ENTROPY IN ERGODIC THEORY

TRISHA SABADRA

Abstract. Entropy is a fundamental concept in information theory and ergodic theory, characterizing the unpredictability or complexity of a system. The seminal work by Claude Shannon established the quantitative measure of information known as Shannon entropy; this paper extends the notion of entropy beyond discrete probabilistic frameworks to the measure-theoretic context. The paper highlights the Shannon-McMillan-Breiman theorem, which asserts the convergence of entropy in stationary ergodic processes. Finally, we examine the practical role of entropy in developing loss functions for model optimization in machine learning.

1. Introduction to Entropy

The concept of information entropy was introduced by Claude Shannon in his 1948 paper "A Mathematical Theory of Communication" [Sha48] and is also referred to as Shannon entropy. The core idea of information theory is that the "informational value" of a message depends on the degree to which the content of the message is surprising. If a highly likely event occurs, the message carries very little information. On the other hand, if a highly unlikely event occurs, the message is much more informative.

Entropy is a way to quantify the measure of information, measuring the expected or average amount of information conveyed by identifying the outcome of a random trial. The *information content*, also called the *surprisal* or *self-information*, of an event E is a function which increases as the probability $p(E)$ of an event decreases. This relationship can be described by the function

$$-\log\left(\frac{1}{p(E)}\right).$$

In N (independent) events, where the ith event has probability p_i, we will get total information I of

$$I = -\sum_{i=1}^{n}(N \cdot p_i) \cdot \log(1/p_i).$$

But then, the average information we get per symbol observed will be

$$\frac{I}{N} = \left(\frac{1}{N}\right)\sum_{i=1}^{n}(N \cdot p_i) \cdot \log(1/p_i) = \sum_{i=1}^{n} p_i \cdot \log(1/p_i).$$

This brings us to a fundamental definition of entropy [Car14].

Definition 1. *Let the probability distribution $P = \{p_1, p_2, \ldots, p_n\}$. We define the* ***entropy*** *of the distribution P by:*

$$H(P) = -\sum_{i=1}^{n} p_i \cdot \log(1/p_i).$$

If we have a continuous rather than discrete probability distribution $P(x)$:

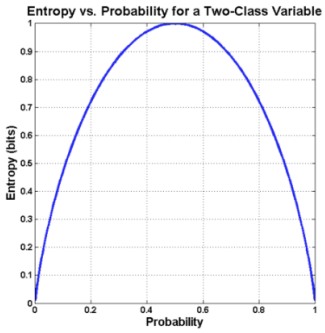
$$H(P) = -\int P(x) \cdot \log(1/P(x))\, dx.$$

Example. Consider tossing a coin with probabilities p and q for heads and tails, respectively. The entropy of the unknown result of the next toss of the coin is maximized if the coin is fair ($p = q = 1/2$) because there is maximum uncertainty: it is most difficult to predict the outcome of the next toss.

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) = -\sum_{i=1}^{2} \frac{1}{2} \log_2 \frac{1}{2} = -2 \cdot \frac{1}{2} \cdot (-1) = 1$$

For an unfair coin, entropy decreases since one outcome is more probable; for example, if $p = 0.7$:

$$H(X) = -0.7 \log_2(0.7) - 0.3 \log_2(0.3) = 0.8816 < 1.$$



2. Measure-Theoretic Entropy

Entropy can be formally defined in the language of measure theory [Fie19].

Definition 2. *A **partition** of a measure space $(X, \mathcal{B}, m)$ is a pairwise disjoint countable collection of sets $\{P_1, P_2, \ldots\}$ such that $\bigcup_k P_k = X$.*

Definition 3. *The **join** of the partitions $P = \{P_1, P_2, \ldots\}$ and $Q = \{Q_1, Q_2, \ldots\}$, is $P \vee Q = \{P_i \cap Q_j\}$. It can be extended to more than two partitions as follows:*

$$P_1 \vee \ldots \vee P_n = \bigvee_{i=1}^{n} P_i = \{P_{r_1} \cap P_{r_2} \cap \ldots \cap P_{r_n}\}$$

where $P_{r_j} \in P_i, r_j \geq 1$.

The result of the join is another partition. We will mostly concern ourselves with finite partitions, each of which generates a finite σ-algebra. If after a certain n, the partition stops becoming more refined, then we can say that the transformation is no longer adding "information" to the system.

Definition 4. *Define the **information function** for measure μ and partition P*

$$I_P(x) = -\log \mu(P(x)) = -\sum_{P_i \in P} 1_{P_i}(x) \log \mu(P_i).$$

This brings us to the fundamental definition of entropy in measure theory.

Definition 5. *The* ***entropy*** *of a finite partition* $P = \{P_1, P_2, \ldots, P_k\}$ *is*

$$H(P) = \mathbb{E}(I_P) = -\sum_{i=1}^{k} \mu(P_i)\log(\mu(P_i)).$$

Since $\lim_{x\to 0} x\log(x) = 0$, *we let* $0\log(0) = 0$.

Definition 6. *The entropy of a finite partition* P *and transformation* T, *is*

$$H(P,T) = \lim_{n\to\infty} \frac{1}{n} H\left(\bigvee_{i=0}^{n-1} T^{-i}(P)\right).$$

The entropy $H(P,T)$ quantifies the new information obtained per application of a transformation T to a partition P, averaged over n applications. To isolate T's intrinsic effect, we define T's entropy as the supremum of $H(P,T)$ over all finite partitions, thereby removing dependence on any particular partition.

Definition 7. *The entropy of a transformation* T *is* $H(T) = \sup(H(P,T))$ *over all finite partitions* P.

Just from this definition, we draw a few conclusions about the entropy of certain transformations. First, the entropy of the identity transformation is 0. Given that the identity does not change the original partition at all, we have $\lim_{n\to\infty} \frac{1}{n} H(P, \mathrm{Id}^n) = 0$, which is 0 for any partition, so the supremum is 0. Indeed, the same holds for any periodic ($T = T^k$ for some k) transformation.

3. Shannon-McMillan-Breiman Theorem

The Shannon-McMillan-Breiman theorem is a fundamental result that extends Claude Shannon's concept of entropy to the realm of random processes. Essentially, the theorem states that for a stationary ergodic process, the average unpredictability of a single symbol converges almost surely to a constant value as the length of the sequence of symbols goes to infinity. This constant value is equal to the entropy rate of the process, which is the average entropy per symbol of the process [UWs].

Theorem 1 (Shannon-McMillan-Breiman Theorem)**.** *Let* $(X, \mathcal{B}, \mu, T)$ *be a measure-preserving transformation and* P *a finite partition with* $H(P) < \infty$. *Let* $P_n = \bigvee_{k=0}^{n-1} T^{-k}(P)$ *and* $P_n(x)$ *the element of* P_n *containing* x. *Then*

$$h(P,T) = \lim_{n\to\infty} -\frac{1}{n}\log(\mu(P_n(x)))$$

Before we begin the proof of this theorem, we must first define some terms.

Definition 8. *For a measure preserving system* $(X, \mathcal{B}, \mu, T)$, *some measurable function* $f : X \to \mathbb{R}$ *and* σ*-algebra* $\mathcal{C}$, *we define the* ***conditional expectation*** $E_\mu(f|\mathcal{C})$ *as the unique* $\mathcal{C}$*-measurable function* $\widetilde{f}$ *such that*

$$\int_C \widetilde{f} d\mu = \int_C f d\mu \text{ for all } C \in \mathcal{C}.$$

$\mathcal{C}$-measurable means that $\widetilde{f}^{-1}((t,\infty)) \in \mathcal{C}$ for all $t \in \mathbb{R}$, and therefore $\widetilde{f}$ must be constant on all atoms of $\mathcal{C}$. It is the function $\widetilde{f}$ such that for each atom C,

$$\widetilde{f}(x) = \frac{1}{\mu(C)} \int_C f d\mu \text{ for } \mu\text{-a.e. } x \in C.$$

The finer the σ-algebra $\mathcal{C}$, the more $\widetilde{f}$ looks like f. This is expressed in the following theorem.

Theorem 2 (Martingale Convergence Theorem)**.** *If $(\mathcal{C}_n)_n$ is a sequence of σ-algebras such that $\mathcal{C}_{n+1}$ refines $\mathcal{C}_n$ and $\mathcal{C} = \lim_{n\to\infty} \mathcal{C}_n := \bigvee_{n=1}^{\infty} \mathcal{C}_n$, then for every $f \in L^1(\mu)$*

$$E_\mu(f|\mathcal{C}_n) \to E_\mu(f|\mathcal{C}) \quad \text{as } n \to \infty.$$

An elegant and short proof can be found here [Isa].

Definition 9. *We define **conditional entropy** of a measure μ and partitions P and Q as*

$$H_\mu(P|Q) = -\sum_{Q_j \in Q} \mu(Q_j) \sum_{P_i \in P} \frac{\mu(P_i \cap Q_j)}{\mu(Q_j)} \log \frac{\mu(P_i \cap Q_j)}{\mu(Q_j)}.$$

Theorem 3. *Given measures μ, μ_i and two partitions P and Q, these properties follow:*

1. *$H_\mu(P \vee Q) \le H_\mu(P) + H_\mu(Q)$;*
2. *$H_\mu(Q) = H_\mu(P) + H_\mu(Q|P)$, and hence $h_\mu(T, Q) = h_\mu(T, P) + H_\mu(Q|P)$.*
3. *$\sum_{i=1}^n p_i H_{\mu_i}(P) \le H_{\sum_{i=1}^n p_i \mu_i}(P)$ for each probability vector $(p_1, \dots, p_n)$.*

Definition 10. *Similarly to conditional entropy, we define the **conditional information function***

$$I_{P|Q}(x) := -\sum_{P_i \in P} \sum_{Q_j \in Q} (1)_{P_i \cap Q_j}(x) \log \frac{\mu(P_i \cap Q_j)}{\mu(Q_j)}.$$

Comparing this to conditional entropy, we get

$$\int_X I_{P|Q} d\mu = -\sum_{P_i \in P} \sum_{Q_j \in Q} \mu(P_i \cap Q_j) \log \frac{\mu(P_i \cap Q_j)}{\mu(Q_j)} = H_\mu(P_i|Q_j).$$

By the previous definition and Theorem 3, we can show that

$$I_{P \vee Q} = I_P + I_{Q|P}. \tag{1}$$

By the definition of conditional expectation and because $1_{P\cap Q} = 1_P 1_Q$ we have

$$\begin{aligned} -\log \mathbb{E}_\mu(1_P(x)|Q) &= -\log \mathbb{E}_\mu \left(\sum_{P \in \mathcal{P}} 1_P | Q \right) \\ &= -\log \sum_{Q \in \mathcal{Q}} \frac{1}{\mu(Q)} \int_Q \sum_{P \in \mathcal{P}} 1_P d\mu \\ &= -\log \sum_{P \in \mathcal{P}} \sum_{Q \in \mathcal{Q}} 1_{P \cap Q} \frac{\mu(P \cap Q)}{\mu(Q)} \\ &= l_{P|Q}(x). \end{aligned}$$

We are now ready to prove the Shannon-Breiman-McMillan Theorem.

Proof. Let $g_k(x) = I_{P|V_{j=1}^{k-1}T^{-j}P}(x)$ for $k \geq 2$ and $g_1(x) = I_P$. Then by (1),

$$\begin{aligned}
I_{V_{j=0}^{n-1}T^{-j}P}(x) &= I_{V_{j=1}^{n-1}T^{-j}P}(x) + I_{P|V_{j=1}^{n-1}T^{-j}P}(x) \\
&= I_{V_{j=0}^{n-2}T^{-j}P}(Tx) + g_n(x) \\
&= I_{V_{j=0}^{n-2}T^{-j}P}(Tx) + I_{P|V_{j=1}^{n-2}T^{-j}P}(Tx) + g_n(x) \\
&= I_{V_{j=0}^{n-3}T^{-j}P}(T^2x) + g_{n-1}(Tx) + g_n(x) \\
&\vdots \\
&= g_1(T^{n-1}(x)) + \cdots + g_{n-1}(T(x)) + g_n(x) \\
&= \sum_{j=0}^{n-1} g_{n-j}(T^j x).
\end{aligned}$$

Let $g = \lim_{n\to\infty} g_n$, which belongs to $L^1(\mu)$ because of the Martingale convergence theorem. We write the previous equaIity as

$$\frac{1}{n} I_{V_{j=0}^{n-1}T^{-j}P}(x) = \frac{1}{n}\sum_{j=0}^{n-1} g(T^j x) + \frac{1}{n}\sum_{j=0}^{n-1}(g_{n-j} - g)(T^j x).$$

Since μ is ergodic, the first sum converges almost everywhere with respect to the measure μ to $\int_X g d\mu$, which is equal to $H_\mu(P|V_{j=1}^{\infty}T^{-j}P)$ by (1), which in turn is equal to $h(P,T)$.

For the second sum, we define

$$G_N = \sup_{k\geq N} |g_k - g| \quad \text{and} \quad g^* = \sup_{n\geq 1} g_n.$$

Then $0 \leq G_N \leq g + g^*$ and $g + g^* \in L^1(\mu)$ because $\int_X g_n d\mu = H_\mu(P|V_{j=1}^{n-1}T^{-j}P)$ is decreasing in n. Moreover, $G_N \to 0$ μ-a.e., so by the dominated convergence theorem,

$$\lim_{N\to\infty} \int_X G_N d\mu = \int_X \lim_{N\to\infty} G_N d\mu = 0$$

Now for any $N \geq 1$ and $n \geq N$ we split the second sum:

$$\begin{aligned}
\frac{1}{n}\sum_{j=0}^{n-1}(g_{n-j} - g)(T^j x) &= \frac{1}{n}\sum_{j=0}^{n-N-1}(g_{n-j} - g)(T^j x) + \frac{1}{n}\sum_{j=n-N}^{n-1}(g_{n-j} - g)(T^j x) \\
&\leq \frac{1}{n}\sum_{j=0}^{n-N-1} G_N(T^j x) + \frac{1}{n}\sum_{j=n-N}^{n-1}(g_{n-j} - g)(T^j x).
\end{aligned}$$

First take the limit $n \to \infty$. The second sum tends to zero, and by the ergodic theorem, the first sum tends to $\int_X G_N d\mu$. Finally, taking $N \to \infty$, also $\int_X G_N d\mu \to 0$. Hence

$$\frac{1}{n}\sum_{j=0}^{n-1} I_{V_{j=0}^{n-1}T^{-j}P}(x) \to h(P,T) \quad \mu\text{-a.e.},$$

as required. This finishes the proof. □

This theorem is powerful because it connects the concept of entropy, a theoretical measure of information content, with the practical occurrence of sequences in a random process, showing that there is a predictable pattern in how information is distributed across different sequences in the long run.

4. Application of Shannon entropy in Machine Learning

Cross entropy is a key component in creating loss functions that measure the accuracy of predictions made by logistic regression models and neural networks, types of predictive models used in machine learning. It is used in both binomial and multinomial classification scenarios [MMZ23].

Definition 11. *The* ***cross entropy*** *compares two probability distributions p and q*

$$H(p,q) = E_p[-\log(q)] = -\int p(x)\cdot\log(q(x))dx.$$

$H(p,q)$ gives us the average number of bits required to code an event from q if we use the "wrong" coding scheme q instead of p. In machine learning, it is a very useful measure for the similarity of probability distributions and serves as a loss function, an equation that calculates how far the prediction deviates from the actual values. Usually, p is used for the true (or empirical) distribution (i.e., the distribution of the training set), and q is the distribution described by a model.

Let's take the binary logistic regression as an example. The two classes are labeled 0 and 1, and the logistic model assigns the probabilities $q_{y=1} = \hat{y}$ and $q_{y=0} = 1-\hat{y}$ to each input x. This can be concisely written as $q \in \{\hat{y}, 1-\hat{y}\}$. Using this notation, the cross entropy between empirical and estimated distribution for a single sample is

$$H(p,q) = -\sum_i p_i \log(q_i) = -y\log(\hat{y}) - (1-y)\log(1-\hat{y}).$$

When used as a loss function, the average of all cross entropies from all N samples is used,

$$L = -\frac{1}{N}\sum_{j=1}^{N}\sum_i p_{ij}\log(q_{ij}) = -\frac{1}{N}\sum_{j=1}^{N} y_j\log(\hat{y}_j) - (1-y_j)\log(1-\hat{y}_j).$$

In essence, the cross-entropy framework underpins a critical loss function in machine learning algorithms, an exciting connection between entropy and the practical endeavors of predictive modeling.

References

[Car14] Tom Carter. An introduction to information theory and entropy. *CSU Stanislaus*, pages 15–25, September 3, 2014.

[Fie19] Jacob Fielder. Ergodic theory and entropy. *University of Chicago*, pages 17–18, May, 2019.

[Isa] Richard Isaac. A proof of the martingale convergence theorem. *American Mathematical Society.*

[MMZ23] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications. *PMLR Volume 202: International Conference on Machine Learning*, June 20, 2023.

[Sha48] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 1948.

[UWs] The shannon-mcmillan-breiman theorem. *Universitat Wien.*

Euler Circle, Mountain View, CA 94040