Information theory[1]

A course instructed by Amir Yehudayoff, Department of Mathematics, Technion-IIT

[1] An apology: this text probably contains errors.

Contents

1 Goals 5

2 Entropy 7

3 Applications 19

4 Shearer's lemma 29

5 Mutual information 33

6 Harmonic functions 41

7 The second law of thermodynamics 47

8 Mutual information and sampling 51

9 Graph entropy 55

Chapter 1

Goals

The course is an introduction for information theory and will focus on its applications in mathematics.

Information theory started with the works of Shannon. Its high-level goal is to quantify the amount of information that is conveyed in communication. But it has evolved and found applications in many other areas:

- Communication: capacities of channels.
- Physics: thermodynamics.
- Computer science: communication complexity, data structures, etc.
- Economics.
- Mathematics: group theory, random walks, dynamical systems, geometry, etc.

They basic definitions follow from simple and natural “axiomatic” requirements. This yields tools for analyzing complex systems.

Chapter 2

Entropy

The first notion we discuss is *entropy*. It captures a fundamental and important idea. But it takes some time to digest it.

2.1 A bit of history

It started in thermodynamics (physics). Joule (1840) studied systems focussing on pressure, temperature, etc. Several "laws" were discovered, including the second law of thermodynamics: the entropy in the system always increases. (More on this later on.) Later Bolzmann (1880) and Gibbs gave statistical interpretation to this elusive notion. More recently Shannon (1950) studied communication, identified natural axioms for "entropy" and gave a formal definition following the axioms (read his paper!).

2.2 The definition

We shall see many applications later on! But we start with the most basic definition, and go relatively slow here.

What is a "system"? Think a cup of water, the sky, the stock market, etc.

What is the math model for it? The standard choice for a system is a **probability space**. For most of the course, we shall deal with finite probability spaces. There is a distribution p on $[n] = \{1, 2, \dots, n\}$. This is our "system". The system takes finitely many states (n), and state i is "seen" with probability $p(i) \in [0, 1]$. As all models, this is not a perfect model. **What are its pros and cons?**

Goal. The goal is to assign a number to each system; this number should capture *"the amount of disorder in the system"*. The larger it is, the less order there is. We want to define

$$H : \text{all distribution} \to \mathbb{R}.$$

The term that was chosen for this H is **entropy**; this name was suggested to Shannon by John von Neumann.

What are the properties of "disorder"? What should be the properties of the entropy H?

1. Does not depend on ordering: for every distribution p on $[n]$ and permutation π of $[n]$,

   $$H(p) = H(p \circ \pi).$$

2. Continuous: $H(p)$ is continuous function in p (what is the natural topology?). Make this assumption for each n separately. That is, for every n and $\epsilon > 0$, there is $\delta > 0$ so that if p, q are distribution on $[n]$ so that $\|p - q\|_1 \leq \delta$ then $|H(p) - H(q)| < \epsilon$. (All norms are equivalent here.)

3. Monotonicity: If U_n is the uniform distribution on n, **what should be the order on $H(U_1), H(U_2), \ldots$?** The larger n is, the larger the entropy.

4. Splitting: for $p = (p_1, \ldots, p_{n-1}, p_n)$ and $p_n = q_1 + q_2$ with $q_1, q_2 \geq 0$, if $p' = (p_1, \ldots, p_{n-1}, q_1, q_2)$ then **what should be $H(p')$?**

   The choice that was made is

   $$H(p') = H(p) + p_n H(\tfrac{q_1}{p_n}, \tfrac{q_2}{p_n}).$$

   This corresponds to viewing the larger system p' in two steps. First we view p. If the value is less than n, then this is also the state of p'. If the value is n, then we view the system $(\frac{q_1}{p_n}, \frac{q_2}{p_n})$ to determine the state of p'. The chance of seeing n in p is p_n.

5. Normalization:

   $$H(U_2) = 1.$$

   This is "suitable for bits (binary digits)".

Theorem 1. *If H satisfies the above axioms then*

$$H(p) = \sum_i p_i \log(\tfrac{1}{p_i})$$

where $\log = \log_2$ and $0 \log(\frac{1}{0}) = 0$ (agrees with $x \log x \to 0$ when $x \to 0$).

Claim 2 (warm up)**.** $H(1, 0) = 0$.

Proof. By splitting rule,

$$H(0.5, 0.5, 0) = H(0.5, 0.5) + 0.5H(1, 0).$$

By symmetry,

$$= H(0, 0.5, 0.5) = H(0, 1) + H(0.5, 0.5).$$ □

Corollary 3. *$H(p_1, \dots, p_n, 0) = H(p_1, \dots, p_n)$.*

Proof. Exercise (now). □

Claim 4. *For $m \geq 1$,*

$$H(p_1, \dots, p_{n-1}, q_1, \dots, q_m) = H(p) + p_n H(\tfrac{q_1}{p_n}, \dots, \tfrac{q_m}{p_n}),$$

where $q_j > 0$ and $\sum_j q_j = p_n$.

Proof. By induction (left as an exercise). □

Claim 5 (products)**.** *$H(p \times q) = H(p) + H(q)$.*

Remark. *This is a hint toward the general formula. Which functions $f : \mathbb{R}_{>0} \to \mathbb{R}$ satisfy $f(xy) = f(x) + f(y)$? There are many such f's; all $\log_b$ for $b > 0$. If we add normalization $f(2) = 1$, does it help? No, we know nothing of $f(3)$ for example. If we also add monotonicity, then there is a unique function $\log_2(x)$.*

Here is a proof. First,

$$f(2^k) == k.$$

What can we say about $f(3)$? We can write

$$f(3^k) = kf(3).$$

Let k be large, and let ℓ be the smallest integer so that $2^\ell \leq m^k$;

$$\ell = \lfloor k \log_2(m) \rfloor.$$

By monotonicity,

$$f(2^\ell) \leq f(3^k) \leq f(2^{\ell+1})$$

so

$$\frac{\ell}{k} \leq f(3) \leq \frac{\ell+1}{k}.$$

And for k tending to ∞, using sandwich,

$$f(3) = \log_2(3).$$

A similar argument hold for all x.

Proof.

$$
\begin{aligned}
H(p \times q) &= H(p_1q_1, p_1q_2, \ldots, p_{n-1}q_m, \underline{p_nq_1, \ldots, p_nq_m}) \\
&= H(p_1q_1, p_1q_2, \ldots, p_{n-1}q_m, p_n) + p_nH(q) \\
&= p_nH(q) + H(p_n, p_1q_1, p_1q_2, \ldots, p_{n-2}q_m, \underline{p_{n-1}q_1, \ldots, p_{n-1}q_m}) \\
&= p_nH(q) + H(p_n, p_1q_1, p_1q_2, \ldots, p_{n-2}q_m, p_{n-1}) + p_{n-1}H(q) \\
&= \ldots \\
&= (p_1 + p_2 + \ldots + p_n)H(q) + H(p).
\end{aligned}
$$ □

Corollary 6. $H(U_{nm}) = H(U_n) + H(U_m)$.

Corollary 7. $H(U_n) = \log_2(n)$.

Proof. Similar to the discussion concerning $f(n) = H(U_n)$ above. □

Proof of main formula. By continuity, assume that p_i is rational. Write $p_i = \frac{q_i}{s}$ for q_i, s positive integers. By splitting and induction (similarly to product claim):

$$
\begin{aligned}
\log_2(s) &= H(U_s) \\
&= H(p) + \sum_i p_iH(U_{q_i}) \\
&= H(p) + \sum_i p_i \log_2(q_i).
\end{aligned}
$$ □

Remark. *Axiomatic definitions are extremely powerful. For example, determinant has an axiomatic definition and it is one of the most important functions in math.*

Remark. *Gromov found an axiomatic definition based on the product property*

$$H(p \times q) = H(p) + H(q).$$

But the notion of continuity (the underlying topology) is more complicated and we do not discuss it in detail.

Remark. *We can also write*

$$H(p) = -\mathbb{E}_{X\sim p} \log p(X).$$

It is related to statistical physics in a way. Temperature is "average velocity" and entropy is a different average.

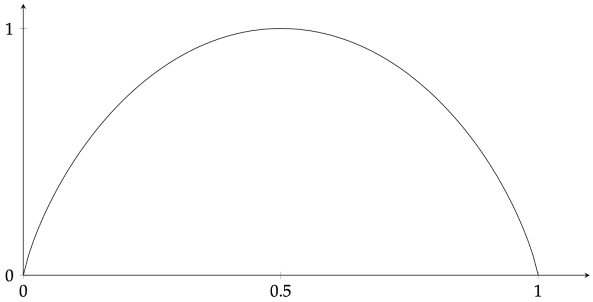
Remark. *There are other definitions of "entropy" that are useful in different contexts. The formula we saw is known as Shannon's entropy.*

2.3 Examples

A random coin. If X takes the value 1 w.p. p and 0 w.p. $1-p$, then

$$H(X) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p} := h(p)$$

is called the binary entropy function:



It is concave with maximum at one half. It is symmetric around $p = 0.5$. The property $h(0) = h(1) = 0$ means that "no randomness means no entropy". The most difficult thing is to guess the output of a uniformly random coin. Around the maximum we have

$$h(\tfrac{1}{2} - \epsilon) \approx 1 - c\epsilon^2.$$

Around the minimum:

$$h(\epsilon) \approx \epsilon \log \tfrac{1}{\epsilon};$$

specifically, the derivative at zero is ∞.

n-bits. If X is uniform on $\{0,1\}^n$ then

$$H(X) = H(U_{2^n}) = n.$$

Intuitively, to describe X we need n bits. This is the first connection between entropy and description length (more on this later on).

Three bits. If $X \in \{0,1\}^3$ is so that $X_1 + X_2 + X_3 = 0 \mod 2$ then

$$H(X) = 2$$

and

$$H(X_1) = 1$$

and

$$H(X_1, X_2) = 2 = H(X).$$

If we known (X_1, X_2) then we know X. This is true also for 13 and for 23.

2.4 Convexity

The basic properties of entropy hold due to convexity. In fact, entropy can be thought of as a clean and intuitive framework for using convexity.

Definition 8. *A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for every $x \in \mathbb{R}^n$ and $p \in [0,1]$,*

$$f(px + (1-p)y) \leq pf(x) + (1-p)f(y).$$

Remark. *In one-dimensional space this means that the function is "smiling". In general, the area above the graph of a convex function is convex.*

Example 9. *$x^2, e^x, -\log(x), x\log(x)$ are convex. Pay attention to the domain of the function.*

Lemma 10 (Jensen's inequality)**.** *If $f : \mathbb{R} \to \mathbb{R}$ is convex and X is a real-valued random variable then*

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

Remark. *A function f is concave if $-f$ is convex. The function $\log(x)$ is concave.*

2.5 Properties

Entropy has simple and intuitive properties. This makes it an extremely powerful tool for counting.

Claim 11 (non-negative)**.** *For every random variables X,*

$$H(X) \geq 0.$$

Equality holds iff X is constant.

Proof. If $p(x) > 0$ then $\log \frac{1}{p(x)} > 0$. □

Claim 12 (uniform has most entropy)**.** *If X is a random variable taking values in $[n]$ then*

$$H(X) \leq \log n.$$

Equality holds iff X is uniform.

Proof.

$$\begin{aligned} H(X) &= \sum_x p(x) \log \tfrac{1}{p(x)} \\ &\leq \log \Big(\sum_x p(x) \tfrac{1}{p(x)} \Big) = \log n. \end{aligned} \qquad \square$$

Equality holds iff $\frac{1}{p(x)} = c$ because log is strictly concave. Because a probability distribution sums to one, it follows that $c = n$.

The entropy of a system that is composed of two parts is at most the sum of the entropies of the parts.

Claim 13 (subadditivity)**.** *If X, Y are two jointly distributed[1] random variables then*

$$H(X, Y) \leq H(X) + H(Y).$$

Equality holds iff X, Y are independent.

Proof.

$$\begin{aligned} H(X,Y) - H(X) - H(Y) &= \sum_{x,y} p(x,y) \big(\log \tfrac{1}{p(x,y)} + \log p(x) + \log p(y) \big) \\ &= \sum_{x,y} p(x,y) \log \tfrac{p(x)p(y)}{p(x,y)} \\ &\leq \log \Big(\sum_{x,y} p(x,y) \tfrac{p(x)p(y)}{p(x,y)} \Big) = 0. \end{aligned} \qquad \square$$

Equality holds iff $\frac{p(x,y)}{p(x)p(y)} = c$. Because a probability distribution sums to one, it follows that $c = 1$.

2.6 Encoding

The source of Shannon's definition was communication. We now see that entropy is deeply related to description length; $H(X)$ is (almost) equal to the amount of bits that are typically needs to describe X.

Definition 14. *A prefix free encoding of $[n]$ is a map E from $[n]$ to the leaves of a rooted binary tree. The size $|E(i)|$ is defined to be the depth of $E(i)$ in the tree.*

Remark. *Every leaf in a rooted binary tree corresponds to a word in $\{0,1\}^* = \bigcup_k \{0,1\}^k$. The size $|E(i)|$ is the length of the word encoding i. A code is prefix free iff for every $i \neq j$*

[1] We shall omit the "jointly distributed" part in the future.

the words $E(i), E(j)$ are not prefixes of each other. The prefix free condition says that we can not confuse two encodings when we read them from left to write.

Theorem 15. *Let X be a random variable taking values in $[n]$.*

1. *There is a prefix free encoding E so that*
$$\mathbb{E}|E(X)| \leq H(X) + 1.$$

2. *If E is a prefix free encoding then*
$$\mathbb{E}|E(X)| \geq H(X).$$

Remark. *n has no real significance in the theorem.*

Proof. Without loss of generality, assume that $p(1) \geq p(2) \geq \ldots \geq p(n) > 0$. Encode $x \in [n]$ with
$$\ell_x = \lceil \log \tfrac{1}{p(x)} \rceil$$
bits as follows. Consider the full infinite binary tree. Choose a vertex v_1 of depth ℓ_1, assign it to be the encoding of 1, and remove all its children from the tree. Then, choose a vertex $v_2 \neq v_1$ of depth ℓ_2, and so forth.

Why can we continue until all of $[n]$ is encoded? When we are about to assign vertex j, we have deleted $\sum_{i<j} 2^{\ell_j - \ell_i}$ vertices from the 2^{ℓ_j} vertices at depth ℓ_j. There are vertices left because
$$\sum_{i<j} 2^{\ell_j - \ell_i} \leq 2^{\ell_j} \sum_{i<j} p(i) < 2^{\ell_j}.$$

We have thus found an encoding so that
$$\mathbb{E}|E(X)| = \sum_x p(x)|E(x)| \leq \sum_x p(x)(\log(\tfrac{1}{p(x)}) + 1) \leq H(X) + 1.$$

It remains to prove the other direction. Let E be a prefix free encoding of $[n]$. Denote by ℓ_x the encoding length of $x \in [n]$.

We claim that
$$\sum_x 2^{-\ell_x} \leq 1.$$
Imagine a random walk on the tree that starts at the root and moves left or right uniformly. It stops when it hits a leaf. The probability that it hits the leaf encoding x is $2^{-\ell_x}$. Because these event are disjoint the claim holds.

Now, write
$$\mathbb{E}\ell_X = H(X) - \sum_x p(x) \log \frac{2^{-\ell_x}}{p(x)}.$$

Because log is concave and increasing,

$$\sum_x p(x) \log \frac{2^{-\ell_x}}{p(x)} \leq \log \Big(\sum_x p(x) \frac{2^{-\ell_x}}{p(x)} \Big) \leq 0. \qquad \square$$

Remark. *This shows that deep connection between "disorder" and description length. They are basically equivalent; a way to formally define "disorder" is as the cost of the best encoding.*

Remark. *This help to understand the formula for entropy. The number $\log \frac{1}{p(x)}$ is the number of bits in the "optimal" encoding of x. It can be thought of as the "amount of surprise when we see x". The entropy is the expected amount of surprise.*

2.7 An application

The application we consider is concentration of measure (Chernoff and related inequalities) and estimation of binomial coefficients. Let X be uniformly distributed in $\{0,1\}^n$. Think of X as a set as well. The size of X is sharply concentrated around $n/2$. It is typically $n/2 \pm O(\sqrt{n})$. This follows by Chebyshev's inequality. The central limit theorem says something stronger for $n \to \infty$. We want to obtain a concrete bound for finite n.

For $\epsilon > 0$, what is

$$\Pr[|X| \leq (\tfrac{1}{2} - \epsilon)n] =?$$

In other words, what is the size of the set

$$S = \{x \in \{0,1\}^n : |x| \leq (\tfrac{1}{2} - \epsilon)n\}?$$

It can be expressed as a sum of binomial coefficients. But this sum is too complicated for applications.

Theorem 16.

$$\Pr[|X| \leq (\tfrac{1}{2} - \epsilon)n] \leq 2^{(h(\frac{1}{2}-\epsilon)-1)n} \leq e^{-2\epsilon^2 n},$$

where $h(\cdot)$ is the binary entropy function.

Remark. *The second inequality holds by a Taylor expansion of $h(\cdot)$ around $\frac{1}{2}$:*

$$h(\tfrac{1}{2} - \epsilon) \leq 1 - 2\epsilon^2 \log e.$$

This is left as an exercise.

Remark. *The probability for being more than $(\frac{1}{2} + \epsilon)n$ is the same.*

Proof. Let X be uniformly distributed in the set S defined above. So,

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

By subadditivity and symmetry,

$$H(X) \leq \sum_{i=1}^{n} H(X_i) = nH(X_1).$$

What is the distribution of X_1? By linearity of expectation,

$$n\Pr[X_1 = 1] = n\mathbb{E}X_1 = \mathbb{E}|X| \leq (\tfrac{1}{2} - \epsilon)n.$$

By monotonicity of the binary entropy function,

$$H(X_1) \leq h(\tfrac{1}{2} - \epsilon).$$

The probability is

$$\frac{|S|}{2^n} \leq 2^{nH(X_1)-n}. \qquad \square$$

Remark. *For $k \leq \frac{n}{2}$ we proved*

$$\binom{n}{k} \leq 2^{nh(\frac{k}{n})}.$$

This is a pretty good estimate; if $\alpha = \frac{k}{n} \in [0,1]$ is fixed, and n is large, this upper bound is essentially sharp and we shall discuss this later on.

Remark. *There are other methods for proving concentration of measure. The standard proof of Chernoff inequality looks at the exponential moment $\mathbb{E}e^{tX}$ and relies on Markov's inequality. The higher the moment we can estimate, the better the end result is (and the exponential moment is "highest"). The parameter t is chosen to optimize the end result.*

2.8 Conditional entropy

When we perform an experiment and have two possible observables X and Y, we can observe X and ask how much entropy does Y still possess.

Definition 17. *The conditional entropy is defined as*

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

Remark. *The conditional entropy is the entropy of the whole system (X,Y) minus the entropy of the observed part X.*

Remark. *We have the following intuitive chain rule:*

$$H(X,Y) = H(X) + H(Y|X).$$

Notation 18. $H(Y|X = x) = \sum_y p(y|x)\log\frac{1}{p(y|x)}$.

Remark. *The expression $H(Y|X)$ is a number while $H(Y|X=x)$ is a random variable.*

Remark. *The conditional entropy is the average over x, of the entropy of Y conditioned on the event $X=x$:*

$$H(Y|X)=\sum_x p(x)\sum_y p(y|x)\log\tfrac{1}{p(y|x)}=\mathbb{E}_x H(Y|X=x).$$

Try to prove.

Remark. *It follows that $H(Y|X)\geq 0$, as the average of non-negative numbers.*

Remark. *$H(Y|X)=H(Y)$ iff X,Y are independent.*

Remark. *One of the most useful properties of entropy is the chain rule:*

$$H(X_1,X_2,\dots,X_n)=\sum_{i=1}^{n} H(X_i|X_{<i}),$$

where $X_{<i}=(X_1,\dots,X_{i-1})$.

Example 19. *Let $X,Y,Z\in\{0,1\}$ be uniform conditioned on their XOR being zero:*

$$\begin{aligned}&H(X,Y,Z)=\log 4=2\\&H(X)=1\\&H(Y|X)=1\\&H(Z|X,Y)=0.\end{aligned}$$

Remark. *A minor generalization: $H(X_1,X_2,\dots,X_n|Y)=\sum_{i=1}^n H(X_i|X_{<i},Y)$.*

Remark. *$H(X|Y)\neq H(Y|X)$ in general.*

Claim 20 (conditioning reduces entropy)**.** *$H(Y|X)\leq H(Y)$.*

Proof. Subadditivity. □

Remark. *While $H(Y|X)\leq H(Y)$, it is not necessarily true that for each x, we have $H(Y|X=x)\leq H(Y)$. E.g., if $X,Y\in\{0,1\}$ are uniform so that $X+Y>0$, then*

$$\begin{aligned}&H(Y)=h(\tfrac{2}{3})<1\\&H(Y|X=1)=h(\tfrac{1}{2})=1\\&H(Y|X=0)=0.\end{aligned}$$

2.9 Entropy and predictions

We performed an experiment and saw Y. But our goal is to estimate an unknown X as accurately as we possibly can. If $X = f(Y)$ then we can compute X exactly. Conditional entropy allows to control the chance of a correct guess.

Theorem 21 (Fano). *Let X, Y be random variables so that X takes values in $[n]$. Let g be a function so that $g(Y)$ is a random variable. Let*

$$p_e = \mathbb{P}[g(Y) \neq X].$$

Then,

$$h(p_e) + p_e \log n \geq H(X|g(Y)) \geq H(X|Y).$$

(We already know the right inequality.)

Remark. *The theorem says that if X has entropy conditioned on Y then we can not hope to always guess it correctly.*

Proof. Let E be the indicator random variable for the event $g(Y) = X$.

$$H(E, X|g(Y)) = H(X|g(Y)) + H(E|X, g(Y)) = H(X|g(Y))$$
$$H(E, X|g(Y)) = H(E|g(Y)) + H(X|E, g(Y)) \leq H(E) + p_e H(X|E = 1, g(Y)). \qquad \square$$

Chapter 3

Applications

3.1 Data processing

Assume that a system produced outcome X, and that this outcome was processed to $f(X)$. What is the relation between $H(X)$ and $H(f(X))$?

Exercise 22 (data processing i)**.** *$H(X) \geq H(f(X))$.*

Proof.

$$\begin{aligned} H(X, f(X)) &= H(X) + H(f(X)|X) = H(X) \\ H(X, f(X)) &= H(f(X)) + H(X|f(X)) \geq H(f(X)). \end{aligned}$$ □

Remark. *This can be proved in a "straightforward" way, but conditioning makes the proof cleaner.*

Exercise 23 (data processing ii)**.** *$H(X|Y) \leq H(X|f(Y))$.*

Proof.

$$\begin{aligned} H(X, f(Y), Y) &= H(Y) + H(X|Y) + H(f(Y)|Y, X) \\ &= H(Y) + H(X|Y) + H(f(Y)|Y) \\ &= H(Y, f(Y)) + H(X|Y) \\ H(X, f(Y), Y) &= H(f(Y)) + H(X, Y|f(Y)) \\ &\leq H(f(Y)) + H(X|f(Y)) + H(Y|f(Y)) \\ &= H(Y, f(Y)) + H(X|f(Y)); \end{aligned}$$

the inequality is sub-additivity. □

3.2 Cauchy-Schwarz

We prove the well-known

$$\sum_i u_i v_i \leq \sum_i |u_i||v_i| \leq \sqrt{\sum_i u_i^2 \cdot \sum_j v_j^2}.$$

This is an important inequality in linear algebra, geometry and has many applications in most areas in math.

Theorem 24. *If $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$ are positive integers then*

$$\sum_i u_i v_i \leq \sqrt{\sum_i u_i^2 \cdot \sum_j v_j^2}.$$

Equality holds iff $v = cu$ for some $c > 0$.

Remark. *The theorem implies the same statement for the rationals, and by continuity for the reals.*

Remark. *This is perhaps to most complicated proof of this inequality. But it's a good exercise.*

Proof. Let $A_1, \ldots, A_n$ be pairwise disjoint subsets of $\mathbb{N}$ where $|A_i| = u_i$. Let $B_1, \ldots, B_n$ be pairwise disjoint subsets of $\mathbb{N}$ where $|B_i| = v_i$. Choose $X = (X_1, X_2)$ and $Y = (Y_1, Y_2)$ uniformly and independently in $\bigcup_i A_i \times B_i$.

$$2 \log \Big(\sum_i u_i v_i \Big) = H(X, Y).$$

This is the l.h.s. that we are interested in. The goal is to reach the r.h.s. Let I be the unique index so that $X \in A_I \times B_I$. Let J be the unique index so that $Y \in A_J \times B_J$. Let X_1' be uniform in A_I, and let Y_2' be uniform in B_J. Some observations:

- The indices I, J are identically distributed.
- The distribution of X_1 conditioned on $I = 1$ is the same as the distribution of Y_1 conditioned on $J = 1$.
- Conditioned on I, the random variables X_1 and X_2 are independent.
- The distribution of (Y_1, J) is identical to that of (X_1', I), and similarly for (X_2, I) and (Y_2', J).

$$\begin{aligned}
&H(X,Y)\\
&\qquad = H(X,Y,I,J)\\
&\qquad = H(I) + H(X_1|I) + H(X_2|X_1,I) + H(J) + H(Y_1|J) + H(Y_2|Y_1,I)\\
&\qquad = H(I) + H(X_1|I) + H(X_2|I) + H(J) + H(Y_1|J) + H(Y_2|J)\\
&\qquad = H(I) + H(X_1|I) + H(Y_2'|J) + H(J) + H(X_1'|I) + H(Y_2|J)\\
&\qquad = H(X_1,X_1') + H(Y_2,Y_2')\\
&\qquad \leq \log\Big(\sum_i u_i^2\Big) + \log\Big(\sum_j v_j^2\Big).
\end{aligned}$$

The only inequality is the last step. There is equality iff (X_1, Y_1') is uniform on its support iff for all i,

$$\Pr[I=i] = \frac{u_i^2}{\sum_j u_j^2};$$

the r.h.s. is the only distribution on I that yield the uniform distribution on (X_1, Y_1') iff there is $c > 0$ so that $u_i v_i = c u_i^2$ for all i iff $v = cu$. □

Remark. *The equality case holds also if u, v are not positive, because if $u_i v_i < 0$ for some i then there is inequality.*

3.3 Counting perfect matchings

Here is an application is combinatorics. Let $G = (A \cup B, E)$ be a bipartite graph with $|A| = |B| = n$.

Definition 25. *A perfect matching in G is a collection of n disjoint edges.*

Remark. *A perfect matching defines a bijection between A and B.*

Remark. *The problem of deciding if G has a perfect matching can be solved in polynomial time (and is important; discuss examples).*

Definition 26. *Let M be the adjacency matrix of G; that is, $M_{a,b} = 1$ iff $\{a,b\} \in E$. The permanent of M is*

$$\mathsf{perm}(M) = \sum_\sigma \prod_a M_{a,\sigma(a)}$$

where σ is a bijection from A to B.

Remark. *It is like determinant but with no signs.*

Claim 27. $\mathsf{perm}(M)$ *is the number of perfect matchings in G.*

Remark. *Computing* $\mathsf{perm}(M)$ *is believed to be difficult (it is #P-complete; harder than NP). This is in contrast to deciding if* $\mathsf{perm}(M) > 0$*, which is in P.*

Exercise 28. $\mathsf{perm}(M) \leq \prod_a d_a$*, where* d_a *is the degree of* $a \in A$*.*

Theorem 29 (Bregman)**.** $\mathsf{perm}(M) \leq \prod_a (d_a!)^{1/d_a}$.

Remark. *The theorem is sharp for the complete bipartite graph with* $n!$ *matchings, or for a disjoint union of such graphs.*

Proof by Radhakrishnan. Let $A = [n]$. Let ρ be a uniformly random matching in G (if there are no matchings, the theorem is trivial). Denote by $\rho(a) \in B$ the neighbor of a with respect to μ. The "simple" bound:

$$\log \mathsf{perm}(M) = H(\rho) \leq \sum_a H(\rho(a)) \leq \sum_a \log d_a.$$

To get a more accurate bound, use the chain rule:

$$H(\rho) = \sum_{a=1}^{n} H(\rho(a)|\rho(1), \dots, \rho(a-1)).$$

What is the correct way to order A? Not so clear. Choose a random order. Let π be a uniform permutation of A.

$$\log(\mathsf{perm}(M)) = H(\rho) = \mathbb{E}_\pi \sum_{a=1}^{n} H(\rho(\pi(a))|\rho(\pi(1)), \dots, \rho(\pi(a-1))).$$

For $a \in [n]$, let $k = k_a$ be defined by $k = \pi^{-1}(a)$ or $a = \pi(k)$; in other words, a is the k'th element in the sum for this π. Collect terms as

$$H(\rho) = \sum_{a=1}^{n} \mathbb{E}_\pi H(\rho(a)|\rho(\pi(1)), \dots, \rho(\pi(k-1))).$$

Now, fix a and focus on

$$\mathbb{E}_\pi H(\rho(a)|\rho(\pi(1)), \dots, \rho(\pi(k-1))).$$

For each π, the conditional entropy is at most log of the number of "free" neighbors of a. That is, if we denote by $F = F(\pi, \rho, a)$ the number of neighbors of a that are not in $\rho(\pi([k-1]))$, then

$$\mathbb{E}_\pi H(\rho(a)|\rho(\pi(1)), \dots, \rho(\pi(k-1))) \leq \sum_{j=1}^{d_a} \Pr[F = j] \log j.$$

The last step is proving that

$$\Pr[F = j] = \frac{1}{d_a}.$$

Write

$$\Pr[F = j] = \mathbb{E}_\rho \Pr[F = j|\rho].$$

For fixed ρ, each neighbor b of a is matched to $\rho^{-1}(b)$. The value of F is the number of neighbors b of a so that $\rho^{-1}(b)$ is at least a in the ordered defined by π. Because π is uniform, the number of elements of $\rho^{-1}(N(a))$ that are at least a with respect to π is uniform in $[d_a]$. □

3.4 The local lemma

The probabilistic method (initiated by Erdös) tells us that if we want to prove that some object x exists then we can find a random variable X and prove that $\Pr[X = x] > 0$. This simple idea is extremely powerful.

Remark. *One of the central goals is to find* explicit *good objects. The probabilistic method typically does not provide constructions.*

A typical proof using this method shows that $\Pr[X = x]$ is very close to 1. In other words, that most objects are "good". The local lemma is one of the few methods that gives only $\Pr[X = x] > 0$.

Lemma 30 (Lovasz's local lemma)**.** *Let $A_1, \ldots, A_n$ be events (think of them as "bad"). Define a "dependency" graph G as follows. The vertices are $[n]$. There is an edge between i and j iff A_i and A_j are not independent. Assume that*

- *all degrees in the graph are at most d,*
- *there is $p > 0$ so that $\Pr[A_i] \leq p$ for all i, and*
- *$ep(d+1) \leq 1$.*

Then,

$$\Pr\left[\bigcap_i A_i^c\right] > 0.$$

(There is something "good".)

Remark. *There are three assumptions. The first is that there isn't much "dependency" between the events. The second is that each individual bad event is not likely. The third quantitively relates between the first two.*

Theorem 31 (Moser)**.** *Under the same assumptions, an $x \in \bigcap_i A_i^c$ can be found efficiently.*

Remark. *We won't prove general statement, and focus on CNF formulas.*

Consider the formula

$$f = C_1 \wedge C_2 \wedge \ldots \wedge C_s,$$

where each clause C_i is of the form

$$C_i = \ell_{i,1} \vee \ell_{i,2} \vee \ldots \vee \ell_{i,k},$$

where each literal $\ell_{i,j}$ is a variable or its negation.

Remark. *Finding x so that $f(x) = 1$ is a constraint satisfaction problem. There is a list of constraints defined by the clauses, and we need to satisfy all of them (or as many as we can). It is NP-hard in general.*

Definition 32. *For each $i \in [s]$, let $\Gamma(i)$ be the set of $j \in [s]$ so that C_i and C_j share a variable (so that $i \in \Gamma(i)$).*

Claim 33. *If for all i,*

$$|\Gamma(i)| \leq 2^{k-3}$$

then there is x so that $f(x) = 1$.

Remark. *Comparison to the local lemma:*

1. *The bad events are $A_i = \{x : C_i(x) = 0\}$. If X is uniformly random then for all i,*

$$\Pr[A_i] = 2^{-k}.$$

2. *The set $\Gamma(i)$ describes the clauses that are "not independent of C_i". The assumption in the claim implies that for all i,*

$$|\Gamma(i)| 2^{-k} \leq \frac{1}{8}.$$

Remark. *The number of variables n does not really matter.*

Proof of claim. We shall construct "better and better" assignments to f. The basic building block is the following procedure $FIX(C_i)$:

1. Choose k random bits and substitute them into C_i.
2. Go over all $j \in \Gamma(i)$ so that C_j is not satisfied (after the new assignment), and run $FIX(C_j)$.

The algorithm is simple:

> *Start with a uniformly random assignments X_0, and as long as there is an unsatisfied C_i, run $FIX(C_i)$.*

Remark. *It is not clear that the algorithm terminates.*

Remark. *If the algorithm terminates then we found a satisfying assignment.*

The argument proceeds by keeping track of two lists.

indices of FIX	*changes*
	X_0
i_1	R_1
i_2	R_2
$\dots$	$\dots$
i_t	R_t
X_t	

where the i_j's are the names of the clauses FIX was applied on, X_0 is the initial assignment, $R_j \in \{0,1\}^k$ is the k random bits that were chosen in FIX, and X_t is the assignment after running FIX for t times.

Claim 34. *If we know $i_1, i_2, \dots, i_t, X_t$ then we know $X_0, R_1, R_2, \dots, R_t$.*

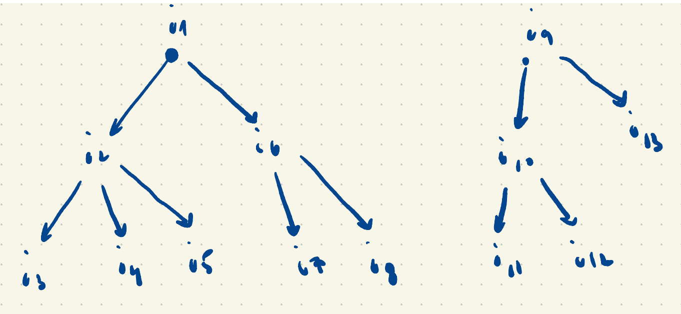
Proof. Let's see what happens for $t = 1$. If we know i_1, X_1, then in the positions outside C_{i_1} we know that X_0 and X_1 agree. In the positions inside C_{i_1}, we know X_0 because there is a single unsatisfying assignment. The positions inside C_{i_1} of X_1 are R_1. The rest is similar. □

Claim 35. *For every t, we can encode $i_1, \dots, i_t, X_t$ using at most*

$$n + t(k-3) + 2t + s\lceil \log(s) \rceil$$

bits.

Proof. The algorithm run can be described using a tree; for example,



This example can be encoded by

$$i_1, 1, i_2, 1, i_3, 0, i_4, 0, i_5, 0, 0, i_6, 1, i_7, 0, i_8, 0, 0, 0, i_9, \dots$$

where 1 indicates "go down" and 0 indicates "go up". The important facts are:

1. When we go down we can describe the next i_j using $k-3$ bits, because there are at most 2^{k-3} options.

2. The number of "direction bits" after t steps is at most $2t$.

3. There are the "roots" of the trees. No root appears twice, because when a tree is "completed" all relevant clauses are satisfied. Each root can be described using $\lceil \log s + 1 \rceil$ bits.

□

We are almost done. If the algorithm runs t steps then

$$n + kt = H(X_0, R_1, \ldots, R_t) \leq H(i_1, \ldots, i_t, X_t) \leq n + t(k-3) + 2t + s\lceil \log(s+1) \rceil$$

so that

$$t \leq s\lceil \log(s+1) \rceil.$$

□

Remark. *What have we proved?*

Exercise 36. *Describe a (probabilistic) algorithm that gets as input a k-CNF formula $f = \bigwedge_{i \in [s]} C_i$ so that $|\Gamma(C_i)| \leq 2^{k-3}$ for all i. The algorithm should output, with probability at least $2/3$, a satisfying assignment to f. Explain how to change the algorithm to get probability of success at least $1 - \delta$ for a given $\delta > 0$. How does this affect the running time?*

Remark. *The proof uses the formula f to compress information. The data $X_0, R_1, \ldots, R_t$ is just a sequence of $n + kt$ uniformly random bits. We saw that there is a (deterministic) procedure that uses f to encode this data using $s \log s + n + (k-1)t$ bits. This is so amazing that it can't be true. This can happen only for a short amount of time, only as long as $f(X_t) = 0$.*

Remark. *We can "generically" amplify the success probability. If we have running time T with success probability $\frac{2}{3}$, then we can get running time $O(T \log \frac{1}{\delta})$ with success probability $1 - \delta$. We just run the algorithm $O(\log \frac{1}{\delta})$ times, and see if one of the runs generated x so that $f(x) = 1$. But in some cases we can do better, we can obtain running time of the form $T + O(\log \frac{1}{\delta})$. This is not generic, but it works if we have "additional structure". In this case, we have this additional structure.*

3.5 Isoperimetry

In Euclidean space, the shape that minimizes surface area for fixed volume is a ball. Physics tells us that this is the reason that soap bubbles are round. A similar question can be asked in any space where volume and surface area are defined. A central example of such spaces are finite graphs.

The Boolean cube is a central graph in math and CS. Its vertices are the element of $V = \{0,1\}^n$. Two vertices are connected by an edge if they differ in a single coordinate. There are 2^n vertices and $n2^{n-1}$ edges. The volume of $A \subseteq V$ is its size. The surface area is measured by the number $\delta(A)$ of edges between A and $V \setminus A$. The isoperimetric problem is to determine $\min \delta(A)$ for fixed $|A|$. It was solved by Harper in 1966, but we shall see a proof by Samorodnitsky from 2017.

Theorem 37. *$\delta(A) \geq |A|(n - \log |A|)$.*

Remark. *If $|A| = 2^k$ then this is sharp for a subcube A of dimension k.*

Proof. Let $\delta = \delta(A)$ and let e be the number of edges whose two end-points are in A. Because the cube is n-regular,

$$\delta = n|A| - 2e.$$

We shall prove an upper bound on e.

Let X be uniform in A. For $x \in \{0,1\}^n$ and $i \in [n]$, let $x_{-i} \in \{0,1\}^{n-1}$ be x after deleting the i'th coordinate. The main observation is that for all i and $x \in \{0,1\}^n$ so that $\Pr[X = x] > 0$,

$$H(X_i|X_{-i} = x_{-i}) = \begin{cases} 1 & \text{both } x \text{ and the } i\text{'th neighbor of } x \text{ are in } A, \\ 0 & \text{otherwise.} \end{cases}$$

It follows that

$$\begin{aligned} \sum_i H(X_i|X_{-i}) &= \sum_i \frac{1}{|A|} \sum_{x \in A} H(X_i|X_{-i} = x_{-i}) \\ &= \frac{1}{|A|} \sum_{x \in A} \sum_i H(X_i|X_{-i} = x_{-i}) \\ &= \frac{1}{|A|} 2e. \end{aligned}$$

Finally,

$$\begin{aligned} \log |A| &= H(X) \\ &= \sum_i H(X_i|X_{<i}) \\ &\geq \sum_i H(X_i|X_{-i}) \\ &= \frac{1}{|A|} 2e. \end{aligned}$$

□

Chapter 4

Shearer's lemma

Shearer found a useful generalization of sub-additivity. It allows to control entropy under projections.

Remark. *Sub-additivity can be stated as follows. If I is uniform in $[n]$ and independent of X then*

$$\tfrac{1}{n}H(X) \leq \mathbb{E}_I H(X_I).$$

If X has a lot of entropy, then a random coordinate has entropy.

Notation 38. *Let $X = (X_1, \ldots, X_n)$ be a random variable. For $S \subseteq [n]$, let $X_S = (X_i : i \in S)$. For $i \in [n]$, let $X_{<i} = X_{[i-1]}$.*

Theorem 39. *If S is a random subset of $[n]$ distributed independently of X so that for each $i \in [n]$, we have $\Pr[i \in S] \geq \mu$ then*

$$\mathbb{E}_S H(X_S) \geq \mu H(X).$$

Proof by Radhakrishnan. For each $s = \{s_1 < s_2 < \ldots < s_k\}$,

$$\begin{aligned} H(X_s) &= H(X_{s_1}) + H(X_{s_2}|X_{s_1}) + H(X_{s_3}|X_{s_1}, X_{s_2}) + \ldots \\ &\geq H(X_{s_1}|X_{<s_1}) + H(X_{s_2}|X_{<s_2}) + H(X_{s_3}|X_{<s_3}) \ldots \\ &= \sum_{i \in [n]} 1_{i \in S} H(X_i|X_{<i}). \end{aligned}$$

In expectation,

$$\begin{aligned}\mathbb{E}_S H(X_S) &\geq \mathbb{E}_S \sum_{i\in[n]} 1_{i\in S} H(X_i|X_{<i}) \\ &= \sum_{i\in[n]} \mathbb{E}_S 1_{i\in S} H(X_i|X_{<i}) \\ &= \sum_{i\in[n]} H(X_i|X_{<i}) \mathbb{E}_S 1_{i\in S} \\ &\geq \sum_{i\in[n]} H(X_i|X_{<i})\mu \\ &= \mu H(X). \qquad \square\end{aligned}$$

4.1 Application: Loomis-Whitney

Let A be a compact subset in Euclidean space $\mathbb{R}^d$. We can try to understand properties of A by looking at its projections on lower dimensional spaces. Let's focus on $d = 3$ for now.

If we project A to the three main axis, and we see lengths $\lambda_1, \lambda_2, \lambda_3$, then what can we say on the volume of A? It is at most

$$|A| \leq \lambda_1\lambda_2\lambda_3$$

because A is contained in the "box" defined by the projections.

If we project A to the three main planes, and we see areas $\alpha_{12}, \alpha_{13}, \alpha_{23}$, then what can we say on the volume of A? This seems much more complicated; instead of a box we get a cylinder intersection. Now we can bound

$$|A|^2 \leq \alpha_{12}\alpha_{13}\alpha_{23}.$$

Remark. *Such an inequality must be homogeneous.*

Loomis and Whitney proved this inequality for general d. Let's use Shearer lemma to prove that.

Theorem 40. *If A is a finite subset of $\mathbb{Z}^d$ and $\pi_i(A)$ is the projection of A to the coordinates not in $i \in [d]$, then*

$$|A|^{d-1} \leq \prod_i |\pi_i(A)|.$$

Remark. *This is a discrete statement but it implies the continuous statement by approximation.*

Proof. Let X be uniformly random in A so that

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

Let I be uniform in $[d]$ of size $d-1$ chosen independently of X. For each $j \in [d]$, we have $\Pr[j \in [d] \setminus \{I\}] = \frac{d-1}{d}$. By Shearer's lemma,

$$\begin{aligned} \tfrac{d-1}{d} H(X) &\leq \mathbb{E}_I H(X_{[d]\setminus\{I\}}) \\ &\leq \mathbb{E}_I \log|\pi_I(A)|. \end{aligned}$$ □

Remark. *The only case of equality is that $X_{[d]\setminus\{i\}}$ is uniform on its support which means that A is a cube.*

Exercise 41. *Prove that now.*

Remark. *We can ask for a stable version of this inequality; if A is so that the inequality is close to being satisfied does it mean that A is close to a box? Together with Ellis, Friedgut and Kindler, we proved that. For every d, there is $c > 0$ so that for every finite $A \subset \mathbb{Z}^d$ and every $\epsilon > 0$, if*

$$|A| \geq (1-\epsilon) \prod_i |\pi_i(A)|^{\frac{1}{d-1}}$$

then there is a box $B \subset \mathbb{Z}^d$ so that

$$|A \triangle B| \leq c\epsilon |B|.$$

You can try to think how to find this box.

Remark. *The Loomis-Whitney inequality depends on the coordinate system we use. If we rotate the body, the volume is fixed but the projections may change. The equality case is achieve by an axis-oriented box. Imagine that we have a fixed body, and we randomly rotate it and then apply the d projections. Can we get a better bound on the volume? Together with Milman, we showed that we can (using Petty's inequality). This inequality turns out to be stronger than the classical isoperimetric inequality, and the equality case is achieved by a ball.*

Chapter 5

Mutual information

Entropy measures the amount of unpredictability of a system. The goal now is to measure the mutual information between two parts X and Y of the system. Mutual information is defined via entropy. It is the difference between the entropy of X and the entropy of X when we know Y.

Definition 42. *The mutual information between X and Y is*

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

Example 43. $I(X;X) = H(X)$.

Example 44. $I(X;Y) = 0$ *iff X, Y are independent.*

5.1 Data processing

If $X \to Y \to Z$ is a Markov chain (i.e., conditioned on Y, the two variable X, Z are independent) then what's the connection between $I(X;Y)$ and $I(X;Z)$?

Remark. *We can think of Z as a processing of Y that may involve some extra randomness (that is independent of X).*

Remark. *The Markovian property appears in many systems in physics. Examples?*

Claim 45. *If $X \to Y \to Z$ then*

$$I(X;Y) \geq I(X;Z).$$

Proof.

$$I(X;Z) = H(X) - H(X|Z) \leq H(X) - H(X|Y,Z) = H(X) - H(X|Y) = I(X;Y). \quad \square$$

5.2 Lower bound for indexing

Remark. *This example shall guide us through many useful ideas.*

There are two players: Alice and Bob. Alice gets a uniformly random X in $\{0,1\}^n$ and Bob gets an independent I that is uniform in $[n]$. Their goal is that Bob will know the bit X_I. To achieve this goal, Alice sends Bob some message $M = M(X,R)$ where R is some extra randomness that Bob knows as well (and is independent of X). Alice can send the n bits of X to Bob.

Can she do much better?

The first observation is that if Alice sends k bits then

$$I(X;M) \leq H(M) \leq k.$$

This amount of information, intuitively, is spread between the n possible values of I. The average information on an average i is $\frac{k}{n}$. If $k \ll n$ then Bob got very little info on X_I, and so he can't expect to know it. Now, let's formalize this.

Exercise 46 (general chain rule)**.**

$$I(X,Y;Z) = I(X;Z) + I(Y;Z|X),$$

where

$$I(Y;Z|X) = H(Y|X) - H(Y|Z,X).$$

Claim 47 (chain rule for independence)**.** *If $X_1,\ldots,X_n$ are independent then*

$$\sum_i I(X_i;Y) \leq \sum_i I(X_i;YX_{<i}) = I(X;Y).$$

Remark. *This is surprising and powerful; the information contained in Y is at most "evenly split" between the coordinates of X, due to independence.*

Proof. By independence and the chain rule,

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= \sum_i H(X_i) - H(X_i|YX_{<i}) \\ &= \sum_i I(X_i;YX_{<i}). \end{aligned}$$

And

$$\sum_i H(X_i) - H(X_i|YX_{<i}) \geq \sum_i H(X_i) - H(X_i|Y)$$
$$= \sum_i I(X_i;Y). \qquad \square$$

We can start formalizing the intuiting above; if $k \ll n$ then M tells Bob very little on X_I:

$$I(X_I;M|I) = \sum_i \tfrac{1}{n} I(X_i;M) \leq \tfrac{k}{n}.$$

That's a start. But how can we show that Bob can't predict X_I? For this, we introduce two measures of "distance" between distributions.

KL divergence

Definition 48. *The KL divergence between two distribution p,q over the same set is*

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

where $0 \log \frac{0}{0} = 0$.

Remark. *The KL divergence can be infinite.*

Remark. *It is part of the family of f-divergences that we shall discuss later on:*

$$D(p||q) = \sum_x q(x) \frac{p(x)}{q(x)} \log \frac{p(x)}{q(x)} = \mathbb{E}_{x\sim q} f(\tfrac{p(x)}{q(x)})$$

where $f(z) = z \log z$ is so that f is convex in $[0,\infty)$ and $f(1) = 0$.

Claim 49. $D(p||q) \geq 0$.

Remark. *The claim is sometimes called Gibbs's inequality.*

Proof.

$$D(p||q) = \sum_x q(x) f(\tfrac{p(x)}{q(x)}) \geq f\Big(\sum_x q(x) \tfrac{p(x)}{q(x)}\Big) = 0,$$

where $f(z) = z \log z$. $\square$

Remark. *Some intuition for $D(p||q) \geq 0$. If we know that $X \sim q$ then we can encode X using approximately $H(q)$ bits. Each x is encoded using $\approx \log \frac{1}{q(x)}$ bits. But what if we are* *wrong? What if we think that the input distribution is q, but the true distribution is p? This creates some inefficiency. The loss is*

$$D(p||q) = \sum_x p(x) \log \frac{1}{q(x)} - H(p).$$

Remark. *In general $D(p||q) \neq D(q||p)$.*

Remark. *If p_X is the distribution of $X \in [n]$ and u is the uniform distribution on $[n]$ then*

$$H(X) = \log(n) - D(p_X||u).$$

In words, the divergence from the uniform distribution is (up to a constant) the entropy.

Remark. *The KL divergence is deeply related to mutual information. The mutual information is the divergence between the true distribution and the product of the marginals:*

$$I(X;Y) = D(p_{X,Y}||p_X \times p_Y).$$

And the mutual information is also the divergence between the (posterior) distribution of Y conditioned on $X = x$, and the (prior) distribution of Y:

$$I(X;Y) = \mathbb{E}_x D(p_{Y|x}||p_Y).$$

Remark. *Entropy is defined only for discrete random variables. Divergence and hence mutual information are defined more generally. The expression $p(x)/q(x)$ is the Radon-Nikodym derivative; for example, if we have two densities f, g on $\mathbb{R}^d$ then*

$$D(f||g) = \int_x f(x) \log(\tfrac{f(x)}{g(x)}) dx.$$

Statistical distance

Definition 50. *The statistical distance between two distribution p, q is*

$$|p - q| := \max\{p(E) - q(E) : E \textit{ is an event}\}.$$

In words, the probabilities of events in p and q are the same up to $|p - q|$.

Exercise 51. $2|p - q| = \|p - q\|_1 = \sum_x |p(x) - q(x)|$.

Remark. *The maximizing event can be identified from the two histograms: $E = \{x : p(x) > q(x)\}$. The event E is the part of the world that the histogram of p is above that of q.*

Remark. *If $X \sim p$ and $Y \sim q$ then there is a coupling of X, Y (a joint distribution on X, Y so that each is distributed correctly) so that*

$$\mathbb{P}[X \neq Y] \leq 2|p - q|.$$

Here's a sketch of the construction. Draw the two histograms in $[n] \times [0,1]$, where $[n]$ is the support of both p and q. Let $(z_1, r_1), (z_2, r_2), \ldots$ be an infinite sequence of i.i.d. uniform points in $[n] \times [0,1]$. Let

$$T_X = \min\{t : r_t \leq p(z_t)\}$$

and

$$T_Y = \min\{t : r_t \leq q(z_t)\}.$$

Set

$$X = z_{T_X} \quad \& \quad Y = z_{T_Y}.$$

The first claim is that both X, Y are correctly distributed. The second claim is that

$$\Pr[X \neq Y] \leq \Pr[T_X \neq T_Y] \leq \Pr[T_X < T_Y] + \Pr[T_X > T_Y] \leq 2|p - q|.$$

Intuitively, $\Pr[T_X < T_Y]$ *means that the dart we threw at time T_X fell under the histogram of p but not under that of q. This means that out of the area of $1/n$ under the histogram of p, the arrow landed in the "wrong region" that has area $|p - q|/n$.*

Remark. *This coupling can be viewed as a game between two players. Alice known p and Bob knows q. Alice's goal is to sample X from p, Bob's is to sample Y from q, but they also want $X = Y$ with as high chance as possible. The above allows to achieve this with no communication, using just shared public randomness. In fact, this is done when Alice does not know q and Bob does not know p.*

Relating the two

Remark. *We defined $D(p||q)$ and $|p - q|$. The former is useful because it is related to information and satisfies chain rules:*

$$D(p_{X,Y}||q_{X,Y}) = D(p_X||q_X) + \mathbb{E}_{x \sim p_X} D(p_{Y|x}||q_{Y|x}).$$

The latter is useful because it is related to probabilities of events. The following important inequality connects between the two.

Theorem 52 (Pinsker)**.** *$D(p||q) \geq 2|p - q|^2$.*

Proof sketch. If $D(p||q) = \infty$ then we are done. So we can assume that if $q(x) = 0$ then $p(x) = 0$. Let E be the event that maximizes $p - q$. Think of X as a random variable that is distributed as p or q:

$$D(p||q) = D(p_X||q_X).$$

Let $Y \in \{0,1\}$ be the indicator of $X \in E$. By the chain rule and because divergence is non-negative:

$$\begin{aligned} D(p_{X,Y}||q_{X,Y}) &= D(p_X||q_X) + \mathbb{E}_{x\sim p} D(p_{Y|x}||q_{Y|x}) \\ &= D(p_X||q_X) + 0 \end{aligned}$$

and

$$D(p_{X,Y}||q_{X,Y}) \geq D(p_Y||q_Y).$$

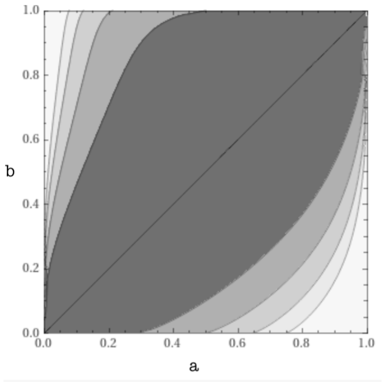
It remains to prove

$$D(p_Y||q_Y) \geq 2(p(Y=1) - q(Y=1))^2.$$

In other words, we need to prove that for all $a, b \in [0,1]$,

$$a \log \tfrac{a}{b} + (1-a)\log \tfrac{1-a}{1-b} - 2(a-b)^2 \geq 0.$$

This is a two-dimensional statement and can be verified via elementary calculus. Here's the two-dimensional picture (lighter color is higher value):



□

Remark. *The square in Pinsker's inequality must be there (and is sharp). The divergence between the distribution of $X \sim Ber(\frac{1}{2}+\epsilon)$ and $Y \sim Ber(\frac{1}{2})$ is $O(\epsilon^2)$ and the statistical distance is ϵ. This is related to the fact that two distinguish between X and Y via sampling we must toss the coin order $\frac{1}{\epsilon^2}$ times.*

Putting it together

Going back to indexing, recall that

$$I(X_I; M|I) \leq \tfrac{k}{n}.$$

Write the l.h.s. as

$$\mathbb{E}_I \mathbb{E}_M D(p_{X_I|M} || p_{X_I}).$$

The distribution p_{X_I} is uniform in $\{0,1\}$; denote it by u. By Pinsker and convexity,

$$\mathbb{E}_I \mathbb{E}_M D(p_{X_I|M} || u) \geq \mathbb{E}_I \mathbb{E}_M 2 |p_{X_I|M} - u|^2 \geq 2(\mathbb{E}_I \mathbb{E}_M |p_{X_I|M} - u|)^2.$$

Together,

$$\mathbb{E}_I \mathbb{E}_M |p_{X_I|M} - u| \leq \sqrt{\tfrac{k}{2n}}.$$

In other words, if $k \leq \frac{n}{50}$ then even after Bob knows M, the bit X_I is $\frac{1}{10}$-close to uniform, so Bob can't expect to guess it correctly with probability higher than $\frac{1}{2} + \frac{1}{10}$.

Remark. *Even with no communication Bob can output a correct guess with probability* $\frac{1}{2}$.

Remark. *A summary of the approach: First, use the chain rule for mutual information to move from n coordinates to a single random coordinate. Second, use Pinsker to move from mutual information to statistical distance.*

Chapter 6

Harmonic functions

Remark. *Harmonic functions are the kernel of the laplacian. There is a rich theory with connections to many areas in math. The term "harmonic" comes from the wave equation (imagine a string of a guitar). Here we see how information theory can help to analyze the space of harmonic functions. The model for "geometry" we use is that of Cayley graphs.*

Notation 53. *Let $G = \langle S \rangle$ be a group generated by a finite set S. Assume that S is symmetric*

$$S = S^{-1} = \{s^{-1} : s \in S\}.$$

Definition 54. *The Cayley graph of G, S has vertex-set G, and its edges are of the form $\{g, sg\}$ for $s \in S$. It is denoted by $Cay(G, S)$.*

Remark. *The graph is undirected because S is symmetric.*

Remark. *Our focus is on infinite groups.*

Example 55. *The Cayley graph of $(\mathbb{Z}, \pm 1)$ is the line. The Cayley graph of $(\mathbb{Z}^2, (\pm 1, \pm 1))$ is "the standard square lattice". The Cayley graph of the free group with k generators is the $(2k)$-regular infinite tree.*

Remark. *Cayley graphs provide a powerful framework for studying the group.*

Definition 56. *A function $f : G \to \mathbb{R}$ is harmonic (with respect to S) if for every $g \in G$, the value $f(g)$ is equal to the average of f on the neighbors of g:*

$$f(g) = \frac{1}{|S|} \sum_{s \in S} f(sg).$$

Example 57. *In a star with four vertices, the middle value is zero, and the values of the four neighbors are $-1, -1, 2$.*

Example 58. *The constant functions are harmonic.*

Example 59. *A linear function on* $(\mathbb{Z}, \pm 1)$ *is harmonic. This is true also for all* S *and for higher dimensions.*

Example 60. *Build a bounded harmonic function on the 4-regular tree ("there is always a possibility to continue").*

Remark. *The space of harmonic functions is a vector space over* $\mathbb{R}$*. Its dimension teaches us something about the "geometry".*

Remark. *For complex functions* $f : \mathbb{C} \to \mathbb{C}$*, if* f *is holomorphic (i.e. differentiable everywhere) and* $f = u + iv$*, then both* u, v *are harmonic.*

Remark. *Liouville's theorem for complex functions states that a bounded holomorphic function is constant. It can be extended to all harmonic functions on* $\mathbb{R}^n$*. (The theorem was actually proved by Cauchy.) The property of* $\mathbb{R}^n$ *that is key in the proof is that the boundary size of balls in* $\mathbb{R}^n$ *are much smaller than their volume (this follows from commutativity or polynomial-growth; more on this below).*

Definition 61. *A graph is called Liouville if every bounded harmonic function on it is constant.*

Definition 62. *A random walk on the Cayley graph starts at* $X_0 \in G$*, and* X_{t+1} *is defined from* X_t *by choosing a uniformly random element* s_{t+1} *of* S*, independently of all previous choices, and setting* $X_{t+1} = s_{t+1} X_t$*.*

Remark. *The behavior of the random walk tells us a lot about the graph and the group.*

Theorem 63 (Kaimanovich-Vershik, Avez)**.** *Let* (X_t) *be a random walk on* $Cay(G, S)$ *started at* $X_0 = id$*. If*

$$\lim_{t \to \infty} \frac{H(X_t)}{t} = 0$$

then $Cay(G, S)$ *is Liouville.*

Remark. *The theorem is actually an "if and only if" statement.*

Remark. *A central open problem in this area is "is Liouville a group property?" Is it true that the Liouville property holds for all generating sets, or for none?*

Remark. *Because we are dealing with a Cayley graph, the entropy* $H(X_t | X_0 = x_0)$ *does not depend on* x_0*.*

Exercise 64. *Prove that the limit of* $\frac{H(X_t)}{t}$ *exists.*

Example 65. *The Cayley graph of* $\mathbb{Z}^2$ *with* $(0, \pm 1), (\pm 1, 0)$ *can be thought of as a discrete version of* $\mathbb{C}$ *or* $\mathbb{R}^2$*. Because* G *is abelian, it has polynomial growth; the support of* X_t *is of size at most* t^2*. It follows that*

$$H(X_t | X_0) \leq 2 \log t.$$

Every bounded harmonic function on this graph is therefore constant.

Example 66. *If G is the free group generated by two generators, then the Cayley graph is a four regular infinite tree, and there are bounded harmonic functions that are not constant. What is the entropy of X_t? It must be $\Omega(t)$; in fact it is roughly $\log(3)t$.*

Remark. *Let us start building the bridge between harmonicity and entropy, between analysis and information theory.*

Lemma 67. *Let X, Y be random variables and let r be a function. If $\mathbb{E}r(X) = 0$ and $\|r\|_\infty \leq 1$ then*

$$\mathbb{E}_Y\big|\mathbb{E}[r(X)|Y]\big| \leq \sqrt{2I(X;Y)}.$$

Proof. For every y,

$$\begin{aligned}\Big|\sum_x p(x|y)f(x)\Big| &= \Big|\sum_x (p(x|y)-p(x))f(x)\Big| && (\mathbb{E}f = 0)\\ &\leq \sum_x |p(x|y)-p(x)| && (\|f\|_\infty \leq 1)\\ &\leq \sqrt{2D(P_{X|y}||p_X)}. && (\text{Pinsker})\end{aligned}$$

Taking expectation over y completes the proof (using convexity). □

Remark. *Mutual information allows to control an analytic quantity.*

Remark. *The worst-case condition $\|r\|_\infty \leq 1$ can be replaced by the average-case condition $\mathbb{E}r^2 \leq 1$. A typical way to do so is to replace the $|\langle u, v\rangle| \leq \|u\|_\infty \|v\|_1$ inequality by Cauchy-Schwarz.*

Remark. *We shall use the lemma with $X = X_1$ and $Y = X_t$. Because we are dealing with a Cayley graph,*

$$I(X;Y) = I(X_1;X_t) = H(X_t) - H(X_t|X_1) = H(X_t) - H(X_{t-1}).$$

We need to control these differences.

Claim 68. *The map $t \mapsto H(X_t)$ is increasing and concave.*

Remark. *The differences $H(X_t) - H(X_{t-1})$ get smaller as t grows, so that*

$$H(X_t) - H(X_{t-1}) \leq \frac{H(X_t)}{t}.$$

Remark. *A crucial property behind the claim is the* sub-modularity *of entropy (also known as "strong sub-additivity"). For three random variables A, B, C,*

$$H(A,B,C) + H(C) \leq H(A,C) + H(B,C).$$

It can be proved as follows:

$$\begin{aligned} H(A,C)+H(B,C)-H(A,B,C) &= H(C)+H(A|C)+H(B|C)-H(A,B|C) \\ &= H(C)+I(A;B|C). \end{aligned}$$

Sub-modularity is important in economics and computer science. A function $f : 2^X$ is sub-modular if

$$f(A)+f(B) \geq f(A\cup B)+f(A\cap B).$$

An equivalent formulation: if $A \subseteq B$ and $x \notin B$ then

$$f(A\cup\{x\})-f(A) \geq f(B\cup\{x\})-f(B).$$

This is the "diminishing returns" property; if we think of f as "gain" then our gain from x when we have A is at least as large than when we have $B \supseteq A$. Stated differently,

$$f(A\cup\{x\})+f(B) \geq f(B\cup\{x\})+f(A),$$

which follows from the first condition with $A\cup\{x\}$ and B because $A\cup\{x\}\cup B = B\cup\{x\}$ and $(A\cup\{x\})\cap B = A$. The other direction can be proved by induction.

Proof of claim. Monotonicity holds because

$$H(X_{t+1}) \geq H(X_{t+1}|X_1) = H(X_t).$$

Let's prove

$$2H(X_t) \geq H(X_{t-1})+H(X_{t+1}).$$

Write $X_{t-1}=g_1$, $X_t=g_2g_1$ and $X_{t+1}=g_3g_2g_1$, where g_1 is the position after $t-1$ steps, and g_2,g_3 are uniform in S (and all are independent). Then,

$$H(X_t)=H(g_2g_1)=H(g_1g_3)$$

and

$$H(X_{t+1})=H(g_2g_1g_3).$$

By sub-modularity, because g_2g_1, g_3, g_1g_3, g_2 determine all of g_1,g_2,g_3,

$$H(g_2g_1,g_3)+H(g_1g_3,g_2) \geq H(g_1,g_2,g_3)+H(g_2g_1g_3).$$

So,

$$H(X_t)+H(X_t)+H(g_3)+H(g_2) \geq H(X_{t-1})+H(g_3)+H(g_2)+H(X_{t+1}). \qquad \square$$

Remark. *The properties of $e(t)=H(X_t|X_0)$ hold also for finite groups. For a finite Cayley graph, $e(t) \leq \log|G|$, and $e(t)$ monotonically tends to $\log|G|$ in a concave manner (more on this in the next chapter).*

Lemma 69 (Benjamini-(Duminil-Copin)-Kozma-Yadin)**.** *If f is harmonic, then almost surely over X_0,*

$$\mathbb{E}|f(X_1) - f(X_0)| \leq \sqrt{\frac{2H(X_t)}{t}} \cdot \sup |f(X_t) - f(X_0)|.$$

Remark. *The lemma implies that for a (connected) Cayley graph with $\frac{H(X_t)}{t} \to 0$, every bounded harmonic function is constant. The right hand side tends to zero, which means that $f(X_1) = f(X_0)$ a.s.*

Remark. *The limit is zero for all finite groups, so every harmonic function on a finite connected Cayley graph is constant. This is true for all finite connected graphs (V, E). If f is harmonic and non-constant, and if v is the maximizer of f then the value of f on the neighbors of v must be $f(v)$ as well, and so forth. This is a finite version of the "maximum principle".*

Remark. *Going back to infinite Cayley graphs, if f is a non-constant harmonic function then there is $c > 0$ and g_0 so that*

$$\sup\{|f(g)| : dist(g, g_0) \leq t\} \geq c\sqrt{\frac{t}{H(X_t)}}.$$

On $\mathbb{Z}^2$ as before we see that every non-constant harmonic function grows at least as quickly as order $\sqrt{\frac{t}{\log t}}$. This bound can be improved with a more careful analysis.

Definition 70. *A sequence of random variables $M_0, M_1, \ldots$ is a Martingale with respect to $X_0, X_1, \ldots$ if for every t we have*

$$\mathbb{E}[M_{t+1} | X_{\leq t}] = M_t.$$

Remark. *In particular, the random variable M_t is measurable with respect to the σ-algebra generated by $X_{\leq t}$.*

Exercise 71. *If f is harmonic then $M_t = f(X_t)$ is a Martingale with respect to $(X_t)_{t=0}^{\infty}$.*

Remark. *Martingales have many useful properties; concentration of measure, anti-concentration, etc.*

Proof of lemma. Fix X_0, and assume without loss of generality that $f(X_0) = 0$. Because f is harmonic, for all $t \geq 1$,

$$\mathbb{E}f(X_t) = \mathbb{E}\ \mathbb{E}[f(X_t)|X_{<t}] = \mathbb{E}f(X_{t-1}) = \ldots = f(X_0) = 0$$

and similarly

$$\mathbb{E}[f(X_t)|X_1] = f(X_1).$$

By Lemma 67,

$$\begin{aligned}\mathbb{E}_{X_1}|f(X_1)| &= \mathbb{E}_{X_1}\big|\mathbb{E}[f(X_t)|X_1]\big| \\ &\leq \sqrt{2I(X_1;X_t)} \cdot \sup|f(X_t)|.\end{aligned}$$

And as we saw

$$I(X_1;X_t) = H(X_t) - H(X_t|X_1) = H(X_t) - H(X_{t-1}) \leq \frac{H(X_t)}{t}. \qquad \square$$

Chapter 7

The second law of thermodynamics

The second law of thermodynamic says that the entropy increases with time. (We shall not go into the history or physics.)

Remark. *This is unique in that most equations in physics are oblivious to changing the direction of time.*

Question 72. *Is it surprising?*

Remark. *The model we use for the "states of the world" is the vertices of a finite graph, and the transitions between states ("laws of physics") are the edges of the graph.*

Question 73. *Does the second law always hold?*

Example 74. *Consider a random walk on a path of length two. Start it in the middle vertex. At even times the entropy of X_t is zero, and at odd times it is one. The second law does not hold.*

Example 75. *Consider a random walk on a triangle. The entropy of X_0 is zero, of X_1 is one, and of $X_2, \ldots$? The entropy increases in this case.*

Question 76. *What is the difference between the two systems?*

Brief intro to random walks

Let G be a connected and non-bipartite graph.

Definition 77. *Consider the matrix*

$$M_{x,y} = \frac{1_{\{x,y\} \in E}}{deg(x)}.$$

In words, $M_{x,y}$ is the probability that a (simple) random walk moves from x to y.

Remark. *If at time t the random walk has distribution p_t, then at time $t+1$ the distribution is $p_{t+1} = p_t M_t$:*

$$p_{t+1}(y) = \sum_x p_t(x) M_{x,y}.$$

In matrix form, $p_{t+1} = p_t M$.

Remark. *It follows that $p_t = p_0 M^t$. This indicates that spectral properties of M are important in understanding the behavior of the random walk. Because p_t is a probability distribution, and not a general vector, not only the spectral properties determine the behavior.*

Remark. *An important notion is "stationary" distribution; a distribution π so that*

$$\pi = \pi M.$$

If we start with π, then we remain with π for all times.

Claim 78. *The random walk on G has a stationary distribution π.*

Remark. *Intuitively, the higher the degree of x is (compared to others), the more likely we are to visit x.*

Proof. The stationary measure is defined by

$$\pi_x = \frac{deg(x)}{2|E|},$$

because

$$(\pi M)_y = \sum_x \frac{deg(x)}{2|E|} 1_{\{x,y\} \in E} \frac{1}{deg(x)} = \frac{deg(y)}{2|E|}.$$ □

Remark. *Because G is not bipartite, the stationary measure is unique. We shall not prove this now; e.g., the Perron-Frobenius theorem.*

Remark. *For every initial distribution p_0, we have that $p_t \to \pi$ as $t \to \infty$.*

Remark. *If π is not uniform (i.e., the graph is not regular) then the entropy can decrease, because if p_0 is uniform (with maximum entropy) then $H(p_t) < H(p_0)$ for some large enough t.*

Remark. *We saw that entropy is basically the divergence from the uniform distribution. This leads to the following general second law.*

The second law

Theorem 79. *For every initial distribution p_0, the map $t \mapsto D(p_t || \pi)$ is decreasing.*

Corollary 80. *If π is uniform then the entropy of p_t is increasing.*

The theorem follows from the following more general lemma.

Lemma 81. *Let p_0, q_0 be two initial distributions on $V = V(G)$. Denote by p_t, q_t the distributions of the random walks at time t. Then,*

$$D(p_{t+1}||q_{t+1}) \leq D(p_t||q_t).$$

Remark. *The lemma is more general because if $q_t = \pi$ is stationary then $q_{t+1} = \pi$.*

Proof of lemma. Let X be the state at time t and Y be the state at time $t+1$, so that

$$a_{X,Y}(x,y) = p_t(x)M_{x,y}$$

and

$$b_{X,Y}(x,y) = q_t(x)M_{x,y}.$$

By the chain rule for divergence (twice):

$$D(a_{X,Y}||b_{X,Y}) = D(a_X||b_X) + \mathbb{E}_{x\sim a_X} D(a_{Y|x}||b_{Y|x}) = D(p_t||q_t) + \mathbb{E}_{x\sim p_X} 0$$

and

$$D(a_{X,Y}||b_{X,Y}) = D(a_Y||b_Y) + \mathbb{E}_{y\sim a_Y} D(a_{X|y}||b_{X|y}) \geq D(p_{t+1}||q_{t+1}) + \mathbb{E}_{x\sim p_X} 0. \qquad \square$$

Example 82. *Consider a star with k leaves $v_1, \ldots, v_k$ and a root v_0. This graph is bipartite. If the random walk starts at v_0 then at even times it is on v_0 and at odd time it is on the leaves. So p_t alternates between $1_{x=v_0}$ and the uniform distribution on the k leaves. A stationary distribution π assigns v_0 weight $\frac{1}{2}$ and each leaves weight $\frac{1}{2k}$, so that*

$$D(p_0||\pi) = 1 \log \frac{1}{1/2} = 1$$

and

$$D(p_1||\pi) = \sum_{x=1}^{k} \frac{1}{k} \log \tfrac{1/k}{1/(2k)} = 1.$$

This second law still holds, but in a not so interesting way.

Remark. *If G is regular then π is uniform. In this case, there is symmetry in time when $p_0 = \pi$. The law of (X_0, X_1) is the same as the law of (X_1, X_0). At the same time, both $H(X_t|X_1)$ and $H(X_t)$ are increasing in time. Even in systems that are symmetric in time, some quantities grow with time.*

Remark. *Under the stationary distribution, the information about the future gets smaller with time. If p_0 is stationary then*

$$t \mapsto I(X_t; X_0)$$

is decreasing:

$$\begin{aligned} I(X_{t+1};X_0) &= H(X_{t+1}) - H(X_{t+1}|X_0) \\ &\leq H(X_t) - H(X_{t+1}|X_0, X_1) \\ &= H(X_t) - H(X_{t+1}|X_1) \\ &= H(X_t) - H(X_t|X_0) = I(X_t;X_0). \end{aligned}$$

Remark. *The second law for entropy holds for "regular" systems where the uniform distribution is the stationary one. What does it say about the world?*

Chapter 8

Mutual information and sampling

Remark. *If $I(X;Y) = 0$ then X, Y are independent. What if $I(X;Y) = 1$? It is tempting to guess that the following is true: If $I(X;Y) = 1$ then there is a random variable Z so that $H(Z) = 1$ and conditioned on $Z = z$, the variables X and Y become independent. This turns out to be too good to be true.*

Example 83. *For every k, there is a distribution on X, Y so that $I(X;Y) \leq 1$ but for every Z so that conditioned on $Z = z$, the variables X, Y are independent, it holds that $H(Z) \geq k$. The distribution of X, Y is constructed as follows. Let M be an $n \times n$ Boolean matrix so that the largest monochromatic sub-matrix of M has size at most $O(n)$. A random matrix satisfies this with high probability. Let (X, Y) be a random entry in M among the one entries. The mutual information $I(X;Y)$ is*

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \lesssim 2\log(n) - \log(n^2/2) = 1.$$

If Z is so that conditioned on $Z = z$, the variables X, Y are independent so that

$$H(X,Y|Z) \leq \log(n) + O(1)$$

and

$$2\log(n) \lesssim H(X,Y) \leq H(X,Y,Z) = H(Z) + H(X,Y|Z).$$

So that

$$H(Z) \gtrsim 2\log(n) - \log(n).$$

Remark. *What is the correct statement?*

Theorem 84. *For every jointly distributed random variables X, Y, there is a distribution on three random variables (X, Y, Z) so that*

1. *(X, Y) is distributed correctly.*

2. *Z is independent of X.*

3. Y is a deterministic function of (X, Z).

4. $H(Y|Z) \leq I(X,Y) + log(I(X,Y)+1) + C$, where $C > 1$ is some universal constant.

Remark. *The theorem can be interpreted as via the following story. Alice wants to sample X and Bob wants to sample Y, but X, Y are jointly distributed. Alice samples X, and wishes to send a message M to Bob that tells him what is the appropriate Y. They have access to public randomness Z. Alice can send Bob $\approx I(X;Y)$ bits to achieve this task.*

Remark. *This was proved by Harsha-Jain-McAllester-Radhakrishnan, and by Braverman-Garg.*

Remark. *In a paper with Bassily, Moran, Nachum and Shafer we found applications to learning theory.*

Remark. *The proof actually works in a more general scenario.*

Remark. *There are two distributions on p and q on Y. Alice knows both of them, and Bob just knows q. The players have public randomness, and Alice want to send Bob a message that allows him to sample from p. Alice can send Bob a message of expected length*

$$D(q||p) + \log(D(q||p) + 1) + C.$$

This is a stronger "point-wise" guarantee; in the theorem above, Alice first choses x and then there are $p = p_{Y|x}$ and $q = p_Y$.

Proof. What is the random variable Z? It is an infinite sequence $(z_1, \alpha_1), (z_2, \alpha_2), \ldots$ of i.i.d. samples that are uniform in $[n] \times [0,1]$, where Y takes values in $[n]$.

Remark. *This is reminiscent of the coupling we saw earlier.*

Alice samples x from the correct distribution. Then she finds

$$T = \min\{t : \alpha_t < p(z_t)\}.$$

Alice sends T to Bob, who now knows y. As we saw, the eventual (x, y) are properly distributed.

How should Alice convey T to Bob?

The data T is encodes in three parts:

$$K = \left\lceil \frac{T}{n} \right\rceil, \quad Q = \left\lceil \frac{\alpha_T}{q(z_T)} \right\rceil$$

and the index I of T among all n samples in $\{(K-1)n+1, \ldots, Kn\}$ that are consistent with Q. That is, among the indices t in that interval so that

$$Q = \left\lceil \frac{\alpha_t}{q(z_t)} \right\rceil .$$

We need to bound the entropy of each part separately.

Claim 85. *$H(K) \leq O(1)$.*

Proof idea. The entropy of K is bounded because

$$\Pr[T = 1] = \tfrac{1}{n}.$$

This means that $\Pr[K > 1] \leq \frac{3}{4}$ and in general $\Pr[K > k] \leq (\frac{3}{4})^k$. □

Claim 86. *$\mathbb{E}[\log Q] \leq I(X;Y) + 2$.*

Proof idea. First, bound $\mathbb{E}[\log Q]$ from above. By the choice of T and Q,

$$\begin{aligned}
\mathbb{E}[\log Q] &\leq \mathbb{E}_{z\sim p}\mathbb{E}\Big[\log\Big(\frac{\alpha}{q(z)}+1\Big)\Big|\alpha < p(z)\Big] \\
&= \sum_{z\in[n]} p(z)\cdot\frac{1}{p(z)}\int_0^{p(z)} \log\Big(\frac{\alpha}{q(z)}+1\Big)d\alpha \\
&\leq \sum_{z\in[n]} \int_0^{p(z)} \log\Big(\frac{p(z)}{q(z)}+1\Big)d\alpha \\
&= \sum_{z\in[n]} p(z)\log\Big(\frac{p(z)}{q(z)}+1\Big).
\end{aligned}$$

Use that $\log(\xi+1) - \log(\xi) \leq \frac{2}{\xi}$ to bound

$$\begin{aligned}
\mathbb{E}[\log Q] &\leq \sum_{z\in[n]} p(z)\Big(\log\Big(\frac{p(z)}{q(z)}\Big) + \frac{2\Pr[Y=z]}{p(z)}\Big) \\
&= D(q||p) + 2.
\end{aligned}$$

□

Claim 87. *$H(I) \leq O(1)$.*

Proof idea. There are n possible indices we care about. For each index t,

$$\begin{aligned}\Pr[t \in I] &= \Pr[Q - 1 < \frac{\alpha_t}{q(z_t)} \leq Q] \\ &= \Pr[(Q-1)\Pr[Y = z_t] < \alpha_t \leq Qq(z_t)] \\ &\leq \sum_z \Pr[z_t = z] \cdot \Pr[(Q-1)q(z) < \alpha_t \leq Qq(z)] \\ &\leq \sum_z \frac{1}{n} \cdot q(z) = \frac{1}{n}.\end{aligned}$$

In words, among the n indices, typically only a constant number are consistent with Q. It follows that $H(I)$ is constant (the details are left as an exercise). □

□

Exercise 88. *For a random variable M taking values in $\mathbb{N}$,*

$$H(M) \leq \mathbb{E}[\log(M)] + 2\log(\mathbb{E}[\log(M+1)]) + C.$$

Exercise 89. *There is a random variable M taking value in $\mathbb{N}$ so that*

$$H(M) > \mathbb{E}[\log(M)] + \log(\mathbb{E}[\log(M)]).$$

Remark. *Entropy is deeply related to prefix-free encoding. There is a prefix-free encoding E_0 of $\mathbb{N}$ so that for each n we have $|E_0(n)| \leq n$; this is the unary encoding. There is a prefix-free encoding E_1 of $\mathbb{N}$ so that for each n we have $|E_1(n)| \leq 2\log(n)$; this is obtained by adding a bit after each bit of the binary encoding, and the additional bits are* $000\ldots001$. *By first encoding $\lceil \log n \rceil$ using E_1, we get a prefix-free encoding E_2 so that for each n we have $|E_2(n)| \leq \log(n) + 1 + 2\log(\log(n)+1)$. The code E_2 is slightly better. We can continue...*

Exercise 90. *If E is a prefix-free encoding of $\mathbb{N}$ so that $|E(n)|$ decreases with n then $|E(n)| \geq \log(n) + \log(\log(n)) + \omega(1)$.*

Chapter 9

Graph entropy

Remark. *There are many natural communication scenarios, and each leads to different mathematics. Körner suggested the following scenario, which lead to the definition of graph entropy. We have a source of randomness $X \sim p^T$ where p is the uniform distribution on a finite set V. (We focus on the uniform distribution for concreteness.) We want to understand how efficiently we can encode X. But there is a catch. Some of the symbols in V are distinguishable and some are not.*

This is captured by the edges of a graph $G = (V, E)$. If $\{v, u\} \in E$ then the two alphabet symbols v, u are distinguishable. If $\{v, u\} \notin E$ then they are indistinguishable. Two strings $x, x' \in V^T$ are indistinguishable if all of their coordinates are indistinguishable. The T'th power G^T of G is the graph with vertex-set V^T where $x \neq x'$ in V^T are connected by an edge iff $\{x_i, x'_i\} \in E$ for some $i \in [T]$.

Our goal is to find an encoding f of X with maximum entropy. The encoding must respect the structure; that is, if x, x' are distinguishable then $f(x) \neq f(x')$. We also allow an ϵ-fraction of errors. For $U \subseteq V^T$ so that $p^T(U) \geq 1 - \epsilon$, an encoding f that is "proper" on U defines a proper coloring of the vertices in U. The encoding length is therefore at most

$$\log \chi(G^T|_U)$$

where $G^T|_U$ is the induced graph on U. This leads to the following definition.

Definition 91. *The ϵ asymptotic rate of G is defined to be*

$$R_\epsilon(G) = \lim_{T \to \infty} \min_U \frac{1}{T} \log \chi(G^T|_U),$$

where $U \in V(G^T)$ is so that $p^T(U) > 1 - \epsilon$.

Remark. *Körner proved that the limit exists and the result is given by the* graph *entropy.*

Definition 92. *The graph entropy of G is*

$$H(G) = \min_Y I(X; Y),$$

where the minimum is over a random independent set Y in the graph that must contain X.

Theorem 93 (Körner). *For all $\epsilon > 0$,*

$$R_\epsilon(G) = H(G).$$

Remark. *We shall not prove that, but we shall see some examples and applications.*

Example 94. *If G is the empty graph, we can choose $Y = V$ and $H(G) = 0$.*

Example 95. *If G is the complete graph, the only choice we have is $Y = \{X\}$, and $H(G) = \log n$.*

Example 96. *If G is bipartite then*

$$H(G) \leq 1$$

be choosing Y to be the color class of X, because

$$H(X) = \log n$$

and

$$H(X|Y) = q \log(qn) + (1-q)\log((1-q)n) = \log n - h(q)$$

where q is the fractional size of one of the color classes.

Remark. *Here is another equivalent definition (we shall not prove the equivalence).*

Lemma 97. *The* independent set *polytope $ISP(G) \subset \mathbb{R}^V$ of G is the convex hull of the indicators of independent sets in G. It is a compact convex set. The graph entropy of G is*

$$H(G) = \inf_{a \in ISP(G)} \frac{1}{|V|} \sum_{x \in V} \log \tfrac{1}{a_x},$$

where a is assumed to be positive everywhere.

Properties

Remark. *It is possible to control the growth of graph entropy under several operations. We mention a couple.*

Lemma 98. *If G_1, G_2 have the same vertex set and $G_1 \subseteq G_2$ then*

$$H(G_1) \leq H(G_2).$$

Proof. An independent set in G_2 is also an independent set in G_1. □

Lemma 99. *If G_1, G_2 have the same vertex set then*

$$H(G_1 \cup G_2) \leq H(G_1) + H(G_2).$$

Proof. Let Y_1 be the minimizer for G_1, and Y_2 be the minimizer for G_2. The set $Y_1 \cap Y_2$ is independent in $G_1 \cup G_2$, and contains X. Conditioned on the value of X, the two sets Y_1, Y_2 are independent. So,

$$\begin{aligned} H(G_1 \cup G_2) &\leq I(X; Y_1 \cap Y_2) \\ &\leq I(X; Y_1, Y_2) \\ &= H(Y_1, Y_2) - H(Y_1, Y_2 | X) \\ &= H(Y_1, Y_2) - H(Y_1 | X) + H(Y_2 | X) \\ &\leq H(Y_1) + H(Y_2) - H(Y_1 | X) + H(Y_2 | X). \end{aligned}$$ □

Exercise 100. *If G_1, G_2 are graphs on disjoint vertex-sets, and G is their union, then*

$$H(G) = \frac{|V(G_1)|}{|V(G)|} H(G_1) + \frac{|V(G_2)|}{|V(G)|} H(G_2).$$

Remark. *We move to two applications.*

Covering graphs

Remark. *The problem of covering a target graph using graphs from some family of graphs is natural and has applications. One specific example is covering K_n by bipartite graphs.*

Theorem 101. *If $B_1, \ldots, B_t$ are bipartite graphs with vertex-set $[n]$, and their union is the complete graph, then $t \geq \log n$.*

Proof.

$$\log n = H\Big(\bigcup_i B_i\Big) \leq \sum_i H(B_i) \leq t.$$ □

Remark. *This can also be proved by analyzing the chromatics number ($\chi(G_1 \cup G_2) \leq \chi(G_1)\chi(G_2)$).*

Computational complexity

Remark. *The main goal is to understand the minimum number of operations that are needed to achieve some goal. Here we focus on a specific situation (studied by Krichevskii, and we shall follow proofs by Newman, Ragde and Wigderson and by Radhakrishnan).*

Definition 102. *A monotone boolean formula is a rooted binary tree whose leaves are labelled by variables $x_1, \ldots, x_n$ and inner nodes by $\vee, \wedge$.*

Remark. *A boolean formula computes a function is the obvious way. The cost of the computation is the size of the formula; i.e., the number of leaves in it.*

Remark. *A general formula can also have negated variables at the leaves, but we do not address this model here.*

Remark. *Given a monotone function $f : \{0,1\}^n \to \{0,1\}$, its monotone formula complexity $M(f)$ is the least size of a monotone formula computing f.*

Remark. *On a high-level, devices have costs and functions have complexities.*

Remark. *The monotone formula complexity of a monotone n-variate function is at most $n2^n$.*

Theorem 103. *Let $T = T_{n,2}$ be the threshold function that is 1 iff $\sum_i x_i \geq 2$. Then $M(T) \geq n \log n$.*

Remark. *The proof proceed by constructing a "progress" measure μ. This is a measure whose value on the leaves is small, and does not grow much during the operations. So if $\mu(f)$ is large then f must require many operations. This is a standard and natural method for proving complexity lower bounds. The hard part is to come up with the "correct" measure.*

Definition 104. *For a monotone boolean function f and an integer k, denote by f_k the set of inputs x of weight $|x| = k$ so that $f(x) = 1$ and if $y < x$ then $f(x) = 0$. The set f_1 can be thought of as a subset of the coordinates $[n]$. The set f_2 can be thought of as a graph. Define*

$$\mu(f) = H(f_2) + \frac{|f_1|}{n}.$$

Proof. The function T has large measure

$$\mu(T) = H(K_n) + 0 = \log n.$$

If f is computed on a leaf then

$$\mu(f) = 0 + \frac{1}{n}.$$

If $f = h \vee g$ then

$$f_k \subset h_k \cup g_k.$$

The subadditivity of graph entropy yields

$$\mu(f) \leq \mu(h) + \mu(g).$$

If $f = g \wedge h$ then the argument is a bit more complicated, because

$$f_1 \subseteq g_1 \cap h_1$$

but

$$f_2 \subseteq g_2 \cup h_2 \cup E$$

where

$$E = \{\{i, j\} : i \in g_1 \setminus h_1, j \in h_1 \setminus g_1\}.$$

(If $x \in f_2$ then $g(x) = h(x) = 1$, but it could be that $x \notin g_2 \cup h_2$, because it is not "minimal".) We need to understand $H(E)$. The graph E has two parts. One part is a complete bipartite graph with sides $g_1 \setminus h_1$ and $h_1 \setminus g_1$. Its entropy is at most one, and the number of its vertices is $|g_1 \cup h_1| - |g_1 \cap h_1|$. The other part is the empty graph. We can conclude that

$$H(E) \leq 0 + \frac{|g_1 \cup h_1| - |g_1 \cap h_1|}{n} \cdot 1.$$

Therefore,

$$\mu(f) \leq \frac{|h_1 \cap g_1|}{n} + H(g_2) + H(h_2) + \frac{|g_1 \cup h_1| - |g_1 \cap h_1|}{n} \leq \mu(g) + \mu(h).$$

We can finally conclude that if T is computed by a formula of size s then

$$\log n = \mu(T) \leq s \cdot \frac{1}{n}.$$

□