EE376A/STATS376A Information Theory Lecture 4 - 01/18/2018

Lecture 4: Asymptotic Equipartition Property

Lecturer: Tsachy Weissman Scribe: Alexandros Anemogiannis, Dongyuan Mao, Mirae Parker

In this lecture, we discuss the asymptotic equipartition property (AEP) regarding the sequences output by a stochastic source. We'll find that virtually all sequences generated by the source are confined to an exponentially small subset of the set of all possible source sequences, which we use to define a near-lossless fixed length compression scheme.

1 Asymptotic Equipartition Property

1.1 Notation

We briefly describing some of the relevant terms and notations used in this section

1. **Memoryless source:** $U_1, U_2, \ldots$ iid $\sim U$. Note that "memoryless" is used here because samples are drawn iid and have no dependence on past realizations.
2. **Alphabet:** $\mathcal{U} = \{1, 2, \ldots, r\}$ specifies the possible values that each symbol U_i can take on. The size of $\mathcal{U}$ is denoted $|\mathcal{U}|$.
3. **Source sequence:** $U^n = (U_1, \ldots, U_n)$ denotes the n-tuple that specifies a sequence of n source symbols. Further note that $\mathcal{U}^n$ indicates the set of all possible source sequences of length n.
4. **Probability:** The probability assigned to a source sequence U^n is given by $P(U^n) = \prod_{i=1}^n P_{\mathcal{U}}(U_i)$. Since we implicitly evaluate the probabilities over the alphabet $\mathcal{U}$, we may also write

$$P(U^n) = \prod_{i=1}^{n} P(U_i).$$

1.2 The ϵ-typical set

Definition 1. *For some $\epsilon > 0$, the source sequence U^n is ϵ-**typical** if,*

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

Let $A_\epsilon^{(n)}$ denote the "ϵ-typical set", that is the set of all source sequences U^n that are ϵ-typical. Furthermore, note the following equivalent way of defining ϵ-typicality:

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

$$\iff -n(H(u) + \epsilon) \leq \log(P(U^n)) \leq -n(H(u) - \epsilon) \tag{2}$$

$$\iff 2^{-n(H(u)+\epsilon)} \leq P(U^n) \leq 2^{-n(H(u)-\epsilon)}. \tag{3}$$

Theorem 2. $\forall \epsilon > 0$, $P(U^n \in A_\epsilon^{(n)}) \xrightarrow{n \to \infty} 1$.

Proof Observe the following reformulation

$$P(U^n \in A_\epsilon^{(n)}) = P\left(\left|-\frac{1}{n}\log P(U^n) - H(U)\right| \leq \epsilon\right) \tag{4}$$

$$= P\left(\left|-\frac{1}{n}\log\left[\prod_{i=1}^{n} P(U_i)\right] - H(U)\right| \leq \epsilon\right) \tag{5}$$

$$= P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\log\frac{1}{P(U_i)} - H(U)\right| \leq \epsilon\right), \tag{6}$$

Noting that:
$H(U) \triangleq E\left(\log\frac{1}{P(U)}\right)$, and $\log\frac{1}{P(U_i)}$ are iid, since U_i are iid.
Then by the weak law of large numbers (LLN),

$$P(U^n \in A_\epsilon^{(n)}) = P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\log\frac{1}{P(U_i)} - H(U)\right| \leq \epsilon\right) \xrightarrow{n\to\infty} 1.$$

□

Note: Since $P(U^n \in A_\epsilon^{(n)}) \approx 1$ and $A_\epsilon^{(n)}$ is comprised of sequences each with probability roughly $2^{-nH(U)}$ of being observed, then $\left|A_\epsilon^{(n)}\right| \approx 2^{nH(U)}$. We'll provide more rigorous bounds on $\left|A_\epsilon^{(n)}\right|$ in the following theorem.

Theorem 3. *$\forall\epsilon > 0$ and n sufficiently large, $(1-\epsilon)\cdot 2^{n(H(U)-\epsilon)} \leq \left|A_\epsilon^{(n)}\right| \leq 2^{n(H(u)+\epsilon)}$.*

Proof

Upper bound:

$$\begin{aligned}
1 &\geq P(U^n \in A_\epsilon^{(n)}) \\
&\geq \sum_{u^n \in A_\epsilon^{(n)}} 2^{-n(H(U)+\epsilon)} \\
&= 2^{-n(H(U)+\epsilon)} \cdot \left|A_\epsilon^{(n)}\right| \\
&\Rightarrow \left|A_\epsilon^{(n)}\right| \leq 2^{n(H(U)+\epsilon)}
\end{aligned}$$

Lower bound:

$$\begin{aligned}
1-\epsilon &\leq P(U^n \in A_\epsilon^{(n)}) \\
&\leq \sum_{u^n \in A_\epsilon^{(n)}} 2^{-n(H(U)-\epsilon)} \\
&= 2^{-n(H(U)-\epsilon)} \cdot \left|A_\epsilon^{(n)}\right| \\
&\Rightarrow \left|A_\epsilon^{(n)}\right| \geq (1-\epsilon)\cdot 2^{n(H(U)-\epsilon)}.
\end{aligned}$$

The starting equation in the lower bound proof is a consequence of Theorem 2. Since $P(U^n \in A_\epsilon^{(n)}) \xrightarrow{n\to\infty} 1$, we can choose a sufficiently large n such that $P(U^n \in A_\epsilon^{(n)}) \geq 1-\epsilon\ \forall\epsilon > 0$. □

1.3 Some perspective

The space of all possible source sequences $\mathcal{U}^n$ has exponential size $|\mathcal{U}^n| = r^n$, and the ϵ-typical set $A_\epsilon^{(n)}$ comprises a tiny fraction of $\mathcal{U}^n$ with size $|A_\epsilon^{(n)}| \approx 2^{nH(U)}$. In fact, $A_\epsilon^{(n)}$ is an exponentially smaller than $\mathcal{U}^n$, as indicated by the ratio of their sizes, except when U is uniformly distributed.

$$\frac{\left|A_\epsilon^{(n)}\right|}{|\mathcal{U}^n|} \approx \frac{2^{nH(U)}}{r^n} = \frac{2^{nH(U)}}{2^{n\log r}} = 2^{-n(\log r - H(U))}.$$

Despite the small size of $A_\epsilon^{(n)}$, the probabilistic mass in $\mathcal{U}^n$ is almost entirely concentrated in $A_\epsilon^{(n)}$. The forthcoming theorem illustrates the point that any subset of $\mathcal{U}^n$ that's smaller than $A_\epsilon^{(n)}$ fails to capture almost all of its probabilistic mass.

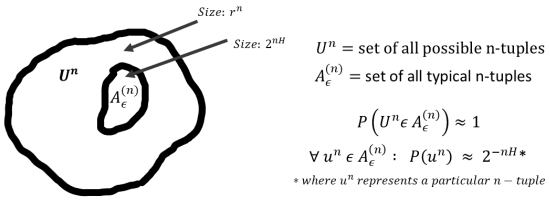


Figure 1: ϵ-typical set: almost all the probability mass is concentrated in an exponentially small set

Theorem 4. *Fix $\delta > 0$ and $B^{(n)} \subseteq \mathcal{U}^n$ such that $\left|B^{(n)}\right| \leq 2^{n(H(U)-\delta)}$. Then*

$$\lim_{n\to\infty} P(U^n \in B^{(n)}) = 0.$$

Proof

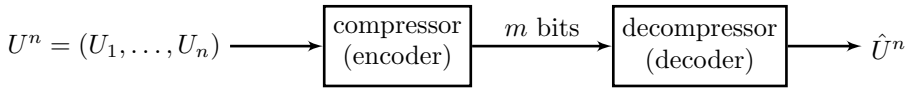
$$\begin{aligned}
P(U^n \in B^{(n)}) &= P(U^n \in B^{(n)} \cap A_\epsilon^{(n)}) + P(U^n \in B^{(n)} \cap (A_\epsilon^{(n)})^c) && (7)\\
&\leq P(U^n \in B^{(n)} \cap A_\epsilon^{(n)}) + P(U^n \notin A_\epsilon^{(n)}) && (8)\\
&= \sum_{u^n \in B^{(n)} \cap A_\epsilon^{(n)}} P(u^n) + P(U^n \notin A_\epsilon^{(n)}) && (9)\\
&\leq \sum_{u^n \in B^{(n)} \cap A_\epsilon^{(n)}} 2^{-n(H(U)-\epsilon)} + P(U^n \notin A_\epsilon^{(n)}) && (10)\\
&= \left|B^{(n)} \cap A_\epsilon^{(n)}\right| \cdot 2^{-n(H(U)-\epsilon)} + P(U^n \notin A_\epsilon^{(n)}) && (11)\\
&\leq \left|B^{(n)}\right| \cdot 2^{-n(H(U)-\epsilon)} + P(U^n \notin A_\epsilon^{(n)}) && (12)\\
&\leq 2^{n(H(U)-\delta)} \cdot 2^{-n(H(U)-\epsilon)} + P(U^n \notin A_\epsilon^{(n)}) && (13)\\
&= \underbrace{2^{-n(\delta-\epsilon)}}_{\to 0 \text{ as } n\to\infty} + \underbrace{P(U^n \notin A_\epsilon^{(n)})}_{\to 0 \text{ as } n\to\infty} && (14)
\end{aligned}$$

□

From the theorems proven above, we understand $A_\epsilon^{(n)}$ as a subset of $\mathcal{U}^n$ that most efficiently contains virtually all of the source sequences that can be drawn from $\mathcal{U}^n$. The following section confirms the intuition that, when developing a scheme which encodes sequences from $\mathcal{U}^n$, we should focus our efforts towards the sequences that lie in $A_\epsilon^{(n)}$.

2 Near-lossless (fixed-length/block) compression

In this section, we consider the problem of developing a *scheme* that allows us to encode and decode a source sequence $U^N = (U_1, \ldots, U_N)$ iid $\sim U$.



2.1 Notation

We briefly describing some of the relevant terms and notations used in this section

1. **Probability of error:** $P_e = P(\hat{U}^N \neq U^N)$ denotes the probability of error.
2. **Rate:** The rate of a scheme is the average number of bits it uses to encode source symbols.
3. **"Near lossless"** indicates that $P_e \ll 1$.
4. **"Fixed length"** or **"block"** indicates that source symbols are encoded with a fixed number of bits.
5. **"Scheme"** is a compressor (encoder) and its corresponding decompressor (decoder).

2.2 Encoder & decoder design

Theorem 5 (Direct theorem)**.** *$\forall R > H(U)$ and $\forall \delta > 0$, $\exists$ a large enough N and a scheme with rate $r < R$ and $P_e < \delta$.*

Proof Fix $\epsilon > 0$ such that $H(U) + \epsilon < R$ and enumerate the elements in $A_\epsilon^{(N)}$, where $\mathcal{I}(u^N)$ indicates index of $u^N \in A_\epsilon^{(N)}$. Then use the following encoding approach

$$\begin{cases} \text{output a m-bit representation of } \mathcal{I}(u^N), & \text{if } u^N \in A_\epsilon^{(N)} \\ \text{output an arbitrary sequence of bits,} & \text{if } u^N \notin A_\epsilon^{(N)}. \end{cases}$$

For decoding, simply let $\hat{U}^N$ be the sequence in $A_\epsilon^{(N)}$ whose index corresponds to the m received bits. Note that this scheme makes an error whenever u^N is not in $A_\epsilon^{(N)}$. Such a scheme requires $m = \log \left|A_\epsilon^{(N)}\right|$ bits per source symbol, which achieves the following rate and probability of error

$$\begin{aligned} r = \frac{m}{N} = \frac{\log \left|A_\epsilon^{(N)}\right|}{N} \leq \frac{N(H(U)+\epsilon)}{N} = H(U) + \epsilon < R \\ P_e = P(U^N \notin A_\epsilon^{(N)}) < \delta \text{ for } N \text{ sufficiently large.} \end{aligned}$$

□

Theorem 6 (Converse theorem)**.** *If $R < H(U)$ then for all sequences of schemes with rate $r \leq R$, $P_e \xrightarrow{n\to\infty} 1$.*

Proof A scheme with rate r can at most represent 2^{Nr} different sequences without error. Let's define $B^{(N)}$ as the set of these sequences. Because $r \leq R < H(u)$, $\exists \epsilon > 0$ *s.t.* $r + \epsilon < H(U)$.

$$P(U^N \in B^{(N)}) = P(U^N \in B^{(N)} \cap A_\epsilon^{(N)}) + P(U^N \in B^{(N)} \cap (A_\epsilon^{(N)})^c) \tag{15}$$

$$\tag{16}$$

By Theorem 4 we know: (17)

$$\tag{18}$$

$$P_e = 1 - P(U^N \in B^{(N)}) \xrightarrow{N\to\infty} 1 \tag{19}$$

□

Note: Drawbacks of this framework:

1. Nonzero error probability (which can be resolved by using a variable length scheme).
2. Enumerating an exponentially large set and then searching through the resulting list is incredibly inefficient.
3. Large block length N introduces significant delay in encoding and decoding.

A simple variable length scheme to solve the first problem above is to first send 1 bit to indicate whether the sequence is typical. If typical, we use the above encoding scheme. In the unlikely event that the sequence is not typical, we can just send the N bits without encoding.