EE376A Information Theory Lecture 4 - 01/15/2015

Lecture 4: Asymptotic Equipartition Properties

Lecturer: Tsachy Weissman Scribe: Phan Minh Nguyen, Haomiao Jiang, George Supaniratisai

1 Recap on Mutual Information

The mutual information between two random variables X and Y with joint probability mass function P_{XY} and marginal mass functions P_X and P_Y, respectively, is given by:

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X;Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) - H(X|Y) \\ &= \mathcal{D}(P_{XY} \| P_X \times P_Y) \end{aligned}$$

This quantity measures a certain kind of dependency between X and Y. When X and Y are independent, the mutual information is zero. We will see later that the mutual information emerges as the answer to some fundamental questions.
Some properties of the mutual information:

1. $I(X;Y) \geq 0$, coming from the fact that $H(Y) \geq H(Y|X)$.
2. $I(X;Y) \leq \min\{H(X), H(Y)\}$, since the conditional entropies are non-negative. The equality occurs iff there exists a deterministic function f s.t. $Y = f(X)$ or $X = f(Y)$ (so that either $H(Y|X)$ or $H(X|Y)$, respectively, is zero).

We introduce the notation $X - Y - Z$ to reflect that

$$\begin{aligned} &X \text{ and } Z \text{ are conditionally independent given } Y \\ \Leftrightarrow &(X,Y,Z) \text{ is a Markov triplet} \\ \Leftrightarrow &p(x,z|y) = p(x|y)p(z|y) \\ \Leftrightarrow &p(x|y,z) = p(x|y) \\ \Leftrightarrow &p(z|y,x) = p(z|y) \end{aligned}$$

For example, let X, W_1, W_2 be three independent Bernoulli random variables, with $Y = X \oplus W_1$ and $Z = Y \oplus W_2$. Then, X and Z are conditionally independent given Y, i.e., $X - Y - Z$. Intuitively, Y is a noisy measurement of X, and Z is a noisy measurement of Y. Since the noise variables W_1 and W_2 are independent, we only need Y to infer X.
We can also show that if $X - Y - Z$, then

1. $H(X|Y) = H(X|Y,Z)$
2. $H(Z|Y) = H(Z|X,Y)$
3. $H(X|Y) \leq H(X|Z)$
4. $I(X;Y) \geq I(X;Z)$, and $I(Y;Z) \geq I(X;Z)$

Intuitively, $X - Y - Z$ indicates that X and Y are more closely related than X and Z. Therefore $I(X;Y)$ (i.e., the dependency between X and Y) is no smaller than $I(X;Z)$, and $H(X|Y)$ (the uncertainty in X given knowledge Y) is no greater than $H(X|Z)$.

2 Asymptotic Equipartition Property (AEP)

Reading: Chapter 3 of Cover and Thomas.

Some notations:

- For a set $\mathcal{S}$, $|\mathcal{S}|$ denotes its cardinality (number of elements contained on the set). For example, let $\mathcal{U} = \{1, 2, \ldots, M\}$, then $|\mathcal{U}| = M$.
- $u^n = (u_1, \ldots, u_n)$ is an n-tuple of u.
- $\mathcal{U}^n = \{u^n |\ u_i \in \mathcal{U};\ i = 1, \ldots, n\}$. It is easy to see that $|\mathcal{U}^n| = |\mathcal{U}|^n$.
- "U_i generated by a memoryless source U" means $U_1, U_2, \ldots$ i.i.d. according to U (or P_U). That is,

$$p(u^n) = \prod_{i=1}^{n} P_U(u_i)$$

Definition 1. *The sequence u^n is ϵ-typical for a memoryless source U for $\epsilon > 0$, if*

$$\left| -\frac{1}{n} \log p(u^n) - H(U) \right| \leq \epsilon$$

or equivalently,

$$2^{-n(H(U)+\epsilon)} \leq p(u^n) \leq 2^{-n(H(U)-\epsilon)}$$

Let $A_\epsilon^{(n)}$ denote the set of all ϵ-typical sequences, called the typical set.

So a length-n typical sequence would assume a probability approximately equal to $2^{-nH(U)}$. Note that this applies to memoryless sources, which will be the focus on this course[1].

Theorem 2 (AEP)**.** *$\forall \epsilon > 0$, $P\left(U^n \in A_\epsilon^{(n)}\right) \to 1$ as $n \to \infty$.*

Proof This is a direct application of the Law of Large Numbers (LLN).

$$\begin{aligned}
P\left(U^n \in A_\epsilon^{(n)}\right) &= P\left(\left| -\frac{1}{n} \log p(U^n) - H(U) \right| \leq \epsilon\right) \\
&= P\left(\left| -\frac{1}{n} \log \prod_{i=1}^{n} p(U_i) - H(U) \right| \leq \epsilon\right) \\
&= P\left(\left| \frac{1}{n} \left[\sum_{i=1}^{n} -\log p(U_i)\right] - H(U) \right| \leq \epsilon\right) \\
&\to 1 \text{ as } n \to \infty
\end{aligned}$$

where the last step is due to the Law of Large Numbers (LLN), in which $-\log p(U_i)$'s are i.i.d. and hence their arithmetic average converges to their expectation $H(U)$. □

This theorem tells us that with very high probability, we will generate a typical sequence. But how large is the typical set $A_\epsilon^{(n)}$?

[1] For a different definition of typicality, see e.g. [1]. For treatment of non-memoryless sources, see e.g. [2], [3].

Theorem 3. $\forall \epsilon > 0$ *and sufficiently large* n*,*

$$(1-\epsilon)2^{n(H(U)-\epsilon)} \leq \left|A_\epsilon^{(n)}\right| \leq 2^{n(H(U)+\epsilon)}$$

Proof The upper bound:

$$1 \geq P\left(U^n \in A_\epsilon^{(n)}\right) = \sum_{u^n \in A_\epsilon^{(n)}} p(u^n) \geq \sum_{u^n \in A_\epsilon^{(n)}} 2^{-n(H(U)+\epsilon)} = \left|A_\epsilon^{(n)}\right| 2^{-n(H(U)+\epsilon)},$$

which gives the upper bound. For the lower bound, by the AEP theorem, for any $\epsilon > 0$, there exists sufficiently large n such that

$$1-\epsilon \leq P\left(U^n \in A_\epsilon^{(n)}\right) = \sum_{u^n \in A_\epsilon^{(n)}} p(u^n) \leq \sum_{u^n \in A_\epsilon^{(n)}} 2^{-n(H(U)-\epsilon)} = \left|A_\epsilon^{(n)}\right| 2^{-n(H(U)-\epsilon)}.$$

□

The intuition is that since all typical sequences assume a probability about $2^{-nH(U)}$ and their total probability is almost 1, the size of the typical set has to be approximately $2^{nH(U)}$. Although $\left|A_\epsilon^{(n)}\right|$ grows exponentially with n, notice that it is a relatively small set compared to $\mathcal{U}^n$. For some $\epsilon > 0$, we have

$$\frac{\left|A_\epsilon^{(n)}\right|}{|\mathcal{U}^n|} \leq \frac{2^{n(H(U)+\epsilon)}}{2^{n\log|\mathcal{U}|}} = 2^{-n(\log|\mathcal{U}|-H(U)-\epsilon)} \to 0 \text{ as } n \to \infty$$

given that $H(U) < \log|\mathcal{U}|$ (with strict inequality!), i.e., the fraction that the typical set takes up in the set of all sequences vanishes exponentially. Note that $H(U) = \log|\mathcal{U}|$ only if the source is uniformly distributed, in which case all the possible sequences are typical.

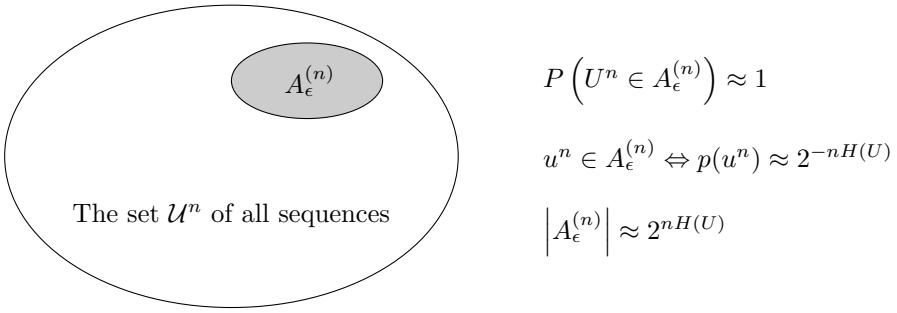

Figure 1: Summary of AEP

In the context of lossless compression of the source U, the AEP tells us that we may only focus on the typical set, and we would need about $nH(U)$ bits, or $H(U)$ bits per symbol, for a decent representation of the typical sequences.

References

[1] A. E. Gamal and Y.-H. Kim, *Network Information Theory*, Cambridge Univ. Press, UK, 2012.

[2] R. M. Gray, *Entropy and Information Theory*, Springer-Verlag, New York, 1990.

[3] T. S. Han, *Information-Spectrum Methods in Information Theory*, Springer, 2003.