Asymptotic equipartition property

P1 For any nonnegative random variable X and $t > 0$ show that:

$$P(X \geq t) \leq \frac{E[X]}{t}.$$ (Markov's inequality)

Sol: From the definition of expectation:

$$E(X) = \sum_{x} x \cdot P(X=x) \geq \sum_{x \geq t} x \cdot P(X=x),$$ because

X is nonnegative and so if $x<0$ then $P(X=x)=0$, so for all x we have $x \cdot P(X=x) \geq 0$, and hence we just discarded some nonnegative terms from the sum. Now if $x \geq t$ then $x P(X=x) \geq t \cdot P(X=x)$, so:

$$\sum_{x \geq t} x \cdot P(X=x) \geq \sum_{x \geq t} t \cdot P(X=x) = t \cdot \sum_{x \geq t} P(X=x) = t \cdot P(X \geq t),$$

because $$\sum_{x \geq t} P(X=x) \underset{\substack{\text{mutually}\\\text{exclusive}\\\text{events}}}{=} P\Big(\bigcup_{x \geq t} X=x\Big) = P(X \geq t).$$

Hence $E(X) \geq t \cdot P(X \geq t)$, so $$P(X \geq t) \leq \frac{E(X)}{t}.$$

P2 Let Y be a random variable with expected value μ and variance σ^2 (variance is the

expected value of $(Y-\mu)^2$ i.e. it is $E[(Y-\mu)^2]$).
Show that for any $\varepsilon > 0$

$$P(|Y-\mu| > \varepsilon) \le \frac{\sigma^2}{\varepsilon^2}. \quad \left(\text{Chebyshev's inequality}\right)$$

Sol: $|Y-\mu| > \varepsilon$ iff $(Y-\mu)^2 > \varepsilon^2$, and so setting X to be $(Y-\mu)^2$, and $t = \varepsilon^2$ in Markov's inequality we have:

$$P(|Y-\mu|>\varepsilon) = P((Y-\mu)^2 > \varepsilon^2) = P(X > t) \le P(X \ge t);$$

$$P(X \ge t) \underset{\text{Markov's inequality}}{\le} \frac{E[X]}{t} = \frac{E[(Y-\mu)^2]}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2}.$$

P3 (Weak law of large numbers) Let $Z_1, Z_2, \dots, Z_n$ be a sequence of independent identically distributed random variables with mean μ and variance σ^2.
Let $M_n = \frac{1}{n} \cdot \sum_{i=1}^{n} Z_i$ be the sample mean.
Show that:

$$P(|M_n - \mu| > \varepsilon) \le \frac{\sigma^2}{n \cdot \varepsilon^2}.$$

Sol: We want to use Chebyshev's inequality to prove this, hence we need to find $E[M_n]$ and the variance of M_n.

$$E[M_n] = \sum_{\omega \in \Omega} P(\omega) \cdot M_n(\omega) = \sum_{\omega \in \Omega} P(\omega) \cdot \left(\frac{1}{n} \sum_{i=1}^{n} Z_i(\omega)\right) =$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{\sigma \in \Omega} P(\sigma) \cdot Z_i(\sigma) \right) = \frac{1}{n} \cdot \sum_{i=1}^{n} E[Z_i] = \frac{1}{\cancel{n}} \cdot \cancel{n} \cdot \mu = \mu .$$

And:

$$E((M_n - \mu)^2) = E\left(\frac{1}{n^2} \left(\sum_{i=1}^{n} (Z_i - \mu) \right)^2 \right) = \frac{1}{n^2} \sum_{i=1}^{n} E\left((Z_i - \mu)^2 \right)$$

$$+ \sum_{i \neq j} E\left((Z_i - \mu)(Z_j - \mu) \right).$$

Now, if X and Y are independent then $E(X \cdot Y) = E(X) \cdot E(Y)$, so

$$E((Z_i - \mu) \cdot (Z_j - \mu)) = E(Z_i - \mu) \cdot E(Z_j - \mu) = 0,$$

since Z_i's are independent and since μ is the mean of Z_i's, hence $E(Z_i - \mu) = 0$. Using the fact that $E((Z_i - \mu)^2) = \sigma^2$ for all i's we get:

$$E((M_n - \mu)^2) = \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}.$$

This means that the expectation of M_n is μ, and its variance is $\frac{\sigma^2}{n}$, so when we plug this in Chebyshev's inequality we get:

$$P\left(\left| \frac{1}{n} \sum_{i=1}^{n} Z_i - \mu \right| > \varepsilon \right) = P(|M_n - \mu| > \varepsilon) \leq \frac{\sigma^2}{n \cdot \varepsilon^2}.$$

P4 A cake is being sliced and the largest piece is chosen each time and the other discarded. We will assume that a random cut creates pieces of proportions:

$$R = \begin{cases} \left(\frac{2}{3}, \frac{1}{3}\right) & \text{with prob. } \frac{3}{4} \\ \left(\frac{2}{5}, \frac{3}{5}\right) & \text{with prob. } \frac{1}{4} \end{cases}$$

Approximately, how large is the piece of cake after n cuts (assuming that the slices are independent)?

Sol: Let Y_i be the random variable representing of the chosen piece on the i-th cut, i.e. Y_i takes value $\frac{2}{3}$ with prob. $\frac{3}{4}$ and $\frac{3}{5}$ with prob. $\frac{1}{4}$. Then the product:

$$\prod_{i=1}^{n} Y_i$$

is the random variable representing the size of cake after n cuts. Let X_i be the random variable $\log_2 Y_i$. The expected value of X_i is:

$$E[X_i] = \frac{3}{4} \cdot \log_2 \frac{2}{3} + \frac{1}{4} \log_2 \frac{3}{5} = \frac{3}{4} - \frac{1}{2}\log_2 3 - \frac{1}{4}\log_2 5 = \mu$$

From the weak law of large numbers, we know that for large enough n,

$$\frac{1}{n} \sum_{i=1}^{n} X_i$$

will be approximately μ. But:

$$\frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{n} \sum_{i=1}^{n} \log_2 Y_i = \frac{1}{n} \log_2 \prod_{i=1}^{n} Y_i, \text{ hence}$$

$\prod_{i=1}^{n} Y_i$ will approximately be equal to $2^{n \cdot \mu}$, i.e.

after n cuts the size (the fraction actually) of the piece will approximately be $2^{n\cdot\left(\frac{3}{4}-\frac{1}{2}\log_2 3-\frac{1}{4}\log_2 5\right)} = 2^{(-0.623)\cdot n} = (0.649)^n$.

P5 A discrete memoryless source emits a sequence of binary digits with probabilities:

$$P(1) = 0.005 \quad \text{and} \quad P(0) = 0.995 .$$

The digits are taken 100 at a time and a binary codeword is provided for every sequence containing three or fewer ones.

a) Ignoring the probabilities, find the minimum length required to provide codewords for all sequences with ≤ 3 ones.

b) Use Chebyshev's inequality to estimate the probability of observing a sequence for which no codeword has been assigned.

c) Calculate the exact probability of observing a sequence for which no codeword has been assigned. Compare the result with the estimate obtained in b).

Sol: a) We need to count the number of sequences with ≤ 3 ones. We have 1 with 0 ones, 100 with one 1, $\binom{100}{2} = 4950$ with two ones, and $\underbrace{\binom{100}{3}}_{=161700}$ with 3, so we have

$$1 + 100 + 4950 + 161700 = 166751$$

sequences with ≤ 3 ones, so we need $\lceil \log_2 166751 \rceil = 18$ bits to provide codewords for sequences with ≤ 3 ones.

b) Let X_i be the random variable representing the i-th emitted bit. We are interested in the random variable $Y = \sum_{i=1}^{100} X_i$. To apply Chebyshev's ineq. we need the expectation and variance of Y.

$$E(Y) = E\left(\sum_{i=1}^{100} X_i\right) = \sum_{i=1}^{100} E(X_i) = 100 \cdot 0.005 = 0.5$$

$$Var(Y) = Var\left(\sum_{i=1}^{100} X_i\right) = (\text{because } X_i\text{'s are independent}) = \sum_{i=1}^{100} Var(X_i)$$

$$= \sum_{i=1}^{100}\left(E(X_i^2) - E(X_i)^2\right) = 100 \cdot 0.005 \cdot 0.995$$

The set representing 4 or more ones is $|Y - 0.5| \geq 3.5$, so from Chebysev's inequality:

$$P(|Y - 0.5| \geq 3.5) \leq \frac{100 \cdot 0.005 \cdot 0.995}{(3.5)^2} = 0.04$$

c) The exact probability of observing a sequence with ≤ 3 ones is:

$$\binom{100}{0}(0.005)^0(0.995)^{100} + \binom{100}{1} 0.005 \cdot (0.995)^{99} + \binom{100}{2}(0.005)^2(0.995)^{98}$$

$$+ \binom{100}{3}(0.005)^3 \cdot (0.995)^{97} = 0.99833$$

So the exact prob. of observing a sequence with no codeword is $1 - 0.99833 = 0.00167$. (Note that it is less than 0.04 as predicted by Chebyshev's inequality, but the gap between the two is also substantial.)

Asymptotic equipartition property:

A.E.P. If $X_1, X_2, \ldots$ are independent identically distributed random variables, then

$$\frac{1}{n}\log_2 \frac{1}{P(X_1,\ldots,X_n)} \longrightarrow H(X_1) \quad \text{in probability}$$

[which means that for every $\varepsilon > 0$ and $\varepsilon' > 0$ there is a natural number N, such that $\forall n > N$ we have

$$P\left(\left|\frac{1}{n}\log_2 \frac{1}{P(X_1,\ldots,X_n)} - H(X_1)\right| > \varepsilon\right) < \varepsilon' \quad]$$

The **typical set** $A_\varepsilon^{(n)}$ is the set of all sequences $(x_1,\ldots,x_n)$ with the property:

$$2^{-n(H(X_1)+\varepsilon)} \leq P(x_1,\ldots,x_n) \leq 2^{-n(H(X_1)-\varepsilon)}. \quad (T)$$

(Since X_i's are identically distributed $H(X_1) = H(X_i)$ for all i, so instead of $H(X_1)$ we can take any $H(X_i)$, hence we often omit the index and just write $H(X)$)

P6 Prove the following properties of typical sets:

I) If $(x_1,\ldots,x_n) \in A_\varepsilon^{(n)}$ then:

$$H(X) - \varepsilon \leq -\frac{1}{n}\log P(x_1,\ldots,x_n) \leq H(X) + \varepsilon$$

II) $P(A_\varepsilon^{(n)}) > 1 - \varepsilon$ for n sufficiently large.

III) $|A_\varepsilon^{(n)}| \leq 2^{n(H(X)+\varepsilon)}$ ($|\,|$ for a set is the number of elements in it)

IV) $|A_\varepsilon^{(n)}| \geq (1-\varepsilon)\cdot 2^{n(H(X)-\varepsilon)}$ for n sufficiently large.

Sol: I) Follows from the definition (T) by taking $\log_2$ and multiplying with $-\frac{1}{n}$.

II) If in the statement of asymptotic equipartition property (A.E.P.) we set $\varepsilon = \varepsilon'$, then there is some $N_\varepsilon \in \mathbb{N}$, such that for all $n \geq N_\varepsilon$ we have:

$$P\left(\left|-\frac{1}{n}\log_2 P(X_1,\dots,X_n) - H(X)\right| > \varepsilon\right) < \varepsilon, \text{ and so}$$

$$P(A_\varepsilon^{(n)}) = 1 - P\left(\left|-\frac{1}{n}\log_2 P(X_1,\dots,X_n) - H(X)\right| > \varepsilon\right)$$
$$\geq 1-\varepsilon \quad \text{for all } n \geq N_\varepsilon.$$

III) If $(x_1,\dots x_n) \in A_\varepsilon^{(n)}$, then from (T) $P(x_1,\dots,x_n) \geq 2^{-n(H+\varepsilon)}$, hence:

$$2^{-n(H+\varepsilon)}\cdot |A_\varepsilon^{(n)}| \leq \sum_{(x_1\dots x_n)\in A_\varepsilon^{(n)}} P(x_1,\dots,x_n) \leq P(\Omega) = 1, \text{ so}$$

multiplying with $2^{n(H+\varepsilon)}$ we get: $|A_\varepsilon^{(n)}| \leq 2^{n(H+\varepsilon)}$.

IV) Assume that n is larger than N_ε from part II) of the problem. Then $P(A_\varepsilon^{(n)}) \geq 1-\varepsilon$. On the other hand from the definition (T) we have that for $(x_1,\dots,x_n)\in A_\varepsilon^{(n)}$:

$P(x_1,\dots,x_n) \leq 2^{-n(H-\varepsilon)}$, hence:

$$2^{-n(H-\varepsilon)}\cdot |A_\varepsilon^{(n)}| \geq \sum_{(x_1\dots x_n)\in A_\varepsilon^{(n)}} P(x_1,\dots,x_n) = P(A_\varepsilon^{(n)}) \geq 1-\varepsilon, \text{ and}$$

multiplying with $2^{n(H-\varepsilon)}$ we get $|A_\varepsilon^{(n)}| \geq (1-\varepsilon)2^{n(H-\varepsilon)}$.

P7 For each $n = 1, 2, \dots$ let $B_\delta^{(n)} \subseteq A^n$ (so $B_\delta^{(n)}$ is a subset of the set of all sequences of length n of symbols in our alphabet) be the smallest set with $P(B_\delta^{(n)}) \geq 1 - \delta$. Show that for any $\delta < \frac{1}{2}$ and $\varepsilon > 0$ we have:

$$\frac{1}{n} \log |B_\delta^{(n)}| > H - \varepsilon \quad \text{for } n \text{ sufficiently large.}$$

Sol:

From P6 we know that for any $\varepsilon' > 0$, if we take n large enough we will have:

$$P(A_{\varepsilon'}^{(n)}) > 1 - \varepsilon'.$$

Now let's look at the set $A_{\varepsilon'}^{(n)} \cap B_\delta^{(n)}$. We have:

$$P(A_{\varepsilon'}^{(n)} \cap B_\delta^{(n)}) = P(A_{\varepsilon'}^{(n)}) + P(B_\delta^{(n)}) - P(A_{\varepsilon'}^{(n)} \cup B_\delta^{(n)}) \geq$$

$$\geq (1 - \varepsilon') + (1 - \delta) - 1 = 1 - \varepsilon' - \delta \geq 1 - \Delta \quad \left(\text{where we set } \varepsilon' + \delta = \Delta\right)$$

For elements in $A_{\varepsilon'}^{(n)}$ we also know that their prob. is less than $2^{-n(H-\varepsilon')}$, so:

$$2^{-n(H-\varepsilon')} \cdot |A_{\varepsilon'}^{(n)} \cap B_\delta^{(n)}| \geq P(A_{\varepsilon'}^{(n)} \cap B_\delta^{(n)}) \geq 1 - \Delta, \text{ so}$$

$$|B_\delta^{(n)}| \geq |A_{\varepsilon'}^{(n)} \cap B_\delta^{(n)}| \geq (1 - \Delta) 2^{n(H-\varepsilon')}, \text{ so}$$

$$\frac{1}{n} \log_2 |B_\delta^{(n)}| \geq H - \left(\varepsilon' - \frac{\log(1-\Delta)}{n}\right).$$ If we take ε' small and n large enough such that $\varepsilon' - \frac{\log(1-\Delta)}{n} < \varepsilon$, then we have that $\frac{1}{n} \log_2 |B_\delta^{(n)}| > H - \varepsilon$.

Typical Set Problem Solution

Problem Statement

Consider a sequence of i.i.d. binary random variables $X_1, X_2, \ldots, X_n$, where

$$P(X_i = 1) = 0.6, \quad P(X_i = 0) = 0.4.$$

(a) Calculate $H(X)$.

(b) With $n = 10$ and $\epsilon = 0.1$, determine which sequences fall in the typical set $A_\epsilon^{(n)}$. What is the probability of the typical set? How many elements are in the typical set?

Solution

(a) Entropy $H(X)$

The entropy of a binary random variable is given by:

$$H(X) = -p \log_2 p - (1-p) \log_2(1-p)$$

Substituting $p = 0.6$:

$$H(X) = -0.6 \log_2(0.6) - 0.4 \log_2(0.4)$$

$$\log_2(0.6) \approx -0.737, \quad \log_2(0.4) \approx -1.322$$

$$H(X) \approx 0.6(0.737) + 0.4(1.322) = 0.971 \text{ bits}$$

(b) Typical Set

A sequence x^n is in the typical set if:

$$2^{-n(H+\epsilon)} \leq P(x^n) \leq 2^{-n(H-\epsilon)}$$

With $n = 10$, $H \approx 0.971$, and $\epsilon = 0.1$:

$$\text{Lower bound} = 2^{-10(0.971+0.1)} = 2^{-10.71} \approx 0.0006$$

$$\text{Upper bound} = 2^{-10(0.971-0.1)} = 2^{-8.71} \approx 0.00238796929$$

The probability of a sequence with k ones is:

$$P(k) = (0.6)^k (0.4)^{10-k}$$

We evaluate which values of k satisfy the bounds:

k	$P(k)$
5	0.000797
6	0.001194
7	0.001791
8	0.002687 (excluded)

Thus, the typical set consists of sequences with:

$$k = 5, 6, 7$$

Number of Elements

$$|A_\epsilon^{(10)}| = \binom{10}{5} + \binom{10}{6} + \binom{10}{7}$$

$$= 252 + 210 + 120 = 582$$

Probability of Typical Set

$$P(A_\epsilon^{(10)}) = \sum_{k=5}^{7} \binom{10}{k} (0.6)^k (0.4)^{10-k}$$

$$\approx 0.201 + 0.251 + 0.215 = 0.667$$

Final Answers

- $H(X) \approx 0.971$ bits
- Typical set: sequences with $5, 6, 7$ ones
- Number of elements: 582
- Probability: ≈ 0.667