## Class Road Map - IT-I

- **L1 (9/25):** Overview, Communications, Information, Entropy
- **L2 (9/30):** Entropy, Mutual Information, KL-Divergence
- **L3 (10/2):** More KL, Jensen, more Venn, Log Sum, Data Proc. Inequality
- **L4 (10/7):** Data Proc. Ineq., thermodynamics, Stats, Fano,
- **L5 (10/9):** M. of Conv, AEP,
- **L6 (10/14):** AEP, Source Coding, Types
- **L7 (10/16):** Types, Univ. Src Coding, Stoc. Procs, Entropy Rates
- **L8 (10/21):**
- **L9 (10/23):**

- **L10 (10/28):**
- **L11 (10/30):**
- **L12 (11/4):**
- **LXX (11/6):** In class midterm exam
- **LXX (11/11):** Veterans Day holiday
- **L13 (11/13):**
- **L14 (11/18):**
- **L15 (11/20):**
- **L16 (11/25):**
- **L17 (11/27):**
- **L18 (12/2):**
- **L19 (12/4):**
- **LXX (12/10):** Final exam

**Finals Week: December 9th–13th.**
Cumulative Outstanding Reading

- Read chapters 1 and 2 in our book (Cover & Thomas, “Information Theory”) (including Fano’s inequality).
- Read chapters 3 and 4 in our book (Cover & Thomas, “Information Theory”).
- Read sections 11.1 through 11.3 in our book (Cover & Thomas, “Information Theory”).
- Read chapter 4 in our book (Cover & Thomas, “Information Theory”).
Homework

- Homework 1 on our assignment dropbox (https://canvas.uw.edu/courses/1319497/assignments), was due Tuesday, Oct 8th, 11:55pm.

- Homework 2 on our assignment dropbox (https://canvas.uw.edu/courses/1319497/assignments), due Friday 10/18/2019, 11:45pm.
Division of set of all sequences into type classes

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

$\mathcal{X}^n$: partitioned into blocks within which all sequences have the same type
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.
Proposition 7.2.3

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

(7.29)

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 \(n\), \(|\mathcal{X}|^n\).
Another summary thus far (don’t say we didn’t remind you)

- \( P_{x_1:n} \) is the type (empirical distribution) of the sequence \( x_{1:n} \), with 
  \[ P_{x_1:n}(a) = \frac{N(a|x_{1:n})}{n} \]
  for all \( a \in \mathcal{X} \).

- \( \mathcal{P}_n(\mathcal{X}) \) (or just \( \mathcal{P}_n \)) is the set of types with denominator \( n \).

- \( T(P) \), for type \( P \in \mathcal{P}_n \), is the set of all sequences with type \( P \), i.e.,
  \[ T(P) = \{ x_{1:n} \in \mathcal{X}^n : P_{x_1:n} = P \} \]

- **Theorem:** Number of types bounded by poly in \( n \),
  \[ |\mathcal{P}_n| \leq (n + 1)|\mathcal{X}| \]

- **Fact:** number of sequences of length \( n \) is exponential in \( n \),
  \[ |\mathcal{X}|^n \]
Probability depends only on the type

**Theorem 7.2.3**

- 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

$$Q^n(x_1:n) = 2^{-n[H(P_{x_1:n})+D(P_{x_1:n}\|Q)]} \quad (7.29)$$

- 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.
Proposition 7.2.3

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

$$\frac{1}{(n + 1)|x|}2^{nH(P)} \leq |T(P)| \leq 2^{nH(P)} \quad (7.36)$$

Proof.

$$1 \geq P^n(T(P)) = \sum_{x_{1:n} \in T(P)} P^n(x_{1:n}) = \sum_{x_{1:n} \in T(P)} 2^{-nH(P)} \quad (7.37)$$

$$= |T(P)|2^{-nH(P)} \quad (7.38)$$

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

Before doing the lower bound, lets do a lemma . . .
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 most similar 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 7.2.3**

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

$$P^n(T(P)) \geq P^n(T(\hat{P})), \forall \hat{P} \in \mathcal{P}_n \quad (7.36)$$

- Note: Non-negative integers $m$ and $n$, then $\frac{m!}{n!} \geq n^{m-n}$ since if $m > n$, then $\frac{m!}{n!} = m(m-1) \ldots (n+1) \geq n^{m-n}$, and if $m < n$, then $\frac{m!}{n!} = \frac{1}{n(n-1) \ldots (m+1)} \geq \frac{1}{n^{n-m}}$, and if $m = n$ then obvious.
Proof of Lemma 6.5.4.

\[
\frac{P^n(T(P))}{P^n(T(\hat{P}))}
\]
Type class with highest probability

Proof of Lemma 6.5.4.

\[
\frac{P^n(T(P))}{P^n(T(\hat{P}))} = \frac{|T(P)| \prod_{a \in \mathcal{X}} P(a)^{nP(a)}}{|T(\hat{P})| \prod_{a \in \mathcal{X}} P(a)^{n\hat{P}(a)}}
\]  

(7.1)

\[
\geq \prod_{a \in \mathcal{X}} \left[ \frac{n\hat{P}(a)}{nP(a)} \right] \left[ \frac{n\hat{P}(a)}{nP(a)} \right]
\]

(7.6)
Type class with highest probability

Proof of Lemma 6.5.4.

\[
\frac{P^n(T(P))}{P^n(T(\hat{P}))} = \frac{|T(P)| \prod_{a \in \mathcal{X}} P(a)^{nP(a)}}{|T(\hat{P})| \prod_{a \in \mathcal{X}} P(a)^{n\hat{P}(a)}}
\]

(7.1)

\[
= \frac{\binom{n \cdot P(a_1)}{nP(a_1)} \cdots \binom{n \cdot P(a_n)}{nP(a_n)}}{\binom{n \cdot \hat{P}(a_1)}{n\hat{P}(a_1)} \cdots \binom{n \cdot \hat{P}(a_n)}{n\hat{P}(a_n)}} \prod_{a \in \mathcal{X}} P(a)^{nP(a)}
\]

(7.2)

\[
\geq \prod_{a \in \mathcal{X}} P(a)^{nP(a)} - \hat{P}(a)
\]

(7.4)

\[
= \prod_{a \in \mathcal{X}} \left[ n \cdot \hat{P}(a) - \sum_{a \in \mathcal{X}} P(a) - \sum_{a \in \mathcal{X}} \hat{P}(a) \right]
\]

(7.5)

\[
= n \cdot (1 - 1) = 1
\]

(7.6)
Proof of Lemma 6.5.4.

\[
\frac{P^n(T(P))}{P^n(T(\hat{P}))} = \frac{|T(P)| \prod_{a \in \mathcal{X}} P(a)^{nP(a)}}{|T(\hat{P})| \prod_{a \in \mathcal{X}} P(a)^{n\hat{P}(a)}}
\]

(7.1)

\[
= \frac{\left(nP(a_1) \ nP(a_2) \ \ldots \ \ nP(a_n)\right) \prod_{a \in \mathcal{X}} P(a)^{nP(a)}}{\left(n\hat{P}(a_1) \ n\hat{P}(a_2) \ \ldots \ \ n\hat{P}(a_n)\right) \prod_{a \in \mathcal{X}} P(a)^{n\hat{P}(a)}}
\]

(7.2)

\[
= \prod_{a \in \mathcal{X}} \frac{[n\hat{P}(a)]!}{nP(a)!} \frac{P(a)^{nP(a)-\hat{P}(a)}}{P(a)^{nP(a)}}
\]

(7.3)

\[
\geq \prod_{a \in \mathcal{X}} \frac{[n\hat{P}(a)]!}{nP(a)!} \frac{P(a)^{nP(a)-\hat{P}(a)}}{P(a)^{nP(a)}}
\]

(7.4)

\[
= \prod_{a \in \mathcal{X}} \frac{[n\hat{P}(a)]!}{nP(a)!} \frac{P(a)^{nP(a)}}{P(a)^{nP(a)}}
\]

(7.5)

\[
= \prod_{a \in \mathcal{X}} \frac{[n\hat{P}(a)]!}{nP(a)!} \frac{P(a)^{nP(a)}}{P(a)^{nP(a)-\hat{P}(a)}}
\]

(7.6)
Type class with highest probability

Proof of Lemma 6.5.4.

\[
\frac{P^n(T(P))}{P^n(T(\hat{P}))} = \frac{|T(P)| \prod_{a \in \mathcal{X}} P(a)^{nP(a)}}{|T(\hat{P})| \prod_{a \in \mathcal{X}} P(a)^{n\hat{P}(a)}}
\]

(7.1)

\[
= \frac{\left(\begin{array}{c}nP(a_1) \\ nP(a_2) \\ \vdots \\ nP(a_n) \end{array}\right) \prod_{a \in \mathcal{X}} P(a)^{nP(a)}}{\left(\begin{array}{c}n\hat{P}(a_1) \\ n\hat{P}(a_2) \\ \vdots \\ n\hat{P}(a_n) \end{array}\right) \prod_{a \in \mathcal{X}} P(a)^{n\hat{P}(a)}}
\]

(7.2)

\[
= \prod_{a \in \mathcal{X}} \left[\frac{n\hat{P}(a)!}{nP(a)!}\right] P(a)^{nP(a) - \hat{P}(a)}
\]

(7.3)

\[
\geq \prod_{a \in \mathcal{X}} (nP(a))^{n(\hat{P}(a) - P(a))} P(a)^{nP(a) - \hat{P}(a)}
\]

(7.4)

(7.6)
Type class with highest probability

Proof of Lemma 6.5.4.

\[
\frac{P^n(T(P))}{P^n(T(\hat{P}))} = \frac{|T(P)| \prod_{a \in \mathcal{X}} P(a)^{nP(a)}}{|T(\hat{P})| \prod_{a \in \mathcal{X}} P(a)^{n\hat{P}(a)}}
\]

\[
= \frac{\left(\sum_{a \in \mathcal{X}} nP(a) \right)}{\left(\sum_{a \in \mathcal{X}} n\hat{P}(a) \right)} \prod_{a \in \mathcal{X}} P(a)^{nP(a)}
\]

\[
\geq \prod_{a \in \mathcal{X}} \left( \sum_{a \in \mathcal{X}} nP(a) - n\hat{P}(a) \right) \prod_{a \in \mathcal{X}} P(a)^{nP(a) - n\hat{P}(a)}
\]

\[
= \prod_{a \in \mathcal{X}} n^n(\hat{P}(a) - P(a))
\]
Type class with highest probability

Proof of Lemma 6.5.4.

\[
\frac{P^n(T(P))}{P^n(T(\hat{P}))} = \frac{|T(P)| \prod_{a \in \mathcal{X}} P(a)^nP(a)}{|T(\hat{P})| \prod_{a \in \mathcal{X}} P(a)^n\hat{P}(a)}
\]

\[= \frac{\binom{nP(a_1)}{nP(a_2)} \cdots \binom{n}{nP(a_n)}}{\binom{n\hat{P}(a_1)}{n\hat{P}(a_2)} \cdots \binom{n}{n\hat{P}(a_n)}} \prod_{a \in \mathcal{X}} P(a)^nP(a)
\]

\[= \prod_{a \in \mathcal{X}} \frac{[n\hat{P}(a)]!}{nP(a)!} P(a)^n(P(a) - \hat{P}(a))
\]

\[\geq \prod_{a \in \mathcal{X}} (nP(a))^{n(\hat{P}(a)-P(a))} P(a)^n(P(a) - \hat{P}(a))
\]

\[= \prod_{a \in \mathcal{X}} n^n(\hat{P}(a)-P(a)) = n^n[\sum_{a \in \mathcal{X}} \hat{P}(a) - \sum_{a \in \mathcal{X}} P(a)]
\]
Proof of Lemma 6.5.4.

\[
\frac{P^n(T(P))}{P^n(T(\hat{P}))} = \frac{|T(P)| \prod_{a \in \mathcal{X}} P(a)^{nP(a)}}{|T(\hat{P})| \prod_{a \in \mathcal{X}} P(a)^{n\hat{P}(a)}}
\]

\[
= \frac{\left(\binom{nP(a_1)}{nP(a_2)} \cdots \binom{nP(a_n)}{nP(a_n)}\right) \prod_{a \in \mathcal{X}} P(a)^{nP(a)}}{\left(\binom{n\hat{P}(a_1)}{n\hat{P}(a_2)} \cdots \binom{n\hat{P}(a_n)}{n\hat{P}(a_n)}\right) \prod_{a \in \mathcal{X}} P(a)^{n\hat{P}(a)}}
\]

\[
= \prod_{a \in \mathcal{X}} \left[\frac{n\hat{P}(a)!}{nP(a)!}\right] P(a)^{nP(a)-\hat{P}(a)}
\]

\[
\geq \prod_{a \in \mathcal{X}} \left(nP(a)\right)^{nP(a)-\hat{P}(a)} P(a)^{nP(a)-\hat{P}(a)}
\]

\[
= \prod_{a \in \mathcal{X}} n^{nP(a)-\hat{P}(a)} = n^n \left[\sum_{a \in \mathcal{X}} \hat{P}(a) - \sum_{a \in \mathcal{X}} P(a)\right]
\]

\[
= n^n (1-1)
\]
Type class with highest probability

Proof of Lemma 6.5.4.

\[
\frac{P^n(T(P))}{P^n(T(\hat{P}))} = \frac{|T(P)| \prod_{a \in \mathcal{X}} P(a)^{nP(a)}}{|T(\hat{P})| \prod_{a \in \mathcal{X}} P(a)^{n\hat{P}(a)}} = \frac{\left(\begin{array}{c}nP(a_1) \\ nP(a_2) \\ \vdots \\ nP(a_n) \end{array}\right) \prod_{a \in \mathcal{X}} P(a)^{nP(a)}}{\left(\begin{array}{c}n\hat{P}(a_1) \\ n\hat{P}(a_2) \\ \vdots \\ n\hat{P}(a_n) \end{array}\right) \prod_{a \in \mathcal{X}} P(a)^{n\hat{P}(a)}}
\]

(7.1)

\[
= \prod_{a \in \mathcal{X}} \frac{[n\hat{P}(a)]!}{[nP(a)]!} P(a)^{n(P(a) - \hat{P}(a))}
\]

(7.2)

\[
\geq \prod_{a \in \mathcal{X}} \left(\begin{array}{c}nP(a) \\ \hat{P}(a) \\ P(a) \end{array}\right)^n(P(a) - \hat{P}(a)) P(a)^{n(P(a) - \hat{P}(a))}
\]

(7.3)

\[
= \prod_{a \in \mathcal{X}} n^n(\hat{P}(a) - P(a)) = n^n[\sum_{a \in \mathcal{X}} \hat{P}(a) - \sum_{a \in \mathcal{X}} P(a)]
\]

(7.4)

\[
= n^n(1-1) = 1
\]

(7.5)
Type class with highest probability

Proof of Lemma 6.5.4.

\[
\frac{P^n(T(P))}{P^n(T(\hat{P}))} = \frac{|T(P)| \prod_{a \in \mathcal{X}} P(a)^{nP(a)}}{|T(\hat{P})| \prod_{a \in \mathcal{X}} P(a)^{n\hat{P}(a)}}
\]

\[= \frac{\left(\begin{array}{c} nP(a_1) \\ nP(a_2) \\ \vdots \\ nP(a_n) \end{array}\right)}{\left(\begin{array}{c} n\hat{P}(a_1) \\ n\hat{P}(a_2) \\ \vdots \\ n\hat{P}(a_n) \end{array}\right)} \prod_{a \in \mathcal{X}} P(a)^{nP(a)}
\]

\[= \prod_{a \in \mathcal{X}} \left[ \frac{n\hat{P}(a)!}{nP(a)!} \right] P(a)^{n(P(a) - \hat{P}(a))}
\]

\[\geq \prod_{a \in \mathcal{X}} (nP(a))^{n(\hat{P}(a) - P(a))} P(a)^{n(P(a) - \hat{P}(a))}
\]

\[= \prod_{a \in \mathcal{X}} n^n(\hat{P}(a) - P(a)) = n^n \left[ \sum_{a \in \mathcal{X}} \hat{P}(a) - \sum_{a \in \mathcal{X}} P(a) \right]
\]

\[= n^n(1 - 1) = 1
\]

thus, 

\[P^n(T(P)) \geq P^n(T(\hat{P})).\]
Proof of lower bound

Proof of lowerbound of theorem 6.5.3.

1

(7.11)
Proof of lower bound of theorem 6.5.3.

\[ 1 = \sum_{Q \in \mathcal{P}_n} P^n(T(Q)) \]  

(7.11)
Proof of lower bound of theorem 6.5.3.

\[ 1 = \sum_{Q \in \mathcal{P}_n} P^n(T(Q)) \leq \sum_{Q \in \mathcal{P}_n} \max_{R \in \mathcal{P}_n} P^n(T(R)) \quad (7.7) \]

\[ (7.11) \]
Proof of lower bound

Proof of lower bound of theorem 6.5.3.

\[ 1 = \sum_{Q \in \mathcal{P}_n} P^n(T(Q)) \leq \sum_{Q \in \mathcal{P}_n} \max_{R \in \mathcal{P}_n} P^n(T(R)) \quad (7.7) \]

\[ = \sum_{Q \in \mathcal{P}_n} P^n(T(P)) \]

\[ = (n+1)|X| \sum_{x_1:n \in T(P)} P^n(x_1:n) \]

\[ \leq (n+1)|X| \left( 2^{-nH(P)} \right) \quad (7.9) \]

\[ = (n+1)|X| \left( \frac{1}{2} \right)^{2^{2-nH(P)}} \]

\[ = (n+1)|X| \left( \frac{1}{2} \right)^{2^{-nH(P)}} \quad (7.10) \]

which gives us our result that.

\[ |T(P)| \geq \frac{1}{2} (n+1)|X| \left( \frac{1}{2} \right)^{2^{-nH(P)}} \quad (7.12) \]
Proof of lower bound

Proof of lowerbound of theorem 6.5.3.

\[ 1 = \sum_{Q \in \mathcal{P}_n} P^n(T(Q)) \leq \sum_{Q \in \mathcal{P}_n} \max_{R \in \mathcal{P}_n} P^n(T(R)) \quad (7.7) \]

\[ = \sum_{Q \in \mathcal{P}_n} P^n(T(P)) \leq (n + 1)|\mathcal{X}| P^n(T(P)) \quad (7.8) \]

\[ (7.11) \]
Proof of lower bound of theorem 6.5.3.

\[
1 = \sum_{Q \in \mathcal{P}_n} P^n(T(Q)) \leq \sum_{Q \in \mathcal{P}_n} \max_{R \in \mathcal{P}_n} P^n(T(R)) \leq \sum_{Q \in \mathcal{P}_n} P^n(T(P)) \leq (n + 1)|\mathcal{X}| P^n(T(P))
\]

(7.7)

\[
= (n + 1)|\mathcal{X}| \sum_{x_{1:n} \in T(P)} P^n(x_{1:n})
\]

(7.8)

\[
= \frac{(n + 1)|\mathcal{X}|}{2^n} |T(P)|
\]

(7.11)
Proof of lower bound of theorem 6.5.3.

\[
1 = \sum_{Q \in \mathcal{P}_n} P^n(T(Q)) \leq \sum_{Q \in \mathcal{P}_n} \max_{R \in \mathcal{P}_n} P^n(T(R)) \\
= \sum_{Q \in \mathcal{P}_n} P^n(T(P)) \leq (n + 1)|\mathcal{X}| P^n(T(P)) \\
= (n + 1)|\mathcal{X}| \sum_{x_1:n \in T(P)} P^n(x_{1:n}) = (n + 1)|\mathcal{X}| \sum_{x_1:n \in T(P)} 2^{-nH(P)}
\]
Proof of lower bound of theorem 6.5.3.

\[ 1 = \sum_{Q \in \mathcal{P}_n} P^n(T(Q)) \leq \sum_{Q \in \mathcal{P}_n} \max_{R \in \mathcal{P}_n} P^n(T(R)) \]  
(7.7)

\[ = \sum_{Q \in \mathcal{P}_n} P^n(T(P)) \leq (n + 1)^{|\mathcal{X}|} P^n(T(P)) \]  
(7.8)

\[ = (n + 1)^{|\mathcal{X}|} \sum_{x_1:n \in T(P)} P^n(x_1:n) = (n + 1)^{|\mathcal{X}|} \sum_{x_1:n \in T(P)} 2^{-nH(P)} \]  
(7.9)

\[ = (n + 1)^{|\mathcal{X}|} |T(P)| 2^{-nH(P)} \]  
(7.10)

\[ = (n + 1)^{|\mathcal{X}|} |T(P)| 2^{-nH(P)} \]  
(7.11)
Proof of lower bound of theorem 6.5.3.

\[
1 = \sum_{Q \in \mathcal{P}_n} P^n(T(Q)) \leq \sum_{Q \in \mathcal{P}_n} \max_{R \in \mathcal{P}_n} P^n(T(R))
\]

\[
= \sum_{Q \in \mathcal{P}_n} P^n(T(P)) \leq (n+1)^{|X|} P^n(T(P))
\]

\[
= (n+1)^{|X|} \sum_{x_{1:n} \in T(P)} P^n(x_{1:n}) = (n+1)^{|X|} \sum_{x_{1:n} \in T(P)} 2^{-nH(P)}
\]

\[
= (n+1)^{|X|} |T(P)| 2^{-nH(P)}
\]

which gives us our result that.

\[
|T(P)| \geq \frac{1}{(n+1)^{|X|}} 2^{nH(P)}
\]
For binary case, $\mathcal{X} = \{0, 1\}$ we have the interesting bound

$$\frac{1}{(n+1)^2}2^n H\left(\frac{k}{n}\right) \leq \binom{n}{k} \leq 2^n H\left(\frac{k}{n}\right) \quad (7.13)$$
For binary case, $\mathcal{X} = \{0, 1\}$ we have the interesting bound

$$\frac{1}{(n + 1)^2} 2^{nH\left(\frac{k}{n}\right)} \leq \binom{n}{k} \leq 2^{nH\left(\frac{k}{n}\right)}$$

(7.13)

Information theory can often be used to produce bounds on combinatorial functions. See chapter 16 in 1st edition of book or chapter 17 in 2nd edition.
For binary case, \( \mathcal{X} = \{0, 1\} \) we have the interesting bound

\[
\frac{1}{(n+1)^2} 2^{nH\left(\frac{k}{n}\right)} \leq \binom{n}{k} \leq 2^{nH\left(\frac{k}{n}\right)}
\] (7.13)

Information theory can often be used to produce bounds on combinatorial functions. See chapter 16 in 1st edition of book or chapter 17 in 2nd edition.

The bound in fact can be tightened a bit in the binary case:

\[
\frac{1}{(n+1)^2} 2^{nH\left(\frac{k}{n}\right)} \leq \binom{n}{k} \] (7.14)

(exercise: show this)
Equal to first order in the exponent

- Notation: \( a_n \approx b_n \) if \( \lim_{n \to \infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0. \)
Equal to first order in the exponent

- **Notation:** $a_n \asymp b_n$ if $\lim_{n \to \infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0$.
- **Example:** $a_n = \text{poly}_1(n)$ and $b_n = \text{poly}_2(n)$ then $a_n \asymp b_n$. 
Equal to first order in the exponent

Notation: \( a_n \overset{\cdot}{=} b_n \) if \( \lim_{n \to \infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0 \).

Example: \( a_n = \text{poly}_1(n) \) and \( b_n = \text{poly}_2(n) \) then \( a_n \overset{\cdot}{=} b_n \).

Example: \( a_n = \alpha 2^n + \text{poly}_1(n) \) and \( b_n = \beta 2^n + \text{poly}_2(n) \) then \( a_n \overset{\cdot}{=} b_n \).
Equal to first order in the exponent

Notation: \( a_n \asymp b_n \) if \( \lim_{n \to \infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0 \).

Example: \( a_n = \text{poly}_1(n) \) and \( b_n = \text{poly}_2(n) \) then \( a_n \asymp b_n \).

Example: \( a_n = \alpha 2^n + \text{poly}_1(n) \) and \( b_n = \beta 2^n + \text{poly}_2(n) \) then \( a_n \asymp b_n \).

Example: \( a_n = \alpha 2^n \text{poly}_1(n) \) and \( b_n = \beta 2^n \text{poly}_2(n) \) then \( a_n \asymp b_n \).
Equal to first order in the exponent

- **Notation:** $a_n \overset{\cdot}{=} b_n$ if $\lim_{n \to \infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0$.
- **Example:** $a_n = \text{poly}_1(n)$ and $b_n = \text{poly}_2(n)$ then $a_n \overset{\cdot}{=} b_n$.
- **Example:** $a_n = \alpha 2^n + \text{poly}_1(n)$ and $b_n = \beta 2^n + \text{poly}_2(n)$ then $a_n \overset{\cdot}{=} b_n$.
- **Example:** $a_n = \alpha 2^n \text{poly}_1(n)$ and $b_n = \beta 2^n \text{poly}_2(n)$ then $a_n \overset{\cdot}{=} b_n$.
- **Example:** $a_n = \alpha 2^{sn} \text{poly}_1(n)$ and $b_n = \beta 2^{tn} \text{poly}_2(n)$ then $a_n \not\overset{\cdot}{=} b_n$ if $s \neq t$. 
How probable is each type class?

**Theorem 7.3.1**

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

$$\frac{1}{(n+1)|X|}2^{-nD(P\|Q)} \leq Q^n(T(P)) \leq 2^{-nD(P\|Q)} \quad (7.15)$$
How probable is each type class?

**Theorem 7.3.1**

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

$$\frac{1}{(n + 1)|\mathcal{X}|} 2^{-nD(P||Q)} \leq Q^n(T(P)) \leq 2^{-nD(P||Q)}$$ \hspace{1cm} (7.15)

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.
How probable is each type class?

**Theorem 7.3.1**

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

\[
\frac{1}{(n+1)|\mathbf{x}|} 2^{-nD(P\|Q)} \leq Q^n(T(P)) \leq 2^{-nD(P\|Q)}
\]  

(7.15)

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.**

\[
Q^n(T(P))
\]  

(7.17)
How probable is each type class?

**Theorem 7.3.1**

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

$$\frac{1}{(n+1)|\mathcal{X}|}2^{-nD(P||Q)} \leq Q^n(T(P)) \leq 2^{-nD(P||Q)} \quad (7.15)$$

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.**

$$Q^n(T(P)) = \sum_{x_1:n \in T(P)} Q^n(x_{1:n}) \quad (7.17)$$
How probable is each type class?

Theorem 7.3.1

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

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

(7.15)

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.

$$Q^n(T(P)) = \sum_{x_1:n \in T(P)} Q^n(x_1:n) = \sum_{x_1:n \in T(P)} 2^{-n(D(P||Q)+H(P))}$$

(7.16)

(7.17)
How probable is each type class?

**Theorem 7.3.1**

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

$$\frac{1}{(n + 1)|X|}2^{-nD(P\|Q)} \leq Q^n(T(P)) \leq 2^{-nD(P\|Q)} \quad (7.15)$$

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.**

$$Q^n(T(P)) = \sum_{x_{1:n} \in T(P)} Q^n(x_{1:n}) = \sum_{x_{1:n} \in T(P)} 2^{-n(D(P\|Q) + H(P))} \quad (7.16)$$

$$= |T(P)|2^{-n(D(P\|Q) + H(P))} \quad (7.17)$$
How probable is each type class?

**Theorem 7.3.1**

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

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

(7.15)

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.**

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

(7.16)

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

(7.17)

and then use

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

- Number of types with denominator \( n \)

\[
|\mathcal{P}_n| \leq (n + 1)^{|X|} \quad (7.18)
\]
Summary of basic theorems

- Number of types with denominator $n$

  $$|P_n| \leq (n + 1)|X|$$  \hspace{1cm} (7.18)

- $p(x_1:n)$ depends only on the type (prob. indep. of sample given type)

  $$Q^n(x_1:n) = 2^{-n}[H(P_{x_1:n}) + D(P_{x_1:n} || Q)]$$  \hspace{1cm} (7.19)
Summary of basic theorems

- Number of types with denominator $n$
  \[ |\mathcal{P}_n| \leq (n + 1)|\mathcal{X}| \]  \hspace{1cm} (7.18)

- $p(x_1:n)$ depends only on the type (prob. indep. of sample given type)
  \[ Q^n(x_1:n) = 2^{-n}[H(P_{x_1:n}) + D(P_{x_1:n} || Q)] \]  \hspace{1cm} (7.19)

- Size of the type class, $|T(P)| \doteq 2^{nH(P)}$, meaning
  \[ \frac{1}{(n + 1)|\mathcal{X}|} 2^{nH(P)} \leq |T(P)| \leq 2^{nH(P)} \]  \hspace{1cm} (7.20)
Summary of basic theorems

- Number of types with denominator $n$
  \[ |\mathcal{P}_n| \leq (n + 1)|\mathcal{X}| \quad (7.18) \]

- $p(x_{1:n})$ depends only on the type (prob. indep. of sample given type)
  \[ Q^n(x_{1:n}) = 2^{-n}[H(P_{x_{1:n}}) + D(P_{x_{1:n}} || Q)] \quad (7.19) \]

- Size of the type class, $|T(P)| \approx 2^{nH(P)}$, meaning
  \[ \frac{1}{(n + 1)|\mathcal{X}|} 2^{nH(P)} \leq |T(P)| \leq 2^{nH(P)} \quad (7.20) \]

- Probability of a type class, $Q^n(T(P)) \approx 2^{-nD(P||Q)}$ meaning
  \[ \frac{1}{(n + 1)|\mathcal{X}|} 2^{-nD(P||Q)} \leq Q^n(T(P)) \leq 2^{-nD(P||Q)} \quad (7.21) \]
Q: Which types will have the most probability?

Intuitively, these are sequences that come from types that are $\epsilon$-close to $Q$ in the KL-sense.
Types with the most probability

Q: Which types will have the most probability?
A: Clearly, the ones that are closest to the true distribution.
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 $Q^n(T(P)) \sim 2^{-nD(P||Q)}$ says that the ones that are farther away will have exponentially smaller probability than the others, as $n \to \infty$. 

\begin{align*}
Q^n(T(P)) \sim 2^{-nD(P||Q)}
\end{align*}
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,
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 $Q^n(T(P)) \doteq 2^{-nD(P||Q)}$ 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,
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 $Q^n(T(P)) \doteq 2^{-nD(P||Q)}$ 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 7.3.2 (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

$$T^\epsilon_Q = \{x_{1:n} : D(P_{x_{1:n}}||Q) \leq \epsilon\} \quad (7.22)$$
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 $Q^n(T(P)) \overset{\triangle}{=} 2^{-nD(P\|Q)}$ 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 7.3.2 (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

$$T_Q^\epsilon = \{x_{1:n} : D(P_{x_{1:n}} \| Q) \leq \epsilon\} \quad (7.22)$$

- Intuitively, these are sequences that come from types that are $\epsilon$-close to $Q$ in the KL-sense.
Theorem 7.3.3

Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\forall i, x_i \sim Q(x)$. Then $\forall \epsilon > 0$, the probability of the complement of the typical set $\bar{T}_Q^\epsilon$ has expression:

$$Q(\bar{T}_Q^\epsilon) = Q(\{x_{1:n} : D(P_{x_{1:n}} || Q) > \epsilon\}) \leq 2^{-n(\epsilon - |X| \log(n+1))}$$  \hspace{1cm} (7.23)

and therefore,

$$D(P_{X_{1:n}} || Q) \xrightarrow{p} 0 \text{ as } n \to \infty$$ \hspace{1cm} (7.24)

- Intuitively, this means that types that are more than $\epsilon$ away from $Q$ have decreasing probability, $Q(\bar{T}_Q^\epsilon) \to 0$ as $n \to \infty$. 
**Theorem 7.3.3**

Let \( X_1, X_2, \ldots, X_n \) be i.i.d. \( \forall i, x_i \sim Q(x) \). Then \( \forall \epsilon > 0 \), the probability of the complement of the typical set \( \bar{T}_Q^\epsilon \) has expression:

\[
Q(\bar{T}_Q^\epsilon) = Q(\{x_1:n : D(P_{x_1:n} \| Q) > \epsilon\}) \leq 2^{-n(\epsilon - |X| \log(n+1)/n)} \tag{7.23}
\]

and therefore,

\[
D(P_{X_1:n} \| Q) \xrightarrow{p} 0 \text{ as } n \to \infty \tag{7.24}
\]

- Intuitively, this means that types that are more than \( \epsilon \) away from \( Q \) have decreasing probability, \( Q(\bar{T}_Q^\epsilon) \to 0 \) as \( n \to \infty \).
- 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 7.3.3.

\[ Q(\overline{T}_{\bar{Q}}^\epsilon) \]

(7.28)
Proof of Theorem 7.3.3.

\[ Q(\bar{T}_Q) = 1 - Q^n(T_Q) \]
Proof of Theorem 7.3.3.

\[ Q(\bar{T}_Q^\epsilon) = 1 - Q^n(T_Q^\epsilon) = \sum_{P \in \mathcal{P}_n : D(P||Q) > \epsilon} Q^n(T(P)) \]  

(7.25)

\[ \leq \sum_{P \in \mathcal{P}_n : D(P||Q) > \epsilon} Q^n(T(P)) \]  

(7.26)

\[ \leq 2 - n\epsilon \]  

(7.27)

\[ \leq (n + 1)|X| - n\epsilon = 2 - n\epsilon = 2 - n(\epsilon - |X|\log(n+1)/n) \]  

(7.28)

and the r.h.s. → 0 as n → ∞, and thus the probability of the typical set → 1 as n → ∞.
Proof of Theorem 7.3.3.

\[ Q(\bar{T}^\epsilon_Q) = 1 - Q^n(T^\epsilon_Q) = \sum_{P \in \mathcal{P}_n: D(P||Q) > \epsilon} Q^n(T(P)) \]  
\leq \sum_{P \in \mathcal{P}_n: D(P||Q) > \epsilon} 2^{-nD(P||Q)} \]  
\leq (n + 1) |X|^2 - n \epsilon \]  
\quad \Rightarrow \quad \text{r.h.s.} \rightarrow 0 \quad \text{as} \quad n \rightarrow \infty, \]  
and thus the probability of the typical set \( \rightarrow 1 \) as \( n \rightarrow \infty \).
Proof of Theorem 7.3.3.

\[ Q(\overline{T^\epsilon_Q}) = 1 - Q^n(T^\epsilon_Q) = \sum_{P \in \mathcal{P}_n : D(P||Q) > \epsilon} Q^n(T(P)) \]  \hspace{1cm} (7.25)

\[ \leq \sum_{P \in \mathcal{P}_n : D(P||Q) > \epsilon} 2^{-nD(P||Q)} \]  \hspace{1cm} (7.26)

\[ \leq \sum_{P \in \mathcal{P}_n : D(P||Q) > \epsilon} 2^{-n \epsilon} \]  \hspace{1cm} (7.27)

\[ \leq (n + 1)|X|^2 - n \epsilon \]  \hspace{1cm} (7.28)

and the r.h.s. → 0 as \( n \to \infty \), and thus the probability of the typical set → 1 as \( n \to \infty \).
Proof of Theorem 7.3.3.

\[ Q(\bar{T}_Q^\epsilon) = 1 - Q^n(T_Q^\epsilon) = \sum_{P \in \mathcal{P}_n : D(P||Q) > \epsilon} Q^n(T(P)) \quad (7.25) \]

\[ \leq \sum_{P \in \mathcal{P}_n : D(P||Q) > \epsilon} 2^{-nD(P||Q)} \quad (7.26) \]

\[ \leq \sum_{P \in \mathcal{P}_n : D(P||Q) > \epsilon} 2^{-n\epsilon} \quad (7.27) \]

\[ \leq (n + 1)|\mathcal{X}|2^{-n\epsilon} \quad (7.28) \]
Proof of Theorem 7.3.3.

\[ Q(\bar{T}_Q^e) = 1 - Q^n(T_Q^e) = \sum_{P \in \mathcal{P}_n : D(P || Q) > \epsilon} Q^n(T(P)) \]  \hspace{1cm} (7.25)

\[ \leq \sum_{P \in \mathcal{P}_n : D(P || Q) > \epsilon} 2^{-nD(P || Q)} \]  \hspace{1cm} (7.26)

\[ \leq \sum_{P \in \mathcal{P}_n : D(P || Q) > \epsilon} 2^{-n\epsilon} \]  \hspace{1cm} (7.27)

\[ \leq (n + 1)|\mathcal{X}|2^{-n\epsilon} = 2^{-n\left(\epsilon - |\mathcal{X}| \frac{\log(n+1)}{n}\right)} \]  \hspace{1cm} (7.28)
Proof of Theorem 7.3.3.

\[ Q(T_Q^\epsilon) = 1 - Q^n(T_Q^\epsilon) = \sum_{P \in \mathcal{P}_n : D(P \| Q) > \epsilon} Q^n(T(P)) \]  
\[ \leq \sum_{P \in \mathcal{P}_n : D(P \| Q) > \epsilon} 2^{-nD(P \| Q)} \]  
\[ \leq \sum_{P \in \mathcal{P}_n : D(P \| Q) > \epsilon} 2^{-n\epsilon} \]  
\[ \leq (n + 1)|\mathcal{X}|2^{-n\epsilon} = 2^{-n\left(\epsilon - |\mathcal{X}|\frac{\log(n+1)}{n}\right)} \]

and the r.h.s. \( \to 0 \) as \( n \to \infty \),
Proof of Theorem 7.3.3.

\begin{align*}
Q(\overline{T}^\epsilon_Q) &= 1 - Q^n(T^\epsilon_Q) = \sum_{P \in \mathcal{P}_n : D(P \| Q) > \epsilon} Q^n(T(P)) \quad (7.25) \\
&\leq \sum_{P \in \mathcal{P}_n : D(P \| Q) > \epsilon} 2^{-nD(P \| Q)} \quad (7.26) \\
&\leq \sum_{P \in \mathcal{P}_n : D(P \| Q) > \epsilon} 2^{-n\epsilon} \quad (7.27) \\
&\leq (n + 1)|\mathcal{X}|2^{-n\epsilon} = 2^{-n\left(\epsilon - |\mathcal{X}| \frac{\log(n+1)}{n}\right)} \quad (7.28)
\end{align*}

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
Also, $D(P_{X_1:n} \| Q) \to 0$ with probability 1.
Also, $D(P_{X_1:n} \parallel Q) \to 0$ with probability 1.

If not, i.e., if $\Pr(D(P_{X_1:n} \parallel Q) \to \gamma) > 0$ with $\gamma > 0$, then, say $\bar{T}_Q^{\gamma_1}$ 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 random sequence gets closer and closer to $Q$ as $n$ grows, so the random sequence's type becomes less and less random.
Also, $D(P_{X_1:n} \| Q) \rightarrow 0$ with probability 1.

If not, i.e., if $\Pr(D(P_{X_1:n} \| Q) \rightarrow \gamma) > 0$ with $\gamma > 0$, then, say $\bar{T}^\gamma_Q$ with $\gamma_1 = \gamma/2$ would not converge to 0 with $n \rightarrow \infty$.

Note this is very much like the AEP we saw before.
Also, $D(P_{X_1:n} \parallel Q) \to 0$ with probability 1.

If not, i.e., if $\Pr(D(P_{X_1:n} \parallel Q) \to \gamma) > 0$ with $\gamma > 0$, then, say $\bar{T}_Q^{\gamma_1}$ 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 random sequence gets closer and closer to $Q$ as $n$ grows, so the random sequence’s type becomes less and less random.
How often does an atypical event occur

Since \( p(\bar{T}^\epsilon) \leq 2^{-n\left(\epsilon + |\mathcal{X}| \log \frac{n+1}{n}\right)} \) is an exponentially decreasing sequence in \( n \), it is summable,
How often does an atypical event occur

Since $p(T_Q^\epsilon) \leq 2^{-n \left( \epsilon + |X| \log \frac{n+1}{n} \right)}$ is an exponentially decreasing sequence in $n$, it is summable,

and in fact

$$\sum_{n=1}^{\infty} p(D(P_{X_1:n} \parallel Q) > \epsilon) = E \left[ \sum_{n=1}^{\infty} 1\{D(P_{X_1:n} \parallel Q) > \epsilon\} \right]$$

(7.29)
How often does an atypical event occur

- Since \( p(T_{Q}^{\epsilon}) \leq 2^{-n\left(\epsilon + |\mathcal{X}| \log \frac{n+1}{n}\right)} \) is an exponentially decreasing sequence in \( n \), it is summable,

- and in fact

\[
\infty > \sum_{n=1}^{\infty} p(D(P_{X_1:n} || Q) > \epsilon) = E \left[ \sum_{n=1}^{\infty} 1\{D(P_{X_1:n} || Q) > \epsilon\} \right] \quad (7.29)
\]

- 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.
How often does an atypical event occur

- Since $p(T^\epsilon_Q) \leq 2^{-n\left(\epsilon + |X|\log\frac{n+1}{n}\right)}$ is an exponentially decreasing sequence in $n$, it is summable,

and in fact

$$\sum_{n=1}^{\infty} p(D(P_{X_1:n} \| Q) > \epsilon) = E \left[ \sum_{n=1}^{\infty} 1\{D(P_{X_1:n} \| Q) > \epsilon\} \right]$$

(7.29)

- 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).
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.
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 $p(x)$?
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?
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)

---

Prof. Jeff Bilmes
EE514a/Fall 2019/Info. Theory I – Lecture 7 - Oct 16th, 2019
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)
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.
Basic idea is similar to the typical set $A^{(n)}_{\epsilon}$ 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.
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.
Basic idea is similar to the typical set $A^{(n)}_{\epsilon}$ 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 $|A^{(n)}_{\epsilon}|$ code words for which we can index with $nH$ bits.
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 $P$. 
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.
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.
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.
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.
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.
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. Why not more than one?
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. Why not more than one?
- If $P \approx Q$, and $H(Q) < R$, then all types that actually “occur” can be represented in $R$ bits per source symbol.
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. Why not more than one?
- 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 can not 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 \(R\).
$(M, n)$ codes

- Fixed rate block code of rate $R$.
- There are $M$ code words, $M =$ number of possible messages.
**$(M, n)$ 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.
(\(M, n\)) codes

- Fixed rate block code of rate \(R\).
- There are \(M\) code words, \(M = \) 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.

\[
\begin{align*}
\text{n source symbols} & \quad \text{Encoder} \quad \text{m bits in each code word} \\
(x^{(1)}_1, x^{(1)}_2, \ldots, x^{(1)}_n) & \quad (y^{(1)}_1, y^{(1)}_2, \ldots, y^{(1)}_m) \\
(x^{(2)}_1, x^{(2)}_2, \ldots, x^{(2)}_n) & \quad (y^{(2)}_1, y^{(2)}_2, \ldots, y^{(2)}_m) \\
(x^{(3)}_1, x^{(3)}_2, \ldots, x^{(3)}_n) & \quad (y^{(3)}_1, y^{(3)}_2, \ldots, y^{(3)}_m) \\
\vdots & \quad \vdots \\
(x^{(M)}_1, x^{(M)}_2, \ldots, x^{(M)}_n) & \quad (y^{(M)}_1, y^{(M)}_2, \ldots, y^{(M)}_m)
\end{align*}
\]
(M, n) codes

- Fixed rate block code of rate R.
- There are M code words, M = 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

\[ R = \frac{\log M}{n} = \frac{\log(\text{# of code words})}{\text{# of source symbols}} \] (7.30)
Fixed rate block code of rate $R$

- An $(M, n)$ code is one that uses $M$ code words for $n$ source symbols.
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.
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:
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:

\begin{align}
\text{Encoder: } f_n &: X_n \rightarrow \{1, 2, \ldots, 2^{nR}\} \\
\text{Decoder: } \phi_n &: \{1, 2, \ldots, 2^{nR}\} \rightarrow X_n
\end{align}

(7.32)

Notation: $(M, n) = (2^{nR}, n)$ designates a series (in $n$) of such codes.
**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: \(f_n : \mathcal{X}^n \rightarrow \{1, 2, \ldots, 2^{nR}\}\)  
(7.31)

Decoder: \(\phi_n : \{1, 2, \ldots, 2^{nR}\} \rightarrow \mathcal{X}^n\)  
(7.32)
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: $f_n : \mathcal{X}^n \rightarrow \{1, 2, \ldots, 2^{nR}\}$ (7.31)

Decoder: $\phi_n : \{1, 2, \ldots, 2^{nR}\} \rightarrow \mathcal{X}^n$ (7.32)
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: $f_n : \mathcal{X}^n \rightarrow \{1, 2, \ldots, 2^{nR}\}$ \hspace{1cm} (7.31)

Decoder: $\phi_n : \{1, 2, \ldots, 2^{nR}\} \rightarrow \mathcal{X}^n$ \hspace{1cm} (7.32)

and probability of error

$$P_e^{(n)} = Q^n(\{x_{1:n} : \phi_n(f_n(x_{1:n})) \neq x_{1:n}\}) \hspace{1cm} (7.33)$$
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: $f_n : \mathcal{X}^n \rightarrow \{1, 2, \ldots, 2^{nR}\}$ \hspace{1cm} (7.31)

Decoder: $\phi_n : \{1, 2, \ldots, 2^{nR}\} \rightarrow \mathcal{X}^n$ \hspace{1cm} (7.32)

and probability of error

$$P_e^{(n)} = Q^n(\{x_{1:n} : \phi_n(f_n(x_{1:n})) \neq x_{1:n}\})$$ \hspace{1cm} (7.33)

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

$$P_e^{(n)} \to 0 \text{ as } n \to \infty \text{ whenever } H(Q) < R$$ (7.34)
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

$$P_e^{(n)} \to 0 \text{ as } n \to \infty \text{ whenever } H(Q) < R$$

(7.34)

- 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.
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

$$P_e^{(n)} \to 0 \text{ as } n \to \infty \text{ whenever } H(Q) < R \quad (7.34)$$

- 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.
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

$$P_e^{(n)} \to 0 \text{ as } n \to \infty \text{ whenever } H(Q) < R$$  \hspace{1cm} (7.34)

- 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.
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

$$P_e^{(n)} \to 0 \text{ as } n \to \infty \text{ whenever } H(Q) < R$$

(7.34)

- 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)

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

Proof.
- Fix \( R > H(Q) \) to be strictly greater than entropy.
Source Coding Theorem

Theorem 7.4.3 (Shannon’s Source Coding Theorem)

∃ a sequence \((2^nR, n)\) of universal source codes such that \(P_e(n) \to 0\) for all source distributions \(Q\) 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.,

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

(7.35)
Source Coding Theorem

Theorem 7.4.3 (Shannon’s Source Coding Theorem)

∃ a sequence \((2^{nR}, n)\) of universal source codes such that \(P_e^{(n)} \to 0\) for all source distributions \(Q\) 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.,

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

(7.35)

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

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

(7.36)

\[
= \bigcup_{P \in P_n} T(P) : H(P) \leq R_n
\]

(7.37)
Source Coding Theorem

... Proof of theorem 7.4.3 continued.

- Then

\[
| A_n | = \sum_{P \in P^n} H(P) \leq Rn \leq \sum_{P \in P^n} H(P) \leq Rn^2 n H(P) \leq (n+1)|X| 2\frac{Rn}{n} \quad (7.39)
\]

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

Let the encoder be:

\[
f_n(x_1:n) = \begin{cases} \text{lexicographic index of } x_1:n \text{ in } A_n \text{ (i.e., if } H(P_{x_1:n}) \leq Rn) \\ 0 \text{ (i.e., if } H(P_{x_1:n}) > Rn) \end{cases} \quad (7.41)
\]

...
Source Coding Theorem

... Proof of theorem 7.4.3 continued.

Then $|A_n|$
Then $|A_n| = \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} |T(P)|$ (7.40)
Source Coding Theorem

... Proof of theorem 7.4.3 continued.

Then $|A_n| = \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} |T(P)| \leq \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} 2^n H(P)$ \hspace{1cm} (7.38)

Since $|A_n| \leq 2^n R_n$, we can index $A_n$ with $nR_n$ bits.

Let the encoder be:

$\begin{align*}
  f_n(x_1:n) &= \begin{cases} 
    \text{lexicographic index} \\
    0 
  \end{cases} \\
  &\text{if } x_1:n \in A_n \quad \text{(i.e., if } H(P_{x_1:n}) \leq R_n) \\
  &\text{else} \quad \text{(i.e., if } H(P_{x_1:n}) > R_n) 
\end{align*}$ \hspace{1cm} (7.41)
Source Coding Theorem

... Proof of theorem 7.4.3 continued.

Then
\[
|A_n| = \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} |T(P)| \leq \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} 2^{nH(P)} \tag{7.38}
\]

\[
\leq \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} 2^{nR_n} \tag{7.40}
\]
Then $|A_n| = \sum_{P \in \mathcal{P}_n: H(P) \leq R_n} |T(P)| \leq \sum_{P \in \mathcal{P}_n: H(P) \leq R_n} 2^{nH(P)} \leq \sum_{P \in \mathcal{P}_n: H(P) \leq R_n} 2^{nR_n} \leq (n + 1)|\mathcal{X}|2^{nR_n}$ \hspace{1cm} (7.38)

(7.39)

(7.40)
Then \( |A_n| = \sum_{P \in \mathcal{P}_n: H(P) \leq R_n} |T(P)| \leq \sum_{P \in \mathcal{P}_n: H(P) \leq R_n} 2^{nH(P)} \) \hspace{1cm} (7.38)

\[
\leq \sum_{P \in \mathcal{P}_n: H(P) \leq R_n} 2^{nR_n} \leq (n + 1)|\mathcal{X}|2^{nR_n} \hspace{1cm} \text{(7.39)}
\]

\[
= 2^n(R_n + |\mathcal{X}| \frac{\log(n+1)}{n}) \hspace{1cm} \text{(7.40)}
\]
Then

\[ |A_n| = \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} |T(P)| \leq \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} 2^{nH(P)} \quad (7.38) \]

\[ \leq \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} 2^{nR_n} \leq (n + 1)|X|2^{nR_n} \quad (7.39) \]

\[ = 2^n(R_n + |X|\frac{\log(n+1)}{n}) = 2^nR \quad (7.40) \]
Proof of theorem 7.4.3 continued.

Then

\[ |A_n| = \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} |T(P)| \leq \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} 2^{nH(P)} \quad (7.38) \]

\[ \leq \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} 2^{nR_n} \leq (n+1)|\mathcal{X}|2^{nR_n} \quad (7.39) \]

\[ = 2^n(R_n + |\mathcal{X}|rac{\log(n+1)}{n}) = 2^nR \quad (7.40) \]

Since \( |A_n| \leq 2^{nR} \), we can index \( A_n \) with \( nR \) bits.
Source Coding Theorem

... Proof of theorem 7.4.3 continued.

- Then
  \[|A_n| = \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} |T(P)| \leq \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} 2^{nH(P)} \] (7.38)

  \[\leq \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} 2^{nR_n} \leq (n + 1)|\mathcal{X}|2^{nR_n} \] (7.39)

  \[= 2^n(R_n + |\mathcal{X}| \frac{\log(n+1)}{n}) = 2^{nR} \] (7.40)

- Since \(|A_n| \leq 2^{nR}\), we can index \(A_n\) with \(nR\) bits.
- Let the encoder be:

  \[f_n(x_{1:n}) = \begin{cases} 
  \text{lexicographic index} \\
  \text{of } x_{1:n} \text{ in } A_n \\
  0 \\
  \end{cases} \quad \begin{array}{l}
  \text{if } x_{1:n} \in A_n \\
  \text{(i.e., if } H(P_{x_{1:n}}) \leq R_n) \\
  \text{else} \\
  \text{(i.e., if } H(P_{x_{1:n}}) > R_n) \\
  \end{array} \] (7.41)
Proof of theorem 7.4.3 continued.

Note: $f_n(\cdot)$ does not depend on the source distribution, only on the type, the lexicographic ordering, $R_n$, and $H$. 
Source Coding Theorem

... Proof of theorem 7.4.3 continued.

- Note: $f_n(\cdot)$ does not depend on the source distribution, only on the type, the lexicographic ordering, $R_n$, and $H$.
- Error occurs if $x_{1:n} \notin A_n$. 

...
... Proof of theorem 7.4.3 continued.

- Note: $f_n(\cdot)$ does not depend on the source distribution, only on the type, the lexicographic ordering, $R_n$, and $H$.
- 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
... 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 $n$).
... 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.

\[
H(P) < R_n
\]

\[
H(P) > R_n
\]

\[
H(P) = R_n
\]

Set of sequences that are encoded correctly.
... 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.
... Proof of theorem 7.4.3 continued.

- We upper bound $P_e^{(n)}$ and show it $\rightarrow 0$ as $n \rightarrow \infty$ when $R > H(Q)$. 
Source Coding Theorem

... Proof of theorem 7.4.3 continued.

- We upper bound $P_e^{(n)}$ and show it $\to 0$ as $n \to \infty$ when $R > H(Q)$.
- An error occurs when the sequence is not in $A_n$, thus

$$P_e^{(n)} = 1 - Q_n(A_n) = \sum_{P: H(P) > R} n Q_n(T(P)) \leq (n+1)|X|$$

(7.44)

...
Source Coding Theorem

... Proof of theorem 7.4.3 continued.

- We upper bound $P_e^{(n)}$ and show it $\to 0$ as $n \to \infty$ when $R > H(Q)$.
- An error occurs when the sequence is not in $A_n$, thus

$$P_e^{(n)}$$

(7.44)
Source Coding Theorem

... Proof of theorem 7.4.3 continued.

- We upper bound $P_e^{(n)}$ and show it $\to 0$ as $n \to \infty$ when $R > H(Q)$.
- An error occurs when the sequence is not in $A_n$, thus

$$P_e^{(n)} = 1 - Q^n(A_n)$$

(7.44)
Proof of theorem 7.4.3 continued.

- We upper bound $P_e^{(n)}$ and show it → 0 as $n \to \infty$ when $R > H(Q)$.
- An error occurs when the sequence is not in $A_n$, thus

$$P_e^{(n)} = 1 - Q^n(A_n) = \sum_{P: H(P) > R} Q^n(T(P))$$  \hspace{1cm} (7.42)

$$\leq (n+1)|X|^\max P: H(P) > R_n$$  \hspace{1cm} (7.43)

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 $H(Q) < R^n$.

$$\phi(Q) \phi(P) \text{ } R$$
Source Coding Theorem

... Proof of theorem 7.4.3 continued.

- We upper bound $P_e^{(n)}$ and show it $\to 0$ as $n \to \infty$ when $R > H(Q)$.
- An error occurs when the sequence is not in $A_n$, thus

$$P_e^{(n)} = 1 - Q^n(A_n) = \sum_{P : H(P) > R_n} Q^n(T(P)) \quad (7.42)$$

$$\leq (n + 1)^{|X|} \max_{P : H(P) > R_n} Q^n(T(P)) \quad (7.43)$$

$$\leq (n + 1)^{|X|} \max_{P : H(P) > R_n} Q^n(T(P)) \quad (7.44)$$
Source Coding Theorem

... Proof of theorem 7.4.3 continued.

- We upper bound $P_e^{(n)}$ and show it $\to 0$ as $n \to \infty$ when $R > H(Q)$.
- An error occurs when the sequence is not in $A_n$, thus

$$P_e^{(n)} = 1 - Q^n(A_n) = \sum_{P:H(P)>R_n} Q^n(T(P)) \quad (7.42)$$

$$\leq (n + 1)|\mathcal{X}| \max_{P:H(P)>R_n} Q^n(T(P)) \quad (7.43)$$

$$\leq (n + 1)|\mathcal{X}|2^{-n\left[\min_{P:H(P)>R_n} D(P||Q)\right]} \quad (7.44)$$
We upper bound $P_e^{(n)}$ and show it $\to 0$ as $n \to \infty$ when $R > H(Q)$.

An error occurs when the sequence is not in $A_n$, thus

$$P_e^{(n)} = 1 - Q^n(A_n) = \sum_{P: H(P) > R_n} Q^n(T(P))$$  \hspace{1cm} (7.42)

$$\leq (n + 1)|\mathcal{X}| \max_{P: H(P) > R_n} Q^n(T(P))$$  \hspace{1cm} (7.43)

$$\leq (n + 1)|\mathcal{X}| 2^{-n \left[ \min_{P: H(P) > R_n} D(P || Q) \right]}$$  \hspace{1cm} (7.44)

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

...
We upper bound $P_e^{(n)}$ and show it $\to 0$ as $n \to \infty$ when $R > H(Q)$.

An error occurs when the sequence is not in $A_n$, thus

$$P_e^{(n)} = 1 - Q^n(A_n) = \sum_{P : H(P) > R_n} Q^n(T(P)) \leq (n + 1)|\mathcal{X}| \max_{P : H(P) > R_n} Q^n(T(P)) \leq (n + 1)|\mathcal{X}| 2^{-n \min_{P : H(P) > R_n} D(P||Q)}$$

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 $H(Q) < R_n$
In Eq. (7.44) chose a \( P \) : \( H(P) > R_n \) for the current \( n \) (assuming there is one, if not we immediately get \( P_e = 0 \) in previous frame).
### Source Coding Theorem

... Proof of theorem 7.4.3 continued.

<table>
<thead>
<tr>
<th>$R_{n_0-2}$</th>
<th>$R_{n_0-1}$</th>
<th>$R_{n_0}$</th>
<th>$R_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H(Q)$</td>
<td>$H(P)$</td>
<td>$R$</td>
<td></td>
</tr>
</tbody>
</table>

- In Eq. (7.44) chose a $P : H(P) > R_n$ for the current $n$ (assuming there is one, if not we immediately get $P_e = 0$ in previous frame).
- $\Rightarrow H(P) > R_n > H(Q)$
Proof of theorem 7.4.3 continued.

- In Eq. (7.44) chose a $P : H(P) > R_n$ for the current $n$ (assuming there is one, if not we immediately get $P_e = 0$ in previous frame).
- $\Rightarrow H(P) > R_n > H(Q)$
- $\Rightarrow P \neq Q$
In Eq. (7.44) chose a $P : H(P) > R_n$ for the current $n$ (assuming there is one, if not we immediately get $P_e = 0$ in previous frame).

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

$\Rightarrow P \neq Q$

$\Rightarrow D(P\|Q) > 0$ for the chosen $P$
In Eq. (7.44) chose a $P : H(P) > R_n$ for the current $n$ (assuming there is one, if not we immediately get $P_e = 0$ in previous frame).

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

$\Rightarrow P \neq Q$

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

Thus, we get

$$P_e^{(n)} \leq (n + 1)|\mathcal{X}| \cdot 2^{-n \left[ \min_{P: H(P) > R_n} D(P||Q) \right]}$$ (7.45)
In Eq. (7.44) chose a $P : H(P) > R_n$ for the current $n$ (assuming there is one, if not we immediately get $P_e = 0$ in previous frame).

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

$\Rightarrow P \neq Q$

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

Thus, we get

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

(7.45)
\( R_{n_0-2} \quad R_{n_0-1} \quad R_{n_0} \quad R_n \)

\[ H(Q) \quad H(P) \quad R \]

- In Eq. (7.44) chose a \( P : H(P) > R_n \) for the current \( n \) (assuming there is one, if not we immediately get \( P_e = 0 \) in previous frame).
- \( \Rightarrow H(P) > R_n > H(Q) \)
- \( \Rightarrow P \neq Q \)
- \( \Rightarrow D(P \| Q) > 0 \) for the chosen \( P \)
- Thus, we get

\[
P_e^{(n)} \leq (n + 1)|\mathcal{X}| 2^{-n \left[ \min_{P : H(P) > R_n} D(P \| Q) \right]} \quad (7.45)
\]

\( \text{poly in } n \quad \text{exp. decreasing as } n \to \infty \)
In Eq. (7.44) chose a $P : H(P) > R_n$ for the current $n$ (assuming there is one, if not we immediately get $P_e = 0$ in previous frame).

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

$\Rightarrow P \neq Q$

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

Thus, we get

$$P_e^{(n)} \leq (n + 1)|\mathcal{X}| \cdot 2^{-n\left[\min_{P : H(P) > R_n} D(P\|Q)\right]}$$

(7.45)

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

Prof. Jeff Bilmes
Conversely, if $R < H(Q)$ then $P_e^{(n)} \rightarrow 1$ (left to HW)
Conversely, if $R < H(Q)$ then $P_e^n \rightarrow 1$ (left to HW)

So entropy is a good measure of the “information” in a source.
Conversely, if $R < H(Q)$ then $P_e^{(n)} \to 1$ (left to HW)

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

Can you compress a compressed file? $\text{gzip(\text{gzip(\text{gzip(...\text{gzip}(x))))})}$
Conversely, if $R < H(Q)$ then $P_e^{(n)} \to 1$ (left to HW)

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

Can you compress a compressed file? \( \text{gzip(gzip(gzip(...gzip(x))))} \)

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

- Conversely, if $R < H(Q)$ then $P_e^n \to 1$ (left to HW)
- So entropy is a good measure of the “information” in a source.
- Can you compress a compressed file? gzip(gzip(gzip(\ldots gzip(x))))
- Once a source is compressed to the the entropy rate, it is uncompressible, it is at its entropy limit.
- It looks like a random string of bits.
Source Coding Theorem

- Conversely, if $R < H(Q)$ then $P_e^n \rightarrow 1$ (left to HW)
- So entropy is a good measure of the “information” in a source.
- Can you compress a compressed file? $\text{gzip(gzip(gzip(\ldots \text{gzip}(x))))}$
- 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.
So far we’ve been talking about i.i.d. random variables, $X_1, X_2, \ldots$. 
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.
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?
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.
Stochastic Process

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

$$p(X_{1:n} = x_{1:n}) = p(X_{1+\ell:n+\ell} = x_{1:n})$$  (7.46)

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

\[
p(X_1:n = x_{1:n}) = p(X_{1+\ell}:n+\ell = x_{1:n})
\]  
(7.46)

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

\[
p(X_{n+1} = x_{n+1} | X_{1:n} = x_{1:n}) = p(X_{n+1} = x_{n+1} | X_{n} = x_{n})
\]  
(7.47)
### 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

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

(7.46)

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

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

(7.47)

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

$$p(X_{n+1} = b|X_n = a) = p(X_2 = b|X_1 = a) \ \forall a, b, n \quad (7.48)$$
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

$$p(X_{n+1} = b|X_n = a) = p(X_2 = b|X_1 = a) \; \forall a, b, n \quad (7.48)$$

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.
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.
**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.

- Recall, vector-matrix multiply $\mu_{n+1}^T = \mu_n^T P$ for state probability at time $n + 1$ given state probability at time $n$.

\[
p(x_{n+1}) = \sum_{x_n} p(x_n)p_{x_n,x_{n+1}} \tag{7.49}
\]
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.

- Recall, vector-matrix multiply $\mu_{n+1}^\top = \mu_n^\top P$ for state probability at time $n + 1$ given state probability at time $n$.

\[
p(x_{n+1}) = \sum_{x_n} p(x_n)p_{x_n,x_{n+1}} \quad (7.49)
\]

- (first order) Markov chain is stationary if $p(x_{n+1}) = p(x_n)$
A Markov chain is periodic if $d(i) > 1$ with

$$d(i) = \gcd\{n : p_{ii}(n) > 0\} \quad (7.50)$$

and $d(i)$ is called the period of state $i$.

- Note that this is the gcd (greatest common divisor) of the epochs at which return to the same state is possible.
- State $i$ is called periodic if $d(i) > 1$ and aperiodic if $d(i) = 1$.
- If $p_{ii}(n)$ is the probability of returning to state $i$ after $n$ steps starting from state $i$, then $p_{ii}(n) = 0$ unless $n$ is a multiple of $d(i)$. The value $d(i)$ is maximal with this property.
**Example:**

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

- Example:

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

- If \( \mu = [p_1 \, p_2]^T \) is stationary distribution then we must have that \( \mu^T P = \mu^T \).
Stochastic Process

Example:

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

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

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

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

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

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.
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. Hence, immediate past is sufficient to render more distant past irrelevant.
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. Hence, immediate past is sufficient to render more distant past irrelevant.
- (time) homogeneous: parameters of Markov chain don’t change with time.
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. Hence, 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.
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. Hence, 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.
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. Hence, 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.
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.
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
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

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

(7.51)

when it exists.
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

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

(7.51)

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

\[
H(X) = \lim_{n \to \infty} \frac{H(X_{x1:n})}{n} = \sum_{i=1}^{n} \frac{H(X_i)}{n} = H(X_1)
\]  

(7.52)
Examples

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

$$H(X) = \lim_{n \to \infty} \frac{H(X_{x1:n})}{n} = \frac{\sum_{i=1}^{n} H(X_i)}{n} = H(X_1) \quad (7.52)$$

- Independent but not identically distributed:

$$\lim_{n \to \infty} \frac{\sum_{i=1}^{n} H(X_i)}{n} = ? \quad (7.53)$$
in this case it might not exist.

Note, $2^k < \log \log i \leq 2^k + 1$ means that $\log \log i$ is between even and next greater odd integer, while $2^k + 1 < \log \log i \leq 2^k + 2$ means $\log \log i$ is between an odd and next greater even integer.
Examples

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

$$H(X) = \lim_{n \to \infty} \frac{H(X_{x1:n})}{n} = \frac{\sum_{i=1}^{n} H(X_i)}{n} = H(X_1)$$  \hspace{1cm} (7.52)

- Independent but not identically distributed:

$$\lim_{n \to \infty} \frac{\sum_{i=1}^{n} H(X_i)}{n} = ?$$  \hspace{1cm} (7.53)

in this case it might not exist.

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

$$p_i = \begin{cases} 0.5 & \text{if } 2k < \log \log i \leq 2k + 1 \\ 0 & \text{if } 2k + 1 < \log \log i \leq 2k + 2 \end{cases}$$  \hspace{1cm} (7.54)

Note, $2k < \log \log i \leq 2k + 1$ means that $\log \log i$ is between even and next greater odd integer, while $2k + 1 < \log \log i \leq 2k + 2$ means $\log \log i$ is between an odd and next greater even integer.
**Alternative Definition**

**Definition 7.6.2**

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

\[
H'(\mathcal{X}) \triangleq \lim_{n \to \infty} H(X_n | X_{n-1}, X_{n-2}, \ldots, X_1) \tag{7.55}
\]

assuming it exists.
Alternative Definition

Definition 7.6.2

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

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

(7.55)

assuming it exists.

Theorem 7.6.3

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

**Definition 7.6.2**

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

$$H'(\mathcal{X}) \triangleq \lim_{n \to \infty} H(X_n | X_{n-1}, X_{n-2}, \ldots, X_1)$$  \hspace{1cm} (7.55)

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.**

$$H(X_{n+1} | X_1, \ldots, X_n) \leq H(X_{n+1} | X_2, \ldots, X_n)$$  \hspace{1cm} (7.56)

$$= H(X_n | X_1, \ldots, X_{n-1})$$  \hspace{1cm} (7.57)
Alternative Definition

**Definition 7.6.2**

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

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

(7.55)

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 \( H'(X) \).

**Proof.**

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

(7.56)

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

(7.57)

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 \( b_n \to a \)
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).
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

\[
\lim_{n \to \infty} H(X_n | X_{n-1}, X_{n-2}, \ldots, X_1) \triangleq H'(\mathcal{X}) \tag{7.58}
\]

\[
= H(\mathcal{X}) \triangleq \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n) \tag{7.59}
\]
Entropy rates or entropy rate

**Theorem 7.6.4**

*We have that for stationary stochastic processes*

\[
\lim_{n \to \infty} H(X_n | X_{n-1}, X_{n-2}, \ldots, X_1) \triangleq H'(X) \quad (7.58)
\]

\[
= H(X) \triangleq \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n) \quad (7.59)
\]

**Proof.**

\[b_n = \frac{H(X_1, X_2, \ldots, X_n)}{n} = \frac{1}{n} \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1) = a_i \quad (7.60)\]

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

Note that for any stationary ergodic (loosely, time and ensemble averages are the same) process, we have

\[
-\frac{1}{n} \log p(x_1, \ldots, x_n) \to H(X) \quad (7.61)
\]
Note that for any stationary ergodic (loosely, time and ensemble averages are the same) process, we have

$$\frac{1}{n} \log p(x_1, \ldots, x_n) \to H(X)$$

(7.61)

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.

\begin{align*}
H(X) &= H(X_2 | X_1) = \beta \alpha + \beta H(\alpha) + \alpha \alpha + \beta H(\beta).
\end{align*}
When the process is a stationary Markov chain, entropy rate has a nice form.

That is

\[ H(X) = H'(X) = \lim_{n \to \infty} H(X_n|X_{n-1}, \ldots, X_1) \]  
\[ = \lim_{n \to \infty} H(X_n|X_{n-1}) = H(X_2|X_1) \]  
\[ = - \sum_{x_2, x_1} p(x_2, x_1) \log p(x_2|x_1) = \sum_{i} \mu_i \left[ - \sum_{j} p_{ij} \log p_{ij} \right] \]

where again \( \mu \) is the stationary distribution and \( p_{ij} \) is the transition probability from \( i \) to \( j \).
Entropy rate and stationary Markov chain

- When the process is a stationary Markov chain, entropy rate has a nice form.

  That is

  \[
  H(X) = H'(X) = \lim_{n \to \infty} H(X_n | X_{n-1}, \ldots, X_1) = \lim_{n \to \infty} H(X_n | X_{n-1}) = H(X_2 | X_1)
  \] (7.62)

  \[
  = - \sum_{x_2, x_1} p(x_2, x_1) \log p(x_2 | x_1) = \sum_i \mu_i \left[ - \sum_j p_{ij} \log p_{ij} \right]
  \] (7.63)

  where again \( \mu \) is the stationary distribution and \( p_{ij} \) is the transition probability from \( i \) to \( j \).

- Ex: previous figure \( H(X) = H(X_2 | X_1) = \frac{\beta}{\alpha + \beta} H(\alpha) + \frac{\alpha}{\alpha + \beta} H(\beta) \).