Logistics

Class Road Map - IT-I

- L1 (9/26): Overview, Communications, Information, Entropy
- L2 (10/1): Props. Entropy, Mutual Information,
- L5 (10/10): AEP, Compression
- L6 (10/15): Compression, Method of Types,
- L8
- L9
- L10
- L11
- L12
- L13
- L14
- L15
- L16
- L17
- L18
- L19

Finals Week: December 12th–16th.
Cumulative Outstanding Reading

- Read chapters 1 and 2 in our book (Cover & Thomas, “Information Theory”) (including Fano’s inequality).
- Chapter 3 in our book (Cover & Thomas, “Information Theory”).
- Section 11.1 (method of types).

Homework

- Homework 2 is posted on our assignment dropbox (https://canvas.uw.edu/courses/847774/assignments), due tonight, Tuesday, Oct 15th, at 11:45pm.
Office hours, every week, Tuesdays 4:30-5:30pm. Can also reach me at that time via a canvas conference.

**Division of set of all sequences into type classes**

- $\mathcal{P}_n = \{P_1, P_2, \ldots, P_{|P_n|}\}$ is the set of all types,
- Thus, $\bigcup_{P \in \mathcal{P}_n} T(P) = \mathcal{X}^n$.
- The space of all sequences.
Division of set of all sequences into type classes

- \( P_n = \{ P_1, P_2, \ldots, P_{|P_n|} \} \) is the set of all types,
- Thus, \( \bigcup_{P \in P_n} T(P) = \mathcal{X}^n \).
- The space of all sequences.

Bound on number of type classes

**Proposition 7.2.2**

\[ |P_n| \leq (n + 1)^{|\mathcal{X}|} \tag{7.19} \]

**Proof.**

- Note that numerator of each entry of a type may take on at most \((n + 1)\) possible values,
- And there are \( |\mathcal{X}| \) numerators.
- The numerator values interact (they must sum to \( n \)) but we can upper bound, pretending no interaction, leading to the upper bound.
- Key point: there are, thus, only a polynomial in \( n \) number of types of sequences of length \( n \).
- However, \( \exists \) an exponential number of sequences of length \( n \), \( |\mathcal{X}|^n \).
Probability depends only on the type

**Theorem 7.2.2**

- Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim Q(x)$,
- with extension $Q^n(x_{1:n}) = \prod_i Q(x_i)$, with $Q$ otherwise arbitrary.
- The probability of the sequence depends only on the type
- restated, the probability is “independent” of the sequence given the type and $Q$
- That is

$$Q^n(x_{1:n}) = 2^{-n[H(P_{x_{1:n}}) + D(P_{x_{1:n}} || Q)]} \quad (7.19)$$

- So, probability doesn’t depend on the sequence, once we are given the type
- Compare with sufficient statistics
- all sequences with the same type have the same probability.

Type class with highest probability

- What type class has the highest probability, when the generating distribution is $P \in \mathcal{P}_n$?
- Consider AEP, the typical sequences are the the ones closest to the real distribution, and they have all the probability.
- Thus, we’ll guess that $T(P)$ has the highest probability under distribution $P$.
- I.e., our lemma becomes

**Lemma 7.2.2**

for $P \in \mathcal{P}_n$, then $T(P)$ has the highest probability. That is

$$P^n(T(P)) \geq P^n(T(\hat{P})), \quad \forall \hat{P} \in \mathcal{P}_n \quad (7.25)$$

- Note: Non-negative integers $m$ and $n$, then $\frac{m^l}{n!} \geq n^{m-n}$ since if $m > n$, then $\frac{m^l}{n!} = m(m-1) \ldots (n+1) \geq n^{m-n}$, and if $m < n$, then $\frac{m^l}{n!} = \frac{1}{n(n-1) \ldots (m+1)} \geq \frac{1}{n^{n-m}}$, and if $m = n$ then obvious.
Size of type class

**Proposition 7.2.2**

For any type $P \in \mathcal{P}_n$, we have

$$\frac{1}{(n+1)|\mathcal{X}|} 2^{nH(P)} \leq |T(P)| \leq 2^{nH(P)} \quad (7.25)$$

**Proof.**

$$1 \geq P^n(T(P)) = \sum_{x_{1:n} \in T(P)} P^n(x_{1:n}) = \sum_{x_{1:n} \in T(P)} 2^{-nH(P)} \quad (7.26)$$

$$= |T(P)| 2^{-nH(P)} \quad (7.27)$$

Note that in the sums, $P_{x_{1:n}} = P$. This gives the upper bound.

Before doing the lower bound, let’s do a lemma . . .

How probable is each type class?

- **Notation:** $a_n \doteq b_n$ if $\lim_{n \to \infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0$.

**Theorem 7.2.2**

For any $P \in \mathcal{P}_n$, and any distribution $Q$, the probability of type class $T(P)$ under $Q^n$ is such that $Q^n(T(P)) \doteq 2^{-nD(P||Q)}$. Specifically,

$$\frac{1}{(n+1)|\mathcal{X}|} 2^{-nD(P||Q)} \leq Q^n(T(P)) \leq 2^{-nD(P||Q)} \quad (7.39)$$

Note: so any type less close than the “closest” type to $Q$ will decrease in probability exponentially (in $n$) faster than the most probable type.

**Proof.**

$$Q^n(T(P)) = \sum_{x_{1:n} \in T(P)} Q^n(x_{1:n}) = \sum_{x_{1:n} \in T(P)} 2^{-n(D(P||Q)+H(P))} \quad (7.40)$$

$$= |T(P)| 2^{-n(D(P||Q)+H(P))} \quad (7.41)$$

and then use

$$\frac{1}{(n+1)|\mathcal{X}|} 2^{nH(P)} \leq |T(P)| \leq 2^{nH(P)} \quad \square$$
Summary of basic theorems

- Number of types with denominator $n$

\[ |\mathcal{P}_n| \leq (n + 1)|\mathcal{X}| \quad (7.39) \]

- $p(x_{1:n})$ depends only on the type (prob. indep. of sample given type)

\[ Q^n(x_{1:n}) = 2^{-n[H(P_{x_{1:n}}) + D(P_{x_{1:n}} || Q)]} \quad (7.40) \]

- Size of the type class

\[ |T(P)| := 2^{nH(P)} \quad (7.41) \]

- Probability of a type class

\[ Q^n(T(P)) := 2^{-nD(P || Q)} \quad (7.42) \]

Types with the most probability

- Q: Which types will have the most probability?
- A: Clearly, the ones that are closest to the true distribution.
- The property $Q^n(T(P)) = 2^{-nD(P || Q)}$ says that the ones that are farther away will have exponentially smaller probability than the others, as $n \to \infty$.
- This suggests that “typical set of sequences” applies here as well, in fact

**Definition 7.2.2 (typical set of sequences)**

Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\forall i, x_i \sim Q(x)$. Then the typical set is defined as

\[ T^\epsilon_Q = \{x_{1:n} : D(P_{x_{1:n}} || Q) \leq \epsilon\} \quad (7.39) \]

- Intuitively, these are sequences that come from types that are $\epsilon$-close to $Q$ in the KL-sense.
Theorem 7.3.1

Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\forall i, x_i \sim Q(x)$. Then the probability of the complement of the typical set $\overline{T^c_Q}$ has expression:

$$Q(\overline{T^c_Q}) = Q(\{x_1:n : D(P_{x_1:n}||Q) > \epsilon\}) \leq 2^{-n(\epsilon - |X|\log(n+1))}$$  \hspace{1cm} (7.1)

and therefore,

$$D(P_{X_1:n}||Q) \xrightarrow{p} 0 \text{ as } n \to \infty$$  \hspace{1cm} (7.2)

- Intuitively, this means that types that are more than $\epsilon$ away from $Q$ have decreasing probability.
- Moreover, the typical set, which ends up for large $n$ being the only thing that occurs without vanishingly small probability, is such that the KL divergence gets between the type and $Q$ quickly gets arbitrarily small.

Proof of Theorem 7.3.1.

$$Q(\overline{T^c_Q}) = 1 - Q^n(T^c_P) = \sum_{P \in \mathcal{P}_n : D(P||Q) > \epsilon} Q^n(T(P)) \hspace{1cm} (7.3)$$

$$\leq \sum_{P \in \mathcal{P}_n : D(P||Q) > \epsilon} 2^{-nD(P||Q)} \hspace{1cm} (7.4)$$

$$\leq \sum_{P \in \mathcal{P}_n : D(P||Q) > \epsilon} 2^{-n\epsilon} \hspace{1cm} (7.5)$$

$$\leq (n + 1)^{|X|} 2^{-n\epsilon} = 2^{-n(\epsilon - |X|\log(n+1))}$$  \hspace{1cm} (7.6)

and the r.h.s. $\to 0$ as $n \to \infty$, and thus the probability of the typical set $\to 1$ as $n \to \infty$. \qed
Probability of typical set: further notes

- Also, $D(P_{X_1:n} \| Q) \to 0$ with probability 1.
- If not, i.e., if $p(D(P_{X_1:n} \| Q) \to \gamma) > 0$ with $\gamma > 0$, then, say $T_{Q_1}^{\gamma_1}$ with $\gamma_1 = \gamma/2$ would not converge to 0 with $n \to \infty$.
- Note this is very much like the AEP we saw before.
- the type of the sequence gets closer and closer to $Q$ as $n$ grows.

How often does an atypical event occur

- Since $p(T_{Q}^{\epsilon}) \leq 2^{-n(\epsilon + |X| \log \frac{n+1}{n})}$ is an exponentially decreasing sequence in $n$, it is summable,
  - and in fact
    \[
    \infty > \sum_{n=1}^{\infty} p(D(P_{X_1:n} \| Q) > \epsilon) = E \left[ \sum_{n=1}^{\infty} 1_{D(P_{X_1:n} \| Q) > \epsilon} \right] \tag{7.7}
    \]
  - So expected number of times the event $D(P_{X_1:n} \| Q) > \epsilon$ occurs is finite, out of an infinite set of possible times.
  - This has probability 0 by the Borel-Cantelli lemma (see the book Billingsly, Probability & Measure, page 59 for more details).
Universal Source Coding

- If we know \( p(x) \), then we will be able to develop a code to compress sources generated by \( p(x) \). Huffman, Lempel-Ziv, etc. are codes that, as we will soon see, do that.
- What if we don’t know \( p(x) \)?
- Q: do there exist codes that can compress without knowing \( p(x) \) and that do so down to the entropy limit?
- Q: can we compress down to the rate \( R \) (in units of bits per source symbol) if \( R > H(Q) \)? (this is Shannon’s source coding theorem)
- What happens if \( R < H(Q) \)? (this is the converse of Shannon’s source coding theorem)
- We’ll formally prove this theorem using the method of types.

Universal Source Coding: intuitive idea from AEP

- Basic idea is similar to the typical set \( A_\epsilon^{(n)} \) we’ve already seen: when \( n \) is long enough, the only sequences that occur (with non-vanishingly small probability) will be typical.
- If we encounter such a sequence, it “must” be typical since the only things that occur are typical.
- Thus, we only encode the things we see, and we count them along the way.
- In the end, we’ll need at most \( |A_\epsilon^{(n)}| \) code words for which we can index with \( nH \) bits.
- We want to formalize Shannon’s theorem and its converse using the method of types.
Universal Source Coding: intuitive idea from types

- There are $2^{nH(P)}$ sequences of type $P$.
- Thus, we use $nH(P)$ bits to represent such sequences.
- If $R > H(P)$, we are easily able to use $nR$ bits to represent such sequences.
- As $n$ gets big, only the types $P$ that are “close” to $Q$ will occur.
- There are an exponential (in $n$) number of sequences.
- There are only a polynomial (in $n$) number of types.
- Thus, one type must eventually get “all” of the probability.
- If $P \approx Q$, and $H(Q) < R$, then all types that actually “occur” can be represented in $R$ bits per source symbol.
- If $P \approx Q$, and $H(Q) > R$, then types that occur cannot be represented in $R$ bits per source symbol.

Our encoder setup

- Recall from earlier our $x$ to $y$ encoder setup.

Source messages
$\{X_1, X_2, \ldots, X_n\}$

$X_i \in \{a_1, a_2, \ldots, a_K\}$

$K^n$ possible messages

$n$ source letters in each source msg

Encoder

Code words
$\{Y_1, Y_2, \ldots, Y_m\}$

$Y_i \in \{0, 1\}$

$2^m$ possible messages

$m$ total bits
(M, n) codes

- Fixed rate block code of rate $R$.
- There are $M$ code words, $M = \text{number of possible messages}$.
- There are $n$ source symbols encoded at a time in each code word.
- An encoder maps from length-$n$ strings of source symbols to length-$m$ bit strings.

The rate $R$ of the code depends on $M$ and $n$

$$R = \frac{\log M}{n} = \frac{\log(\text{# of code words})}{\text{# of source symbols}}$$ (7.8)

Fixed rate block code of rate $R$

- An $(M, n)$ code is one that uses $M$ code words for $n$ source symbols.
- Such a code thus has rate of $R = \frac{\log M}{n}$ bits per source symbol. So we need $\log M = nR$ bits to index this code.
- Then a code is defined as follows:

**Definition 7.4.1 (fixed rate block code of rate $R$)**

Let $X_1, X_2, \ldots, X_n \sim Q$, i.i.d. but $Q$ unknown. We have encoder and decoder functions as follows:

Encoder: $f_n : \mathcal{X}^n \rightarrow \{1, 2, \ldots, 2^{nR}\}$ (7.9)

Decoder: $\phi_n : \{1, 2, \ldots, 2^{nR}\} \rightarrow \mathcal{X}^n$ (7.10)

and probability of error

$$P_e(n) = Q^n(\{x_{1:n} : \phi_n(f_n(x_{1:n})) \neq x_{1:n}\})$$ (7.11)

- Notation: $(M, n) = (2^{nR}, n)$ designates a series (in $n$) of such codes.
**Universal Code**

**Definition 7.4.2 (Universal rate \(R\) block code)**

A rate \(R\) block code for a source is universal if the functions \(f_n\) and \(\phi_n\) do not depend (rely directly) on the source distribution \(Q\) and if

\[
P_e^n \to 0 \text{ as } n \to \infty \text{ whenever } H(Q) < R \tag{7.12}
\]

- So we require the “ability to code” at rate \(R\), which really means code without error, or the error goes to zero for larger block length.
- We next state and prove one of Shannon’s main theorems.
- If \(R > H(Q)\), then there exists a sequence (in \(n\)) of codes with the error of becoming vanishingly small.
- Conversely, if \(R < H(Q)\), then the error goes to 1.

---

**Source Coding Theorem**

**Theorem 7.4.3 (Shannon’s Source Coding Theorem)**

\[
\exists \text{ a sequence } (2^{nR}, n) \text{ of universal source codes such that } P_e^n \to 0 \text{ for all source distributions } Q \text{ such that } H(Q) < R.
\]

**Proof.**

- Fix \(R > H(Q)\) to be strictly greater than entropy.
- Define a rate for \(n\) that is “fixed up” with a polynomial factor. I.e.,

\[
R_n \triangleq R - |X| \frac{\log(n + 1)}{n} < R \tag{7.13}
\]

- Define set of sequences that have type entropy less than this rate.

\[
A_n \triangleq \{x_{1:n} \in X^n : H(P_{x_{1:n}}) \leq R_n\} \tag{7.14}
\]

\[
= \left\{ \bigcup_{P \in \mathcal{P}_n} T(P) : H(P) \leq R_n \right\} \tag{7.15}
\]
Source Coding Theorem

... Proof of theorem 7.4.3 continued.

- Then
  \[ |A_n| = \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} T(P) \leq \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} 2^n H(P) \]  
  \[ \leq \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} 2^n R_n \leq (n + 1)^{|X|} 2^n R_n \]  
  \[ = 2^n (R_n + |X| \frac{\log(n + 1)}{n}) = 2^n R \]  

- Since \(|A_n| \leq 2^n R\), we can index \(A_n\) with \(nR\) bits.

- Let the encoder be:

  \[ f_n(x_{1:n}) = \begin{cases} 
  \text{lexicographic index} & \text{if } x_{1:n} \in A_n \\
  \text{of } x_{1:n} \text{ in } A_n & (\text{i.e., if } H(P_{x_{1:n}}) \leq R_n) \\
  0 & (\text{i.e., if } H(P_{x_{1:n}}) > R_n) 
  \end{cases} \]

Note: \(f_n(\cdot)\) does not depend on the source distribution, only on the ordering and on \(R_n\).

- Error occurs if \(x_{1:n} \notin A_n\).

- We can represent this by placing types within the probability simplex, to indicate which types may be encoded. E.g., if \(|\mathcal{X}| = 3\), then

\[ (1,0,0) \quad (0,1,0) \quad (0,0,1) \]
Proof of theorem 7.4.3 continued.

- Within the simplex, each point is potentially a type (the points with rational values with denominator \( n \) and numerator between 0 and \( n \)).
- Yellow region corresponds to types \( P \in P_n \) whose sequences can be encoded correctly, as the rate constraint is satisfied.
- Light blue corresponds to types whose sequences will result in an error.

\[
\begin{align*}
H(P) < R_n \\
H(P) > R_n \\
H(P) = R_n
\end{align*}
\]

So we have \( R_n \uparrow R \Rightarrow R_n < R \) for all \( n \), and \( H(Q) < R \).

Thus, for some \( n_0 \), \( \forall n > n_0 \), we have \( H(Q) < R_n \).
Source Coding Theorem

... Proof of theorem 7.4.3 continued.

In Eq. (7.22) we chose $P : H(P) > R_n$ for the current $n$ (assuming there is one).

$\Rightarrow H(P) > R_n > H(Q)$

$\Rightarrow P \neq Q$

$\Rightarrow D(P||Q) > 0$ for the chosen $P$

Thus, we get

$$P_e^{(n)} \leq \frac{(n + 1)|X|}{2^n \left[ \min_{P : H(P) > R_n} D(P||Q) \right]} \text{poly in } n$$

exp. decreasing as $n \to \infty$ (7.23)

Which implies that $P_e^{(n)} \to 0$ as $n \to \infty$.

Conversely, if $R < H(Q)$ then $P_e^{(n)} \to 1$.

So entropy is a good measure of the “information” in a source.

Can you compress a compressed file? gzip(gzip(gzip(...gzip(x))))

Once a source is compressed to the entropy rate, it is uncompressible, it is at its entropy limit.

It looks like a random string of bits.

How can we tell the difference? The code is the key, both figuratively and literally.
Stochastic Processes

- So far we’ve been talking about i.i.d. random variables, $X_1, X_2, \ldots$
- In such case, each random variable has the same entropy.
- When the random variables are no longer i.i.d., how can we talk about the entropy of a process?
- We start to address that here.

### Definition 7.5.1 ((strict-sense) Stationary stochastic Process)

A sequence of r.v.s, $X_1, X_2, \ldots, X_n$ governed by a probability distribution is strict sense stationary if it is the case that

$$p(X_{1:n} = x_{1:n}) = p(X_{1+\ell:n+\ell} = x_{1:n})$$

(7.24)

for all $\ell$, for all $n$, and for all $x_{1:n} \in \mathcal{X}^n$.

### Definition 7.5.2 (Markov process)

A stochastic process is first-order Markov if

$$p(X_{n+1} = x_{n+1} | X_{1:n} = x_{1:n}) = p(X_{n+1} = x_{n+1} | X_n = x_n)$$

(7.25)

In this latter case, it means that $p(x_{1:n}) = p(x_1)p(x_2|x_1) \ldots p(x_n|x_{n-1})$
### Stochastic Process

**Definition 7.5.3 (homogeneous)**

A Markov chain is time-invariant (or time-homogeneous, or just homogeneous) if \( p(x_{n+1}|x_n) \) does not depend on time. i.e., if

\[
p(X_{n+1} = b| X_n = a) = p(X_2 = b| X_1 = a) \quad \forall a, b, n \tag{7.26}
\]

In such case, there is a fixed transition matrix \( P = [p_{ij}]_{ij} \) with \( p_{ij} = p(X_{n+1} = j| X_n = i) \) that can be drawn as a directed graph with arrows pointing between states that have non-zero transition.

![Directed Graph Example](image)

### Stochastic Process

**Definition 7.5.4 (irreducible)**

A Markov chain is irreducible if \( p_{ij}(n) > 0 \) for all \( i, j \) and for some \( n \) where \( p_{ij}(n) = p(X_n = j| X_0 = i) \).

This is if it is possible to get from all states to all others with non-zero probability.

- Also recall note, matrix-vector for state probability at time \( n + 1 \) given that at time \( n \).

\[
p(x_{n+1}) = \sum_{x_n} p(x_n)p_{x_n,x_{n+1}} \tag{7.27}
\]

- (first order) Markov chain is stationary if \( p(x_{n+1}) = p(x_n) \)
Definition 7.5.5

A Markov chain is periodic if \( d(i) > 1 \) with

\[
d(i) = \gcd\{n : p_{ii}(n) > 0\}
\]  
(7.28)

Note that this is the \( \gcd \) of the epochs at which return to the same state is possible.

Example:

\[
P = \begin{pmatrix}
1 - \alpha & \alpha \\
\beta & 1 - \beta
\end{pmatrix}
\]

If \( \mu = [p_1 p_2]^\top \) is stationary distribution then we must have that \( \mu^\top P = \mu^\top \).

In fact, in this case \( \mu = \left[ \frac{\beta}{\alpha + \beta}, \frac{\alpha}{\alpha + \beta} \right] \).

More facts about Markov chains and stochastic processes: Great source is the text: see “Probability and Random Processes”, Grimmett and Stirzaker.
Stochastic Processes: definition brief summary

- Stationary stochastic process, statistics don’t change when we shift time.
- Markov process (model): future and past independent given the present, or immediate past is sufficient to render more distant past irrelevant.
- (time) homogeneous: parameters of Markov chain don’t change with time.
- irreducible: it is possible to get from any state to any other state eventually.
- periodic: greatest common divisor of the time intervals where return is possible is > 1.
- More facts about Markov chains and stochastic processes: Great source is the text: see “Probability and Random Processes”, Grimmett and Stirzaker.

Entropy rates

- Stochastic processes have entropy rates, which intuitively is the amount of new information, on average, that is provided by the stochastic process at each time step.
- More formally

**Definition 7.6.1**

The entropy rate of a stochastic process \( \{X_i\}_i \) is defined as

\[
H(X) \triangleq \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n)
\]

when it exists.

- So, as can be seen, it is the per symbol entropy given by the stochastic process when \( n \) gets large.
Examples

- I.i.d. set of r.v.s all \( \sim p(x) \) then

\[
H(\mathcal{X}) = \lim_{n \to \infty} \frac{H(X_{x_1:n})}{n} = \frac{\sum_{i=1}^{n} H(X_i)}{n} = H(X_1)
\]  
(7.30)

- Independent but not identically distributed:

\[
\lim_{n \to \infty} \frac{\sum_{i=1}^{n} H(X_i)}{n} = ?
\]  
(7.31)

in this case it might not exist.

- Example when it doesn’t exist. Let \( p_i = P(X_i = 1) \). Define it as

\[
p_i = \begin{cases} 
0.5 & \text{if } 2k < \log \log i \leq 2k + 1 \\
0 & \text{if } 2k + 1 < \log \log i \leq 2k + 2 
\end{cases}
\]  
(7.32)

Alternative Definition

**Definition 7.6.2**

Again, assume a stochastic process and define the following rate:

\[
H'(\mathcal{X}) \triangleq \lim_{n \to \infty} H(X_n|X_{n-1}, X_{n-2}, \ldots, X_1)
\]  
(7.33)

assuming it exists.

**Theorem 7.6.3**

For stationary stochastic process, \( H(X_n|X_{n-1}, X_{n-2}, \ldots, X_1) \) is decreasing in \( n \) and has a limit, lets call it \( H'(\mathcal{X}) \).

**Proof.**

\[
H(X_{n+1}|X_1, \ldots, X_n) \leq H(X_{n+1}|X_2, \ldots, X_n) = H(X_n|X_1, \ldots, X_{n-1})
\]  
(7.34)

Since decreasing sequence with lower bound 0, it has a limit \( H' \).
Entropy rates or entropy rate

- Cesáro mean: if $a_n \to a$ and $b_n = \frac{1}{n} \sum_{i=1}^{n} a_i$ then $b_n \to a$
- Key idea is that most of the terms in the sum are close to $a$, so the average is also close to $a$ (formal proof in book).
- This then gives the next theorem.

**Theorem 7.6.4**

*We have that for stationary stochastic processes*

\[
\lim_{n \to \infty} H(X_n \mid X_{n-1}, X_{n-2}, \ldots, X_1) \triangleq H'(\mathcal{X}) \\
= H(\mathcal{X}) \triangleq \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n)
\]

(7.36)

(7.37)

**Proof.**

\[
b_n = \frac{H(X_1, X_2, \ldots, X_n)}{n} = \frac{1}{n} \sum_{i=1}^{n} H(X_i \mid X_{i-1}, \ldots, X_1) = a_i
\]

(7.38)

and $a_n \to H'(\mathcal{X})$ so $b_n \to H'(\mathcal{X})$ but by definition $b_n \to H(\mathcal{X})$
Entropy rate

- Note that for any stationary ergodic (loosely, time and ensemble averages are the same) process, we have
  \[
  -\frac{1}{n} \log p(x_1, \ldots, x_n) \to H(\mathcal{X}) \quad (7.39)
  \]

- With this, we can prove AEP-like theorems and prove the source coding theorem for such processes, but we need more machinery to do so. This is done in section 16.8 Shannon-McMillan-Breiman Theorem (General AEP) (page 644) in our book.

Entropy rate and stationary Markov chain

- When the process is a stationary Markov chain, entropy rate has a nice form.
  - That is
  \[
  H(\mathcal{X}) = H'(\mathcal{X}) = \lim_{n \to \infty} H(X_n|X_{n-1}, \ldots, X_1) = \lim_{n \to \infty} H(X_n|X_{n-1}) = \lim_{n \to \infty} H(X_2|X_1) \quad (7.40)
  \]
  \[
  = -\sum_{x_2, x_1} p(x_2, x_1) \log p(x_2|x_1) = \sum_{i} \mu_i \left[ -\sum_{j} p_{ij} \log p_{ij} \right] \quad (7.42)
  \]
  where again \( \mu \) is the stationary distribution and \( p_{ij} \) is the transition probability from \( i \) to \( j \).

- Ex: previous figure \( H(\mathcal{X}) = H(X_2|X_1) = \frac{\beta}{\alpha+\beta} H(\alpha) + \frac{\alpha}{\alpha+\beta} H(\beta) \).
Assume irreducible and aperiodic so unique stationary distribution.

Graph $G = (V, E)$ with $m$ nodes labeled $\{1, 2, \ldots, m\}$ and edges with weight $w_{ij} \geq 0$.

Random walk: start at a node, say $i$, and choose next node with probability proportional to edge weight, i.e., $p_{ij}$ as

$$p_{ij} = \frac{w_{ij}}{\sum_j w_{ij}} = \frac{w_{ij}}{w_i}$$

where $w_i \triangleq \sum_j w_{ij}$.

Guess that stationary distribution has probability proportional to $w_i$.

If $w = \sum_{i,j:j>i} w_{ij}$ then $\sum_i w_i = 2w$, so guess as stationary distribution $\mu$ with $\mu_i = w_i/2w$.

This is stationary since

$$\forall j, \mu'_j = \sum_i \mu_i P_{ij} = \sum_i w_i \frac{w_{ij}}{2w} \frac{w_{ij}}{w_i} = \sum_i \frac{1}{2w} w_{ij} = \frac{w_j}{2w} = \mu_j$$

Can swap edges elsewhere (i.e., edges between nodes not including $i$), does not change the stationary condition which is local.

Note chain is aperiodic since $w_{ii} = 0$. This is because

$$2w = \sum_i w_i = \sum_{ij} w_{ij} = \sum_{i:j=i} w_{ij} + \sum_{i:j > i} w_{ij} + \sum_{i:j < i} w_{ij}$$

$$= \sum_{i:j=i} w_{ij} + w + \sum_{i:j < i} w_{ij}$$

and $2w - w = \sum_{i:j=i} w_{ij} + \sum_{i:j < i} w_{ij} \Rightarrow \sum_{i:j=i} w_{ij} = 0 \Rightarrow w_{ii} = 0$. 
What is entropy of this random walk

\[ H(\mathcal{X}) = H(X_2 | X_1) = - \sum_i \mu_i \sum_j p_{ij} \log p_{ij} \]
\[ = - \sum_i \frac{w_i}{2w} \sum_j \frac{w_{ij}}{w_i} \log \frac{w_{ij}}{w_i} = - \sum_i \frac{w_{ij}}{2w} \log \frac{w_{ij}}{w_i} \]
\[ = - \sum_{ij} \frac{w_{ij}}{2w} \log \left[ \frac{w_{ij}}{2w w_i} \right] \]
\[ = - \sum_{ij} \frac{w_{ij}}{2w} \log \frac{w_{ij}}{2w} + \sum_{ij} \frac{w_{ij}}{2w} \log \frac{w_i}{2w} \]
\[ = H(\ldots, \frac{w_{ij}}{2w}, \ldots) - H(\ldots, \frac{w_i}{2w}, \ldots) \]
\[ = (\text{overall edge uncertainty}) \]
\[ - (\text{overall node uncertainty in stationary condition}) \]

So, the entropy of the random walk is

\[ H(\mathcal{X}) = (\text{overall edge uncertainty}) \]
\[ - (\text{overall node uncertainty in stationary condition}) \]

Intuition: As node entropy decreases while keeping edge uncertainty constant, the network becomes more concentrated,

- fewer nodes are hubs, and the hubs that remain are widely connected (since edge entropy is fixed).

In such case (few well connected hubs), it is likely one will land on such a hub (in a random walk) and then will be faced with a wide variety of choice as to where to go next \( \Rightarrow \) increase in overall uncertainty of the walk.

If node entropy goes up with edge entropy fixed, then many nodes are hubs all with relatively low connectivity, so hitting them doesn’t provide much choice \( \Rightarrow \) random walk entropy goes down.
Hidden Markov models (HMMs)

An HMM is a distribution $p(X_{1:n}, Y_{1:n})$ over $2n$ random variables that factors in a particular way.

Easiest way to depict all of the factorization properties is to use a graphical model, as in the below, where $n = 5$:

Let $Y_1, Y_2, \ldots, Y_n$ be a stationary Markov chain.
Let $X_{1:n}$ be a random function of this Markov chain. I.e.,

$$X_i = \begin{cases} 
\phi_1(Y_i) & \text{with probability } p_1 \\
\phi_2(Y_i) & \text{with probability } p_2 \\
\vdots \\
\phi_m(Y_i) & \text{with probability } p_m
\end{cases} = \phi_N(X_i) \quad (7.55)$$

where $N \in \{1, 2, \ldots, m\}$ itself is a random variable.

HMMs

Note that the stochastic process $X_1, X_2, \ldots$ does not form a Markov chain in general. Why? because it does not satisfy the first order Markov assumption, nor any order Markov assumption in general.

If $\{Y_i\}_i$ is stationary, then is $\{X_i\}_i$ a stationary stochastic process? Yes. Possible HW problem, so no more given here.

We can compute the entropy rate of $\{X_i\}_i$, i.e.,

$$H(X) = \lim_{n \to \infty} H(X_n|X_{n-1}, \ldots, X_1)$$

but it is ugly, so instead we compute upper and lower bounds.

Upper bound(s):}

$$H(X_n|X_{n-1}, \ldots, X_1) = H(X_{n+1}|X_n, \ldots, X_2) \geq H(X_{n+1}|X_1, \ldots, X_n) \quad (7.56)$$

$$\geq H(X_{n+2}|X_{n+1}, \ldots, X_1) \geq \cdots \geq H(X) \quad (7.57)$$
A lower bound is given by $H(X_n|X_{n-1}, \ldots, X_2, Y_1) \leq H(\mathcal{X})$ because

\begin{align}
H(X_n|X_{n-1}, \ldots, X_2, Y_1) &= H(X_n|X_{n-1}, \ldots, X_2, X_1, Y_1) \quad (7.58) \\
&= H(X_n|X_{n-1}, \ldots, X_1, Y_1, Y_0, Y_{-1}, \ldots, Y_{-k}) \quad (7.59) \\
&= H(X_n|X_{n-1}, \ldots, X_1, Y_1, Y_0, Y_{-1}, \ldots, Y_{-k}, X_0, \ldots, X_{-k}) \quad (7.60) \\
&= H(X_{n+k+1}|X_{n+k}, \ldots, X_1) \quad (7.61)
\end{align}

So summarizing the bounds on the HMM information rates, we have

\begin{align}
H(\mathcal{X}) &\leq H(X_n|X_{n-1}, \ldots, X_1, Y_1) \leq H(X_n|X_{n-1}, \ldots, X_1) \quad (7.62)
\end{align}

---

**Lemma 7.7.1 (ever shrinking sandwich)**

\begin{align}
H(X_n|X_{n-1}, \ldots, X_1) - H(X_n|X_{n-1}, \ldots, X_1, Y_1) &\to 0 \quad (7.63)
\end{align}

**Proof.**

\begin{align}
H(X_n|X_{n-1}, \ldots, X_1) - H(X_n|X_{n-1}, \ldots, X_1, Y_1) &= I(X_n; Y_1|X_{n-1}, \ldots, X_1) \leq H(Y_1) \quad (7.64) \\
Also, \quad I(Y_1; X_1, \ldots, X_n) &\leq H(Y_1)\text{ for all } n.
\end{align}

Now,

\begin{align}
\lim_{n \to \infty} I(Y_1; X_1, \ldots, X_n) &= \lim_{n \to \infty} \sum_{i=1}^{n} I(Y_1; X_i|X_1:i-1) \quad (7.65) \\
&= \sum_{i=1}^{\infty} I(Y_1; X_i|X_1:i-1) \leq H(Y) < \infty \quad (7.66)
\end{align}

So an infinite sum is constant, must mean the terms $\to 0$ as $n \to \infty$. Thus, each of the terms $I(Y_1; X_i|X_1:i-1) \to 0$ as $n \to \infty$. 

\[\square\]
Summarizing, we have

$$\lim_{n \to \infty} H(X_n | X_{n-1}, \ldots, X_1, Y_1) = H(\mathcal{X}) = \lim_{n \to \infty} H(X_n | X_{n-1}, \ldots, X_1)$$

(7.67)