15-859: Information Theory and Applications in TCS **CMU: Spring 2013**

Lecture 6: Source coding, Typicality, and Noisy channels and capacity

January 31, 2013

Lecturer: Mahdi Cheraghchi *Scribe: Tongbo Huang*

1 Recap

Universal Source Coding: Lempel-Ziv Algorithm and its optimality. Please see notes of last lecture.

2 Source Coding by Typicality

2.1 AEP and Typical Set

We have learned AEP in lecture 4, and here is the definition once again:

Theorem 1 (AEP) *If $x_1, x_2, ..., x_n$ are i.i.d $p(x)$, then $-\frac{\log p(x_1,x_2,...,x_n)}{n} \to H(x)$ in probability.*

Now, we can define a *typical set* given the definition of AEP:

Definition 2 (Typical Set) *The typical set $A_\epsilon^{(n)}$ with respect to $p(x)$ is the set of sequences $(x_1, x_2, ..., x_n) \in \chi^{(n)}$ with the property*

$$2^{-nH(X)+\epsilon} \leq p(x_1, x_2, ..., x_n) \leq 2^{-nH(X)-\epsilon} \tag{1}$$

Theorem 3 (Properties of Typical Set) *As a consequence of AEP, we can show that the set $A_\epsilon(n)$ comes with following properties:*

1. *If n is large enough, then $Pr(x_1^n \in A_\epsilon(n)) \geq 1 - \epsilon$.*
2. *$|A_\epsilon(n)| \leq 2^{-nH(X)+\epsilon}$, where $|A|$ denotes the number of elements in the typical set.*

2.2 Data Compression with AEP

We can divide all sequences in χ^n into two sets: those in the typical set and those in its complement. This is illustrated in 2.2: Figure 6.1.

Encode each $x_1^n \in A_\epsilon(n)$ by its index in $A_\epsilon(n)$ (with a leading zero). Also, encode $x_1^n \notin A_\epsilon(n)$ trivially (with a leading one). Such encoding with ensure uniqueness since the decoder knows the distribution and can tell the elements in typical set with the leading bit.

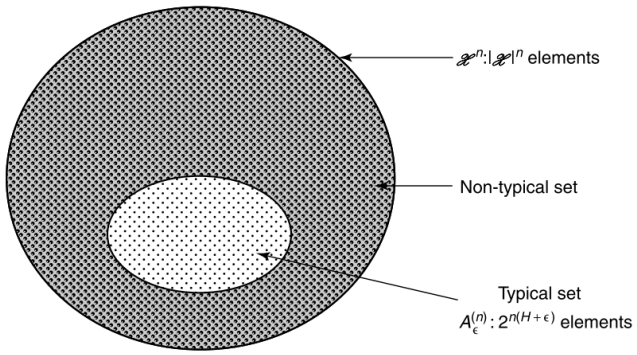

Figure 1: Typical sets and source coding

Now we can calculate the expected length.

$$\begin{aligned} E(l(x_1^n)) &= \Sigma_{x_1^n \in \chi^n} P(x_1^n) * l(x_1^n) \\ &= \Sigma_{x_1^n \in A_\epsilon(n)} P(x_1^n) l(x_1^n) + \Sigma_{x_1^n \in A_\epsilon^{(n)c}} P(x_1^n) l(x_1^n) \\ &\leq \Sigma_{x_1^n \in A_\epsilon(n)} P(x_1^n) l(n(H(x)+\epsilon)+2) + \Sigma_{x_1^n \in A_\epsilon^{(n)c}} P(x_1^n) l(nlog|\chi|+2) \\ &\leq n(H+\epsilon) + \epsilon n log|\chi| + 2 \\ &\leq n(H(x)+\epsilon') \end{aligned} \tag{2}$$

where $\epsilon' = \epsilon + \epsilon log|\chi| + \frac{2}{n}$ can be made arbitrarily small by appropriate choice of ϵ and n.

This can be in terms turned into a universal coding scheme for all source χ with bounded $H(x) \leq h$. (Roughly speaking, this is Csiszar-Körner Universal Source Coding.) Here, we consider only source x $s.t.$ $\forall x\ p(x) = \frac{int}{|X|}$, which is used to limit the number of unique distribution for union bound. Also, we need a bounded initial source, so the typical set also works for small entropy sets.

3 Joint Typicality

Definition 4 (Jointly Typical Set) *The set $A_\epsilon^{(n)}$ of jointly typical sequences (x^n, y^n) with respect to the distribution $p(x,y)$ is the set of n-sequences with empirical entropies $\epsilon -$ close to the true entropies:*

$$\begin{aligned} A_\epsilon^{(n)} &= \{(x^n, y^n) \in X^n \times Y^n : \\ &\quad |-\frac{\log p(x_1^n, y_1^n)}{n} - H(x,y)| \leq \epsilon, \\ &\quad |-\frac{\log p(x_1^n)}{n} - H(x)| \leq \epsilon, \\ &\quad |-\frac{\log p(y_1^n)}{n} - H(y)| \leq \epsilon\} \end{aligned} \tag{3}$$

where $p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i)$.

Theorem 5 (Joint AEP) *From the definition of jointly typical set, we have following properties:*

1. $Pr(x_1^n, y_1^n) \in A_\epsilon^{(n)} \to 1$ *as* $n \to \infty$
2. *For large n,* $(1-\epsilon)2^{n(H(x,y)-\epsilon)} \leq |A_\epsilon^{(n)}| \leq 2^{n(H(x,y))+\epsilon}$
3. *If* $(\tilde{x_1^n}, \tilde{y_1^n}) \sim p(x_1^n)p(y_1^n)$ *independent and has the same marginal distribution, but not joint distribution, then* $(1-\epsilon)2^{-n(I(x;y)+3\epsilon)} \leq Pr(\tilde{x_1^n}, \tilde{y_1^n}) \in A_\epsilon^{(n)} \leq 2^{n(H(x,y))+\epsilon}$.

Proof:

1. According to the weak law of large numbers: $\forall \epsilon > 0$, $\exists n_0$, *s.t.* $\forall n > n_0$, $Pr(|-\frac{\log p(x_1^n)}{n} - H(x)| \geq \epsilon) \leq \frac{\epsilon}{3}$, $Pr(|-\frac{\log p(y_1^n)}{n} - H(y)| \geq \epsilon) \leq \frac{\epsilon}{3}$, $Pr(|-\frac{\log p(x_1^n, y_1^n)}{n} - H(x,y)| \geq \epsilon) \leq \frac{\epsilon}{3}$. Property is easily proved with union bound.
2. The upper bound follow from the one variable AEP.
   To prove the lower bound, we take n large enough so that:
$$\begin{aligned} 1-\epsilon &\leq Pr(A_\epsilon^{(n)}) \\ &= \Sigma_{(x_1^n, y_1^n) \in A_\epsilon^{(n)}} p(x_1^n, y_1^n) \\ &\leq |A_\epsilon^{(n)}| \cdot 2^{-n(H(x,y)-\epsilon)} \end{aligned} \tag{4}$$
3. For property 3,
$$\begin{aligned} Pr(\tilde{x_1^n}, \tilde{y_1^n}) \in A_\epsilon^{(n)} &= \Sigma_{(x_1^n, y_1^n) \in A_\epsilon^{(n)}} p(x_1^n, y_1^n) \\ &= 2^{n(H(x,y)+\epsilon)} \cdot 2^{-n(H(x)-\epsilon)} \cdot 2^{-n(H(y)-\epsilon)} \ (\textit{by property 2}) \\ &= 2^{-n(I(x,y)-3\epsilon)} \end{aligned} \tag{5}$$
   And the proof is similar for lower bound.

■

4 Channel Coding

4.1 Definition

We now prove what is perhaps the basic theorem of information theory, the achievablility of channel capacity, first stated and essentially proved by Shannon in his original 1948 paper. The result is rather counterintuitive; if the channel introduces errors, how can one correct them all? Any correction process is also subject to error, ad infinitum. Shannon used a number of new ideas to prove that information can be sent reliably over a channel at all rates up to the channel capacity.

Let's take a look at how channel work first, please see 4.1: Figure 6.2.

A discrete channel can be viewed as transforming input x to output y through probability $p(y|x)$. Meanwhile, channel can also be viewed as a "probability transformation matrix" as in 4.1: Figure 6.3.

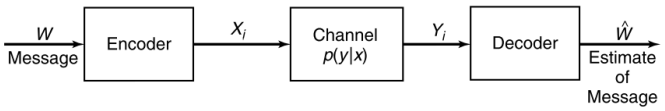

Figure 2: Illustration of channel coding

$$\mathcal{C} = \begin{bmatrix} x_1(1) & x_2(1) & \cdots & x_n(1) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(2^{nR}) & x_2(2^{nR}) & \cdots & x_n(2^{nR}) \end{bmatrix}$$

Figure 3: A probability transformation matrix

4.2 Examples

4.2.1 Noiseless Channel

Input	Output	Probability
1	1	1
0	0	1

Transformation Matrix:

$$p(y|x) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

4.2.2 Noiseless Channel, Unpredictable Output

Input	Output	Probability
0	1	0.5
0	2	0.5
1	3	0.5
1	4	0.5

Transformation Matrix:

$$p(y|x) = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 \end{pmatrix}$$

4.2.3 Binary Symmetric Channel (BSC(p))

Input	Output	Probability
0	0	1-p
0	1	p
1	0	p
1	1	1-p

Transformation Matrix:

$$p(y|x) = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$$

4.2.4 Binary Erasure Channel (BEC(p))

Input	Output	Probability
0	0	1-p
0	e	p
1	e	p
1	1	1-p

Transformation Matrix:

$$p(y|x) = \begin{pmatrix} 1-p & p & 0 \\ 0 & 1-p & p \end{pmatrix}$$

4.3 Channel Capacity

4.4 Definition

Definition 6 *We define channel capacity of a discreet memoryless channel as:*

$$C = \max_{p(x)} I(x;y) \tag{6}$$

Which is taking the best mutual information among all distribution in the transformation matrix.

4.5 Examples

4.5.1 Noiseless Channel

$I(x;y) = H(x) = 1$ for uniform x (lower bound).

4.5.2 Noiseless Channel, Unpredictable Output

$I(x;y) = H(x) - H(x|y) = 1$, since y determines x.

4.5.3 BSC(p)

$$\begin{aligned} I(x;y) &= H(y) - H(y|x) \\ &= H(y) - \Sigma p(x) H(y|X=x) \\ &= H(y) - h(p) \\ &\leq 1 - h(p) \end{aligned} \tag{7}$$

For uniform x, uniform y, $C = 1 - h(p)$.

4.5.4 BEC(p)

$C = \max_{p(x)} H(y) - h(p)$

$$\begin{aligned} H(y) &= H(y, \text{“erasure”}) \\ &= H(\text{“erasure”}) + H(y|\text{“erasure”}) \\ &= h(p) + (1-p)h(p_1) \text{where } p_1 = Pr(x-1), \text{which means no info gain if erasure happens} \\ &\rightarrow C = \max_{p_1} (1-p)h(p_1) = (1-p)(p_1 = 1/2) \end{aligned} \tag{8}$$