Fundamentals of Information Science

Shandong University
2025 Spring

Definition of Entropy

Information of a random variable X?

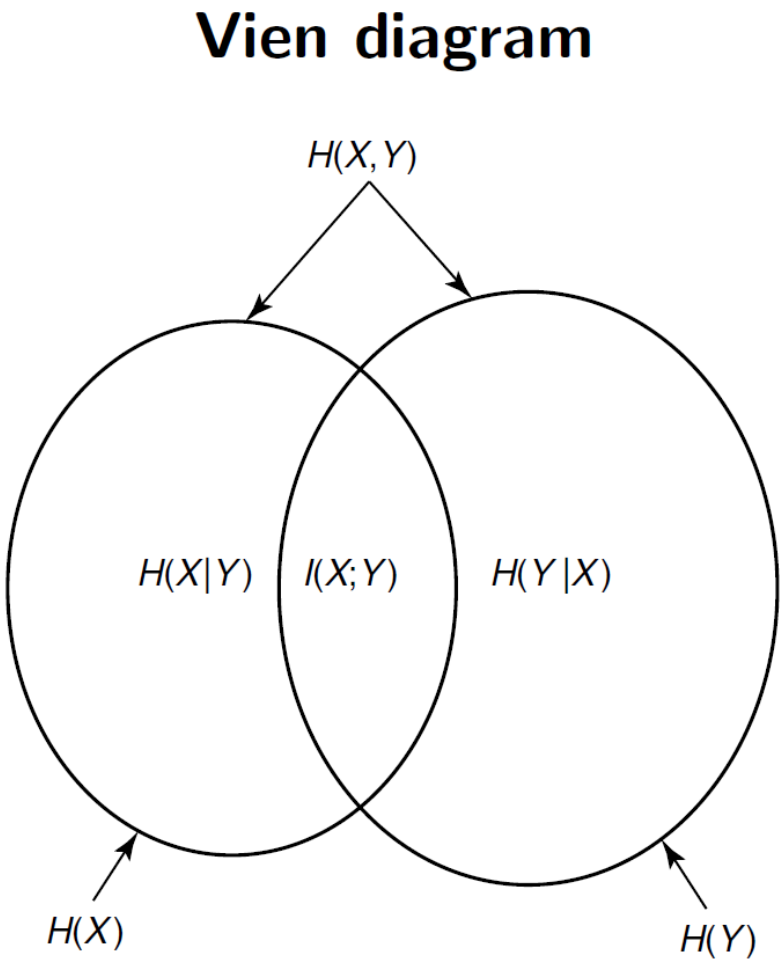
Let X be a random variable taking on a finite number M of different values $x_1, \cdots, x_M$

With probability $p_1, \cdots, p_M$, $p_i > 0$, $\sum_{i=1}^{M} p_i = 1$

Information of X = Expected Information of All Outcomes (Entropy)

$$H(p_1, \cdots, p_M) = -\sum_{i=1}^{M} p_i \log_2 p_i$$

Vien Diagram



$I(X;Y)$ is the intersection of information in X with information in Y

Lecture 2.2: Properties of Entropy

Chain rule for entropy

- Last time, simple chain rule $H(X,Y) = H(X) + H(Y|X)$
- No matter how we play with chain rule, we get the same answer

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

Chain rule for entropy

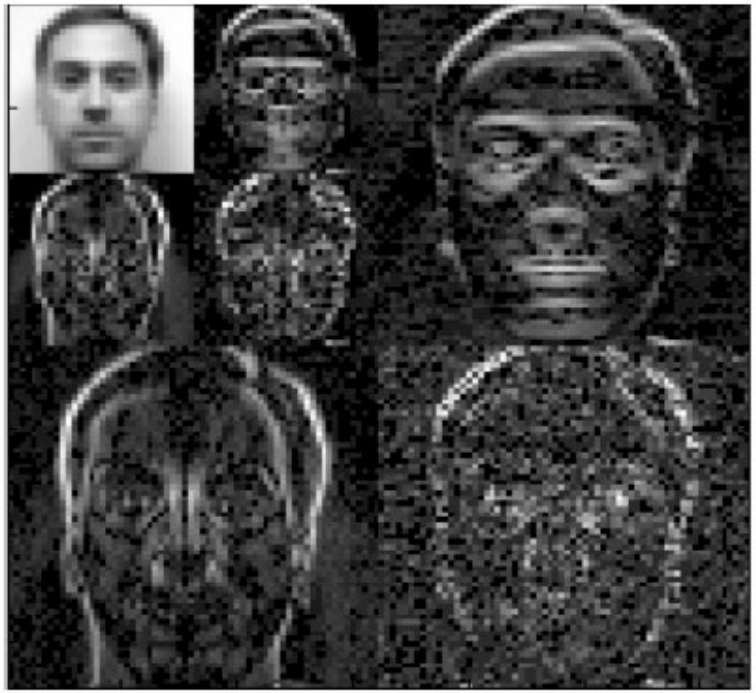
- Entropy for a collection of RV's is the sum of the conditional entropies
- More generally: $H(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \cdots, X_1)$

  Proof:

$$H(X_1, X_2) = H(X_1) + H(X_2|X_1)$$
$$H(X_1, X_2, X_3) = H(X_3, X_2|X_1) + H(X_1)$$
$$= H(X_3|X_2, X_1) + H(X_2|X_1) + H(X_1)$$
$$\vdots$$

Chain rule for entropy

$$H(X^n) = \sum_{i=1}^{n} H(X_i | \underbrace{X_{-i}}_{\text{everything seen before}})$$



Convexity

- A function $f(x)$ is convex over an interval (a, b) if for every $x, y \in (a, b)$ and $0 \leq \lambda \leq 1$,

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y).$$

  Strictly convex if equality holds only if $\lambda = 0$.

Convexity

- If a function f has second order derivative $\geq 0 (> 0)$, the function is convex (strictly convex).
- Examples: x^2, e^x, $|x|$, $x \log x (x \geq 0)$, $\|\boldsymbol{x}\|^2$.
- A function f is concave if $-f$ is convex.
- Linear function $ax + b$ is both convex and concave.

Convexity of Entropy

Let $f(x) = -x \log(x)$ then

$$\begin{aligned} f'(x) &= -x \log(e)\frac{1}{x} - \log(x) \\ &= -\log(x) - \log(e) \end{aligned}$$

and

$$f''(x) = -\log(e)\frac{1}{x} < 0$$

for $x > 0$.

$H(X) = \sum_{x \in |\mathcal{X}|} f(P_X(x))$

thus the entropy of X is concave in the value of $P_X(x)$ for every x.

Jensen's Inequality

- Due to Danish mathematician Johan Jensen, 1906
- Widely used in mathematics and information theory
- Convex transformation of a mean
  $\leq$ mean after convex transformation

Jensen's Inequality

Theorem. *(Jensen's inequality) If f is a convex function,*

$$Ef(X) \geq f(EX).$$

If f strictly convex, equality holds when

$$X = constant.$$

Proof: Let $x^\star = EX$. Expand $f(x)$ by Taylor's Theorem at $x^\star$:

$$f(x) = f(x^\star) + f'(x^\star)(x - x^\star) + \frac{f''(z)}{2}(x - x^\star)^2, \quad z \in (x, x^\star)$$

f convex: $f''(z) \geq 0$. So $f(x) \geq f(x^\star) + f'(x^\star)(x - x^\star)$. Take expectation on both size: $Ef(X) \geq f(x^\star)$.

Jensen's Inequality

Consequences

- $f(x) = x^2$, $EX^2 \geq [EX]^2$: variance is nonnegative
- $f(x) = e^x$, $Ee^x \geq e^{E(x)}$

Information Inequality

$$D(p||q) \geq 0,$$

equality iff $p(x) = q(x)$ for all x.

Proof:

$$\begin{aligned} D(p||q) &= \sum_x p(x) \log \frac{p(x)}{q(x)} \\ &= -\sum_x p(x) \log \frac{q(x)}{p(x)} \\ &\geq -\log \sum_x p(x) \frac{q(x)}{p(x)} \\ &= -\log \sum_x q(x) = 0. \end{aligned}$$

Information Inequality

- $I(X;Y) \geq 0$, equality iff X and Y are independent.
  Since $I(X;Y) = D(p(x,y)||p(x)p(y))$.

- Conditional relative entropy and mutual information are also nonnegative

Information Inequality

Conditioning reduces entropy

Information cannot hurt:

$$H(X|Y) \leq H(X)$$

- Since $I(X;Y) = H(X) - H(X|Y) \geq 0$
- Knowing another RV Y only reduces uncertainty in X on average
- $H(X|Y = y)$ may be larger than $H(X)$: in court, knowing a new evidence sometimes can increase uncertainty

Example 2.6.1 Let (X, Y) have the following joint distribution:

Y \ X	1	2
1	0	$\frac{3}{4}$
2	$\frac{1}{8}$	$\frac{1}{8}$

Then $H(X) = H(\frac{1}{8}, \frac{7}{8}) = 0.544$ bit, $H(X|Y = 1) = 0$ bits, and $H(X|Y = 2) = 1$ bit. We calculate $H(X|Y) = \frac{3}{4}H(X|Y = 1) + \frac{1}{4}H(X|Y = 2) = 0.25$ bit. Thus, the uncertainty in X is increased if $Y = 2$ is observed and decreased if $Y = 1$ is observed, but uncertainty decreases on the average.

Information Inequality

Independence bound on entropy

$$H(X_1, \cdots, X_n) \leq \sum_{n=1}^{n} H(X_i).$$

equality iff X_i independent.

- From chain rule:

$$H(X_1, \cdots, X_n) = \sum_{i=1}^{n} H(X_i | X_{i-1}, \cdots, X_1) \leq \sum_{i=1}^{n} H(X_i).$$

Maximum Entropy

Maximum entropy

Uniform distribution has maximum entropy among all distributions with finite discrete support.

Theorem. $H(X) \leq \log |\mathcal{X}|$, *where* $\mathcal{X}$ *is the number of elements in the set. Equality iff* X *has uniform distribution.*

Proof: Let U be a uniform distributed RV, $u(x) = 1/|\mathcal{X}|$

$$0 \leq D(p||u) = \sum p(x) \log \frac{p(x)}{u(x)} \tag{1}$$

$$= \sum p(x) \log |\mathcal{X}| - (-\sum p(x) \log p(x)) = \log |\mathcal{X}| - H(X) \tag{2}$$

Maximum Entropy

Maximum entropy

Uniform distribution has maximum entropy among all distributions with finite discrete support.

Theorem. $H(X) \leq \log |\mathcal{X}|$, *where* $\mathcal{X}$ *is the number of elements in the set. Equality iff* X *has uniform distribution.*

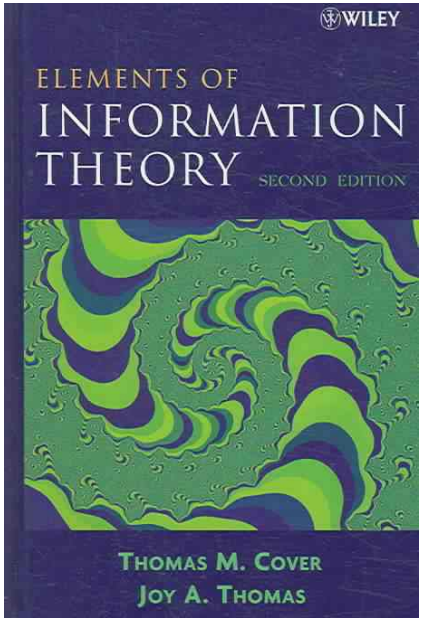
Proof: Let U be a uniform distributed RV, $u(x) = 1/|\mathcal{X}|$

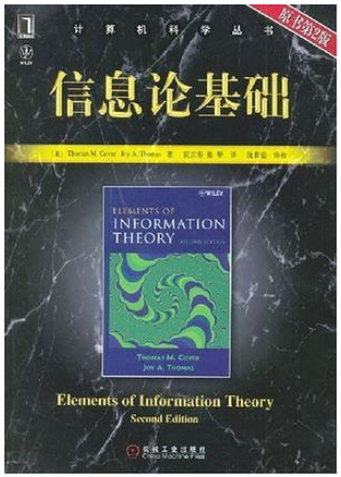
$$0 \leq D(p||u) = \sum p(x) \log \frac{p(x)}{u(x)} \tag{1}$$

$$= \sum p(x) \log |\mathcal{X}| - (-\sum p(x) \log p(x)) = \log |\mathcal{X}| - H(X) \tag{2}$$

Convexity

- $D(p||q)$ convex in (p, q)
- Entropy $H(p)$ concave in p
- Mutual information $I(X;Y)$ concave in $p(x)$ (fixing $p(y|x)$), and convex in $p(y|x)$ (fixing $p(x)$)





Elements of Information Theory
Thomas M. Cover

Data Processing System

Nature

Camera

CD

Markov Chain

- Definition: We say X, Y, Z is a Markov chain in this order, denoted

$$X \to Y \to Z,$$

  if we can write

$$p(x, y, z) = p(z|y)p(y|x)p(x).$$

- Special case:

$$X \to Y \to g(Y)$$

Markov Chain

- Definition: We say X, Y, Z is a Markov chain in this order, denoted

$$X \to Y \to Z,$$

  if we can write

$$p(x, y, z) = p(z|y)p(y|x)p(x).$$

- Special case:

$$X \to Y \to g(Y)$$

Markov Chain

- Examples
  - X is binary, you change w.p. p becomes Y, and you further corrupt it and it becomes Z.
  - Bent coin: probability of getting a head is θ. Generate a sequence of independent tosses $X_1, X_2, \cdots$ (Bernoulli(θ) process).

$$\bar{X}_n = \sum_{i=1}^{n} X_n$$

is Markov:

$$\theta \to \{X_1, \cdots, X_n\} \to \bar{X}_n$$

Markov Chain

Simple consequences

- $X \to Y \to Z$ iff X and Z are conditionally independent given Y.

  Proof:

$$p(x,z|y) = \frac{p(x,y,z)}{p(y)} = \frac{p(x)p(y|x)p(z|y)}{p(y)} = \frac{p(y,x)p(z|y)}{p(y)} = p(x|y)p(z|y)$$

- This characterization is true for general n-dimensional Markov field.
- Useful for checking Markovity

Markov Chain

Best definition of Markovity:

Past and future are conditionally independent given the present.

Data-Processing Inequality

- No clever manipulation of the data can improve inference

  Theorem. *If* $X \to Y \to Z$, *then the*

$$I(X;Y) \geq I(X;Z), \quad I(Y;Z) \geq I(X;Z).$$

  Equality iff $I(X;Y|Z) = 0$.

- Discouraging: we process information, then we will loose information
- Encouraging: sometimes we throw away something, equality still holds.

Data-Processing Inequality

Proof:

$$\begin{aligned} I(X;Y,Z) &= I(X;Z) + I(X;Y|Z) \\ &= I(X;Y) + \underbrace{I(X;Z|Y)}_{0} \end{aligned}$$

Since X and Z are cond. indept. given Y. So

$$I(X;Y) \geq I(X;Z).$$

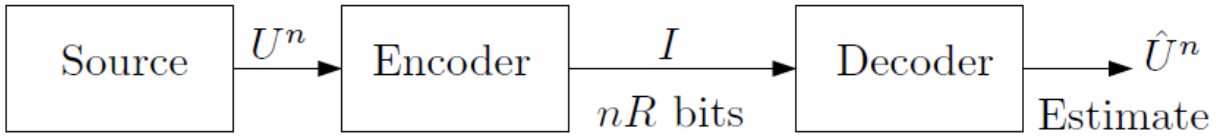
Equality iff $I(X;Y|Z) = 0$: $X \to Z \to Y$ form a Markov chain. Similarly, can also prove

$$I(Y;Z) \geq I(X;Z).$$

Data-Processing Inequality

Modeling data-compression systems

Compression system model:



- Encode message W from source using $X^n = (X_1, X_2, \cdots, X_n)$ (sequence of RVs)
- Through a channel, get Y^n,
- Decode to obtain $\hat{W}$.

$$I(W; \hat{W}) \leq I(X; Y).$$

Summary

- Mutual information is nonnegative
- Conditioning reduces entropy
- Uniform distribution maximizes entropy
- Properties
  - $D(p||q)$ convex in (p, q)
  - Entropy $H(p)$ concave in p
  - Mutual information $I(X;Y)$ concave in $p(x)$ (fixing $p(y|x)$), and convex in $p(y|x)$ (fixing $p(x)$)
- Data-processing inequality: data processing may (or may not)

  lose information

Lecture 2.3: Asymptotic Equipartition Property

(Weak) Law of Large Number

Theorem. *For independent, identically distributed (i.i.d.) random variables* X_i,

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \to EX, \quad \textit{in probability.}$$

- Convergence *in probability* if for every $\epsilon > 0$,

$$P\{|X_n - EX| > \epsilon\} \to 0.$$

- Proof by Markov inequality.
- So this means

$$P\{|\bar{X}_n - EX| \leq \epsilon\} \to 1, \quad n \to \infty.$$

Asymptotic Equipartition Property (AEP)

- LLN states that

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to EX$$

- AEP states that most sequences

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

$$p(X_1, X_2, \ldots, X_n) \approx 2^{-nH(X)}$$

- Analyze using LLN for product of random variables

Asymptotic Equipartition Property (AEP)

- LLN states that

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to EX$$

- AEP states that most sequences

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

$$p(X_1, X_2, \ldots, X_n) \approx 2^{-nH(X)}$$

- Analyze using LLN for product of random variables

Asymptotic Equipartition Property (AEP)

AEP lies in the heart of information theory.

- Proof for lossless source coding
- Proof for channel capacity
- and more...

Asymptotic Equipartition Property (AEP)

Theorem. *If* $X_1, X_2, \ldots$ *are i.i.d.* $\sim p(x)$, *then*

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \to H(X), \quad \textit{in probability.}$$

Proof:

$$\begin{aligned} -\frac{1}{n}\log p(X_1, X_2, \cdots, X_n) &= -\frac{1}{n}\sum_{i=1}^{n}\log p(X_i) \\ &\to -E\log p(X) \\ &= H(X). \end{aligned}$$

There are several consequences.

Typical Set

A typical set

$$A_\epsilon^{(n)}$$

contains all sequences $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ with the property

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

Typical Set

- Coin tossing example: $X \in \{0, 1\}$, $p(1) = 0.8$

$$p(1, 0, 1, 1, 0, 1) = p^{\sum X_i}(1-p)^{5-\sum X_i} = p^4(1-p)^2 = 0.0164$$

$$p(0, 0, 0, 0, 0, 0) = p^{\sum X_i}(1-p)^{5-\sum X_i} = p^4(1-p)^2 = 0.000064$$

- In this example, if

$$(x_1, \ldots, x_n) \in A_\epsilon^{(n)},$$

$$H(X) - \epsilon \leq -\frac{1}{n} \log p(X_1, \ldots, X_n) \leq H(X) + \epsilon.$$

- This means a binary sequence is in typical set is the frequency of heads is approximately k/n

Typical Set

$p = 0.6,\ n = 25,\ k =$ number of "1"s

k	$\binom{n}{k}$	$\binom{n}{k} p^k (1-p)^{n-k}$	$-\frac{1}{n} \log p(x^n)$
0	1	0.000000	1.321928
1	25	0.000000	1.298530
2	300	0.000000	1.275131
3	2300	0.000001	1.251733
4	12650	0.000007	1.228334
5	53130	0.000054	1.204936
6	177100	0.000227	1.181537
7	480700	0.001205	1.158139
8	1081575	0.003121	1.134740
9	2042975	0.013169	1.111342
10	3268760	0.021222	1.087943
11	4457400	0.077801	1.064545
12	5200300	0.075967	1.041146
13	5200300	0.267718	1.017748
14	4457400	0.146507	0.994349
15	3268760	0.575383	0.970951
16	2042975	0.151086	0.947552
17	1081575	0.846448	0.924154
18	480700	0.079986	0.900755
19	177100	0.970638	0.877357
20	53130	0.019891	0.853958
21	12650	0.997633	0.830560
22	2300	0.001937	0.807161
23	300	0.999950	0.783763
24	25	0.000047	0.760364
25	1	0.000003	0.736966

Typical Set

Theorem. *1. If* $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}$, *then for* n *sufficiently large:*

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

2. $P\{A_\epsilon^{(n)}\} \geq 1 - \epsilon$.

3. $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$.

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

Typical Set

Property 1

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

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

- Proof from definition:

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

  if

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

- The number of bits used to describe sequences in typical set is approximately $nH(X)$.

Typical Set

Property 2

$P\{A_\epsilon^{(n)}\} \geq 1 - \epsilon$ for n sufficiently large.

- Proof: From AEP: because

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

in probability, this means for a given $\epsilon > 0$, when n is sufficiently large

$$p\{\underbrace{\left|-\frac{1}{n}\log p(X_1, \ldots, X_n) - H(X)\right| \leq \epsilon}_{\in A_\epsilon^{(n)}}\} \geq 1 - \epsilon.$$

  - High probability: sequences in typical set are "most typical".
  - These sequences almost all have same probability - "equipartition".

Typical Set

Property 3 and 4: size of typical set

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

- Proof:

$$\begin{aligned}
1 &= \sum_{(x_1,\ldots,x_n)} p(x_1,\ldots,x_n) \\
&\geq \sum_{(x_1,\ldots,x_n)\in A_\epsilon^{(n)}} p(x_1,\ldots,x_n) \\
&\geq \sum_{(x_1,\ldots,x_n)\in A_\epsilon^{(n)}} p(x_1,\ldots,x_n)2^{-n(H(X)+\epsilon)} \\
&= |A_\epsilon^{(n)}|2^{-n(H(X)+\epsilon)}.
\end{aligned}$$

Typical Set

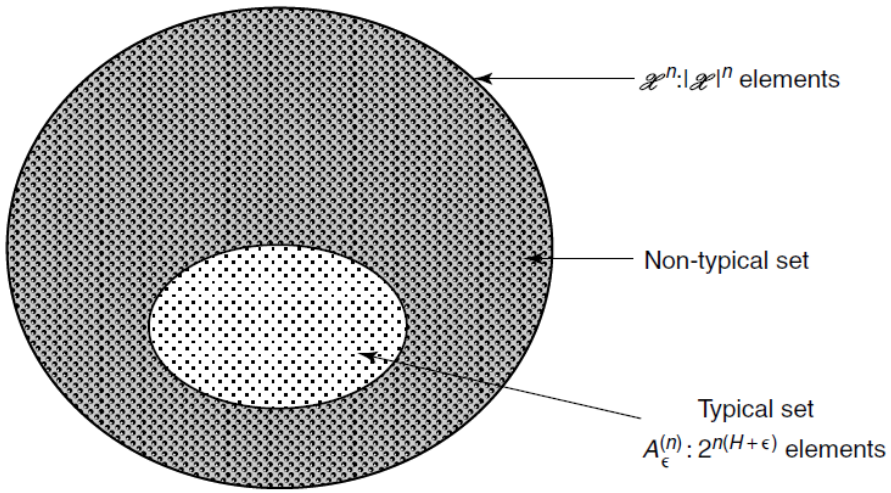
On the other hand, $P\{A_\epsilon^{(n)}\} \geq 1-\epsilon$ for n, so

$$\begin{aligned} 1-\epsilon &< \sum_{(x_1,\ldots,x_n)\in A_\epsilon^{(n)}} p(x_1,\ldots,x_n) \\ &\leq |A_\epsilon^{(n)}| 2^{-n(H(X)-\epsilon)}. \end{aligned}$$

- Size of typical set depends on $H(X)$.
- When $p = 1/2$ in coin tossing example, $H(X) = 1$, $2^{nH(X)} = 2^n$: all sequences are typical sequences.

Typical Set DIAGRAM

- This enables us to divide all sequences into two sets
  - Typical set: high probability to occur, sample entropy is close to entropy
    so we will focus on analyzing sequences in typical set
  - Non-typical set: small probability, can ignore in general

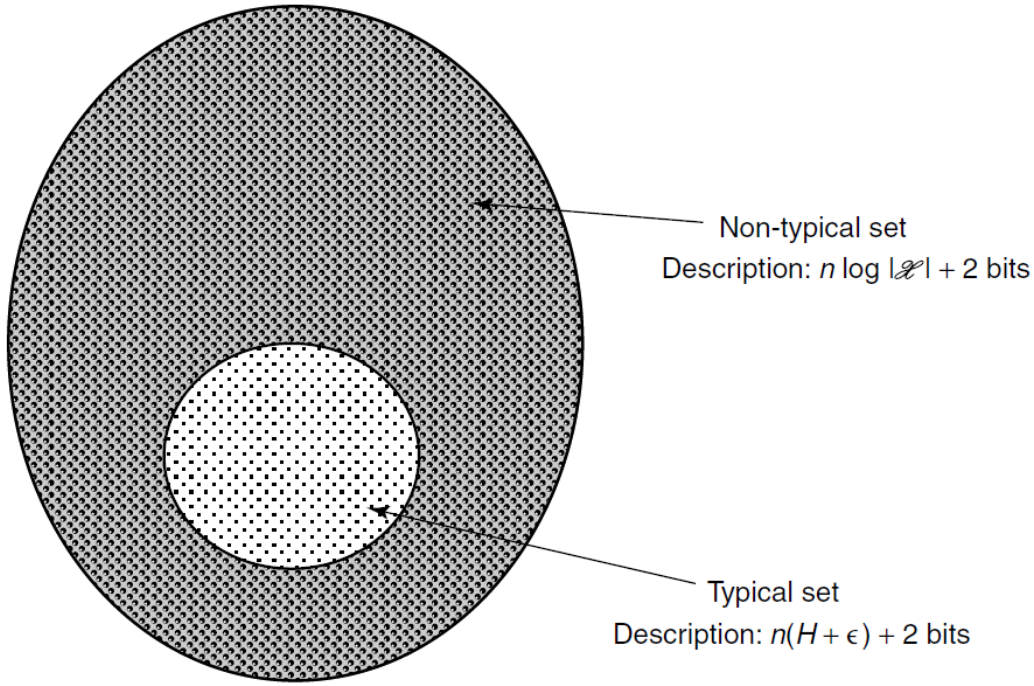


Data Compression (Source Coding) from AEP

- Let $X_1, X_2, \ldots, X_n$ be i.i.d. RV drawn from $p(x)$
- We wish to find short descriptions for such sequences of RVs

Data Compression from AEP

- Divide all sequences in $\mathcal{X}^n$ into two sets



Data Compression (Source Coding) from AEP

- Use one bit to indicate which set
  - Typical set $A_\epsilon^{(n)}$ use prefix "1"
    Since there are no more than $2^{n(H(X)+\epsilon)}$ sequences, indexing requires no more than $\lceil (H(X)+\epsilon) \rceil + 1$ (plus one extra bit)
  - Non-typical set use prefix "0"
    Since there are at most $|\mathcal{X}|^n$ sequences, indexing requires no more than $\lceil n \log |\mathcal{X}| \rceil + 1$

- Notation: $x^n = (x_1, \ldots, x_n)$, $l(x^n) =$ length of codeword for x^n

- We can prove

$$E\left[\frac{1}{n} l(X^n)\right] \leq H(X) + \epsilon$$

Summary

Almost everything is almost equally probable.

- Reasons that AEP has $H(X)$
  - $-\frac{1}{n}\log p(x^n) \to H(X)$, in probability
  - $n\big(H(X) \pm \epsilon\big)$ suffices to describe that random sequence on average
  - $2^{H(X)}$ is the effective alphabet size
  - Typical set is the smallest set with probability near 1
  - Size of typical set $2^{nH(X)}$
  - The distance of elements in the set nearly uniform