Electrical Engineering 229A Lecture 6 Notes

Daniel Raban

September 14, 2021

1 The Asymptotic Equipartition Property and Data Compression

1.1 The asymptotic equipartition property

Last time, we discussed the asymptotic equipartition property (AEP). Given an iid sequence of random variables $X_1, X_2, \dots \sim (p(x), x \in \mathscr{X})$ with $\mathscr{X}$ finite, the weak law of large numbers applied to the sequence $\log \frac{1}{p(X_1)}, \log \frac{1}{p(X_2)}, \dots$ tells us that for every $\varepsilon > 0$,

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^n \log \frac{1}{p(X_i)} - \mathbb{E}\left[\log \frac{1}{p(X)} \right] \right| < \varepsilon \right) \xrightarrow{n \to \infty} 1.$$

Note that $\frac{1}{n} \sum_{i=1}^n \log \frac{1}{p(X_i)} = \frac{1}{n} \log \frac{1}{p^n(X_1^n)}$ because $p^n(X_1^n) = \prod_{i=1}^n p(X_i)$ from the iid assumption. Also note that $\mathbb{E}[\log \frac{1}{p(X)}] = H(X)$. In other words,

$$\mathbb{P}\left(-\varepsilon < \frac{1}{n} \log \frac{1}{p(X_1^n)} - H(X) < \varepsilon \right) \xrightarrow{n \to \infty} 1.$$

We can also write this as

$$\mathbb{P}\left(2^{-nH} 2^{-n\varepsilon} < p^n(X_1^n) < 2^{-nH} 2^{n\varepsilon} \right) \xrightarrow{n \to \infty} 1.$$

We define the set of **ε-weakly typical sequences** $A_\varepsilon^{(n)} \subseteq \mathscr{X}^n$ as

$$A_\varepsilon^{(n)} := \{ x_1^n \in \mathscr{X}^n : 2^{-nH} 2^{-n\varepsilon} < p^n(x_1^n) < 2^{-nH} 2^{n\varepsilon} \}.$$

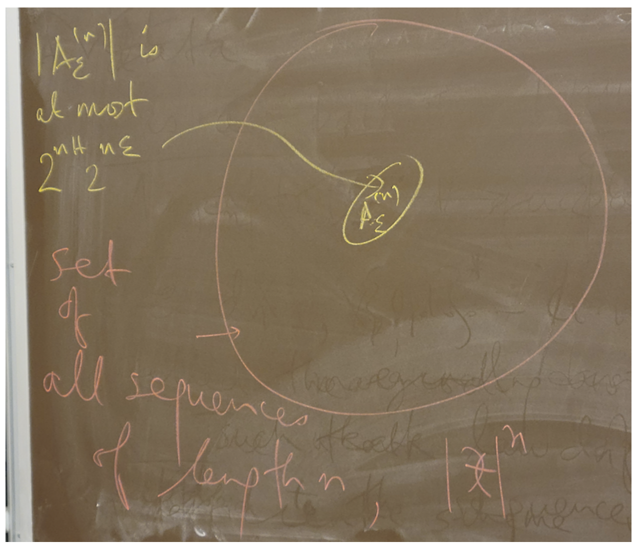
We learn that

1. For all $\varepsilon > 0$,
$$\mathbb{P}(X_1^n \in A_\varepsilon^{(n)}) \xrightarrow{n \to \infty} 1.$$

2. For all $\varepsilon > 0$, $|A_\varepsilon^{(n)}| \le 2^{nH} 2^{n\varepsilon}$ because
$$\mathbb{P}(X_1^n \in A_\varepsilon^{(n)}) = \sum_{x_1^n \in A_\varepsilon^{(n)}} p^n(x_1^n) \ge \sum_{x_1^n \in A_\varepsilon^{(n)}} 2^{-nH} 2^{-n\varepsilon} = |A_\varepsilon^{(n)}| 2^{-nH} 2^{-n\varepsilon}.$$

3. For any $\varepsilon > 0$ and $\delta > 0$, for all sufficiently large n (how large depending on (ε, δ)),
$$|A_\varepsilon^{(n)}| > (1-\delta) 2^{nH} 2^{-n\varepsilon}$$
because if n is large enough,
$$1 - \delta < \mathbb{P}(X_1^n \in A_\varepsilon^{(n)}) = \sum_{x_1^n \in A_\varepsilon^{(n)}} p^n(x_1^n) \le \sum_{x_1^n \in A_\varepsilon^{(n)}} 2^{-nH} 2^{n\varepsilon} = |A_\varepsilon^{(n)}| 2^{-nH} 2^{n\varepsilon}.$$

Together, these three statements comprise the **asymptotic equipartition property**.



1.2 Data compression

From the point of view of data compression, the AEP says that there is a data compression scheme where you assign shorter length bit strings to more commonly occurring sequences. On average, you will end up compressing the data with such a scheme.

Definition 1.1. A **lossless data compression scheme at block length** n is a pair of maps (e_n, d_n) called the **encoding** and **decoding maps**
$$e_n : \mathscr{X}^n \to \{0,1\}^* \setminus \{\varnothing\}, \qquad d_n : \{0,1\}^* \setminus \{\varnothing\} \to \mathscr{X}^n$$
(with $\{0,1\}^*$ denoting the set of binary sequences of finite length) such that $d_n \circ e_n : \mathscr{X}^n \to \mathscr{X}^n$ is the identity map.

An efficient scheme will try to minimize $\mathbb{E}[\ell(e_n(X_1^n))]$, where $\ell : \{0,1\}^* \to \mathbb{N}$ denotes the length of the string and the expectation is for $X_1, \dots, X_n \overset{\text{iid}}{\sim} (p(x), x \in \mathscr{X})$.

The AEP suggests the following scheme:

1. Use 1 bit to declare if $x_1^n \in A_\varepsilon^{(n)}$ or not.
2. If $x_1^n \in A_\varepsilon^{(n)}$, we can represent it by at most
$$\lceil \log |A_\varepsilon^{(n)}| \rceil \leq \lceil 2^{nH} 2^{n\varepsilon} \rceil \leq nH + n\varepsilon + 1$$
bits.
3. If $x_1^n \notin A_\varepsilon^{(n)}$, we can represent it by $\lceil \log |\mathscr{X}^n| \rceil \leq n \log |\mathscr{X}| + 1$ bits.

With this data compression scheme,

$$\mathbb{E}[\ell(e_n(X_1^n))] \leq 1 + \mathbb{P}(X_1^n \in A_\varepsilon^{(n)})(nH + n\varepsilon + 1) + (1 - \mathbb{P}(X_1^n \in A_\varepsilon^{(n)}))(n \log |\mathscr{X}| + 1),$$

so

$$\limsup_{n \to \infty} \frac{1}{n} \mathbb{E}[\ell(e_n(X_1^n))] \leq H(X) + \varepsilon$$

because $\mathbb{P}(X_1^n \in A_\varepsilon^{(n)}) \to 1$. This scheme is lossless, as well.

1.3 Asymptotic optimality of the AEP compression scheme

It turns out that asymptotically compressing below $H(X) - \varepsilon$ bits per symbol via a lossless scheme is impossible for any $\varepsilon > 0$. To see this, let $B_\delta^{(n)} \subseteq \mathscr{X}^n$ be any set with $\mathbb{P}(X_1^n \in B_\delta^{(n)}) \geq 1 - \delta$. Then

$$\mathbb{P}(X_1^n \in B_\delta^{(n)} \cap A_\varepsilon^{(n)}) \geq 1 - 2\delta$$

for all large enough n because $\mathbb{P}(X_1^n \in A_\varepsilon^{(n)}) > 1 - \delta$ (and using a union bound). So

$$1 - 2\delta \leq \sum_{x_1^n \in B_\delta^{(n)} \cap A_\varepsilon^{(n)}} p^n(x_1^n) \leq |B_\delta^{(n)} \cap A_\varepsilon^{(n)}| 2^{-nH} 2^{n\varepsilon}.$$

This tells us that

$$|B_\delta^{(n)} \cap A_\varepsilon^{(n)}| \geq (1 - 2\delta) 2^{nH} 2^{-n\varepsilon}$$

for all large enough n.

Suppose we have a probability distribution on a finite set giving probability $2^{-nH} 2^{n\varepsilon}$ to each of $\lfloor (1-2\delta) 2^{nH} 2^{-n\varepsilon} \rfloor$ elements of the set and giving an arbitrary distribution to the rest of the sequences. We claim that the expected length under any lossless binary encoding of such a distribution is "approximately" bounded below by $nH - n\varepsilon - 1$. To see this, consider a binary tree of depth L. The total number of nodes is $2 + 2^2 + \cdots + 2^L = 2^{L+1} - 2$. The total depth of all the nodes is

$$1 \cdot 2 + 2 \cdot 2^2 + 3 \cdot 2^3 + \cdots + L2^L = (L-1)2^{L+1} + 2.$$

So the average depth is

$$\frac{(L-1)2^{L+1} + 2}{2^{L+1} - 2} \geq L - 1$$

The precise lower bound is

$$\log\left(\lfloor (1-2\delta)2^{nH}2^{n\varepsilon} \rfloor + 2\right) - 2.$$

This is further lower bounded by

$$\log((1-2\delta)2^{nH}2^{-n\varepsilon}) - 2 = \log(1-2\delta) + n(H-\varepsilon) - 2.$$

So

$$\frac{1}{n}\text{expected depth} \geq \frac{1}{n}(\log(1-2\delta) - 2) + H - \varepsilon$$

A lossless compression scheme $\mathscr{X}^n \to \{0,1\}^* \backslash \{\varnothing\}$ must use at least this many bits/symbols because $\mathbb{P}(X_1^n \in B_\delta^{(n)} \cap A_\varepsilon^{(n)}) > 1 - 2\delta$ and each $x_1^n \in B_\delta^{(n)} \cap A_\varepsilon^{(n)}$ has $p^n(x_1^n) \leq 2^{-nH}2^{n\varepsilon}$.