ENTROPY OF X, $|\mathcal{X}| = M$, $\Pr(X{=}j) = p_j$

$$\mathsf{H}(X) = \sum_j -p_j \log p_j = \mathsf{E}[-\log p_X(X)]$$

$-\log p_X(X)$ **is a rv, called the log pmf.**

$\mathsf{H}(X) \geq 0$**; Equality if** X **deterministic.**

$\mathsf{H}(X) \leq \log M$**; Equality if** X **equiprobable.**

If X **and** Y **are independent random symbols, then the random symbol** XY **takes on sample value** xy **with probability** $p_{XY}(xy) = p_X(x)p_Y(y)$**.**

$$\begin{aligned}\mathsf{H}(XY) &= \mathsf{E}[-\log p_{XY}(XY)] = \mathsf{E}[-\log p_X(X)p_Y(Y)] \\ &= \mathsf{E}[-\log p_X(X) - \log p_Y(Y)] = \mathsf{H}(X) + \mathsf{H}(Y)\end{aligned}$$

For a discrete memoryless source, a block of n random symbols, $X_1, \ldots, X_n$, can be viewed as a single random symbol X^n taking on the sample value $\mathrm{x}^n = x_1 x_2 \cdots x_n$ with probability

$$p_{\mathbf{X}^n}(\mathbf{x}^n) = \prod_{j=1}^{n} p_X(x_j)$$

The random symbol X^n has the entropy

$$\begin{aligned} \mathsf{H}(\mathbf{X}^n) &= \mathsf{E}[-\log p_{\mathbf{X}^n}(\mathbf{X}^n)] = \mathsf{E}[-\log \prod_{j=1}^{n} p_X(X_j)] \\ &= \mathsf{E}\left[\sum_{j=1}^{n} -\log p_X(X_j)\right] = n\mathsf{H}(X) \end{aligned}$$

Fixed-to-variable prefix-free codes

Segment input into n-blocks $\mathbf{X}^n = X_1X_2\cdots X_n$.

Form min-length prefix-free code for $\mathbf{X}^n$.

This is called an n-to-variable-length code

$$\mathbf{H}(\mathbf{X}^n) = n\mathbf{H}(X)$$

$$\mathbf{H}(\mathbf{X}^n) \le \mathbf{E}[L(\mathbf{X}^n)]_{\text{min}} < \mathbf{H}(\mathbf{X}^n) + 1$$

$$\overline{L}_{\text{min},n} = \frac{\mathbf{E}[L(X^n)]_{\text{min}}}{n} \qquad \text{bpss}$$

$$\mathbf{H}(X) \le \overline{L}_{\text{min},n} < \mathbf{H}(X) + 1/n$$

WEAK LAW OF LARGE NUMBERS (WLLN)

Let $Y_1, Y_2, \ldots$ be sequence of rv's with mean $\overline{Y}$ and variance σ_Y^2.

The sum $S = Y_1 + \cdots + Y_n$ has mean $n\overline{Y}$ and variance $n\sigma_Y^2$

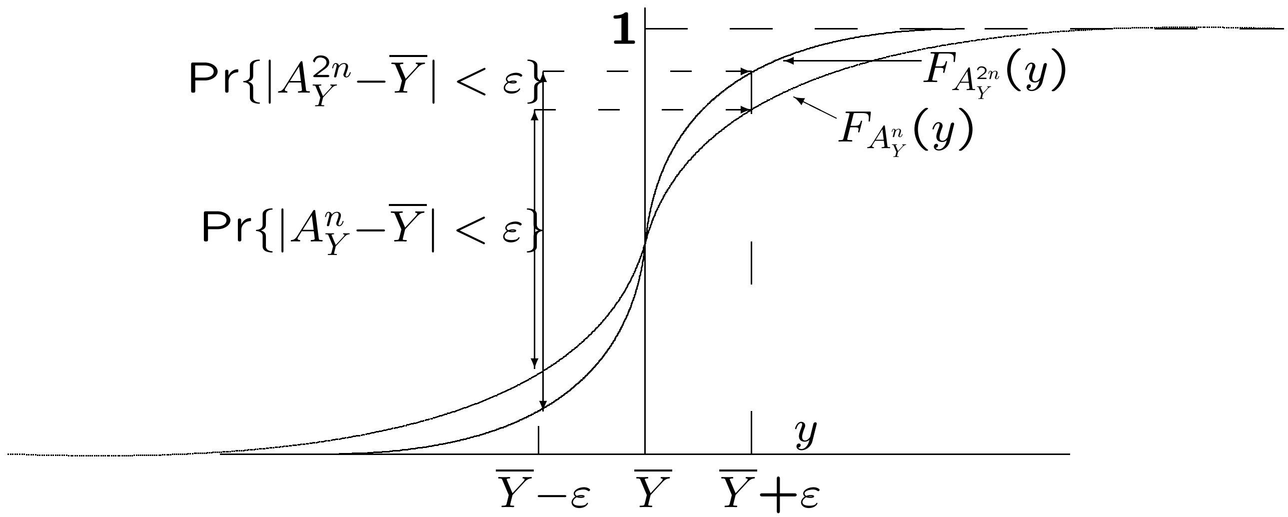
The sample average of $Y_1, \ldots, Y_n$ is

$$A_Y^n = \frac{S}{n} = \frac{Y_1 + \cdots + Y_n}{n}$$

It has mean and variance

$$\mathsf{E}[A_y^n] = \overline{Y}; \qquad \mathsf{VAR}[A_Y^n] = \frac{\sigma_Y^2}{n}$$

Note: $\lim_{n\to\infty} \mathsf{VAR}[S] = \infty \qquad \lim_{n\to\infty} \mathsf{VAR}[A_Y^n] = 0$.



The distribution of A_Y^n clusters around $\overline{Y}$, clustering more closely as $n \to \infty$.

Chebyshev: for $\varepsilon > 0$, Pr$\{|A_Y^n - \overline{Y}| \geq \varepsilon\} \leq \frac{\sigma_Y^2}{n\varepsilon^2}$

For any $\varepsilon, \delta > 0$, **large enough** n,

$$\Pr\{|A_Y^n - \overline{Y}| \geq \varepsilon\} \leq \delta$$

ASYMPTOTIC EQUIPARTITION PROPERTY (AEP)

Lete $X_1, X_2, \ldots,$ be output from DMS.

Define log pmf as $w(x) = -\log p_X(x)$.

$w(x)$ maps source symbols into real numbers.

For each j, $W(X_j)$ is a rv; takes value $w(x)$ for $X_j = x$. Note that

$$\mathsf{E}[W(X_j)] = \sum_x p_X(x)[-\log p_X(x)] = H(X)$$

$W(X_1), W(X_2), \ldots$ sequence of iid rv's.

For $X_1 = x_1, X_2 = x_2$, the outcome for $W(X_1)+W(X_2)$ is

$$\begin{aligned} w(x_1) + w(x_2) &= -\log p_X(x_1) - \log p_X(x_2) \\ &= -\log\{p_X(x_1)p_X(x_2)\} \\ &= -\log\{p_{X_1X_2}(x_1x_2)\} = w(x_1x_2) \end{aligned}$$

where $w(x_1x_2)$ is -log pmf of event $X_1X_2 = x_1x_2$

$$W(X_1X_2) = W(X_1) + W(X_2)$$

X_1X_2 is a random symbol in its own right (takes values x_1x_2). $W(X_1X_2)$ is -log pmf of X_1X_2

Probabilities multiply, log pmf's add.

For $\mathbf{X}^n = \mathbf{x}^n$; $\mathbf{x}^n = (x_1, \ldots, x_n)$, the outcome for $W(X_1) + \cdots + W(X_n)$ is

$$\sum_{j=1}^{n} w(x_j) = -\sum_{j=1}^{n} \log p_X(x_j) = -\log p_{\mathbf{X}^n}(\mathbf{x}^n)$$

Sample average of log pmf's is

$$S_W^n = \frac{W(X_1) + \cdots W(X_n)}{n} = \frac{-\log p_{\mathbf{X}^n}(\mathbf{X}^n)}{n}$$

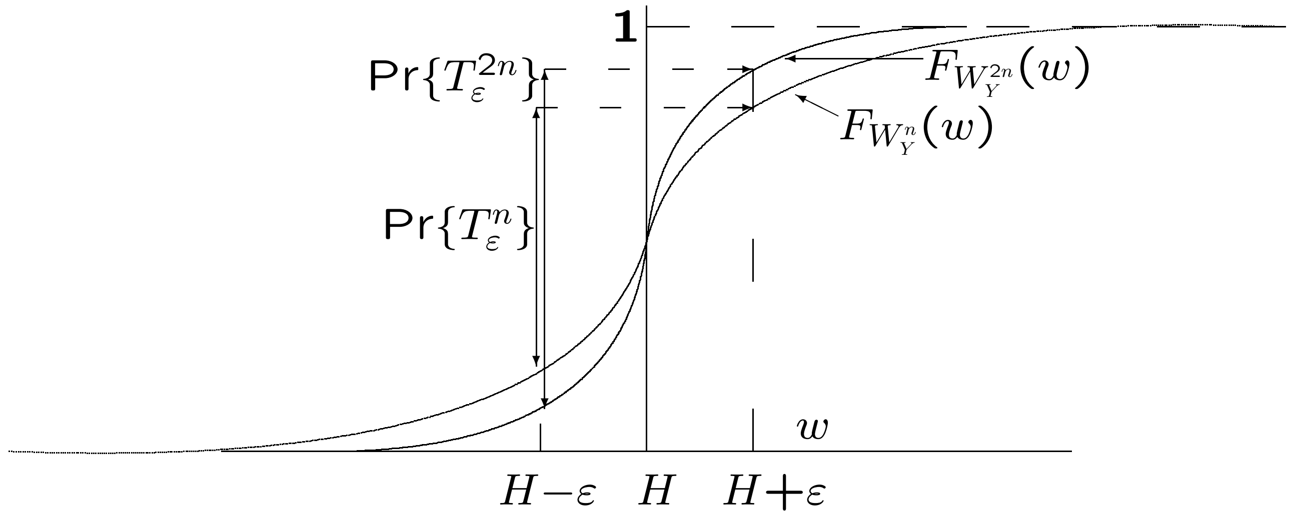
WLLN applies and is

$$\Pr\left(\left| A_W^n - \mathsf{E}[W(X)] \right| \geq \varepsilon \right) \leq \frac{\sigma_W^2}{n\varepsilon^2}$$

$$\Pr\left(\left| \frac{-\log p_{\mathbf{X}^n}(\mathbf{X}^n)}{n} - H(X) \right| \geq \varepsilon \right) \leq \frac{\sigma_W^2}{n\varepsilon^2}.$$

Define typical set as

$$T_\varepsilon^n = \left\{ \mathbf{x}^n : \left| \frac{-\log p_{\mathbf{X}^n}(\mathbf{x}^n)}{n} - H(X) \right| < \varepsilon \right\}$$



As $n \to \infty$, typical set approaches probability 1:

$$\Pr(\mathbf{X}^n \in T_\varepsilon^n) \geq 1 - \frac{\sigma_W^2}{n\varepsilon^2}$$

We can also express T_ε^n as

$$T_\varepsilon^n = \left\{ \mathbf{x}^n : n(H(X)-\varepsilon) < -\log p_{\mathbf{X}^n}(\mathbf{x}^n) < n(H(X)+\varepsilon) \right\}$$

$$T_\varepsilon^n = \left\{ \mathbf{x}^n : \; 2^{-n(H(X)+\varepsilon)} < p_{\mathbf{X}^n}(\mathbf{x}^n) < 2^{-n(H(X)-\varepsilon)} \right\}.$$

Typical elements are approximately equiprobable in the strange sense above.

The complementary, atypical set of strings, satisfy

$$\Pr[(T_\varepsilon^n)^c] \leq \frac{\sigma_W^2}{n\varepsilon^2}$$

For any $\varepsilon, \delta > 0$, large enough n, $\Pr[(T_\varepsilon^n)^c] < \delta$.

For all $\mathbf{x}^n \in T_\varepsilon^n$**,** $p_{\mathbf{X}^n}(\mathbf{x}^n) > 2^{-n[H(X)+\varepsilon]}$**.**

$$1 \geq \sum_{\mathbf{x}^n \in T_\varepsilon^n} p_{\mathbf{X}^n}(\mathbf{x}^n) > |T_\varepsilon^n| \, 2^{-n[H(X)+\varepsilon]}$$

$$|T_\varepsilon^n| < 2^{n[H(X)+\varepsilon]}$$

$$1-\delta \leq \sum_{\mathbf{x}^n \in T_\varepsilon^n} p_{\mathbf{X}^n}(\mathbf{x}^n) < |T_\varepsilon^n| 2^{-n[H(X)-\varepsilon]}$$

$$|T_\varepsilon^n| > (1-\delta) 2^{n[H(X)-\varepsilon]}$$

Summary: $\Pr[(T_\varepsilon^n)^c] \approx 0, \quad |T_\varepsilon^n| \approx 2^{n\mathbf{H}(X)},$

$$p_{\mathbf{X}^n}(\mathbf{x}^n) \approx 2^{-n\mathbf{H}(X)} \quad \textbf{for } \mathbf{x}^n \in T_\varepsilon^n.$$

EXAMPLE

Consider binary DMS with $\Pr[X{=}1] = p_1 < 1/2$**.**

$$\mathbf{H}(X) = -p_1 \log p_1 - p_0 \log(p_0)$$

Consider a string $\mathbf{x}^n$ **with** n_1 **ones and** n_0 **zeros.**

$$p_{\mathbf{X}^n}(\mathbf{x}^n) = p_1^{n_1}\, p_0^{n_0}$$

$$\frac{-\log p_{\mathbf{X}^n}(\mathbf{x}^n)}{n} = -\frac{n_1}{n}\log p_1 - \frac{n_0}{n}\log p_0$$

The typical set T_ε^n **is the set of strings for which**

$$\mathbf{H}(X) \approx \frac{-\log p_{\mathbf{X}^n}(\mathbf{x}^n)}{n} = -\frac{n_1}{n}\log p_1 - \frac{n_0}{n}\log p_0$$

In the typical set, $n_1 \approx p_1 n$**. For this binary case, a string is typical if it has about the right relative frequencies.**

$$\mathbf{H}(X) \approx \frac{-\log p_{\mathbf{X}^n}(\mathbf{x}^n)}{n} = -\frac{n_1}{n}\log p_1 - \frac{n_0}{n}\log p_0$$

The probability of a typical n-tuple is about

$$p_1^{p_1 n} p_0^{p_0 n} = 2^{-n\mathbf{H}(X)}.$$

The number of n-tuples with $p_1 n$ ones is

$$\frac{n!}{(p_1 n)!(p_0 n)!} \approx 2^{n\mathbf{H}(X)}$$

Note that there are 2^n binary strings. Most of them are collectively very improbable.

The most probable strings have almost all zeros, but there aren't enough of them to matter.

Fixed-to-fixed-length source codes

For any $\varepsilon, \delta > 0$, and any large enough n, assign fixed length codeword to each $\mathrm{x}^n \in T_\varepsilon^n$.

Since $|T_\varepsilon^n| < 2^{n[H(X)+\varepsilon]}$, $\quad \overline{L} \leq H(X) + \varepsilon + \frac{1}{n}$.

$$\Pr\{\text{failure}\} \leq \delta.$$

Conversely, take $\overline{L} \leq H(X) - 2\varepsilon$, and n large.

Probability of failure will then be almost 1.

For any $\varepsilon > 0$, the probability of failure will be almost 1 if $\overline{L} \leq H(X) - 2\varepsilon$ and n is large enough:

We can provide codewords for at most $2^{n\mathbf{H}(X)-2\varepsilon n}$ source n-tuples. Typical n-tuples have at most probability $2^{-n\mathbf{H}(X)+\varepsilon n}$.

The aggregate probability of typical n-tuples assigned codewords is at most $2^{-\varepsilon n}$.

The aggregate probability of typical n-tuples not assigned codewords is at least $1 - \delta - 2^{-n\varepsilon}$.

$$\Pr\{\text{failure}\} > 1 - \delta - 2^{-\varepsilon n} \ \rightarrow \ 1$$

General model: Visualize any kind of mapping from the sequence of source symbols X^∞ into a binary sequence Y^∞.

Visualize a decoder that observes encoded bits, one by one. For each n, let D_n be the number of observed bits required to decode X^n (decisions are final).

The rate rv, as a function of n, is D_n/n.

In order for the rate in bpss to be less than H(X) in any meaningful sense, we require that D_n/n be smaller than H(X) with high probability as $n \to \infty$.

Theorem: For a DMS and any coding/decoding technique, let $\varepsilon, \delta > 0$ be arbitrary. Then for large enough n,

$$\Pr\{D_n \leq n[\mathbf{H}(X) - 2\varepsilon]\} \; < \; \delta + 2^{-\varepsilon n}.$$

Proof: For given n, let $m = \lfloor n[\mathbf{H}(X) - 2\varepsilon]\rfloor$. Suppose that x^{n} is decoded upon observation of y^{j} for some $j \leq m$. Only x^{n} can be decoded from y^{m}. There are only 2^m source n-tuples (and thus at most 2^m typical n-tuples) that can be decoded by time m. Previous result applies.

Questions about relevance of AEP and fixed-to-fixed length source codes:

1) Are there important real DMS sources? No, but DMS model provides memory framework.

2) Are fixed-to-fixed codes at very long length practical? No, but view length as product lifetime to interpret bpss.

3) Do fixed-to-fixed codes with rare failures solve queueing issues? No, queueing issues arise only with real-time sources, and discrete sources are rarely real time.

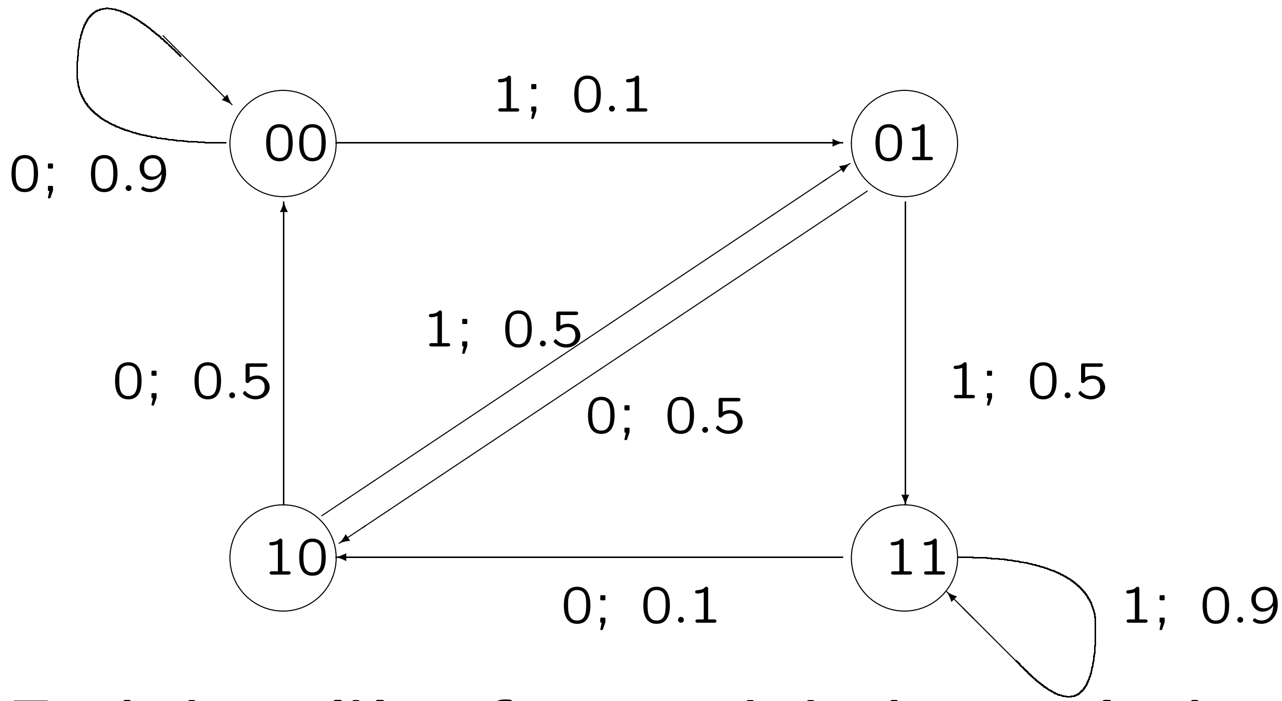
MARKOV SOURCES

A finite state Markov chain is a sequence $S_0, S_1, \ldots$ of discrete cv's from a finite alphabet $\mathcal{S}$ where $q_0(s)$ is a pmf on S_0 and for $n \geq 1$,

$$
\begin{aligned}
Q(s|s') &= \Pr(S_n{=}s|S_{n-1}{=}s') \\
&= \Pr(S_n{=}s|S_{n-1}{=}s', S_{n-2} = s_{n-2} \ldots, S_0{=}s_0)
\end{aligned}
$$

for all choices of $s_{n-2} \ldots, s_0$, We use the states to represent the memory in a discrete source with memory.

Example: Binary source $X_1, X_2, \ldots$; $S_n = (X_{n-1}X_n)$



Each transition from a state has a single and distinct source letter.

Letter specifies new state, new state specifies letter.

Transitions in graph imply positive probability.

A state s is accessible from state s' if graph has a path from $s' \to s$.

The period of s is gcd of path lengths from s back to s.

A finite state Markov chain is ergodic if all states are aperiodic and accessible from all other states.

A Markov source $X_1, X_2, \ldots$ is the sequence of labeled transitions on an ergodic Markov chain.

Ergodic Markov chains have steady state probabilities given by

$$\begin{aligned} q(s) &= \sum_{s' \in \mathcal{S}} q(s') Q(s|s'); \qquad s \in \mathcal{S} && (1) \\ \sum_{s \in \mathcal{S}} q(s) &= 1 \end{aligned}$$

Steady-state probabilities are approached asymptotically from any starting state, i.e., for all $s, s' \in \mathcal{S}$**,**

$$\lim_{n \to \infty} \Pr(S_n{=}s | S_0{=}s') = q(s) \qquad (2)$$

Coding for Markov sources

Simplest approach: use separate prefix-free code for each prior state.

If $S_{n-1}{=}s$, then encode X_n with the prefix-free code for s. The codeword lengths $l(x,s)$ are chosen for the pmf $p(x|s)$.

$$\sum_x 2^{-l(x,s)} \leq 1 \qquad \textbf{for each } s$$

It can be chosen by Huffman algorithm and satisfies

$$\mathbf{H}[X|s] \leq \overline{L}_{min}(s) < \mathbf{H}[X|s] + 1$$

where

$$\mathbf{H}[X|s] = \sum_{x\in\mathcal{X}} -P(x|s) \log P(x|s)$$

If the pmf on S_0 is the steady state pmf, $\{q(s)\}$, then the chain remains in steady state.

$$\mathsf{H}[X|S] \le \overline{L}_{\mathsf{min}} < \mathsf{H}[X|S] + 1, \qquad (3)$$

where

$$\begin{aligned} \overline{L}_{\mathsf{min}} &= \sum_{s \in \mathcal{S}} q(s)\overline{L}_{\mathsf{min}}(s) \qquad \textbf{and} \\ \mathsf{H}[X|S] &= \sum_{s \in \mathcal{S}} q(s)\mathsf{H}[X|s] \end{aligned}$$

The encoder transmits s_0 followed by codeword for x_1 using code for s_0.

This specifies s_1 and x_2 is encoded with code for s_1, etc.

This is prefix free and can be decoded instantaneously.

Conditional Entropy

$\mathbf{H}[X|S]$ for Markov is like $\mathbf{H}[X]$ for DMS.

$$\mathbf{H}[X|S] = \sum_{s\in\mathcal{S}} \sum_{x\in\mathcal{X}} q(s)P(x|s) \log \frac{1}{P(x|s)}$$

Note that

$$\begin{aligned} \mathbf{H}[XS] &= \sum_{s,x} q(s)P(x|s) \log \frac{1}{q(s)P(x|s)} \\ &= \mathbf{H}[S] + \mathbf{H}[X|S] \end{aligned}$$

Recall that

$$\mathbf{H}[XS] \le \mathbf{H}[S] + \mathbf{H}[X]$$

Thus,

$$\mathbf{H}[X|S] \le \mathbf{H}[X]$$

Suppose we use n-to-variable-length codes for each state.

$$\mathsf{H}[S_1, S_2, \dots S_n | S_0] = n\mathsf{H}[X|S]$$

$$\mathsf{H}[X_1, X_2, \dots X_n | S_0] = n\mathsf{H}[X|S]$$

By using n-to-variable length codes,

$$\mathsf{H}[X|S] \leq \overline{L}_{\mathsf{min},n} < \mathsf{H}[X|S] + 1/n$$

Thus, for Markov sources, $\mathsf{H}[X|S]$ is asymptotically achievable.

The AEP also holds for Markov sources.

$\overline{L} \leq \mathsf{H}[X|S] - \varepsilon$ can not be achieved, either in expected length or fixed length, with low probability of failure.