Chapter 3: Asymptotic Equipartition Property
Chapter 3 outline

• Strong vs. Weak Typicality
• Convergence
• Asymptotic Equipartition Property Theorem
• High-probability sets and the “typical set”
• Consequences of the AEP: data compression
Strong versus Weak Typicality

- Intuition behind typicality?

- \( \mathcal{X} = \{\spadesuit, \lozenge, \heartsuit, \clubsuit\} \) with pmf \( p_X = [0.5; 0.25; 0.125; 0.125] \)
  \( \Rightarrow H(X) = 1.75 \) bits.

- Sample sequences consisting of eight i.i.d samples

- strongly typical \( \Rightarrow \) correct proportions
  \( \spadesuit\spadesuit\spadesuit\spadesuit\lozenge\lozenge\heartsuit \) \( - \log p(x) = 14 = 8 \times 1.75 \)

- not typical at all \( \Rightarrow \log p(x) \neq nH(X) \)
  \( \spadesuit\spadesuit\spadesuit\spadesuit\spadesuit\spadesuit\spadesuit \) \( - \log p(x) = 24 \neq 8 \times 1.75 \)

- weakly typical \( \Rightarrow \log p(x) = nH(X) \)
  \( \spadesuit\spadesuit\lozenge\lozenge\lozenge\lozenge\lozenge\lozenge \) \( - \log p(x) = 14 = 8 \times 1.75 \)
Another example of Typicality

• Bit-sequences of length \( n = 8 \), \( \text{prob}(1) = p \) (\( \text{prob}(0) = (1-p) \))

• There are \( 2^8 \) sequences, which for \( p \neq 1/2 \) have different probabilities

• \( \text{Pr}(00000000) = (1 - p)^8 \)

• \( \text{Pr}(10001100) = p^3(1 - p)^5 \)

• \( \text{Pr}(11111111) = p^8 \)

• Strong typicality?
  - All sequences with about \( p/n \) 1’s

• Weak typicality?
  - All sequences with probability about \( 2^{-nH(p)} \)

• What if \( p=0.5 \)?
Convergence of random variables

We say a sequence \( \{X_n, n \geq 0\} \) and \( n \) integer, converges to a limit \( X \) if for every \( \epsilon > 0 \), \( \exists m \) such that \( \forall n > m, \ |X_n - X| < \epsilon \).

A sequence of random variables, \( \{X_n, n \geq 0\} \) converge

1. In mean square if \( E(X_n - X)^2 \rightarrow 0 \)

2. In probability if for every \( \epsilon > 0 \), \( \Pr\{|X_n - X| > \epsilon\} \rightarrow 0 \)

3. With probability 1 (also called almost surely) if \( \Pr\{\lim_{n \rightarrow \infty} X_n = X\} = 1 \)
Weak Law of Large Numbers + the AEP

- Let $X_1, X_2, \ldots$, be i.i.d distributed with mean $\mu$ and variance $\sigma^2 < \infty$. Let

$$S_n \triangleq \frac{1}{n} [X_1 + X_2 + \ldots + X_n]$$

- **Theorem: Weak Law of Large Numbers**

$$S_n \rightarrow \mu \quad \text{in probability}$$

- **Theorem: Asymptotic Equipartition Property (AEP):**

  If $X_1, X_2, \ldots \overset{iid}{\sim} p(x)$, then

  $$-\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) \rightarrow H(X) \quad \text{in probability}.$$
Typical sets intuition

• What’s the point?
• Consider iid bit-strings strings of length N=100, prob(1) = p_1=0.1
• Probability of a given string X with \( r \) ones is

\[
P(X) = p_1^r (1 - p_1)^{N-r}
\]

• Number of strings with \( r \) ones is

\[
\binom{N}{r} = \frac{N!}{r!(N - r)!}
\]

• Distribution of \( r \), the # of ones in a string of length N is thus

\[
P(r) = \binom{N}{r} p_1^r (1 - p_1)^{N-r}
\]
Typical sets intuition

\[ P(r) = \binom{N}{r} p_1^r (1-p_1)^{N-r} \]

\[ n(r)P(x) = \binom{N}{r} p_1^r (1-p_1)^{N-r} \]

- Consider iid bit-strings of length N, prob(1) = \(p_1 = 0.1\)
Typical sets intuition

- What’s the point?
- Consider iid bit-strings strings of length $N=100$, $\text{prob}(1) = p_1 = 0.1$

<table>
<thead>
<tr>
<th>$x$</th>
<th>$\log_2(P(x))$</th>
</tr>
</thead>
<tbody>
<tr>
<td>........................................</td>
<td>-50.1</td>
</tr>
<tr>
<td>........................................</td>
<td>-37.3</td>
</tr>
<tr>
<td>........................................</td>
<td>-65.9</td>
</tr>
<tr>
<td>........................................</td>
<td>-56.4</td>
</tr>
<tr>
<td>........................................</td>
<td>-53.2</td>
</tr>
<tr>
<td>........................................</td>
<td>-43.7</td>
</tr>
<tr>
<td>........................................</td>
<td>-46.8</td>
</tr>
<tr>
<td>........................................</td>
<td>-56.4</td>
</tr>
<tr>
<td>........................................</td>
<td>-37.3</td>
</tr>
<tr>
<td>........................................</td>
<td>-43.7</td>
</tr>
<tr>
<td>........................................</td>
<td>-56.4</td>
</tr>
<tr>
<td>........................................</td>
<td>-37.3</td>
</tr>
<tr>
<td>........................................</td>
<td>-56.4</td>
</tr>
<tr>
<td>........................................</td>
<td>-59.5</td>
</tr>
<tr>
<td>........................................</td>
<td>-46.8</td>
</tr>
<tr>
<td>........................................</td>
<td>-15.2</td>
</tr>
<tr>
<td>........................................</td>
<td>-332.1</td>
</tr>
</tbody>
</table>

Figure 4.10. The top 15 strings are samples from $X^{100}$, where $p_1 = 0.1$ and $p_0 = 0.9$. The bottom two are the most and least probable strings in this ensemble. The final column shows the log-probabilities of the random strings, which may be compared with the entropy $H(X^{100}) = 46.9$ bits.
Definition: weak typicality

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

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

- All sequence in the typical set have roughly equal probabilities. Clearly, a new notion of approximation is used in such a statement. We call that the “exponential approximation.”

- If $(x_1, x_2, \ldots, x_n) \in A_{\epsilon}^{(n)}$, then

$$H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon.$$
The typical set visually

**FIGURE 3.1.** Typical sets and source coding.

*Cover+Thomas pg. 60*

---

Most + least likely sequences

NOT in typical set!!

*Mackay pg. 81*
Properties of the typical set

1. If \((x_1, x_2, \cdots, x_n) \in A_\epsilon^{(n)}\) then \(H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \cdots, x_n) \leq H(X) + \epsilon\)

2. \(\Pr\{A_\epsilon^{(n)}\} > 1 - \epsilon\) for \(n\) sufficiently large.

3. \((1 - \epsilon)2^{n(H(X) - \epsilon)} \leq |A_\epsilon^{(n)}| \leq 2^{n(H(X) + \epsilon)}\) for \(n\) sufficiently large.

[Cover+Thomas pg. 60]
Consequences of the AEP

Typical set contains almost all the probability!

How many are in this set useful for source coding (compression)!

FIGURE 3.1. Typical sets and source coding.

FIGURE 3.2. Source code using the typical set.
Consequences of the AEP

Let $x^n$ denote $(x_1, x_2, \ldots, x_n)$, and let $l(x^n)$ be the length of the codeword corresponding to $x^n$.

**Coding Scheme:**

- if $x^n \in A_\epsilon^{(n)}$: ‘0’ + at most $1 + n(H(X) + \epsilon)$
- if $x^n \notin A_\epsilon^{(n)}$: ‘1’ + at most $1 + n \log |\mathcal{X}|$

If $n$ is sufficiently large so that $\Pr\{A_\epsilon^{(n)}\} \geq 1 - \epsilon$, the expected codeword length is

$$E[l(X^n)] = \sum_{x^n} p(x^n) l(x^n)$$

$$\leq n(H + \epsilon) + \epsilon n \log |\mathcal{X}| + 2$$

$$= n(H + \epsilon')$$
AEP and data compression

**Theorem: Data Compression** Let $X^n \overset{iid}{\sim} p(x)$ and let $\epsilon > 0$. Then there exists a code that maps sequences $x^n$ of length $n$ into binary strings such that the mapping is one-to-one (and therefore invertible) and

$$E \left[ \frac{1}{n} l(X^n) \right] \leq H(X) + \epsilon$$

for $n$ sufficiently large.
``High-probability set” vs. “typical set”

- Typical set: small number of outcomes that contain most of the probability
- Is it the smallest such set?

**Definition:** Smallest set for a given probability. For each \( n = 1, 2, \ldots \), let \( B^{(n)}_\delta \subset \mathcal{X}^n \) be the smallest set with

\[
\Pr\{B^{(n)}_\delta\} \geq 1 - \delta.
\]

**Theorem:** Problem 3.3.11 Let \( X_1, X_2, \ldots, X_n \) \( \text{iid} \) \( p(x) \). For \( \delta < 1/2 \) and \( \delta' > 0 \), if \( \Pr\{B^{(n)}_\delta\} > 1 - \delta \), then for sufficiently large \( n \) we have

\[
\frac{1}{n} \log |B^{(n)}_\delta| > H - \delta'.
\]

So (to first order) \( B^{(n)}_\delta \) has at least \( 2^{nH} \) elements.
Some notation

**Definition:** The notation \( a_n \div b_n \) means \( \lim_{n \to \infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0 \). Thus, \( a_n \) and \( b_n \) are equal to the first order in the exponent.

Thus, if \( \delta_n \to 0 \) and \( \epsilon_n \to 0 \), the “high-probability set” \( B^{(n)}_{\delta_n} \) and the “typical set” \( A^{(n)}_{\epsilon_n} \) are related as

\[
|B^{(n)}_{\delta_n}| \div |A^{(n)}_{\epsilon_n}| \div 2^{nH}
\]
What’s the difference?

*Theorem:* Let $X_1, X_2, \cdots, X_n$ be i.i.d. $\sim p(x)$. For $\delta > \frac{1}{2}$ and any $\delta' > 0$, if $\Pr\{B^{(n)}_\delta\} > 1 - \delta$, then

$$\frac{1}{n} \log |B^{(n)}_\delta| > H - \delta'$$

for $n$ sufficiently large.

• Why use the “typical set” rather than the “high-probability set”?