EE5585 Data Compression April 02, 2013

Lecture 18

Instructor: Arya Mazumdar *Scribe: Shashanka Ubaru*

Solutions for problems 1 - 4 and 6 of HW2 were provided in the previous lecture and also the concepts of Turing machines and Kolmogorov Complexity were introduced . In this lecture,

1. Solution for the fifth problem of HW2 is provided.
2. Kolmogorov Complexity is defined.
3. Properties (Theorems related to) Kolmogorov Complexity are stated and proved.
4. Concept of Incompressible Sequences is introduced.

(Reference for these topics : Chapter 14, "Elements of Information Theory" 2nd ed - T. Cover, J. Thomas (Wiley, 2006).)

Solution for Problem 5 of Homework 2

Given: A source X uniformly distributed on the set $\{1, 2, \ldots\ldots, m\}$. That is, $\Pr(x = i) = \frac{1}{m}$.
$R(D) =?$ with Hamming distortion,

$$d(x, \widehat{x}) = \begin{cases} 0 & if\; x = \widehat{x} \\ 1 & if\; x \neq \widehat{x} \end{cases}$$

We know that the rate distortion is given by,

$$R(D) = \min_{p(\widehat{x}|x):E\{d(x,\widehat{x})\leq D\}} I(X; \widehat{X})$$

This optimization equation seems difficult to solve. So, a good trick is to find a lower bound for $I(X; \widehat{X})$ subjected to the constraint mentioned and come up with an example that achieves this lower bound given the constraint on $p(\widehat{x} \mid x)$. (Recall : This technique was used to find $R(D)$ for Binary and Gaussian random variables also)

By definition,

$$\begin{aligned} I(X; \widehat{X}) &= H(X) - H(X \mid \widehat{X}) \\ &= \log m - H(X \mid \widehat{X}) \end{aligned}$$

For binary random variable, we had equaled $H(X \mid \widehat{X})$ to $H(X - widehatX \mid \widehat{X})$, but here this is not true. So, we define a new random variable Y,

$$Y = \begin{cases} 0 & if\; X = \widehat{X} \\ 1 & if\; X \neq \widehat{X} \end{cases}$$

$H(X \mid \widehat{X})$ is the uncertainity in X if $\widehat{X}$ is known and we have,

$$\begin{aligned} H(X \mid \widehat{X}) &\leq H(X, Y \mid \widehat{X}) \\ &= H(Y \mid \widehat{X}) + H(X \mid \widehat{X}, Y). \end{aligned}$$

Substituting,

$$\begin{aligned} I(X;\widehat{X}) &\geq H(X) - H(Y \mid \widehat{X}) - H(X \mid \widehat{X}, Y) \\ &\geq \log m - H(Y) - H(X \mid \widehat{X}, Y) \end{aligned}$$

as $H(Y) \geq H(Y \mid \widehat{X})$. Now consider $H(X \mid \widehat{X}, Y)$,

$$H(X \mid \widehat{X}, Y) = \Pr(Y=0)H(X \mid \widehat{X}, Y=0) + \Pr(Y=1)H(X \mid \widehat{X}, Y=1)$$

If $Y = 0 \Rightarrow X = \widehat{X} \Rightarrow H(X \mid \widehat{X}, Y = 0)$ there is no uncertainity and for a given $\widehat{X}$, there are only M-1 choices for X.

$$\begin{aligned} H(X \mid \widehat{X}, Y) &= \Pr(Y=1)\log(M-1) \\ &= \Pr(X \neq \widehat{X})\log(M-1) \end{aligned}$$

and $H(Y) = h\left(\Pr(X \neq \widehat{X})\right)$ then,

$$\begin{aligned} I(X;\widehat{X}) &\geq \log m - h\left(\Pr(X \neq \widehat{X})\right) - \Pr(X \neq \widehat{X})\log(M-1) \\ Ed(x;\widehat{x}) &= 1.\Pr(X \neq \widehat{X}) + 0.\Pr(X = \widehat{X}) \\ &= \Pr(X \neq \widehat{X}) \leq D \\ I(X;\widehat{X}) &\geq \log m - \underbrace{D\log(M-1) - h(D)}_{both\,are\,increasing\,functions} \end{aligned}$$

Example to show that, this lower bound is achieved

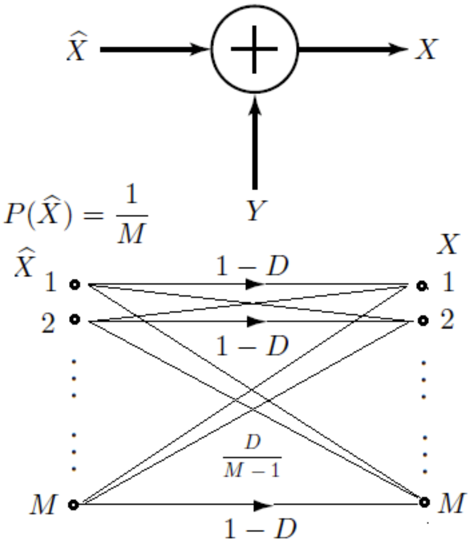

Figure 1: System that achieves the lower bound

Consider the system shown in Figure 1. $\widehat{X} \in \{1, 2, \ldots\ldots, M\}$ with $\Pr(\widehat{X}) = \frac{1}{M}$. If the distortion is D then, $\Pr(X = i \mid \widehat{X} = i) = 1 - D$ and $\Pr(X = i \mid \widehat{X} = j) = \frac{D}{M-1}$ $\forall i \neq j$ as shown in the figure. The probability X is,

$$
\begin{aligned}
\Pr(X=i) &= \Pr(\widehat{X}=i)\Pr(X=i\mid\widehat{X}=i)+\sum_{i\neq j}\Pr(\widehat{X}=j)\Pr(X=i\mid\widehat{X}=j)\\
&= \frac{1}{M}(1-D)+\sum_{i\neq j}\frac{1}{M}\frac{D}{M-1}\\
&= \frac{1-D}{M}+\frac{D}{M}\\
&= \frac{1}{M}
\end{aligned}
$$

So, X are equiprobable. The mutual information is given by,

$$
\begin{aligned}
I(X;\widehat{X}) &= \log m - H(X\mid\widehat{X})\\
H(X\mid\widehat{X}) &= -\sum_{i=1}^{M}\Pr(\widehat{X}=i)H(X\mid\widehat{X}=i)\\
&= \frac{1}{M}\sum_{i=1}^{M}H(X\mid\widehat{X}=i)\\
&= H(X\mid\widehat{X}=i)
\end{aligned}
$$

We have, $\Pr(X=i\mid\widehat{X}=i)=1-D$ and $\Pr(X=i\mid\widehat{X}=j)=\frac{D}{M-1}\ \forall i\neq j$. So,

$$
\begin{aligned}
H(X\mid\widehat{X}=i) &= -(1-D)\log(1-D)-\sum_{i\neq j}\frac{D}{M-1}\log\left(\frac{D}{M-1}\right)\\
&= -(1-D)\log(1-D)-D\log\left(\frac{D}{M-1}\right)\\
&= h(D)+D\log(M-1)
\end{aligned}
$$

Substituting,

$$
\begin{aligned}
I(X;\widehat{X}) &= \log m - h(D) - D\log(M-1)\\
R(D) &= \log m - h(D) - D\log(M-1)
\end{aligned}
$$

Digression:

What happens, if we use scalar quantizer for the above mentioned system (quantize $X\in\{1,2,\ldots\ldots,M\}$)?

Suppose we use an uniform quantizer.

$$
\Big|\underbrace{1,\ 2,}_{\leftarrow\Delta\rightarrow}\Big|\ \Big|\underbrace{\ldots\ldots}_{\leftarrow\Delta\rightarrow}\Big|\ \ldots\ldots\ \Big|\underbrace{,\ M-1,\ M}_{\leftarrow\Delta\rightarrow}\Big|
$$

The reconstruction points will be $i\left(\frac{\triangle+1}{2}\right)$ for i=1,3,....M-1 odd no.s

Find average distortion D: This will be same for each points (uniform quantizer). We are using hamming distortion. $d=\begin{cases}1 & \Delta\neq i\\ 0 & \Delta=i\end{cases}$. Thus,

$$\begin{aligned} D &= \frac{\Delta - 1}{\Delta}.1 + \frac{1}{\Delta}.0 \\ &= \frac{\Delta - 1}{\Delta} \end{aligned}$$

Rate-Distortion trade-off: we have $\frac{M}{\Delta}$ possible outputs. So, $\log \frac{M}{\Delta}$ number of bits. Then,

$$R(D) = \log \frac{M}{\Delta}$$

But, $D = \frac{\Delta-1}{\triangle} = 1 - \frac{1}{\Delta}$ implies, $\frac{1}{\Delta} = 1 - D$
Thus the rate-distortion for this scheme is,

$$\begin{aligned} R &= \log M(1 - D) \\ &= \log M - \log \frac{1}{1 - D} \end{aligned}$$

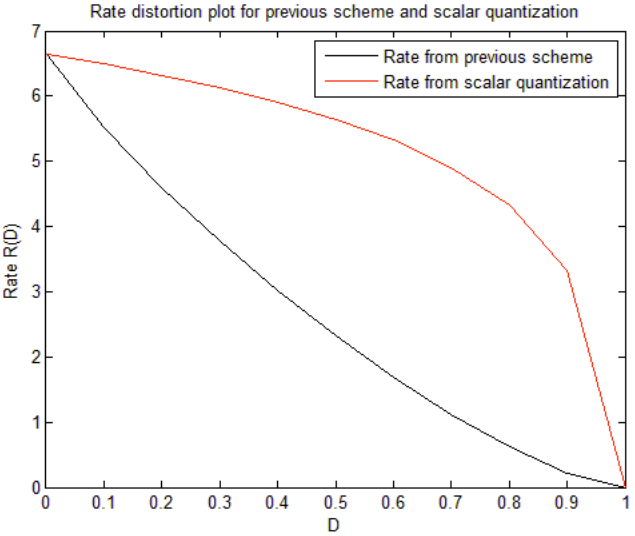

Figure 2: Rate distortion function and performance of and scalar quantization

From Figure 2 , it is evident that, the rate for scalar quantizer is always higher than the rate-distortion function. Thus, this scheme is suboptimal.

Kolmogorov Complexity

Introduction

So far, a given sequence of data (object) X was treated as a random variable with probability mass function $p(x)$. And the attributes (properties) defined for the sequence like entropy $H(X)$, average length $L(C)$, relative entropy divergence $D(p \parallel q)$, Rate-distortion $R(D)$ etc., depended on the probability distribution of the sequence. Most of the coding techniques and quantization techniques that we saw, also depended on the probability distribution of the sequence. We can define a descriptive complexity of the event $X = x$ as $\log \frac{1}{p(x)}$. But, Andrey Kolmogorov (Soviet mathematician) defined an algorithmic (descriptive) complexity of an object X to be the length of the shortest binary computer program that describes the object. He also observed that, this definition of complexity is essentially computer independent. Here, the object X is treated as strings of data that enter a computer and Kolmogorov Complexity is analogous to Entropy of this sequence.

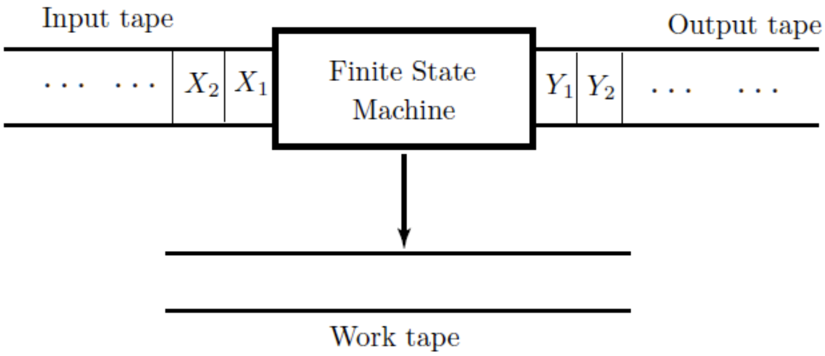

Figure 3: A Turing machine

An acceptable model for computers that is universal, in the sense that they can mimic the actions of other computers is 'Turing Machine' model. This model considers a computer as a finite-state machine operating on a finite symbol set. A computer program is fed left to right into this finite-state machine as a program tape (shown in Figure 2). The machine inspects the program tape, writes some symbols on a work tape and changes its state according to its transition table and outputs a sequence Y. In this model, we consider only a program tape containing a halt command (ie,. when to stop the program). No program leading to a halting computation can be the prefix of another program. This forms a prefix-free set of programs. Now the question is, given a string/sequence, can we compress it or not?

Answer: Kolmogorov Complexity and its properties.

Kolmogorov Complexity: Definitions and Properties

Definition:

The Kolmogorov complexity $K_{\mathcal{U}}(x)$ of a string x with respect to a universal computer $\mathcal{U}$ is defined as

$$K_{\mathcal{U}}(x) = \min_{p:\mathcal{U}(p)=x} l(p)$$

the minimum length over all programs that print x and halt. Thus, $K_{\mathcal{U}}(x)$ is the shortest description length of x over all descriptions interpreted by computer $\mathcal{U}$.

Conditional Kolmogorov complexity knowing $l(x)$ is defined as

$$K_{\mathcal{U}}(x \mid l(x)) = \min_{p:\mathcal{U}(p,l(x))=x} l(p)$$

This is the shortest description length if the computer $\mathcal{U}$ has the length of x made available to it.

Property 1:

If $\mathcal{U}$ is a universal computer, for any other computer $\mathcal{A}$ there exists a constant $\mathcal{C}$ such that,

$$K_{\mathcal{U}}(x) \leq K_{\mathcal{A}}(x) + \mathcal{C}$$

The constant $\mathcal{C}$ does not depend on x. All universal computers have same $K(x)$ so they differ by $\mathcal{C}$.

Property 2:

$$K(x|l(x)) \leq l(x) + c.$$

Conditional complexity is less than the length of the sequence. That is, the length of a program will be atmost length of our string x.

Example: "Print the following l length sequence : $x_1 \ldots\ldots x_l$"

Here l is given so the program knows when to stop.

Property 3:

$$K(x) \leq K(x|l(x)) + \log^*(l(x)) + c.$$

where $\log^*(n) = \log n + \log\log n + \log\log\log n + \cdots\cdots$

Here we do not know the length $l(x)$ of the sequence. The length of a sequence is represented as $\log l(x)$. But $\log l(x)$ is unknown. So, we need $\log\log l(x)$ and so on. Hence the term $\log^* l(x)$.

Property 4:

The number of binary sequence x with complexity $K(x) < k$ is $< 2^k$,

$$|\{x \in \{0,1\} : K(x) < k\}| < 2^k$$

The total length of any program of binary sequence is,

$$1 + 2 + 2^2 + 2^4 + ... + 2^{k-1} = 2^k - 1 < 2^k$$

Property 5:

The Kolmogorov complexity of a binary string X is bounded by

$$K(x_1 x_2 \cdots\cdots x_n \mid n) \leq nH\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) + \log^* n + c$$

Suppose our sequence X has 'k' ones. Can we compress a sequence of n bits with k ones? Given a table of X with k ones, our computer produces an index of length, $\log\binom{n}{k}$. But we do not know 'k'. So, to know 'k' we need $\log^* k$. So, the worst case length will be,

$$\log\binom{n}{k} + \log^* n + c$$

By Sterling's approximation,

$$\log\binom{n}{k} \leq nH\left(\frac{k}{n}\right) = nH\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)$$

and we know, $K(X) \leq l(X)$,

$$K(x_1x_2\cdots\cdots x_n \mid n) \leq nH\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) + \log^* n + c$$

Property 6:

The halting programs form a prefix-free set, and their lengths satisfy the Kraft inequality,

$$\sum_{p:p \text{ halts}} 2^{-l(p)} \leq 1$$

Theorem:

Suppose $\{X_i\}$ is an iid sequence with $X \in \mathcal{X}$. Then,

$$\frac{1}{n}\sum_{x_1x_2\cdots\cdots x_n} K(x_1x_2\cdots\cdots x_n \mid n)\Pr(x_1x_2\cdots\cdots x_n) \longrightarrow H(X)$$

For large sequence, the Kolmogorov complexity approaches entropy.

Proof:

$$\begin{aligned}\sum_{x_1x_2\cdots\cdots x_n} K(x_1x_2\cdots\cdots x_n \mid n)\Pr(x_1x_2\cdots\cdots x_n) &\geq H(x_1x_2\cdots\cdots x_n) \\ &= nH(X)\end{aligned}$$

Here $K(x_1x_2\cdots\cdots x_n \mid n)$ is the smallest length for any program and because the programs are prefix free, this is the length of prefix-free codes. Then the LHS of above equation is nothing but the average length of the symbol. And we have $L(C) \geq H(X)$. Thus,

$$\frac{1}{n}\sum_{x_1x_2\cdots\cdots x_n} K(x_1x_2\cdots\cdots x_n \mid n)\Pr(x_1x_2\cdots\cdots x_n) \geq H(X)$$

next we have to prove this is less than $H(X)$.
From property 5, we have

$$\begin{aligned}\frac{1}{n}K(x_1x_2\cdots\cdots x_n \mid n) &\leq H\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) + \frac{1}{n}\log^* n + \frac{c}{n} \\ E\left\{\frac{1}{n}K(x_1x_2\cdots\cdots x_n \mid n)\right\} &\leq E\left[H\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\right] + \frac{1}{n}\log^* n + \frac{c}{n}\end{aligned}$$

$H(X)$ is a concave function, so using Jensen's inequality,

$$\begin{aligned}E\left\{\frac{1}{n}K(x_1x_2\cdots\cdots x_n \mid n)\right\} &\leq H\left[\frac{1}{n}\left(E\sum_{i=1}^{n} x_i\right)\right] + \frac{1}{n}\log^* n + \frac{c}{n} \\ &= H\left[\frac{1}{n}\left(\sum_{i=1}^{n} Ex_i\right)\right] + \frac{1}{n}\log^* n + \frac{c}{n} \\ &= H\left[E(X)\right] + \frac{1}{n}\log^* n + \frac{c}{n}\end{aligned}$$

but

$$\begin{aligned} E(X) &= 1.\Pr(x=1) + 0.\Pr(x=0) \\ &= \Pr(x=1) \end{aligned}$$

then

$$\begin{aligned} E\left\{\frac{1}{n}K(x_1x_2\cdot\cdots\cdot x_n \mid n)\right\} &\leq H\left[\Pr(x=1)\right] + \frac{1}{n}\log^* n + \frac{c}{n} \\ &= H\left[X\right] + \frac{1}{n}\log^* n + \frac{c}{n} \\ &\longrightarrow H(X) \end{aligned}$$

as $n \longrightarrow \infty$, Kolmogorov complexity approaches entropy

Incompressible Sequences

There are certain large numbers that are simple to describe like,

$$2^{2^{2^{2^2}}} \ or \ (100!)$$

But most of such large sequences do not have a simple description. That is, such sequences are incompressible. Given below is the condition for incompressible sequence.

A sequence $X = \{x_1x_2x_3\ldots x_n\}$is incompressible if and only if,

$$\underset{n\to\infty}{lim}\frac{K(x_1x_2x_3\ldots x_n|n)}{n} = 1.$$

Thus, Kolmogorov complexity tells us given a sequence, how much we can compress. (Answering our question posted in the Introduction). That is, if $K(X)$ is of the order of length n then clearly, the sequence is incompressible.

Theorem:

For binary incompressible sequence $X = \{x_1, x_2, x_3, \ldots\ldots, x_n\}$,

$$\frac{1}{n}\sum_{i=1}^{n} x_{_i} \longrightarrow \frac{1}{2}$$

i.e., approximately same # of 1's and 0's or the proportions of 0's and 1's in any incompressible string are almost equal.

Proof:
We have by definition,

$$K(x_1x_2x_3\ldots x_n|n) \quad \geq \quad n - c_n$$

where c_n is some number. Then by Property 5, we have

$$\begin{aligned}
n - c_n &\leq K(x_1 x_2 \cdot \cdots \cdot x_n \mid n) \leq nH\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) + \log^* n + c \\
1 - \frac{c_n}{n} &\leq H\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) + \frac{\log^* n}{n} + \frac{c}{n} \\
H\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) &\geq 1 - \frac{(c_n + c + \log^* n)}{n} \\
&= 1 - \varepsilon_n
\end{aligned}$$

and $\varepsilon_n \longrightarrow 0$ as $n \longrightarrow \infty$,

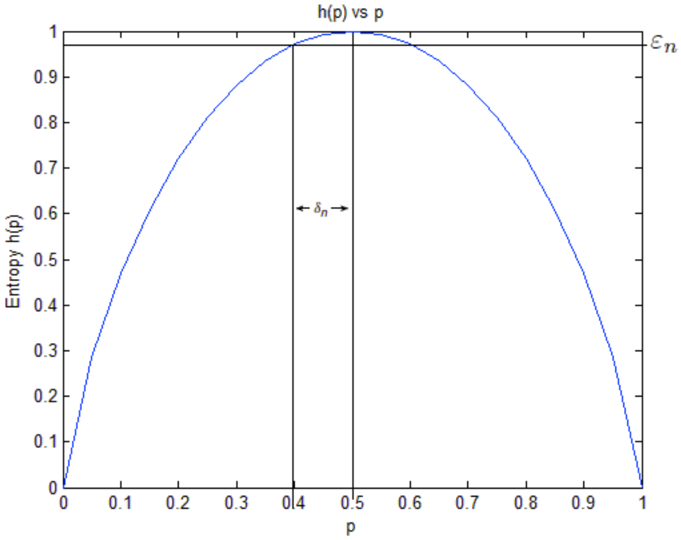

Figure 4: H(p) vs p

By inspecting the above graph, we see that,

$$\frac{1}{n}\sum_{i=1}^{n} x_i \in \left\{\frac{1}{2} - \delta_n, \frac{1}{2} + \delta_n\right\}$$

where δ_n is chosen such that,

$$H\left(\frac{1}{2} - \delta_n\right) = 1 - \varepsilon_n$$

this implies, $\delta_n \longrightarrow 0 \,\text{as}\, n \longrightarrow \infty$ and

$$\frac{1}{n}\sum_{i=1}^{n} x_i \longrightarrow \frac{1}{2}$$

References

[1] Chapter 14, "Elements of Information Theory" 2nd ed - T. Cover, J. Thomas (Wiley, 2006)

[2] http://en.wikipedia.org/wiki/Kolmogorov_complexity