Lecture 4: The AEP

Biology 497
Carl Bergstrom

January 15, 2008

Sources: This lecture largely follows the presentation in R. W. Yeung (2002), *A First Course in Information Theory*, though I have adapted some of the approach and terminology to be consistent with that in Cover and Thomas 2nd edition. Again, some of the text and equations are taken directly from those sources.

At the core of information theory is a theorem called the Asymptotic Equipartition (AEP) theorem. If you understand this theorem and why it is true, you'll understand intuitively most of the rest of the course. You'll know why we are able to prove most of the theorems that we will prove, and design most of the coding schemes that we will design. Let us start by stating and proving the theorem, and then we can consider some of its consequences and implications.

Theorem 1 *The asymptotic equipartition theorem. Let $X_1, X_2, X_3, \ldots$ be independent and identically distributed (iid) random variables drawn from some distribution with probability function $p(x)$, and let $H(X)$ be the entropy of this distribution. Then as $n \to \infty$,*

$$-\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) \to H(X)$$

in probability.

Once we see what the statement above means, it is extremely easy to prove. Yet as we will see, this theory it is almost uncannily powerful when coupled with the notion of a "typical set" which we will develop shortly.

To unpack the statement of the theorem above note that on the left hand side we are computing the probability of the realized sequence of random variables. This seems like a funny thing to do — once we know what the realized sequence is, isn't its probability 1? This is not what we mean. We are not computing the probability of the realized sequence conditional that is has been realized, but rather the probability that is assigned by the probability function to that realized sequence. Bayesian language makes it a bit clearer what this probability is: it is the prior probability of the realized sequence.

By "converges in probability", we mean that as n gets large, $-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n)$ gets arbitrarily close to $H(x)$ with arbitrarily high probability. Formally, for n sufficiently large

$$\Pr\left[\left|-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) - H(X)\right| \leq \epsilon\right] > 1-\epsilon$$

With the meaning of the theorem explained, the result is an immediate consequence of the (weak) law of large numbers.

Theorem 2 *The weak law of large numbers. The sample mean of n sequential i.i.d. draws of a random variable X_1, X_2, … converges in probability to the expected value μ of the random variable as n gets large. That is, for n sufficiently large,*

$$Pr\left[\left|\frac{1}{n}(X_1 + X_2 + \ldots, X_n) - \mu\right| \leq \epsilon\right] > 1-\epsilon$$

Proof of the AEP: The X_i are i.i.d., so

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) = -\frac{1}{n}\sum_{k=1}^{n}\log p(X_k)$$

But the $\log p(X_k)$ terms on the right hand side are themselves random variables, and those random variables are also i.i.d. Therefore, by the weak law of large numbers,

$$-\frac{1}{n}\sum_{k=1}^{n}\log p(X_k) \rightarrow -E\log p(X)$$

in probability as $n \rightarrow \infty$. Now we simply recall that $H(X) = E[\log p(x)]$ and the proof is complete.

Corresponding to this theorem is the notion of a "typical set" of sequences; here the idea is that as n gets large, only some sorts of sequences are likely to be observed.

Definition 1 *The weakly typical set (hereafter "typical set") of sequences A_ϵ^n for a probability distribution $p(x)$ is the set of sequences $x_1, x_2, \ldots$ with average probability very close to the entropy of the random variable X:*

$$\left| -\frac{1}{n} \log p(x) - H(X) \right| \leq \epsilon$$

With this theorem and definition in place, a number of results follow directly as the length of a sequence gets large ($n \to \infty$).

1. **The probability of any sequence in the typical set is given by the entropy rate.**

   For $x \in A_\epsilon^n$,

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

   This follows directly from the definition of a typical sequence.

2. **The probabilities of all sequences in the typical set are the same.** This follows from the tight (converging to zero-width) bound in the previous item.

3. **Almost all realized sequences will be members of the typical set:**

   $$\Pr\Big[X \in A_\epsilon^n\Big] > 1 - \epsilon$$

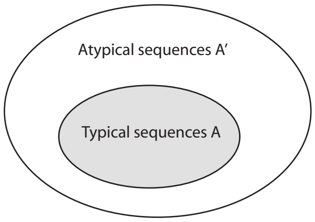
   This follows from the AEP theorem.

4. **The typical set has approximately are $2^{nH(X)}$ members.** This follows from the tight bound on probability (1) of a typical sequence and the fact that almost all realized sequences are in the typical set (3) above.

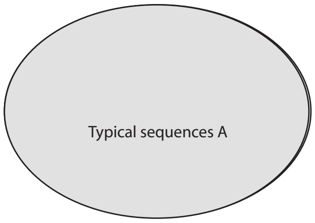
5. **Most sequences are not weakly typical**. In fact, the fraction of possible sequences that are typical is vanishingly small, unless all sequences are equally likely in which case all sequences are typical sequences. This follows because

   $$\frac{|A_\epsilon^n|}{|X|^n} \equiv \frac{2^{nH(X)}}{2^{n \log |X|}} = 2^{-n(\log |X| - H(x))} \to 0$$

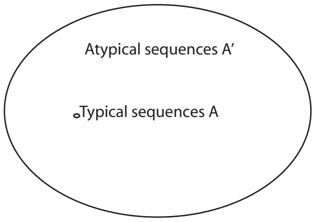
Interestingly, notice that the most likely sequence is often not a member of the typical set! Yeung provides one example: Suppose that X is an i.i.d. Bernoulli random variable with $p(0) = 0.1$ and $p(1) = 0.9$. Now the sequence $(1, 1, \ldots, 1)$ of all 1's is the most likely, but it is not in the typical set because its average probability is not close to the entropy of the random variable X. How can the typical set approach probability 1 while excluding the most likely sequences?This is possible only because those sequences become vanishingly infrequent as $n \to \infty$. In other words, they have higher probability than those sequences in the typical set, but the number of such items becomes extremely small compared to the number of items in the typical set.



Cover and Thomas illustrate the set of all sequences as shown at left.



If we represent areas according to probability, the set of typical sequences is (almost) as large as the set of the all sequences.



If we represent areas according to the size (number of members) of each set, the set of typical sequences is vanishingly small within the set of all sequences.

Example: Introduction to Shannon's Source Coding theorem. Loosely speaking, Shannon's Source Coding theorem states that given a source S that emits symbols X_1, X_2, X_3, etc., one can transmit this information through a noiseless channel at a rate arbitrarily close to the entropy rate $H(X)$ of the source.

The AEP suggests a way for us to do so. For now, we'll keep things simple and assume that our X_i's are iid. (Later we can generalize this). Suppose we use a code with *block-length* n: each keyword represents n symbols X_i from the source. The AEP tells us that as n gets large, there are approximately $2^{nH(X)}$ typical sequences, each with probability close to $2^{-nH(X)}$. Moreover, for any ϵ we can choose n such that the probability of observing an atypical sequence is less than ϵ.

So we use the following simple code. Suppose we want our rate to be within δ of $H(X)$. Pick a block length n such that the probability of observing an atypical sequence is less than $\epsilon = 2^{-n \log |\mathcal{X}|}\delta$. Now simply encode each typical sequence by enumerating its position within an ordered list of typical sequences (we know can order such a list because we are working with a discrete symbol set). This will require $2^{nH(X)}$ bits since there are $2^{nH(X)}$. Encode each atypical sequence by enumerating its position within an ordered list of all possible sequences. This will require $2^{-n \log |\mathcal{X}|}$ bits. The average number of bits required to transmit a block of size n will therefore be $(1 - \epsilon)2^{nH(X)} + \epsilon\, 2^{n \log |\mathcal{X}|} = (1-\epsilon)2^{nH(X)} + \delta < 2^{nH(X)} + \delta$. Thus we are transmitting at a rate within δ of the entropy rate $H(X)$.

While Cover and Thomas do not do this, one can strengthen the notion of a typical set for random variables with a finite number of values. This gives us the notion of a *strongly typical set*, treated in Chapter 5 of Yeung.

Definition 2 *The strongly typical set of sequences T_ϵ^n for a probability distribution $p(x)$ is the set of sequences $x_1, x_2, \ldots$ for which each possible value of the random variable X is represented close to its expected number of times.*

$$\sum_x \left| \frac{1}{n} N\Big(x; (X_1, X_2, \ldots, X_i)\Big) - \log p(x) \right| \le \epsilon$$

All of the analogous properties then hold as $n \to \infty$:

- The probability of each strongly typical set is close to the entropy rate.
- The probabilities of all sequences in the strongly typical set are the same.
- Almost all realized sequences will be members of the strongly typical set.
- The strongly typical set has approximately $2^{nH(X)}$ members.
- Most sequences are not members of the strongly typical set.

Moreover, the strongly typical set is contained within the typical set, but the converse does not hold. For example, consider a random variable $p(0) = .25, p(1) = .25, p(2) = .5$. The sequence $(0, 2, 0, 2, 0, 2, \ldots, 0, 2)$ is a member of the typical set but not a member of the strongly typical set.