Asymptotic Equipartition Property

An Analogous of the Law of Large Numbers

Radu Trîmbiţaş

October 2012

1 Asymptotic Equipartition Property

1.0.1 Introduction

Asymptotic equipartition property - Introduction

- The *law of large numbers* states that for i.i.d. RVs $\frac{1}{n}\sum X_i$ is close to $E(X)$ for large values of n.
- The *AEP* states that states that $\frac{1}{n}\log\frac{1}{p(X_1,X_2,\ldots,X_n)}$ is close to the entropy H.
- The probability $p(X_1, X_2, \ldots, X_n)$ assigned to an observed sequence will be close to 2^{-nH}.
- This enables us to divide the set of all sequences into two sets, the typical set, where the sample entropy is close to the true entropy, and the nontypical set, which contains the other sequences.

Definition 1 (Convergence of random variables)**.** Given a sequence of RVs, $X_1, X_2, \ldots$ we say that the sequences $X_1, X_2, \ldots$ *converges* to a random variable X

1. *In probability* ($X_n \xrightarrow{p} X$) if for every $\varepsilon > 0$, $P(|X_n - X| > \varepsilon) \to 0$.
2. *In mean square* if $E(X_n - X)^2 \to 0$.
3. With probability 1 (also called *almost surely* $X_n \xrightarrow{a.s.} X$) if $P(\lim_{n\to\infty} X_n = X) = 1$.

1.1 Asymptotic equipartition property theorem

Asymptotic equipartition property theorem

Theorem 2 (AEP). *If $X_1, X_2, \ldots$ are i.i.d.$\sim p(x)$, then*

$$-\frac{1}{n}\log p\left(X_1, X_2, \ldots, X_n\right) \xrightarrow{p} H(X) \tag{1}$$

Proof. X_i i.i.d $\Longrightarrow \log p(X_i)$ i.i.d, by the weak law of large numbers

$$\begin{aligned} -\frac{1}{n}\log p\left(X_1, X_2, \ldots, X_n\right) &= -\frac{1}{n}\sum_i \log p(X_i) \\ &\xrightarrow{p} -E(\log p(X) \\ &= H(X). \end{aligned}$$

□

Definition 3. The *typical set* $A_\varepsilon^{(n)}$ with respect to $p(x)$ is the set of sequences $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ with the property

$$2^{-n(H(X)+\varepsilon)} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-n(H(X)-\varepsilon)}. \tag{2}$$

Theorem 4. *(1) If $(x_1, x_2, \ldots, x_n) \in A_\varepsilon^{(n)}$, then*

$$H(X) - \varepsilon \leq -\frac{1}{n}\log p(x_1, x_2, \ldots, x_n) \leq H(X) + \varepsilon.$$

(2) $P\left(A_\varepsilon^{(n)}\right) > 1 - \varepsilon$ for n sufficiently large.

(3) $\left|A_\varepsilon^{(n)}\right| \leq 2^{n(H(X)+\varepsilon)}$, where $|A|$ denotes the cardinal of A.

(4) $\left|A_\varepsilon^{(n)}\right| \geq (1-\varepsilon)\, 2^{n(H(X)-\varepsilon)}$ for n sufficiently large.

Thus, the typical set has probability nearly 1, all elements of the typical set are nearly equiprobable, and the number of elements in the typical set is nearly 2^{nH}.

Proof. (1) from definition of $A_\varepsilon^{(n)}$.

(2) follows from Theorem 2, since $P((X_1, X_2, \ldots, X_n) \in A_\varepsilon^{(n)} \to 1$ as $n \to \infty$. Thus, $\forall \delta > 0 \exists n_0$ such that for all $n \geq n_0$

$$P\left(\left|-\frac{1}{n}\log p\left(X_1, X_2, \ldots, X_n\right) - H(X)\right| < \varepsilon\right) > 1 - \delta.$$

Setting $\delta = \varepsilon$ complets the proof of (2).

□

Proof - continuation. (3)

$$1 = \sum_{x \in \mathcal{X}^n} p(x) \geq \sum_{x \in A_\varepsilon^{(n)}} p(x) \geq \sum_{x \in A_\varepsilon^{(n)}} 2^{-n(H(X)+\varepsilon)} = 2^{-n(H(X)+\varepsilon)} \left| A_\varepsilon^{(n)} \right|$$

$$\text{Hence } \left| A_\varepsilon^{(n)} \right| \leq 2^{n(H(X)+\varepsilon)}.$$

(4) For sufficiently large n, $P\left(A_\varepsilon^{(n)}\right) > 1 - \varepsilon$, so that

$$1 - \varepsilon < P\left(A_\varepsilon^{(n)}\right) \leq \sum_{x \in A_\varepsilon^{(n)}} 2^{-n(H(X)-\varepsilon)} = 2^{-n(H(X)-\varepsilon)} \left| A_\varepsilon^{(n)} \right|$$

which yields to

$$\left| A_\varepsilon^{(n)} \right| \geq (1 - \varepsilon)\, 2^{n(H(X)-\varepsilon)}.$$

□

1.2 Consequences of the AEP: Data compression

Consequences of the AEP: Data compression

- $X_1, X_2, \ldots, X_n \sim p(x)$, i.i.d
- We wish short descriptions of such sequences
- We divide sequences in $\mathcal{X}^n$ in two sets: the typical set $A_\varepsilon^{(n)}$ and its complement (Figure 1)
- All elements in each set are ordered (e.g. lexicografic order)
- We represent each sequence of $A_\varepsilon^{(n)}$ by its index of the sequence in the set.
- Since there are $\leq 2^{n(H+\varepsilon)}$ sequences in $A_\varepsilon^{(n)}$, the indexing requires $\leq n(H+\varepsilon)+1$ bits
- We prefix each sequence by 0: $\leq n(H+\varepsilon)+2$ bits required (Figure 2)
- Similarly, for the complement of $A_\varepsilon^{(n)}$ we need $n \log |\mathcal{X}| + 1$; we prefix each sequence by 1
- Features of the coding scheme
  - one-to-one and easy decodadle; the initial bit acts as a flag to indicate the length of the codeword
  - For complement of $A_\varepsilon^{(n)}$ we have a brute-force enumeration, without taking into account that the number of sequences in $\left(A_\varepsilon^{(n)}\right)^C \leq |\mathcal{X}|^n$

– Typical sequences have short descriptions of length$\approx nH$

- Notations x^n - a sequence $x_1, x_2, \ldots, x_n$; $\ell(x^n)$ is the length of the codeword corresponding to x^n

Theorem 5. *Let X^n i.i.d $\sim p(x)$, and $\varepsilon > 0$. The there exists a code that maps sequences x^n of length n into binary strings such that the mappings is one-to-one (and therefore invertible) and*

$$E\left[\frac{1}{n}\ell\left(X^n\right)\right] \leq H(X) + \varepsilon \tag{3}$$

for n sufficiently large.

We can represent sequences X^n using $nH(X)$ bits on the average.

Proof. If n is sufficiently large so that $P\left(A_\varepsilon^{(n)}\right) \geq 1 - \varepsilon$, the expected length of codeword is

$$\begin{aligned}
E\left(\ell\left(X^n\right)\right) &= \sum_{x^n} p(x^n)\ell(x^n) \\
&= \sum_{x^n \in A_\varepsilon^{(n)}} p(x^n)\ell(x^n) + \sum_{x^n \in \left(A_\varepsilon^{(n)}\right)^C} p(x^n)\ell(x^n) \\
&\leq \sum_{x^n \in A_\varepsilon^{(n)}} p(x^n)\left(n\left(H+\varepsilon\right)+2\right) + \sum_{x^n \in \left(A_\varepsilon^{(n)}\right)^C} p(x^n)\left(n\log|\mathcal{X}|+2\right) \\
&= P\left(A_\varepsilon^{(n)}\right)\left(n\left(H+\varepsilon\right)+2\right) + P\left(\left(A_\varepsilon^{(n)}\right)^C\right)\left(n\log|\mathcal{X}|+2\right) \\
&\leq n\left(H+\varepsilon\right) + \varepsilon n\log|\mathcal{X}| + 2 = n\left(H+\varepsilon'\right)
\end{aligned}$$

But, $\varepsilon' = \varepsilon + \varepsilon\log|\mathcal{X}| + \frac{2}{n}$ can be made arbitrarily small. □

Typical sets and source coding

Source coding using the typical sets

High-probability sets and the typical sets

It is not clear that $A_\varepsilon^{(n)}$ is *the smallest set* that contains most of the probability.

Definition 6. For each $n = 1, 2, \ldots$, let $B_\delta^{(n)} \subset \mathcal{X}^n$ with

$$P\left(B_\delta^{(n)}\right) \geq 1 - \delta. \tag{4}$$

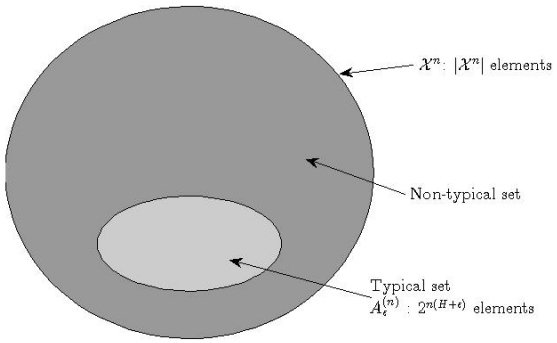

Figure 1: Typical sets and source coding

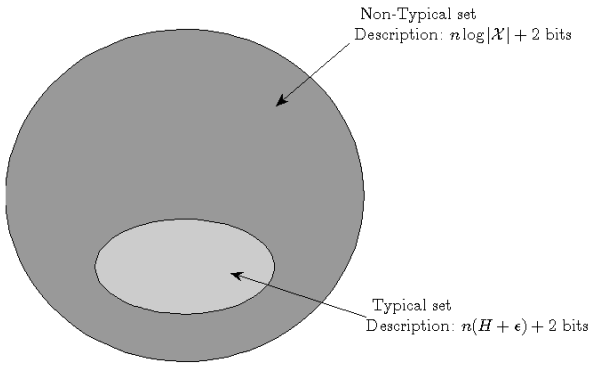

Figure 2: Source code using the typical set

Theorem 7. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim p(x)$. For $\delta < \frac{1}{2}$ and any $\delta' > 0$, if $P\left(B_\delta^{(n)}\right) > 1 - \delta$, then*

$$\frac{1}{n} \log \left| B_\delta^{(n)} \right| > H - \delta' \tag{5}$$

for n sufficiently large.

Definition 8. The notation $a_n \doteq b_n$ means

$$\lim_{n \to \infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0.$$

Thus, $a_n \doteq b_n$ implies a_n and b_n are equal to the first order in the exponent. We can restate the above theorem: if $\delta_n \to 0$ and $\varepsilon_n \to 0$, then

$$\left| B_{\delta_n}^{(n)} \right| \doteq \left| A_{\varepsilon_n}^{(n)} \right| \doteq 2^{nH}.$$

References

References

[1] Thomas M. Cover, Joy A. Thomas, *Elements of Information Theory*, 2nd edition, Wiley, 2006.

[2] David J.C. MacKay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 2003.

[3] Robert M. Gray, *Entropy and Information Theory*, Springer, 2009