^* = \Gamma_3$.

Note that, Γ_n is determined by a collection of linear inequalities and since $h(\emptyset) = 0$ all of these inequalities are homogeneous. Hence Γ_n forms a convex cone. Zhang and Yeung showed that this is also true of $\overline{\Gamma}^*_n$.

Theorem 5.2. *$\overline{\Gamma}^*_n$ is a convex cone.*

Proof. If we let X_i be random variables taking the constant value 1, then $h = 0$, and so Γ^*_n contains the origin. Furthermore, suppose we have two vectors $x, x' \in \Gamma_n$. By assumption there are random variables $(X_i \colon i \in [n])$ and $(X'_i \colon i \in [n])$ witnessing that x and x' are in Γ^*_n. We may assume that $(X_i \colon i \in [n])$ and $(X'_i \colon i \in [n])$ are independent, and define random variables $Y_i = (X_i, X'_i)$ for each $i \in [n]$.

Then, for any subset $I \subseteq [n]$ we have

$$H(Y_I) = H(X_I, X'_I) = H(X_I) + H(X'_I) = h(I) + h'(I) = x_I + x'_I.$$

Hence, the vector $x + x' \in \Gamma^*_n$, as witnessed by the family of random variables $(Y_i \colon i \in [n])$. It follows that Γ^*_n is closed under taking integer multiples.

To show that the closure $\overline{\Gamma}^*_n$ is a convex cone it thus suffices to show that for every $x, x' \in \Gamma_n$ and every $\lambda \in (0, 1)$ the convex combination $\lambda x + (1 - \lambda)x' \in \overline{\Gamma}^*_n$.

To do so, let us suppose we have families of random variables $(X_i \colon i \in [n])$ and $(X'_i \colon i \in [n])$ witnessing that $x, x' \in \Gamma^*_n$ with $(X_i \colon i \in [n])$ and $(X'_i \colon i \in [n])$ independent. Let us take $(Y_i \colon i \in [n])$ and $(Z_i \colon i \in [n])$ where each Y_i is the joint distribution of k independent copies o f X_i and similarly Z_i is the joint distribution of k independent copies of X'_i, and let U be a discrete random variable independent of the rest such that

$$\begin{aligned} \mathbb{P}(U = -1) &= \varepsilon; \\ \mathbb{P}(U = 0) &= 1 - \delta - \varepsilon; \\ \mathbb{P}(U = 1) &= \delta. \end{aligned}$$

Finally, let us consider the random variables $(\hat{X}_i \colon i \in [n])$ given by

$$\hat{X}_i = \begin{cases} Z_i \text{ if U = -1} \\ 0 \text{ if U = 0} \\ Y_i \text{ if U = 1} \end{cases} \tag{5.1}$$

Now, for any non-empty subset $I \subseteq [n]$

$$\begin{aligned} H(\hat{X}_I) &\leq H(\hat{X}_I, U) \\ &= H(U) + H(X_I|U) \\ &= H(U) + \sum_{\mu \in \{-1,0,1\}} \mathbb{P}(U = \mu) H(\hat{X}_I | U = \mu) \\ &= H(U) + \delta k H(X_I) + \varepsilon k H(X'_I). \end{aligned}$$

On the other hand

$$H(\hat{X}_I) \geq H(\hat{X}_I|U) = \delta k H(X_I) + \varepsilon H(X'_I).$$

and so

$$0 \leq H(\hat{X}_I) - \big(\delta k H(X_I) + \varepsilon H(X'_I)\big) \leq H(U).$$

Letting $\delta = \frac{\lambda}{k}$ and $\varepsilon = \frac{(1-\lambda)}{k}$ we see that

$$0 \leq H(\hat{X}_I) - \big(\lambda H(X_I) + (1-\lambda) H(X'_I)\big) \leq H(U).$$

However, as $\varepsilon, \delta \to 0$, $H(U) \to 0$, and hence by taking k arbitrarily large we can find $\hat{X}$ such that $\hat{h}_I$ is arbitrarily close to $\lambda x_I + (1-\lambda)x'_I$. Hence $\lambda x + (1-\lambda)x' \in \overline{\Gamma}^*_n$. □

However, Zhang and Yeung showed a year later that in fact $\overline{\Gamma}^*_n \neq \Gamma_n$ for all $n \geq 4$, by exhibiting non-Shannon-inequalities that are satisfied by any set of discrete random variables. Below we give such an inequality for five random variables.

Theorem 5.3. *For any discrete random variables A, B, C, D, E*

$$2I(A;B) \leq I(A,E;C|B) + I(A;B|E) + I(A;B) - I(A;C) + I(A;E|B) + I(B;E|A) + I(A;B|D) + I(C;D).$$

Taking $C = E$ we get the following inequality for 4 random variables

Corollary 5.4. *For any discrete random variables A, B, C, D*

$$2I(A;B) \leq 3I(A;B|C) + I(A;B|D) + I(A,B;C) + I(C;D).$$

Let us first show that this is not a Shannon-inequality. A little trick to make this easier, and an interesting thing to note anyway, is that given a collection of random variables $(X_i \colon i \in [n])$ the function $h : 2^{[n]} \to \mathbb{R}$ we defined above is a *polymatroid*.

That is

1. $h(\emptyset) = 0$;
2. $h(I) \leq h(J)$ if $I \subseteq J$;
3. $h(I) + h(J) \geq h(I \cap J) + h(I \cup J)$.

The first two are apparent and for the third let us write $A = I \cap J$, $B = I \setminus J$ and $C = J \setminus I$. Then we need to show

$$H(X_A, X_B) + H(X_B, X_C) \geq H(X_A) + H(X_A, X_B, X_C).$$

However, by Lemma 2.4

$$\begin{aligned} H(X_A) + H(X_A, X_B, X_C) &= 2H(X_A) + H(X_B, X_C|X_A) \\ &\leq 2H(X_A) + H(X_B|X_A) + H(X_C|X_A) \\ &= H(X_A, X_B) + H(X_A, X_C) \end{aligned}$$

as claimed.

In fact, the converse is also true, in that every polymatroid satisfies the Shannon-inequalities. Indeed, if h is a polymatroid function and $U, V, W \subseteq [n]$ then by 2. and 3. we have hat

$$h(U \cup W) + h(V \cup W) \geq h(U \cup V \cup W) + h((U \cup W) \cap (V \cup W)) \geq h(U \cup V \cup W) + h(W)$$

and hence (See Lemma 5.1)

$$h(U \cup W) + h(V \cup W) - h(W) - h(U \cup V \cup W) \geq 0.$$

So, to show that Corollary 5, and hence Theorem 5.3, is a non-Shannon-inequality it will be sufficient to give a polymatroidal h which doesn't satisfy it, or more precisely, if we let $X_1 = A, X_2 = B, \ldots, X_5 = E$, then Theorem 5.3 can be rewritten as a inequality

$$\sum_{I \subset [n]} \lambda_i \mathbb{H}(X_I) \leq 0$$

and it will be sufficient to give a polymatroid $h : 2^{[n]} \to \mathbb{R}$ such that

$$\sum_{I \subset [n]} \lambda_i h(I) > 0.$$

Let h be as follows:

- $h(i) = 2$ for $i \in [4]$;
- $h(i, j) = 3$ for $\{i, j\} \neq \{3, 4\}$;
- $h(I) = 4$ otherwise.

It is a simple check that h is a polymatroid, and that h does not satisfy the inequality from Corollary 5. Indeed, $I(A; B) = H(A) + H(B) - H(A, B) = 2 + 2 - 3 = 1$. However, the two terms of the form $I(X; Y|Z)$ can be seen to be zero, since all pairs but C, D have $H(X, Y) = 3$ and so $I(X; Y|Z) = H(X, Y) + H(Y, Z) - H(X, Y, Z) - H(Z) = 3 + 3 - 4 - 2 = 0$ for those two terms. Also, $I(A, B; C) = H(A, B) + H(C) - H(A, B, C) = 1$ and finally $I(C; D) = H(C) + H(D) - H(C, D) = 2 + 2 - 4 = 0$. Putting this all together we see that

$$2 = 2I(A; B) > 3I(A; B|C) + I(A; B|D) + I(A, B; C) + I(C; D) = 1.$$

Proof of Theorem 5.3. Let us first note that E only appears in terms of the inequality that we wish to prove together with A and B. This allows us to 'redefine' E so that it is independent of C and D, conditioned on A and B. More precisely we can define a random variable E' such that the following holds

- $I(A, B; E) = I(A, B; E')$ and $I(A; B|E) = I(A; B|E')$;
- $I(C, D; E|A, B) = 0$.

Formally, we can do this by defining the conditional distribution of E' given A, B, C, D to be the conditional distribution of E given A and B. That is, for every a, b, c, d, e

$$\mathbb{P}(E' = e|A = a, B = b, C = c, D = d) = \mathbb{P}(E = e|A = a, B = b).$$

It is then a simple check that the two conditions above hold for E', and hence to prove the inequality for arbitrary A, B, C, D, E it will be sufficient to prove it under the additional assumption that $I(C, D; E|A, B) = 0$. Note that this is reason we prove Corollary 5 via Theorem 5.3.

The theorem now reduces to verifying a series of inequalities, although hopefully we can split it into a number of steps that will make sense. Firstly we note that, since $I(C, D; E|A, B) = 0$ we have that $I(C, D; E) \leq I(A, B; E)$. Indeed firstly we note that for any X, Y, Z the following holds:

$$\begin{aligned} I(X, Y; Z) &= \mathbb{H}(X, Y) + \mathbb{H}(Z) - \mathbb{H}(X, Y, Z) \\ &= \mathbb{H}(X|Y) + \mathbb{H}(Y) + \mathbb{H}(Z) - \mathbb{H}(X, Z|Y) - \mathbb{H}(Y) \\ &= I(X; Z|Y) + \mathbb{H}(Z) - \mathbb{H}(Z|Y) \\ &= I(X; Z|Y) + I(Y; Z). \end{aligned}$$

Hence we can calculate $I(A, B, C, D; E)$ in two ways. Firstly

$$I(A, B, C, D; E) = I(C, D; E|A, B) + I(A, B; E) = I(A, B; E),$$

and hence

$$I(A, B; E) = I(A, B, C, D; E) = I(A, B; E|C, D) + I(C, D; E) \geq I(C, D; E). \tag{5.2}$$

We will use this to verify the next inequality

$$I(C; E) + I(D; E) \leq I(A, B; E) + I(C; D). \tag{5.3}$$

By (5.2) it will be sufficient to show that

$$I(C; E) + I(D; E) \leq I(C, D; E) + I(C, D)$$

However

$$\begin{aligned} I(C; E) + I(D; E) &= \mathbb{H}(C) + \mathbb{H}(E) - \mathbb{H}(C, E) + \mathbb{H}(D) + \mathbb{H}(E) - \mathbb{H}(D, E) \\ &= \mathbb{H}(C, D) + I(C; D) + 2\mathbb{H}(E) - \mathbb{H}(C, E) - \mathbb{H}(D, E) \\ &= I(C, D; E) + I(C, D) + \mathbb{H}(E) + \mathbb{H}(C, D, E) - \mathbb{H}(C, E) - \mathbb{H}(D, E) \\ &= I(C, D; E) + I(C, D) + \mathbb{H}(D|C, E) - \mathbb{H}(D|E) \\ &\leq I(C, D; E) + I(C, D). \end{aligned}$$

Finally we claim the follow inequality holds

$$I(A; B) \leq I(A; B|C) + I(A; B|E) + I(C; E), \tag{5.4}$$

and so by symmetry also

$$I(A; B) \leq I(A; B|D) + I(A; B|E) + I(D; E), \tag{5.5}$$

Note that, by adding (5.4) and (5.5) we see that

$$2I(A;B) \leq I(A;B|C) + I(A;B|D) + 2I(A;B|E) + I(C;E) + I(D;E),$$

and by (5.3) we could conclude that

$$2I(A;B) \leq I(A;B|C) + I(A;B|D) + 2I(A;B|E) + I(A,B;E) + I(C;D).$$

as claimed.

So, it remains to prove (5.4). Note that, $I(C,D;E|A,B) = 0$ implies that $I(C;E|A,B) = 0$. Re-writing this in terms of entropy we get the following equality

$$\mathbb{H}(A,B,C) + \mathbb{H}(A,B,E) = \mathbb{H}(A,B) + \mathbb{H}(A,B,C,E)$$

So now we can calculate

$$\begin{aligned}
&I(A;B|C) + I(A;B|E) + I(C;E) \\
&= \mathbb{H}(A,C) + \mathbb{H}(B,C) - \mathbb{H}(C) - \mathbb{H}(A,B,C) + \mathbb{H}(A,E) + \mathbb{H}(B,E) - \mathbb{H}(E) - \mathbb{H}(A,B,E) + \mathbb{H}(C) + \mathbb{H}(E) - \mathbb{H}(E,C) \\
&= \mathbb{H}(A,C) + \mathbb{H}(B,C) + \mathbb{H}(A,E) + \mathbb{H}(B,E) - \mathbb{H}(E,C) - \mathbb{H}(A,B,C,E) - \mathbb{H}(A,B) \\
&= I(A;B|E,C) - \mathbb{H}(A,E,C) - \mathbb{H}(B,E,C) + \mathbb{H}(A,C) + \mathbb{H}(B,C) + \mathbb{H}(A,E) + \mathbb{H}(B,E) - \mathbb{H}(A,B) \\
&\geq \mathbb{H}(A,C) + \mathbb{H}(B,C) + \mathbb{H}(A,E) + \mathbb{H}(B,E) - \mathbb{H}(A,B) - \mathbb{H}(A,E,C) - \mathbb{H}(B,E,C) \\
&= \mathbb{H}(A,C) + \mathbb{H}(A,E) - \mathbb{H}(A) - \mathbb{H}(A,C,E) + \mathbb{H}(B,C) + \mathbb{H}(B,E) - \mathbb{H}(B) - \mathbb{H}(B,E,C) + \mathbb{H}(A) + \mathbb{H}(B) - \mathbb{H}(A,B) \\
&= I(C;E|A) + I(C;E|B) + I(A;B) \geq I(A;B).
\end{aligned}$$

□