Typicality and the Source Coding Theorem

Uri Shaham

1 Asymptotic Equipartition Property

1.1 Motivation

Lecture 1 showed that a single symbol can be encoded with expected length between $H(X)$ and $H(X)+1$. The extra 1 bit can be significant for a single symbol, but it disappears when we encode long blocks.

Typicality explains why. For a long IID sequence $X^n = (X_1, \ldots, X_n)$, most probability mass lies on a set of about $2^{nH(X)}$ typical sequences. Therefore roughly $nH(X)$ bits are enough to name the likely sequence.

1.2 IID sources and entropy rate

Let $X_1, X_2, \ldots$ be independent copies of $X \sim p$. Then

$$p(x^n) = \prod_{i=1}^{n} p(x_i),$$

and by the chain rule,

$$H(X^n) = H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i) = nH(X).$$

The quantity $H(X)$ is therefore the entropy per symbol for an IID source.

1.3 Asymptotic Equipartition Property (AEP)

Theorem 1.1 (Asymptotic Equipartition Property)**.** *Let $X_1, X_2, \ldots$ be i.i.d. random variables taking values in a finite alphabet $\mathcal{X}$, with distribution $p(x)$. Let*

$$X^n = (X_1, \ldots, X_n).$$

Then

$$-\frac{1}{n} \log_2 p(X^n) \longrightarrow H(X).$$

Equivalently, for large n, most sequences x^n generated by the source have probability approximately

$$p(x^n) \approx 2^{-nH(X)}.$$

Proof. Since the variables $X_1, \ldots, X_n$ are i.i.d.,

$$p(X^n) = \prod_{i=1}^{n} p(X_i).$$

Taking logarithms,

$$-\frac{1}{n} \log_2 p(X^n) = \frac{1}{n} \sum_{i=1}^{n} \log_2 \frac{1}{p(X_i)}.$$

Now define

$$Z_i = \log_2 \frac{1}{p(X_i)}.$$

Since $X_1, X_2, \ldots$ are i.i.d., the random variables $Z_1, Z_2, \ldots$ are also i.i.d. Moreover,

$$\mathbb{E}[Z_i] = \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{1}{p(x)} = H(X).$$

Therefore, by the strong law of large numbers,

$$\frac{1}{n} \sum_{i=1}^{n} Z_i \longrightarrow \mathbb{E}[Z_1] = H(X)$$

almost surely.

But

$$\frac{1}{n} \sum_{i=1}^{n} Z_i = -\frac{1}{n} \log_2 p(X^n).$$

Hence

$$-\frac{1}{n} \log_2 p(X^n) \longrightarrow H(X)$$

almost surely.

□

The AEP says that a random sequence generated by the source has probability approximately

$$p(X^n) \approx 2^{-nH(X)}.$$

This does not mean all sequences have this probability. It means that the sequences that actually occur with high probability have roughly this probability.

2 Typical sets

Definition 2.1 (Typical set)**.** For $\varepsilon > 0$, the typical set is

$$\mathcal{T}_\varepsilon^{(n)} = \left\{ x^n : \left| -\frac{1}{n} \log p(x^n) - H(X) \right| \leq \varepsilon \right\}.$$

If $x^n \in \mathcal{T}_\varepsilon^{(n)}$, then

$$2^{-n(H(X)+\varepsilon)} \leq p(x^n) \leq 2^{-n(H(X)-\varepsilon)}.$$

Corollary 2.2 (Typical-set form of the AEP)**.** *For every $\varepsilon > 0$,*

$$\mathbb{P}\left(X^n \in A_\varepsilon^{(n)}\right) \to 1.$$

Moreover, for sufficiently large n, the typical set has size approximately

$$|A_\varepsilon^{(n)}| \approx 2^{nH(X)}.$$

More precisely,

$$|A_\varepsilon^{(n)}| \le 2^{n(H(X)+\varepsilon)}$$

and, for large n,

$$|A_\varepsilon^{(n)}| \ge (1-\varepsilon)2^{n(H(X)-\varepsilon)}.$$

Proof. By the AEP,

$$-\frac{1}{n}\log_2 p(X^n) \longrightarrow H(X)$$

in probability. Therefore,

$$\mathbb{P}\left(\left|-\frac{1}{n}\log_2 p(X^n) - H(X)\right| \le \varepsilon\right) \to 1.$$

This is exactly

$$\mathbb{P}\left(X^n \in A_\varepsilon^{(n)}\right) \to 1.$$

Now suppose $x^n \in A_\varepsilon^{(n)}$. Then

$$p(x^n) \ge 2^{-n(H(X)+\varepsilon)}.$$

Since total probability is at most 1,

$$1 \ge \sum_{x^n \in A_\varepsilon^{(n)}} p(x^n) \ge |A_\varepsilon^{(n)}|\, 2^{-n(H(X)+\varepsilon)}.$$

Thus

$$|A_\varepsilon^{(n)}| \le 2^{n(H(X)+\varepsilon)}.$$

For the lower bound, for large n we have

$$\mathbb{P}\left(X^n \in A_\varepsilon^{(n)}\right) \ge 1-\varepsilon.$$

Also, for every $x^n \in A_\varepsilon^{(n)}$,

$$p(x^n) \le 2^{-n(H(X)-\varepsilon)}.$$

Therefore

$$1-\varepsilon \le \sum_{x^n \in A_\varepsilon^{(n)}} p(x^n) \le |A_\varepsilon^{(n)}|\, 2^{-n(H(X)-\varepsilon)}.$$

Hence

$$|A_\varepsilon^{(n)}| \ge (1-\varepsilon)2^{n(H(X)-\varepsilon)}.$$

□

3 Source coding for long blocks

The source coding theorem says that $H(X)$ is the sharp asymptotic compression limit.

Theorem 3.1 (Lossless source coding theorem, informal)**.** *Let $X_1, X_2, \ldots$ be IID with entropy $H(X)$. For every $\delta > 0$ and sufficiently large block length n, there are codes with rate less than $H(X) + \delta$ bits per source symbol whose decoding error probability is arbitrarily small. Conversely, any sequence of codes with rate less than $H(X) - \delta$ has decoding error probability bounded away from zero.*

Here "rate" means number of encoded bits per source symbol.

3.1 Achievability via typicality

Fix $\varepsilon > 0$. For large n, the typical set has size at most $2^{n(H+\varepsilon)}$. A compressor can do the following:

1. If x^n is typical, send its index in the typical set using about $n(H + \varepsilon)$ bits.
2. If x^n is atypical, declare an error or use a longer escape code.

The probability of the atypical event goes to zero by the AEP. Hence a rate just above $H(X)$ is achievable with vanishing error probability.

For strictly lossless coding with no error, one can add an escape symbol and a raw encoding for atypical sequences. The expected rate still approaches $H(X)$.

3.2 Converse intuition

Suppose a compressor uses fewer than $n(H(X) - \delta)$ bits. It can assign distinct binary strings to at most

$$2^{n(H(X)-\delta)}$$

source sequences. But the typical set has roughly $2^{nH(X)}$ sequences, each with comparable probability. The compressor cannot uniquely name enough typical sequences. Therefore a non-negligible error probability is unavoidable.

The proof is a counting argument: high probability mass is spread across exponentially many typical sequences.

3.3 Method of Types: A More Concrete View of Typicality

For finite alphabets, typicality can be described very explicitly using empirical distributions. The main idea is to group sequences according to how often each symbol appears.

Definition 3.2 (Type)**.** Let $x^n = (x_1, \ldots, x_n) \in \mathcal{X}^n$. The *type* of x^n is its empirical distribution

$$\hat{p}_{x^n}(a) = \frac{1}{n} \left| \{ i : x_i = a \} \right|, \qquad a \in \mathcal{X}.$$

Thus, $\hat{p}_{x^n}(a)$ is the fraction of times the symbol a appears in the sequence.

For example, if $\mathcal{X} = \{0, 1\}$ and

$$x^n = 00101,$$

then

$$\hat{p}_{x^n}(0) = \frac{3}{5}, \qquad \hat{p}_{x^n}(1) = \frac{2}{5}.$$

The usefulness of types comes from the following observation: under an IID source, the probability of a sequence depends only on its type, not on the order of its symbols.

Indeed, suppose x^n has type q. This means that symbol a appears exactly $nq(a)$ times. Therefore

$$p(x^n) = \prod_{i=1}^{n} p(x_i) = \prod_{a \in \mathcal{X}} p(a)^{nq(a)}.$$

Taking logarithms, we get

$$\log_2 p(x^n) = n \sum_{a \in \mathcal{X}} q(a) \log_2 p(a).$$

Hence

$$p(x^n) = 2^{-n \sum_{a \in \mathcal{X}} q(a) \log_2 \frac{1}{p(a)}}.$$

Now we rewrite the exponent in a more meaningful way. Using

$$H(q) = \sum_{a \in \mathcal{X}} q(a) \log_2 \frac{1}{q(a)}$$

and

$$D_{\mathrm{KL}}(q \| p) = \sum_{a \in \mathcal{X}} q(a) \log_2 \frac{q(a)}{p(a)}.$$

Adding these two quantities gives

$$H(q) + D_{\mathrm{KL}}(q \| p) = \sum_{a \in \mathcal{X}} q(a) \log_2 \frac{1}{p(a)}.$$

Therefore, if x^n has type q, then

$$p(x^n) = 2^{-n(H(q) + D_{\mathrm{KL}}(q \| p))}.$$

This formula separates two effects.

First, $H(q)$ measures how many sequences have empirical distribution q. Indeed, the number of sequences of type q is approximately

$$2^{nH(q)}.$$

This is because high-entropy empirical distributions can be arranged in many ways, while low-entropy empirical distributions can be arranged in fewer ways.

Second, $D_{\mathrm{KL}}(q \| p)$ measures how different the empirical distribution q is from the true distribution p. Since

$$D_{\mathrm{KL}}(q \| p) \geq 0,$$

with equality if and only if $q = p$, empirical distributions far from p are exponentially unlikely.

Putting these two facts together, the total probability of all sequences of type q is approximately

$$\underbrace{2^{nH(q)}}_{\text{number of sequences of type } q} \cdot \underbrace{2^{-n(H(q)+D_{\mathrm{KL}}(q\|p))}}_{\text{probability of each such sequence}}.$$

The factors involving $H(q)$ cancel, leaving

$$\mathbb{P}(\text{type} = q) \approx 2^{-nD_{\mathrm{KL}}(q\|p)}.$$

This is the key message of the method of types:

$$\boxed{\mathbb{P}(\hat{p}_{X^n} \approx q) \approx 2^{-nD_{\mathrm{KL}}(q\|p)}.}$$

Thus, empirical distributions close to the true distribution p are likely, because $D_{\mathrm{KL}}(q\|p)$ is small. Empirical distributions far from p are exponentially unlikely, because $D_{\mathrm{KL}}(q\|p)$ is large.

This gives a more concrete interpretation of typicality. A typical sequence is one whose empirical distribution is close to the true distribution p. For such sequences, $q \approx p$, and therefore

$$D(q\|p) \approx 0, \qquad H(q) \approx H(p).$$

Consequently,

$$p(x^n) \approx 2^{-nH(p)},$$

which is exactly the statement of the AEP.