Lecture 3. Relation with Information Theory and Symmetry of Information

- Shannon entropy of random variable X over sample space S:

  $H(X) = \sum P(X=x) \log 1/P(X=x)$,

  the sum taken over x in S.

  Interpretation: H(X) bits are necessary on P- average to describe the outcome x in a prefix-free code.

Example. For P is uniform over finite S, we have $H(X)=\sum (1/|S|)\log |S| = \log |S|$.

- C(x), the Kolmogorov complexity, is the minimum description (smallest program) for one fixed x. It is a beautiful fact that H(X) and the P-expectation of C(x) converge to the same thing.
- The expected complexity and symmetry of information are examples that the two concepts approximately coincide.

Prefix (free) codes and the ubiquitous Kraft Inequality

- Prefix (free) code is a code such that no code word is a proper prefix of another one.
- Example: 1, 01, 000, 001 with length set l_1=1, l_2=2, l_3=3, l_4=3.

Kraft Inequality: $\sum_x 2^{-l_x} \leq 1.$ (*).

(a) (*) holds for {l_x} is the length set of a prefix code.

(b) If {l_x} satisfies (*) then ∃ prefix code with that length set.

Proof. Consider binary rooted tree with 0 labeling left edge and 1 labeling right edge. The prefix code words are leafs labeled with the concatenated labels of the edges on path from root to leaf. The code word length is # edges. Put weight 1 on root, weight ½ on each of its sons, ¼ on each of their sons, ... So the prefix code leaves x have weight 2^{-l_x}, and the sum of the weights ≤ 1. QED

Kullback-Leibler Divergence

- The KL-divergence between two probability mass functions P and Q is

- $D(P\|Q) = \sum P(x) \log (P(x)/Q(x))$.
  - Asymmetric
  - Always ≥ 0 and $=0$ only if $P = Q$

Noiseless Coding

- The expected code word length with code c for random variable X is $l(X,c)=\sum P(X=x)\,|c(x)|$.

Noiseless Coding Theorem (Shannon):

- $H(X) \leq \min\{l(X,c): c \text{ is prefix code}\} \leq H(X)+1$.

Proof: (left ≤) Let P be as above, and let $\{l_x : x \text{ in } S\}$ be a length set of a prefix code. Let Q be a probability mass function defined by $Q(x)=2^{-l_x}$ (Q is probability by Kraft inequality).

Then, $-D(P\|Q) = \sum P(x) \log 1/P(x) - \sum P(x) \log 1/Q(x)$

$= H(P)-\sum P(x) \log 1/Q(x) = H(P)-\sum P(x)\, l_x) \leq 0$ and $= 0$ only if $Q=P$.

(right ≤) Shannon-Fano code achieves this by coding x as c(x) with length $\leq \log 1/P(x)+1$. (Code word length = $\log 1/P(x)$ rounded upwards).

QED

Asymptotic Equivalence Entropy and Expected Complexity

- String x=y_1 ... y_m (l(y_i)=n)

p_k= d({i: y_i = k}) / m for k = 1,...,2^n=N.

Theorem (2.8.1 in book). C(x) ≤ m(H + ε(m))
with H= ∑ p_k log 1/p_k , ε(m) = 2^{n+1} l(m)/m
(ε(m)→0 for m→∞ and n fixed).

Proof. C(x) ≤ 2l(mp_1)+...+ 2l(mp_N)+l(j) with j ≤ $\binom{m}{mp\_1 \ldots mp\_N}$

Since mp_k ≤ m we have C(x)≤2N l(m) + l(j), and writing multinomial as factorials and using Stirling's approximation, the theorem is proved. ■

Continued

- For x is an outcome of a sequence of independent trials, a random variable X, the inequality can be replaced by an asymptotic equality w.h.p. Namely, X uniform with 2^n outcomes:
- $H(X)= \sum P(X=x) \log 1/P(X=x)$ and $E=\sum P(X=x) C(x)$ ($l(x)=n$). There are at least $2^n(1-2^{-c+1})$ many x's with $C(x) \geq n-c$.
- Hence, $n/(n+O(1)) \leq H(X)/E \leq n/(1-2^{-c+1})(n-c))$. Substitute $c=\log n$ to obtain

$$\lim H(X)/E = 1 \text{ for } n \rightarrow \infty.$$

Symmetry of Information

- In Shannon information theory, the symmetry of information is well known:

- $I(X;Y)=I(Y;X)$ with

$I(X;Y)=H(Y)-H(Y|X)$ (information in random variable X about random variable Y)

Here X,Y are random variables, and probability $P(X=x) = p_x$ and the entropy

$H(X) = \sum p_x \log 1/p_x$.

- The proof is by simple rewriting.

Algorithmic Symmetry of Information

- In Kolmogorov complexity, the symmetry of information is : I(x;y)=I(x;y) with I(x;y)=C(y)-C(y|x) up to an additive log term. The proof is totally different, as well as the meaning, from Shannon's concept.
- The term C(y)-C(y|x) is known as "the information x knows about y". That information is symmetric was first proved by Levin and Kolmogorov (in Zvonkin-Levin, Russ. Math Surv, 1970)

Theorem (2.8.2 book). C(x)-C(x|y)=C(y)-C(y|x), up to an additive log-term.

Proof. Essentially we will prove:

$$C(x,y)=C(y|x)+C(x) + O(\log C(x,y)).$$

(Since $C(x,y)=C(x|y)+C(y) + O(\log C(x,y))$, Theorem follows).

(≤). It is trivial that $C(x,y) \le C(y|x)+C(x) + O(\log C(x,y))$ is true.

(≥). We now need to prove $C(x,y) \ge C(y|x)+C(x) + O(\log C(x,y))$.

Proving: $C(x,y) \geq C(y|x)+C(x) + O(\log C(x,y))$.

Assume to the contrary: for each $c \geq 0$, there are x and y s.t.

$$C(x,y)<C(y|x)+C(x) - c \log C(x,y) \qquad (1)$$

Let $A=\{(u,z): C(u,z) \leq C(x,y)\}$. Given C(x,y), the set A can be recursively enumerated.

Let $A_x=\{z: C(x,z) \leq C(x,y)\}$. Given C(x,y) and x, we have a simple algorithm to recursively enumerate A_x. One can describe y, given x, using its index in the enumeration of A_x , and C(x,y). Hence

$$C(y|x) \leq \log |A_x| + 2\log C(x,y) + O(1) \qquad (2)$$

By (1) and (2), for each c, there are x, y s.t.

$|A_x|>2^e$, where $e=C(x,y)-C(x)+(c-2) \log C(x,y)-O(1)$.

But now, we obtain a too short description for x as follows. Given C(x,y) and e, we can recursively enumerate the strings u which are candidates for x by satisfying condition

$$A_u=\{z: C(u,z) \leq C(x,y)\}, \text{ and } 2^e <|A_u|. \qquad (3)$$

Denote the set of such u by U. Clearly, $x \in U$. Also

$$\{(u,z) : u \in U \ \& \ z \in A_u\} \subseteq A \qquad (4)$$

The number of elements in A cannot exceed the available number of programs that are short enough to satisfy its definition:

$$|A| \leq 2^{C(x,y)+O(1)} \qquad (5)$$

Note that $\{u\} \times A_u$ is a disjoint subset of A for every different u in U. Using (3), (4), (5),

$$|U| \leq |A| / \min \{ |\{A_u| : u \text{ in } U\}| < |A| /2^e \leq 2^{C(x,y)+O(1)} / 2^e$$

Hence we can reconstruct x from C(x,y), e, and the index of x in the enumeration of U. Therefore

$$C(x) < 2\log C(x,y) + 2\log e + C(x,y) - e +O(1)$$

substituting e as given above yields C(x)<C(x), for large c, contradiction. QED

Symmetry of Information is sharp

Example.

Let n be random: $C(n) = |n|+O(1)$.

Let x be random of length n: $C(x|n)=n+O(1)$

and $C(x) = n+O(1)$.

Then:

$C(n)-C(n|x) = |n|+O(1) = \log n+O(1)$

$C(x)-C(x|n) = n-n+O(1) = O(1)$.

So $I(x:n) = I(n:x)+\log n +O(1)$.

Kamae Theorem

For each natural number m, there is a string x such that for all but finitely many strings y,

$I(y;x) = C(x) - C(x|y) \geq m$

That is: there exist finite objects x such that almost all finite objects y contain a large amount of information about them.