Information Theory

2nd Module

Agenda

- Markov Chain and Entropy
- Sufficient Statistics
- Fano's Inequality
- Different Types of Convergences
- Asymptotic Equipartition Property
- High Probable Set vs. Typical Set
- Homeworks

Markov Chain

- Consider random variables X, Y, and Z.
- A chain of random variables $X \rightarrow Y \rightarrow Z$ is called Markov chain if

$$p(z \mid x, y) = p(z \mid y) .$$

- Note it implies $p(x, z|y) = p(x|y)\, p(z|x, y) = p(x|y)\, p(z|y)$
  - The first equality is due to conditional probability.
  - The second is due to Markov chain.
  - Thus, a MC $X \rightarrow Y \rightarrow Z$ implies, conditional independence between X and Z knowing Y.
- ***Conditioning on current, future and past are independent.***

Data Processing Inequality

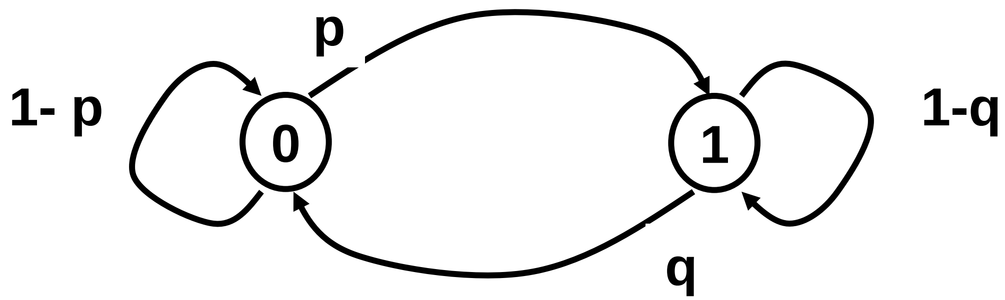
- If $X \rightarrow Y \rightarrow Z$, then $I(X; Y) \geq I(X; Z)$
- Proof:

$$I(X; Y, Z) = I(X ; Y) + I(X; Z| Y)$$

or $$= I(X; Z) + I(X; Y| Z)$$

  – We know $I(X; Z| Y) = 0$ and $I(X; Y | Z) \geq 0$ (why?)
  – Thus, $I(X; Y) \geq I(X; Z)$
  – Equality *iff* $I(X; Y|Z) = 0$, i.e., $X \rightarrow Z \rightarrow Y$ is a Markov chain.
- Let's use $Z:=g(Y)$, a function of Y.
- The function implies an arbitrary data processing on Y.
- The inequality implies then any data processing will not help us understand X any better.

Markov Chain



- Consider a Markov chain, $X_0, X_2, \ldots, X_n$
  - Transition matrix **P** = [1-p q ; p 1-q]
  - Initial distr. $\boldsymbol{\pi} = [\alpha; 1-\alpha]$;
  - Stationary distr. $s_0 = q/(p+q)$, $s_1 = p/(p+q)$, $\mathbf{s} = [s_0; s_1]$
  - $[\Pr\{X_1=0\}; \Pr\{X_1=1\}] = \mathbf{P}\,\boldsymbol{\pi}$
  - $\Pr\{X_1=0\} = \Pr\{X_1=0|X_0=0\}\Pr\{X_0=0\}+\Pr\{X_1=0|X_0=1\}\Pr\{X_0=1\}$
  - $\Pr\{X_1=1\} = \Pr\{X_1=1|X_0=0\}\Pr\{X_0=0\}+\Pr\{X_1=1|X_0=1\}\Pr\{X_0=1\}$

Markov Chain and Entropy

- Distr. at any n is $\mathbf{t}_n := [\Pr\{X_n=0\}; \Pr\{X_n=1\}] = \mathbf{P}^n\boldsymbol{\pi}$
- The stationary distr. is $\mathbf{s} = \lim_{n \to \infty} \mathbf{t_n}$
  - Or, simply solve $\mathbf{s} = \mathbf{P}\mathbf{s}$.
- Ex) p = 0.1, q=0.3, P = [0.9 0.3;0.1 0.7], P^{∞} = [0.75 0.75;0.25 0.25], **s** = [0.75; 0.25]
- Consider the following cases
  - $\boldsymbol{\pi}$ ~ uniform, **s** ~ non-uniform: $H(\mathbf{t_n})$ is decreasing toward $H(\mathbf{s})$
  - $\boldsymbol{\pi}$ ~ non-uniform, **s** ~ uniform: $H(\mathbf{t_n})$ is increasing toward $H(\mathbf{s})$

The Second Law of Thermodynamics

- *Entropy of an isolated system is non-decreasing*.
- This comes from the notion that the micro states in a thermodynamic system reach equally likely states in equilibrium (uniform stationary distr.)
  - If started off with non-uniform initial distr., then, entropy increases.
  - If started off with uniform initial distr. → then, entropy stays the same.

Sufficient Statistics

- Suppose an index set {θ: 1, 2, …, n} and a family of pmf's parameterized by θ, $\{f_1(x), f_2(x), \ldots, f_n(x)\}$.
- Let
  - X be a sample from *a* distribution in this family and
  - T(X) be a function of the sample (a statistic) for inference of θ.
- MC: $\theta \rightarrow X \rightarrow T(X)$
- Thus, in general $I(\theta; X) \geq I(\theta; T(X))$.
- When the equality is achieved, we call T(X)

  a sufficient statistic for inference on θ.
  - Basically, it implies that T(X) contains all the information for θ.
  - No loss of information for θ.

Example on Sufficient Statistics

- Consider a sequence of coin tosses, $X_1, X_2, \ldots, X_n$, iid with $X_i \in \{0,1\}$, with an unknown parameter $\theta = \Pr\{X_i = 1\}$.
- Given n, the number of 1's in n-trials is a *sufficient statistic* for θ.
  - $T(X_1, \ldots, X_n) = \sum_{i=1}^{n} X_i$
  - $\Pr\{X_1=1, X_2=1, \ldots, X_n=0$, i.e. k 1's$\} = \theta^k (1-\theta)^{n-k}$, for any $k \in \{0, 1, \ldots, n\}$
- Also $\hat{\theta} = \frac{T}{n}$ is the sufficient statistic for θ.
- Thus, we note that $\Pr\{X_1=x_1, X_2=x_2, \ldots, X_n=x_n \mid T = k\}$

$$= \begin{cases} 1/(n \text{ choose } k) & \text{if } \sum_{i=1}^{n} x_i = k \\ 0 & \text{o.w.} \end{cases}$$

- θ is independent of the sequence $\{X_i\}$ given T. Thus, $\theta \rightarrow T \rightarrow \{X_i, i=1,\ldots,n\}$ forms a MC. Thus, T is sufficient statistic for θ.

Sufficient Statistics (2nd Ex)

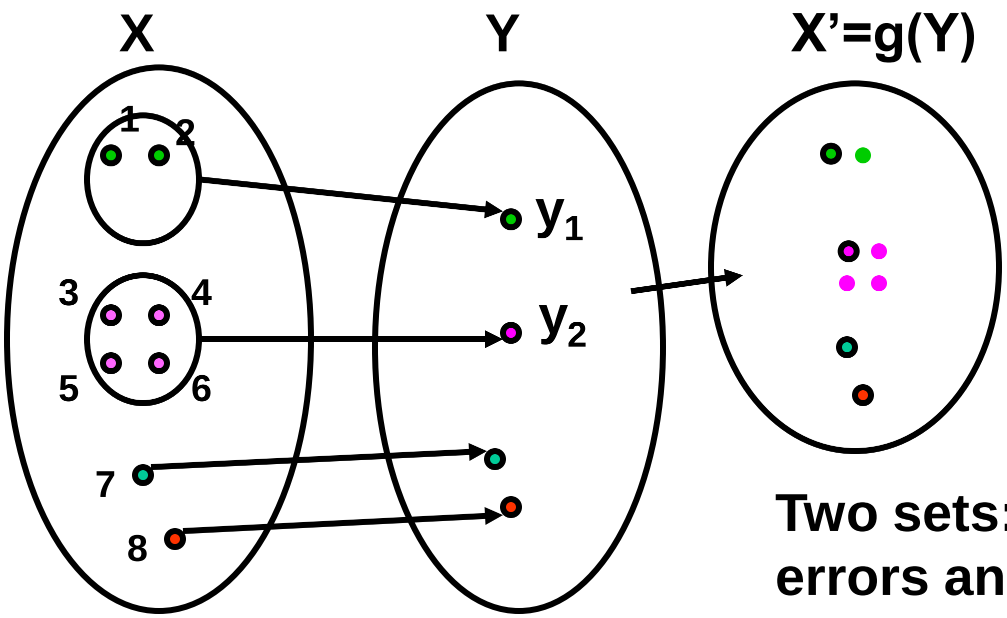
- Other examples of sufficient statistics

Fano's Inequality

- Consider the problem of "send X, observe Y, and make a guess g(Y) on X."
- Note that X → Y → X'=g(Y) forms a MC.
- FI relates the P_e := Pr{X':=g(Y) ≠ X} with H(X|Y).
- We already know H(X|Y) ≥ 0 with "=" iff X is a func. of Y:
  - Pr{X'(Y) ≠ X} = 0 iff H(X|Y) = 0
- Thus, we expect "*small* P_e for *small* H(X|Y)."

Fano's Inequality

- A thought experiment
- y_1 observed: two possibilities on X
  - P_e is 1/2
- y_2 observed: 4 possibilities on X
  - P_e is ¾
- We can divide the set {X = x} into two disjoint sets
  - {X' = X} = {1, 3, 7, 8}
  - {X'≠ X} = {2, 4, 5, 6}



Two sets: errors and corrects

Fano's Inequality (2)

- $H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|Y)$
- Or a weaker version is

  $1 + P_e \log|\mathcal{X}| \geq H(X|Y)$ or

  $P_e \geq (H(X|Y) - 1)/\log|\mathcal{X}|$
- Proof:

  Consider $E := \begin{cases} 1 & \text{if } X' \neq X \\ 0 & \text{o.w.} \end{cases}$

  Chain rule gives $H(E, X|\, Y) = H(X \mid Y) + H(E \mid X, Y)$

  $= H(E \mid Y) + H(X \mid Y, E)$

Fano's Inequality (3)

0

$$H(X \mid Y) + H(E \mid X, Y) = H(E \mid Y) + H(X \mid Y, E)$$

$\leq H(E) = H(P_e) \leq 1.0$

The last term can be bounded as

0

$$H(X|Y, E) = \Pr\{E=1\}\, H(X|Y, E=1\} + \Pr\{E=0\}\, H(X|Y, E=0\}$$

$$= P_e \sum_y p(y)\, H(X|Y=y, E=1)$$

---- But, we know $H(X|Y=y, E=1) \leq \log(|\mathcal{X}| - 1)$

for any y (There is at least one ω $X'(\omega) = X(\omega)$)

$$\leq P_e \log(|\mathcal{X}| - 1)$$

Therefore,

$$H(X|Y) \leq H(P_e) + P_e \log(|\mathcal{X}| - 1) \leq 1 + P_e \log(|\mathcal{X}| - 1) \qquad \text{Q.E.D.}$$

Types of Convergences

- *In distribution*: $X_n \Rightarrow X$ in distribution if

  $$F_n(x) = \Pr\{X_n \leq x\} \rightarrow F(x) = \Pr\{X \leq x\} \text{ as } n \rightarrow \infty$$

  – *Ex)* Let $X_1, X_2, \ldots$ iid fair binary {-1,+1} rvs. Then, $S_n = (1/\text{sqrt}(n)) \sum_{i=1}^{n} X_i$. Then, $F_n(y) := \Pr(S_n \leq y) \rightarrow \mathcal{N}(0, 1)$ (C.L.T.)
- *In probability*: $X_n \Rightarrow X$ *in probability* as $n \rightarrow \infty$ if $\forall\, \varepsilon > 0$

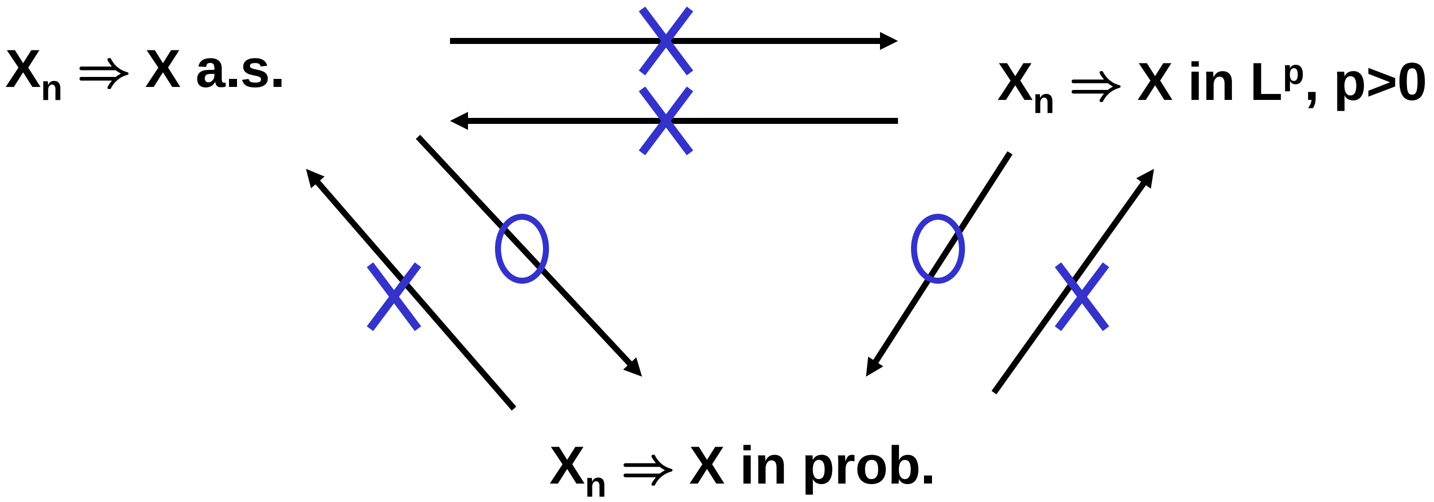
  $$\Pr\{\omega: |X_n(\omega) - X(\omega)| > \varepsilon\} \rightarrow 0 \text{ as } n \rightarrow \infty$$
- *In almost sure, almost everywhere sense, or with prob. 1*:

  $X_n \Rightarrow X$ a.s. as $n \rightarrow \infty$, if

  -- $\Pr\{\omega: \lim X_n(\omega) = X(\omega)\} = 1$, or

  -- For $\forall\, \varepsilon$, $\Pr\{\omega: |X_n(\omega) - X(\omega)| > \varepsilon, \text{i.o.}\} = 0$, as $n \rightarrow \infty$
- *In L^2* : $X_n \Rightarrow X$ in L^2, if $E\{|X_n - X|^2\} \rightarrow 0$, as $n \rightarrow \infty$

Relationship Between Different Types



Richard Durrett, Probability: Theory and Examples, 1991, Wadsworth

"$X_n \Rightarrow X$ a.s." $\not\Rightarrow$ "$X_n \Rightarrow X$ in prob."

- $X_n \Rightarrow X$ a.s. implies that for $\forall\, \varepsilon > 0$

  $$\lim_{k \to \infty} P\{\bigcup_{n \geq k} [|X_n - X| > \varepsilon]\} = 0$$

- Since $\{|X_k - X| > \varepsilon\} \subseteq \bigcup_{n \geq k} \{|X_n - X| > \varepsilon\}$,

  $$\Pr\{|X_k - X| > \varepsilon\} \leq \Pr(\bigcup_{n \geq k} \{|X_n - X| > \varepsilon\})$$

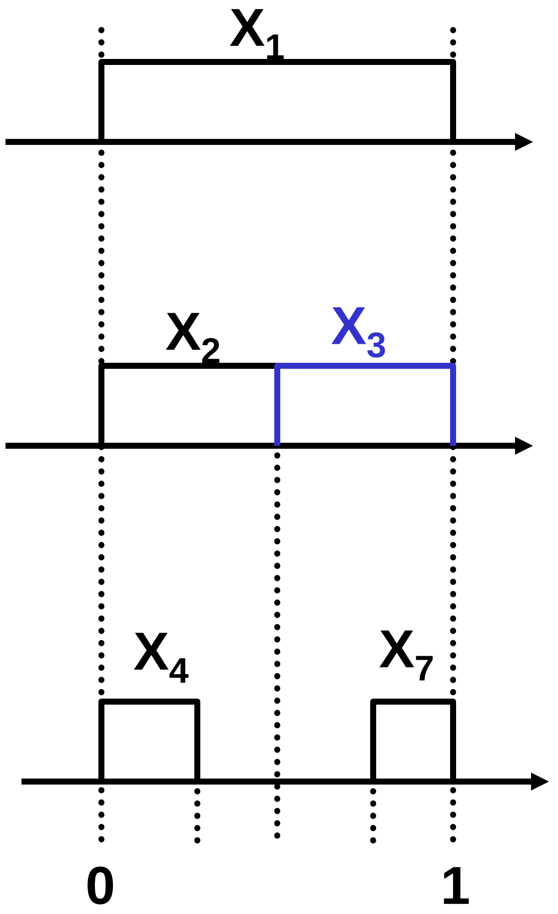
- Taking the limit on both sides,

  $$\lim_{k \to \infty} \Pr\{|X_k - X| > \varepsilon\} \leq \lim_{k \to \infty} \Pr(\bigcup_{n \geq k} \{|X_n - X| > \varepsilon\}) = 0$$

  Q.E.D.

$X_n \Rightarrow X$ in prob. $\not\Rightarrow$ $X_n \Rightarrow X$ a.s. (Converse is not true)

- Consider a series of r.v.'s $X_n := 1_{A_n}$ where A_n are defined as
  $A_1 = [0, 1]$;
  $A_2 = [0, 1/2)$, $A_3=[1/2, 1]$;
  $A_4 = [0, 1/4)$, $A_5 = [1/4, 1/2)$, $A_6 = [1/2, 3/4)$, $A_7 = [3/4, 1]$;
  …
- Let $\Pr\{X_n = 1\} = \text{length}(A_n)$ (Lebesque)
- Now, let $X = 0$. Then,
- For $\forall\ \varepsilon > 0$, $\Pr(|X_n - X| > \varepsilon) \to 0$ as $n \to \infty$
- But, $\{\omega: \lim X_n(\omega) = X(\omega)\} = \emptyset$
  Thus, $\Pr\{\omega: \lim X_n(\omega) = X(\omega)\} = 0.$
  Q.E.D.



Example for both "in prob." and "a.s."

- Consider a series of r.v. $X_n = 1_{An}$ where $A_1 = [0\ 1]$; $A_n = [0, 1/n]$, with the Lebesque measure as the prob.
- Let X = 0.
- With this example, we note that $X_n \Rightarrow X$ in both "in prob" and "a.s." senses

Laws of Large Numbers

- *Weak Law* of Large Numbers: Let $X_1, X_2, \ldots$ be i.i.d. with $E|X_1| < \infty$ and $E\{X_1\} = \mu$, and as $n \to \infty$,

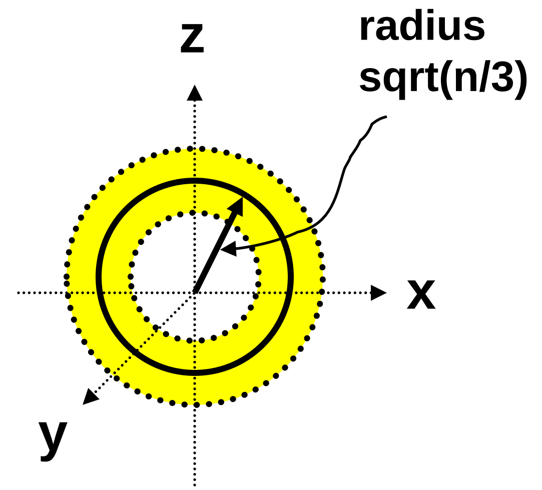
  $$S_n/n \Rightarrow \mu \text{ in probability}$$

  where $S_n = X_1 + X_2 + \ldots + X_n$.
- *Strong Law* of Large Numbers: $S_n/n \Rightarrow \mu$ a.s. as $n \to \infty$.
  - That is, it is in fact a.s.

- L^2 Weak Law: Let $X_1, X_2, \ldots, X_n$ be uncorrelated r.v.'s with $E\{X_i\} = \mu$ and $\text{var}(X_i) \le C < \infty$. Then, as $n \to \infty$

  $$S_n/n \Rightarrow \mu \text{ in } L^2$$

Surface Hardening



- A high-dimensional cube $[-1, 1]^n$ is almost the boundary of a ball.
- Let $X_1, X_2, \ldots$ be independent uniformly distributed on [-1, 1].
  - Then, $EX_i^2 = 1/3$.
- Then, the WLLN implies

  $(X_1^2 + \ldots + X_n^2)/n \to 1/3$ in probability as $n \to \infty$
- Consider an *n*-dimensional random vector $\mathbf{X} := (X_1, \ldots, X_n)$, and its length $||\mathbf{X}|| = \mathrm{sqrt}(X_1^2 + \ldots + X_n^2)$
- Thus, for $\forall\, \varepsilon > 0$, you can always find a large enough *n*, such that $\Pr\{|\, ||\mathbf{X}||^2/n - 1/3 \,| > \varepsilon\} \approx 0$
- $\Pr\{\mathbf{X} \in R^n: 1/3-\varepsilon < ||\mathbf{X}||^2/n < 1/3+\varepsilon\} \approx 1$

Length^2 = norm^2
$= \sum x_i^2$

$$Pr\{\mathbf{X} \in R^n : \sqrt{n(1/3-\epsilon)} < ||\mathbf{X}|| < \sqrt{n(1/3+\epsilon)}\} \approx \mathbf{1}$$

Asymptotic Equi-partition Property

- Let X_1, X_2, …, i.i.d. with p(x).
- The sample entropy
  - $H_n' = -(1/n) \log p(X_1=x_1, \ldots, X_n=x_1) = -(1/n) \sum_i \log p(X_i=x_i)$

  Converges in prob. to

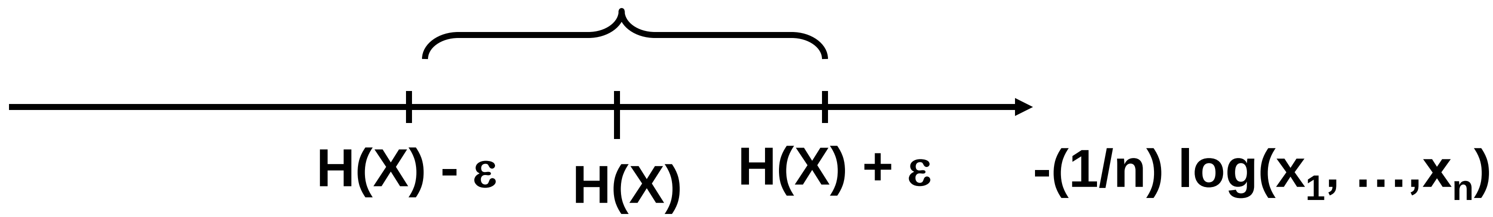
  the true entropy $H(X) = -\sum_i p(x_i) \log p(X_1=x_i)$.
- As $n \to \infty$, Ω can be divided into two mutually exclusive sets: The typical set and the non-typical set.
  - The sequences in the typical set have the sample entropy $\approx H(X)$
  - Those in the non-typical set have the sample entropy $\neq H(X)$
- From WLLN, Pr{Typical set} $\approx$ 1.0 as $n \to \infty$

Asymptotic Equi-partition Property (2)

- AEP: If $X_1, X_2, \ldots$ iid with p(x), then
  $H_n' := -(1/n) \log p(X_1, X_2, \ldots, X_n) = -(1/n) \sum_i \log p(X_i)$
  $\Rightarrow - E(\log p(X_1)) = H(X)$ *in prob.*
  (due to WLLN)
- This means, for $\forall\ \varepsilon > 0$

  $\Pr\{(x_1, \ldots, x_n): |H_n' - H(X)| > \varepsilon\} \rightarrow 0$ as $n \rightarrow 0$
  - Prob. of the *atypical* set goes to zero
  - Prob. of the *typcial* set goes to 1
- We can divide the entire set Ω, the set of all possible sequences of length *n*, into two mutually exclusive sets
  - Typical set $A_\varepsilon^{(n)} := \{(x_1, \ldots, x_n): |H_n' - H(X_1)| \leq \varepsilon\}$
  - Atypical set $\Omega - A_\varepsilon^{(n)}$

A sequence in the Typical Set $A_\varepsilon^{(n)}$



- For any sequence $(x_1, \ldots, x_n) \in A_\varepsilon^{(n)} := \{(x_1, \ldots, x_n): | -(1/n) \log p(x_1, \ldots, x_n) - H(X) | \leq \varepsilon \}$, the prob. of the sequence must have the following property

$$| -(1/n) \log p(x_1, \ldots, x_n) - H(X) | \leq \varepsilon$$

$$H(X) - \varepsilon \leq -(1/n) \log p(x_1, \ldots, x_n) \leq H(X) + \varepsilon$$

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

- Since we can choose a very small ε, the prob. of a sequence can be made very close to $2^{-nH(X)}$, as $n \to \infty$.

$\Pr\{A_\varepsilon^{(n)}\} > 1 - \varepsilon$, for *n* sufficiently large

- For any $\varepsilon > 0$ and $\delta > 0$, there exists an n_o such that $n > n_o$,
  $\Pr\{ \, | -(1/n) \log[p(x_1, \ldots, x_n)] - H(X) | \leq \varepsilon \, \} > 1 - \delta$.
- Choose $\delta = \varepsilon$.

The Size of the Typical Set $A_\varepsilon^{(n)}$

- The size of the typical set satisfies

1. $|A_\varepsilon^{(n)}| \le 2^{n(H(X) + \varepsilon)}$
2. $(1-\varepsilon)\, 2^{n(H(X) - \varepsilon)} \le |A_\varepsilon^{(n)}|$

- Proof of 1: $1 = \sum_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x})$

$$\ge \sum_{\mathbf{x} \in A_\varepsilon} p(\mathbf{x})$$

$$\ge \sum_{\mathbf{x} \in A_\varepsilon} 2^{-n(H(X)+\varepsilon)}$$

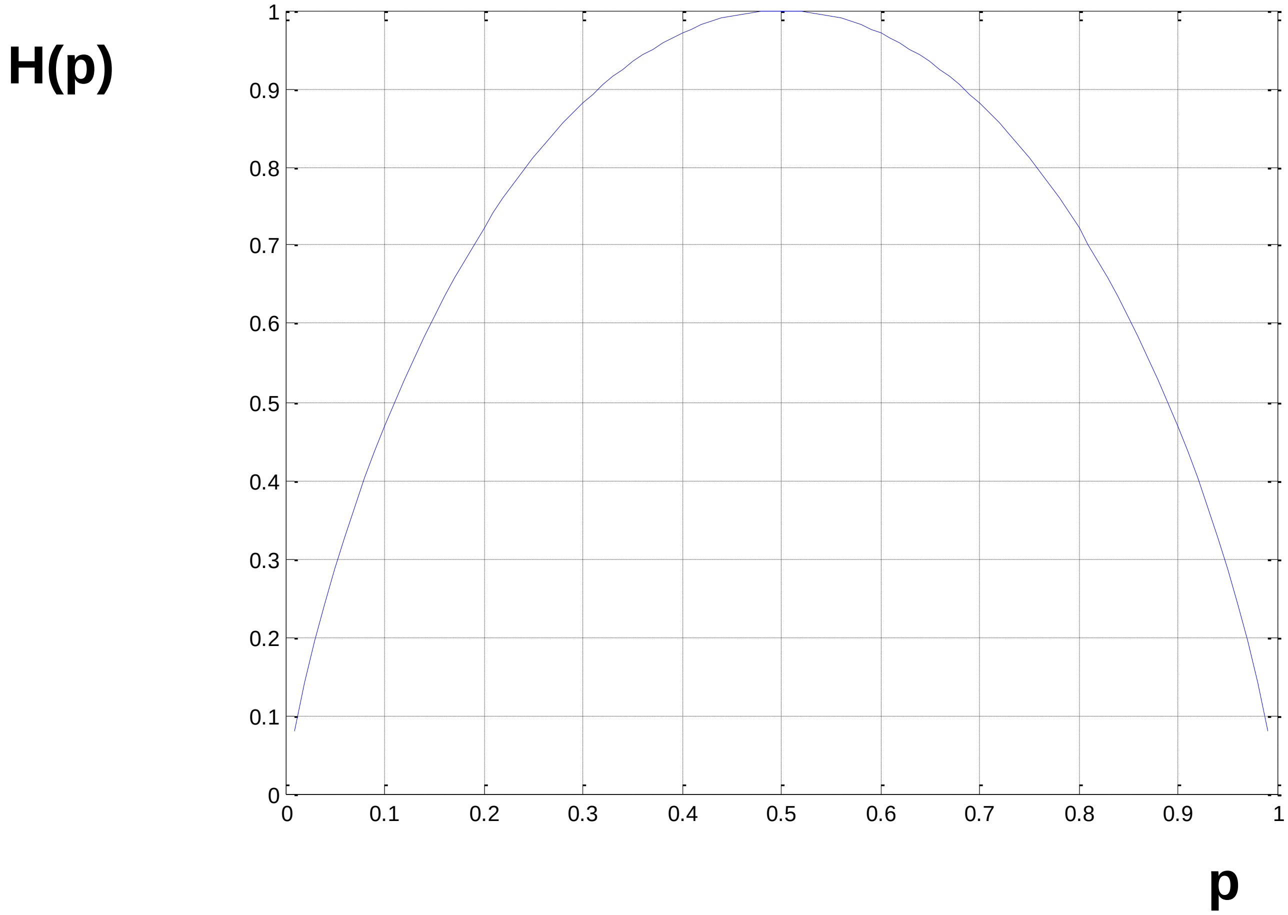
$$= |A_\varepsilon^{(n)}|\; 2^{-n(H(X)+\varepsilon)}$$ Q.E.D.

- Proof of 2: $1 - \varepsilon \le \Pr\{A_\varepsilon^{(n)}\} \le |A_\varepsilon^{(n)}|\, 2^{-n(H(X)-\varepsilon)}$

Example

- X_1 ~ binary r.v. taking 1 or 0, with prob. p and (1-p)
- Let $X_1, X_2, \ldots, X_n$ i.i.d.
- Ex. with n=6, p=2/3
  - The most typical sequences have 4 ones (np = 6*2/3 = 4).
  - The prob. of any sequence with 4 ones is $p^4 (1-p)^2$. There are (6 choose 4) number of such sequences.
  - There are total of 2^6 possible sequences.
- We can divide the complete set into the typical and the non-typical sets.
- In a trial, the sequences in the non-typical set occur rarely while those in the typical set occur very often.

H(p)



Example (2)

- Consider p = 0.5
  - Then, we note H(X) = 1; the size of typical set is 2^6; each and every sequences happens equally likely with prob. $1/2^6$

Example (3) at n=6

- Consider p = 0.061, H(0.061) = 0.33; the size of typical set is $2^{6*(0.33+1/6)} = 7.88$; compared to $2^6 = 64$
- A sequence in the typical set is expected to have *np* = 0.061*6 =0.37 number of 1's
- Exact calculation:

  - a seq. with no 1: $(1-p)^6 = 0.6855$

  (The most probable sequence and also most typical)

  - seq.'s with a single 1: $C^6_1(1-p)^5\, p = (6)\, 0.0445 = 0.2672$

  - These two kinds of sequences (7 seq's) account for 95% occurrences.

Example (4): n=10

- Consider n = 10 with p = 0.14. Then, H(0.14) = 0.58; nH = 5.8; the size of typical set is $2^{10*(0.58+1/10)} \approx 111$; prob.=(1/111) = 0.009; $2^{10} = 1024$
- Exact calculation:

Most probable

- a seq. with no 1: $(1-p)^{10} = 0.22$
- seq.'s with a single 1: $C^{10}_1 (1-p)^9 p = 0.036 (x 10) = 0.36$
- seq's with two 1's: $C^{10}_2 (1-p)^8 p^2 = (45) 0.0059 = 0.27$ **85%**
- seq's with three 1's: $C^{10}_3 (1-p)7 p^3 = (120) 9.5e-4 (120) = 0.11$ **96%**
- size of the 96% occurrence set is 1 + 10 + 45 + 120 = 176

Most typical set

Example (5): n = 100

- Consider n = 100 with p = 0.02. Then, H(0.02) = 0.1414; nH ≈ 14; the size of typical set is $2^{14} \approx 18054$; prob.=1/(18K) = 5.538e-5; $2^{100} = (1024)^{10}$
- Exact calculation:

  - a seq. with no 1: $(1-p)^{100}$ = 0.1326
  - seq.'s with a single 1: $(1-p)^{99}$ p = 0.0027, (x 100) = 0.27
  - seq's with two 1's: $(1-p)^{98} p^2$ = 5.25e-5, (x 4950) = 0.2734
  - seq's with three 1's: $(1-p)^{97} p^3$ = 1.12e-6, (x 161700) = 0.1823
  - seq's with four 1's: $(1-p)^{96} p^4$ = 2.3e-8, (x 3.9M) = 0.09
  - size of the 95% occurrence set is about 4 Million

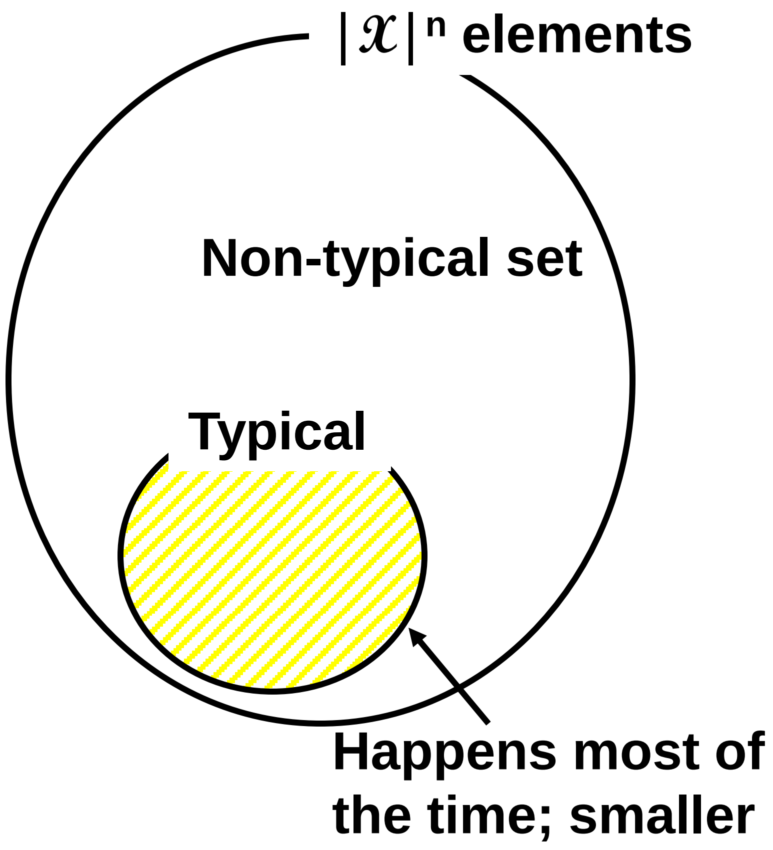
Most probable (→ 0.1326)

95% (brace around the five rows above)

Most typical set (→ 0.2734)

Consequences of AEP: Data Compression

- The size of the typical set is $2^{n(H(X) + \varepsilon)}$
- Data Compression Scheme:
- Seq.'s in typical set: In general, we need $(nH(X)+\varepsilon) + 1$ bits to represent them
  - Let's use 0 as prefix to denote membership to the typical set
  - $n(H(X) + \varepsilon) + 2$ bits in total
- Seq.'s in atypical set:
  - $n \log_2 |\mathcal{X}| + 1$ bits (Use prefix 1)



High Probability Sets and the Typical Sets

- Typical set is a small set that accounts for the most of the probability.
- But, is there a set smaller than the typical set, that accounts for the most of the probability?
- Theorem 3.3.1 states that the size of the typical set is the same as the size of the high probability set, to the first order in the exponent
  - The proof is easy, and outlined in prob. 3.11

High Probability Sets and the Typical Sets

- High Probability Set $B_\delta^{(n)} \subset \mathcal{X}^n$ is defined as a set

  $$\Pr\{B_\delta^{(n)}\} \geq 1 - \delta, \qquad \text{for } 1/2 > \delta > 0.$$

- The theorem indicates that the size of this set is

  $$\lim_{n \to \infty} (1/n) \log (|B_\delta^{(n)}|/|A_\delta^{(n)}|) = 0$$

- At a finite *n*, $(1/n) \log (|B_\delta^{(n)}|/|A_\delta^{(n)}|) = \varepsilon > 0$

  $$|B_\delta^{(n)}| = |A_\delta^{(n)}|\, 2^{n\varepsilon}$$

  – Both sizes grow exponentially fast
  – But the exponent of the growth is linear, *nH*

- Using Example (5), we note that the most probable set must include the all 0 sequence by definition; but the typical set may not include it (the most typical set include all the sequences with two ones).

Homework #2, #3

- HW#2
  - P2.6 (Conditional vs. unconditional mutual information)
  - P2.23 (Conditional MI)
  - P2.26 (Relative entropy is non negative)
  - P2.29 (Inequalities)
  - P2.34 (Entropy of initial condition)
  - P2.40 (Discrete Entropies)
  - P2.43 (MI of heads and tails)
  - P2.48 (Sequence length)
- HW#3
  - P2.21 (Markov inequality)
  - P2.30 (Maximum entropy)
  - P2.32, P2.33 (Fano's inequality)
  - P3.1 (Markov and Chebyshev inequalities)
  - P3.2 (AEP and MI)
  - P3.4 (AEP)
  - P3.10 (Random box size)
  - P3.13 (Calculation of typical set) Note the table on pg. 69 might have some errors. Generate your own and do the problem.