Let $X_1, \ldots, X_n$ be i.i.d. random variables drawn from a distribution $p(\cdot)$.

Data compression involves finding short descriptions for such sequences.

For this, we treat sequences in the typical set differently from the other ‘non-typical’ sequences.

Typical set $A_{\epsilon}^{(n)}$ with $2^{n(H+\epsilon)}$ elements

Non-typical set

$\mathcal{X}^n$
Describing Typical and Non-Typical Sequences

- Partition the alphabet-sequence set $\mathcal{X}^n$ into two: typical and non-typical.
- Order the elements in each set.
  - Any total order suffices (e.g., a lexicographical ordering).
- Then, every sequence in $A_{\epsilon}^{(n)}$ can be identified with its index in this ordering.
- Since there are at most $2^{n(H+\epsilon)}$ sequences in $A_{\epsilon}^{(n)}$, the indexing requires at most $n(H + \epsilon) + 1$ bits.
  - Prefix all these sequences with 0, giving a maximum length of $n(H + \epsilon) + 2$ bits.
- Similarly, every sequence in the non-typical set can be indexed using at most $n \log |\mathcal{X}| + 1$ bits.
  - Prefix these indices by 1, giving a maximum length of $n \log |\mathcal{X}| + 2$ bits.
- The collection of all these sequences offers a description of the entire set $\mathcal{X}^n$.
  - Each sequence thus obtained is also called a code for a sequence in $\mathcal{X}^n$. 
Features of the Coding Scheme

- The code is a bijection and therefore, decodable. The first bit acts as a flag to indicate the length of the code that follows.
- All typical sequences have short description length ($\approx nH$).

- **But what is the length of any sequence description?**
  - We want to compute the *expected value* of the description lengths.
Average Length of Descriptions

- Let \( x^n \) denote a sequence of \( n \)-length: \( x_1, \ldots, x_n \).
- Let \( l(x^n) \) denote the length of the code of \( x^n \).

**Theorem**

Let \( X^n \) be i.i.d. with distribution \( p(x) \), 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 a bijection and

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

for large enough value of \( n \).

- Thus, we can represent sequences \( X^n \) using \( nH(X) \) bits on the average.
Small Typical Sets

- Typical sets are highly *probable* sets (hence the name).
- However, in terms of cardinality, they are small sets.
- The question we now ask is: is $A_{\epsilon}^{(n)}$ the smallest set containing most of the probability?
  - To answer this, we are going to define a *parameterized* set, the probability of which is defined in terms of that parameter.
High Probability Sets

- For \( n = 1, 2, \ldots \), a high-probability set \( B_\delta^{(n)} \subset \mathcal{X}^n \) is the smallest set with
  \[
  Pr \left\{ B_\delta^{(n)} \right\} \geq 1 - \delta
  \]

- The intuition is that if two sets on the same (finite) domain have high probability, then they must have “significant” intersection.

Theorem

Let \( X_1, \ldots, X_n \) be i.i.d. with distribution \( p(x) \). For any \( \delta < 1/2 \) and \( \delta' > 0 \),

\[
Pr \left\{ B_\delta^{(n)} \right\} \geq 1 - \delta
\]

\[
\Rightarrow \frac{1}{n} \log \left| B_\delta^{(n)} \right| > H - \delta'
\]

for sufficiently large \( n \).
The implication of the theorem is that $B_{\delta}^{(n)}$ has at least $2^{nH}$ elements.

But we also know that $A_{\epsilon}^{(n)}$ has $2^{n(H \pm \epsilon)}$ elements.

Therefore, the size of the typical set is about the same as the smallest high-probability set.

This is not an equality, but an agreement to the first order in the exponent, that is

$$A_{\epsilon}^{(n)} \approx B_{\delta}^{(n)} \approx 2^{nH}$$