Shannon Entropy

Suppose we have X, a classical random variable, Ω a sample space, and $p : \Omega \to [0, 1]$ a probability distribution of X.

Definition 1 (Shannon Entropy) *Given X, Ω, and p as above, we define the* Shannon entropy *as*

$$H(X) \equiv H(p) := -\sum_{x \in \Omega} p(x) \log_2 p(x).$$

For example, 1 fair coin has entropy $H = 1$ (representing 1 unit of information).

NB: $H(X) = \mathbb{E}_x(-\log p(x))$, where $\mathbb{E}$ denotes *expectation*.

For another example consider n fair coins, then it is easy to check that the Shannon entropy is $H = n$, since we have $p(x) = 2^{-n}$ for each $x \in \Omega = \{\text{Tails}, \text{Heads}\}^n$.

Also, if $p(x) = \dfrac{1}{2^x}$, $x = 1, 2, \ldots$, then the Shannon entropy is $H = 2$.

Intuition:

1. Uncertainty before learning the value of x.
2. Information gained from learning the value of x.

Note that the definition allows for the quantity of information to be additive.

Typical Sequences

Consider n random variables $X_1, X_2, \ldots, X_n$, i.i.d. source, each $X_i \sim X$. The outcomes are $x^n := x_1 x_2 \cdots x_n \in \Omega^n$. Let $f(a|x^n) = \dfrac{1}{n}$(no. times a source a occurs in x^n), where the f stands for *frequency*.

Expect: For most x^n that $f(a|x^n) \approx p(a)$.

Expect: $p(x^n) = \Pi_a p(a)^{n \cdot f(a|x^n)} \approx \Pi_a p(a)^{n \cdot p(a)} = 2^{-n \cdot H(X)}$

Definition 2 (Typical Set) *For $\epsilon > 0$, we define the* typical set *as*

$$T_{n,\epsilon}^{(S)} := \left\{ x^n : \forall a \ |f(a|x^n) - p(a)| < \frac{\epsilon}{|\Omega|} \right\}.$$

We say that $x^n \in T_{n,\epsilon}^{(S)}$ is strongly typical.

Definition 3 (Another Typical Set) *We define another typical set as below.*

$$T_{n,\epsilon} := \left\{ x^n : \left| -\frac{1}{n} \log p(x) - H(X) \right| \leq \epsilon \right\}.$$

We say that $x^n \in T_{n,\epsilon}$ is ϵ-typical.

An exercise is to show that

$$x^n \in T_{n,\epsilon}^{(S)} \Rightarrow x^n \in T_{n,\epsilon'},$$

and to find ϵ' in terms of ϵ.

Theorem 4 (Theorem 12.2 in Nielsen and Chuang, Asymptotic Equipartition Theorem.)
$\forall \epsilon > 0, \forall \delta > 0, \exists n_0$ such that $\forall n \geq n_0$:

1. $\displaystyle\sum_{x^n \in T_{n,\epsilon}} p(x^n) \geq 1 - \delta,$
2. $(1-\delta)2^{n(H(X)-\epsilon)} \leq |T_{n,\epsilon}| \leq 2^{n(H(X)+\epsilon)}$.

Proof :

1. Let $Y := -\log p(X)$ be a random variable, *e.g.*,if $X = a$, $Y = y = -\log p(a)$. Let $Y_i = -\log p(X_i)$ be i.i.d. $x^n \in T_{n,\epsilon} \iff |\frac{1}{n}\sum_i y_i - \mathbb{E}(Y)| \leq \epsilon$.

   Law of Large Numbers:
   $$\Pr_{Y_1,\ldots,Y_n}\left(\left|\frac{1}{n}\sum_{i=1}^{n} y_i - \mathbb{E}(Y)\right| \geq \epsilon\right) \leq \frac{\mathrm{Var}\,(Y)}{n\epsilon^2}.$$
   Choose $n_0 := \dfrac{\mathrm{Var}\,(Y)}{\epsilon^2\delta}$, then $\dfrac{\mathrm{Var}\,(Y)}{n\epsilon^2} \leq \delta$. Therefore,
   $$\Pr_{X_1,\ldots,X_n}(x^n \in T_{n,\epsilon}) \geq 1 - \delta,$$

   since this is a complementary event than that in the Law of Large Numbers.

2. For one inequality, we have that
   $$1 \geq \sum_{x^n \in T_{n,\epsilon}} p(x^n) \geq |T_{n,\epsilon}| \cdot \min_{x^n \in T_{n,\epsilon}} p(x^n) = |T_{n,\epsilon}| \cdot 2^{-n(H(X)+\epsilon)}.$$

   This implies $|T_{n,\epsilon}| \leq 2^{n(H(X)+\epsilon)}$. For the other inequality, we have
   $$(1-\delta) \leq \sum_{x^n \in T_{n,\epsilon}} p(x^n) \leq \sum_{x^n \in T_{n,\epsilon}} 2^{-n(H(X)-\epsilon)} = |T_{n,\epsilon}| 2^{-n(H(X)-\epsilon)},$$

   where the first inequality holds by part 1. This implies that $|T_{n,\epsilon}| \geq (1-\delta)2^{n(H(X)-\epsilon)}$ as required.

Data Compression:

Let X_i be an i.i.d. source. Then $\forall R > H(X)$, $\forall \delta > 0$, $\exists n_0$ such that $\forall n \geq n_0$, there exists an encoding map, $\mathcal{E}_n : \Omega^n \to \{0,1\}^{nR}$, and a decoding map, $\mathcal{D}_n : \{0,1\}^{nR} \to \Omega^n$, such that

$$\Pr_{X_1,\ldots,X_n}(\mathcal{D} \circ \mathcal{E}_n(x^n) \neq x^n) \leq \delta,$$

where R is the number of bits we are willing to spend per copy of X for the compression and δ is the fidelity parameter.

We usually choose our typical set to have $\epsilon = R - H(X)$.

Let $\mathcal{L}_n : T_{n,\epsilon} \to \{0,1\}^{nR}$ be a labeling map for $T_{n,\epsilon}$, and $P_n(x^n) = x^n$ if $x^n \in T_{n,\epsilon}$ and $P_n(x^n) = \text{ERROR}$ otherwise. Then, $\mathcal{E}_n = \mathcal{L}_n \circ P_n$ works.

Quantum Analogue:

Definition 5 (Von Neumann Entropy) *Suppose ρ is a density matrix with spectral decomposition $\rho = \sum_\lambda p(\lambda)|\lambda\rangle\langle\lambda|$. We define the* Von Neumann entropy *of ρ as*

$$S(\rho) := -\mathrm{Tr}\,(\rho \log \rho) = H(\Lambda),$$

where Λ is a random variable with values λ. Recall that $\log(\rho) := \sum_\lambda \log(p(\lambda))|\lambda\rangle\langle\lambda|$.

The idea is that if we look at the eigenbasis of ρ we can treat it classically.

Let $\{q_x, |\psi_x\rangle\}$ be an ensemble where the $|\psi_x\rangle$'s are not necessarily orthogonal and let ρ be the corresponding density matrix. Again, write the spectral decomposition of $\rho = \sum_x q_x|\psi_x\rangle\langle\psi_x|$ as $\sum_\lambda p(\lambda)|\lambda\rangle\langle\lambda|$. We will apply data compression in the eigenbasis of ρ. Notice that $\rho^{\otimes n} = \sum_{\lambda^n} p(\lambda^n)|\lambda^n\rangle\langle\lambda^n|$, where $|\lambda^n\rangle := |\lambda_1\rangle|\lambda_2\rangle\cdots|\lambda_n\rangle$ is a tensor product of eigenvectors of ρ. We let $T_{n,\epsilon}$ be the typical set for Λ^n. Define

$$P_{n,\epsilon} = \sum_{\lambda^n \in T_{n,\epsilon}} |\lambda^n\rangle\langle\lambda^n|,$$

which is the projector onto the *typical subspace* $\mathcal{H}_T$, where $\mathcal{H}_T = \mathrm{span}\,(T_{n,\epsilon})$. Note that $\mathrm{Rank}\,(P_{n,\epsilon}) = \dim(\mathcal{H}_T) = |T_{n,\epsilon}|$. Also, we have the identity

$$\mathrm{Tr}\,(P_{n,\epsilon}\rho^{\otimes n}) = \sum_{\lambda^n \in T_{n,\epsilon}} p(\lambda^n) \geq 1 - \delta. \tag{1}$$

<u>Quantum Sources and Data Compression</u>

Let $\{q_x, |\psi_x\rangle\}$ be an i.i.d. source, $|\psi_x\rangle \in \mathbb{C}_d$, and let X be a classical random variable with distribution $q(x) := q_x$.

Theorem 6 (Data Compression Theorem) *$\exists n_0$ such that $\forall n \geq n_0$, $\exists \mathcal{E}_n, \mathcal{D}_n$, such that*

$$\sum_{x^n \in T_{n,\epsilon}} q(x^n) \cdot F(|\psi_{x^n}\rangle\langle\psi_{x^n}|, \mathcal{D}_n \circ \mathcal{E}_n(|\psi_{x^n}\rangle\langle\psi_{x^n}|)) \geq 1 - \delta,$$

where $\mathcal{E}_n : \mathcal{H}_d^{\otimes n} \to \mathcal{H}_2^{\otimes nR}$, and $R > S(\rho) = S(\sum_x q(x)|\psi_x\rangle\langle\psi_x|)$.

Proof : Let $\epsilon = R - S(\rho) > 0$, and let $P_{n,\epsilon}$ and $\mathcal{H}_T$ be as defined earlier. Let $\mathcal{L}_n$ be the change of basis from $\mathcal{H}_T$ to $\mathcal{H}_2^{\otimes nR}$. $\mathcal{E}_n(\rho) = \mathcal{L}_n \circ (P_{n,\epsilon}(\rho)P_{n,\epsilon}) + |e\rangle\langle e|\mathrm{Tr}\,((I - P_{n,\epsilon})(\rho)(I - P_{n,\epsilon}))$, where $|e\rangle$ is some state we don't care about. Notice we have

$$|\psi_{x^n}\rangle = P_{n,\epsilon}|\psi_{x^n}\rangle + (I - P_{n,\epsilon})|\psi_{x^n}\rangle.$$

We can also write ρ_{out}, being the output of data compression when the input state is $|\psi_{x^n}\rangle\langle\psi_{x^n}|$, as

$$\rho_{out} = \mathcal{D}_n \circ \mathcal{E}_n(|\psi_{x^n}\rangle\langle\psi_{x^n}|) = P_{n,\epsilon}|\psi_{x^n}\rangle\langle\psi_{x^n}|P_{n,\epsilon} + |e\rangle\langle e|\mathrm{Tr}\,((I - P_{n,\epsilon})|\psi_{x^n}\rangle\langle\psi_{x^n}|(I - P_{n,\epsilon})).$$

Then we have

$$\begin{aligned}
1 - \delta &\leq \mathrm{Tr}\,(P_{n,\epsilon}\rho^{\otimes n}) \qquad \text{from Eq. (1)} \\
&= \sum_{x^n \in T_{n,\epsilon}} q(x^n)\mathrm{Tr}\,(|\psi_{x^n}\rangle\langle\psi_{x^n}|P_{n,\epsilon}) \\
&= \sum_{x^n \in T_{n,\epsilon}} q(x^n)\langle\psi_{x^n}|P_{n,\epsilon}|\psi_{x^n}\rangle \\
&= \sum_{x^n \in T_{n,\epsilon}} q(x^n)F(|\psi_{x^n}\rangle\langle\psi_{x^n}|, \rho_{out}).
\end{aligned}$$

This proves the result.

An exercise is to check that the converse holds, that is, if $R \leq S(\rho)$, then no noiseless data compression is possible.

Definition 7 (Conditional Entropy) *Let X and Y be random variables. We define the* conditional entropy *as*

$$H(X|Y) = \sum_y p(y) H(X|Y = y) = -\sum_{x,y} p(x,y) \log \left(\frac{p(x,y)}{p(y)} \right).$$

Note that $H(X|Y) = H(X,Y) - H(Y) = \sum_y p(y) H(X|Y = y)$.

Definition 8 (Mutual Information) *We define* mutual information *as*

$$I(X:Y) = H(X) - H(X|Y).$$

By the above note, we have that

$$I(X:Y) = H(X) - H(X|Y) = H(X) - H(X,Y) + H(Y) = H(Y) - H(Y|X) = I(Y:X),$$

so mutual information is symmetric.

Definition 9 (Relative Entropy) *We define the* relative entropy *as*

$$H(p||q) = \sum_x p(x) \log \left(\frac{p(x)}{q(x)} \right).$$

Note that relative entropy is *not* symmetric, that is, $H(p||q) \neq H(q||p)$ in general.
Letting u be the uniform distribution over Ω, $|\Omega| = n$, we have $H(X) = \log n - H(p||u)$.
We also have

$$I(X:Y) = H(p(x,y)||p(x)p(y)).$$

Theorem 10 *$H(p||q) \geq 0$, with equality if and only if $p = q$.*

The proof of the above result is in Nielsen and Chuang.

Corollary 11

- $H(X) \leq \log(n)$,
- $I(X:Y) \geq 0$, *with equality if and only if X and Y are independent,*
- $H(X,Y) \leq H(X) + H(Y)$ *Subadditivity,*
- $H(X|Y) \leq H(X)$,
- $\sum_y p(y) H(X|Y = y) \leq H\left(\sum_y (X|Y = y) \right)$ *Concavity.*