ECE 515
Information Theory

Typical Sequences

Typical Sequences

- Consider a binary discrete memoryless source (DMS) X = {0,1} with symbol probabilities

  p(1) = 1/4 p(0) = 3/4

- Sequences of N = 20 symbols

  1. 1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1
  2. 1,0,1,0,1,0,0,0,0,0,0,0,0,0,1,1,0,0,0,1
  3. 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0

Tchebycheff Inequality

$$\eta_X \equiv E[X] = \sum_{k=1}^{K} p(x_k) \times x_k$$

$$\sigma_X^2 \equiv E[(X - \eta_X)^2] = \sum_{k=1}^{K} p(x_k) \times (x_k - \eta_X)^2$$

$$Pr\{|X - \eta_X| \geq \delta\} \leq \frac{\sigma_X^2}{\delta^2}$$

Weak Law of Large Numbers

- Sequence of *N* i.i.d. RVs

$$\overline{X} = X_1, \ldots, X_n, \ldots, X_N$$

- Define a new RV

$$Y_N \equiv \frac{1}{N} \sum_{n=1}^{N} X_n$$

$$\eta_{Y_N} = \eta_X \quad \sigma_{Y_N}^2 = \frac{\sigma_X^2}{N}$$

Weak Law of Large Numbers

$$\lim_{N\to\infty} Pr\left\{\left|\left[\frac{1}{N}\sum_{n=1}^{N} X_n\right] - \eta_X\right| \geq \delta\right\} = 0$$

$$\lim_{N\to\infty} Pr\left\{\left|\left[\frac{1}{N}\sum_{n=1}^{N} X_n\right] - \eta_X\right| < \delta\right\} = 1$$

The sample average approaches the statistical mean

Asymptotic Equipartition Property

- N i.i.d. random variables X_1, …, X_N

$$\mathrm{p}(X_1, X_2, \ldots, X_N) = \mathrm{p}(X_1)\mathrm{p}(X_2)\ldots\mathrm{p}(X_N)$$

$$-\frac{1}{N}\log \mathrm{p}(X_1, X_2, \ldots, X_N) = -\frac{1}{N}\sum_{n=1}^{N}\log \mathrm{p}(X_n) \rightarrow -\mathrm{E}\left[\log \mathrm{p}(X)\right] = \mathrm{H}(X)$$

$$\text{as } N \rightarrow \infty$$

Typical Sequences

- RV X where $p(x_k) = p_k$
- Consider a sequence **x** of length N where x_k appears approximately Np_k times

$$
\begin{aligned}
\mathrm{p}(\mathbf{x}) &\approx p_1^{Np_1} p_2^{Np_2} \cdots p_K^{Np_K} \\
&= \prod_{k=1}^{K} p_k^{Np_k} = \prod_{k=1}^{K} ((2^{\log_2 p_k})^{Np_k} \\
&= \prod_{k=1}^{K} 2^{Np_k \log_2 p_k} = 2^{N \sum_{k=1}^{K} p_k \log_2 p_k} \\
&= 2^{-N\mathrm{H}(\mathrm{X})}
\end{aligned}
$$

Typical Sequences

- Binary RV X where p(x_1) = p and p(x_2) = 1-p
- The number of sequences **x** of length N with Np x_1's is

$$\binom{N}{Np} = \frac{N!}{(Np)!(N(1-p))!}$$

- Stirling's approximation

$$n! \approx \sqrt{2\pi n} n^n \mathrm{e}^{-n}$$

Typical Sequences

$$\binom{N}{Np} \approx \frac{N^N e^{-N}}{(Np)^{Np} e^{-Np} (N(1-p))^{N(1-p)} e^{-N(1-p)}}$$

$$= \frac{1}{p^{Np} (1-p)^{N(1-p)}}$$

$$= p^{-Np} (1-p)^{-N(1-p)}$$

$$= 2^{-Np \log p - N(1-p) \log(1-p)}$$

$$= 2^{N(-p \log p - (1-p) \log(1-p))}$$

$$= 2^{N \mathrm{H}(\mathrm{X})}$$

Typical Sequences

- Consider a binary discrete memoryless source (DMS) $X = \{0,1\}$ with symbol probabilities

  $p(1) = 1/4 \qquad p(0) = 3/4$

- $H(X) = 0.811$ bit
- Sequences of $N = 20$ symbols
- $2^{-NH(X)} = 1.3050 \times 10^{-5}$
- $2^{NH(X)} = 76627$

Summary

- The Tchebycheff inequality was used to prove the weak law of large numbers (WLLN)
  – the sample average approaches the statistical mean as $N \to \infty$

- The WLLN was used to prove the AEP

$$-\frac{1}{N}\sum_{n=1}^{N}\log \mathrm{p}(\mathrm{X}_n) \to \mathrm{H}(\mathrm{X}) \text{ as } N \to \infty$$

- A typical sequence has probability $\mathrm{p}(\mathbf{x}) \approx 2^{-N\mathrm{H}(\mathrm{X})}$

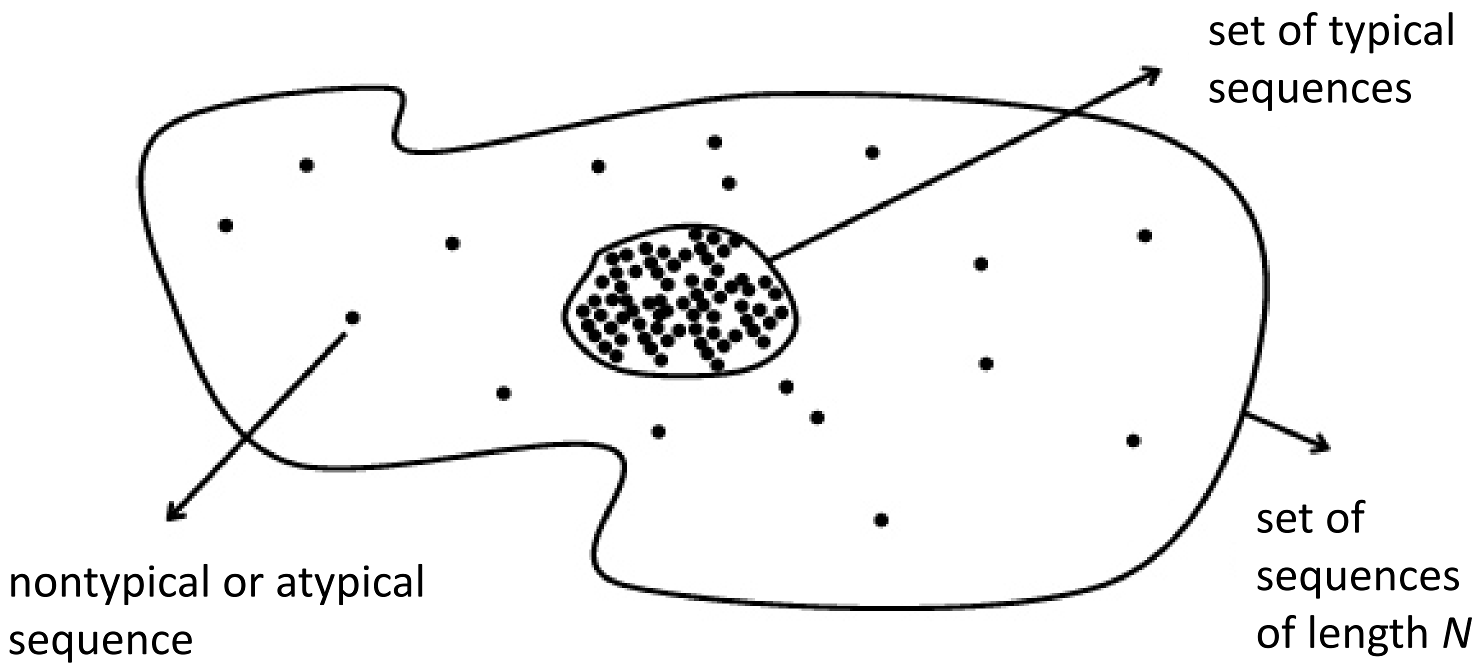
- There are about $2^{N\mathrm{H}(\mathrm{X})}$ typical sequences of length N

Typical Sequences

$$\mathcal{T}_X(\delta) \equiv \{\mathbf{x} : \left| -\frac{1}{N} \log_b p(\mathbf{x}) - H(X) \right| < \delta\}$$

$$\mathcal{T}_X^c(\delta) \equiv \{\mathbf{x} : \left| -\frac{1}{N} \log_b p(\mathbf{x}) - H(X) \right| \geq \delta\}$$

Typical Sequences



Interpretation

- Although there are very many results that may be produced by a random process, the one actually produced is most probably from a set of outcomes that all have approximately the same chance of being the one actually realized.
- Although there are individual outcomes which may have a higher probability than outcomes in this set, the vast number of outcomes in the set almost guarantees that the outcome will come from the set.
- ``Almost all events are almost equally surprising''

  Cover and Thomas

Typical Sequences

- From the definition, the probability of occurrence of a typical sequence p(**x**) is

$$b^{-N[H(X)+\delta]} < p(\mathbf{x}) < b^{-N[H(X)-\delta]}$$

Example

- p(x_1) = p(1) = 1/4   p(x_2) = p(0) = 3/4
- H(X) = 0.811 bit
- N = 3
- p(x_1,x_1,x_1) = 1/64
- p(x_1,x_1,x_2) = p(x_1,x_2,x_1) = p(x_2,x_1,x_1) = 3/64
- p(x_1,x_2,x_2) = p(x_2,x_2,x_1) = p(x_2,x_1,x_2) = 9/64
- p(x_2,x_2,x_2) = 27/64

Example

- H(X) = 0.811 bit $N = 3$ $b = 2$

$$2^{-3[.811+\delta]} < p(x_1, x_2, x_3) < 2^{-3[.811-\delta]}$$

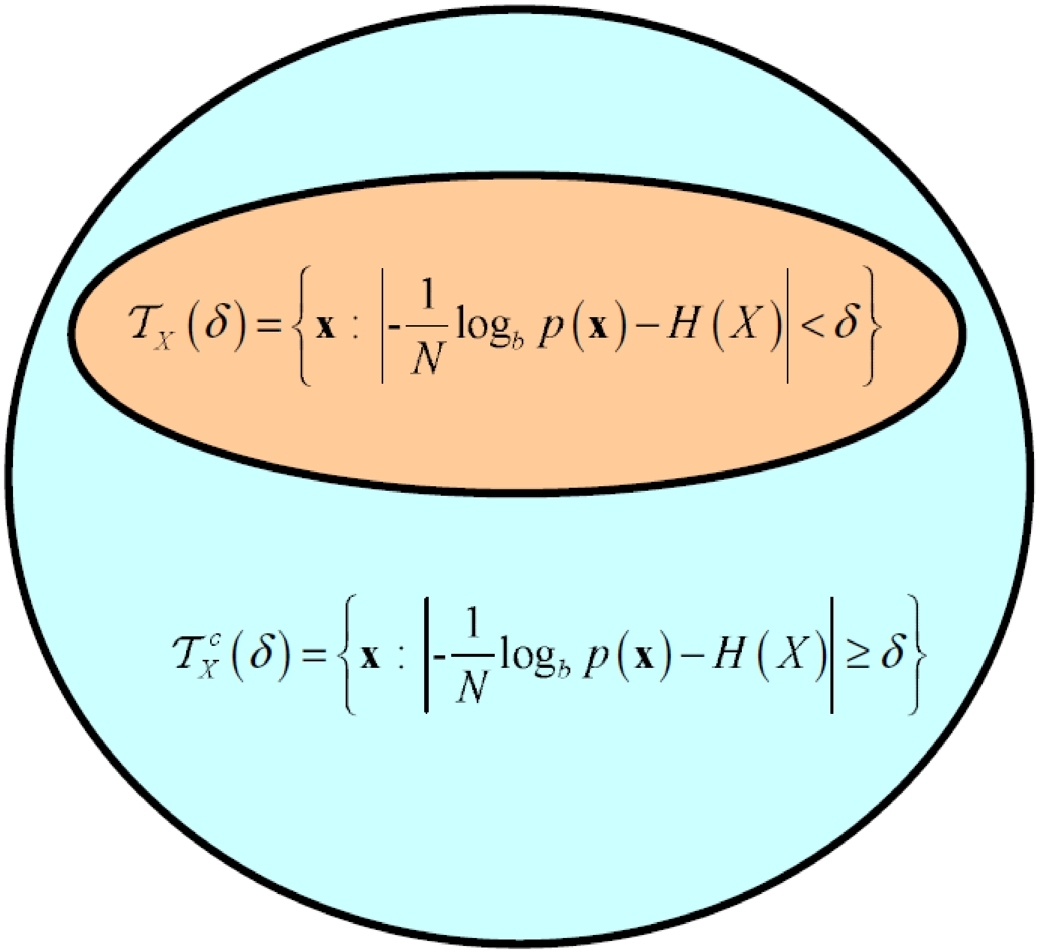
- x_1,x_1,x_1 $1/64 = 2^{-3[.811+1.199]}$
- x_1,x_1,x_2 $3/64 = 2^{-3[.811+0.661]}$
- x_1,x_2,x_2 $9/64 = 2^{-3[.811+0.132]}$
- x_2,x_2,x_2 $27/64 = 2^{-3[.811-0.395]}$

Example

- If $\delta = 0.2$ the typical sequences are
  - (x_1,x_2,x_2), (x_2,x_1,x_2), (x_2,x_2,x_1)
    with probability 0.422
    (1,0,0), (0,1,0), (0,0,1)
- If $\delta = 0.4$ the typical sequences are
  - (x_1,x_2,x_2), (x_2,x_1,x_2), (x_2,x_2,x_1), (x_2,x_2,x_2)
    with probability 0.844
    (1,0,0), (0,1,0), (0,0,1), (0,0,0)

Occurrences of x_1 n	Number of sequences $\binom{N}{n}$	Probability of each sequence $p(x_1)^n\ p(x_2)^{N-n}$	Probability of all sequences $\binom{N}{n}\ p(x_1)^n\ p(x_2)^{N-n}$
0	1	$3,171 \times 10^{-3} = 2^{-20\times 0,415}$	0,003171
1	20	$1,057 \times 10^{-3} = 2^{-20\times 0,494}$	0,021141
2	190	$3,524 \times 10^{-4} = 2^{-20\times 0,574}$	0,066948
3	1140	$1,175 \times 10^{-4} = 2^{-20\times 0,653}$	0,133896
4	4845	$3,915 \times 10^{-5} = 2^{-20\times 0,732}$	0,189685
5	15504	$1,305 \times 10^{-5} = 2^{-20\times 0,811}$	0,202331
6	38760	$4,350 \times 10^{-6} = 2^{-20\times 0,891}$	0,168609
7	77520	$1,450 \times 10^{-6} = 2^{-20\times 0,970}$	0,112406
8	125970	$4,833 \times 10^{-7} = 2^{-20\times 1,049}$	0,060887
9	167960	$1,611 \times 10^{-7} = 2^{-20\times 1,128}$	0,027061
10	184756	$5,370 \times 10^{-8} = 2^{-20\times 1,208}$	0,009922
11	167960	$1,790 \times 10^{-8} = 2^{-20\times 1,287}$	0,003007
12	125970	$5,967 \times 10^{-9} = 2^{-20\times 1,366}$	0,000752
13	77520	$1,989 \times 10^{-9} = 2^{-20\times 1,445}$	0,000154
14	38760	$6,630 \times 10^{-10} = 2^{-20\times 1,525}$	0,000026
15	15504	$2,210 \times 10^{-10} = 2^{-20\times 1,604}$	0,000003
16	4845	$7,367 \times 10^{-11} = 2^{-20\times 1,683}$	0,000000
17	1140	$2,456 \times 10^{-11} = 2^{-20\times 1,762}$	0,000000
18	190	$8,185 \times 10^{-12} = 2^{-20\times 1,842}$	0,000000
19	20	$2,728 \times 10^{-12} = 2^{-20\times 1,921}$	0,000000
20	1	$9,095 \times 10^{-13} = 2^{-20\times 2,000}$	0,000000

Typical Sequences



- Random variable X
- Alphabet size *K*
- Entropy H(X)
- Arbitrary number δ>0
- Sequences **x** of blocklength $N \geq N_0$ and probability p(**x**)
- $\left\| \mathcal{T}_X(\partial) \right\| + \left\| \mathcal{T}_X^{\,c}(\partial) \right\| = K^N$

Shannon-McMillan Theorem

a) The probability that a particular sequence $\mathbf{x}$ of blocklength N belongs to the set of atypical sequences $\mathcal{T}_X^c(\delta)$ is upperbounded as:

$$Pr[\mathbf{x} \in \mathcal{T}_X^c(\delta)] < \epsilon$$

b) If a sequence $\mathbf{x}$ is in the set of typical sequences $\mathcal{T}_X(\delta)$ then its probability of occurrence $p(\mathbf{x})$ is approximately equal to $b^{-NH(X)}$, that is:

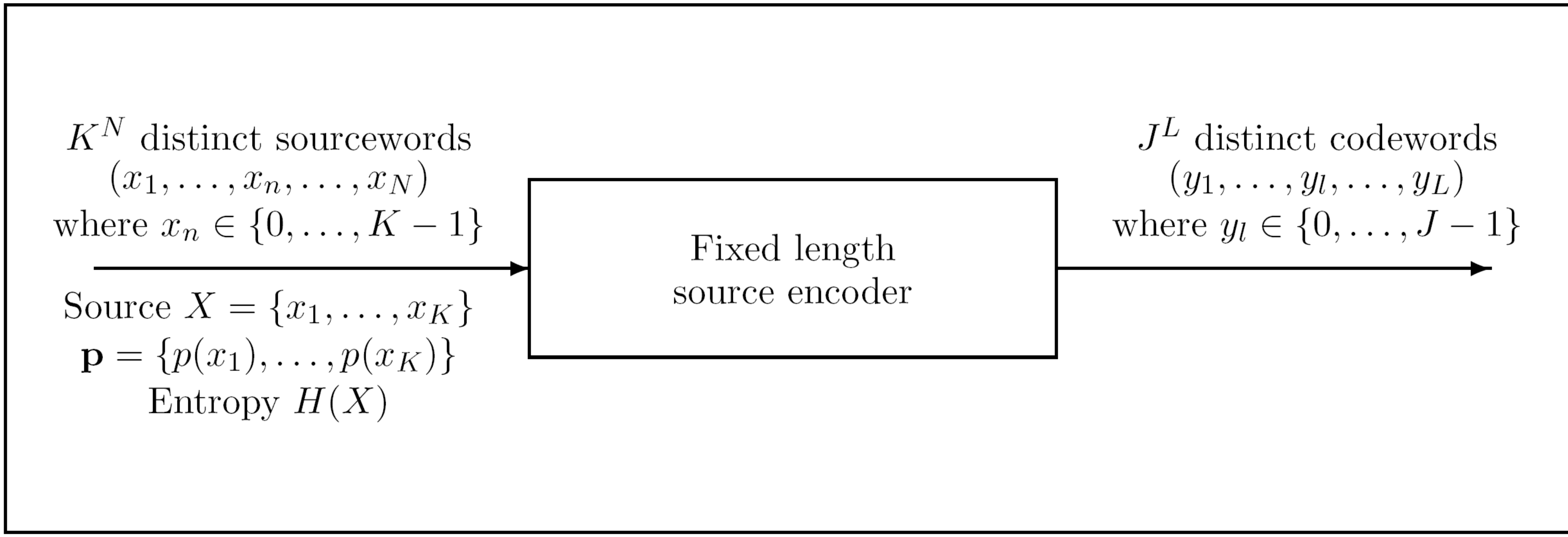
$$b^{-N[H(X)+\delta]} < p(\mathbf{x}) < b^{-N[H(X)-\delta]}$$

c) The number of typical, or likely, sequences $\|\mathcal{T}_X(\delta)\|$ is bounded by:

$$(1-\epsilon)b^{N[H(X)-\delta]} < \|\mathcal{T}_X(\delta)\| < b^{N[H(X)+\delta]}$$

- The essence of source coding or data compression is that as $N \to \infty$, atypical sequences almost never appear as the output of the source.
- Therefore, one can focus on representing typical sequences with codewords and ignore atypical sequences.
- Since there are only about $2^{N\mathrm{H}(X)}$ typical sequences of length N, and they are approximately equiprobable, it takes about $N\mathrm{H}(X)$ bits to represent them.
- On average it takes H(X) bits to represent a source symbol.

Fixed Length Source Compaction Codes



Fixed Length Source Compaction Codes

- If $J^L < K^N$ we cannot uniquely encode all source words with length L codewords
- Two questions
  1. How small can J^L be such that performance is acceptable?
  2. How should sourcewords be encoded to length L codewords for unique decodability?

The number of typical sequences satisfies

$$\left\| T_X(\delta) \right\| < b^{N[\mathrm{H(X)}+\delta]}$$

so encoding all typical sequences with length *L* codewords requires that

$$J^L \geq b^{N[\mathrm{H(X)}+\delta]}$$

- Although the set of atypical sequences may be large, the Shannon-McMillan Theorem ensures that

$$Pr[\mathbf{x} \in \mathcal{T}_X^c(\delta)] < \epsilon$$

- Thus it is possible to encode sourcewords with an arbitrarily small block decoding failure probability P_e provided that
  - $L\log_b J > N\mathrm{H}(X)$
  - N is sufficiently large

Example

- $K = J = 2$
- $p(x_1) = 0.1$ $p(x_2) = 0.9$ $H(X) = 0.469$ bit
- Choose $N = 4$, $L = 3$

$$R = \frac{L}{N} = \frac{3}{4} > \mathrm{H(X)}$$

- Partition the 16 sourcewords into 7 typical sequences and 9 atypical sequences

$$
\begin{aligned}
p(x_1)^4 &= 0.0001 & \binom{4}{4} &= 1 \text{ sourceword} \\
p(x_1)^3 p(x_2) &= 0.0009 & \binom{4}{3} &= 4 \text{ sourcewords} \\
p(x_1)^2 p(x_2)^2 &= 0.0081 & \binom{4}{2} &= 6 \text{ sourcewords} \\
p(x_1) p(x_2)^3 &= 0.0729 & \binom{4}{1} &= 4 \text{ sourcewords} \\
p(x_2)^4 &= 0.6561 & \binom{4}{0} &= 1 \text{ sourceword}
\end{aligned}
$$

The Code

Typical Sequence	Codeword
$x_2x_2x_2x_2$	000
$x_1x_2x_2x_2$	100
$x_2x_1x_2x_2$	010
$x_2x_2x_1x_2$	001
$x_2x_2x_2x_1$	110
$x_1x_1x_2x_2$	101
$x_1x_2x_1x_2$	011

The Code

Atypical Sequence	Codeword
$x_1x_2x_2x_1$	111 0000
$x_2x_1x_1x_2$	111 1000
$x_2x_1x_2x_1$	111 0100
$x_2x_2x_1x_1$	111 0010
$x_1x_1x_1x_2$	111 0001
$x_1x_1x_2x_1$	111 1100
$x_1x_2x_1x_1$	111 1010
$x_2x_1x_1x_1$	111 1001
$x_1x_1x_1x_1$	111 0110

Code Rate

- The actual code rate is

$$R = \frac{.9639 \times 3 + .0361 \times 7}{4} = \frac{3}{4} + .0361 = .7861$$

Example

- $K = J = 2$
- $p(x_1) = 0.1$ $p(x_2) = 0.9$ H(X) = 0.469 bit
- Choose $N = 8$, $L = 6$

$$R = \frac{L}{N} = \frac{6}{8} = \frac{3}{4} > \mathrm{H(X)}$$

- Partition the 256 sourcewords into 63 typical sequences and 193 atypical sequences

$$
\begin{aligned}
p(x_1)^8 &= 1.0000 \times 10^{-8} & \tbinom{8}{8} &= 1 \text{ sourceword} \\
p(x_1)^7 p(x_2) &= 9.0000 \times 10^{-8} & \tbinom{8}{7} &= 8 \text{ sourcewords} \\
p(x_1)^6 p(x_2)^2 &= 8.1000 \times 10^{-7} & \tbinom{8}{6} &= 28 \text{ sourcewords} \\
p(x_1)^5 p(x_2)^3 &= 7.2900 \times 10^{-6} & \tbinom{8}{5} &= 56 \text{ sourcewords} \\
p(x_1)^4 p(x_2)^4 &= 6.5610 \times 10^{-5} & \tbinom{8}{4} &= 70 \text{ sourcewords} \\
p(x_1)^3 p(x_2)^5 &= 5.9049 \times 10^{-4} & \tbinom{8}{3} &= 56 \text{ sourcewords} \\
p(x_1)^2 p(x_2)^6 &= 5.3144 \times 10^{-3} & \tbinom{8}{2} &= 28 \text{ sourcewords} \\
p(x_1) p(x_2)^7 &= 4.7830 \times 10^{-2} & \tbinom{8}{1} &= 8 \text{ sourcewords} \\
p(x_2)^8 &= 4.3047 \times 10^{-1} & \tbinom{8}{0} &= 1 \text{ sourceword}
\end{aligned}
$$

Code Rate

- For N = 8, L = 6 the actual code rate is

$$R = \frac{.9773 \times 6 + .0227 \times 14}{8} = \frac{3}{4} + .0227 = .7727$$

Theorem *(Converse of the Source Coding Theorem)*

Let $\epsilon > 0$. Given a memoryless source X of entropy $H(X)$, a codeword alphabet size J and a codeword length L, if:

a) $L \log_b J < NH(X)$ and

b) $N \geq N_0$

then the probability of decoding failure P_e is lower bounded by:

$$P_e \;\; > \;\; 1 - \epsilon$$

Example

- $K = J = 2$
- $p(x_1) = 0.3$ $p(x_2) = 0.7$ $H(X) = 0.881$ bit
- Choose $N = 4$, $L = 3$

$$R = \frac{L}{N} = \frac{3}{4} < H(X)$$

- Partition the 16 sourcewords into 7 typical sequences and 9 atypical sequences

$$
\begin{aligned}
p(x_1)^4 &= 2.4010 \times 10^{-1} & \binom{4}{4} &= 1 \text{ sourceword} \\
p(x_1)^3 p(x_2) &= 1.0290 \times 10^{-1} & \binom{4}{3} &= 4 \text{ sourcewords} \\
p(x_1)^2 p(x_2)^2 &= 4.4100 \times 10^{-2} & \binom{4}{2} &= 6 \text{ sourcewords} \\
p(x_1) p(x_2)^3 &= 1.8900 \times 10^{-2} & \binom{4}{1} &= 4 \text{ sourcewords} \\
p(x_2)^4 &= 8.1000 \times 10^{-3} & \binom{4}{0} &= 1 \text{ sourceword}
\end{aligned}
$$

Example

- $K = J = 2$
- $p(x_1) = 0.3 \quad p(x_2) = 0.7 \quad H(X) = 0.881$ bit
- Choose $N = 8$, $L = 6$

$$R = \frac{L}{N} = \frac{6}{8} = \frac{3}{4} < H(X)$$

- Partition the 256 sourcewords into 63 typical sequences and 193 atypical sequences

$$
\begin{aligned}
p(x_1)^8 &= 6.5610 \times 10^{-5} & \binom{8}{8} &= 1 \text{ sourceword} \\
p(x_1)^7 p(x_2) &= 1.5309 \times 10^{-4} & \binom{8}{7} &= 8 \text{ sourcewords} \\
p(x_1)^6 p(x_2)^2 &= 3.5721 \times 10^{-4} & \binom{8}{6} &= 28 \text{ sourcewords} \\
p(x_1)^5 p(x_2)^3 &= 8.3349 \times 10^{-4} & \binom{8}{5} &= 56 \text{ sourcewords} \\
p(x_1)^4 p(x_2)^4 &= 1.9448 \times 10^{-3} & \binom{8}{4} &= 70 \text{ sourcewords} \\
p(x_1)^3 p(x_2)^5 &= 4.5379 \times 10^{-3} & \binom{8}{3} &= 56 \text{ sourcewords} \\
p(x_1)^2 p(x_2)^6 &= 1.0588 \times 10^{-2} & \binom{8}{2} &= 28 \text{ sourcewords} \\
p(x_1) p(x_2)^7 &= 2.4706 \times 10^{-2} & \binom{8}{1} &= 8 \text{ sourcewords} \\
p(x_2)^8 &= 5.7648 \times 10^{-2} & \binom{8}{0} &= 1 \text{ sourceword}
\end{aligned}
$$

Fixed Length Source Compaction Codes

- If $R > H(X)$, as $N \to \infty$ $P_e \to 0$
- If $R < H(X)$, as $N \to \infty$ $P_e \to 1$