Lecture 4 — Jan 19

Lecturer: David Tse *Scribe: Jaydeep S, Dennis Wu, David W, LC Tao, Vivek B.*

4.1 Outline

- Asymptotic Equipartition Property (AEP)
- Markov Chains

4.1.1 Readings

- Shannon [1]: 5,6,7
- CT [2]: 5.1-5.8

4.2 Recap

Relative entropy (or *Kullback–Leibler* divergence) between two distributions p and q is defined as

$$D(p||q) = \mathbb{E}\left[\log \frac{p(X)}{q(X)}\right],$$

where $X \sim p$. It is a **non-negative** quantity and is equal to zero iff $p = q$. Mutual information between a pair of random variables X and Y can be expressed as the relative entropy between the joint distribution p_{XY} and the product distribution $p_X.p_Y$, i.e.,

$$I(X;Y) = D\big(p_{XY}||p_X.p_Y\big).$$

This formulation immediately implies that mutual information is a non-negative quantity.

In the previous lecture, we also considered the problem of compressing a random sequence $\{X_i\}_{i=1}^n \overset{i.i.d}{\sim} Bern(p)$. We "loosely" argued that out of the 2^n possible sequences, there are $2^{nH(X_1)}$ "typical" sequences, each with probability close to $2^{-nH(X_1)}$. Therefore, $\{X_i\}_{i=1}^n$ can be encoded using only $nH(X_1)$ bits. In this lecture, we will make this argument rigorous by using AEP.

4.3 Asymptotic Equipartition Property (AEP)

The asymptotic equiparitition property (AEP) is central in information theory and is the consequence of the *law of large numbers*.

Theorem 1. (AEP) If $X_1, X_2, \cdots, X_n \overset{i.i.d}{\sim} p$, then

$$-\frac{1}{n}\log p(X_1, X_2, \cdots, X_n) \xrightarrow{p} H(X).$$

Proof. Random variables $p(X_i)$ are i.i.d and hence by using the law of large numbers, we have

$$\begin{aligned} -\frac{1}{n}\log p(X_1, X_2, \cdots, X_n) &= -\frac{1}{n}\sum_{i=1}^{n}\log p(X_i) \\ \big(\text{using L.L.N}\big) &\xrightarrow{p} -\mathbb{E}\big[\log p(X)\big] \\ &= H(X) \end{aligned}$$

□

AEP enables us to partition (refer to Figure 4.1) the sequence $\{X_i\}_{i=1}^n \sim p$ into

1. Typical set - Containing sequences with probability close to $2^{-nH(X)}$.
2. Atypical set - Containing the other sequences.

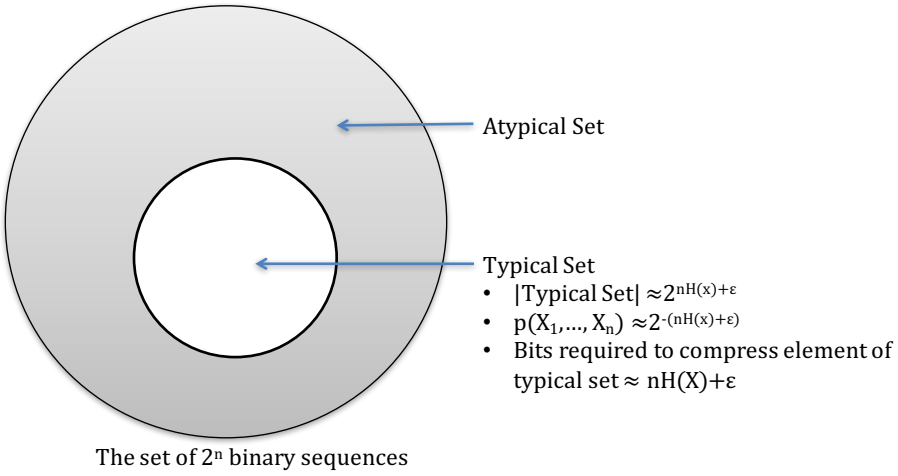

Figure 4.1: Typical set and Aypical set.

As shown in the next section, sequences from the typical set **occur with high probability** and therefore, any property true on the typical set, will be true with high probability. Using this line of argument, we will propose a scheme which acheives a good compression rate on the typical set, thereby achieving a good compression rate with high probability! We will formalize this idea in the next section.

4.4 Typical set

Definition 1. A *typical set* $A_\epsilon^{(n)}$ with respect to a distribution p is the set of sequences $(x_1, x_2, \cdots x_n) \in \mathcal{X}^n$ with probability

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

4.4.1 Properties

1. For a sequence of random variables $\{X_i\}_{i=1}^n \overset{i.i.d}{\sim} p$ and sufficiently large n
$$Pr\big((X_1, \ldots, X_n) \in A_\epsilon^{(n)}\big) \geq 1 - \epsilon.$$
2. $\big|A_\epsilon^{(n)}\big| \leq 2^{n(H(X)+\epsilon)}$.
3. $\big|A_\epsilon^{(n)}\big| \geq 2^{n(H(X)-\epsilon)}$ for a sufficiently large n.

The above properties can be proved by using AEP (1). This is left as an exercise to the reader.

4.4.2 Data compression

Theorem 2. Let $\big\{X_i\big\}_{i=1}^n \overset{i.i.d}{\sim} Bern(p)$ and $\epsilon > 0$. Then there exists an invertible mapping of the sequences $(x_1, \cdots, x_n)$ into binary codewords of length $l(x_1, \cdots, x_n)$ and
$$\mathbb{E}\Big[l(X_1, \cdots X_n)\Big] \leq n\big(H(X) + 2\epsilon\big)$$

Proof. Consider a scheme which encodes the typical set $A_\epsilon^{(n)}$ using codewords of length $n(H(X)+\epsilon)$ [Property 2] and the atypical set using codewords of length n. The expected length of the codeword under this scheme is
$$\begin{aligned} \mathbb{E}\Big[l(X_1, \cdots X_n)\Big] &= n(H(X)+\epsilon)Pr(A_\epsilon^{(n)}) + n\big(1 - Pr(A_\epsilon^{(n)})\big) \\ \big(\text{From Property 1}\big) &\leq n(H(X)+\epsilon) + n\epsilon \\ &\leq n(H(X)+2\epsilon) \end{aligned}$$

Refer to Figure 4.2 for a diagrammatic proof. □

Remark 1. Theorem 2 can be extended for $\big\{X_i\big\}_{i=1}^n \overset{i.i.d}{\sim} p$ (for a general distribution p). This is left as an exercise to the reader.

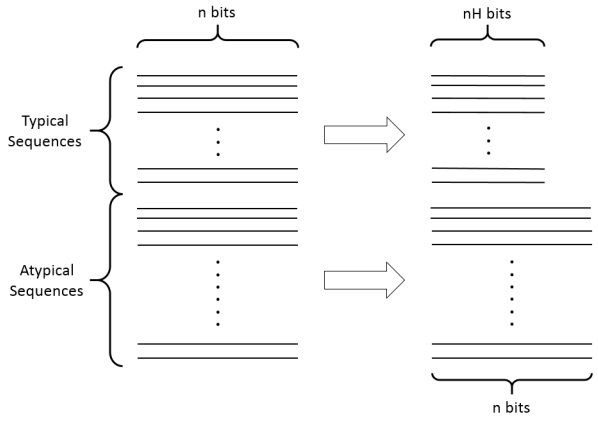

Figure 4.2: Compression scheme for typical and atypical sequences.

Now let us ask the converse - Is the compression rate achieved in Theorem 2 is **optimal**? The answer is **Yes**! We will prove this later in the course.

4.5 Markov chains

4.5.1 Modeling English Text

In 1948 , Shannon [1] considered the problem of modeling and compressing *English text*. A sentence can be modeled as a realization of a sequence of random variables $\{X_i\}_{i=1}^n$, where X_i denotes the i^{th} letter in the sentence. For example, consider a sentence - *'quick brown fox jumped over the lazy dogs'*. Here, the random variable X_1 takes on value *'q'* and the random variable X_2 takes on value *'u'* and so on. We will discuss three models of the random variables X_i's with increasing level of complexity:

- 0^{th} **order Model -** This is the simplest model where random variables are $X_i \overset{i.i.d}{\sim} Unif\{`a', \cdots, `z'\}$. The entropy is $H(X) = \log 26 = 4.7$ bits.
- 1^{st} **order Model -** Here, we consider a non-uniform distribution p on alphabet $\{`a', \cdots, `z'\}$ and $\{X_i\}_{i=1}^n \overset{i.i.d}{\sim} p$. Here, the entropy $H(X) \leq \log 26 = 4.7$ bits. Note that the distribution p can be empirically estimated.

While the previous model captures that certain letters appear more frequently than the others, it fails to capture that certain pairs of letters are more common. For example one might imaging that the sequence *'th'* appears more often than *'tz'*. This feature is captured in the following model.

- 2^{nd} **order Model -** In this model, random variables $X_i's$ have dependence with the previous letters $\{X_j\}_{j<i}$ in the sequence. The dependency with the past observations can be modeled via Markov chains.

4.5.2 A short introduction to Markov chains

A stochastic process is a sequence of random variables $X_1, \dots X_n \sim p$, where X_n is the **state** of the process at time n.

Definition 2. A stochastic process $X_1, \dots X_n \sim p$ is a *Markov process or chain,* if

$$\begin{aligned} & Pr\big(X_{n+m+1} = x_{n+m-1}, \cdots, X_{n+1} = x_{n+1} | X_n = x_n, \cdots X_1 = x_1\big) \\ = \; & Pr(X_{n+m+1} = x_{n+m+1}, \cdots, X_{n+1} = x_{n+1} | X_n = x_n\big) \qquad \forall m > n > 0 \end{aligned} \tag{4.1}$$

In other words, conditioned on the observation $\{X_n = x_n\}$, the states after time n, $\{X_{n+m}, \dots X_n\}$ are independent of the states before time n, $\{X_1, \dots X_{n-1}\}$. This directly implies that the joint distribution of $X_i's$ can be split into product of "simple" conditional distributions,

$$Pr(X_1 = x_1, \dots X_n = x_n) = Pr(X_1 = x_1) Pr(X_2 = x_2 | X_1 = x_1) \dots Pr(X_n = x_n | X_{n-1} = x_{n-1}).$$

Definition 3. Markov chain is *time invariant*, with *transition probability matrix* P if

$$P_{ij} \triangleq Pr(X_2 = i | X_j = j) = Pr(X_{n+1} = i | X_n = j) \qquad \forall n > 0$$

In other words, P_{ij} is the probability of moving from state j to i, in one time step.

A *time invariant* Markov chain is completely characterized by its transition probability matrix and the distribution of the initial state. We will assume that Markov chain is time invariant unless otherwise stated.

4.5.3 Mickey Mouse Markov Chain

We consider an example of two state Markov chain in order to understand basic properties of Markov chains.

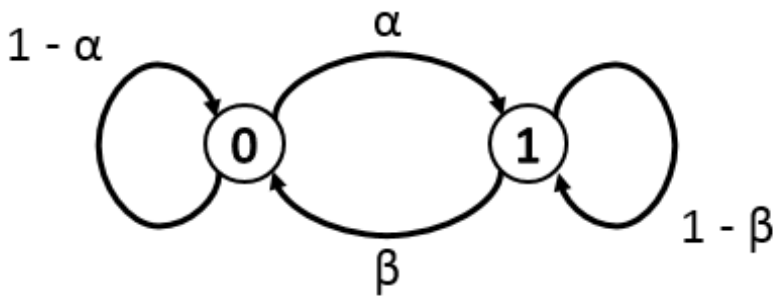

Figure 4.3: Mickey Mouse Markov Chain

Definition 4. (*Mickey Mouse Markov Chain [MMMC]*) Let $X_1 \to X_2 \ldots \to X_n$ be Markov chain (Figure 4.3) with two states - 0 and 1, and with transition probability matrix

$$P = \begin{bmatrix} 1-\alpha & \beta \\ \alpha & 1-\beta \end{bmatrix}.$$

Characterizing MMMC

We will characterize the MMMC in terms of the probability transition matrix and the initial state. Let the initial state be represented by a column vector $\vec{\pi}_1 \triangleq \begin{bmatrix} p(X_1 = 0) \; p(X_1 = 1) \end{bmatrix}^T$. Using the rule of total probability, the state at time $n = 2$ is,

$$\begin{aligned} \vec{\pi}_2[0] &\triangleq Pr(X_2 = 0) = Pr(X_2 = 0 | X_1 = 0)\vec{\pi}_1[0] + Pr(X_2 = 0 | X_1 = 1)\vec{\pi}_1[1] \\ \vec{\pi}_2[1] &\triangleq Pr(X_2 = 1) = Pr(X_2 = 1 | X_1 = 0)\vec{\pi}_1[0] + Pr(X_2 = 1 | X_1 = 1)\vec{\pi}_1[1] \end{aligned} \tag{4.2}$$

Equations (4.2) can be expressed via transition probability matrix,

$$\vec{\pi}_2 = \begin{bmatrix} 1-\alpha & \beta \\ \alpha & 1-\beta \end{bmatrix} \vec{\pi}_1 = P\vec{\pi}_1$$

Extending this to the state at time n,

$$\vec{\pi}_n = \begin{bmatrix} 1-\alpha & \beta \\ \alpha & 1-\beta \end{bmatrix}^{n-1} \vec{\pi}_1 = P^{n-1}\vec{\pi}_1$$

Observation: For $\alpha, \beta < 1$, $\lim_{n\to\infty} \pi_n = [\frac{\beta}{\alpha+\beta} \quad \frac{\alpha}{\alpha+\beta}]$. We make this idea rigorous in the following section.

4.5.4 Stationary distribution

For Markov chain, the transition probability matrix P is a row-stochastic matrix (entries in every row, sum up to 1). From results in linear algebra, P has at least one eigenvalue equal to 1 (other eigenvalues have magnitude < 1).

Definition 5. *Stationary distribution* of a transition probability matrix P is the eigenvector π, associated with eigenvalue 1. Mathematically, $\pi = P\pi$.

For the sake of simplicity, we will restrict ourselves to *irreducible* and *aperiodic* markov chains (refer CT [2] for definition). These Markov chains have a unique eigenvector associated to eigenvalue 1.

Property 2. *A irreducible* and *aperiodic* Markov chain has a unique stationary distribution π. As a consequence, for all the initial vector $\vec{\pi}_1$, Markov chain converges to the unique vector i.e, $\lim_{n\to\infty} \vec{\pi}_n = P^{n-1}\vec{\pi}_1 \longrightarrow \pi$

4.5.5 Entropy of stationary Markov chains

The entropy of Markov chain $\{X_i\}_{i=1}^n$ is

$$\begin{aligned} H(X_1, \dots X_n) &= H(X_1) + H(X_2|X_1) + \dots + H(X_n|X_{n-1}, \dots, X_1) \\ &\overset{(a)}{=} H(X_1) + H(X_2|X_1) + H(X_3|X_2) + \dots + H(X_n|X_{n-1}) \\ &\overset{(b)}{=} H(X_1) + (n-1)H(X_2|X_1), \end{aligned}$$

where (a) follows from the definition of Markov chain and (b) follows from the time invariant property of Markov chain. Therefore,

$$\lim_{n\to\infty} \frac{1}{n} H(X_1, \dots X_n) = H(X_2|X_1) \tag{4.3}$$

In the next lecture will prove a stronger result than (4.3) which is

$$\lim_{n\to\infty} -\frac{1}{n} \log p(X_1, X_2, \cdots, X_n) \xrightarrow{p} H(X_2|X_1)$$

Bibliography

[1] Shannon, Claude. "A mathematical theory of communication" ACM SIGMOBILE Mobile Computing and Communications Review 5.1 (2001): 3-55.

[2] Cover, Thomas M., and Joy A. Thomas. Elements of information theory. John Wiley & Sons, 2012.