(Classical) Information Theory II: Source coding

Sibasish Ghosh

The Institute of Mathematical Sciences
CIT Campus, Taramani, Chennai 600 113, India.

Abstract

The information content of a random variable X, with associated probability mass density $p_x \equiv \mathrm{Prob}(X = x)$, is described by Shannon as the entropy $H(X)$ bits. In the present lecture, we will see that in order to store a string of values of X of arbitrarily large length in the most economic way, one would require, on an average, a bit string of length $H(X)$ for storing the value of X.

Asymptotic equipartition property guarantees that, on an average, at least $nH(X)$-bit string is a must in order to reliably store the information about an arbitrary sequence (of length n) of values of X, n being very large. For achievability of this limit, one should look for proper encoding of the strings – Huffman codes provide such a solution.

Outline

- **Thermodynamic entropy versus information theoretic entropy**
- **Asymptotic equipartition property**
- **Data compression and encoding**

Thermodynamic versus information theoretic entropies

Entropy in statistical thermodynamics: If p_i is the probability of the i-th state of a thermodynamic system (e.g., p_i may be the probability that in an idael gas of several colliding molecules, some of the gas particles are having velocity v_i at an instant of observing the motion of the gas molecules). The thermodynamic entropy is then given by: $S = k \times \sum_i p_i \times \log_e p_i$, k being Boltzman constant and the logarithm is the natural one.

Microcanonical ensemble: For any microcanonical ensemble, all the microstates are taken to be equally probable. So $S = k \times \log_e n$, where n is the total no. of microstates.

Thermodynamic versus ....

2nd law of thermodynamics: Entropy of an isolated thermodynamic system is non-decreasing.

- **Question: Does the same feature happen for information theoretic entropy?**
- **Modelling isolated systems: For this purpose, we need to model an isolated system from the perspective of information theory. For any isolated system, the state of the system at some time t depends solely on the governing dynamical laws and the state of the system immediately before the time t. This indicates to a *Markov chain* type evolution of the system with the transition probabilities being provided by the governing physical laws.**

Thermodynamic versus ....

- **Stochastic process:** $\{X_\alpha : \alpha \in \Lambda\}$ **is a collection of random variables each having the same value set,** Λ **being either a discrete or a continuous index set, and where the joint probability mass density** $\mathrm{Prob}(X_{\alpha_1} = x_{\alpha_1}, X_{\alpha_2} = x_{\alpha_2}, \ldots, X_{\alpha_n} = x_{\alpha_n})$ **always exits for any choice of** n **indices** $\alpha_1, \alpha_2, \ldots, \alpha_n$ **from** Λ **with** $n = 1, 2, \ldots$**.**
- **Discrete stochastic process: A stochastic process for which** Λ **is discrete. In this case, we can take** $\Lambda = \{1, 2, \ldots\}$**.**
- **Markov chain: It is a discrete stochastic process for which** $\mathrm{Prob}(X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1) = \mathrm{Prob}(X_{n+1} = x_{n+1} | X_n = x_n)$ **for all** $x_{n+1}, x_n, \ldots, x_1$ **and for all** n**. So** $p_{x_1, x_2, \ldots, x_n} = p_{x_1} p_{x_2|x_1} p_{x_3|x_2} \cdots p_{x_n|x_{n-1}}$**.**

Thermodynamic versus ....

- **Time invariance, states, stationary distribution, etc.:**

(i) A Markov chain $\{X_1, X_2, \ldots\}$ **is time-invariant iff**

$\text{Prob}(X_{n+1} = a | X_n = b) = \text{Prob}(X_2 = a | X_1 = b)$ **for all** n, a, b**.**

(ii) X_n**: state of the chain at time** n**.**

(iii) For all $n \geq 2$**, stationary distribution:**

$\text{Prob}(X_{n+1} = a) = \text{Prob}(X_n = a)$ **for all** a**.**

(iv) Transition probability matrix:

$P_{ij} = \text{Prob}(X_{n+1} = j | X_n = i)$**.**

(v) Thus $p_{x_{n+1}} = \sum_{x_n} p_{x_n} \times P_{x_n x_{n+1}}$**. So, for the same Markov chain,** p_{x_n} **will be different if** p_{x_1}**'s are different.**

(vi) Call the prob. mass density p_{x_n} **as** p_n**.**

Thermodynamic versus ....

- **Relative entropy decreases: For two different probability mass functions p_n and p'_n of the same Markov chain, $D(p_{n+1}||p'_{n+1}) \leq D(p_n||p'_n)$ for all n. (Use chain rule of relative entropy)**

- **Entropy decreases for non-uniform stationary distribution: Intial probability mass density p_1 be uniform. So $H(X_1)$ is maximum. Under the time evolution this initial uniform prob. mass density will tend to a stationary distribution p_n for all $n \geq 2$, and hence $H(X_n)$ can not be maximum, as p_n is non-uniform.**

Thermodynamic versus ....

- **Entropy increases for uniform stationary distribution:** p'_n**: uniform stationary distribution,** p_n**: any other distribution at time** n**. Then** $D(p_n||p'_n) = \log_2 d - H(p_n) = \log_2 d - H(X_n)$**,** d**: size of the value set of the chain. As** $D(p_n||p'_n)$ **decreases with** n**,** $H(X_n)$ **will increase.**

- **Above scenario is somewhat closed to the case of microcaconical ensemble in statistical thermodynamics.**

Asymptotic equipartition property

String of random variables: **The height h_i of a flight at time t_i, while moving from one place A to another place B, will not be far apart from its height at time t_{i-1} if $t_i - t_{i-1}$ is small enough. But during an entire year, the height X_i may vary within the interval $[h_i^{min}, h_i^{max}]$ for each i with associated probability $\mathrm{Prob}(X_i = h_i) \equiv p_i(h_i)$. So the random variables X_i, X_2, $\ldots$ are not independent. We need to know the joint probabilities $\mathrm{Prob}(X_1 = h_1, X_2 = h_2, \ldots, X_n = h_n)$ to get the information content about the heights of the flight.**

Asymptotic equipartition ....

i.i.d. random variables: L **random variables** $X_1, X_2, \ldots, X_L$**, having same value set, are i.i.d. iff (i)** $\text{Prob}(X_1 = a) = \text{Prob}(X_2 = a) = \ldots = \text{Prob}(X_L = a)$ **(**$\equiv \text{Prob}(X = a)$**, say) for all** a **and** $\text{Prob}(X_1 = a_1, X_2 = a_2, \ldots, X_L = a_L) = \text{Prob}(X_1 = a_1) \times \text{Prob}(X_2 = a_2) \times \ldots \times \text{Prob}(X_L = a_L)$ **for all** $a_1, a_2, \ldots, a_L$**.**

Letters, alphabet, messages: X_1**,** X_2**,** $\ldots$**,** X_L **be i.i.d. random variables distributed as** X**. For any random variable** X**, each of its values** x_1**,** x_2**,** $\ldots$**,** x_n **is called a** *letter*; **the set** $\{x_1, x_2, \ldots, x_n\}$ **is called** *alphabet*; **any string** $x_{i_1} x_{i_2} \ldots x_{i_L}$ **of length** L **is called a** *message* **where** L **can be any positive integer.**

$\text{Prob}(X^L = x_{i_1} x_{i_2} \ldots x_{i_L}) = p(x_{i_1}) p(x_{i_2}) \ldots p(x_{i_L})$**.**

Asymptotic equipartition ....

Weak law of large numbers: **For all n, if $X_1, X_2, \ldots, X_n$ are i.i.d. random variables distributed as X and if $E(X) < \infty$, then for any given $\epsilon > 0$: $\mathrm{Prob}(|(1/n)\sum_{i=1}^{n} X_i - E(X)| > \epsilon)$ tends to 0 as n goes to ∞ – 'convergence in probability'.**

Asymptotic equipartition property: **If $X_1, X_2, \ldots$ are i.i.d. random variables distributed as X (with prob. mass density $p(x) \equiv \mathrm{Prob}(X = x)$ and joint prob. mass density $p(x_1, x_2, \ldots) = p(x_1)p(x_2)\ldots$) then the random variable $-(1/n) \times \log_2 p(X_1, X_2, \ldots) \rightarrow H(X)$ in probability for finite $H(X)$.**

Proof: **$\log_2 p(X_1)$, $\log_2 p(X_2), \ldots$ are i.i.d. random variables distributed as $\log_2 p(X)$ with $E(\log_2 p(X)) = \sum_x p(x) \log_2 p(x) = -H(X)$. So, by WLLN, $-(1/n) \times \log_2 p(X_1, X_2, \ldots) \rightarrow H(X)$ in probability.**

Asymptotic equipartition ....

Typical set and sequence: **Let $\mathcal{H}$ be the value set of the random variable X with prob. mass density $p(x)$. Given any $\epsilon > 0$ and n, the *typical set* w.r.t. $p(x)$ is:** $A(n,\epsilon) \equiv \{(x_1, x_2, \ldots, x_n) \in \mathcal{H}^n | 2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)}\}$**. Each element of $A(n,\epsilon)$ is called as a *typical sequence*.**

Asymptotic equipartition ....

Property (1): $(x_1, x_2, \ldots, x_n) \in A(n, \epsilon)$ **implies that** $H(X) - \epsilon \leq -(1/n)\log_2 p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon$**.**

Property (2): $\mathrm{Prob}(A(n, \epsilon)) > 1 - \epsilon$ **for sufficiently large** n**.**

Property (3): Total no. of elements in $A(n, \epsilon) \leq 2^{n(H(X)+\epsilon)}$**.**

Property (4): Total no. of elements in $A(n, \epsilon) \geq (1 - \epsilon) \times 2^{n(H(X)-\epsilon)}$ **for sufficiently large** n**.**

Proof: Follows from the definition of $A(n, \epsilon)$ **and AEP.**

Data compression and encoding

- **Aim:** $X_1, X_2, \ldots, X_n$ **be i.i.d. random variables distributed according to** X **with prob. mass density** $p(x)$**,** $\mathcal{H}$ **being the value set. We want to find out shortest description for the sequences** $(x_1, x_2, \ldots, x_n)$**.**
- **How to do that?: (1) Devide** $\mathcal{H}^n$ **into typical and atypical sets** $A(n, \epsilon), (A(n, \epsilon))^c$ **respectively.**

(2) Order the elements in each of these two sets according to lexicographic order. We can then represent each element of these sets by the corresponding index of the ordering. By property (3), this indexing requires not more than $n(H(X) + \epsilon) + 1$ **bits for** $A(n, \epsilon)$**. We prefix all these bit strings by a** 0**, giving a total length:** $n(H(X) + \epsilon) + 2$ **bits.**

Data compression ....

(3) Similarly, encode the indices of the ordered set $(A(n,\epsilon))^c$ by means of strings of $\log_2|\mathcal{H}|^n + 1$ bits. Prefix all these bit strings by an 1, giving a total length: $n\log_2|\mathcal{H}| + 2$. Thus encoding of $\mathcal{H}^n$ is completed now.

(3) Let $x^n \equiv (x_1, x_2, \ldots, x_n)$. $l(x^n)$: length of the codeword. Then, for sufficiently large n: $E(l(X^n)) = \sum_{x^n} p(x^n)l(x^n) = \sum_{x^n \in A(n,\epsilon)} p(x^n)l(x^n) + \sum_{x^n \in (A(n,\epsilon))^c} p(x^n)l(x^n) \leq \sum_{x^n \in A(n,\epsilon)} p(x^n)[n(H(X)+\epsilon)+2] + \sum_{x^n \in (A(n,\epsilon))^c} p(x^n)[n\log_2|\mathcal{H}|+2] = \mathrm{Prob}(A(n,\epsilon))[n(H(X)+\epsilon)+2] + \mathrm{Prob}((A(n,\epsilon))^c)[n\log_2|\mathcal{H}|+2] \leq [n(H(X)+\epsilon)+2] + \epsilon[n\log_2|\mathcal{H}|+2] = n(H(X)+\epsilon+\epsilon\log_2|\mathcal{H}|+2/n) \rightarrow nH(X)$.

Data compression ....

- **Thus, for sufficiently large n, we can store the information about all the strings $(x_1, x_2, \ldots, x_n)$ by represrenting them using $nH(X)$ bits, on an average.**
- **Is it optimum?**
- **Shannon has shown, using Huffman codes (which involves block coding), that this is optimum.**

Variable vs. fixed length coding

- **Eight 'letters' 1, 2, 3, ..., 8 are produced by a source with respective probabilities 1/2, 1/4, 1/8, 1/16, 1/64, 1/64, 1/64, 1/64.**

Fixed lenght coding: $C(1) = 000$, $C(2) = 001$, $C(3) = 010$, $C(4) = 011$, $C(5) = 100$, $C(6) = 101$, $C(7) = 110$, $C(8) = 111$. **Average no. of bits per letter is 3.**

Variable lenght coding: $C(1) = 0$, $C(2) = 10$, $C(3) = 110$, $C(4) = 1110$, $C(5) = 111100$, $C(6) = 111101$, $C(7) = 111110$, $C(8) = 111111$. **Average no. of bits per letter is** $2 = H(X)$.

Information about the next lecture

- **In the next lecture, we will discuss about Shannon's noisy channel coding theorem, according to which, one can send any large string of values of a random variable through a noisy chhanel, almost in an error-free manner at a maximum rate of $I(X;Y)$ bits per single use of the channel, X and Y being respectively the input and output random variables of the channel and $I(X;Y)$ is their mutual information.**