Entropy

February 11, 2015

In this lecture we will discuss the concept of Entropy. We will begin by studying entropy through simple finite state, discrete time, processes. Then work up to Markov chains with an infinite state space.

1 History

[From `wikipedia.org`]

Consider a Carnot cycle as pictured below. The heat bath on the left is at temperature

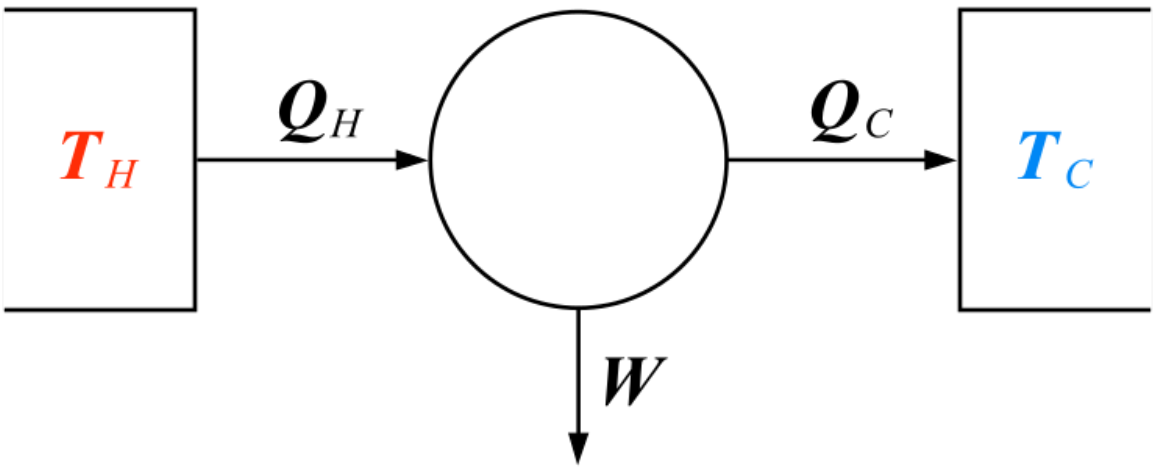

Figure 1: A diagram of a Carnot cycle from `http://en.wikipedia.org/wiki/File:Carnot_heat_engine_2.svg`.

T_H (hot) and the bath at the left is at temperature T_C (cold). The two baths, when the switch is opened, will want to revert to equilibrium. So thermal energy will be transfered from the hot bath to the switch Q_H then to the switch to Q_C. The maximum amount of work that can be done is then

$$W_{max} = \underbrace{\left(1 - \frac{T_C}{T_H}\right)}_{\text{efficiency}} Q_H.$$

This was calculated by Kelvin. It was first hypothesized when the incorrect assumption that $Q_H = Q_C$ was implied. The work can be easily calculated by finding

$$W = Q_H - Q_C.$$

This equation is valid over the entire cycle. Thus

$$Q_H - Q_C < \left(1 - \frac{T_C}{T_H}\right) Q_H$$

$$\frac{Q_C}{T_C} > \frac{Q_H}{T_H}$$

So the entropy is defined as $S_H = Q_H/T_H$ and similarly for S_C. There is a increase of entropy in the system. There is an irreversible process which prevents the optimal amount of work. Entropy is so named because it is close to (en)ergy and tropē, which means transformation.

2 Entropy for Random Variables

Boltzmann was the first to define entropy for a statistical mechanical system. Let $X \in A$ be a random variable on $(\Omega, \mathcal{F}, P)$ with $A = \{1, 2, ..., r\}$ a finite state space. Let $P_X(j) = p_j$ be the pmf of X for $j = 1, ..., r$. Then the *entropy* of X, denoted $H(X)$, is defined as

$$H(X) = -\sum_{i=1, p_j \neq 0}^{r} p_j \log(p_j)$$

and we define any term of the sum zero if $p_j = 0$. It is the logarithmic measure of the number of states. The connection with heat can be made, but it is not very exciting.

2.1 Properties and Examples

What are some properties of H. For example, what are the min and max of H? A uniform distribution on A gives the maximum entropy. This is proved later in Theorem 2. By inspection, we get the smallest when $p_i = 1$ for some $i = 1, ..., r$. With $H(X) = 0$.

1. $H(X) \geq 0$ for all RV X.
2. $H(X)$ achieves its maximum of $\log(r)$ for $p_i = 1/r$ for all $i = 1, ..., r$.
3. $H(X)$ is its minimum of 0 when X is a constant random variable.

Let's do a couple of examples that demonstrate some intuition about entropy.
Example 1: Find the entropy of Y, a geometrically distributed random variable with parameters (r, p).
Solution: Recall that the pmf of Y is

$$P_Y(j) = (1-p)^{j-1} p.$$

Thus the entropy is

$$H(Y) = -\sum_{i=1}^{r}(1-p)^{j-1}p\left[(j-1)\log(1-p)+\log(p)\right]$$

This example does not give us much intuition. Let's do a couple of very simple examples.
Example 2: Consider $A = \{1, 2\}$ as a coin flip between heads and tails. Suppose it's a weighted coin, giving you state 1 with probability p. What is the entropy of this system?
Solution: We have

$$H(X) = p\log(p) + (1-p)\log(1-p) = \log(1-p) - p\log(p-p^2).$$

The more sure of what state we are in, the lower the entropy. It maxes out at the uniform distribution of $p = 1/2$. The next example demonstrates the meaning of uncertainty that entropy captures.
Example 3: Consider $X \in A = \{1, 2, 3\}$. Then the entropy is

$$H(X) = \frac{1}{4}\log(4) + \frac{1}{2}\log(2) + \frac{1}{4}\log(4).$$

The logarithmic factor provides a weight to the usual probability distribution.

Next, we prove a theorem using the Law of Large Numbers and entropy.

Theorem 1 (From [1])**.** *Let X_i be IID random variables on $(\Omega, \mathcal{F}, P)$ with state space $A = \{1, 2, 3, ..., r\}$ with probability mass function $p_X(j)$. Define*

$$\nu_j^n = \sum_{i=1}^{n}\chi_{\{X_i=j\}}$$

as the number of times state j appears in n trials. Then, for all $\epsilon > 0$, and large n, one can find a subset $\Omega_n \subseteq \Omega^n$ such that

1. $\lim_{n\to\infty} P(\Omega_n) = 1$.
2. $e^{-n(H(X)+\epsilon)} \leq p(\omega) \leq e^{-n(H(X)-\epsilon)}$ *for all* $\omega \in \Omega^n$.
3. $e^{n(H(X)-\epsilon)} \leq |\Omega_n| \leq e^{n(H(X)+\epsilon)}$.

Proof. of 1. Let

$$\Omega_n = \left\{\omega : \left|\frac{\nu_j^n}{n} - p_X(j)\right| \leq \delta, \quad 1 \leq j \leq r\right\},$$

where $\delta = \delta(\epsilon)$ will be chosen later. By the Law of Large Numbers, $P(\Omega_n) \to 1$ as $n \to \infty$.

of 2. Assume all $p_X(j) > 0$. If not, we can disregard the jth site. Then for all $\omega \in \Omega^n$, by independence,

$$P(\omega) = p_X(1)^{\nu_1^n} p_X(2)^{\nu_2^n} \cdots p_X(r)^{\nu_r^n}.$$

Next we use a common technique in entropy and statistical mechanics. We write the probability as an exponential. Therefore,

$$P(\omega) = \exp\left(\sum_{j=1}^{r} \nu_j^n \log(p_X(j))\right)$$

Now we add and subtract by $n\sum p_x(j)\log(p_x(j))$ in the exponential to obtain

$$\begin{aligned} P(\omega) =& \exp\left(n\sum_{j=1}^{r} p_X(j)\log(p_X(j))\right)\exp\left(\sum_{j=1}^{r}\left(\frac{\nu_j^n}{n} - p_X(j)\right)\log(p_X(j))\right) \\ =& \exp\left(n\left[-H(X) + \sum_{j=1}^{r}\left(\frac{\nu_j^n}{n} - p_X(j)\right)\log(p_X(j))\right]\right) \end{aligned}$$

Now we choose δ small enough such that

$$\left|\sum_{j=1}^{r}\left(\frac{\nu_j^n}{n} - p_X(j)\right)\log(p_X(j))\right| \leq \epsilon.$$

of 3. We first write,

$$1 \geq P(\Omega_n).$$

From the proof of 2, we can bound this from below by

$$1 \geq P(\Omega_n) \geq e^{-n(H+\epsilon)}|\Omega_n|.$$

This gives the upper bound in 3. The lower bound is trickier. Note that, by 1, there exists n large enough such that $P(\Omega_n) \geq 1/2$. Then we use the upper bound from the proof of 2,

$$\frac{1}{2} \leq P(\Omega_n) \leq e^{-n(H-\epsilon/2)}|\Omega_n|$$

We have the factor of $\epsilon/2$ in the exponential because we will need to choose a new n to rid of the factor of $1/2$ as the lower bound. This gives the lower bound

$$|\Omega_n| \geq \frac{1}{2}e^{n(H-\epsilon/2)}.$$

We now choose a sufficiently large $\tilde{n}$ such that

$$|\Omega_{\tilde{n}}| \geq \frac{1}{2}e^{\tilde{n}(H-\epsilon/2)} \leq e^{n(H-\epsilon)}.$$

□

2.2 Relative Entropy

[From [2]] In this section, we discuss relative entropy and how it can be used as a measure of distance between two distributions. However, it is not a full metric because it lacks the symmetry property.

Relative entropy is a way to measure how close are two probability distributions μ and ν. Define the relative entropy as

$$H(\mu||\nu) = \sum_{i=1}^{r} \mu_i \log\left(\frac{\mu_i}{\nu_i}\right) = -H(X) - \sum_{i=1}^{r} \mu_i \log(\nu_i)$$

Note that relative entropy can be infinite if $\nu_i = 0$ where $\mu_i \neq 0$.

To show the utility of relative entropy we prove the following theorem.

Theorem 2 (Non-negativity)**.** *The relative entropy is non-negative and is zero if and only if $\nu = \mu$.*

Proof. First, note that $\log(x) \leq x - 1$ for all $x > 0$. Thus

$$H(\mu||\nu) = -\sum_{i=1}^{r} \mu_i \log\frac{\nu_i}{\mu_i} \geq \sum \mu_i \left(1 - \frac{\nu_i}{\mu_i}\right)$$

Summing over all i gives

$$H(\mu||\nu) \geq -\sum_{i=1}^{r} \mu_i - \nu_i = 0$$

Note that the relative entropy is zero in the first equality if and only if $\mu_i = \nu_i$ for every $i = 1, ..., r$. □

Now we can use relative entropy to prove that the uniform distribution provides a maximum for the entropy of a RV.

Theorem 3. *The maximum entropy for a distribution on a finite state space $A = \{1, 2, 3, ..., r\}$ is the uniform distribution. Let μ be the probability distribution of X over the r out comes of A. Let ν be the uniform distribution over r outcomes. I.e. $\nu_i = 1/r$ for all i. Then the entropy of μ is bounded above by the entropy of ν.*

Proof. Let us compute the relative entropy of μ and ν. By the definition,

$$H(\mu||\nu) = \sum_{i=1}^{r} \mu_i \log(\mu_i/\nu_i) = \sum_{i=1}^{r} \mu_i \log(\mu_i d).$$

Using the multiplicative property of logs we have

$$H(\mu||\nu) = \sum_{i=1}^{r} \mu_i \log(\mu_i) + \mu_i \log(d).$$

Using the definition of entropy of X and summing over all i for the term on the right, we have

$$H(\mu||\nu) = \log(d) - H(X).$$

Now we use the non-negativity of relative entropy to give our result,

$$\log(d) - H(X) \geq 0 \implies \log(d) \geq H(X).$$

□

The above proof also lets us interpret the entropy of X as the "distance" away from the uniform distribution. I.e.

$$H(X) = -\sum_i \mu_i \log(\mu_i) = -\sum_i \mu_i \log(\mu_i \frac{d}{d}) = \log(d) - \sum_i \mu_i \log(\frac{\mu_i}{\frac{1}{d}}).$$

However, the relative entropy is not quite a metric because it is not symmetric. Let $A = \{1, 2\}$ and $\mu_1 = \alpha$ and $\nu_2 = \beta$ with $\alpha \neq \beta$. Then

$$H(\mu||\nu) = \alpha \log(\alpha/\beta) + (1-\alpha)\log((1-\alpha)/(1-\beta)) \neq \beta \log(\beta/\alpha) + (1-\beta)\log((1-\beta)/(1-\alpha)).$$

We can also use relative entropy to prove the Markov Chain Convergence Theorem. That is, given a Markov chain with a unique stationary distribution ν, then given any initial distribution μ, $\mu P^n \to \nu$ as $n \to \infty$.

Theorem 4 (From [3])**.** *Let X_n be an irreducible and aperiodic Markov chain with finite state space $A = \{1, 2, 3, ..., r\}$, transition matrix P, and unique stationary distribution ν. Then given an arbitrary initial condition μ,*

$$\mu P^n \to \nu \quad \text{as } n \to \infty.$$

Proof. First we show that for any vector ν, $H(\nu P||\pi) \leq H(\nu||\pi)$. To do so, we use the (weak) convexity of the function $f(x) = x\log(x)$. Note that

$$f''(x) = \frac{1}{x} > 0$$

for all $x \in (0, 1]$. Thus,

$$H(\mu P||\pi) = \sum_i \left(\sum_j \mu_j P_{j,i} \right) \log \frac{\left(\sum_j \mu_j P_{j,i} \right)}{\pi_i} \leq \sum_i \sum_j P_{ji} \mu_j \log \frac{\mu_j}{\pi_i}.$$

Now we break the log apart using the multiplicative property of logs, to get

$$H(\mu P||\pi) \leq \sum_i \sum_j \mu_j P_{ji} \log(\mu_j) - \sum_i \sum_j \mu_j P_{ji} \log(\pi_i)$$

We compute the sum on the left over i first, using the property that $\sum_i P_{ji} = 1$. Next for the sum on the right, note that $\log(x)$ is a concave function and thus

$$\sum_i P_{ji} \log(\pi_i) \leq \log(\sum_i P_{ji} \pi_i) = \log(\pi_j)$$

by the fact that π is the stationary distribution. Therefore

$$H(\mu P||\pi) \leq \sum_j \mu_j \log(\mu_j) - \sum_j \mu_j \log(\pi_j) = H(\mu||\pi).$$

Note that by the definition of irreducible Markov chain, any state is reachable from any other state given enough time. That is, for $i \neq j$, there exists $n_1 \in \mathbb{N}$ such that

$$P(X_{n_1} = i | X_0 = j) > 0.$$

By the definition of aperiodic, there exists an n_2 such that for all $n \geq n_2$ and $X_0 = i$, state i can reach itself with non–zero probability. Therefore

$$P(X_n = i | X_0 = i) > 0.$$

Thus there exists $m > \max(n_1, n_2)$ such that

$$P_{ij}^m > 0$$

for all i, j. This implies that in the convexity and concavity arguments given before, because $x \log(x)$ is strict convex and $\log(x)$ is strictly concave, then for

$$P_{ij}^m > 0$$

and $\sum_i P_{ij}^m = \sum_j P_{ij}^m = 1$, then

$$H(\mu P^m||\pi) < H(\mu||\pi)$$

for $\mu \neq \pi$.

This implies that there exists a subsequence $H(\mu P^{n_k}||\pi)$ that is monotonic and bounded, by $H(\mu||\pi)$ and 0. Therefore $H(\mu P^{n_k}||\pi)$ converges and thus so does $H(\mu P^n||\pi)$. Now we must justify that the limit is 0. Note that by continuity of relative entropy, μP^n converges as well. Let ν be a limit point of μP^n. Suppose that $H(\nu||\pi) > 0$. If we apply the transition matrix P^m to ν we see that

$$0 \leq H(\nu P^m||\pi) < H(\nu||\pi)$$

Thus ν must not be the limit point unless $\nu = \pi$ which implies $H(\nu||\pi) = 0$. We already have that this implies $\nu = \pi$. □

Now we can ask, what if the Markov chain we are working with does not have a stationary distribution? Or what if we are working with a process that is not Markov? Can we define entropy in a certain manner for these processes?

2.3 Entropy for a Process

[From [4]]

Note that we can also write the entropy of a random variable as an expectation. That is

$$H(X) = -E[\log p_X(i)] = -\sum_i p_X(i) \log(p_X(i)).$$

We will define joint and conditional entropy in a similar manner.

Let $X, \tilde{X}$ be two random variables on a state space $A = \{1, 2, ..., r\}$ with joint pmf $p_{X,\tilde{X}}(i, j)$. Then the joint entropy is given as

$$H(X, \tilde{X}) = -E[\log p_{X,\tilde{X}}(i, j)] = -\sum_{i,j} p_{X,\tilde{X}} \log(p_{X,\tilde{X}}(i, j)).$$

This makes it natural to define the conditional entropy as

$$H(X|\tilde{X}) = -\sum_{i,j} p_{X,\tilde{X}}(i, j) \log p_{X,\tilde{X}}(i|j) = -\sum_{i,j} p_{X,\tilde{X}}(i, j) \log \frac{p_{X,\tilde{X}}(i, j)}{p_{\tilde{X}}(j)}.$$

Our intuition tells us that if we are given information $\tilde{X}$, then the system is a little more ordered. The analysis of this intuition gives the inequality,

$$H(X|\tilde{X}) \leq H(X)$$

Proof. Note again, that $x \log x$ is a convex function. Thus

$$H(X|\tilde{X}) = -\sum_i \sum_j p_{X,\tilde{X}}(i, j) \log \frac{p_{X,\tilde{X}}(i, j)}{p_{\tilde{X}}(j)}$$

First we multiply and divide each term of H by $p_{\tilde{X}}(j)$.

$$H(X|\tilde{X}) = -\sum_i \sum_j p_{X,\tilde{X}}(i, j) \frac{p_{\tilde{X}}(j)}{p_{\tilde{X}}(j)} \log \frac{p_{X,\tilde{X}}(i, j)}{p_{\tilde{X}}(j)}.$$

We write the joint pmf in terms of the conditional. Then, again, we use the concave property of $f(x) = -x \log(x)$ to take the sum over j inside the log. Thus we get

$$H(X|\tilde{X}) \leq -\sum_i \left(\sum_j p_{\tilde{X}}(j) p_{X|\tilde{X}}(i|j) \right) \log \left(\sum_j p_{\tilde{X}}(j) p_{X|\tilde{X}}(i|j) \right).$$

multiply by $p_{\tilde{X}}(j)/p_{\tilde{X}}(j)$ and pull the sum inside. Now we can write the conditional pmfs as joint pmfs and sum over j to obtain

$$H(X|\tilde{X}) \leq -\sum_i p_X(i) \log p_X(i) = H(X).$$

□

For the reader:

- Next, consider the n random variables $X_1, X_2, ...$ Show that $H(X_n, ..., X_0)/n$ is the average of $H(X_i)$.
- Use the previous exercise to show that for a stationary process, $\{H(X_n|X_{n-1}, ..., X_0); n \geq 1\}$ is a monotonically decreasing and non–nonegative. Therefore we define the entropy of the process as

$$H(X) = \lim_{n\to\infty} H(X_n|X_{n-1}, ..., X_0)$$

References

[1] Leonid B. Koralov and Yakov G. Sinai. *Theory of probability and random processes*. Universitext. Springer, Berlin, second edition, 2007.

[2] Michael A. Nielsen and Isaac L. Chuang. *Quantum computation and quantum information*. Cambridge University Press, Cambridge, 2000.

[3] Timo Seppäläinen and Firas Rassoul-Agha. *A Course on Large Deviations with an Introduction to Gibbs Measures*. 2014.

[4] Joseph C. Watkins. *Discrete Time Stochastic Processes*. 2007.