Asymptotic Equipartition

Recall: X random variable:

$$H(X) = -\sum_{x\in\mathcal{X}} p_X(x)\log_2 p_X(x) = \sum_{x\in\mathcal{X}} p_X(x)\underbrace{(-\log_2 p_X(x))}_{\mathbb{E}(-\log_2 p_X(X)}$$

Example. *$\mathcal{X} = \{0,1\}$ and $\mathbb{P}[X=1] = p(=\theta)$ and $\mathbb{P}[X=0] = 1-p(=1-\theta)$ So, $p_X(1) = p$ and $p_X(0) = 1-p$.*

$$p_X(X) = \begin{cases} p & \textit{if } X = 1 \\ 1-p \textit{ if } X = 0 \end{cases}$$

$$h = \lim \frac{1}{n} H(X_1,\ldots,X_n)$$
$$\lim \frac{1}{n}\,\mathbb{E}(-\log_2 p_{X_1,\ldots,X_n}(X_1,\ldots,X_n)$$

Definition. *The $\mathcal{X}$-valued stochastic process $(X_n)_{n\in\mathbb{N}}$ has the asymptotic equipartition property if*

$$-\frac{1}{n}\log_2 p_n(X_1,\ldots,X_n) \to h \textit{ almost surely}$$

Remark. *We have in the definition above $p_n = p_{X_1,\ldots,X_n}$ and $p_n(x_1,\ldots,x_n) = \mathbb{P}[X_1 = x_1,\ldots,X_n = x_n]$*

1. Specify classes of stochastic processes which do have the AEP
2. How can it be used?

Lemma 1.1. *If The X_n are iid $[h = H(X_1)]$, then the AEP holds.*

Proof.

$$\begin{aligned} p_n(x_1,\ldots,x_n) &= p(x_1)p(x_2)\ldots p(x_n) \\ -\frac{1}{n}\log_2 p_n(X_1,\ldots,X_n) &= -\frac{1}{n}\log_2 p(X_1)p(X_2)\ldots p(X_n) \\ &= -\frac{1}{n}\left[\log_2 p(X_1) + \log_2 p(X_2) + \cdots + \log_2 p(X_n)\right] \end{aligned}$$

$\implies \mathbb{E}(-\log_2 p(X_1)) = H(X_1) = h$

□

Example. *Continued X_n are iid Bernoulli, $\mathcal{X} = \{0,1\}$ and $(X_1,\ldots,X_n)$ are values in $\{0,1\}^n$*

$$p_n(x_1,\ldots,x_n) = p(x_1)p(x_2)\ldots p(X_n) = \theta^k(1-\theta)^{n-k}$$

where $k = |\{j : x_j = 1\}|$ and $n-k = |\{j : x_j = 0\}|$.

$$p_n(X_1,\ldots,X_n) = \theta^K(1-\theta)^{n-K}$$

where $K = |\{j : X_j = 1\}|$ (is random since value of rvs).

$$-\frac{1}{n}\log_2 \theta^K(1-\theta)^{n-K} simh$$

$\frac{K}{n} = \frac{X_1 + \cdots + X_n}{n} \to \theta$ almost surely

Definition. *Let $\varepsilon > 0$ and $n \in \mathbb{N}$. The* typical set *is*

$$A_\varepsilon^{(n)} = \{(x_1, \dots, x_n) \in \mathcal{X}^n : \left| -\frac{1}{n} \log_2 p_n(x_1, \dots, x_n) - h \right| < \varepsilon\}$$

Theorem 1.2. *If the AEP holds, then*

1. $\mathbb{P}[(X_1, \dots, X_n) \in A_\varepsilon^{(n)}] > 1 - \varepsilon$ *for all* $n \geq N_\varepsilon$
2. $(1-\varepsilon)2^{n(h-\varepsilon)} < (\leq) \left| A_\varepsilon^{(n)} \right| < (\leq) 2^{n(h+\varepsilon)}$ *for* $n \geq N_\varepsilon$

0 *For* $(x_1, \dots, X_n) \in A_\varepsilon^{(n)}$

$$2^{-n(h+\varepsilon)} < p_n(x_1, \dots, x_n) < 2^{-n(h-\varepsilon)}$$

Proof. 1.

$$\mathbb{P}[\left| -\frac{1}{n} \log_2 p_n(X_1, \dots, X_n) - h \right| \geq \varepsilon] \overset{n \to \infty}{\to} 0$$

$$\mathbb{P}[< \varepsilon] \overset{n \to \infty}{\to} 1$$

Thus,

$$\mathbb{P}[(X_1, \dots, X_n) \in A_\varepsilon^{(n)}] > 1 - \varepsilon$$

for $n \geq N_\varepsilon$

2.

$$1 \geq \sum_{(x_1, \dots, x_n) \in A_\varepsilon^{(n)}} p_n(x_1, \dots, x_n) > 2^{-n(h+\varepsilon)} \left| A_\varepsilon^{(n)} \right|$$

$$1 - \varepsilon < \mathbb{P}[(X_1, \dots, X_n) \in A_\varepsilon^{(n)}] = \sum_{(x_1, \dots, x_n) \in A_\varepsilon^{(n)}} p_n(x_1, \dots, x_n) < 2^{-n(h-\varepsilon)} \left| A_\varepsilon^{(n)} \right|$$

$\square$

Remark. *Let ε be very small, then*

$$|[|A_\varepsilon^{(n)} \approx 2^{nh}$$

on $A_\varepsilon^{(n)}$:

$$p_n(x_1, \dots, x_n) \approx 2^{-nh}$$

This is the reason why it is called asymptotic equipartition distribution.

Coding / Data compression

Let n be large, $p_n(x_1, \ldots, x_n)$. Elements to be encoded are $\mathcal{X}^n$, look for binary code $C(x_1, \ldots, x_n)$ such that

$$\sum_{(x_1,\ldots,x_n)} p_n(x_1, \ldots, x_n) \cdot \text{length}(C(x_1, \ldots, x_n))$$

is small.

Without probability: $l \equiv n \log_2 |\mathcal{X}|\rceil$ (ever sequence is encoded by a word in $\{0, 1^l)$

AEP: $\overline{(x_1, \ldots, x_n)} \in A_\varepsilon^{(n)} \to$ words over $\{0, 1\}$, length: $1 + \lceil n(h + \varepsilon)\rceil$ where the 1 is an initial 1, indicating "typical"

$(x_1, \ldots, x_n) \in \mathcal{X}^n \setminus A_\varepsilon^{(n)}$ length: $1 + \lceil \log_2 |\mathcal{X}|\rceil$ where the 1 is an initial 0 for non-typical.

Expected length:

$$\begin{aligned}
\mathbb{E}(l(X_1, \ldots, X_n)) = & \sum_{(x_1,\ldots,x_n)\in\mathcal{X}^n} p_n(x_1, \ldots, x_n) l(x_1, \ldots, x_n) = \sum_{(x_1,\ldots,x_n)\in A_\varepsilon^{(n)}} \cdots + \sum_{(x_1,\ldots,x_n)\in\mathcal{X}^n\setminus A_\varepsilon^{(n)}} \cdots \\
& \leq \sum_{(x_1,\ldots,x_n)\in A_\varepsilon^{(n)}} p_n(x_1, \ldots, x_n)(n(h - \varepsilon) + 2) + \sum_{(x_1,\ldots,x_n)\in\mathcal{X}^n\setminus A_\varepsilon^{(n)}} p_n(x_1, \ldots, x_n)(n \log_2 |\mathcal{X}| + \\
& \leq n(h + \varepsilon) + 2 + (n \log_2 |\mathcal{X}| + 2)\varepsilon = n(h + \varepsilon')
\end{aligned}$$

where $\varepsilon' = \varepsilon + \varepsilon \log_2 |\mathcal{X}| + \frac{2+2\varepsilon}{n}$ which is small for large n. Thus,

$$\mathbb{E}(l(X_1, \ldots, X_n)) \approx nh$$

Shannon-McMillan-Breiman

If the stochastic process $(X_n)_{n\in\mathbb{N}}$ is stationary and ergodic, then it has the AEP.

Nice version: If $(X_n)_{n\geq 0}$ is an irreducible, time homogeneous Markov chain, then it has the AEP.