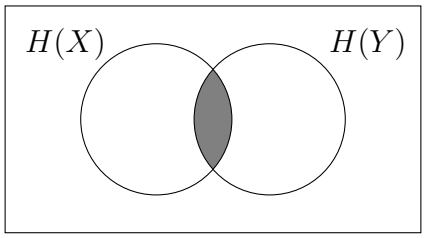
Information Theory with Applications, Math6397
Lecture Notes from September 11, 2014

taken by Dax Mahoney

2 previous notes continued

2.1 Warm-Up

2.1.1 Question. Recall the definition of mutual information.



The mnemonic device using Venn diagrams shows you several ways to write it:

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

We are going to examine another form of data processing for Markov chains. As you move along in the Markov Chain, the only thing that can happen is you lose information. So as time progresses you might be forgetting things and mutual information between the inital and current state goes down. We only really need to look at three moments in time because we are comparing between an intitial point in time, an intermediate time, and a final time.

2.2 A Data Processing Inequality for Markov Chains

2.2.2 Proposition. *Given a Markov chain consisting of three discrete r.v.'s $\{X, Y, Z\}$ (ordered in sequence of "time"), then $I(X;Y) \geq I(X:Z)$.*

Proof. We use the property that if I condition ona random variable at the intermediate time (in this case Y), then X and Z are independent if I conditioned on Y: $I(X;Z|Y)=0$.

By additivity, reversing the order and conditioning on Z gives

$$\begin{aligned} I(X;Z)+I(X;Y|Z) &= I(X;Y|Z) \\ &= I(X;Y,Z) \\ \text{(flip the role of Y and Z)} &= I(X;Y)+I(X:Z|Y) \\ \text{(additivity)} &= I(X;Y) \end{aligned}$$

By $I(X;Y|Z)\geq 0$, we obtain $I(X;Y)\geq I(X;Z)$.
We also know that equality holds if and only if $I(X;Y|Z)=0$ □

What does equality mean? Since the conditional mutual information averages the mutual information for the conditional distributions over the outcomes of Z, for any outcome for which Z has a nonzero probability, either X determines Y or Y determines X.

Additionally, I can use the same proof here in order to make another statement.

2.2.3 Theorem. *Given a Markov chain X,Y,Z as above, the $I(X;Y|Z)\leq I(X;Y)$. In words, the conditional mutual information is less than or equal to the mutual information.*

Proof.

$$\text{From the proof of the previous assertion, } \underbrace{I(X;Z)}_{\geq 0}+I(X;Y|Z)\ \underbrace{=I(X;Y)}_{\text{by the additivity}}$$

□

2.2.4 Remark. Before we had concluded that conditioning reduced entropy, so conditioning reduces uncertainty. But here, conditioning reduces mutual information. This can be interpreted that the additional information about Y gained from observing Z makes X and Y appear more independent. The above comparison is generally only true if you have a Markov chain, otherwise it may or may not hold.

2.2.5 Remark. For example, if $\{X,Y,Z\}$ are random variables that have outcomes

$$\begin{array}{cc} (0,0,0) & (0,1,1) \\ (1,0,1) & (1,1,0) \end{array}$$

with equal probability. You can observe that if $Z=0$ then we have $(0,0)$ or $(1,1)$ with equal probability. So knowing Z does not give any information about X alone. However, knowing Z and Y determines X. On the other hand, just looking at X and Y shows that all possible outcomes have equal probability $1/4$, so they are independent, $I(X;Y)=0$. We see that the conditional mutual information is

$$\begin{aligned} I(X;Y|Z) &= \textstyle\sum_{(a,b,c)\in\{0,1\}} \mathbb{P}_Z(c) I(X;Y|Z=c) \\ &= \tfrac{1}{2}(\tfrac{1}{2}\ln\tfrac{\frac{1}{2}}{\frac{1}{4}}+\tfrac{1}{2}\ln\tfrac{\frac{1}{2}}{\frac{1}{4}}) \\ &\quad +\tfrac{1}{2}(\tfrac{1}{2}\ln\tfrac{\frac{1}{2}}{\frac{1}{4}}+\tfrac{1}{2}\ln\tfrac{\frac{1}{2}}{\frac{1}{4}}) \\ &= \ln 2\,, \end{aligned}$$

larger than $I(X;Y)$.

This is the end of the prelude and of the introduction of the main players related to entropy. We are at the first few pages of Khinchin.

3 Sources and Coding

3.1 Asymptotic Equipartitioning

We start with the simplest type of source, a very trivial Markov chain. It spits something out and immediately forgets what it produced. What information theory does is relates how sequences behave in an asymptotic way in terms of entropy, in terms of something that only looks at the individual distruibution.

3.1.1 Definition. A discrete memoryless source (DMS) is a stochastic process $\{X_j\}_{j=1}^{\infty}$ consisting of independently and identically distributed (i.i.d) random variables with an at most countable alphabet $\mathbb{A}$.

3.1.2 Theorem. *Wimpy Asymptotic Equipartitioning Property (not in the book)*
Given a DMS (discrete memoryless source) $\{x_j\}_{j=1}^{\infty}$, then

$$-\frac{1}{n}\ln \mathbb{P}_{X_1,X_2,\ldots,X_n}(X_1,X_2,\ldots,X_n) \xrightarrow{n\to\infty} H(X_1) \text{ in probability}$$

3.1.3 Remark. This converges to the entropy of any one of those copies. What it really says is that if we look at this probability, it decays exponentially, and the rate of exponential decay is $H(X_1)$. So with probability one in the limit the asymptotic rate is within $\pm\epsilon$ of the number $H(X_1)$. The rate of decay is fixed. The asymptotic rate of decay is as close to $H(X_1)$ as we want.

Proof. We note $-\ln \mathbb{P}_{X_j}(X_j)$ are independent and identically distributed. So by the Weak Law of Large Numbers, by making n sufficiently large, we can make them as close as I want to their expected value.

$$\mathbb{P}(|\frac{1}{n}\sum_{j=1}^{n}\ln \mathbb{P}_{X_j}(X_j) - \underbrace{\mathbb{E}[ln\mathbb{P}_{X_1}(X_1)]|}_{\substack{\text{since they are all i.i.d we can take} \\ X_1 \text{ and this is now } -H(X_1)}} < \epsilon) \xrightarrow{n\to\infty} 1$$

□

This is an overwhelming theme in information theory, we split between things that happen with overwhelming probability and ignore the remaining outcomes (outside of of an ϵ range of the decay rate $H(X_1)$. So this motivates another definition.

3.1.4 Definition. We call a sequence $\{x_1, x_2, \ldots, x_n\}$ (note these are lower case x's) **weakly ϵ-typical**, if it is exactly in this range of probabilities $|\frac{1}{n}\ln \mathbb{P}_{X_1,X_2,\ldots,X_n}(x_1,x_2,dots,x_n)+H(X_1)| < \epsilon$ and we combine all such x in the set A_ϵ^n

This is where we get to the real statement on asymptotic equipartitioning.

3.1.5 Theorem. *Shannon-McMillan-Breiman Theorem Let $\epsilon > 0$, for a DMS $\{X_j\}_{j=1}^{\infty}$ with $\mathbb{E}_{X_1}[\ln \mathbb{P}(X_1)^2)] < \infty$. Consider $\{A_\epsilon^n\}_{n=1}^{\infty}$ then the following holds*

1. $\mathbb{P}((X_1, X_2, \ldots, X_n) \notin A_\epsilon^n) < \epsilon \, \forall$ *sufficiently large* n
2. $$\underbrace{|A_\epsilon^n|}_{\text{size of} A_\epsilon^n} > (1-\epsilon)e^{n(H(X_1)-\epsilon)} \, \forall \text{ sufficiently large } n$$
$$|A_\epsilon^n| < (1-\epsilon)e^{n(H(X_1)+\epsilon)} \, \forall \, n \in \mathbb{N}$$
3. *If $x \in A_\epsilon^n$, then for sufficiently large n,* $e^{-n(H(X_1)+\epsilon)} < \mathbb{P}(X_1, \ldots, X_n) < e^{-n(H(X_1)-\epsilon)}$

So the importance for coding, when you send data with a cell phone you give yourself some slack probability of encountering an atypical sequence. Accordingly for typical sequences, you ennumerate the sequences. You can say something about the size of these typical messages and can now send a compressed message. The only quantity that becomes relevant is the entropy. On the whole what you are talking about is squences, long sequences with a little bit of slack. So you can make a statement about the whole coding scheme based on the entropy.

Proof. By Chebyshev, for $n > \frac{\sigma^2}{\epsilon^3}$ with $\sigma^2 = \mathbb{E}_x[(\ln \mathbb{P}_{x_1})^2 - H(x_1)^2]$
we have $\mathbb{P}((X_1, \ldots, X_n) \notin A_\epsilon^n))$
$= \mathbb{P}(|-\frac{1}{n}\ln \mathbb{P}(X_1, \ldots, X_n) - H(X_1)| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} < \epsilon$

Using the above, for $n > \frac{\sigma^2}{\epsilon^3}$,

$$1 - \epsilon \leq 1 - \frac{\sigma^2}{n\epsilon^3} \leq \sum_{x \in A_\epsilon^n} \mathbb{P}_{x_1, x_2, \ldots, x_n}(x)$$

By exponentiating

$$< \sum_{x \in A_\epsilon^n} e^{-n(H(X_1)-\epsilon)}$$

from the defintion of the typical set, the sum is summed over a constant so

$$= \underbrace{|A_{\epsilon|}^n}_{\text{size of} A_\epsilon^n} > (1-\epsilon)e^{n(H(X_1)-\epsilon)}$$

This holds for all sufficiently large n For all $n \in \mathbb{N}$,

$$1 \geq \sum_{x \in A_\epsilon^n} \mathbb{P}_{x_1, x_2, \ldots, x_n}(x) > \sum_{x \in A_\epsilon^n} e^{-n(H(x_1)+\epsilon)}$$

so

$$= |A_\epsilon^n| e^{-n(H(X_1)+\epsilon)}$$

The next part is just by the defintion of the set A_ϵ^n, Inequalitites for probabiliity of any $x \in A_\epsilon^n$follows from the defintion of A_ϵ^n □

The main point of this enumeration business is this bound in the size of the set. We have control over the size of A_n then each element has the same probability, and we only have to deal with the uniform distribution.

In the near future, we will discuss typical sequences and the consequences of asymptotic equipartitioning for coding.

3.2 Block Codes for Discrete Memory Sources

3.2.6 Definition. An (n, m) block code is a set $\mathcal{C}$ of size n together with a map $\phi : \mathbb{A}^n \to \mathcal{C}$

As a consequence of the Shannon-McMillan-Breiman Theorem we have the following

3.2.7 Theorem. *Block Coding Theorem Let $\{X_j\}_{j=1}^{\infty}$ be a DMS with (marginal since all copies of one r.v.) entropy $H(X_1)$ and $\epsilon > 0$ then there is $0 < \delta < \epsilon$ and a sequence of codes $\{(\mathcal{C}_n, \phi_n)\}_{n=1}^{\infty}$ with block sizes $\{m_n{}_{n=1}^{\infty}\}, |\mathcal{C}_n| = m_n$such that $\frac{1}{n} \ln m_n < H(x) + \delta$ and there exist maps $\{\psi\}_{n=1}^{\infty}\}$ where $\psi : \mathcal{C}_n \to \mathbb{A}^n$ such that for all sufficiently large n, the failure probability of decoding is bounded by $\mathbb{P}(\psi_n \circ \phi_n(X_1, X_2, X_3, \ldots, X_n)) \neq (X_1, X_2, X_3, \ldots, X_n)) < \epsilon$*

The ϕ_n is for encoding, the ψ_n for decoding. The standard assumption is that the ϵ is the tolerance for errors. Th bound on the size of $\mathcal{C}_n$ tells you how long the blocks (for any given code alphabet) are going to be. If the source has low entropy, then we can compress, meaning achieve small block sizes. However, this theorem does not tell you *how* to do it, it gives you a performance bound.