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, Source Coding
- L6 (10/14):
- L7 (10/16):
- 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”).
Homework

- Homework 1 on our assignment dropbox (https://canvas.uw.edu/courses/1319497/assignments), was due Tuesday, Oct 8th, 11:55pm.
- Homework 2 will be out by early next week.
Fano’s Inequality: Summary

- Consider the following situation where we send $X$ through a noisy channel, receive $Y$, and do further processing.

$$
\begin{array}{c}
X \xrightarrow{\text{Noisy Channel}} Y \xrightarrow{\text{processing } g(\cdot)} \hat{X}
\end{array}
$$

$\hat{X}$ is an estimate of $X$.

- An error if $X \neq \hat{X}$. How do we measure the error? With probability, $P_e \triangleq p(X \neq \hat{X})$.

- Intuitively, conditional entropy should tell us something about the error possibilities, in fact, we have

$$
H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|\hat{X}) \geq H(X|Y) \quad (4.34)
$$

Theorem 4.2.7 (Fano’s Inequality)
Fano’s Inequality

**Theorem 4.2.7**

\[
H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|\hat{X}) \geq H(X|Y)
\] (4.28)

- So \( P_e = 0 \) requires that \( H(X|Y) = 0 \! \)!
- Note, the theorem simplifies (and implies)
  \[
  1 + P_e \log(|\mathcal{X}|) \geq H(X|Y), \text{ or}
  \]
  \[
P_e \geq \frac{H(X|Y) - 1}{\log |\mathcal{X}|}
\] (4.29)

yielding a lower bound on the error.

- This will be used to prove the converse to Shannon’s coding theorem, i.e., that any code with probability of error \( \rightarrow 0 \) as the block length increases must have a rate \( R < C \) = the capacity of the channel (to be defined).
An interesting bound on probability of equality

**Lemma 4.3.1**

Let $X, X'$ be two independent r.v.s with $X \sim p(x)$ and $X' \sim r(x)$, with $x, x' \in X$ (same alphabet). L. bound on cross collision probability:

$$p(X = X') \geq \max \left( 2^{-H(p) - D(p||r)}, 2^{-H(r) - D(r||p)} \right) \quad (4.1)$$

**Proof.**

$$2^{-H(p) - D(p||r)} = 2 \sum_x p(x) \log p(x) + \sum_x p(x) \log \frac{r(x)}{p(x)} \quad (4.2)$$
An interesting bound on probability of equality

Lemma 4.3.1

Let \( X, X' \) be two independent r.v.s with \( X \sim p(x) \) and \( X' \sim r(x) \), with \( x, x' \in X \) (same alphabet). L. bound on cross collision probability:

\[
p(X = X') \geq \max \left( 2^{-H(p) - D(p||r)}, 2^{-H(r) - D(r||p)} \right) \tag{4.1}
\]

Proof.

\[
2^{-H(p) - D(p||r)} = 2^{\sum_x p(x) \log p(x) + \sum_x p(x) \log \frac{r(x)}{p(x)}} = 2^{\sum_x p(x) \log r(x)} \tag{4.2}\]

(4.6)
Lemma 4.3.1

Let $X, X'$ be two independent r.v.s with $X \sim p(x)$ and $X' \sim r(x)$, with $x, x' \in \mathcal{X}$ (same alphabet). L. bound on cross collision probability:

$$p(X = X') \geq \max \left(2^{-H(p) - D(p||r)}, 2^{-H(r) - D(r||p)}\right) \tag{4.1}$$

Proof.

$$2^{-H(p) - D(p||r)} = 2\sum_x p(x) \log p(x) + \sum_x p(x) \log \frac{r(x)}{p(x)} \tag{4.2}$$

$$= 2\sum_x p(x) \log r(x) \tag{4.3}$$

$$\leq \sum_x p(x) 2^{\log r(x)} \tag{4.4}$$

$$E[\phi(x)] \geq \phi(E(x)) \tag{4.6}$$
An interesting bound on probability of equality

Lemma 4.3.1

Let $X, X'$ be two independent r.v.s with $X \sim p(x)$ and $X' \sim r(x)$, with $x, x' \in \mathcal{X}$ (same alphabet). L. bound on cross collision probability:

$$p(X = X') \geq \max \left( 2^{-H(p) - D(p||r)}, 2^{-H(r) - D(r||p)}, \right)$$  \hspace{1cm} (4.1)

Proof.

$$2^{-H(p) - D(p||r)} = 2 \sum_x p(x) \log p(x) + \sum_x p(x) \log \frac{r(x)}{p(x)}$$  \hspace{1cm} (4.2)

$$= 2 \sum_x p(x) \log r(x)$$  \hspace{1cm} (4.3)

$$\leq \sum_x p(x) 2 \log r(x)$$  \hspace{1cm} (4.4)

$$= \sum_x p(x) r(x)$$  \hspace{1cm} (4.5)
An interesting bound on probability of equality

**Lemma 4.3.1**

Let $X, X'$ be two independent r.v.s with $X \sim p(x)$ and $X' \sim r(x)$, with $x, x' \in \mathcal{X}$ (same alphabet). L. bound on cross collision probability:

$$p(X = X') \geq \max \left( 2^{-H(p) - D(p||r)}, 2^{-H(r) - D(r||p)} \right)$$  \hspace{1cm} (4.1)

**Proof.**

$$2^{-H(p) - D(p||r)} = 2^{\sum_x p(x) \log p(x) + \sum_x p(x) \log \frac{r(x)}{p(x)}}$$  \hspace{1cm} (4.2)

$$= 2^{\sum_x p(x) \log r(x)}$$  \hspace{1cm} (4.3)

$$\leq \sum_x p(x) 2^{\log r(x)}$$  \hspace{1cm} (4.4)

$$= \sum_x p(x) r(x)$$  \hspace{1cm} (4.5)

$$= p(X = X')$$  \hspace{1cm} (4.6)
An interesting bound on probability of equality

Thus, taking \( p(x) = r(x) \), the probability that two i.i.d. random variables \( X, X' \) are the same can be bounded as follows:

\[
P(X = X') \geq 2^{-H(p)}
\]  

(4.7)
An interesting bound on probability of equality

- Thus, taking $p(x) = r(x)$, the probability that two i.i.d. random variables $X, X'$ are the same can be bounded as follows:

$$P(X = X') \geq 2^{-H(p)}$$

(4.7)

- $P(X = X') = \sum_x p(x)^2$ known as the probability of collision.
An interesting bound on probability of equality

- Thus, taking \( p(x) = r(x) \), the probability that two i.i.d. random variables \( X, X' \) are the same can be bounded as follows:

\[
P(X = X') \geq 2^{-H(p)} \tag{4.7}
\]

- \( P(X = X') = \sum_x p(x)^2 \) known as the probability of collision.
- Above improves standard collision probability bound

\[
\sum_x p(x)^2 = \sum_x [(p(x) - 1/n) + 1/n]^2 \tag{4.8}
\]

\[
= 1/n + \sum_x [(p(x) - 1/n)]^2 \geq 1/n \tag{4.9}
\]

so equality achieved when \( p(x) = 1/n \) (uniform distribution), also when \( P(X = X') = 2^{-H(p)} = 2^{-\log n} = \sqrt[2]{n} \).
An interesting bound on probability of equality

Thus, taking $p(x) = r(x)$, the probability that two i.i.d. random variables $X, X'$ are the same can be bounded as follows:

$$P(X = X') \geq 2^{-H(p)} \quad (4.7)$$

$P(X = X') = \sum_x p(x)^2$ known as the probability of collision.

Above improves standard collision probability bound

$$\sum_x p(x)^2 = \sum_x [(p(x) - 1/n) + 1/n]^2 \quad (4.8)$$

$$= 1/n + \sum_x [(p(x) - 1/n)]^2 \geq 1/n \quad (4.9)$$

so equality achieved when $p(x) = 1/n$ (uniform distribution), also when $P(X = X') = 2^{-H(p)} = 2^{-\log n}$.

Many other probabilistic quantities can be bounded in terms of entropic quantities as we will see throughout the course.
Weak Law of Large Numbers (WLLN)

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. r.v.s with $EX_i = \mu < \infty \ \forall i$. 

Written as $\frac{1}{n}S_n \xrightarrow{p} \mu$ (converges in probability) 

$\frac{1}{n}S_n$ gets as close to $\mu$ as we want and the variance of $\frac{1}{n}S_n$ gets arbitrarily small if $n$ gets big enough.
Weak Law of Large Numbers (WLLN)

- Let $X_1, X_2, \ldots$ be a sequence of i.i.d. r.v.s with $EX_i = \mu < \infty \ \forall i$.
- Let $S_n = \sum_{i=1}^{n} X_i$ be a partial sum of $n$ terms.
Weak Law of Large Numbers (WLLN)

- Let $X_1, X_2, \ldots$ be a sequence of i.i.d. r.v.s with $EX_i = \mu < \infty \ \forall i$.
- Let $S_n = \sum_{i=1}^{n} X_i$ be a partial sum of $n$ terms.
- Then, $\forall \epsilon > 0$, WLLN says that

$$p\left(\left|\frac{1}{n}S_n - \mu\right| > \epsilon\right) \to 0 \text{ as } n \to \infty \quad (4.10)$$
Weak Law of Large Numbers (WLLN)

- Let $X_1, X_2, \ldots$ be a sequence of i.i.d. r.v.s with $EX_i = \mu < \infty \ \forall i$.
- Let $S_n = \sum_{i=1}^{n} X_i$ be a partial sum of $n$ terms.
- Then, $\forall \epsilon > 0$, WLLN says that
  \[ p\left(\left| \frac{1}{n} S_n - \mu \right| > \epsilon \right) \to 0 \text{ as } n \to \infty \]  
  (4.10)

- Written as $\frac{1}{n} S_n \xrightarrow{p} \mu$ (converges in probability)
Weak Law of Large Numbers (WLLN)

- Let \(X_1, X_2, \ldots\) be a sequence of i.i.d. r.v.s with \(EX_i = \mu < \infty\ \forall i\).
- Let \(S_n = \sum_{i=1}^{n} X_i\) be a partial sum of \(n\) terms.
- Then, \(\forall \epsilon > 0\), WLLN says that
  \[
p(\left| \frac{1}{n} S_n - \mu \right| > \epsilon) \to 0 \text{ as } n \to \infty \tag{4.10}
  \]
- Written as \(\frac{1}{n} S_n \xrightarrow{p} \mu\) (converges in probability)
- \(\frac{1}{n} S_n\) gets as close to \(\mu\) as we want and the variance of \(\frac{1}{n} S_n\) gets arbitrarily small if \(n\) gets big enough.
Weak Law of Large Numbers (WLLN)

- Let $X_1, X_2, \ldots$ be a sequence of i.i.d. r.v.s with $EX_i = \mu < \infty \ \forall i$.
- Let $S_n = \sum_{i=1}^{n} X_i$ be a partial sum of $n$ terms.
- Then, $\forall \epsilon > 0$, WLLN says that
  \[
p\left(\frac{1}{n}S_n - \mu \right) > \epsilon \right) \to 0 \text{ as } n \to \infty \quad (4.10)
  \]

- Written as $\frac{1}{n}S_n \overset{p}{\to} \mu$ (converges in probability)
- $\frac{1}{n}S_n$ gets as close to $\mu$ as we want and the variance of $\frac{1}{n}S_n$ gets arbitrarily small if $n$ gets big enough.
- SLLN: If all r.v.s have finite 2nd moments, i.e., $E(X_i^2) < \infty$, then $\frac{1}{n} \sum_{i=1}^{n} X_i \to \mu \text{ a.s. and in the mean square (r = 2) sense.}$
Aside: modes of convergence

- \( X_n \rightarrow X \) almost surely (\( X_n \xrightarrow{\text{a.s.}} X \)) if the set
  \[
  \{ \omega \in \Omega : X_n(\omega) \rightarrow X(\omega) \text{ as } n \rightarrow \infty \}
  \] (4.11)
is an event with probability 1. So all such events combined together have probability 1, and any event left out has probability zero, according to the underlying probability measure.

- \( X_n \rightarrow X \) in the \( r^{\text{th}} \) mean (\( r \geq 1 \)), (written \( X_n \xrightarrow{r} X \)) if
  \[
  E|X_n^r| < \infty \quad \forall n, \quad \text{and} \quad E(|X_n - X|^r) \rightarrow 0 \text{ as } n \rightarrow \infty
  \] (4.12)

- \( X_n \rightarrow X \) in probability (written \( X_n \xrightarrow{p} X \)) if
  \[
  p(|X_n - X| > \epsilon) \rightarrow 0 \text{ as } n \rightarrow \infty, \quad \epsilon > 0
  \] (4.13)

- \( X_n \rightarrow X \) in distribution (written \( X_n \xrightarrow{D} X \)) if
  \[
  p(X_n \leq x) \rightarrow P(X < x) \text{ as } n \rightarrow \infty
  \] (4.14)
for all points \( x \) at which \( F_X(x) = p(X \leq x) \) is continuous.
Aside: modes of convergence, expanded definition

- Expanded description of convergence in probability.
Aside: modes of convergence, expanded definition

- Expanded description of convergence in probability.
- Let $X_1, X_2, \ldots$ be a sequence of i.i.d. r.v.s, with $EX = \mu < \infty$, with $S_n = \sum_{i=1}^{n} X_i$. 

Equivalently

$$p \left( \frac{1}{n} S_n - \mu \right) < \varepsilon$$

for $n > n_0$ (4.15)
Aside: modes of convergence, expanded definition

- Expanded description of convergence in probability.
- Let $X_1, X_2, \ldots$ be a sequence of i.i.d. r.v.s, with $EX = \mu < \infty$, with $S_n = \sum_{i=1}^{n} X_i$.
- Then, $\forall \epsilon > 0, \forall \delta > 0$, $\exists n_0$ such that for all $n > n_0$, we have

  $$p\left(\left|\frac{1}{n}S_n - \mu \right| > \epsilon\right) < \delta \text{ for } n > n_0$$

(4.15)
Aside: modes of convergence, expanded definition

- Expanded description of convergence in probability.
- Let $X_1, X_2, \ldots$ be a sequence of i.i.d. r.v.s, with $EX = \mu < \infty$, with $S_n = \sum_{i=1}^{n} X_i$.
- Then, $\forall \epsilon > 0, \forall \delta > 0$, $\exists n_0$ such that for all $n > n_0$, we have
  \[ p \left( \left| \frac{1}{n} S_n - \mu \right| > \epsilon \right) < \delta \text{ for } n > n_0 \]  
  \[ (4.15) \]

- Equivalently
  \[ p \left( \left| \frac{1}{n} S_n - \mu \right| \leq \epsilon \right) > 1 - \delta \text{ for } n > n_0 \]  
  \[ (4.16) \]
Some modes of convergence are stronger than others. That is

\[ \text{a.s.} \implies p \implies D \implies r \geq 1 \]
Aside: modes of convergence, implications

- Some modes of convergence are stronger than others. That is a.s.

\[ p \Rightarrow D \quad \forall r \geq 1 \]

- Also, if \( r > s \geq 1 \), then \( r \Rightarrow s \).
Aside: modes of convergence, implications

- Some modes of convergence are stronger than others. That is
  \[ \text{a.s.} \quad p \Rightarrow D \quad \forall r \geq 1 \]

- Also, if \( r > s \geq 1 \), then \( r \Rightarrow s \).

- Different versions of things like the law of large numbers differ only in the strength of their required modes of convergence.
Aside: modes of convergence, implications

- Some modes of convergence are stronger than others. That is
  \[ a.s. \quad r \Rightarrow p \Rightarrow D \quad \forall r \geq 1 \]
  
- Also, if \( r > s \geq 1 \), then \( r \Rightarrow s \).

- Different versions of things like the law of large numbers differ only in the strength of their required modes of convergence.

- Keep this in mind when considering the AEP which we turn to next.
Asymptotic Equipartition Property (AEP)

- At first, we’ll emphasize mostly intuition
At first, we’ll emphasize mostly intuition
Please read chapter 3 if you haven’t yet. If you have read it, read it again.
At first, we’ll emphasize mostly intuition

Please read chapter 3 if you haven’t yet. If you have read it, read it again.

Consider blocks of outcomes of random variables (i.e., random vectors of length $n$). $n = \text{the block length}$.
At first, we’ll emphasize mostly intuition

Please read chapter 3 if you haven’t yet. If you have read it, read it again.

Consider blocks of outcomes of random variables (i.e., random vectors of length $n$). $n =$ the block length.

Let $X_1, X_2, \ldots, X_n$ be i.i.d. r.v.s all distributed according to $p$ (we say $X_i \sim p(x)$.)
Asymptotic Equipartition Property (AEP)

- At first, we’ll emphasize mostly intuition
- Please read chapter 3 if you haven’t yet. If you have read it, read it again.
- Consider blocks of outcomes of random variables (i.e., random vectors of length $n$). $n = \text{the block length}$.
- Let $X_1, X_2, \ldots, X_n$ be i.i.d. r.v.s all distributed according to $p$ (we say $X_i \sim p(x)$).
- As before, $\forall i, X_i \in \{a_1, a_2, \ldots, a_K\}$, so $K$ possible symbols (alphabet or state space of size $K$).
Asymptotic Equipartition Property (AEP)

- At first, we’ll emphasize mostly intuition
- Please read chapter 3 if you haven’t yet. If you have read it, read it again.
- Consider blocks of outcomes of random variables (i.e., random vectors of length \( n \)). \( n \) = the block length.
- Let \( X_1, X_2, \ldots, X_n \) be i.i.d. r.v.s all distributed according to \( p \) (we say \( X_i \sim p(x) \)).
- As before, \( \forall i, X_i \in \{a_1, a_2, \ldots, a_K\} \), so \( K \) possible symbols (alphabet or state space of size \( K \)).
- For the \( n \) random variables \( (X_1, X_2, \ldots, X_n) \), there are \( K^n \) possible outcomes.
Towards AEP

- We wish to encode these $K^n$ outcomes with binary digit strings (i.e., code words) of length $m$. 

Hence, $M = 2^m$ possible code words.

We can represent the encoder as follows:

Example: English letters, would have $K = 26$ (alphabet size), a "source message" consists of $n$ letters.

We want to have a code word for every possible source message, must have what condition?

$$M = 2^m \approx K^n m$$
Towards AEP

- We wish to encode these $K^n$ outcomes with binary digit strings (i.e., code words) of length $m$. Hence, $M = 2^m$ possible code words.
Towards AEP

- We wish to encode these $K^n$ outcomes with binary digit strings (i.e., code words) of length $m$. Hence, $M = 2^m$ possible code words.
- We can represent the encoder as follows:

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
Towards AEP

- We wish to encode these $K^n$ outcomes with binary digit strings (i.e., code words) of length $m$. Hence, $M = 2^m$ possible code words.
- We can represent the encoder as follows:

  **Source messages**
  \[
  \{X_1, X_2, \ldots, X_n\}
  \]

  **Encoder**

  **Code words**
  \[
  \{Y_1, Y_2, \ldots, Y_m\}
  \]

  - $X_i \in \{a_1, a_2, \ldots, a_K\}$
  - $K^n$ possible messages
  - $n$ source letters in each source msg
  - $Y_i \in \{0, 1\}$
  - $2^m$ possible messages
  - $m$ total bits

- Example: English letters, would have $K = 26$ (alphabet size $K$), a “source message” consists of $n$ letters.
Towards AEP

- We wish to encode these $K^n$ outcomes with binary digit strings (i.e., code words) of length $m$. Hence, $M = 2^m$ possible code words.
- We can represent the encoder as follows:

  ![Encoder Diagram]

  - **Source messages**: $\{X_1, X_2, \ldots, X_n\}$
  - **Code words**: $\{Y_1, Y_2, \ldots, Y_m\}$

  - $X_i \in \{a_1, a_2, \ldots, a_K\}$
  - $K^n$ possible messages
  - $n$ source letters in each source msg

  - $Y_i \in \{0, 1\}$
  - $2^m$ possible messages
  - $m$ total bits

- Example: English letters, would have $K = 26$ (alphabet size $K$), a “source message” consists of $n$ letters.
- **We want to have a code word for every possible source message, must have what condition?**
Towards AEP

- We wish to encode these $K^n$ outcomes with binary digit strings (i.e., code words) of length $m$. Hence, $M = 2^m$ possible code words.
- We can represent the encoder as follows:

\[
\begin{align*}
\text{Source messages} & \quad \{X_1, X_2, \ldots, X_n\} \\
\text{Encoder} & \quad \rightarrow \\
\text{Code words} & \quad \{Y_1, Y_2, \ldots, Y_m\}
\end{align*}
\]

\[
\begin{align*}
X_i & \in \{a_1, a_2, \ldots, a_K\} \\
K^n & \text{ possible messages} \\
n & \text{ source letters in each source msg}
\end{align*}
\]

\[
\begin{align*}
Y_i & \in \{0, 1\} \\
2^m & \text{ possible messages} \\
m & \text{ total bits}
\end{align*}
\]

- Example: English letters, would have $K = 26$ (alphabet size $K$), a “source message” consists of $n$ letters.
- We want to have a code word for every possible source message, must have what condition?

\[
M = 2^m \geq K^n \Rightarrow m \geq n \log K
\]  

(4.17)
Towards AEP

- A question on rate: How many bits are used per source letter?
A question on rate: How many bits are used per source letter?

\[ R = \text{rate} = \frac{\log M}{n} = \frac{m}{n} \geq \log K \text{ bits per source letter} \quad (4.18) \]

Not surprising, e.g., for English need \( \lceil \log K \rceil = 5 \) bits.
A question on rate: How many bits are used per source letter?

\[ R = \text{rate} = \frac{\log M}{n} = \frac{m}{n} \geq \log K \text{ bits per source letter} \quad (4.18) \]

Not surprising, e.g., for English need \([\log K] = 5\) bits.

Question: can we use fewer than this bits per source letter (on average) and still have essentially no error?
Towards AEP

- A question on rate: How many bits are used per source letter?

\[ R = \text{rate} = \frac{\log M}{n} = \frac{m}{n} \geq \log K \text{ bits per source letter} \quad (4.18) \]

Not surprising, e.g., for English need \( \lceil \log K \rceil = 5 \) bits.

- Question: can we use fewer than this bits per source letter (on average) and still have “essentially” no error? Yes.
A question on rate: How many bits are used per source letter?

\[ R = \text{rate} = \frac{\log M}{n} = \frac{m}{n} \geq \log K \] bits per source letter (4.18)

Not surprising, e.g., for English need \([\log K] = 5\) bits.

Question: can we use fewer than this bits per source letter (on average) and still have essentially no error? Yes.

How? One way: some source messages would not have a code.
Towards AEP

- A question on rate: How many bits are used per source letter?

\[ R = \text{rate} = \frac{\log M}{n} = \frac{m}{n} \geq \log K \text{ bits per source letter} \quad (4.18) \]

Not surprising, e.g., for English need \([\log K] = 5\) bits.

- Question: can we use fewer than this bits per source letter (on average) and still have essentially no error? Yes.

- How? One way: some source messages would not have a code.

- I.e., code words only assigned to a subset of the source messages!
Towards AEP

- Any source message assigned to garbage would, if we wish to send that message, result in an error.
Towards AEP

- Any source message assigned to garbage would, if we wish to send that message, result in an error.
- Alternatively, and perhaps less distressingly, rather than throw some messages into the trash, we could assign to them long code words, and the non-garbage messages to short code words.

![Diagram showing source messages, short code words, and long code words]

In either case, if \( n \) gets big enough, we make the code such that the probability of getting one of those error source messages (or long-code-word source messages) very small!
Towards AEP

- Any source message assigned to garbage would, if we wish to send that message, result in an error.

- Alternatively, and perhaps less distressingly, rather than throw some messages into the trash, we could assign to them long code words, and the non-garbage messages to short code words.

  ![Source messages diagram]

- In either case, if \( n \) gets big enough, we make the code such that the probability of getting one of those error source messages (or long-code-word source messages) very small!
Probability of Source Words

- The probability of a source word can be expressed

\[ p(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} p(X_i = x_i) \]  

(4.19)
Probability of Source Words

- The probability of a source word can be expressed

\[ p(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} p(X_i = x_i) \]  \hspace{1cm} (4.19)

- Recall, the Shannon/Hartley information about an event \( \{X = x\} \) is 

\[ - \log p(x) = I(x) \], so information of joint event \( \{x_1, x_2, \ldots, x_n\} \) is:

\[ I(x_1, x_2, \ldots, x_n) = - \log p(x_1, x_2, \ldots, x_n) \]  \hspace{1cm} (4.20)

\[ = - \log \prod_{i=1}^{n} p(x_i) = \sum_{i=1}^{n} - \log p(x_i) = \sum_{i=1}^{n} I(x_i) \]  \hspace{1cm} (4.21)
The probability of a source word can be expressed

\[ p(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} p(X_i = x_i) \quad (4.19) \]

Recall, the Shannon/Hartley information about an event \( \{X = x\} \) is \(- \log p(x) = I(x)\), so information of joint event \( \{x_1, x_2, \ldots, x_n\} \) is:

\[ I(x_1, x_2, \ldots, x_n) = - \log p(x_1, x_2, \ldots, x_n) \quad (4.20) \]

\[ = - \log \prod_{i=1}^{n} p(x_i) = \sum_{i=1}^{n} - \log p(x_i) = \sum_{i=1}^{n} I(x_i) \quad (4.21) \]

Also note that \( EI(X) = H(X) \).
Probability of Source Words

The probability of a source word can be expressed as:

$$p(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} p(X_i = x_i)$$

(4.19)

Recall, the Shannon/Hartley information about an event \( \{X = x\}\) is

$$-\log p(x) = I(x),$$

so information of joint event \( \{x_1, x_2, \ldots, x_n\}\) is:

$$I(x_1, x_2, \ldots, x_n) = -\log p(x_1, x_2, \ldots, x_n)$$

(4.20)

$$= -\log \prod_{i=1}^{n} p(x_i) = \sum_{i=1}^{n} -\log p(x_i) = \sum_{i=1}^{n} I(x_i)$$

(4.21)

Also note that \( EI(X) = H(X) \).

The WLLN says that \( \frac{1}{n} S_n \xrightarrow{p} \mu \), where \( S_n \) is the sum of i.i.d. r.v.s with mean \( \mu = EX_i \).
Probability of Source Words

- The probability of a source word can be expressed

\[ p(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} p(X_i = x_i) \quad (4.19) \]

- Recall, the Shannon/Hartley information about an event \( \{X = x\} \) is

\[ -\log p(x) = I(x), \] so information of joint event \( \{x_1, x_2, \ldots, x_n\} \) is:

\[ I(x_1, x_2, \ldots, x_n) = -\log p(x_1, x_2, \ldots, x_n) \quad (4.20) \]

\[ = -\log \prod_{i=1}^{n} p(x_i) = \sum_{i=1}^{n} -\log p(x_i) = \sum_{i=1}^{n} I(x_i) \]

\[ (4.21) \]

- Also note that \( EI(X) = H(X) \).

- The WLLN says that \( \frac{1}{n} S_n \xrightarrow{p} \mu \), where \( S_n \) is the sum of i.i.d. r.v.s with mean \( \mu = EX_i \).

- So, \( I(X_i) \) is also a random variable with mean \( H(X) \).
WLLN and entropy

Combining the above, we get

\[
\frac{1}{n} \sum_{i=1}^{n} I(X_i) \xrightarrow{p} H(X)
\]

(4.22)
Combining the above, we get

\[ \frac{1}{n} \sum_{i=1}^{n} I(X_i) \xrightarrow{p} H(X) \quad (4.22) \]

Thus, if \( n \) is big enough, we have that (this is where it gets cool 😊)

\[ \frac{1}{n} \sum_{i=1}^{n} I(x_i) \approx H(X) \text{ when } \forall i, x_i \sim p(x) \quad (4.23) \]
WLLN and entropy

Combining the above, we get

$$\frac{1}{n} \sum_{i=1}^{n} I(X_i) \xrightarrow{p} H(X) \quad (4.22)$$

Thus, if $n$ is big enough, we have that (this is where it gets cool 😊)

$$\frac{1}{n} \sum_{i=1}^{n} I(x_i) \approx H(X) \text{ when } \forall i, x_i \sim p(x) \quad (4.23)$$

$$\Rightarrow \quad -\frac{1}{n} \sum_{i=1}^{n} \log p(x_i) \approx H(X) \quad (4.24)$$

(4.27)
Combining the above, we get

\[ \frac{1}{n} \sum_{i=1}^{n} I(X_i) \xrightarrow{p \rightarrow \infty} H(X) \quad (4.22) \]

Thus, if \( n \) is big enough, we have that (this is where it gets cool 😊)

\[ \frac{1}{n} \sum_{i=1}^{n} I(x_i) \approx H(X) \text{ when } \forall i, x_i \sim p(x) \quad (4.23) \]

\[ - \frac{1}{n} \sum_{i=1}^{n} \log p(x_i) \approx H(X) \quad (4.24) \]

\[ - \log \prod_{i=1}^{n} p(x_i) \approx nH(X) \quad (4.25) \]

(4.27)
Combining the above, we get
\[
\frac{1}{n} \sum_{i=1}^{n} I(X_i) \xrightarrow{p} H(X) \quad (4.22)
\]

Thus, if \( n \) is big enough, we have that (this is where it gets cool 😊)
\[
\frac{1}{n} \sum_{i=1}^{n} I(x_i) \approx H(X) \text{ when } \forall i, x_i \sim p(x) \quad (4.23)
\]
\[
- \frac{1}{n} \sum_{i=1}^{n} \log p(x_i) \approx H(X) \quad (4.24)
\]
\[
- \log \prod_{i=1}^{n} p(x_i) \approx nH(X) \quad (4.25)
\]
\[
- \log p(x_1, x_2, \ldots, x_n) \approx nH(X) \quad (4.26)
\]
\[
(4.27)
\]
WLLN and entropy

• Combining the above, we get

\[ \frac{1}{n} \sum_{i=1}^{n} I(X_i) \xrightarrow{p} H(X) \quad (4.22) \]

• Thus, if \( n \) is big enough, we have that (this is where it gets cool 😊)

\[ \frac{1}{n} \sum_{i=1}^{n} I(x_i) \approx H(X) \quad \text{when } \forall i, x_i \sim p(x) \quad (4.23) \]

\[ - \frac{1}{n} \sum_{i=1}^{n} \log p(x_i) \approx H(X) \quad (4.24) \]

\[ - \log \prod_{i=1}^{n} p(x_i) \approx nH(X) \quad (4.25) \]

\[ - \log p(x_1, x_2, \ldots, x_n) \approx nH(X) \quad (4.26) \]

\[ p(x_1, \ldots, x_n) \approx 2^{-nH(X)} \quad (4.27) \]
We repeat this last equation: When $n$ is large enough, we have

$$p(x_1, \ldots, x_n) \approx 2^{-nH(X)} \quad \text{when} \quad \forall i, x_i \sim p(x) \quad (4.28)$$

Note, perhaps somewhat startlingly, r.h.s. is the probability and it does not depend on the specific sequence instance $x_1, x_2, \ldots, x_n$!

So, if $n$ gets large enough, pretty much all sequences that happen have the same probability . . . and that probability is equal to $2^{-nH(X)}$. Those sequences that have that probability (which means pretty much all of them that occur) are called the typical sequences, represented by the set $A$. 
We repeat this last equation: When $n$ is large enough, we have

$$p(x_1, \ldots, x_n) \approx 2^{-nH(X)} \quad \text{when} \quad \forall i, x_i \sim p(x)$$

(4.28)

Note, perhaps somewhat startlingly, r.h.s. is the probability and it does not depend on the specific sequence instance $x_1, x_2, \ldots, x_n$!
Towards AEP

- We repeat this last equation: When $n$ is large enough, we have

$$p(x_1, \ldots, x_n) \approx 2^{-nH(X)} \quad \text{when} \quad \forall i, x_i \sim p(x)$$

(4.28)

- Note, perhaps somewhat startlingly, r.h.s. is the probability and it does not depend on the specific sequence instance $x_1, x_2, \ldots, x_n$!

- So, if $n$ gets large enough, pretty much all sequences that happen have the same probability ...
Towards AEP

- We repeat this last equation: When $n$ is large enough, we have

$$ p(x_1, \ldots, x_n) \approx 2^{-nH(X)} \quad \text{when} \quad \forall i, x_i \sim p(x) \quad (4.28) $$

- Note, perhaps somewhat startlingly, r.h.s. is the probability and it does not depend on the specific sequence instance $x_1, x_2, \ldots, x_n$!

- So, if $n$ gets large enough, pretty much all sequences that happen have the same probability . . .

- . . . and that probability is equal to $2^{-nH}$.
We repeat this last equation: When \( n \) is large enough, we have

\[
p(x_1, \ldots, x_n) \approx 2^{-nH(X)} \quad \text{when} \quad \forall i, x_i \sim p(x)
\]  

(4.28)

Note, perhaps somewhat startlingly, r.h.s. is the probability and it does not depend on the specific sequence instance \( x_1, x_2, \ldots, x_n \)!

So, if \( n \) gets large enough, pretty much all sequences that happen have the same probability . . .

. . . and that probability is equal to \( 2^{-nH} \).

Those sequences that have that probability (which means pretty much all of them that occur) are called the typical sequences, represented by the set \( A \).
Said again: Almost all events are almost equally probable

- If $X_1, X_2, \ldots, X_n$ are i.i.d. and $X_i \sim p(x)$ for all $i$, and
Said again: Almost all events are almost equally probable

- If $X_1, X_2, \ldots, X_n$ are i.i.d. and $X_i \sim p(x)$ for all $i$, and
- if $n$ is large enough,
Said again: Almost all events are almost equally probable

- If $X_1, X_2, \ldots, X_n$ are i.i.d. and $X_i \sim p(x)$ for all $i$, and
- if $n$ is large enough,
- then for any sample $x_1, x_2, \ldots, x_n$

\[ x_i \sim p(x) \quad \forall i. \]
Said again: Almost all events are almost equally probable

- If $X_1, X_2, \ldots, X_n$ are i.i.d. and $X_i \sim p(x)$ for all $i$, and
- if $n$ is large enough,
- then for any sample $x_1, x_2, \ldots, x_n$
- The probability of the sample is essentially independent of the sample, i.e.,

$$p(x_1, \ldots, x_n) \approx 2^{-nH(X)} \quad (4.29)$$

where $H(X)$ is the entropy of $p(x)$. 
Said again: Almost all events are almost equally probable

- If $X_1, X_2, \ldots, X_n$ are i.i.d. and $X_i \sim p(x)$ for all $i$, and
- if $n$ is large enough,
- then for any sample $x_1, x_2, \ldots, x_n$
- The probability of the sample is essentially independent of the sample, i.e.,

$$p(x_1, \ldots, x_n) \approx 2^{-nH(X)} \tag{4.29}$$

where $H(X)$ is the entropy of $p(x)$.

- Thus, there can only be $2^{nH}$ such samples, and it may be that $2^{nH} \ll K^n$
Said again: Almost all events are almost equally probable

- If $X_1, X_2, \ldots, X_n$ are i.i.d. and $X_i \sim p(x)$ for all $i$, and
- if $n$ is large enough,
- then for any sample $x_1, x_2, \ldots, x_n$
- The probability of the sample is essentially independent of the sample, i.e.,

\[ p(x_1, \ldots, x_n) \approx 2^{-nH(X)} \tag{4.29} \]

where $H(X)$ is the entropy of $p(x)$.

- Thus, there can only be $2^{nH}$ such samples, and it may be that $2^{nH} \ll K^n$
- Those samples that will happen are called typical, and they are represented by $A_{\epsilon}^{(n)}$. 

---

Prof. Jeff Bilmes
EE514a/Fall 2019/Info. Theory I – Lecture 4 - Oct 7th, 2019
Said again: Almost all events are almost equally probable

- If $X_1, X_2, \ldots, X_n$ are i.i.d. and $X_i \sim p(x)$ for all $i$, and
- if $n$ is large enough,
- then for any sample $x_1, x_2, \ldots, x_n$
- The probability of the sample is essentially independent of the sample, i.e.,

$$p(x_1, \ldots, x_n) \approx 2^{-nH(X)}$$

(4.29)

where $H(X)$ is the entropy of $p(x)$.

- Thus, there can only be $2^{nH}$ such samples, and it may be that
  $2^{nH} \ll K^n$

- Those samples that will happen are called typical, and they are represented by $A_\epsilon^{(n)}$.

- Thus, a large portion of $\mathcal{X}^n$ essentially won’t happen, i.e., could be that $2^{nH} \approx |A_\epsilon^{(n)}| \ll |\mathcal{X}^n| = K^n$. 

Prof. Jeff Bilmes
The Typical Set

Let $A_{\infty}^{(n)}$ be the set of typical sequences (i.e., those with probability $2^{-nH(X)}$.}
The Typical Set

- Let $A_{\epsilon}^{(n)}$ be the set of typical sequences (i.e., those with probability $2^{-nH(X)}$).

- If “all” events have the same probability $p$, then there are $1/p$ of them.
The Typical Set

- Let $A_{\varepsilon}^{(n)}$ be the set of typical sequences (i.e., those with probability $2^{-nH(X)}$).
- If “all” events have the same probability $p$, then there are $1/p$ of them.
- So the number of typical sequences is

$$|A_{\varepsilon}^{(n)}| \approx 2^{nH(X)}. \quad (4.30)$$
The Typical Set

- Let $A^{(n)}_\varepsilon$ be the set of typical sequences (i.e., those with probability $2^{-nH(X)}$).
- If “all” events have the same probability $p$, then there are $1/p$ of them.
- So the number of typical sequences is
  \[ |A^{(n)}_\varepsilon| \approx 2^{nH(X)}. \] (4.30)
- Thus, to represent (or code for) the typical sequences, we need only $nH(X)$ bits, so we take
  \[ m = nH(X) \ll n \log K \] (4.31)
  in the encoder model. Thus, the rate of the code is $H(X)$. 

<table>
<thead>
<tr>
<th>Source messages ${X_1, X_2, \ldots, X_n}$</th>
<th>Encoder</th>
<th>Code words ${Y_1, Y_2, \ldots, Y_m}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_i \in {a_1, a_2, \ldots, a_K}$</td>
<td></td>
<td>$Y_i \in {0, 1}$</td>
</tr>
<tr>
<td>$K^n$ possible messages</td>
<td></td>
<td>$2^m$ possible messages</td>
</tr>
<tr>
<td>$n$ source letters in each source msg</td>
<td></td>
<td>$m$ total bits</td>
</tr>
</tbody>
</table>

Prof. Jeff Bilmes
EE514a/Fall 2019/Info. Theory I – Lecture 4 - Oct 7th, 2019
Code Only The Typical Set

- If we take $m = nH$, then the average number of bits per source alphabet letter is (i.e., the rate becomes):

$$\text{Rate now} = \frac{m}{n} = H,$$

which could be $\leq \log K$ \hspace{1cm} (4.32)
If we take $m = nH$, then the average number of bits per source alphabet letter is (i.e., the rate becomes):

$$\frac{m}{n} = H,$$

which could be $\leq \log K$ \hspace{1cm} (4.32)

So to summarize, we have three uses and interpretations of entropy here for source coding.
If we take $m = nH$, then the average number of bits per source alphabet letter is (i.e., the rate becomes):

$$\frac{m}{n} = H,$$

which could be $\leq \log K$ \hfill (4.32)

So to summarize, we have three uses and interpretations of entropy here for source coding.

1. The probability of a typical sequence is $2^{-nH(X)}$. 
If we take $m = nH$, then the average number of bits per source alphabet letter is (i.e., the rate becomes):

$$\frac{m}{n} = H,$$

which could be $\leq \log K$ \hspace{1cm} (4.32)

So to summarize, we have three uses and interpretations of entropy here for source coding.

1. The probability of a typical sequence is $2^{-nH(X)}$.
2. The number of typical sequences is $2^{nH(X)}$. 
If we take \( m = nH \), then the average number of bits per source alphabet letter is (i.e., the rate becomes):

\[
\frac{m}{n} = H, \quad \text{which could be } \leq \log K \quad (4.32)
\]

So to summarize, we have three uses and interpretations of entropy here for source coding.

1. The probability of a typical sequence is \( 2^{-nH(X)} \).
2. The number of typical sequences is \( 2^{nH(X)} \).
3. Number of bits per source symbol is \( H(X) \), when we code only for the typical set.
AEP setup

- Consider Bernoulli trials $X_1, X_2, \ldots, X_n$ i.i.d. with $p(X_i = 1) = p = 1 - p(X_i = 0)$. 

The probability of a single sequence is 

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i)(1-p_i)^{1-x_i} \quad (4.33)$$

There are $2^n$ possible sequences. Do they all have the same probability? 

No, consider when $p = 0.1$, $(1-p) = 0.9$. The sequence of all zeros is much more likely.

What is the most probable sequence? 

When $p = 0.1$, these sequence of all 0s.

Do the sequences that collectively have "any" probability all have the same probability?

Depends what we mean by "any", but for small $n$, no. But as $n$ gets large, something funny happens and "yes" becomes a more appropriate answer.
AEP setup

- Consider Bernoulli trials $X_1, X_2, \ldots, X_n$ i.i.d. with $p(X_i = 1) = p = 1 - p(X_i = 0)$.
- The probability of a single sequence is

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p^x_i (1 - p_i)^{1-x_i} = p^{\sum_i x_i} (1 - p)^{n-\sum_i x_i}$$

(4.33)
AEP setup

- Consider Bernoulli trials $X_1, X_2, \ldots, X_n$ i.i.d. with $p(X_i = 1) = p = 1 - p(X_i = 0)$.
- The probability of a single sequence is

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p^{x_i} (1 - p_i)^{1-x_i} = p\sum x_i (1 - p)^{n-\sum x_i}$$

(4.33)

- There are $2^n$ possible sequences.
AEP setup

- Consider Bernoulli trials $X_1, X_2, \ldots, X_n$ i.i.d. with $p(X_i = 1) = p = 1 - p(X_i = 0)$.
- The probability of a single sequence is
  
  $$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p^{x_i} (1 - p_i)^{1-x_i} = p^\sum_i x_i (1 - p)^{n-\sum_i x_i}$$

  (4.33)

- There are $2^n$ possible sequences.
- Do they all have the same probability?
AEP setup

- Consider Bernoulli trials $X_1, X_2, \ldots, X_n$ i.i.d. with $p(X_i = 1) = p = 1 - p(X_i = 0)$.
- The probability of a single sequence is

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p^{x_i} (1 - p_i)^{1-x_i} = p^{\sum_i x_i} (1 - p)^{n-\sum_i x_i} \quad (4.33)$$

- There are $2^n$ possible sequences.
- Do they all have the same probability? No, consider when $p = 0.1$, $(1 - p) = 0.9$. The sequence of all zeros is much more likely.
AEP setup

- Consider Bernoulli trials \( X_1, X_2, \ldots, X_n \) i.i.d. with \( p(X_i = 1) = p = 1 - p(X_i = 0) \).
- The probability of a single sequence is

\[
p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p^{x_i} (1 - p_i)^{1-x_i} = p^{\sum_i x_i} (1 - p)^{n-\sum_i x_i}
\]

(4.33)

- There are \( 2^n \) possible sequences.
- Do they all have the same probability? No, consider when \( p = 0.1 \), \( (1 - p) = 0.9 \). The sequence of all zeros is much more likely.
- What is the most probable sequence?
AEP setup

Consider Bernoulli trials $X_1, X_2, \ldots, X_n$ i.i.d. with $p(X_i = 1) = p = 1 - p(X_i = 0)$.

The probability of a single sequence is

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p^{x_i} (1 - p_i)^{1-x_i} = p^\sum_i x_i (1 - p)^{n-\sum_i x_i}$$

(4.33)

There are $2^n$ possible sequences.

Do they all have the same probability? No, consider when $p = 0.1$, $(1 - p) = 0.9$. The sequence of all zeros is much more likely.

What is the most probable sequence? When $p = 0.1$, the sequence of all 0s.
Consider Bernoulli trials $X_1, X_2, \ldots, X_n$ i.i.d. with $p(X_i = 1) = p = 1 - p(X_i = 0)$.

The probability of a single sequence is

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p^{x_i} (1 - p_i)^{1-x_i} = p^{\sum_{i} x_i} (1 - p)^{n-\sum_{i} x_i}$$

There are $2^n$ possible sequences.

Do they all have the same probability? No, consider when $p = 0.1$, $(1 - p) = 0.9$. The sequence of all zeros is much more likely.

What is the most probable sequence? When $p = 0.1$, the sequence of all 0s.

Do the sequences that collectively have “any” probability all have the same probability?
AEP setup

- Consider Bernoulli trials \( X_1, X_2, \ldots, X_n \) i.i.d. with 
  \[ p(X_i = 1) = p = 1 - p(X_i = 0). \]
- The probability of a single sequence is
  \[
p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p^{x_i} (1 - p_i)^{1-x_i} = p^{\sum_i x_i} (1 - p)^{n-\sum_i x_i}
\]
  (4.33)
- There are \( 2^n \) possible sequences.
- Do they all have the same probability? No, consider when \( p = 0.1, (1 - p) = 0.9 \). The sequence of all zeros is much more likely.
- What is the most probable sequence? When \( p = 0.1 \), the sequence of all 0s.
- Do the sequences that collectively have “any” probability all have the same probability? Depends what we mean by “any”, but for small \( n \), no. But as \( n \) gets large, something funny happens and “yes” becomes a more appropriate answer.
**AEP setup**

- Notational reminder: \( H = H(X) \) is the entropy of a single random variable distributed as \( p(x) \).
AEP setup

- Notational reminder: $H = H(X)$ is the entropy of a single random variable distributed as $p(x)$.
- Can we predict the probability that a particular sequence has a particular probability value? I.e.,

$$\Pr(p(X_1, X_2, \ldots, X_n) = \alpha) = ?$$

(4.34)
AEP setup

- Notational reminder: $H = H(X)$ is the entropy of a single random variable distributed as $p(x)$.
- Can we predict the probability that a particular sequence has a particular probability value? I.e.,

$$\Pr(p(X_1, X_2, \ldots, X_n) = \alpha) = ?$$

(4.34)

- Note: “$p(X_1, X_2, \ldots, X_n)$” is a random variable! It is a random probability, and it is a true random variable since it is a probability that is a function of a set of random variables.
Notational reminder: $H = H(X)$ is the entropy of a single random variable distributed as $p(x)$.

Can we predict the probability that a particular sequence has a particular probability value? I.e.,

$$\Pr(p(X_1, X_2, \ldots, X_n) = \alpha) = ?$$  \hspace{1cm} (4.34)

Note: “$p(X_1, X_2, \ldots, X_n)$” is a random variable! It is a random probability, and it is a true random variable since it is a probability that is a function of a set of random variables.

It turns out that

$$\Pr(p(X_1, X_2, \ldots, X_n) \approx 2^{-nH}) \approx 1$$  \hspace{1cm} (4.35)

if $n$ is large enough.
AEP setup

- Notational reminder: $H = H(X)$ is the entropy of a single random variable distributed as $p(x)$.
- Can we predict the probability that a particular sequence has a particular probability value? I.e.,

$$\Pr(p(X_1, X_2, \ldots, X_n) = \alpha) = ?$$  \hspace{1cm} (4.34)

- Note: “$p(X_1, X_2, \ldots, X_n)$” is a random variable! It is a random probability, and it is a true random variable since it is a probability that is a function of a set of random variables.
- It turns out that

$$\Pr(p(X_1, X_2, \ldots, X_n) \approx 2^{-nH}) \approx 1$$  \hspace{1cm} (4.35)

if $n$ is large enough.

- In English, this can be read as: almost all events (that occur collectively with any appreciable probability) are all equally likely.
Ex: Bernoulli trials

Let $S_n \sim \text{Binomial}(n, p)$ with $S_n = X_1 + X_2 + \cdots + X_n$, $X_i \sim \text{Bernoulli}(p)$. 

$np$ is the expected number of 1's.

$nq$ is the expected number of 0's.
Ex: Bernoulli trials

- Let $S_n \sim \text{Binomial}(n, p)$ with $S_n = X_1 + X_2 + \cdots + X_n$, $X_i \sim \text{Bernoulli}(p)$.
- Hence, $ES_n = np$ and $\text{var}(S) = npq$, and

$$p(S_n = k) = \binom{n}{k} p^k q^{n-k} \tag{4.36}$$
Ex: Bernoulli trials

- Let $S_n \sim \text{Binomial}(n, p)$ with $S_n = X_1 + X_2 + \cdots + X_n$, $X_i \sim \text{Bernoulli}(p)$.
- Hence, $ES_n = np$ and $\text{var}(S) = npq$, and

$$p(S_n = k) = \binom{n}{k} p^k q^{n-k} \quad (4.36)$$

- Then, the expression $2^{-nH}$ can be viewed in an intuitive way. I.e.,

$$2^{-nH(p)} = 2^{-n(-p \log p - (1-p) \log(1-p))} \quad (4.37)$$
Ex: Bernoulli trials

- Let $S_n \sim \text{Binomial}(n, p)$ with $S_n = X_1 + X_2 + \cdots + X_n$, $X_i \sim \text{Bernoulli}(p)$.
- Hence, $ES_n = np$ and $\text{var}(S) = npq$, and
  
  $$p(S_n = k) = \binom{n}{k} p^k q^{n-k} \quad (4.36)$$

- Then, the expression $2^{-nH}$ can be viewed in an intuitive way. I.e.,
  
  $$2^{-nH(p)} = 2^{-n(-p \log p - (1-p) \log(1-p))} \quad (4.37)$$
  $$= 2^\log p^{np} + \log(1-p)^n(1-p) \quad (4.38)$$
  $$= 2^{np + \log(1-p)^n(1-p)} \quad (4.39)$$
Ex: Bernoulli trials

- Let $S_n \sim \text{Binomial}(n, p)$ with $S_n = X_1 + X_2 + \cdots + X_n$, $X_i \sim \text{Bernoulli}(p)$.
- Hence, $ES_n = np$ and $\text{var}(S) = npq$, and

$$p(S_n = k) = \binom{n}{k} p^k q^{n-k}$$  \hspace{1cm} (4.36)

- Then, the expression $2^{-nH}$ can be viewed in an intuitive way. I.e.,

$$2^{-nH(p)} = 2^{-n(-p \log p - (1-p) \log (1-p))}$$  \hspace{1cm} (4.37)

$$= 2^{\log p^{np} + \log (1-p)^n (1-p)}$$  \hspace{1cm} (4.38)

$$= p^{np} q^{nq} \text{ where } q = 1 - p$$  \hspace{1cm} (4.39)
Ex: Bernoulli trials

- Let $S_n \sim \text{Binomial}(n, p)$ with $S_n = X_1 + X_2 + \cdots + X_n$, $X_i \sim \text{Bernoulli}(p)$.
- Hence, $E S_n = np$ and $\text{var}(S) = npq$, and

$$p(S_n = k) = \binom{n}{k} p^k q^{n-k}$$ (4.36)

- Then, the expression $2^{-nH}$ can be viewed in an intuitive way. I.e.,

$$2^{-nH(p)} = 2^{-n(-p \log p - (1-p) \log(1-p))}$$ (4.37)

$$= 2^{\log p^{np} + \log(1-p)^n(1-p)}$$ (4.38)

$$= p^{np} q^{nq} \text{ where } q = 1 - p$$ (4.39)

Here $H = H(p)$ the binary entropy with probability $p$. 
Ex: Bernoulli trials

- Let $S_n \sim \text{Binomial}(n, p)$ with $S_n = X_1 + X_2 + \cdots + X_n$, $X_i \sim \text{Bernoulli}(p)$.

- Hence, $ES_n = np$ and $\text{var}(S) = npq$, and

  $$p(S_n = k) = \binom{n}{k} p^k q^{n-k}$$  \hfill (4.36)

- Then, the expression $2^{-nH}$ can be viewed in an intuitive way. I.e.,

  $$2^{-nH(p)} = 2^{-n(-p \log p - (1-p) \log(1-p))}$$ \hfill (4.37)
  $$= 2^{\log p^{np} + \log(1-p)^{n(1-p)}}$$ \hfill (4.38)
  $$= p^{np} q^{nq} \text{ where } q = 1 - p$$ \hfill (4.39)

  here $H = H(p)$ the binary entropy with probability $p$.

- $np$ is the expected number of 1's.
Ex: Bernoulli trials

- Let $S_n \sim \text{Binomial}(n, p)$ with $S_n = X_1 + X_2 + \cdots + X_n$, $X_i \sim \text{Bernoulli}(p)$.

- Hence, $ES_n = np$ and $\text{var}(S) = npq$, and

$$p(S_n = k) = \binom{n}{k} p^k q^{n-k} \quad (4.36)$$

- Then, the expression $2^{-nH}$ can be viewed in an intuitive way. I.e.,

$$2^{-nH(p)} = 2^{-n(p \log p - (1-p) \log(1-p))} \quad (4.37)$$

$$= 2^{\log p^{np} + \log(1-p)^{n(1-p)}} \quad (4.38)$$

$$= p^{np} q^{nq} \text{ where } q = 1 - p \quad (4.39)$$

here $H = H(p)$ the binary entropy with probability $p$.

- $np$ is the expected number of 1’s.

- $nq$ is the expected number of 0’s.
In other words, all sequences that occur are the ones where the number of 1s and 0s are roughly equal to their expected values.
In other words, all sequences that occur are the ones where the number of 1s and 0s are roughly equal to their expected values.

No other sequences have any appreciable probability!
In other words, all sequences that occur are the ones where the number of 1s and 0s are roughly equal to their expected values. No other sequences have any appreciable probability!

The sequence $X_1, X_2, \ldots, X_n$ was assumed i.i.d., but this can be extended to Markov chains, and to ergodic stationary random processes.
In other words, all sequences that occur are the ones where the number of 1s and 0s are roughly equal to their expected values. No other sequences have any appreciable probability!

The sequence $X_1, X_2, \ldots, X_n$ was assumed i.i.d., but this can be extended to Markov chains, and to ergodic stationary random processes.

But before doing any of that, we need more formalism.
**Asymptotic Equipartition Property (AEP)**

**Theorem 4.5.1 (AEP)**

If $X_1, X_2, \ldots, X_n$ are i.i.d. and $X_i \sim p(x)$ for all $i$, then

$$-rac{1}{n} \log p(X_1, X_2, \ldots, X_n) \xrightarrow{p} H(X)$$

(4.40)
Asymptotic Equipartition Property (AEP)

Theorem 4.5.1 (AEP)

If $X_1, X_2, \ldots, X_n$ are i.i.d. and $X_i \sim p(x)$ for all $i$, then

$$-\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) \xrightarrow{p} H(X)$$

(4.40)

Proof.

$$-\frac{1}{n} \log p(X_1, X_2, \ldots, X_n)$$

(4.43)
**Theorem 4.5.1 (AEP)**

If $X_1, X_2, \ldots, X_n$ are i.i.d. and $X_i \sim p(x)$ for all $i$, then

$$\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) \overset{p}{\to} H(X) \quad (4.40)$$

**Proof.**

$$-\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) = -\frac{1}{n} \log \prod_{i=1}^{n} p(X_i) \quad (4.41)$$

$$\quad (4.43)$$
Asymptotic Equipartition Property (AEP)

**Theorem 4.5.1 (AEP)**

If $X_1, X_2, \ldots, X_n$ are i.i.d. and $X_i \sim p(x)$ for all $i$, then

$$\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) \xrightarrow{p} H(X)$$

(4.40)

**Proof.**

$$-\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) = -\frac{1}{n} \log \prod_{i=1}^{n} p(X_i)$$

(4.41)

$$= -\frac{1}{n} \sum_{i} \log p(X_i)$$

(4.43)
Theorem 4.5.1 (AEP)

If $X_1, X_2, \ldots, X_n$ are i.i.d. and $X_i \sim p(x)$ for all $i$, then

$$-rac{1}{n} \log p(X_1, X_2, \ldots, X_n) \xrightarrow{p} H(X)$$ (4.40)

Proof.

$$-rac{1}{n} \log p(X_1, X_2, \ldots, X_n) = -\frac{1}{n} \log \prod_{i=1}^{n} p(X_i)$$ (4.41)

$$= -\frac{1}{n} \sum_{i} \log p(X_i) \xrightarrow{p} -E \log p(X)$$ (4.42)

(4.43)
Asymptotic Equipartition Property (AEP)

Theorem 4.5.1 (AEP)

If $X_1, X_2, \ldots, X_n$ are i.i.d. and $X_i \sim p(x)$ for all $i$, then

$$-rac{1}{n} \log p(X_1, X_2, \ldots, X_n) \xrightarrow{p} H(X) \quad (4.40)$$

Proof.

$$-rac{1}{n} \log p(X_1, X_2, \ldots, X_n) = - \frac{1}{n} \log \prod_{i=1}^{n} p(X_i)$$

$$= - \frac{1}{n} \sum_{i} \log p(X_i) \xrightarrow{p} - E \log p(X) \quad (4.42)$$

$$= H(X) \quad (4.43)$$
Definition 4.5.2 (Typical Set)

The typical set $A_{\epsilon}^{(n)}$ w.r.t. $p(x)$ is the set of sequences $(x_1, x_2, \ldots, x_n) \in X^n$ with the property that

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)}$$

(4.44)
Typical Set

Definition 4.5.2 (Typical Set)

The typical set $A_{\epsilon}^{(n)}$ w.r.t. $p(x)$ is the set of sequences $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ with the property that

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)} \quad (4.44)$$

Equivalently, we may write $A_{\epsilon}^{(n)}$ as

$$A_{\epsilon}^{(n)} = \left\{ (x_1, x_2, \ldots, x_n) : \left| \frac{1}{n} \log p(x_1, \ldots, x_n) - H \right| < \epsilon \right\} \quad (4.45)$$
Definition 4.5.2 (Typical Set)

The typical set $A^{(n)}_{\epsilon}$ w.r.t. $p(x)$ is the set of sequences $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ with the property that

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)} \quad (4.44)$$

Equivalently, we may write $A^{(n)}_{\epsilon}$ as

$$A^{(n)}_{\epsilon} = \left\{ (x_1, x_2, \ldots, x_n) : \left| -\frac{1}{n} \log p(x_1, \ldots, x_n) - H \right| < \epsilon \right\} \quad (4.45)$$

● Typical set are those sequences with log probability within the range $-nH$.
Typical Set

Definition 4.5.2 (Typical Set)

The typical set $A^{(n)}_\varepsilon$ w.r.t. $p(x)$ is the set of sequences $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ with the property that

$$2^{-n(H(X)+\varepsilon)} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-n(H(X)-\varepsilon)}$$  \hspace{1cm} (4.44)

Equivalently, we may write $A^{(n)}_\varepsilon$ as

$$A^{(n)}_\varepsilon = \left\{ (x_1, x_2, \ldots, x_n) : \left| -\frac{1}{n} \log p(x_1, \ldots, x_n) - H \right| < \varepsilon \right\}$$  \hspace{1cm} (4.45)

- Typical set are those sequences with log probability within the range $-nH$

- $A^{(n)}_\varepsilon$ has a number of interesting properties.
Typical Set

- Size of typical set of source sequences is typically much smaller than the size of the set of all source sequences.
Typical Set

- Size of typical set of source sequences is typically much smaller than the size of the set of all source sequences.

- A typical sequence need not have probability close to that of the most probable sequence.
Typical Set

- Size of typical set of source sequences is typically much smaller than the size of the set of all source sequences.
- A typical sequence need not have probability close to that of the most probable sequence.
- Often, the most probable sequence is not in the typical set.
Typical Set

- Size of typical set of source sequences is typically much smaller than the size of the set of all source sequences.
- A typical sequence need not have probability close to that of the most probable sequence.
- Often, the most probable sequence is not in the typical set.
- Yet, the set of typical sequences has all of the probability.
Typical Set

- Size of typical set of source sequences is typically much smaller than the size of the set of all source sequences.
- A typical sequence need not have probability close to that of the most probable sequence.
- Often, the most probable sequence is not in the typical set.
- Yet, the set of typical sequences has all of the probability.
- While this might sound strange now, it will make sense as we get into the details.
Typical Set $A^{(n)}_{\epsilon}$

Theorem 4.5.3 (Properties of $A^{(n)}_{\epsilon}$)

1. If $(x_1, x_2, \ldots, x_n) \in A^{(n)}_{\epsilon}$, then

$$H(X) - \epsilon \leq \frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon$$ (4.46)
Typical Set $A^{(n)}_\epsilon$

### Theorem 4.5.3 (Properties of $A^{(n)}_\epsilon$)

1. If $(x_1, x_2, \ldots, x_n) \in A^{(n)}_\epsilon$, then

\[
H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon
\]  

(4.46)

2. $p(A^{(n)}_\epsilon) = p \left( \left\{ x : x \in A^{(n)}_\epsilon \right\} \right) > 1 - \epsilon$ for large enough $n$, for all $\epsilon > 0$. 

Prof. Jeff Bilmes  
EE514a/Fall 2019/Info. Theory I – Lecture 4 - Oct 7th, 2019
Typical Set $A^{(n)}_\epsilon$

Theorem 4.5.3 (Properties of $A^{(n)}_\epsilon$)

1. If $(x_1, x_2, \ldots, x_n) \in A^{(n)}_\epsilon$, then

$$H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon$$

(4.46)

2. $p(A^{(n)}_\epsilon) = p \left( \left\{ x : x \in A^{(n)}_\epsilon \right\} \right) > 1 - \epsilon$ for large enough $n$, for all $\epsilon > 0$.

3. Upper bound: $|A^{(n)}_\epsilon| \leq 2^{n(H(X)+\epsilon)}$, where $|A|$ is the number of elements in set $A$. 


Typical Set $A^{(n)}_\epsilon$

**Theorem 4.5.3 (Properties of $A^{(n)}_\epsilon$)**

1. If $(x_1, x_2, \ldots, x_n) \in A^{(n)}_\epsilon$, then

   $$H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon$$ \hspace{1cm} (4.46)

2. $p(A^{(n)}_\epsilon) = p\left(\left\{x : x \in A^{(n)}_\epsilon\right\}\right) > 1 - \epsilon$ for large enough $n$, for all $\epsilon > 0$.

3. **Upper bound:** $|A^{(n)}_\epsilon| \leq 2^{n(H(X)+\epsilon)}$, where $|A|$ is the number of elements in set $A$.

4. **Lower bound:** $|A^{(n)}_\epsilon| \geq (1 - \epsilon)2^{n(H(X)-\epsilon)}$ for large enough $n$.
Typical Set $A^{(n)}_\epsilon$

Theorem 4.5.3 (Properties of $A^{(n)}_\epsilon$)

1. If $(x_1, x_2, \ldots, x_n) \in A^{(n)}_\epsilon$, then
   \[
   H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon
   \]  
   (4.46)

2. $p(A^{(n)}_\epsilon) = p \left( \left\{ x : x \in A^{(n)}_\epsilon \right\} \right) > 1 - \epsilon$ for large enough $n$, for all $\epsilon > 0$.

3. **Upper bound:** $|A^{(n)}_\epsilon| \leq 2^{n(H(X)+\epsilon)}$, where $|A|$ is the number of elements in set $A$.

4. **Lower bound:** $|A^{(n)}_\epsilon| \geq (1 - \epsilon)2^{n(H(X)-\epsilon)}$ for large enough $n$

- The typical set has, essentially, probability 1 (something typical will typically occur).
Typical Set $A^{(n)}_{\epsilon}$

Theorem 4.5.3 (Properties of $A^{(n)}_{\epsilon}$)

1. If $(x_1, x_2, \ldots, x_n) \in A^{(n)}_{\epsilon}$, then

\[
H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon \quad (4.46)
\]

2. $p(A^{(n)}_{\epsilon}) = p \left( \left\{ x : x \in A^{(n)}_{\epsilon} \right\} \right) > 1 - \epsilon$ for large enough $n$, for all $\epsilon > 0$.

3. **Upper bound:** $|A^{(n)}_{\epsilon}| \leq 2^{n(H(X) + \epsilon)}$, where $|A|$ is the number of elements in set $A$.

4. **Lower bound:** $|A^{(n)}_{\epsilon}| \geq (1 - \epsilon)2^{n(H(X) - \epsilon)}$ for large enough $n$.

- The typical set has, essentially, probability 1 (something typical will typically occur).
- All items in that set will have the same probability, $\approx 2^{-nH}$.
Typical Set $A^{(n)}_\epsilon$

**Theorem 4.5.3 (Properties of $A^{(n)}_\epsilon$)**

1. If $(x_1, x_2, \ldots, x_n) \in A^{(n)}_\epsilon$, then
   \[
   H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon \quad (4.46)
   \]

2. $p(A^{(n)}_\epsilon) = p \left( \left\{ x : x \in A^{(n)}_\epsilon \right\} \right) > 1 - \epsilon$ for large enough $n$, for all $\epsilon > 0$.

3. **Upper bound**: $|A^{(n)}_\epsilon| \leq 2^n(H(X) + \epsilon)$, where $|A|$ is the number of elements in set $A$.

4. **Lower bound**: $|A^{(n)}_\epsilon| \geq (1 - \epsilon)2^n(H(X) - \epsilon)$ for large enough $n$.

- The typical set has, essentially, probability 1 (something typical will typically occur).
- All items in that set will have the same probability, $\approx 2^{-nH}$.
- The number of elements in that set is $\approx 2^{nH}$. 
Typical Set, example, $K = |\{0, 1\}|$

- Uniform distribution, $K = 2$, $\mathcal{X} = \{0, 1\}$, Bernoulli trials, $p = 0.5$, entropy $H = 1$, and $|\mathcal{A}_c^{(n)}| = 2^{nH} = 2^n = K^n$, so all sequences will occur with equal probability.
Typical Set, example, $K = |\{0, 1\}|$

- Uniform distribution, $K = 2$, $\mathcal{X} = \{0, 1\}$, Bernoulli trials, $p = 0.5$, entropy $H = 1$, and $|A_c^{(n)}| = 2^{nH} = 2^n = K^n$, so all sequences will occur with equal probability.
- Non-uniform distribution, $p = 0.1$, $1 - p = q = 0.9$, the entropy $H \approx 0.469$. 

Consider $n = 100$, then $K^{100} = 2^{100} \approx 10^{30}$, so the representational capacity of the source strings is $10^{30}$. But $|A_c^{(n)}| = 2^{nH} = 2^n = K^n \approx 10^{14} \ll 10^{30}$. So the number of typical sequences is much smaller than the number of possible sequences.

Q: What is $10^{30} \div 10^{14}$?
A: $10^{30} \div 10^{14} \approx 10^{16}$.

Inefficiency: representational capacity is much larger than the things that occur. This means that things are inefficient, source strings are poor codewords (more bits than necessary for compression).

Thought question (I.e., what you might want to think about): Assume $\varepsilon$ is very small, then where did all the mass of those $\approx 10^{30}$ sequences go?

We will answer this shortly.
Typical Set, example, $K = |\{0, 1\}|$

- Uniform distribution, $K = 2$, $\mathcal{X} = \{0, 1\}$, Bernoulli trials, $p = 0.5$, entropy $H = 1$, and $|A^n| = 2^{nH} = 2^n = K^n$, so all sequences will occur with equal probability.

- Non-uniform distribution, $p = 0.1$, $1 - p = q = 0.9$, the entropy $H \approx 0.469$. Consider $n = 100$, then $K^{100} = 2^{100} \approx 10^{30}$, so representational capacity of the source strings is $10^{30}$. 

Question: What is $10^{30} / 10^{14}$? 

Answer: $10^{30} / 10^{14} \approx 10^{16}$.
Typical Set, example, $K = |\{0, 1\}|$

- **Uniform distribution**, $K = 2$, $\mathcal{X} = \{0, 1\}$, Bernoulli trials, $p = 0.5$, entropy $H = 1$, and $|A_{\epsilon}^{(n)}| = 2^{nH} = 2^n = K^n$, so all sequences will occur with equal probability.

- **Non-uniform distribution**, $p = 0.1$, $1 - p = q = 0.9$, the entropy $H \approx 0.469$. Consider $n = 100$, then $K^{100} = 2^{100} \approx 10^{30}$, so representational capacity of the source strings is $10^{30}$.

- But $|A_{\epsilon}^{(n)}| = 2^{nH} \approx 2^{n0.469} \approx 10^{14} \ll 10^{30} \approx K^{100}$. 

```
Typical Set, example, $K = |\{0, 1\}|$

- Uniform distribution, $K = 2$, $\mathcal{X} = \{0, 1\}$, Bernoulli trials, $p = 0.5$, entropy $H = 1$, and $|A^{(n)}_\varepsilon| = 2^{nH} = 2^n = K^n$, so all sequences will occur with equal probability.

- Non-uniform distribution, $p = 0.1$, $1 - p = q = 0.9$, the entropy $H \approx 0.469$. Consider $n = 100$, then $K^{100} = 2^{100} \approx 10^{30}$, so representational capacity of the source strings is $10^{30}$.

- But $|A^{(n)}_\varepsilon| = 2^{nH} \approx 2^{n0.469} \approx 10^{14} \ll 10^{30} \approx K^{100}$. So the number of typical sequences is much smaller than the number of possible sequences.
Typical Set, example, \( K = |\{0, 1\}| \)

- Uniform distribution, \( K = 2, \mathcal{X} = \{0, 1\} \), Bernoulli trials, \( p = 0.5 \), entropy \( H = 1 \), and \( |A_\varepsilon^{(n)}| = 2^{nH} = 2^n = K^n \), so all sequences will occur with equal probability.

- Non-uniform distribution, \( p = 0.1, 1 - p = q = 0.9 \), the entropy \( H \approx 0.469 \). Consider \( n = 100 \), then \( K^{100} = 2^{100} \approx 10^{30} \), so representational capacity of the source strings is \( 10^{30} \).

- But \( |A_\varepsilon^{(n)}| = 2^{nH} \approx 2^{n0.469} \approx 10^{14} \ll 10^{30} \approx K^{100} \). So the number of typical sequences is much smaller than the number of possible sequences.

- Q: What is \( 10^{30} - 10^{14} \)?
Typical Set, example, $K = |\{0, 1\}|$

- Uniform distribution, $K = 2$, $\mathcal{X} = \{0, 1\}$, Bernoulli trials, $p = 0.5$, entropy $H = 1$, and $|A_\epsilon^{(n)}| = 2^{nH} = 2^n = K^n$, so all sequences will occur with equal probability.

- Non-uniform distribution, $p = 0.1$, $1 - p = q = 0.9$, the entropy $H \approx 0.469$. Consider $n = 100$, then $K^{100} = 2^{100} \approx 10^{30}$, so representational capacity of the source strings is $10^{30}$.

- But $|A_\epsilon^{(n)}| = 2^{nH} \approx 2^{n0.469} \approx 10^{14} \ll 10^{30} \approx K^{100}$. So the number of typical sequences is much smaller than the number of possible sequences.

- Q: What is $10^{30} - 10^{14}$? A: $10^{30} - 10^{14} \approx 10^{30}$. 


Typical Set, example, $K = |\{0, 1\}|$

- Uniform distribution, $K = 2$, $\mathcal{X} = \{0, 1\}$, Bernoulli trials, $p = 0.5$, entropy $H = 1$, and $|A_c^{(n)}| = 2^{nH} = 2^n = K^n$, so all sequences will occur with equal probability.

- Non-uniform distribution, $p = 0.1$, $1 - p = q = 0.9$, the entropy $H \approx 0.469$. Consider $n = 100$, then $K^{100} = 2^{100} \approx 10^{30}$, so representational capacity of the source strings is $10^{30}$.

- But $|A_c^{(n)}| = 2^{nH} \approx 2^{n0.469} \approx 10^{14} \ll 10^{30} \approx K^{100}$. So the number of typical sequences is much smaller than the number of possible sequences.

- Q: What is $10^{30} - 10^{14}$? A: $10^{30} - 10^{14} \approx 10^{30}$.

- Inefficiency: representational capacity is much larger than the things that occur. This means that things are inefficient, source strings are poor codewords (more bits than necessary for compression).

*fun* them selves.
Typical Set, example, $K = |\{0, 1\}|$

- Uniform distribution, $K = 2$, $\mathcal{X} = \{0, 1\}$, Bernoulli trials, $p = 0.5$, entropy $H = 1$, and $|A^{(n)}_\epsilon| = 2^{nH} = 2^n = K^n$, so all sequences will occur with equal probability.

- Non-uniform distribution, $p = 0.1$, $1 - p = q = 0.9$, the entropy $H \approx 0.469$. Consider $n = 100$, then $K^{100} = 2^{100} \approx 10^{30}$, so representational capacity of the source strings is $10^{30}$.

- But $|A^{(n)}_\epsilon| = 2^{nH} \approx 2^{n0.469} \approx 10^{14} \ll 10^{30} \approx K^{100}$. So the number of typical sequences is much smaller than the number of possible sequences.

- Q: What is $10^{30} - 10^{14}$? A: $10^{30} - 10^{14} \approx 10^{30}$.

- Inefficiency: representational capacity is much larger than the things that occur. This means that things are inefficient, source strings are poor codewords (more bits than necessary for compression).

- Thought question (i.e., what you might want to think about): Assume $\epsilon$ is very small, then where did all the mass of those $\approx 10^{30} - 10^{14}$ sequences go?
Typical Set, example, $K = |\{0, 1\}|$

- Uniform distribution, $K = 2$, $\mathcal{X} = \{0, 1\}$, Bernoulli trials, $p = 0.5$, entropy $H = 1$, and $|A_\epsilon^{(n)}| = 2^{nH} = 2^n = K^n$, so all sequences will occur with equal probability.

- Non-uniform distribution, $p = 0.1$, $1 - p = q = 0.9$, the entropy $H \approx 0.469$. Consider $n = 100$, then $K^{100} = 2^{100} \approx 10^{30}$, so representational capacity of the source strings is $10^{30}$.

- But $|A_\epsilon^{(n)}| = 2^{nH} \approx 2^{n0.469} \approx 10^{14} \ll 10^{30} \approx K^{100}$. So the number of typical sequences is much smaller than the number of possible sequences.

Q: What is $10^{30} - 10^{14}$? A: $10^{30} - 10^{14} \approx 10^{30}$.

Inefficiency: representational capacity is much larger than the things that occur. This means that things are inefficient, source strings are poor codewords (more bits than necessary for compression).

Thought question (i.e., what you might want to think about): Assume $\epsilon$ is very small, then where did all the mass of those $\approx 10^{30} - 10^{14}$ sequences go? We will answer this shortly.
Typical Sets are Typical

- Curiously, as $n$ gets big,

\[ p(A^{(n)}_\epsilon) > 1 - \epsilon \text{ for any } \epsilon > 0 \]  

(4.47)
Typical Sets are Typical

Curiously, as $n$ gets big,

$$p(A^{(n)}_\epsilon) > 1 - \epsilon \text{ for any } \epsilon > 0 \quad (4.47)$$

So, $A^{(n)}_\epsilon$ has pretty much all of the probability, and each element in $A^{(n)}_\epsilon$ has the same probability, so

$$p(x) \approx 2^{-nH} \quad \forall x \in A^{(n)}_\epsilon \quad (4.48)$$
Typical Sets are Typical

- Curiously, as $n$ gets big,
  
  $$p(A^{(n)}_{\epsilon}) > 1 - \epsilon \text{ for any } \epsilon > 0 \quad (4.47)$$

- So, $A^{(n)}_{\epsilon}$ has pretty much all of the probability, and each element in $A^{(n)}_{\epsilon}$ has the same probability, so
  
  $$p(x) \approx 2^{-nH} \quad \forall x \in A^{(n)}_{\epsilon} \quad (4.48)$$

- Ex: Bernoulli trials: $X_i \sim \text{Bernoulli}(p)$, with
  $$p(X_i = 1) = p = 1 - p(X_i = 0), \text{ and } p > 0.5.$$
Typical Sets are Typical

Curiously, as $n$ gets big,

$$p(A_{\epsilon}^{(n)}) > 1 - \epsilon \text{ for any } \epsilon > 0$$ (4.47)

So, $A_{\epsilon}^{(n)}$ has pretty much all of the probability, and each element in $A_{\epsilon}^{(n)}$ has the same probability, so

$$p(x) \approx 2^{-nH} \forall x \in A_{\epsilon}^{(n)}$$ (4.48)

Ex: Bernoulli trials: $X_i \sim \text{Bernoulli}(p)$, with

$$p(X_i = 1) = p = 1 - p(X_i = 0), \text{ and } p > 0.5.$$  

Probability of $n$ successive 1s is $p^n$ and is the most likely sequence.
Typical Sets are Typical

- Curiously, as \( n \) gets big,
  \[
p(A^{(n)}_\epsilon) > 1 - \epsilon \text{ for any } \epsilon > 0
  \]
  \[(4.47)\]

- So, \( A^{(n)}_\epsilon \) has pretty much all of the probability, and each element in \( A^{(n)}_\epsilon \) has the same probability, so
  \[
p(x) \approx 2^{-nH} \quad \forall x \in A^{(n)}_\epsilon
  \]
  \[(4.48)\]

- Ex: Bernoulli trials: \( X_i \sim \text{Bernoulli}(p) \), with
  \[
p(X_i = 1) = p = 1 - p(X_i = 0), \text{ and } p > 0.5.
  \]
- Probability of \( n \) successive 1s is \( p^n \) and is the most likely sequence.
- Probability of each typical sequence is \( 2^{-nH} \).
Typical Sets are Typical

- Curiously, as \( n \) gets big,
  \[
p(A^{(n)}_\epsilon) > 1 - \epsilon \text{ for any } \epsilon > 0
  \] (4.47)

- So, \( A^{(n)}_\epsilon \) has pretty much all of the probability, and each element in \( A^{(n)}_\epsilon \) has the same probability, so
  \[
p(x) \approx 2^{-nH} \quad \forall x \in A^{(n)}_\epsilon
  \] (4.48)

- Ex: Bernoulli trials: \( X_i \sim \text{Bernoulli}(p) \), with
  \[
p(X_i = 1) = p = 1 - p(X_i = 0), \text{ and } p > 0.5.
  \]
  - Probability of \( n \) successive 1s is \( p^n \) and is the most likely sequence.
  - Probability of each typical sequence is \( 2^{-nH} \).
  - For \( n = 100, p = 0.9 = 1 - q \), most likely sequence has probability
    \[
p^n \approx 2.66 \times 10^{-5},
    \]
Typical Sets are Typical

- Curiously, as $n$ gets big,
  \[ p(A^{(n)}_\epsilon) > 1 - \epsilon \text{ for any } \epsilon > 0 \]  
  (4.47)

- So, $A^{(n)}_\epsilon$ has pretty much all of the probability, and each element in $A^{(n)}_\epsilon$ has the same probability, so
  \[ p(x) \approx 2^{-nH} \quad \forall x \in A^{(n)}_\epsilon \]  
  (4.48)

- Ex: Bernoulli trials: $X_i \sim \text{Bernoulli}(p)$, with $p(X_i = 1) = p = 1 - p(X_i = 0)$, and $p > 0.5$.

- Probability of $n$ successive 1s is $p^n$ and is the most likely sequence.

- Probability of each typical sequence is $2^{-nH}$.

- For $n = 100$, $p = 0.9 = 1 - q$, most likely sequence has probability $p^n \approx 2.66 \times 10^{-5}$, but a typical sequence has probability $2^{-nH} \approx 7.62 \times 10^{-15}$. 
Non-typical sequences are not typical

- Thus, \( p^n \gg 2^{-nH} \) and the most likely sequence is much more probable than a typical one.
Non-typical sequences are not typical

- Thus, $p^n \gg 2^{-nH}$ and the most likely sequence is much more probable than a typical one.
- What happens to the probability of the most probable sequence, on average (per symbol), as $n$ gets big

$$\frac{1}{n} \log p^n = - \log p = \log \frac{1}{p} \xrightarrow{n \to \infty} H$$

(4.49)
Non-typical sequences are not typical

- Thus, $p^n \gg 2^{-nH}$ and the most likely sequence is much more probable than a typical one.
- What happens to the probability of the most probable sequence, on average (per symbol), as $n$ gets big

$$\frac{1}{n} \log p^n = -\log p = \log 1/p \xrightarrow{n \to \infty} H$$  \hspace{1cm} (4.49)

- Typical set, essentially, has all the probability $A^{(n)}_\epsilon > 1 - \epsilon$
Non-typical sequences are not typical

- Thus, $p^n \gg 2^{-nH}$ and the most likely sequence is much more probable than a typical one.
- What happens to the probability of the most probable sequence, on average (per symbol), as $n$ gets big
  \[
  -\frac{1}{n} \log p^n = -\log p = \log \frac{1}{p} \xrightarrow[n \to \infty]{\text{yes or no?}} H
  \]  
  (4.49)
- Typical set, essentially, has all the probability $A_{(n)}^{(\epsilon)} > 1 - \epsilon$
- But is the most likely sequence in the typical set?
Non-typical sequences are not typical

- Thus, $p^n \gg 2^{-nH}$ and the most likely sequence is much more probable than a typical one.
- What happens to the probability of the most probable sequence, on average (per symbol), as $n$ gets big

$$-\frac{1}{n} \log p^n = - \log p = \log \frac{1}{p} \xrightarrow{n \to \infty} H$$

(4.49)

- Typical set, essentially, has all the probability $A_{\epsilon}^{(n)} > 1 - \epsilon$
- But is the most likely sequence in the typical set? No, since the probability of the most likely sequence $p^n$ is not close to $2^{-nH}$
Non-typical sequences are not typical

- Thus, $p^n \gg 2^{-nH}$ and the most likely sequence is much more probable than a typical one.
- What happens to the probability of the most probable sequence, on average (per symbol), as $n$ gets big

\[ -\frac{1}{n} \log p^n = -\log p = \log \frac{1}{p} \xrightarrow{n \to \infty} H \]  

(4.49)

- Typical set, essentially, has all the probability $A^{(n)}_\epsilon > 1 - \epsilon$
- But is the most likely sequence in the typical set? No, since the probability of the most likely sequence $p^n$ is not close to $2^{-nH}$
- Again, for $n = 100, p = 0.9 = 1 - q$, consider a sequence with ninety 1s and ten 0s, probability $p^{90}(1 - p)^{10} \approx 7.62 \times 10^{-15} \approx 2^{-nH}$. 
Non-typical sequences are not typical

- Thus, \( p^n \gg 2^{-nH} \) and the most likely sequence is much more probable than a typical one.
- What happens to the probability of the most probable sequence, on average (per symbol), as \( n \) gets big

\[
-\frac{1}{n} \log p^n = -\log p = \log 1/p = \log 1 \frac{\text{yes or no?}}{n \to \infty} \to H \quad (4.49)
\]
- Typical set, essentially, has all the probability \( A_{\epsilon}^{(n)} > 1 - \epsilon \)
- But is the most likely sequence in the typical set? No, since the probability of the most likely sequence \( p^n \) is not close to \( 2^{-nH} \)
- Again, for \( n = 100 \), \( p = 0.9 = 1 - q \), consider a sequence with ninety 1s and ten 0s, probability \( p^{90}(1 - p)^{10} \approx 7.62 \times 10^{-15} \approx 2^{-nH} \).
- So, for \( n = 100 \), and \( p = 0.9 \),
Non-typical sequences are not typical

- Thus, $p^n \gg 2^{-nH}$ and the most likely sequence is much more probable than a typical one.
- What happens to the probability of the most probable sequence, on average (per symbol), as $n$ gets big

$$-\frac{1}{n} \log p^n = - \log p = \log \frac{1}{p} \xrightarrow[n \to \infty]{} H$$ (4.49)

- Typical set, essentially, has all the probability $A_{\epsilon}^{(n)} > 1 - \epsilon$
- But is the most likely sequence in the typical set? No, since the probability of the most likely sequence $p^n$ is not close to $2^{-nH}$
- Again, for $n = 100$, $p = 0.9 = 1 - q$, consider a sequence with ninety 1s and ten 0s, probability $p^{90}(1 - p)^{10} \approx 7.62 \times 10^{-15} \approx 2^{-nH}$.
- So, for $n = 100$, and $p = 0.9$,
  - most probable sequence has probability $2.66 \times 10^{-5}$
Non-typical sequences are not typical

- Thus, $p^n \gg 2^{-nH}$ and the most likely sequence is much more probable than a typical one.
- What happens to the probability of the most probable sequence, on average (per symbol), as $n$ gets big

$$\frac{1}{n} \log p^n = -\log p = \log \frac{1}{p} \xrightarrow{n \to \infty} H$$ (4.49)

- Typical set, essentially, has all the probability $A_{\epsilon}(n) > 1 - \epsilon$
- But is the most likely sequence in the typical set? No, since the probability of the most likely sequence $p^n$ is not close to $2^{-nH}$
- Again, for $n = 100$, $p = 0.9 = 1 - q$, consider a sequence with ninety 1s and ten 0s, probability $p^{90}(1-p)^{10} \approx 7.62 \times 10^{-15} \approx 2^{-nH}$.
- So, for $n = 100$, and $p = 0.9$,
  - most probable sequence has probability $2.66 \times 10^{-5}$
  - but a typical sequence (such as above) has probability $7.62 \times 10^{-15} \ll 2.66 \times 10^{-5}$. 
Non-typical sequences are not typical

- Thus, $p^n \gg 2^{-nH}$ and the most likely sequence is much more probable than a typical one.
- What happens to the probability of the most probable sequence, on average (per symbol), as $n$ gets big
  $$\frac{1}{n} \log p^n = - \log p = \log 1/p \xrightarrow[n \to \infty]{} H$$

- Typical set, essentially, has all the probability $A_{\epsilon}^{(n)} > 1 - \epsilon$
- But is the most likely sequence in the typical set? No, since the probability of the most likely sequence $p^n$ is not close to $2^{-nH}$
- Again, for $n = 100$, $p = 0.9 = 1 - q$, consider a sequence with ninety 1s and ten 0s, probability $p^{90}(1 - p)^{10} \approx 7.62 \times 10^{-15} \approx 2^{-nH}$.
- So, for $n = 100$, and $p = 0.9$,
  - most probable sequence has probability $2.66 \times 10^{-5}$
  - but a typical sequence (such as above) has probability $7.62 \times 10^{-15} \ll 2.66 \times 10^{-5}$.
- Thus, this very improbable sequence is typical!
Average probability of sequences

- What happens to the probability of the most probable sequence, on average (per symbol), as $n$ gets big?

By the AEP, we have that if $(x_1, \ldots, x_n)$ is typical (i.e., $(x_1, \ldots, x_n) \in A(n)$), then

$$\frac{1}{n} \log p(x_1, \ldots, x_n) \approx H(x).$$

What happens to the probability of the most probable sequence, on average (per symbol), as $n$ gets big, for $p > 0.5$:

$$\frac{1}{n} \log p \approx \log \frac{1}{p}.$$
What happens to the probability of the most probable sequence, on average (per symbol), as $n$ gets big?

By the AEP, we have that if $(x_1, \ldots, x_n)$ is typical (i.e., $(x_1, \ldots, x_n) \in A^{(n)}$, for any $\epsilon > 0$), then

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

(4.50)
Average probability of sequences

- What happens to the probability of the most probable sequence, on average (per symbol), as \( n \) gets big?

- By the AEP, we have that if \((x_1, \ldots, x_n)\) is typical (i.e., \((x_1, \ldots, x_n) \in A^{(n)}_\epsilon\), for any \( \epsilon > 0 \)), then

\[
\frac{1}{n} \log p(x_1, \ldots, x_n) \xrightarrow{n \to \infty} H \tag{4.50}
\]

- What happens to the probability of the most probable sequence, on average (per symbol), as \( n \) gets big, for \( p > 0.5 \),

\[
\frac{1}{n} \log p^n = - \log p = \log \frac{1}{p} \xrightarrow{n \to \infty} - \log p \tag{4.51}
\]
Average probability of sequences

- What happens to the probability of the most probable sequence, on average (per symbol), as $n$ gets big?

- By the AEP, we have that if $(x_1, \ldots, x_n)$ is typical (i.e., $(x_1, \ldots, x_n) \in A^{(n)}_{\epsilon}$, for any $\epsilon > 0$), then

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

(4.50)

- What happens to the probability of the most probable sequence, on average (per symbol), as $n$ gets big, for $p > 0.5$,

$$-\frac{1}{n} \log p^n = -\log p = \log \frac{1}{p} \xrightarrow{n \to \infty} -\log p$$

(4.51)

- So average probability of the most probable sequence is quite different than the typical sequences.
Are typical sequences most probable?

- Again, typical set has, essentially, all the probability \( p(A_{\epsilon}^{(n)}) > 1 - \epsilon \).
Are typical sequences most probable?

- Again, typical set has, essentially, all the probability $p(A_{\epsilon}^{(n)}) > 1 - \epsilon$.
- How can a sequence having the most probability not be typical, but a sequence with much lower probability be typical?
Are typical sequences most probable?

- Again, typical set has, essentially, all the probability $p(A^{(n)}_\epsilon) > 1 - \epsilon$.
- How can a sequence having the most probability not be typical, but a sequence with much lower probability be typical?
- There are exponentially many sequences that are typical, each with less probability than the high probability sequences.
Are typical sequences most probable?

- Again, typical set has, essentially, all the probability $p(A_{\epsilon}^{(n)}) > 1 - \epsilon$.
- How can a sequence having the most probability not be typical, but a sequence with much lower probability be typical?
- There are exponentially many sequences that are typical, each with less probability than the high probability sequences.
- There are exponentially many more typical sequences than there are “high” probability sequences.
Are typical sequences most probable?

- Again, typical set has, essentially, all the probability \( p(A^{(n)}_{\epsilon}) > 1 - \epsilon \).
- How can a sequence having the most probability not be typical, but a sequence with much lower probability be typical?
- There are exponentially many sequences that are typical, each with less probability than the high probability sequences.
- There are exponentially many more typical sequences than there are “high” probability sequences.
- The probability of each individual sequence goes to zero as \( n \to \infty \).
Are typical sequences most probable?

- Again, typical set has, essentially, all the probability $p(A^{(n)}_{\epsilon}) > 1 - \epsilon$.
- How can a sequence having the most probability not be typical, but a sequence with much lower probability be typical?
- There are exponentially many sequences that are typical, each with less probability than the high probability sequences.
- There are exponentially many more typical sequences than there are “high” probability sequences.
- The probability of each individual sequence goes to zero as $n \to \infty$.
- The size of the set of typical sequences grows fast enough, as $n \to \infty$ such that the probability of $A^{(n)}_{\epsilon}$ goes to 1.
Are typical sequences most probable?

- Again, typical set has, essentially, all the probability \( p(A_{\epsilon}^{(n)}) > 1 - \epsilon \).
- How can a sequence having the most probability not be typical, but a sequence with much lower probability be typical?
- There are exponentially many sequences that are typical, each with less probability than the high probability sequences.
- There are exponentially many more typical sequences than there are “high” probability sequences.
- The probability of each individual sequence goes to zero as \( n \to \infty \).
- The size of the set of typical sequences grows fast enough, as \( n \to \infty \) such that the probability of \( A_{\epsilon}^{(n)} \) goes to 1.
- The size of the set of highly probable sequences grows slow enough so that the probability of that set goes to zero, as \( n \to \infty \).
Binomial Distribution

In a Binomial distribution $B(n, p)$, we are interested in the probability of $S_n$ being a particular value.

$$\Pr(S_n = k) = \binom{n}{k} p^k q^{n-k}$$  \hspace{1cm} (4.52)

where $X_1, X_2, \ldots, X_n$ are i.i.d. Bernoulli, $q = 1 - p$, and

$$S_n = X_1 + X_2 + \cdots + X_n, \quad X_i \sim \text{Bernoulli}(p)$$  \hspace{1cm} (4.53)
Binomial Distribution

- In a Binomial distribution $B(n, p)$, we are interested in the probability of $S_n$ being a particular value.

$$\Pr(S_n = k) = \binom{n}{k} p^k q^{n-k}$$  \hfill (4.52)

where $X_1, X_2, \ldots, X_n$ are i.i.d. Bernoulli, $q = 1 - p$, and

$$S_n = X_1 + X_2 + \cdots + X_n, \quad X_i \sim \text{Bernoulli}(p)$$  \hfill (4.53)

- The range of $k$ is $k \in \{0, 1, \ldots, n\}$. 

How to compare two Binomial distributions $B(n_1, p)$ and $B(n_2, p)$ for different values of $n_1 \neq n_2$?

Normalize: let $S_{0n} = S_n / n$

$$\Pr\left(\frac{S_{0n}}{n} = k/n\right) = \Pr\left(S_n = k\right) = \binom{n}{k} p^k q^{n-k}$$  \hfill (4.54)

So $S_{0n} \in [0, 1]$, can plot $S_{0n_1}$ and $S_{0n_2}$ for $n_1 \neq n_2$ and some value $p$.

In particular, what happens to $p\left(\frac{S_{0n}}{n} = \ell\right)$ as $n$ gets large, for various values of $\ell \in [0, 1]$?
Binomial Distribution

- In a Binomial distribution $B(n, p)$, we are interested in the probability of $S_n$ being a particular value.

$$\Pr(S_n = k) = \binom{n}{k} p^k q^{n-k} \quad (4.52)$$

where $X_1, X_2, \ldots, X_n$ are i.i.d. Bernoulli, $q = 1 - p$, and

$$S_n = X_1 + X_2 + \cdots + X_n, \quad X_i \sim \text{Bernoulli}(p) \quad (4.53)$$

- The range of $k$ is $k \in \{0, 1, \ldots, n\}$.

- How to compare two Binomial distributions $B(n_1, p)$ and $B(n_2, p)$ for different values of $n$ ($n_1 \neq n_2$)?
Binomial Distribution

- In a Binomial distribution $B(n, p)$, we are interested in the probability of $S_n$ being a particular value.

\[
\Pr(S_n = k) = \binom{n}{k} p^k q^{n-k}
\]  
(4.52)

where $X_1, X_2, \ldots, X_n$ are i.i.d. Bernoulli, $q = 1 - p$, and

\[
S_n = X_1 + X_2 + \cdots + X_n, \; X_i \sim \text{Bernoulli}(p)
\]  
(4.53)

- The range of $k$ is $k \in \{0, 1, \ldots, n\}$.

- How to compare two Binomial distributions $B(n_1, p)$ and $B(n_2, p)$ for different values of $n$ ($n_1 \neq n_2$)? Normalize: let $S'_n = S_n/n$ in:

\[
\Pr(S'_n = k/n) = \Pr(S_n = k) = \binom{n}{k} p^k q^{n-k}
\]  
(4.54)
Binomial Distribution

- In a Binomial distribution $B(n, p)$, we are interested in the probability of $S_n$ being a particular value.

$$\Pr(S_n = k) = \binom{n}{k} p^k q^{n-k} \quad (4.52)$$

where $X_1, X_2, \ldots, X_n$ are i.i.d. Bernoulli, $q = 1 - p$, and

$$S_n = X_1 + X_2 + \cdots + X_n, \quad X_i \sim \text{Bernoulli}(p) \quad (4.53)$$

- The range of $k$ is $k \in \{0, 1, \ldots, n\}$.

- How to compare two Binomial distributions $B(n_1, p)$ and $B(n_2, p)$ for different values of $n$ ($n_1 \neq n_2$)? Normalize: let $S'_n = S_n / n$ in:

$$\Pr(S'_n = k/n) = \Pr(S_n = k) = \binom{n}{k} p^k q^{n-k} \quad (4.54)$$

- So $S'_n \in [0, 1]$, can plot $S'_{n_1}$ and $S'_{n_2}$ for $n_1 \neq n_2$ and some value $p$. 
Binomial Distribution

- In a Binomial distribution $B(n, p)$, we are interested in the probability of $S_n$ being a particular value.

\[
\Pr(S_n = k) = \binom{n}{k} p^k q^{n-k} \tag{4.52}
\]

where $X_1, X_2, \ldots, X_n$ are i.i.d. Bernoulli, $q = 1 - p$, and

\[
S_n = X_1 + X_2 + \cdots + X_n, \ X_i \sim \text{Bernoulli}(p) \tag{4.53}
\]

- The range of $k$ is $k \in \{0, 1, \ldots, n\}$.

- How to compare two Binomial distributions $B(n_1, p)$ and $B(n_2, p)$ for different values of $n$ ($n_1 \neq n_2$)? Normalize: let $S'_n = S_n/n$ in:

\[
\Pr(S'_n = k/n) = \Pr(S_n = k) = \binom{n}{k} p^k q^{n-k} \tag{4.54}
\]

- So $S'_n \in [0, 1]$, can plot $S'_{n_1}$ and $S'_{n_2}$ for $n_1 \neq n_2$ and some value $p$.

- In particular, what happens to $\Pr(S'_n = \alpha)$ as $n$ gets large, for various values of $\alpha \in [0, 1]$?
Binomial Distribution, when $n$ gets big, $p = 0.5$

$$\Pr(S_n = k) = \binom{n}{k} p^k q^{n-k}, \quad S_n = X_1 + X_2 + \cdots + X_n, \quad X_i \sim \text{Bernoulli}(p)$$

- What happens when $n$ gets big?
Binomial Distribution, when $n$ gets big, $p = 0.5$

$$\Pr(S_n = k) = \binom{n}{k} p^k q^{n-k}, \quad S_n = X_1 + X_2 + \cdots + X_n, \quad X_i \sim \text{Bernoulli}(p)$$

- What happens when $n$ gets big?
- Plot the probability of the normalized values, $S_n/n = k/n$, and see how the distribution changes when $n$ gets large.
Binomial Distribution, when $n$ gets big, $p = 0.5$

$$\Pr(S_n = k) = \binom{n}{k} p^{k} q^{n-k}, \quad S_n = X_1 + X_2 + \cdots + X_n, \quad X_i \sim \text{Bernoulli}(p)$$

- What happens when $n$ gets big?
- Plot the probability of the normalized values, $S_n/n = k/n$, and see how the distribution changes when $n$ gets large.
In previous slide, with $p = 0.5$, all sequences are typical. Why?
Binomial Distribution, $p = 0.5$

- In previous slide, with $p = 0.5$, all sequences are typical. Why?
- ... because all sequences have the same probability, and that is:

$$\Pr(x_1, x_2, \ldots, x_n) = 0.5^n = 2^{-nH} = 2^{-n}$$ (4.55)

regardless of the composition or count of 1’s and 0’s in the binary string $x_1, x_2, \ldots, x_n$ (i.e., $H = 1$ in this case).
Binomial Distribution, \( p = 0.5 \)

- In previous slide, with \( p = 0.5 \), all sequences are typical. Why?
- \( \cdots \) because all sequences have the same probability, and that is:

\[
\Pr(x_1, x_2, \ldots, x_n) = 0.5^n = 2^{-nH} = 2^{-n}
\]  (4.55)

regardless of the composition or count of 1’s and 0’s in the binary string \( x_1, x_2, \ldots, x_n \) (i.e., \( H = 1 \) in this case).

- In previous plot, \( x \)-axis really shows a “type” of sequence corresponding to relative fraction of 1s vs. 0s in the string.
Binomial Distribution, $p = 0.5$

- In previous slide, with $p = 0.5$, all sequences are typical. Why?
- ... because all sequences have the same probability, and that is:

$$
\Pr(x_1, x_2, \ldots, x_n) = 0.5^n = 2^{-nH} = 2^{-n}
$$

(4.55)

regardless of the composition or count of 1’s and 0’s in the binary string $x_1, x_2, \ldots, x_n$ (i.e., $H = 1$ in this case).

- In previous plot, $x$-axis really shows a “type” of sequence corresponding to relative fraction of 1s vs. 0s in the string.
- Each sequence within each “type” has same probability.
Binomial Distribution, $p = 0.5$

- In previous slide, with $p = 0.5$, all sequences are typical. Why?
- ... because all sequences have the same probability, and that is:

$$\Pr(x_1, x_2, \ldots, x_n) = 0.5^n = 2^{-nH} = 2^{-n}$$  \hspace{1cm} (4.55)

regardless of the composition or count of 1’s and 0’s in the binary string $x_1, x_2, \ldots, x_n$ (i.e., $H = 1$ in this case).

- In previous plot, $x$-axis really shows a “type” of sequence corresponding to relative fraction of 1s vs. 0s in the string.
- Each sequence within each “type” has same probability.
- Can partition the strings into $n$ “types”, based on the count of number of 1s in the string, eventually (as $n$ gets big), any type other than the one with strings having $k = n/2$ 1s will have negligible probability.
Binomial Distribution, when $n$ gets big, $p = 0.9$

$$\Pr(S_n = k) = \binom{n}{k} p^k q^{n-k}, \quad S_n = X_1 + X_2 + \cdots + X_n, \; X_i \sim \text{Bernoulli}(p)$$

- Plot the probability of the normalized values, $S_n/n = k/n$, and see how the distribution changes when $n$ gets large.
Binomial Distribution, when $n$ gets big, $p = 0.9$

$$\Pr(S_n = k) = \binom{n}{k} p^k q^{n-k}, \quad S_n = X_1 + X_2 + \cdots + X_n, \quad X_i \sim \text{Bernoulli}(p)$$

- Plot the probability of the normalized values, $S_n/n = k/n$, and see how the distribution changes when $n$ gets large.
- By comparison, number of atoms in observable universe $\approx e^{187} \approx 10^{81}$. 
Typical Set $A_{\epsilon}^{(n)}$

Theorem 4.6.3 (Properties of $A_{\epsilon}^{(n)}$)

1. If $(x_1, x_2, \ldots, x_n) \in A_{\epsilon}^{(n)}$, then

$$H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon \quad (4.46)$$

2. $p(A_{\epsilon}^{(n)}) = p\left(\left\{x : x \in A_{\epsilon}^{(n)}\right\}\right) > 1 - \epsilon$ for large enough $n$, for all $\epsilon > 0$. 
Typical Set $A_{\epsilon}^{(n)}$

Theorem 4.6.3 (Properties of $A_{\epsilon}^{(n)}$)

1. If $(x_1, x_2, \ldots, x_n) \in A_{\epsilon}^{(n)}$, then
   \[
   H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon \tag{4.46}
   \]

2. $p(A_{\epsilon}^{(n)}) = p \left( \{ x : x \in A_{\epsilon}^{(n)} \} \right) > 1 - \epsilon$ for large enough $n$, for all $\epsilon > 0$.

3. Upper bound: $|A_{\epsilon}^{(n)}| \leq 2^{n(H(X) + \epsilon)}$, where $|A|$ is the number of elements in set $A$.

4. Lower bound: $|A_{\epsilon}^{(n)}| \geq (1-\epsilon)2^{n(H(X) - \epsilon)}$ for large enough $n$

- The typical set has, essentially, probability 1 (something typical will typically occur).
- All items in that set will have the same probability, $\approx 2^{-nH}$.
- The number of elements in that set is $\approx 2^{nH}$. 
Proof of Theorem 4.5.3.

1. This is a restatement of the AEP definition.
Proof of Theorem 4.5.3.

1. This is a restatement of the AEP definition.

2. Use the expanded definition of convergence in probability we saw earlier in class

\[
p(A^{(n)}_\epsilon) = p \left( \left| -\frac{1}{n} \sum_i \log p(x_i) - H \right| < \epsilon \right) > 1 - \delta \text{ for } n \text{ big enough}
\]

(4.56)

and we can chose any \( \delta \) we wish, so choose \( \delta = \epsilon \), giving

\[
p(A^{(n)}_\epsilon) > 1 - \epsilon \text{ for } n \text{ big enough } \forall \epsilon
\]

(4.57)
Proof of Theorem 4.5.3.

3. Upper bound size of $A^{(n)}_\epsilon$

\[ |A^{(n)}_\epsilon| \leq 2^n (H(X) + \epsilon) \quad (4.59) \]

...
Proof of Theorem 4.5.3.

3 Upper bound size of $A_{\epsilon}^{(n)}$

1

(4.59)

\[ \sum_{x} p(x) \leq \sum_{x \in A_{\epsilon}^{(n)}} p(x) \leq 2^n \left( H(X) + \varepsilon \right) \]
Proof of Theorem 4.5.3.

3. Upper bound size of $A_\epsilon^{(n)}$

$$1 = \sum_x p(x)$$

(4.59)
Proof of Theorem 4.5.3.

3 Upper bound size of $A^{(n)}_\epsilon$

$$1 = \sum_{x} p(x) \geq \sum_{x \in A^{(n)}_\epsilon} p(x)$$

(4.59)
Proof of Theorem 4.5.3.

3 Upper bound size of $A^{(n)}_\varepsilon$

\[
1 \geq \sum_{x} p(x) \geq \sum_{x \in A^{(n)}_\varepsilon} p(x) \geq \sum_{x \in A^{(n)}_\varepsilon} 2^{-n(H(X)+\varepsilon)}
\] (4.58)

(4.59)

...
Proof of Theorem 4.5.3.

Upper bound size of $A_\epsilon^{(n)}$

\[
1 = \sum_{x} p(x) \geq \sum_{x \in A_\epsilon^{(n)}} p(x) \geq \sum_{x \in A_\epsilon^{(n)}} 2^{-n(H(X) + \epsilon)}
\] (4.58)

\[
= |A_\epsilon^{(n)}| 2^{-n(H(X) + \epsilon)}
\] (4.59)
Proof of Theorem 4.5.3.

3 Upper bound size of $A_{\epsilon}^{(n)}$

\[
1 = \sum_{x} p(x) \geq \sum_{x \in A_{\epsilon}^{(n)}} p(x) \geq \sum_{x \in A_{\epsilon}^{(n)}} 2^{-n(H(X)+\epsilon)}
\]

\[\geq |A_{\epsilon}^{(n)}| 2^{-n(H(X)+\epsilon)}\]  \hspace{1cm} (4.59)

\[
\text{giving } |A_{\epsilon}^{(n)}| \leq 2^{n(H+\epsilon)}.
\]

...
Proof of Theorem 4.5.3.

4. Lower bound size of $A_\epsilon^{(n)}$. For large enough $n$

\begin{align}
\text{(4.60)} \\
= 2^n (H(X) - \epsilon) |A_\epsilon^{(n)}| \\
\end{align}

...
Proof of Theorem 4.5.3.

4 Lower bound size of $A_{\epsilon}^{(n)}$. For large enough $n$

$$1 - \epsilon$$

(4.61)
Proof of Theorem 4.5.3.

Lower bound size of $A^{(n)}_\epsilon$. For large enough $n$

$$1 - \epsilon < p(A^{(n)}_\epsilon)$$

(4.61)
Theorem 4.5.3 Proofs

Proof of Theorem 4.5.3.

4 Lower bound size of $A_\epsilon^{(n)}$. For large enough $n$

$$1 - \epsilon < p(A_\epsilon^{(n)}) \leq \sum_{x \in A_\epsilon^{(n)}} 2^{-n(H(X) - \epsilon)} \quad (4.60)$$

$$= 2^{-n(H(X) - \epsilon)} |A_\epsilon^{(n)}| \quad (4.61)$$

...
Proof of Theorem 4.5.3.

Lower bound size of $A^{(n)}_\epsilon$. For large enough $n$

\[
1 - \epsilon < p(A^{(n)}_\epsilon) \leq \sum_{x \in A^{(n)}_\epsilon} 2^{-n(H(X) - \epsilon)}
\] (4.60)

\[
= 2^{-n(H(X) - \epsilon)} |A^{(n)}_\epsilon|
\] (4.61)

...
Proof of Theorem 4.5.3.

4 Lower bound size of $A_\epsilon^{(n)}$. For large enough $n$

$$1 - \epsilon < p(A_\epsilon^{(n)}) \leq \sum_{x \in A_\epsilon^{(n)}} 2^{-n(H(X) - \epsilon)} \quad (4.60)$$

$$= 2^{-n(H(X) - \epsilon)} |A_\epsilon^{(n)}| \quad (4.61)$$

giving $|A_\epsilon^{(n)}| \geq (1 - \epsilon)2^{n(H(X) - \epsilon)}$

...
Data Compression to the entropy of the source

- An important consequence of this is that we can compress data down to the entropy of the source.
Data Compression to the entropy of the source

- An important consequence of this is that we can compress data down to the entropy of the source.
- **Idea**: Consider $X_1, X_2, \ldots, X_n$ i.i.d. and $\sim p(x)$. 
Data Compression to the entropy of the source

- An important consequence of this is that we can compress data down to the entropy of the source.
- Idea: Consider $X_1, X_2, \ldots, X_n$ i.i.d. and $\sim p(x)$.
- Partition the set of sequences into two blocks:
Data Compression to the entropy of the source

- An important consequence of this is that we can compress data down to the entropy of the source.
- Idea: Consider $X_1, X_2, \ldots, X_n$ i.i.d. and $\sim p(x)$.
- Partition the set of sequences into two blocks:
  - The typical sets $A^{(n)}_\epsilon$, ...
An important consequence of this is that we can compress data down to the entropy of the source.

Idea: Consider $X_1, X_2, \ldots, X_n$ i.i.d. and $\sim p(x)$.

Partition the set of sequences into two blocks:

- The typical sets $A_\epsilon^{(n)}$, ...
- and its complement, the non-typical sets $\mathcal{X}^n \setminus A_\epsilon^{(n)} \triangleq A_\epsilon^{(n)c}$
Data Compression to the entropy of the source

- An important consequence of this is that we can compress data down to the entropy of the source.
- Idea: Consider $X_1, X_2, \ldots, X_n$ i.i.d. and $\sim p(x)$.
- Partition the set of sequences into two blocks:
  - The typical sets $A^{(n)}_\varepsilon$, …
  - and its complement, the non-typical sets $\mathcal{X}^n \setminus A^{(n)}_\varepsilon \triangleq A^{(n)}_{\varepsilon c}$
- A partition, i.e., $A^{(n)}_\varepsilon \cap A^{(n)}_{\varepsilon c} = \emptyset$ and $A^{(n)}_\varepsilon \cup A^{(n)}_{\varepsilon c} = \mathcal{X}^n$:

\[ \mathcal{X}^n \text{ having } |\mathcal{X}^n| = K^n \text{ elements} \]
Typical Set Compression

- We index the elements in each of the sets, the typical set and the non-typical set, separately.
Typical Set Compression

- We index the elements in each of the sets, the typical set and the non-typical set, separately.

- In the typical set, \( \exists |A_{\epsilon}^{(n)}| \leq 2^n(H+\epsilon) \) elements, requiring

\[
\left\lfloor n(H + \epsilon) \right\rfloor \leq n(H + \epsilon) + 1 \text{ bits.} \tag{4.62}
\]
Typical Set Compression

- We index the elements in each of the sets, the typical set and the non-typical set, separately.

- In the typical set, \( |A_\epsilon^{(n)}| \leq 2^{n(H+\epsilon)} \) elements, requiring

\[
\left\lceil n(H + \epsilon) \right\rceil \leq n(H + \epsilon) + 1 \text{ bits.} \tag{4.62}
\]

- We use an extra bit at the beginning to indicate if it is typical or not, i.e., we use

\[
(b_0, b_1, b_2, \ldots, b_{\lceil n(H+\epsilon) \rceil}) \tag{4.63}
\]

which indexes which of the typical set elements it is. \( b_0 = 0 \) indicating that the set is typical.
Typical Set Compression

- We index the elements in each of the sets, the typical set and the non-typical set, separately.
- In the typical set, \( \exists |A_\epsilon^{(n)}| \leq 2^n(H+\epsilon) \) elements, requiring
  \[
  \lceil n(H+\epsilon) \rceil \leq n(H+\epsilon) + 1 \text{ bits.}
  \] (4.62)
- We use an extra bit at the beginning to indicate if it is typical or not, i.e., we use
  \[
  (b_0, b_1, b_2, \ldots, b_{\lceil n(H+\epsilon) \rceil})
  \] (4.63)
  which indexes which of the typical set elements it is. \( b_0 = 0 \) indicating that the set is typical.
- Total number of bits required is \( n(H+\epsilon) + 2 \) for a typical sequence.
Typical Set Compression

- We index the elements in each of the sets, the typical set and the non-typical set, separately.

\[ \text{(4.64)} \]
Typical Set Compression

- We index the elements in each of the sets, the typical set and the non-typical set, separately.
- In the non-typical set, we index everything. I.e., we use $\lfloor \log |\mathcal{X}|^n \rfloor \leq n \log K + 1$ bits.
Typical Set Compression

- We index the elements in each of the sets, the typical set and the non-typical set, separately.
- In the non-typical set, we index everything. I.e., we use \( \lceil \log |X|^n \rceil \leq n \log K + 1 \) bits.
- I.e., we index everything with a bit vector of the form 
  \[
  (b_0, b_1, b_2, \ldots, b_{\lceil \log |X|^n \rceil})
  \]
  where here \( b_0 = 1 \) indicating atypicality.
Typical Set Compression

- We index the elements in each of the sets, the typical set and the non-typical set, separately.
- In the non-typical set, we index everything. I.e., we use \[ \lceil \log |\mathcal{X}|^n \rceil \leq n \log K + 1 \text{ bits.} \]
- I.e., we index everything with a bit vector of the form
  \[(b_0, b_1, b_2, \ldots, b_{\lceil \log |\mathcal{X}|^n \rceil}) \quad (4.64)\]
  where here \(b_0 = 1\) indicating atypicality.
- Total number of bits for an atypical sequence is \(n \log K + 2\).
Typical Set Compression

- We index the elements in each of the sets, the typical set and the non-typical set, separately.
- In the non-typical set, we index everything. I.e., we use $\lceil \log |X|^n \rceil \leq n \log K + 1$ bits.
- I.e., we index everything with a bit vector of the form

$$(b_0, b_1, b_2, \ldots, b_{\lceil \log |X|^n \rceil})$$

where here $b_0 = 1$ indicating atypicality.
- Total number of bits for an atypical sequence is $n \log K + 2$.
- Note, this is our first code for the class! This is called source coding or compression, and entails finding a sequence of bits for each source string so that average length is as short as possible.
Typical Set Compression: Features of our code, and setup

- Code is 1-to-1, so easy to decode and encode, given code book (mapping).
Typical Set Compression: Features of our code, and setup

- Code is 1-to-1, so easy to decode and encode, given code book (mapping).
- Simple but brute force enumeration of atypical set $A^{(n)c}_\epsilon$ means more bits than necessary, since

$$|A^{(n)c}_\epsilon| = |\mathcal{X}^n| - |A^{(n)}_\epsilon| = K^n - |A^{(n)}_\epsilon| < K^n,$$
Typical Set Compression: Features of our code, and setup

- Code is 1-to-1, so easy to decode and encode, given code book (mapping).
- Simple but brute force enumeration of atypical set $A^{(n)c}_\epsilon$ means more bits than necessary, since
  \[
  |A^{(n)c}_\epsilon| = |X^n| - |A^{(n)}_\epsilon| = K^n - |A^{(n)}_\epsilon| < K^n, \text{ but this is not going to matter, as we will see.}
  \]
Typical Set Compression: Features of our code, and setup

- Code is 1-to-1, so easy to decode and encode, given code book (mapping).
- Simple but brute force enumeration of atypical set $A^{(n)c}_\epsilon$ means more bits than necessary, since
  \[ |A^{(n)c}_\epsilon| = |X^n| - |A^{(n)}_\epsilon| = K^n - |A^{(n)}_\epsilon| < K^n, \]
  but this is not going to matter, as we will see.
- typical sequences have a “short” description length, \( \approx nH \).
Typical Set Compression: Features of our code, and setup

- Code is 1-to-1, so easy to decode and encode, given code book (mapping).
- Simple but brute force enumeration of atypical set $A_{\epsilon}^{(n)c}$ means more bits than necessary, since
  \[ |A_{\epsilon}^{(n)c}| = |X^n| - |A_{\epsilon}^{(n)}| = K^n - |A_{\epsilon}^{(n)}| < K^n, \]
  but this is not going to matter, as we will see.
- typical sequences have a “short” description length, $\approx nH$.
- Let $\ell(x_1:n)$ be length of the codeword assigned to sequence $x_1:n$. 
Typical Set Compression: Features of our code, and setup

- Code is 1-to-1, so easy to decode and encode, given code book (mapping).
- Simple but brute force enumeration of atypical set $A_{\epsilon}^{(n)c}$ means more bits than necessary, since
  \[|A_{\epsilon}^{(n)c}| = |\mathcal{X}^n| - |A_{\epsilon}^{(n)}| = K^n - |A_{\epsilon}^{(n)}| < K^n,\]
  but this is not going to matter, as we will see.
- Typical sequences have a “short” description length, $\approx nH$.
- Let $\ell(x_{1:n})$ be length of the codeword assigned to sequence $x_{1:n}$.
- $\ell(X_{1:n})$ is a random variable since $X_{1:n}$ is a random variable.
Typical Set Compression: Features of our code, and setup

- Code is 1-to-1, so easy to decode and encode, given code book (mapping).

- Simple but brute force enumeration of atypical set $A^{(n)c}_\epsilon$ means more bits than necessary, since
  \[ |A^{(n)c}_\epsilon| = |\mathcal{X}^n| - |A^{(n)}_\epsilon| = K^n - |A^{(n)}_\epsilon| < K^n, \]
  but this is not going to matter, as we will see.

- Typical sequences have a “short” description length, $\approx nH$.

- Let $\ell(x_{1:n})$ be length of the codeword assigned to sequence $x_{1:n}$.

- $\ell(X_{1:n})$ is a random variable since $X_{1:n}$ is a random variable.

- Thus, $E\ell(X_{1:n}) = \sum_{x_{1:n}} p(x_{1:n})\ell(x_{1:n})$ is the average, or expected, length of our code. We want this to be as short as possible.
Expected Length

- Suppose that $n$ is large enough so that $p(A_{\epsilon}^{(n)}) > 1 - \epsilon$, 

\[ E[X_{1:n}] = \sum_{x_{1:n}} p(x_{1:n}) \cdot X_{x_{1:n}} \cdot (\eta + 2) \]

\[ E[X_{1:n}] = \sum_{x_{1:n}} p(x_{1:n}) \cdot X_{x_{1:n}} \cdot (\eta + 2) \]

\[ \leq n \cdot (H + \eta) + 2 \]

\[ n \cdot (H + \eta) + 2 \leq n \cdot (H(0) + \eta) + 2 \]

\[ n \cdot (H + \eta) + 2 \leq n \cdot (H + \eta) + 2 \]

\[ n \cdot (H + \eta) + 2 \leq n \cdot (H + \eta) + 2 \]
Expected Length

- Suppose that $n$ is large enough so that $p(A^{(n)}_{\epsilon}) > 1 - \epsilon$, 

\[
\begin{align*}
\mathbb{E} x^{(1:n)} &= x^{(1:n)} p(x^{(1:n)}) 
\end{align*}
\] 

\[
\begin{align*}
x^{(1:n)} &\in A^{(n)}_{\epsilon} \quad \Rightarrow \quad p(x^{(1:n)}) > 1 - \epsilon
\end{align*}
\] 

\[
\begin{align*}
\mathbb{E} x^{(1:n)} &\leq p(A^{(n)}_{\epsilon}) \left[ n \left( H + \varepsilon \right) + 2 \right] + p(A^{(n)}_{\epsilon}) \left[ n \log K + 2 \right] 
\end{align*}
\] 

\[
\begin{align*}
\mathbb{E} x^{(1:n)} &\leq p(A^{(n)}_{\epsilon}) \left[ n \left( H + \varepsilon \right) + 2 \right] + p(A^{(n)}_{\epsilon}) \left[ n \log K + 2 \right] 
\end{align*}
\] 

\[
\begin{align*}
\mathbb{E} x^{(1:n)} &\leq p(A^{(n)}_{\epsilon}) \left[ n \left( H + \varepsilon \right) + 2 \right] + p(A^{(n)}_{\epsilon}) \left[ n \log K + 2 \right] 
\end{align*}
\] 

\[
\begin{align*}
\mathbb{E} x^{(1:n)} &\leq p(A^{(n)}_{\epsilon}) \left[ n \left( H + \varepsilon \right) + 2 \right] + p(A^{(n)}_{\epsilon}) \left[ n \log K + 2 \right] 
\end{align*}
\]
Expected Length

Suppose that \( n \) is large enough so that \( p(A_{\epsilon}^{(n)}) > 1 - \epsilon \), then

\[
E\ell(X_{1:n}) = \sum_{x_{1:n}} p(x_{1:n})\ell(x_{1:n})
\]  

(4.65)

(4.66)

(4.67)

(4.68)

(4.70)
Expected Length

Suppose that \( n \) is large enough so that \( p(A^n) > 1 - \epsilon \), then

\[
E\ell(X_1:n) = \sum_{x_1:n} p(x_1:n)\ell(x_1:n)
\]

\[
= \sum_{x_1:n \in A^n} p(x_1:n)\ell(x_1:n) + \sum_{x_1:n \in A^n_c} p(x_1:n)\ell(x_1:n)
\]

(4.65)

(4.66)

(4.67)

(4.68)

(4.70)
Expected Length

Suppose that $n$ is large enough so that $p(A^{(n)}_\epsilon) > 1 - \epsilon$, then

\[
E\ell(X_{1:n}) = \sum_{x_{1:n}} p(x_{1:n})\ell(x_{1:n})
\]

(4.65)

\[
\begin{align*}
= & \sum_{x_{1:n} \in A^{(n)}_\epsilon} p(x_{1:n})\ell(x_{1:n}) + \sum_{x_{1:n} \in A^{(n)}_\epsilon^c} p(x_{1:n})\ell(x_{1:n}) \\
\leq & \sum_{x_{1:n} \in A^{(n)}_\epsilon} p(x_{1:n})[n(H + \epsilon) + 2] + \sum_{x_{1:n} \in A^{(n)}_\epsilon^c} p(x_{1:n})[n \log K + 2]
\end{align*}
\]

(4.66)

(4.67)

(4.68)

(4.70)
Expected Length

Suppose that $n$ is large enough so that $p(A_{\epsilon}^{(n)}) > 1 - \epsilon$, then

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

(4.65)

$$= \sum_{x_{1:n} \in A_{\epsilon}^{(n)}} p(x_{1:n})\ell(x_{1:n}) + \sum_{x_{1:n} \in A_{\epsilon}^{(n)c}} p(x_{1:n})\ell(x_{1:n})$$

(4.66)

$$\leq \sum_{x_{1:n} \in A_{\epsilon}^{(n)}} p(x_{1:n})[n(H + \epsilon) + 2] + \sum_{x_{1:n} \in A_{\epsilon}^{(n)c}} p(x_{1:n})[n \log K + 2]$$

(4.67)

$$= p(A_{\epsilon}^{(n)})[n(H + \epsilon) + 2] + p(A_{\epsilon}^{(n)c})[n \log K + 2]$$

(4.68)

(4.70)
Expected Length

Suppose that \( n \) is large enough so that \( p(A^{(n)}_\epsilon) > 1 - \epsilon \), then

\[
E\ell(X_{1:n}) = \sum_{x_{1:n}} p(x_{1:n})\ell(x_{1:n})
\]

\[
= \sum_{x_{1:n} \in A^{(n)}_\epsilon} p(x_{1:n})\ell(x_{1:n}) + \sum_{x_{1:n} \in A^{(n)}_\epsilon^c} p(x_{1:n})\ell(x_{1:n})
\]

\[
\leq \sum_{x_{1:n} \in A^{(n)}_\epsilon} p(x_{1:n})[n(H + \epsilon) + 2] + \sum_{x_{1:n} \in A^{(n)}_\epsilon^c} p(x_{1:n})[n \log K + 2]
\]

\[
= p(A^{(n)}_\epsilon)[n(H + \epsilon) + 2] + p(A^{(n)}_\epsilon^c)[n \log K + 2]
\]

\[
\leq 1
\]

(4.65)

(4.66)

(4.67)

(4.68)

(4.69)

(4.70)
Expected Length

Suppose that \( n \) is large enough so that \( p(A^{(n)}_\epsilon) > 1 - \epsilon \), then

\[
E\ell(X_{1:n}) = \sum_{x_{1:n}} p(x_{1:n})\ell(x_{1:n})
\]

(4.65)

\[
= \sum_{x_{1:n} \in A^{(n)}_\epsilon} p(x_{1:n})\ell(x_{1:n}) + \sum_{x_{1:n} \in A^{(n)}_\epsilon^c} p(x_{1:n})\ell(x_{1:n})
\]

(4.66)

\[
\leq \sum_{x_{1:n} \in A^{(n)}_\epsilon} p(x_{1:n})[n(H + \epsilon) + 2] + \sum_{x_{1:n} \in A^{(n)}_\epsilon^c} p(x_{1:n})[n \log K + 2]
\]

(4.67)

\[
= p(A^{(n)}_\epsilon)[n(H + \epsilon) + 2] + p(A^{(n)}_\epsilon^c)[n \log K + 2]
\]

(4.68)

\[
\leq 1 < \epsilon
\]

(4.70)
Expected Length

Suppose that $n$ is large enough so that $p(A^{(n)}_\epsilon) > 1 - \epsilon$, then

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

(4.65)

$$= \sum_{x_{1:n} \in A^{(n)}_\epsilon} p(x_{1:n})\ell(x_{1:n}) + \sum_{x_{1:n} \in A^{(n)}_\epsilon^c} p(x_{1:n})\ell(x_{1:n})$$

(4.66)

$$\leq \sum_{x_{1:n} \in A^{(n)}_\epsilon} p(x_{1:n})[n(H + \epsilon) + 2] + \sum_{x_{1:n} \in A^{(n)}_\epsilon^c} p(x_{1:n})[n \log K + 2]$$

(4.67)

$$= p(A^{(n)}_\epsilon) [n(H + \epsilon) + 2] + p(A^{(n)}_\epsilon^c) [n \log K + 2]$$

(4.68)

$$\leq 1 \cdot [n(H + \epsilon) + 2] + \epsilon \cdot n \log K + \epsilon 2$$

(4.69)

$$\leq n(H + \epsilon) + 2 + \epsilon n \log K + \epsilon 2$$

(4.70)
Expected Length

Suppose that \( n \) is large enough so that \( p(A(n)) > 1 - \epsilon \), then

\[
E\ell(X_{1:n}) = \sum_{x_{1:n}} p(x_{1:n}) \ell(x_{1:n})
\]

(4.65)

\[
= \sum_{x_{1:n} \in A(n)} p(x_{1:n}) \ell(x_{1:n}) + \sum_{x_{1:n} \in A(n)^c} p(x_{1:n}) \ell(x_{1:n})
\]

(4.66)

\[
\leq \sum_{x_{1:n} \in A(n)} p(x_{1:n}) [n(H + \epsilon) + 2] + \sum_{x_{1:n} \in A(n)^c} p(x_{1:n}) [n \log K + 2]
\]

(4.67)

\[
= p(A(n)) [n(H + \epsilon) + 2] + p(A(n)^c) [n \log K + 2]
\]

(4.68)

\[
\leq 1 \cdot [n(H + \epsilon) + 2] + \epsilon \cdot [n \log K + 2]
\]

(4.69)

\[
\leq n(H + \epsilon) + 2 + \epsilon n \log K + \epsilon 2
\]

(4.70)
Suppose that $n$ is large enough so that $p(A_{\epsilon}^{(n)}) > 1 - \epsilon$, then

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

$$= \sum_{x_{1:n} \in A_{\epsilon}^{(n)}} p(x_{1:n})\ell(x_{1:n}) + \sum_{x_{1:n} \in A_{\epsilon}^{(n) c}} p(x_{1:n})\ell(x_{1:n})$$

$$\leq \sum_{x_{1:n} \in A_{\epsilon}^{(n)}} p(x_{1:n})[n(H + \epsilon) + 2] + \sum_{x_{1:n} \in A_{\epsilon}^{(n) c}} p(x_{1:n})[n \log K + 2]$$

$$= p(A_{\epsilon}^{(n)}) [n(H + \epsilon) + 2] + p(A_{\epsilon}^{(n) c}) [n \log K + 2]$$

$$\leq 1$$

$$n(H + \epsilon) + 2 + \epsilon n \log K + \epsilon 2$$

$$= n\left[H + \epsilon + \epsilon \log K + \frac{2}{n} + \frac{2\epsilon}{n}\right] = n(H + \epsilon')$$
Expected Length

- But $\epsilon' = \epsilon + \epsilon \log K + \frac{2}{n} + \frac{2\epsilon}{n}$ can be made as small as we wish by making $\epsilon$ small and $n$ large.
Expected Length

- But $\epsilon' = \epsilon + \epsilon \log K + \frac{2}{n} + \frac{2\epsilon}{n}$ can be made as small as we wish by making $\epsilon$ small and $n$ large.
- Thus, $n(H + \epsilon')$ can be made as close as we want to $nH$ by making $\epsilon$ small and $n$ large.
Expected Length

- But $\epsilon' = \epsilon + \epsilon \log K + \frac{2}{n} + \frac{2\epsilon}{n}$ can be made as small as we wish by making $\epsilon$ small and $n$ large.
- Thus, $n(H + \epsilon')$ can be made as close as we want to $nH$ by making $\epsilon$ small and $n$ large.
- We have just proven the following theorem:
Expected Length

- But \( \epsilon' = \epsilon + \epsilon \log K + \frac{2}{n} + \frac{2\epsilon}{n} \) can be made as small as we wish by making \( \epsilon \) small and \( n \) large.
- Thus, \( n(H + \epsilon') \) can be made as close as we want to \( nH \) by making \( \epsilon \) small and \( n \) large.
- We have just proven the following theorem:

**Theorem 4.7.1**

Let \( X_{1:n} \) be i.i.d. \( \sim p(x) \), \( \epsilon > 0 \), then \( \exists \) a code \( f_n : \mathcal{X}^n \rightarrow \) binary strings and integer \( n_\epsilon \), such that the mapping is one-to-one (so invertible w/o error), and

\[
E\left[\frac{1}{n} \ell(X_{1:n})\right] \leq H(X) + \epsilon
\]

(4.71)

for all \( \epsilon > 0 \) and for all \( n \geq n_\epsilon \).
Expected Length

- But $\epsilon' = \epsilon + \epsilon \log K + \frac{2}{n} + \frac{2\epsilon}{n}$ can be made as small as we wish by making $\epsilon$ small and $n$ large.
- Thus, $n(H + \epsilon')$ can be made as close as we want to $nH$ by making $\epsilon$ small and $n$ large.
- We have just proven the following theorem:

**Theorem 4.7.1**

Let $X_{1:n}$ be i.i.d. $\sim p(x)$, $\epsilon > 0$, then $\exists$ a code $f_n : X^n \rightarrow$ binary strings and integer $n_\epsilon$, such that the mapping is one-to-one (so invertible w/o error), and

$$E\left[\frac{1}{n} \ell(X_{1:n})\right] \leq H(X) + \epsilon \quad (4.71)$$

for all $\epsilon > 0$ and for all $n \geq n_\epsilon$.

- Thus, it takes at most $nH(X)$ bits to represent $X_{1:n}$ on average, or $H(X)$ bits per source alphabet symbol.
Shannon’s source coding theorem

- The previous theorem is Shannon’s first theorem, stating that it is possible (using long block lengths) to compress down arbitrarily close to the entropy limit.
Shannon’s source coding theorem

- The previous theorem is Shannon’s first theorem, stating that it is possible (using long block lengths) to compress down arbitrarily close to the entropy limit.

- An instance of universal source coding, coding without explicitly using the distribution, since whatever happens, once $n$ gets large, is all that will happen.
Shannon’s source coding theorem

- The previous theorem is Shannon’s first theorem, stating that it is possible (using long block lengths) to compress down arbitrarily close to the entropy limit.

- An instance of universal source coding, coding without explicitly using the distribution, since whatever happens, once $n$ gets large, is all that will happen.

- Ex: online coding, code only those things that you encounter knowing that it must be typical if you encounter it, if $n$ is large enough. In such case, you don’t need $p(x)$, only $H(p)$. 

Shannon’s source coding theorem

- The previous theorem is Shannon’s first theorem, stating that it is possible (using long block lengths) to compress down arbitrarily close to the entropy limit.

- An instance of universal source coding, coding without explicitly using the distribution, since whatever happens, once $n$ gets large, is all that will happen.

- Ex: online coding, code only those things that you encounter knowing that it must be typical if you encounter it, if $n$ is large enough. In such case, you don’t need $p(x)$, only $H(p)$.

- Ultimately, we need to prove that we can’t compress to lower than the entropy limit without incurring error, this is the converse of the theorem that we will prove soon.