Chapter 4

Measure of Randomness for Stochastic Processes

Po-Ning Chen

Department of Communications Engineering

National Chiao-Tung University

Hsin Chu, Taiwan 30050

Background

- In the previous chapter, it is shown that the sup-entropy rate is indeed the minimum lossless data compression ratio achievable for block codes.
- Hence, to find an optimal block code becomes a well-defined mission since for any source with well-formulated statistical model, the sup-entropy rate can be computed and such quantity can be used as a criterion to evaluate the optimality of the designed block code.
- In the very recent work of Verdú and Han in 1993, they found that, other than the minimum lossless data compression ratio, the sup-entropy rate actually has another operational meaning, which is called *resolvability*.
- In this chapter, we will explore the new concept in details.

Motivation for resolvability

- In simulations of statistical communication systems, generation of random variables by a computer algorithm is very essential.
- The computer usually has an access to a basic random experiment (through pre-defined Application Programing Interface), which generates equally likely random values, such as **rand**() that generates a real number uniformly distributed over $(0, 1)$.
- Conceptually, random variables with *complex* models are more difficult to generate by computers than random variables with *simple* models.
- **Question** is how to quantify the "complexity" of generating a random variables by computers.
- **Possible solution:** One way to define such "complexity" measurement is:

  Definition 4.1 The complexity of generating a random variable is defined as the number of random bits that the most efficient algorithm requires in order to generate the random variable by computers that has an access to equally likely random experiments.

Example

Example 4.2 Consider the generation of the random variable with probability masses $P_X(-1) = 1/4$, $P_X(0) = 1/2$, and $P_X(1) = 1/4$. An algorithm is written as:

```
Flip-a-fair-coin;                       \\ one random bit
If "Head", then output 0;
else
    {
    Flip-a-fair-coin;                   \\ one random bit
    If "Head", then output −1;
    else output 1;
    }
```

- **average-case**: the above algorithm requires 1.5 coin flips;
- **worst-case**: 2 coin flips are necessary.
- Therefore, the complexity measure can take two fundamental forms: *worst-case* or *average-case* over the range of outcomes of the random variables.

Example

- Note that we did not show in the above example that the algorithm is the most efficient one in the sense of using minimum number of random bits; however, it is indeed an optimal algorithm because it achieves the lower bound of the minimum number of random bits. Later, we will show that such bound for average minimum number of random bits required for generating the random variables is the **entropy**, which is exactly 1.5 bits in the above example.

- As for the worse-case bound, a new terminology, *resolution*, will be introduced. As a result, the above algorithm also achieves the lower bound of the worst-case complexity, which is the *resolution* of the random variable.

Notations and definitions regarding to resolvability

Definition 4.3 (M-type) For any positive integer M, a probability distribution P is said to be M-*type* if

$$P(\omega) \in \left\{0, \frac{1}{M}, \frac{2}{M}, \ldots, 1\right\} \quad \text{for all } \omega \in \Omega.$$

Definition 4.4 (resolution of a random variable) The *resolution* $R(X)$ of a random variable X is the minimum $\log M$ such that P_X is M-type. If P_X is not M-type for any integer M, then $R(X) = \infty$.

- If the base of the logarithmic operation is 2, the resolution is measured in *bits*; however, if natural logarithm is taken, *nats* becomes the basic measurement unit of resolution.

Physical meaning

- As revealed previously, a random source needs to be *resolved* (meaning, can be generated by a computer algorithm with access to equal-probable random experiments).
- As anticipated, a random variable with finite *resolution* is resolvable by computer algorithms.
- Yet, it is possible that the resolution of a random variable is infinity.
- A quick example is the random variable X with distribution $P_X(0) = 1/\pi$ and $P_X(1) = 1 - 1/\pi$. (X does not belong to any M-type for finite M.)
- In such case, one can alternatively choose another computer-resolvable random variable, which resembles the true source within some acceptable range, to simulate the original one.
- One criterion that can be used as a measure of resemblance of two random variables is the *variational distance*.
- As for the same example in the above paragraph, choose a random variable $\widetilde{X}$ with distribution $P_{\widetilde{X}}(0) = 1/3$ and $P_{\widetilde{X}}(1) = 2/3$. Then $\|X - \widetilde{X}\| \approx 0.03$, and $\widetilde{X}$ is 3-type, which is computer-resolvable.

Physical meaning

- A program that generates M-type random variable for any M satisfying $\log_2(M)$ being a positive integer is straightforward.
- A program that generates the 3-type $\widetilde{X}$ is as follows (in C language).

```
even = False;
while (1)
  {Flip-a-fair-coin;                          \\ one random bit
  if (Head)
    {if (even==True) { output 0; even=False;}
     else {output 1; even = True;}
    }
  else
    {if (even==True) even=False;
     else even=True;
    }
  }
```

<u>ε-achievable resolution</u>

Definition 4.5 (variational distance) The *variational distance* (or ℓ_1 distance) between tow distributions P and Q defined on common measurable space $(\Omega, \mathcal{F})$ is

$$\|P - Q\| \triangleq \sum_{\omega \in \Omega} |P(\omega) - Q(\omega)|.$$

(Note that an alternative way to formulate the variational distance is:

$$\|P - Q\| = 2 \cdot \sup_{E \in \mathcal{F}} |P(E) - Q(E)| = 2 \sum_{\{x \in \mathcal{X} \,:\, P(x) \geq Q(x)\}} [P(x) - Q(x)].$$

This two definitions are actually equivalent.)

Definition 4.6 (ε-achievable resolution) Fix $\varepsilon \geq 0$. R is an ε-*achievable resolution* for input X if for all $\gamma > 0$, there exists $\tilde{\boldsymbol{X}}$ satisfies

$$R(\widetilde{X}) < R + \gamma \quad \text{and} \quad \|X - \widetilde{X}\| < \varepsilon.$$

- ε-achievable resolution reveals the possibility that one can choose another computer-resolvable random variable whose variational distance to the true source is within an acceptable range, ε.

ε-resolution rate

- Next we define the *ε-achievable resolution rate* for a sequence of random variables, which is an extension of *ε-achievable resolution* defined for single random variable.
- Such extension is analogous to extending *entropy* for a single source to *entropy rate* for a random source sequence.

Definition 4.7 (ε-achievable resolution rate) Fix $\varepsilon \geq 0$ and input $\boldsymbol{X}$. R is an *ε-achievable resolution rate* for input $\boldsymbol{X}$ if for every $\gamma > 0$, there exists $\tilde{\boldsymbol{X}}$ satisfies

$$\frac{1}{n} R(\widetilde{X}^n) < R + \gamma \quad \text{and} \quad \|X^n - \widetilde{X}^n\| < \varepsilon,$$

for all sufficiently large n.

ε-resolvability

Definition 4.8 (ε-resolvability for $\boldsymbol{X}$) Fix $\varepsilon > 0$. The ε-*resolvability* for input $\boldsymbol{X}$, denoted by $S_\varepsilon(\boldsymbol{X})$, is the minimum ε-achievable resolution rate of the same input, i.e.,

$$S_\varepsilon(\boldsymbol{X}) \triangleq \min\Big\{R : (\forall\, \gamma > 0)(\exists \tilde{\boldsymbol{X}} \text{ and } N)(\forall\, n > N) \; \frac{1}{n}R(\widetilde{X}^n) < R + \gamma \text{ and } \|X^n - \widetilde{X}^n\| < \varepsilon\Big\}.$$

- Here, we define $S_\varepsilon(\boldsymbol{X})$ using the “minimum” instead of a more general “infimum” operation is simply because $S_\varepsilon(\boldsymbol{X})$ indeed belongs to the range of the minimum operation, i.e.,

$$S_\varepsilon(\boldsymbol{X}) \in \Big\{R : (\forall\, \gamma > 0)(\exists \tilde{\boldsymbol{X}} \text{ and } N)(\forall\, n > N) \; \frac{1}{n}R(\widetilde{X}^n) < R + \gamma \text{ and } \|X^n - \widetilde{X}^n\| < \varepsilon\Big\}.$$

- Similar convention will be applied throughout the rest of this chapter.

ε-resolvability

Definition 4.9 (resolvability for $\boldsymbol{X}$) The *resolvability* for input $\boldsymbol{X}$, denoted by $S(\boldsymbol{X})$, is

$$S(\boldsymbol{X}) \triangleq \lim_{\varepsilon \to 0} S_\varepsilon(\boldsymbol{X}).$$

- From the definition of ε-resolvability, it is obvious non-increasing in ε. Hence, the resolvability can also be defined using supremum operation as:

$$S(\boldsymbol{X}) \triangleq \sup_{\varepsilon > 0} S_\varepsilon(\boldsymbol{X}).$$

ε-mean-resolvability

- The resolvability is pertinent to the *worse-case* complexity measure for random variables (cf. Example 4.2, and the discussion following it).
- With the entropy function, the information theorists also define the ε-mean-resolvability and mean-resolvability for input $\boldsymbol{X}$, which characterize the *average-case* complexity of random variables.

Definition 4.10 (ε-mean-achievable resolution rate) Fix $\varepsilon \geq 0$. R is an *ε-mean-achievable resolution rate* for input $\boldsymbol{X}$ if for all $\gamma > 0$, there exists $\tilde{\boldsymbol{X}}$ satisfies

$$\frac{1}{n}H(\widetilde{X}^n) < R + \gamma \quad \text{and} \quad \|X^n - \widetilde{X}^n\| < \varepsilon,$$

for all sufficiently large n.

Definition 4.11 (ε-mean-resolvability for $\boldsymbol{X}$) Fix $\varepsilon > 0$. The *ε-mean-resolvability* for input $\boldsymbol{X}$, denoted by $\bar{S}_\varepsilon(\boldsymbol{X})$, is the minimum ε-mean achievable resolution rate for the same input, i.e.,

$$\bar{S}_\varepsilon(\boldsymbol{X}) \triangleq \min\Big\{R \,:\, (\forall\,\gamma > 0)(\exists \tilde{\boldsymbol{X}} \text{ and } N)(\forall\, n > N) \\ \frac{1}{n}H(\widetilde{X}^n) < R + \gamma \text{ and } \|X^n - \widetilde{X}^n\| < \varepsilon\Big\}.$$

ε-mean-resolvability

Definition 4.12 (mean-resolvability for $\boldsymbol{X}$) The *mean-resolvability* for input $\boldsymbol{X}$, denoted by $\bar{S}(\boldsymbol{X})$, is

$$\bar{S}(\boldsymbol{X}) \stackrel{\triangle}{=} \lim_{\varepsilon \to 0} \bar{S}_{\varepsilon}(\boldsymbol{X}) = \sup_{\varepsilon > 0} \bar{S}_{\varepsilon}(\boldsymbol{X}).$$

- The only difference between resolvability and mean-resolvability is that the former employs *resolution* function, while the latter replaces it by *entropy* function.
- Since entropy is the minimum average codeword length for uniquely decodable codes, an explanation for mean-resolvability is that the new random variable $\widetilde{X}$ can be resolvable through realizing the optimal variable-length code for it.
- You can think of the probability mass of each outcome of $\widetilde{X}$ is $2^{-\ell}$ where ℓ is the codeword length of the optimal lossless variable-length code for $\widetilde{X}$. Such probability mass can actually be generated by flipping fair coins ℓ times, and the average number of fair coin flipping for this outcome is indeed $\ell \times 2^{-\ell}$.
- As you may expect, the mean-resolvability is shown to be the *average complexity* of a random variable.

Ope. meanings of resolvability & mean-resolvability

The operational meanings for the resolution and entropy (a new operational meaning for entropy other than the one from source coding theorem) follow the next theorem.

Theorem 4.13 For a single random variable X,

1. the worse-case complexity is lower-bounded by its resolution $R(X)$ [Han and Verdú 1993];
2. the average-case complexity is lower-bounded by its entropy $H(X)$, and is upper-bounded by entropy $H(X)$ plus 2 bits [Knuth and Yao 1976].

Next, we reveal the operational meanings for resolvability and mean-resolvability in source coding. We begin with some useful lemmas that are useful in characterizing the resolvability.

Lemma 4.14 (bound on variational distance) For every $\mu > 0$,

$$\|P - Q\| \leq 2\mu + 2 \cdot P_X \left[x \in \mathcal{X} \ : \ \log \frac{P(x)}{Q(x)} > \mu \right].$$

Ope. meanings of resolvability & mean-resolvability

Proof:

$$
\begin{aligned}
\|P - Q\| &= 2 \sum_{\left\{x \in \mathcal{X} \,:\, P(x) \geq Q(x)\right\}} [P(x) - Q(x)] \\
&= 2 \sum_{\left\{x \in \mathcal{X} \,:\, \log[P(x)/Q(x)] \geq 0\right\}} [P(x) - Q(x)] \\
&= 2 \left(\sum_{\left\{x \in \mathcal{X} \,:\, \log[P(x)/Q(x)] > \mu\right\}} [P(x) - Q(x)] \right. \\
&\qquad \left. + \sum_{\left\{x \in \mathcal{X} \,:\, \mu \geq \log[P(x)/Q(x)] \geq 0\right\}} [P(x) - Q(x)] \right)
\end{aligned}
$$

Ope. meanings of resolvability & mean-resolvability

$$
\begin{aligned}
&\leq 2\left(\sum_{\left\{x\in\mathcal{X}\,:\,\log[P(x)/Q(x)]>\mu\right\}} P(x)\right.\\
&\qquad\left.+\sum_{\left\{x\in\mathcal{X}\,:\,\mu\geq\log[P(x)/Q(x)]\geq 0\right\}} P(x)\left[1-\frac{Q(x)}{P(x)}\right]\right)\\
&\leq 2\left(P\left[x\in\mathcal{X}\;:\;\log\frac{P(x)}{Q(x)}>\mu\right]\right.\\
&\qquad\left.+\sum_{\left\{x\in\mathcal{X}\,:\,\mu\geq\log[P(x)/Q(x)]\geq 0\right\}} P(x)\left[\log\frac{P(x)}{Q(x)}\right]\right)
\end{aligned}
$$

(by fundamental inequality)

$$
\leq 2\left(P\left[x\in\mathcal{X}\;:\;\log\frac{P(x)}{Q(x)}>\mu\right]+\sum_{\left\{x\in\mathcal{X}\,:\,\mu\geq\log[P(x)/Q(x)]\geq 0\right\}} P(x)\cdot\mu\right)
$$

$$
\begin{aligned}
&= 2\left(P\left[x \in \mathcal{X} \;:\; \log\frac{P(x)}{Q(x)} > \mu\right]\right.\\
&\qquad \left.+\mu \cdot P_X\left[x \in \mathcal{X} \;:\; \mu \geq \log\frac{P(x)}{Q(x)} \geq 0\}\right]\right)\\
&= 2\left(P\left[x \in \mathcal{X} \;:\; \log\frac{P(x)}{Q(x)} > \mu\right] + \mu\right).
\end{aligned}
$$

□

Lemma 4.15

$$
P_{\widetilde{X}^n}\left\{x^n \in \mathcal{X}^n \;:\; -\frac{1}{n}\log P_{\widetilde{X}^n}(x^n) \leq \frac{1}{n}R(\widetilde{X}^n)\right\} = 1,
$$

for every n.

Proof: By definition of $R(\widetilde{X}^n)$,

$$
P_{\widetilde{X}^n}(x^n) \geq \exp\{-R(\widetilde{X}^n)\}
$$

for all $x^n \in \mathcal{X}^n$. Hence, for all $x^n \in \mathcal{X}^n$,

$$
-\frac{1}{n}\log P_{\widetilde{X}^n}(x^n) \leq \frac{1}{n}R(\widetilde{X}^n).
$$

The lemma then holds.

□

Theorem 4.16 The resolvability for input $\boldsymbol{X}$ is equal to its sup-entropy rate, i.e.,

$$S(\boldsymbol{X}) = \bar{H}(\boldsymbol{X}).$$

Proof:

1. $S(\boldsymbol{X}) \geq \bar{H}(\boldsymbol{X})$.

   It suffices to show that $S(\boldsymbol{X}) < \bar{H}(\boldsymbol{X})$ contradicts to Lemma 4.15.

   Suppose $S(\boldsymbol{X}) < \bar{H}(\boldsymbol{X})$. Then there exists $\delta > 0$ such that

$$S(\boldsymbol{X}) + \delta < \bar{H}(\boldsymbol{X}).$$

   Let

$$\mathcal{D}_0 \triangleq \left\{ x^n \in \mathcal{X}^n : -\frac{1}{n} \log P_{X^n}(x^n) \geq S(\boldsymbol{X}) + \delta \right\}.$$

   By definition of $\bar{H}(\boldsymbol{X})$,

$$\limsup_{n\to\infty} P_{X^n}(\mathcal{D}_0) > 0.$$

   Therefore, there exists $\alpha > 0$ such that

$$\limsup_{n\to\infty} P_{X^n}(\mathcal{D}_0) > \alpha,$$

   which immediately implies

$$P_{X^n}(\mathcal{D}_0) > \alpha$$

Resolvability

infinitely often in n.

Select $0 < \varepsilon < \min\{\alpha^2, 1\}$ and observe that $S_\varepsilon(\boldsymbol{X}) \le S(\boldsymbol{X})$, we can choose $\widetilde{X}^n$ to satisfy

$$\frac{1}{n}R(\widetilde{X}^n) < S(\boldsymbol{X}) + \frac{\delta}{2} \quad \text{and} \quad \|X^n - \widetilde{X}^n\| < \varepsilon$$

for sufficiently large n. Define

$$\begin{aligned}\mathcal{D}_1 \triangleq \{x^n \in \mathcal{X}^n \ : \ & P_{X^n}(x^n) > 0 \\ & \text{and } \left|P_{X^n}(x^n) - P_{\widetilde{X}^n}(x^n)\right| \le \sqrt{\varepsilon} \cdot P_{X^n}(x^n)\right\}.\end{aligned}$$

Then

$$\begin{aligned}
P_{X^n}(\mathcal{D}_1^c) &= P_{X^n}\left\{x^n \in \mathcal{X}^n \;:\; P_{X^n}(x^n) = 0 \right.\\
&\qquad\qquad \left. \text{or } \left|P_{X^n}(x^n) - P_{\widetilde{X}^n}(x^n)\right| > \sqrt{\varepsilon}\cdot P_{X^n}(x^n)\right\}\\
&\le P_{X^n}\left\{x^n \in \mathcal{X}^n \;:\; P_{X^n}(x^n) = 0\right\}\\
&\quad + P_{X^n}\left\{x^n \in \mathcal{X}^n \;:\; \left|P_{X^n}(x^n) - P_{\widetilde{X}^n}(x^n)\right| > \sqrt{\varepsilon}\cdot P_{X^n}(x^n)\right\}\\
&= P_{X^n}\left\{x^n \in \mathcal{X}^n \;:\; \left|P_{X^n}(x^n) - P_{\widetilde{X}^n}(x^n)\right| > \sqrt{\varepsilon}\cdot P_{X^n}(x^n)\right\}\\
&= \sum_{\left\{x^n \in \mathcal{X}^n \;:\; P_{X^n}(x^n) < (1/\sqrt{\varepsilon})\left|P_{X^n}(x^n) - P_{\widetilde{X}^n}(x^n)\right|\right\}} P_{X^n}(x^n)\\
&\le \sum_{x^n \in \mathcal{X}^n} \frac{1}{\sqrt{\varepsilon}} \left|P_{X^n}(x^n) - P_{\widetilde{X}^n}(x^n)\right|\\
&\le \frac{\varepsilon}{\sqrt{\varepsilon}} = \sqrt{\varepsilon}.
\end{aligned}$$

Consider that

$$\begin{aligned}
P_{X^n}(\mathcal{D}_1 \cap \mathcal{D}_0) &\ge P_{X^n}(\mathcal{D}_0) - P_{X^n}(\mathcal{D}_1^c)\\
&\ge \alpha - \sqrt{\varepsilon} > 0, \qquad (4.3.1)
\end{aligned}$$

which holds infinitely often in n; and every x_0^n in $\mathcal{D}_1 \cap \mathcal{D}_0$ satisfies

$$P_{\widetilde{X}^n}(x_0^n) \ge (1 - \sqrt{\varepsilon}) P_{X^n}(x_0^n)$$

and

$$
\begin{aligned}
-\frac{1}{n}\log P_{\widetilde{X}^n}(x_0^n) &\geq -\frac{1}{n}\log P_{X^n}(x_0^n) + \frac{1}{n}\log\frac{1}{1+\sqrt{\varepsilon}} \\
&\geq (S(\boldsymbol{X})+\delta) + \frac{1}{n}\log\frac{1}{1+\sqrt{\varepsilon}} \\
&\geq S(\boldsymbol{X}) + \frac{\delta}{2},
\end{aligned}
$$

for $n > (2/\delta)\log(1+\sqrt{\varepsilon})$. Therefore, for those n that (4.3.1) holds,

$$
\begin{aligned}
& P_{\widetilde{X}^n}\left\{x^n \in \mathcal{X}^n \ : \ -\frac{1}{n}\log P_{\widetilde{X}^n}(x^n) > \frac{1}{n}R(\widetilde{X}^n)\right\} \\
\geq\ & P_{\widetilde{X}^n}\left\{x^n \in \mathcal{X}^n \ : \ -\frac{1}{n}\log P_{\widetilde{X}^n}(x^n) > S(\boldsymbol{X}) + \frac{\delta}{2}\right\} \\
\geq\ & P_{\widetilde{X}^n}(\mathcal{D}_1 \cap \mathcal{D}_0) \\
\geq\ & (1-\varepsilon^{1/2})P_{X^n}(\mathcal{D}_1 \cap \mathcal{D}_0) \\
>\ & 0,
\end{aligned}
$$

which contradicts to the result of Lemma 4.15.

Resolvability

2. $S(\boldsymbol{X}) \leq \bar{H}(\boldsymbol{X})$.

   It suffices to show the existence of $\tilde{\boldsymbol{X}}$ for arbitrary $\gamma > 0$ such that

$$\lim_{n\to\infty} \|X^n - \widetilde{X}^n\| = 0$$

   and $\widetilde{X}^n$ is an M-type distribution with

$$M = \left\lfloor \exp\left\{ n(\bar{H}(\boldsymbol{X}) + \gamma) \right\} \right\rfloor .$$

   Let $\widetilde{X}^n = \widetilde{X}^n(X^n)$ be uniformly distributed over a set

$$\mathcal{G} \stackrel{\triangle}{=} \{U_j \in \mathcal{X}^n \ : \ j = 1, \ldots, M\}$$

   which drawn randomly (independently) according to P_{X^n}. Define

$$\mathcal{D} \stackrel{\triangle}{=} \left\{ x^n \in \mathcal{X}^n \ : \ -\frac{1}{n} \log P_{X^n}(x^n) > \bar{H}(\boldsymbol{X}) + \gamma + \frac{\mu}{n} \right\}.$$

Resolvability

For each $\mathcal{G}$ chosen, we obtain from Lemma 4.14 that

$$\begin{aligned}
&\|X^n - \widetilde{X}^n\| \\
\leq\ & 2\mu + 2 \cdot P_{\widetilde{X}^n}\left(x^n \in \boldsymbol{\mathcal{X}}^n : \log \frac{P_{\widetilde{X}^n}(x^n)}{P_{X^n}(x^n)} > \mu\right) \\
=\ & 2\mu + 2 \cdot P_{\widetilde{X}^n}\left(x^n \in \mathcal{G}\ :\ \log \frac{1/M}{P_{X^n}(x^n)} > \mu\right) \quad (\text{since } P_{\widetilde{X}^n}(\mathcal{G}^c) = 0) \\
=\ & 2\mu + P_{\widetilde{X}^n}\left\{x^n \in \mathcal{G}\ :\ -\frac{1}{n}\log P_{X^n}(x^n) > \bar{H}(\boldsymbol{X}) + \gamma + \frac{\mu}{n}\right\} \\
=\ & 2\mu + P_{\widetilde{X}^n}(\mathcal{G} \cap \mathcal{D}) \\
=\ & 2\mu + \frac{1}{M}\left|\mathcal{G} \cap \mathcal{D}\right|.
\end{aligned}$$

Since $\mathcal{G}$ is chosen randomly, we can take the expectation values (w.r.t. the random $\mathcal{G}$) of the above inequality to obtain:

$$E_{\mathcal{G}}\left[\|X^n - \widetilde{X}^n\|\right] \ \leq\ 2\mu + \frac{1}{M} E_{\mathcal{G}}\left[|\mathcal{G} \cap \mathcal{D}|\right].$$

Observe that each U_j is either in $\mathcal{D}$ or not in $\mathcal{D}$, and will contribute weight $1/M$ when it is in $\mathcal{D}$. From the i.i.d. assumption of $\{U_j\}_{j=1}^M$, we can then evaluate $(1/M)E_{\mathcal{G}}[|\mathcal{G}\cap\mathcal{D}|]$ by

$$\begin{aligned}
&\frac{1}{M}E_{\mathcal{G}}[|\mathcal{G}\cap\mathcal{D}|]\\
=\ & P_{X^n}^{M}[\mathcal{D}] + \frac{M-1}{M}P_{X^n}^{M-1}[\mathcal{D}]P_{X^n}[\mathcal{D}^c] + \cdots + \frac{1}{M}P_{X^n}[\mathcal{D}]P_{X^n}^{M-1}[\mathcal{D}^c]\\
=\ & \frac{1}{M}\Big(MP_{X^n}^{M}[\mathcal{D}] + (M-1)P_{X^n}^{M-1}[\mathcal{D}]P_{X^n}[\mathcal{D}^c]\\
&\qquad + \cdots + P_{X^n}[\mathcal{D}]P_{X^n}^{M-1}[\mathcal{D}^c]\Big)\\
=\ & \frac{1}{M}\left(MP_{X^n}[\mathcal{D}]\right)\\
=\ & P_{X^n}[\mathcal{D}].
\end{aligned}$$

Hence,

$$\limsup_{n\to\infty} E_{\mathcal{G}}\left[\|X^n - \widetilde{X}^n\|\right] \le 2\mu + \limsup_{n\to\infty} P_{X^n}[\mathcal{D}] = 2\mu,$$

which implies

$$\limsup_{n\to\infty} E_{\mathcal{G}}\left[\|X^n - \widetilde{X}^n\|\right] = 0 \tag{4.3.2}$$

since μ can be chosen arbitrarily small. (4.3.2) therefore guarantees the existence of the desired $\tilde{\boldsymbol{X}}$.

□

Theorem 4.17 For any $\boldsymbol{X}$,

$$\bar{S}(\boldsymbol{X}) = \limsup_{n\to\infty} \frac{1}{n} H(X^n).$$

Proof:

1. $\bar{S}(\boldsymbol{X}) \le \limsup_{n\to\infty}(1/n)H(X^n)$.

   It suffices to prove that $\bar{S}_\varepsilon(\boldsymbol{X}) \le \limsup_{n\to\infty}(1/n)H(X^n)$ for every $\varepsilon > 0$. This is equivalent to show that for all $\gamma > 0$, there exists $\tilde{\boldsymbol{X}}$ such that

   $$\frac{1}{n} H(\widetilde{X}^n) < \limsup_{n\to\infty} \frac{1}{n} H(X^n) + \gamma \quad \text{and} \quad \|X^n - \widetilde{X}^n\| < \varepsilon$$

   for sufficiently large n. This can be trivially achieved by letting $\tilde{\boldsymbol{X}} = \boldsymbol{X}$, since for sufficiently many n,

   $$\frac{1}{n} H(X^n) < \limsup_{n\to\infty} \frac{1}{n} H(X^n) + \gamma \quad \text{and} \quad \|X^n - X^n\| = 0.$$

2. $\bar{S}(\boldsymbol{X}) \ge \limsup_{n\to\infty}(1/n)H(X^n)$.

   Observe that $\bar{S}(\boldsymbol{X}) \ge \bar{S}_\varepsilon(\boldsymbol{X})$ for any $0 < \varepsilon < 1/2$. Then for any $\gamma > 0$ and all sufficiently large n, there exists $\widetilde{X}^n$ such that

   $$\frac{1}{n} H(\widetilde{X}^n) < \bar{S}(\boldsymbol{X}) + \gamma \tag{4.3.3}$$

and

$$\|X^n - \widetilde{X}^n\| < \varepsilon.$$

Using the fact [1, pp. 33] that $\|X^n - \widetilde{X}^n\| \le \varepsilon \le 1/2$ implies

$$|H(X^n) - H(\widetilde{X}^n)| \le \varepsilon \log \frac{|\mathcal{X}|^n}{\varepsilon},$$

and (4.3.3), we obtain

$$\frac{1}{n}H(X^n) - \varepsilon \log |\mathcal{X}| + \frac{1}{n}\varepsilon \log \varepsilon \le \frac{1}{n}H(\widetilde{X}^n) < \bar{S}(\boldsymbol{X}) + \gamma,$$

which implies that

$$\limsup_{n\to\infty} \frac{1}{n}H(X^n) - \varepsilon \log |\mathcal{X}| < \bar{S}(\boldsymbol{X}) + \gamma.$$

Since ε and γ can be taken arbitrarily small, we have

$$\bar{S}(\boldsymbol{X}) \ge \limsup_{n\to\infty} \frac{1}{n}H(X^n).$$

□

Resolvability and source coding

- In the previous chapter, we have proved that the lossless data compression rate for block codes is lower bounded by $\bar{H}(\boldsymbol{X})$.
- We also show that $\bar{H}(\boldsymbol{X})$ is also the resolvability for source $\boldsymbol{X}$.
- We can therefore conclude that resolvability is equal to the minimum lossless data compression rate for block codes.
- The key to the Shannon's source coding theorem is actually the existence of a set $\mathcal{A}_n = \{x_1^n, x_2^n, \ldots, x_M^n\}$ with $M \approx e^{nH(X)}$ and $P_{X^n}(\mathcal{A}_n^c) \to 0$.
- Thus, if we can find such *typical set*, the Shannon's source coding theorem for *block codes* can actually be generalized to more general sources, such as non-stationary sources.
- Furthermore, extension of the theorems to codes of some specific types becomes feasible.

Definition 4.18 (minimum ε-source compression rate for fixed-length codes) R is the *ε-source compression rate* for fixed-length codes if there exists a sequence of sets $\{\mathcal{A}_n\}_{n=1}^{\infty}$ with $\mathcal{A}_n \subset \mathcal{X}^n$ such that

$$\limsup_{n\to\infty} \frac{1}{n}\log|\mathcal{A}_n| \leq R \quad \text{and} \quad \limsup_{n\to\infty} P_{X^n}[\mathcal{A}_n^c] \leq \varepsilon.$$

$T_\varepsilon(\boldsymbol{X})$ is the minimum of all such rates.

Note that the definition of $T_\varepsilon(\boldsymbol{X})$ is equivalent to the one in Definition 3.2.

Definition 4.19 (minimum source compression rate for fixed-length codes) $T(\boldsymbol{X})$ represents the *minimum source compression rate* for fixed-length codes, which is defined as:

$$T(\boldsymbol{X}) \stackrel{\triangle}{=} \lim_{\varepsilon\to 0} T_\varepsilon(\boldsymbol{X}).$$

Definition 4.20 (minimum source compression rate for variable-length codes) R is an achievable source compression rate for variable-length codes if there exists a sequence of error-free prefix codes $\{\mathscr{C}_n\}_{n=1}^{\infty}$ such that

$$\limsup_{n\to\infty} \frac{1}{n}\ell_n \leq R$$

where ℓ_n is the average codeword length of $\mathscr{C}_n$. $\bar{T}(\boldsymbol{X})$ is the minimum of all such rates.

Resolvability and source coding

- Recall that for a single source, the measure of its uncertainty is entropy. Although the entropy can also be used to characterize the *overall* uncertainty of a random sequence $\boldsymbol{X}$, the source coding however concerns more on the "average" entropy of it.

- So far, we have seen four expressions of "average" entropy:

$$\limsup_{n\to\infty}\frac{1}{n}H(X^n)\stackrel{\triangle}{=}\limsup_{n\to\infty}\frac{1}{n}\sum_{x^n\in\mathcal{X}^n}-P_{X^n}(x^n)\log P_{X^n}(x^n);$$

$$\liminf_{n\to\infty}\frac{1}{n}H(X^n)\stackrel{\triangle}{=}\liminf_{n\to\infty}\frac{1}{n}\sum_{x^n\in\mathcal{X}^n}-P_{X^n}(x^n)\log P_{X^n}(x^n);$$

$$\bar{H}(\boldsymbol{X})\stackrel{\triangle}{=}\inf_{\beta\in\Re}\left\{\beta\ :\ \limsup_{n\to\infty}P_{X^n}\left[-\frac{1}{n}\log P_{X^n}(X^n)>\beta\right]=0\right\};$$

$$\underline{H}(\boldsymbol{X})\stackrel{\triangle}{=}\sup_{\alpha\in\Re}\left\{\alpha\ :\ \limsup_{n\to\infty}P_{X^n}\left[-\frac{1}{n}\log P_{X^n}(X^n)<\alpha\right]=0\right\}.$$

- If

$$\lim_{n\to\infty}\frac{1}{n}H(X^n)=\limsup_{n\to\infty}\frac{1}{n}H(X^n)=\liminf_{n\to\infty}\frac{1}{n}H(X^n),$$

then $\lim_{n\to\infty}(1/n)H(X^n)$ is named the entropy rate of the source.

Resolvability and source coding

- $\bar{H}(\boldsymbol{X})$ and $\underline{H}(\boldsymbol{X})$ are called the sup-entropy rate and inf-entropy rate, which were already introduced.
- Next we will prove that $T(\boldsymbol{X}) = S(\boldsymbol{X}) = \bar{H}(\boldsymbol{X})$ and $\bar{T}(\boldsymbol{X}) = \bar{S}(\boldsymbol{X}) = \limsup_{n\to\infty}(1/n)H(X^n)$ for a source $\boldsymbol{X}$.
- The operational characterization of $\liminf_{n\to\infty}(1/n)H(X^n)$ and $\underline{H}(\boldsymbol{X})$ will be introduced in Chapter 6.

Theorem 4.21 (equality of resolvability and minimum source coding rate for fixed-length codes)

$$T(\boldsymbol{X}) = S(\boldsymbol{X}) = \bar{H}(\boldsymbol{X}).$$

Proof: Equality of $S(\boldsymbol{X})$ and $\bar{H}(\boldsymbol{X})$ is already given in Theorem 4.16. Also, $T(\boldsymbol{X}) = \bar{H}(\boldsymbol{X})$ can be obtained from Theorem 3.5 by letting $\varepsilon = 0$. Here, we provide an alternative proof for $T(\boldsymbol{X}) = S(\boldsymbol{X})$.

1. $T(\boldsymbol{X}) \leq S(\boldsymbol{X})$.

   If we can show that, for any ε fixed, $T_\varepsilon(\boldsymbol{X}) \leq S_{2\varepsilon}(\boldsymbol{X})$, then the proof is completed. This claim is proved as follows.

Resolvability and source coding

- By definition of $S_{2\varepsilon}(\boldsymbol{X})$, we know that for any $\gamma > 0$, there exists $\tilde{\boldsymbol{X}}$ and N such that for $n > N$,

$$\frac{1}{n}R(\widetilde{X}^n) < S_{2\varepsilon}(\boldsymbol{X}) + \gamma \quad \text{and} \quad \|X^n - \widetilde{X}^n\| < 2\varepsilon.$$

- Let $\mathcal{A}_n \stackrel{\triangle}{=} \left\{x^n \ : \ P_{\widetilde{X}^n}(x^n) > 0\right\}$. Since $(1/n)R(\widetilde{X}^n) < S_{2\varepsilon}(\boldsymbol{X}) + \gamma$,

$$|\mathcal{A}_n| \le \exp\{R(\widetilde{X}^n)\} < \exp\{n(S_{2\varepsilon}(\boldsymbol{X}) + \gamma)\}.$$

  Therefore,

$$\limsup_{n\to\infty} \frac{1}{n} \log |\mathcal{A}_n| \le S_{2\varepsilon}(\boldsymbol{X}) + \gamma.$$

- Also,

$$\begin{aligned} 2\varepsilon > \|X^n - \widetilde{X}^n\| &= 2 \sup_{E\subset\mathcal{X}^n} |P_{X^n}(E) - P_{\widetilde{X}^n}(E)| \\ &\ge 2|P_{X^n}(\mathcal{A}_n^c) - P_{\widetilde{X}^n}(\mathcal{A}_n^c)| \\ &= 2P_{X^n}(\mathcal{A}_n^c), \ (\text{since} P_{\widetilde{X}^n}(\mathcal{A}_n^c) = 0). \end{aligned}$$

  Hence, $\limsup_{n\to\infty} P_{X^n}(\mathcal{A}_n^c) \le \varepsilon$.

- Since $S_{2\varepsilon}(\boldsymbol{X}) + \gamma$ is just one of the rates that satisfy the conditions of the minimum ε-source compression rate, and $T_\varepsilon(\boldsymbol{X})$ is the smallest one of such rates,

$$T_\varepsilon(\boldsymbol{X}) \le S_{2\varepsilon}(\boldsymbol{X}) + \gamma \text{ for any } \gamma > 0.$$

2. $T(\boldsymbol{X}) \geq S(\boldsymbol{X})$.

   Similarly, if we can show that, for any ε fixed, $T_\varepsilon(\boldsymbol{X}) \geq S_{3\varepsilon}(\boldsymbol{X})$, then the proof is completed. This claim can be proved as follows.

   - Fix $\alpha > 0$. By definition of $T_\varepsilon(\boldsymbol{X})$, we know that for any $\gamma > 0$, there exists N and a sequence of sets $\{\mathcal{A}_n\}_{n=1}^\infty$ such that for $n > N$,

   $$\frac{1}{n}\log|\mathcal{A}_n| < T_\varepsilon(\boldsymbol{X}) + \gamma \quad \text{and} \quad P_{X^n}(\mathcal{A}_n^c) < \varepsilon + \alpha.$$

   - Choose M_n to satisfy

   $$\exp\{n(T_\varepsilon(\boldsymbol{X}) + 2\gamma)\} \leq M_n \leq \exp\{n(T_\varepsilon(\boldsymbol{X}) + 3\gamma)\}.$$

   Also select one element x_0^n from $\mathcal{A}_n^c$. Define a new random variable $\widetilde{X}^n$ as follows:

   $$P_{\widetilde{X}^n}(x^n) = \begin{cases} 0, & \text{if } x^n \notin \{x_0^n\} \cup \mathcal{A}_n; \\ \dfrac{k(x^n)}{M_n}, & \text{if } x^n \in \{x_0^n\} \cup \mathcal{A}_n, \end{cases}$$

   where

   $$k(x^n) \stackrel{\triangle}{=} \begin{cases} \lceil M_n P_{X^n}(x^n) \rceil, & \text{if } x^n \in \mathcal{A}_n; \\ M_n - \displaystyle\sum_{x^n \in \mathcal{A}_n} k(x^n), & \text{if } x^n = x_0^n. \end{cases}$$

   It can then be easily verified that $\widetilde{X}^n$ satisfies the next four properties:

Resolvability and source coding

(a) $\widetilde{X}^n$ is M_n-type;

(b) $P_{\widetilde{X}^n}(x_0^n) \le P_{X^n}(\mathcal{A}_n^c) < \varepsilon + \alpha$, since $x_0^n \in \mathcal{A}_n^c$;

(c) for all $x^n \in \mathcal{A}_n$,

$$\left|P_{\widetilde{X}^n}(x^n) - P_{X^n}(x^n)\right| = \frac{\lceil M_n P_{X^n}(x^n)\rceil}{M_n} - P_{X^n}(x^n) \le \frac{1}{M_n}.$$

(d) $P_{\widetilde{X}^n}(\mathcal{A}_n) + P_{\widetilde{X}^n}(x_0^n) = 1$.

- Consequently,

$$\frac{1}{n}R(\widetilde{X}^n) \le T_\varepsilon(\boldsymbol{X}) + 3\gamma,$$

and

$$
\begin{aligned}
\|X^n - \widetilde{X}^n\| &= \sum_{x^n \in \mathcal{A}_n} \left|P_{\widetilde{X}^n}(x^n) - P_{X^n}(x^n)\right| + \left|P_{\widetilde{X}^n}(x_0^n) - P_{X^n}(x_0^n)\right| \\
&\quad + \sum_{x^n \in \mathcal{A}_n^c - \{x_0^n\}} \left|P_{\widetilde{X}^n}(x^n) - P_{X^n}(x^n)\right| \\
&\le \sum_{x^n \in \mathcal{A}_n} \left|P_{\widetilde{X}^n}(x^n) - P_{X^n}(x^n)\right| + \left[P_{\widetilde{X}^n}(x_0^n) + P_{X^n}(x_0^n)\right] \\
&\quad + \sum_{x^n \in \mathcal{A}_n^c - \{x_0^n\}} \left|P_{\widetilde{X}^n}(x^n) - P_{X^n}(x^n)\right| \\
&\le \sum_{x^n \in \mathcal{A}_n} \frac{1}{M_n} + P_{\widetilde{X}^n}(x_0^n) + P_{X^n}(x_0^n) + \sum_{x^n \in \mathcal{A}_n^c - \{x_0^n\}} P_{X^n}(x^n) \\
&= \frac{|\mathcal{A}_n|}{M_n} + P_{\widetilde{X}^n}(x_0^n) + \sum_{x^n \in \mathcal{A}_n^c} P_{X^n}(x^n) \\
&\le \frac{\exp\{n(T_\varepsilon(\boldsymbol{X}) + \gamma)\}}{\exp\{n(T_\varepsilon(\boldsymbol{X}) + 2\gamma)\}} + (\varepsilon + \alpha) + P_{X^n}(\mathcal{A}_n^c) \\
&\le e^{-n\gamma} + (\varepsilon + \alpha) + (\varepsilon + \alpha) \\
&\le 3(\varepsilon + \alpha), \text{ for } n \ge -\log(\varepsilon + \alpha)/\gamma.
\end{aligned}
$$

Resolvability and source coding

- Since $T_\varepsilon(\boldsymbol{X})$ is just one of the rates that satisfy the conditions of $3(\varepsilon+\alpha)$-resolvability, and $S_{3(\varepsilon+\alpha)}(\boldsymbol{X})$ is the smallest one of such quantities,

$$S_{3(\varepsilon+\alpha)}(\boldsymbol{X}) \le T_\varepsilon(\boldsymbol{X}).$$

  The proof is completed by noting that α can be made arbitrarily small.

$\square$

This theorem tells us that the minimum source compression ratio for fixed-length code is the resolvability, which in turn is equal to the sup-entropy rate.

Theorem 4.22 (equality of mean-resolvability and minimum source coding rate for variable-length codes)

$$\bar{T}(\boldsymbol{X}) = \bar{S}(\boldsymbol{X}) = \limsup_{n\to\infty} \frac{1}{n} H(X^n).$$

Proof: Equality of $\bar{S}(\boldsymbol{X})$ and $\limsup_{n\to\infty}(1/n)H(X^n)$ is already given in Theorem 4.17.

1. $\bar{S}(\boldsymbol{X}) \le \bar{T}(\boldsymbol{X})$.

   Definition 4.20 states that there exists, for all $\gamma > 0$ and all sufficiently large n, an error-free variable-length code whose average codeword length ℓ_n satisfies

$$\frac{1}{n}\ell_n < \bar{T}(\boldsymbol{X}) + \gamma.$$

   Moreover, the fundamental source coding lower bound for a uniquely decodable code (cf. Theorem 4.18 of Volume I of the lecture notes) is

$$H(X^n) \le \ell_n.$$

   Thus, by letting $\tilde{\boldsymbol{X}} = \boldsymbol{X}$, we obtain $\|X^n - \widetilde{X}^n\| = 0$ and

$$\frac{1}{n}H(\widetilde{X}^n) = \frac{1}{n}H(X^n) < \bar{T}(\boldsymbol{X}) + \gamma,$$

Mean-resolvability and source coding

which concludes that $\bar{T}(\boldsymbol{X})$ is an ε-achievable mean-resolution rate of $\boldsymbol{X}$ for any $\varepsilon > 0$, i.e.,

$$\bar{S}(\boldsymbol{X}) = \lim_{\varepsilon \to 0} \bar{S}_\varepsilon(\boldsymbol{X}) \le \bar{T}(\boldsymbol{X}).$$

2. $\bar{T}(\boldsymbol{X}) \le \bar{S}(\boldsymbol{X})$.

   Observe that $\bar{S}_\varepsilon(\boldsymbol{X}) \le \bar{S}(\boldsymbol{X})$ for $0 < \varepsilon < 1/2$. Hence, by taking γ satisfying $2\varepsilon \log |\mathcal{X}| > \gamma > \varepsilon \log |\mathcal{X}|$ and for all sufficiently large n, there exists $\widetilde{X}^n$ such that

$$\frac{1}{n} H(\widetilde{X}^n) < \bar{S}(\boldsymbol{X}) + \gamma$$

   and

$$\|X^n - \widetilde{X}^n\| < \varepsilon. \tag{4.4.4}$$

   On the other hand, Theorem 4.22 of Volume I of the lecture notes proves the existence of an error-free prefix code for X^n with average codeword length ℓ_n satisfies

$$\ell_n \le H(X^n) + 1 \text{ (bits)}.$$

   By the fact [1, pp. 33] that $\|X^n - \widetilde{X}^n\| \le \varepsilon \le 1/2$ implies

$$|H(X^n) - H(\widetilde{X}^n)| \le \varepsilon \log_2 \frac{|\mathcal{X}|^n}{\varepsilon},$$

and (4.4.4), we obtain

$$\begin{aligned}\frac{1}{n}\ell_n &\le \frac{1}{n}H(X^n)+\frac{1}{n}\\ &\le \frac{1}{n}H(\widetilde{X}^n)+\varepsilon\log_2|\mathcal{X}|-\frac{1}{n}\varepsilon\log_2\varepsilon+\frac{1}{n}\\ &\le \bar{S}(\boldsymbol{X})+\gamma+\varepsilon\log_2|\mathcal{X}|-\frac{1}{n}\varepsilon\log_2\varepsilon+\frac{1}{n}\\ &\le \bar{S}(\boldsymbol{X})+2\gamma,\end{aligned}$$

if $n > (1-\varepsilon\log_2\varepsilon)/(\gamma-\varepsilon\log_2|\mathcal{X}|)$. Since γ can be made arbitrarily small, $\bar{S}(\boldsymbol{X})$ is an achievable source compression rate for variable-length codes; and hence,

$$\bar{T}(\boldsymbol{X}) \le \bar{S}(\boldsymbol{X}).$$

□

Again, the above theorem tells us that the minimum source compression ratio for variable-length code is the mean-resolvability, and the mean-resolvability is exactly $\limsup_{n\to\infty}(1/n)H(X^n)$.

Discussions

- Note that $\limsup_{n\to\infty}(1/n)H(X^n) \leq \bar{H}(\boldsymbol{X})$, which follows straightforwardly by the fact that the mean of the random variable $-(1/n)\log P_{X^n}(X^n)$ is no greater than its right margin of the support.
- Also note that for stationary-ergodic source, all these quantities are equal, i.e.,

$$T(\boldsymbol{X}) = S(\boldsymbol{X}) = \bar{H}(\boldsymbol{X}) = \bar{T}(\boldsymbol{X}) = \bar{S}(\boldsymbol{X}) = \limsup_{n\to\infty} \frac{1}{n} H(X^n).$$

Example

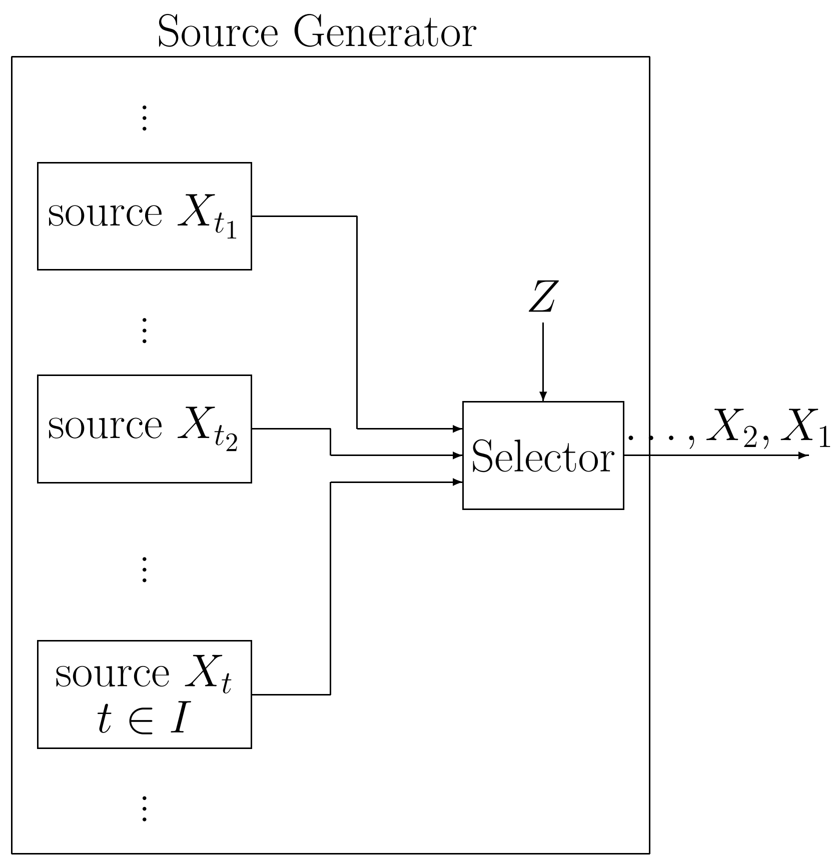
Example 4.23 Consider a binary random source $X_1, X_2, \ldots$ where $\{X_i\}_{i=1}^{\infty}$ are independent random variables with individual distribution

$$P_{X_i}(0) = Z_i \text{ and } P_{X_i}(1) = 1 - Z_i,$$

where $\{Z_i\}_{i=1}^{\infty}$ are pair-wise independent with common uniform marginal distribution over $(0, 1)$.

You may imagine that the source is formed by selecting from infinitely many binary number generators. The selecting process Z is independent for each time instance.

Example



Source generator: $\{X_t\}_{t \in I}$ ($I = (0, 1)$) is an independent random process with $P_{X_t}(0) = 1 - P_{X_t}(1) = t$, and is also independent of the selector Z, where X_t is outputted if $Z = t$. Source generator of each time instance is independent temporally.

Example

- It can be shown that such source is not stationary.
- Nevertheless, by means of similar argument as AEP theorem, we can show that:

$$-\frac{\log P_X(X_1) + \log P_X(X_2) + \cdots + \log P_X(X_n)}{n} \rightarrow h_b(Z) \text{ in probability,}$$

where $h_b(a) \triangleq -a\log_2(a) - (1-a)\log_2(1-a)$ is the binary entropy function.

- To compute the ultimate average entropy rate in terms of the random variable $h_b(Z)$, it requires that

$$-\frac{\log P_X(X_1) + \log P_X(X_2) + \cdots + \log P_X(X_n)}{n} \rightarrow h_b(Z) \text{ in mean,}$$

which is a stronger result than convergence in probability.

Example

- With the fundamental properties for convergence, convergence-in-probability implies convergence-in-mean provided the sequence of random variables is uniformly integrable, which is true for $-(1/n)\sum_{i=1}^{n}\log P_X(X_i)$:

$$\begin{aligned}
&\sup_{n>0} E\left[\left|\frac{1}{n}\sum_{i=1}^{n}\log P_X(X_i)\right|\right] \\
\leq\ & \sup_{n>0}\frac{1}{n}\sum_{i=1}^{n} E\left[|\log P_X(X_i)|\right] \\
=\ & \sup_{n>0} E\left[|\log P_X(X)|\right], \text{because of i.i.d. of } \{X_i\}_{i=1}^{n} \\
=\ & E\left[|\log P_X(X)|\right] \\
=\ & E\left[E\left(|\log P_X(X)| \,\middle|\, Z\right)\right] \\
=\ & \int_0^1 E\left(|\log P_X(X)| \,\middle|\, Z=z\right) dz \\
=\ & \int_0^1 \big(z|\log(z)| + (1-z)|\log(1-z)|\big) dz \\
\leq\ & \int_0^1 \log(2) dz = \log(2).
\end{aligned}$$

Example

- We therefore have:

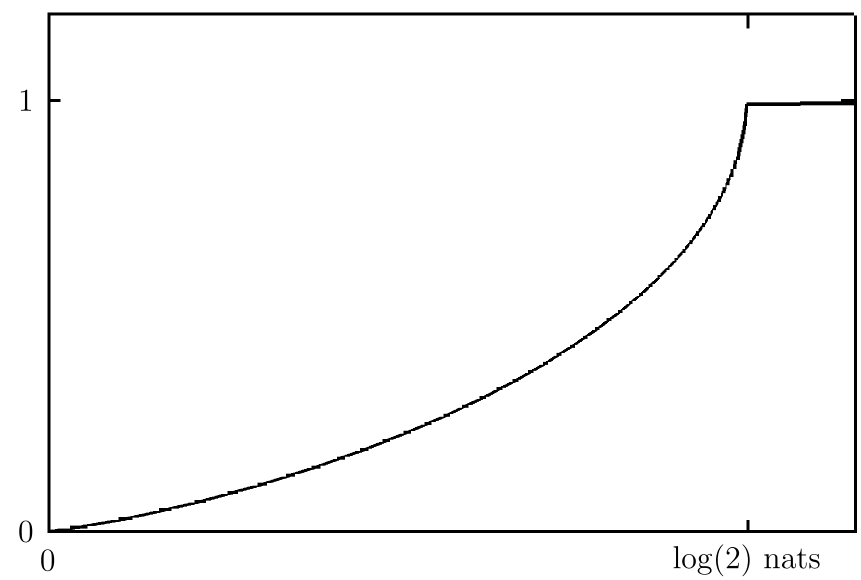
$$\begin{aligned}&\left|E\left[-\frac{1}{n}\log P_{X^n}(X^n)\right]-E[h_b(Z)]\right|\\ \le\ & E\left[\left|-\frac{1}{n}\log P_{X^n}(X^n)-h_b(Z)\right|\right]\to 0.\end{aligned}$$

- Consequently,

$$\begin{aligned}\limsup_{n\to\infty}\frac{1}{n}H(X^n) &= E[h_b(Z)]\\ &= \int_0^1 P_Z[h_b(Z)>t]dt\\ &= 0.596046 \text{ nats} \ \text{ or } \ 0.859912 \text{ bits}.\end{aligned}$$

- However, it can be shown that the ultimate CDF of $-(1/n)\log P_{X^n}(X^n)$ is $P_Z[h_b(Z) \le t]$ for $t \in [0, \log(2)]$. The sup-entropy rate of $\boldsymbol{X}$ should be $\log(2)$ nats or 1 bit (which is the right-margin of the ultimate CDF of $-(1/n)\log P_{X^n}(X^n)$).
- Hence, for this unstationary source, the minimum average codeword length for fixed-length codes and variable-length codes are different, which are 0.859912 bit and 1 bit, respectively.

Example



The ultimate CDF of $-(1/n)\log P_{X^n}(X^n)$: $Pr\{h_b(Z) \le t\}$.