Fixed-Rate Source Coding and the AEP — Lecture 5

Haruka Eshima

September 2026

Abstract

This is a lecture note that I took for the Information Theory class taught by Prof. Amedeo Roberto Esposito at the Okinawa Institute of Science and Technology, 2026. Any errors are my own.
The questions in green are made by me and are not officially related to the mini test in the class.

Conventions. log = $\log_2$ throughout; entropy is in bits. $\mathcal{X}$ is a finite alphabet with $|\mathcal{X}| = r$, and $X_1, X_2, \ldots \overset{\text{iid}}{\sim} P_X$ is a *discrete memoryless source* (DMS). We write $X^n = (X_1, \ldots, X_n)$ and $P_{X^n}(x^n) = \prod_{i=1}^n P_X(x_i)$. Sums over x run over $\text{supp}(P_X) = \{x : P_X(x) > 0\}$.

Contents

Summary 1

1 Information density 2

2 The Asymptotic Equipartition Property 2

3 The typical set 3

3.1 High-probability sets must be large 4

4 Variable-length coding of a block 4

5 Fixed-rate codes 5

6 The converses 6

Summary

- Asymptotic Equipartition Property (AEP).
- The (weakly) typical set, and its three substantive properties.
- Fixed-rate codes.
- Error probability $\varepsilon < 1 \implies$ rate of a fixed-rate code $\geq H(P)$ (*weak converse*).
- Rate of a fixed-rate code $< H(P) \implies$ error probability $\to 1$ (*strong converse*).

1 Information density

Definition 1.1 (Information density)**.** For $x \in \operatorname{supp}(P_X)$,

$$I(x) := -\log P_X(x),$$

the *information density* (also written $i(x)$); it measures "surprise". By definition,

$$H(X) = \mathbb{E}_{P_X}\big[I(X)\big].$$

Example 1.2. Let $X \sim P_X$ on $\{0,1\}$ with $P_X(0) = 0.3$, $P_X(1) = 0.7$. Then

$$I(x) = \begin{cases} -\log_2 0.3 \approx 1.737, & x = 0, \\ -\log_2 0.7 \approx 0.515, & x = 1, \end{cases} \qquad H(X) = 0.3(1.737) + 0.7(0.515) \approx 0.881.$$

Example 1.3 (Why "typical" is not "most likely")**.** Continue Example 1.2 with $n = 10$ and compare

$$\bar{x}_1 = 1101001111 \quad (7 \text{ ones, } 3 \text{ zeros}), \qquad \bar{x}_2 = 1111111111 \quad (10 \text{ ones}).$$

Individually,

$$\mathbb{P}(X^{10} = \bar{x}_2) = 0.7^{10} \approx 0.0282 \;>\; \mathbb{P}(X^{10} = \bar{x}_1) = 0.7^7 0.3^3 \approx 0.00222,$$

so $\bar{x}_2$ is the single most likely sequence. However, as *events*,

$$\mathbb{P}\big(X^{10} \text{ has 10 ones}\big) \approx 0.0282 \;\ll\; \mathbb{P}\big(X^{10} \text{ has 7 ones, 3 zeros}\big) = \binom{10}{3} 0.7^7 0.3^3 \approx 0.267.$$

A *typical* x^n has about $nP_X(1)$ ones and $nP_X(0)$ zeros, and for one such fixed sequence

$$\mathbb{P}(X^n = x^n) = P_X(1)^{nP_X(1)} P_X(0)^{nP_X(0)} = 2^{\,n\left(P_X(1)\log P_X(1) + P_X(0)\log P_X(0)\right)} = 2^{-nH(X)}.$$

Remark 1.4. $\bar{x}_2$ is the most probable *sequence*. However, the typical ones, like $\bar{x}_1$, are overwhelmingly more probable *as a class*, because there are exponentially many of them. A good code should spend its short codewords on the typical class, not on the single most likely string.

2 The Asymptotic Equipartition Property

The key observation is that the normalised information density of a block is a sample average:

$$\frac{I(X^n)}{n} = -\frac{1}{n}\log P_{X^n}(X^n) = -\frac{1}{n}\sum_{i=1}^{n} \log P_X(X_i). \tag{1}$$

Q. *State and prove the asymptotic equipartition property.* (Hint: WLLN.)

Theorem 2.1 (AEP)**.** *If $X_1, \ldots, X_n \overset{iid}{\sim} P_X$ on a finite alphabet, then*

$$-\frac{1}{n}\log \underbrace{P_{X^n}(X^n)}_{\text{a random variable}} \xrightarrow{\;\mathbb{P}\;} H(X) \qquad \text{as } n \to \infty.$$

Proof. Set $Z_i := -\log P_X(X_i)$. The Z_i are i.i.d., and

$$\mathbb{E}[Z_1] = \sum_x P_X(x)\big(-\log P_X(x)\big) = H(X).$$

Because $\mathcal{X}$ is finite, $P_X(x) \geq p_{\min} := \min_{x \in \operatorname{supp} P_X} P_X(x) > 0$ on the support, so $0 \leq Z_i \leq -\log p_{\min} < \infty$; in particular $\mathbb{E}[Z_1^2] < \infty$. By (1) and the weak law of large numbers,

$$-\frac{1}{n}\log P_{X^n}(X^n) = \frac{1}{n}\sum_{i=1}^{n} Z_i \xrightarrow{\mathbb{P}} \mathbb{E}[Z_1] = H(X). \qquad \square$$

3 The typical set

Definition 3.1 ((Weakly) typical set)**.** Fix $\varepsilon > 0$ and $n \in \mathbb{N}$. Given $X \sim P_X$, the *typical set* is

$$A_\varepsilon^n(P_X) := \Big\{x^n \in \mathcal{X}^n \;:\; H(X) - \varepsilon \leq -\tfrac{1}{n}\log P_{X^n}(x^n) \leq H(X) + \varepsilon\Big\}$$

or, equivalently,

$$A_\varepsilon^n(P_X) = \Big\{x^n \;:\; 2^{-n(H(X)+\varepsilon)} \leq P_{X^n}(x^n) \leq 2^{-n(H(X)-\varepsilon)}\Big\}.$$

Proposition 3.2 (Properties of the typical set)**.** *Fix $\varepsilon > 0$ and P_X.*

1) For all n and all $x^n \in A_\varepsilon^n(P_X)$, $H(X) - \varepsilon \leq -\frac{1}{n}\log P_{X^n}(x^n) \leq H(X) + \varepsilon$.

2) If $X_1^n \overset{iid}{\sim} P_X$, then for every $\delta > 0$ there is n large enough that $\mathbb{P}\big(A_\varepsilon^n(P_X)\big) > 1 - \delta$.

3) For all n, $\big|A_\varepsilon^n\big| \leq 2^{n(H(X)+\varepsilon)}$.

4) For every $\delta > 0$ there is n large enough that $\big|A_\varepsilon^n\big| \geq (1-\delta)\,2^{n(H(X)-\varepsilon)}$.

Q. *Prove 1)–4).*

Proof. **1)** This is Definition 3.1.

2) By the AEP (Theorem 2.1), for every $\delta > 0$ and $\varepsilon > 0$ there is $N_{\delta,\varepsilon}$ such that for all $n > N_{\delta,\varepsilon}$,

$$\mathbb{P}\Big(\Big|-\tfrac{1}{n}\log P_{X^n}(X^n) - H(X)\Big| \leq \varepsilon\Big) > 1 - \delta,$$

and the event inside is exactly $\{X^n \in A_\varepsilon^n\}$.

3) Every $x^n \in A_\varepsilon^n$ has $P_{X^n}(x^n) \geq 2^{-n(H(X)+\varepsilon)}$, so

$$1 \;\geq\; \mathbb{P}\big(A_\varepsilon^n\big) = \sum_{x^n \in A_\varepsilon^n} P_{X^n}(x^n) \;\geq\; \big|A_\varepsilon^n\big| \cdot 2^{-n(H(X)+\varepsilon)},$$

which rearranges to $|A_\varepsilon^n| \leq 2^{n(H(X)+\varepsilon)}$.

4) Every $x^n \in A_\varepsilon^n$ has $P_{X^n}(x^n) \leq 2^{-n(H(X)-\varepsilon)}$, so for n large enough that 2) applies,

$$1 - \delta \;\leq\; \mathbb{P}\big(A_\varepsilon^n\big) \;\leq\; \big|A_\varepsilon^n\big| \cdot 2^{-n(H(X)-\varepsilon)},$$

i.e. $|A_\varepsilon^n| \geq (1-\delta)2^{n(H(X)-\varepsilon)}$. $\square$

Remark 3.3 (Exponentially small, yet almost certain)**.** Combining 3) and 4) with $|\mathcal{X}^n| = r^n = 2^{n \log r}$,

$$\frac{|A_\varepsilon^n|}{|\mathcal{X}^n|} \approx 2^{-n(\log r - H(X))} \longrightarrow 0 \qquad \text{whenever } H(X) < \log r,$$

while $\mathbb{P}(A_\varepsilon^n) \to 1$. **The typical set is an exponentially vanishing fraction of all sequences that nonetheless captures essentially all of the probability. This is precisely the room that fixed-rate compression exploits.**

3.1 High-probability sets must be large

Q. *State and prove that a high-probability set has to be large when n is large.* (Hint: $\mathbb{P}(A_\varepsilon^n \cap B^n)$ needs to be large.)

Lemma 3.4 (High-probability sets are big)**.** *Fix $\varepsilon > 0$ and $\delta \in (0,1)$, and let $B^n \subseteq \mathcal{X}^n$ satisfy $\mathbb{P}(B^n) \geq 1 - \delta$. Then for every $\eta > 0$ with $\delta + \eta < 1$ there is N such that for all $n \geq N$,*

$$|B^n| \geq (1 - \delta - \eta)\, 2^{n(H(X)-\varepsilon)}.$$

In particular, taking $\eta = \frac{1-\delta}{2}$,

$$|B^n| \geq \frac{1-\delta}{2}\, 2^{n(H(X)-\varepsilon)} \qquad \textit{for all large } n.$$

Proof. By Proposition 3.2.2 applied with η in place of δ, choose N so that $\mathbb{P}(A_\varepsilon^n) \geq 1 - \eta$ for all $n \geq N$. Then by inclusion exclusion and $\mathbb{P}(A_\varepsilon^n \cup B^n) \leq 1$,

$$\mathbb{P}(A_\varepsilon^n \cap B^n) = \mathbb{P}(A_\varepsilon^n) + \mathbb{P}(B^n) - \mathbb{P}(A_\varepsilon^n \cup B^n) \geq (1 - \eta) + (1 - \delta) - 1 = 1 - \delta - \eta.$$

On the other hand, every $x^n \in A_\varepsilon^n$ has $P_{X^n}(x^n) \leq 2^{-n(H(X)-\varepsilon)}$, so

$$\mathbb{P}(A_\varepsilon^n \cap B^n) \leq 2^{-n(H(X)-\varepsilon)} |A_\varepsilon^n \cap B^n| \leq 2^{-n(H(X)-\varepsilon)} |B^n|.$$

Combining the two displays gives $(1 - \delta - \eta) \leq 2^{-n(H(X)-\varepsilon)} |B^n|$, which is the claim. □

4 Variable-length coding of a block

Let $X \in \mathcal{X}$ with $|\mathcal{X}| = r$, and let $X_1^n \overset{\text{iid}}{\sim} P_X$ be a DMS.

Q. *Prove that for every $\eta > 0$, for n large enough there is a prefix-free code $\mathscr{C}_n$ on $\mathcal{X}^n$ with $H(X) \leq \frac{1}{n}\mathbb{E}[\ell_{\mathscr{C}_n}(X^n)] \leq H(X) + \eta$.*

Construction (via the typical set)

Fix $n \in \mathbb{N}$ and $\varepsilon > 0$.

- Give distinct codewords of a common length ℓ_1 to all sequences in $A_\varepsilon^n(P_X)$, where

  $$\ell_1 := \lceil \log |A_\varepsilon^n| \rceil < \log |A_\varepsilon^n| + 1 \leq n(H(X) + \varepsilon) + 1,$$

  the last step by Proposition 3.2.3.

- Give distinct codewords of a common length ℓ_2 to all sequences *outside* $A_\varepsilon^n(P_X)$, where

$$\ell_2 := \lceil \log |\mathcal{X}^n| \rceil < n \log_2 r + 1.$$

- Finally set

$$\mathscr{C}_n(x^n) = \begin{cases} 0 \cdot (\text{its length-}\ell_1 \text{ codeword}), & x^n \in A_\varepsilon^n, \\ 1 \cdot (\text{its length-}\ell_2 \text{ codeword}), & x^n \notin A_\varepsilon^n, \end{cases}$$

  where $\cdot$ denotes concatenation.

Claim 4.1. *$\mathscr{C}_n$ is prefix-free (as a code on the alphabet $\mathcal{X}^n$).*

Proof. Two codewords with the same first bit have the same total length ($1 + \ell_1$, or $1 + \ell_2$) and are distinct, so neither is a prefix of the other. Two codewords with different first bits disagree in position 1, so again neither is a prefix of the other. □

Upper bound on the expected length

Every typical x^n receives $1 + \ell_1 < n(H(X) + \varepsilon) + 2$ bits and every atypical one receives $1 + \ell_2 < n \log_2 r + 2$ bits. Hence

$$\begin{aligned} \frac{1}{n}\mathbb{E}\big[\ell_{\mathscr{C}_n}(X^n)\big] &< \frac{1}{n}\left(\sum_{x^n \in A_\varepsilon^n} P_{X^n}(x^n)\big(n(H(X)+\varepsilon)+2\big) + \sum_{x^n \notin A_\varepsilon^n} P_{X^n}(x^n)\big(n\log_2 r + 2\big) \right) \\ &= \Big(H(X) + \varepsilon + \tfrac{2}{n}\Big)\mathbb{P}\big(A_\varepsilon^n\big) + \Big(\log_2 r + \tfrac{2}{n}\Big) \underbrace{\big(1 - \mathbb{P}(A_\varepsilon^n)\big)}_{\leq \varepsilon \text{ for large } n} \\ &\leq H(X) + \underbrace{\varepsilon + \tfrac{2}{n} + \varepsilon\Big(\log_2 r + \tfrac{2}{n}\Big)}_{=:\,\eta_{n,\varepsilon}}. \end{aligned}$$

Choosing ε small and then n large makes $\eta_{n,\varepsilon}$ arbitrarily small.

Lower bound on the expected length

By Shannon's noiseless coding theorem (lower bound) applied to the random variable X^n on the alphabet $\mathcal{X}^n$, any prefix-free $\mathscr{C}_n$ satisfies $\mathbb{E}[\ell_{\mathscr{C}_n}(X^n)] \geq H(X^n)$; and $H(X^n) = nH(X)$ by independence. Hence

$$\frac{1}{n}\mathbb{E}\big[\ell_{\mathscr{C}_n}(X^n)\big] \;\geq\; H(X).$$

Proposition 4.2. *For every $\eta > 0$, for n large enough there is a prefix-free code $\mathscr{C}_n$ on $\mathcal{X}^n$ with*

$$H(X) \;\leq\; \frac{1}{n}\mathbb{E}\big[\ell_{\mathscr{C}_n}(X^n)\big] \;\leq\; H(X) + \eta.$$

5 Fixed-rate codes

Now we allow errors outside a specific set B^n.

Definition 5.1 (Fixed-rate code, error probability, rate)**.** A fixed-rate code of block length n is a pair (encoder, decoder) φ_n, ψ_n. Let

$$B^n := \{x^n \in \mathcal{X}^n \ : \ \psi_n(\varphi_n(x^n)) = x^n\}$$

be the set of correctly reproduced sequences. The *error probability* is

$$P_e^n := 1 - \mathbb{P}(B^n),$$

and the *rate* (bits per source symbol) of the code with codebook B^n is

$$R_n := \frac{1}{n}\lceil \log |B^n| \rceil.$$

If $\mathbb{P}(B^n) > 1 - \varepsilon$ then $P_e^n < \varepsilon$, and by Lemma 3.4 (with $\delta = \eta = \varepsilon$),

$$R_n \ \approx \ \frac{1}{n}\log |B^n| \ \geq \ \frac{1}{n}\log\Big((1-2\varepsilon)2^{n(H(X)-\varepsilon)}\Big) \ = \ H(X) - \varepsilon + \frac{1}{n}\log(1-2\varepsilon).$$

6 The converses

Q. *State and prove a lower bound for the rate of fixed-rate coding with error probability $\leq \varepsilon < 1$ for all n.*

Theorem 6.1 (Weak converse)**.** *Let X be a DMS with pmf P_X, and let $(\varphi_n, \psi_n)_{n\geq 1}$ be fixed-rate codes whose decoding error probability satisfies $P_e^n \leq \varepsilon$ for some fixed $\varepsilon < 1$ and all large n. Then*

$$\liminf_{n\to\infty} R_n \ \geq \ H(X).$$

Proof. Let B^n be the codebook, so $\mathbb{P}(B^n) \geq 1 - \varepsilon$ and $|B^n| \leq 2^{nR_n}$.

Fix an arbitrary $\varepsilon' > 0$ — this is the *typical-set width*, and is unrelated to ε. Apply Lemma 3.4 with $\delta = \varepsilon$, typical-set width ε', and $\eta = \frac{1-\varepsilon}{2} > 0$: for all large n,

$$\big|B^n\big| \ \geq \ \frac{1-\varepsilon}{2}\, 2^{\,n(H(X)-\varepsilon')}.$$

Taking logarithms and dividing by n,

$$R_n = \frac{1}{n}\lceil \log |B^n| \rceil \ \geq \ \frac{1}{n}\log |B^n| \ \geq \ H(X) - \varepsilon' + \frac{1}{n}\log\frac{1-\varepsilon}{2}.$$

The last term tends to 0 (the constant $\frac{1-\varepsilon}{2}$ is strictly positive and does not depend on n), so

$$\liminf_{n\to\infty} R_n \ \geq \ H(X) - \varepsilon'.$$

Since $\varepsilon' > 0$ was arbitrary, letting $\varepsilon' \downarrow 0$ gives $\liminf_n R_n \geq H(X)$. □

Q. *State and prove the strong converse.*

Theorem 6.2 (Strong converse)**.** *Let X be a DMS with pmf P_X and let $(\varphi_n, \psi_n)_{n\geq 1}$ be fixed-rate codes with*

$$\limsup_{n\to\infty} R_n \ \leq \ R \ < \ H(X).$$

Then $\lim_{n\to\infty} P_e^n = 1$.

Proof. Let B^n be the codebook. Since $R_n \le R$ for all large n, we have $|B^n| \le 2^{nR_n} \le 2^{nR}$.

Choose the typical-set width

$$\varepsilon := \frac{H(X) - R}{2} \;>\; 0,$$

and split the success probability according to typicality:

$$\mathbb{P}\big(B^n\big) = \mathbb{P}\big(B^n \cap A_\varepsilon^n\big) + \mathbb{P}\big(B^n \setminus A_\varepsilon^n\big) \;\le\; \mathbb{P}\big(B^n \cap A_\varepsilon^n\big) + \mathbb{P}\big(\big(A_\varepsilon^n\big)^c\big).$$

For the first term, every $x^n \in A_\varepsilon^n$ has $P_{X^n}(x^n) \le 2^{-n(H(X)-\varepsilon)}$, so

$$\mathbb{P}\big(B^n \cap A_\varepsilon^n\big) \;\le\; \big|B^n\big| \cdot 2^{-n(H(X)-\varepsilon)} \;\le\; 2^{nR}\, 2^{-n(H(X)-\varepsilon)} \;=\; 2^{-n(H(X)-\varepsilon-R)}.$$

With the choice of ε above,

$$H(X) - \varepsilon - R = H(X) - R - \frac{H(X) - R}{2} = \frac{H(X) - R}{2} > 0,$$

so this term decays exponentially in n. For the second term, $\mathbb{P}(\big(A_\varepsilon^n\big)^c) \to 0$ by Proposition 3.2.2. Hence

$$\mathbb{P}\big(B^n\big) \;\le\; 2^{-n\frac{H(X)-R}{2}} + \mathbb{P}\big(\big(A_\varepsilon^n\big)^c\big) \;\longrightarrow\; 0,$$

and therefore $P_e^n = 1 - \mathbb{P}(B^n) \to 1$. □

Remark 6.3 (Weak versus strong)**.** The weak converse says that error probability bounded away from 1 forces the rate up to at least $H(X)$. The strong converse says more: below $H(X)$ there is no rate–reliability trade-off at all. One cannot buy a slightly smaller rate at the price of a moderate, fixed error probability — the error probability is driven all the way to 1. Together with the achievability construction of the previous section, $H(X)$ is a sharp threshold.

Remark 6.4 (Where each hypothesis is used)**.** The weak converse needs only $\varepsilon < 1$ and gives a bound on $\liminf R_n$; it uses Lemma 3.4, i.e. the lower bound on the size of a high-probability set. The strong converse instead assumes a bound on the rate and uses the upper bound $P_{X^n} \le 2^{-n(H-\varepsilon)}$ on the typical set. The two are dual: one bounds a set's size from its probability, the other bounds a set's probability from its size.