Lecture 5: Asymptotic Equipartition Property

- Law of large number for product of random variables
- AEP and consequences

Stock market

- Initial investment Y_0, daily return ratio r_i, in t-th day, your money is

$$Y_t = Y_0 r_1 \cdot \ldots \cdot r_t.$$

- Now if returns ratio r_i are i.i.d., with

$$r_i = \begin{cases} 4, & \text{w.p. } 1/2 \\ 0, & \text{w.p. } 1/2. \end{cases}$$

- So you think the expected return ratio is $Er_i = 2$,

- and then

$$EY_t = E(Y_0 r_1 \cdot \ldots \cdot r_t) = Y_0 (Er_i)^t = Y_0 2^t?$$

Is "optimized" really optimal?

- With $Y_0 = 1$, actual return Y_t goes like

$$1 \quad 4 \quad 16 \quad 0 \quad 0 \quad 0 \ldots$$

- Optimize expected return is not optimal?

- Fundamental reason: products does not behave the same as addition

(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 $\epsilon > 0$,

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

- Proof by Markov inequality.
- So this means

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

Other types of convergence

- In mean square if as $n \to \infty$

$$E(X_n - X)^2 \to 0$$

- With probability 1 (almost surely) if as $n \to \infty$

$$P\left\{\lim_{n\to\infty} X_n = X\right\} = 1$$

- In distribution if as $n \to \infty$

$$\lim_n F_n \to F,$$

where F_n and F are the cumulative distribution function of X_n and X.

Product of random variables

- How does this behave?

$$\sqrt[n]{\prod_{i=1}^{n} X_i}$$

- Geometric mean $\sqrt[n]{\prod_{i=1}^{n} X_i} \leq$ arithmetic mean $\frac{1}{n}\sum_{i=1}^{n} X_i$

- Examples:
  - Volume V of a random box, each dimension X_i, $V = X_1 \cdot \ldots \cdot X_n$
  - Stock return $Y_t = Y_0 r_1 \cdot \ldots \cdot r_t$
  - Joint distribution of i.i.d. RVs: $p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i)$

Law of large number for product of random variables

- We can write

$$X_i = e^{\log X_i}$$

- Hence

$$\sqrt[n]{\prod_{i=1}^{n} X_i} = e^{\frac{1}{n}\sum_{i=1}^{n} \log X_i}$$

- So from LLN

$$\sqrt[n]{\prod_{i=1}^{n} X_i} \to e^{E(\log X)} \leq e^{\log EX} = EX.$$

- Stock example:

$$E \log r_i = \frac{1}{2}\log 4 + \frac{1}{2}\log 0 = -\infty$$

$$E(Y_t) \to Y_0 e^{E \log r_i} = 0, \quad t \to \infty.$$

- Example

$$X = \begin{cases} a, & \text{w.p. } 1/2 \\ b, & \text{w.p. } 1/2. \end{cases}$$

$$E\left\{\sqrt[n]{\prod_{i=1}^{n} X_i}\right\} \to \sqrt{ab} \leq \frac{a+b}{2}$$

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

AEP lies in the heart of information theory.

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

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

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

Not all sequences are created equal

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

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

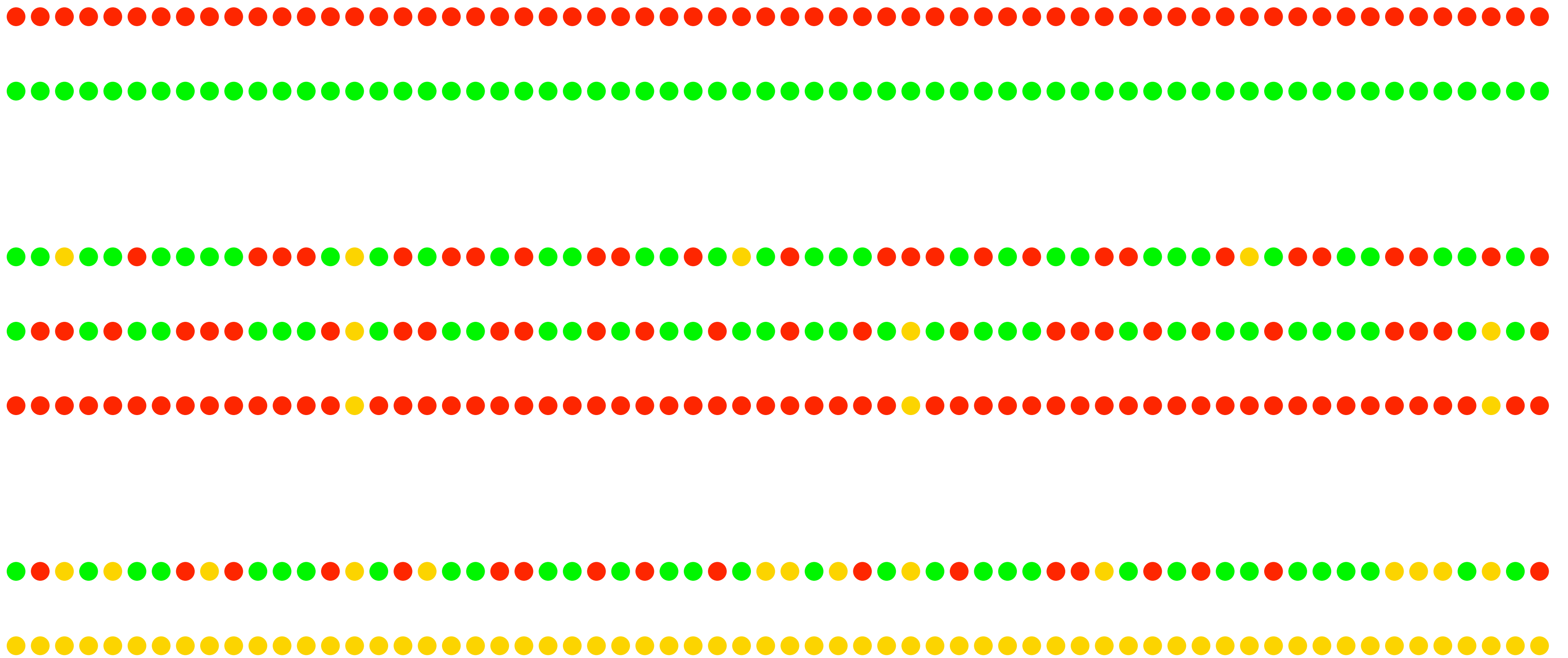
$$p(0,0,0,0,0,0) = p^{\sum X_i}(1-p)^{5-\sum X_i} = p^4(1-p)^2 = 0.000064$$

- In this example, if

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

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

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



$p = 0.6$, $n = 25$, $k =$ 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

Consequences of AEP

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

Property 1

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

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

- Proof from definition:

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

  if

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

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

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”.

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:

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

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

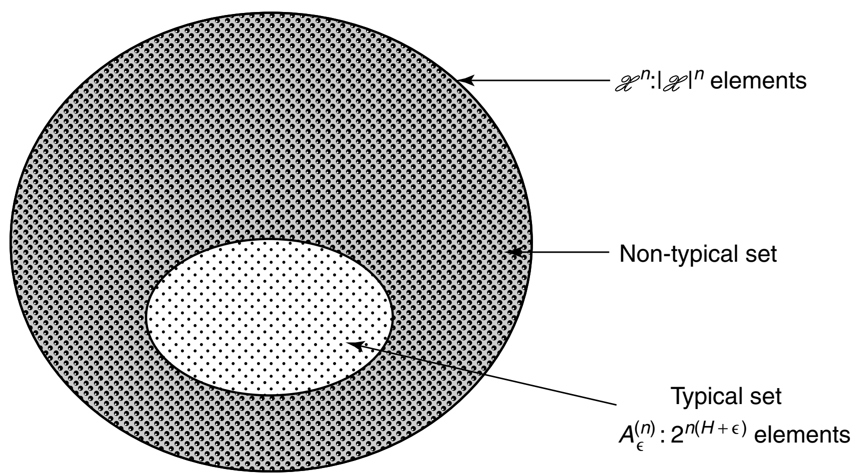
$$\begin{aligned} 1 - \epsilon &< \sum_{(x_1,\ldots,x_n)\in A_\epsilon^{(n)}} p(x_1, \ldots, x_n) \\ &\leq |A_\epsilon^{(n)}| 2^{-n(H(X)-\epsilon)}. \end{aligned}$$

- 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

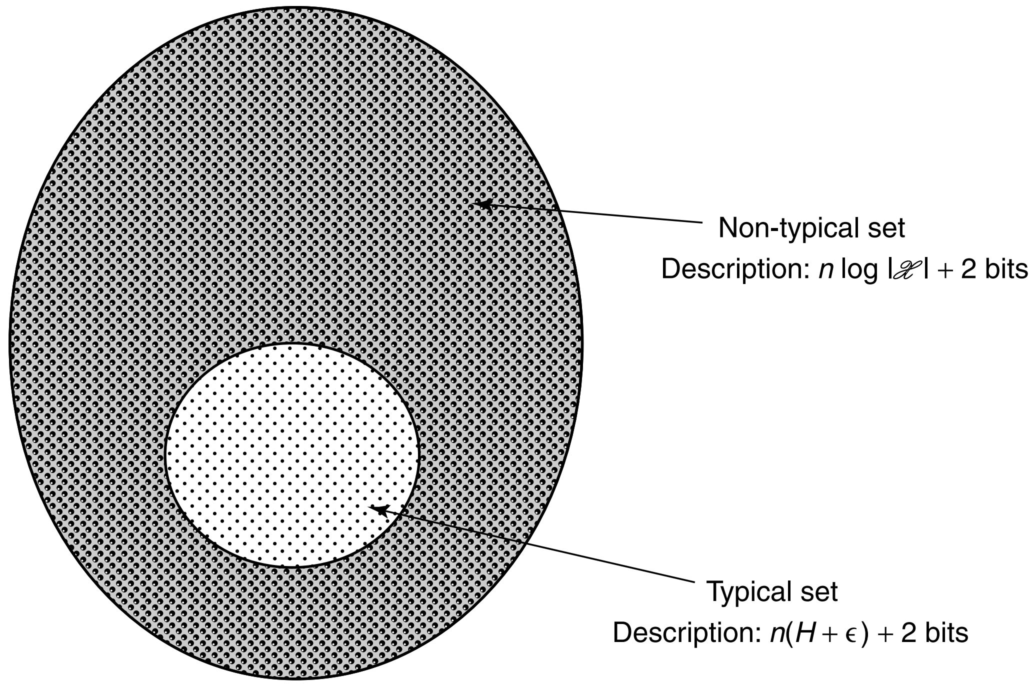
- This enables us to divide all sequences into two sets
  - Typical set: high probability to occur, sample entropy is close to true entropy
    so we will focus on analyzing sequences in typical set
  - Non-typical set: small probability, can ignore in general



Data compression scheme 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

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



- 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 of AEP

Almost everything is almost equally probable.

- Reasons that AEP has $H(X)$
  - $-\frac{1}{n}\log p(x^n) \to H(X)$, in probability
  - $n(H(X) \pm \epsilon)$ 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

Next Time

- AEP is about the property of independent sequences
- What about the dependence processes?
- Answer is entropy rate - next time.