ROMA TRE
UNIVERSITÀ DEGLI STUDI

Introduction to Information Entropy

Adapted from Papoulis (1991)

Federico Lombardo

Papoulis, A., *Probability, Random Variables and Stochastic Processes*, 3rd edition, McGraw Hill, 1991.

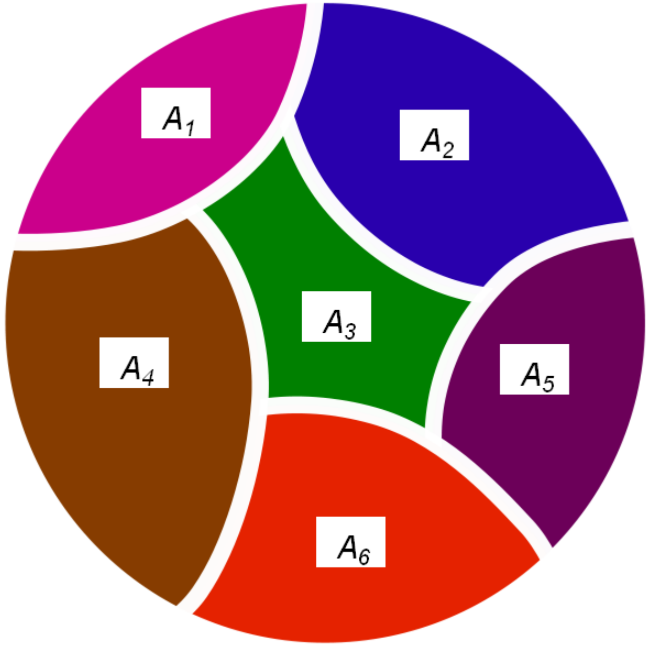
1. INTRODUCTION

Definitions

- The probability P(A) of an event A_i can be interpreted as a measure of our uncertainty about the occurrence or nonoccurrence of A in a single performance of the underlying experiment S (certain event).

- We are interested in assigning a measure of uncertainty to the occurrence or nonoccurrence not of a single event of S, but of any event A_i of a partition **A** of S, where a partition is a collection of mutually exclusive events whose union equals S.

- The measure of uncertainty about **A** will be denoted by H(**A**) and will be called the entropy of the partitioning **A**.



Definitions

- In his landmark paper, Shannon (1948) derived the functional H(**A**) from a number of postulates based on our heuristic understanding of uncertainty. The following is a typical set of such postulates:
  - H(**A**) is a continuous function of p_i = P(A_i).
  - If $p_1 = \ldots = p_N = 1/N$, then H(**A**) is an increasing function of N.
  - If a new partition **B** is formed by subdividing one of the sets of **A**, then H(**B**) ≥ H(**A**).

- It can be shown that the following sum satisfies these postulates and it is unique within a constant factor:

$$\mathrm{H}(\mathbf{A}) = -\sum_{i=1}^{N} p_i \log p_i$$

- The above assertion can be proven, but here we propose to follow the Papoulis (1991) approach by introducing the above formula as the definition of entropy and developing axiomatically all its properties within the framework of probability.

Shannon, C.E., *A Mathematical Theory of Communication*, Bell System Technical Journal, vol. 27, pp. 379–423, 623-656, July, October, 1948.
Papoulis, A., *Probability, Random Variables and Stochastic Processes*, 3rd edition, McGraw Hill, 1991.

Applications

- The applications of entropy can be divided into two categories:

1. Problems involving the determination of unknown distributions:
   - the available information is in the form of known expected values or other statistical functionals, and the solution is based on the principle of maximum entropy;
   - determine the unknown distributions so as to maximize the entropy H(**A**) of some partition **A** subject to the given constraints.

2. Coding theory:
   - in this second category, we are given H(**A**) (source entropy) and we wish to construct various random variables (code lengths) so as to minimize their expected values;
   - the solution involves the construction of optimum mappings (codes) of the random variables under consideration, into the given probability space.

Uncertainty and Information

- In the heuristic interpretation of entropy the number H(**A**) is a measure of our uncertainty about the events A_i of the partition **A** prior to the performance of the underlying experiment.

- If the experiment is performed and the results concerning A_i become known, then the uncertainty is removed.

- We can thus say that the experiment provides information about the events A_i equal to the entropy of their partition.

- Thus uncertainty equals information and both are measured by entropy.

Examples

1. Determine the entropy of the partition **A** = [even, odd] in the fair-die experiment. Clearly, P{even} = P{odd} = 1/2. Hence

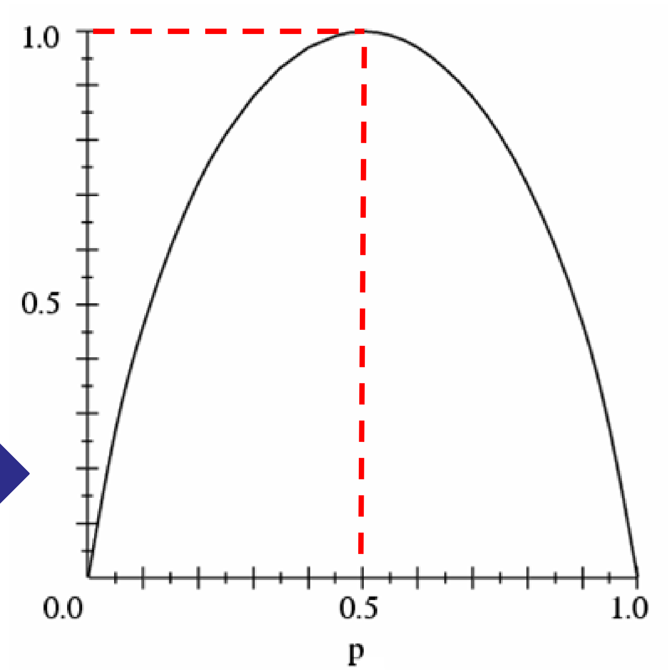
$$\mathrm{H}(\mathbf{A}) = -1/2\log 1/2 - 1/2\log 1/2 = \log 2$$

2. In the same experiment, **G** is the partition consisting of the elementary events $\{f_i\}$. In this case, $\mathrm{P}\{f_i\}$ = 1/6; hence

$$\mathrm{H}(\mathbf{G}) = -\sum_{i=1}^{6} \mathrm{P}\{f_i\}\log \mathrm{P}\{f_i\} = \log 6$$

3. We consider now the coin experiment where P{heads} = *p*. In this case, the entropy of **G** equals

$$\mathrm{H}(\mathbf{G}) = -p\log p - (1-p)\log(1-p) \equiv r(p)$$



Maximum Entropy

- An important application of entropy is the determination of the probabilities p_i of the events of a partition **A** subject to various constraints, with the maximum entropy (ME) method.

- ME principle states that the unknown p_i's must be so chosen as to maximize the entropy of **A** subject to the given constraints (Jaynes, 1957).

- The ME principle is equivalent to the principle of insufficient reason (Bernoulli, 1713): "*In the absence of any prior knowledge, we must assume that the events A_i have equal probabilities*". This conclusion is based on the subjective interpretation of probability as a measure of our state of knowledge about the events A_i.

- Operationally, the ME method simplifies the analysis drastically when, as is the case in most applications, the constraints are phrased in terms of probabilities in the space S^n of repeated trials (i.e., the resulting product space from the experiment S repeated n times).

Jaynes, E. T., *Information Theory and Statistical Mechanics*, Physical Review, 106(4), 620–630, 1957.
Bernoulli, J., *Ars Conjectandi*, 1713.

Examples

1. Determine the probabilities p_i of the six faces of a die, having access to no prior information. The ME principle states that the p_i's must be such as to maximize the sum

$$\mathrm{H}(\mathbf{G}) = -p_1 \log p_1 - \ldots - p_6 \log p_6$$

   Since $p_1 + \ldots + p_6 = 1$, this yields $p_1 = \ldots = p_6 = 1/6$, in agreement with the classical definition.

2. A player places a bet of one euro on "odd" and he wins, on the average, 20 cents per game. We wish again to determine the p_i's using the ME method; however, now we must satisfy the constraints

$$p_1 + p_3 + p_5 = 0.6 \qquad\qquad p_2 + p_4 + p_6 = 0.4$$

   This is a consequence of the available information because an average gain of 20 cents means that P{odd} - P{even} = 0.2. Maximizing H(**G**) subject to the above constraints, we obtain

$$p_1 = p_3 = p_5 = 0.2 \qquad\qquad p_2 = p_4 = p_6 = 0.133$$

Considerations

- The ME method is thus a valuable tool in the solution of applied problems. It can be used, in fact, even in deterministic problems involving the estimation of unknown parameters from insufficient data.
- We should emphasize, however, that as in the case of the classical definition of probability, the conclusions drawn from the ME method must be accepted with skepticism particularly when they involve elaborate constraints.
- Concerning the previous examples, we conclude that all p_i's must be equal in the absence of prior constraints, which is not in conflict with our experience concerning dice. The second conclusion, however, is not as convincing, we would think, even though we have no basis for any other conclusion.
- One might argue that this apparent conflict between the ME method and our experience is due to the fact that we did not make total use of our prior knowledge. This might be true; however, it is not always clear how such constraints can be phrased analytically and, even if they can, how complex the required computations might be.

Interpretation of Probability

1. **<u>Axiomatic</u>**: P(A) is a number assigned to the event A. This number satisfies the following three postulates but is otherwise arbitrary
   - The probability of an event A is a positive number, $P(A) \geq 0$
   - The probability of the certain event S equals 1, $P(S) = 1$
   - If the events A and B are mutually exclusive, $P(A + B) = P(A) + P(B)$

2. **<u>Empirical</u>**: For large n, $P(A) \approx k/n$, where k is the number of times A occurs in n repetitions of the underlying experiment S.

3. **<u>Subjective</u>**: P(A) is a measure of our uncertainty about the occurrence of A in a single performance of S.

4. **<u>Principle of insufficient reason</u>**: If A_i are N events of a partition $\mathbf{A}$ of S and nothing is known about their probabilities, then $P(A_i) = 1/N$.

Interpretation of Entropy

1. **Axiomatic**: H(**A**) is a number assigned to each partition **A** = [A_1, …, A_N] of S. This number equals the sum $-\sum p_i \ln p_i$, where $p_i = \mathrm{P}(A_i)$ and $i = 1, \ldots, N$

2. **Empirical**: This interpretation involves the repeated performance not of the experiment S, but of the experiment S^n of repeated trials. In this experiment, each specific typical sequence $\boldsymbol{t}_j$ = {A_i occurs $n_i \approx np_i$ times in a specific order j} is an event with probability

$$P(\mathbf{t}_j) = p_1^{n_1} \cdots p_N^{n_N} \approx e^{np_1 \ln p_1 + \cdots + np_N \ln p_N} = e^{-n\mathrm{H}(\mathbf{A})}$$

Applying the relative frequency interpretation of probability to this event, we conclude that if the experiment S^n is repeated m times and the event $\boldsymbol{t}_j$ occurs m_j times, then for sufficiently large m,

$$P(\mathbf{t}_j) = e^{-n\mathrm{H}(\mathbf{A})} \approx \frac{m_j}{m}; \quad \text{hence} \quad \mathrm{H}(\mathbf{A}) \approx -\frac{1}{n} \ln \frac{m_j}{m}$$

This relates the theoretical quantity H(**A**) to the experimental numbers m_j and m.

Interpretation of Entropy

3. **Subjective**: H(**A**) is a measure of our uncertainty about the occurrence of the events A_i of the partition **A** in a single performance of *S*.

4. **Principle of maximum enrtropy**: The probabilities p_i = P(A_i) must be such as to maximize H(**A**) subject to the given constraints. Since it can be demonstrated that the number of typical sequences is $n_\mathbf{t} = e^{nH(\mathbf{A})}$, the ME principle is equivalent to the principle of maximizing $n_\mathbf{t}$. If there are no constraints, that is, if nothing is known about the probabilities p_i, then the ME principle leads to the estimates $p_i = 1/N$, H(**A**) = ln*N*, and $n_\mathbf{t} = N^n$.

2. BASIC CONCEPTS

Conditional Entropy

- The entropy H(**A**) of a partition **A** = [A_i] gives us a measure of uncertainty about the occurrence of the events A_i at a given trial.
- If in the definition of entropy we replace the probabilities P(A_i) by the conditional probabilities P(A_i|M), we obtain the conditional entropy H(**A**|M) of **A** assuming M
$$\mathrm{H}(\mathbf{A}|M) = -\sum_i \mathrm{P}(A_i|M) \log \mathrm{P}(A_i|M)$$
- From this it follows that if at a given trial we know that M occurs, then our uncertainty about **A** equals H(**A**|M).
- If we know that the complement M^C of M occurs, then our uncertainty equals H(**A**|M^C).
- Assuming that the binary partition **M** = [M, M^C] is observed, the uncertainty per trial about **A** is given by the weighted sum
$$\mathrm{H}(\mathbf{A}|\mathbf{M}) = \mathrm{P}(M)\mathrm{H}(\mathbf{A}|M) + \mathrm{P}(M^C)\mathrm{H}(\mathbf{A}|M^C)$$

Mutual Information

- If at each trial we observe the partition **B** = [B_j], then we show that the uncertainty per trial about **A** equals H(**A**|**B**)

- Indeed, in a sequence of n trials, the number of times the event B_j occurs equals $n_j \approx n\mathrm{P}(B_j)$; in this subsequence, the uncertainty about **A** equals H(**A**|B_j) per trial. Hence, the total uncertainty about **A** equals
$$\sum_j n_j \, \mathrm{H}\left(\mathbf{A} \middle| B_j\right) \approx \sum_j n \, P\left(B_j\right) \mathrm{H}\left(\mathbf{A} \middle| B_j\right) = n \, \mathrm{H}\left(\mathbf{A} \middle| \mathbf{B}\right)$$
and the uncertainty per trial equals H(**A**|**B**)

- Thus the observation of **B** reduces the uncertainty about **A** from H(**A**) to H(**A**|**B**). The <u>mutual information</u>
$$\mathrm{I}\left(\mathbf{A}, \mathbf{B}\right) = \mathrm{H}\left(\mathbf{A}\right) - \mathrm{H}\left(\mathbf{A} \middle| \mathbf{B}\right)$$
is the reduction of the uncertainty about **A** resulting from the observation of **B**. I(**A**, **B**) can be interpreted as the information about **A** contained in **B**.

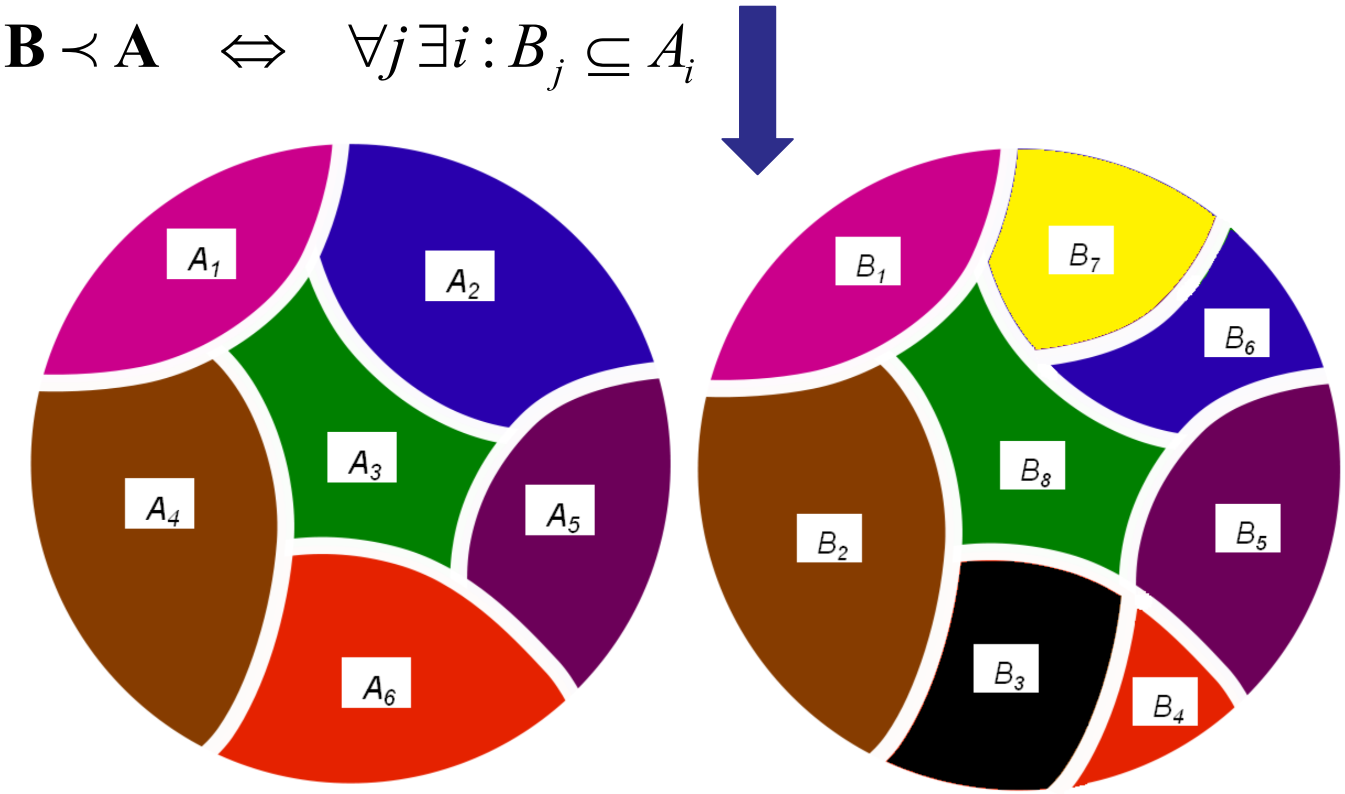
Definitions

1. A partition whose elements are the elementary events $\{\zeta_i\}$ of the space S will be denoted by **G** and will be called the <u>element partition</u>.

2. A <u>refinement</u> of a partition **A** is a partition **B** such that each element B_j of **B** is a subset of some element A_i of **A**. We shall use the following notation:

$$\mathbf{B} \prec \mathbf{A} \quad \Leftrightarrow \quad \forall j \, \exists i : B_j \subseteq A_i$$

3. The <u>product</u> of two partitions **A** and **B** is a partition whose elements are all intersections $A_i \cap B_j$ of the elements of **A** and **B**. This new partition is denoted by **A • B**.



Considerations

1. If **B** is a refinement of **A**, it can be shown that H(**A**) ≤ H(**B**).

   Then, for any **A** we have H(**A**) ≤ H(**G**), where **G** is the element partition.

2. If **B** is a refinement of **A** and **B** is observed, then we know which of the events of **A** occurred. Hence H(**A**|**B**) = 0.

3. Thus, for any **A** we have H(**A**|**G**) = 0.

4. For any **A** and **B**, we have that H(**A** • **B**) ≥ H(**A**) and H(**A** • **B**) ≥ H(**B**), because **A** • **B** is a refinement of both **A** and **B**.

5. If the partitions **A** and **B** are independent (i.e., their events are all independent of each other) and **B** is observed, then no information about **A** is gained. Hence H(**A**|**B**) = H(**A**).

Considerations

6. If we observe **B**, our uncertainty about **A** cannot increase.

   Hence H(**A**|**B**) ≤ H(**A**).

7. To observe **A** • **B**, we must observe **A** and **B**. If only **B** is observed, the information gained equals H(**B**). Therefore, the uncertainty about **A** assuming **B**, equals the remaining uncertainty, H(**A**|**B**) = H(**A** • **B**) – H(**B**).

8. Combining 6 and 7, we conclude that H(**A** • **B**) ≤ H(**A**) + H(**B**).

9. If **B** is observed, then the information that is gained about **A** equals I(**A**, **B**).
   - If **B** is a refinement of **C** and **B** is observed, then **C** is known.
   - But knowledge of **C** yields information about **A** equal to I(**A**, **C**).
   - Hence, if **B** is a refinement of **C**, then I(**A**, **B**) ≥ I(**A**, **C**).
   - Equivalently, we have also that H(**A**|**B**) ≤ H(**A**|**C**).

3. RANDOM VARIABLES AND STOCHASTIC PROCESSES

Random Variables

- We are given an experiment specified by the space S, the field of subsets of S called events, and the probability assigned to these events.

- To every outcome ζ of this experiment, we assign a number $\mathbf{x}(\zeta)$. We have thus created a function $\mathbf{x}$ with domain the set S and range a set of numbers. This function is called <u>random variable</u> (RV) if it satisfies the following conditions but is otherwise arbitrary:
  - The set of experimental outcomes $\{\mathbf{x} \le x\}$ is an event for every x.
  - The probabilities of the events $\{\mathbf{x} = \infty\}$ and $\{\mathbf{x} = -\infty\}$ equal 0.

- The elements of the set S that are contained in the event $\{\mathbf{x} \le x\}$ change as the number x takes various values. The probability $P\{\mathbf{x} \le x\}$ is, therefore, a number that depends on x.

- This number is denoted by $F(x)$ and is called the <u>cumulative distribution function</u> (CDF) of the RV $\mathbf{x}$: $F(x) = P\{\mathbf{x} \le x\}$

Random Variables

- The RV **x** is of continuous type if its CDF F(x) is continuous. In this case, we have: P{**x** = x} = 0.

- The RV **x** is of discrete type if its CDF F(x) is a staircase function. Denoting by x_i the discontinuity points of F(x), we have: P{**x** = x_i} = p_i.

- The derivative f(x) of F(x) is called the probability density function (PDF) of the RV **x**

$$\mathrm{f}(x) = \frac{d\,\mathrm{F}(x)}{dx}$$

- If the RV **x** is of discrete type taking the values x_i with probabilities p_i, then

$$\mathrm{f}(x) = \sum_i p_i\,\delta(x - x_i) \qquad p_i = \mathrm{P}\{\mathbf{x} = x_i\}$$

where $\delta(x)$ is the impulse function. The term $p_i\,\delta(x - x_i)$ can be shown as a vertical arrow at $x = x_i$ with length equal to p_i.

Entropy of RVs

- Entropy is a number assigned to a partition. To define the entropy of an RV we must, therefore, form a suitable partition.

- This is simple if the RV is of discrete type. However, for continuous-type RVs we can do so only indirectly.

- Suppose that the RV **x** is of discrete type taking the values x_i with probabilities $P\{\mathbf{x} = x_i\} = p_i$.
  - The events $\{\mathbf{x} = x_i\}$ are mutually exclusive and their union is the certain event; hence they form a partition.
  - This partition will be denoted by $\mathbf{A_x}$ and will be called the partition of **x**.

- *Definition*: The entropy $H(\mathbf{x})$ of a discrete-type RV **x** is the entropy $H(\mathbf{A_x})$ of its partition $\mathbf{A_x}$:

$$H(\mathbf{x}) = H(\mathbf{A_x}) = -\sum_i p_i \ln p_i$$

Continuous type

- The entropy of a continuous-type RV cannot be so defined because the events $\{\mathbf{x} = x_i\}$ do not form a partition (they are not countable).

- To define H($\mathbf{x}$), we form, first, the discrete-type RV $\mathbf{x}_\delta$ obtained by rounding off $\mathbf{x}$, so as to make it a staircase function: $\mathbf{x}_\delta = n\delta$ if $n\delta - \delta < \mathbf{x} \le n\delta$, hence

$$\mathrm{P}(\mathbf{x}_\delta = n\delta) = \mathrm{P}(n\delta - \delta < \mathbf{x} \le n\delta) = \int_{n\delta-\delta}^{n\delta} \mathrm{f}(x)dx = \delta\,\bar{\mathrm{f}}(n\delta)$$

where $\bar{\mathrm{f}}(n\delta)$ is a number between the maximum and the minimum of f(x) in the interval ($n\delta - \delta$, $n\delta$).

- Applying the definition of the entropy of a discrete-type RV to $\mathbf{x}_\delta$ we obtain

$$\mathrm{H}(\mathbf{x}_\delta) = -\sum_{n=-\infty}^{\infty} \delta\,\bar{\mathrm{f}}(n\delta)\ln\left[\delta\,\bar{\mathrm{f}}(n\delta)\right]$$

where:

$$\sum_{n=-\infty}^{\infty} \delta\,\bar{\mathrm{f}}(n\delta) = \int_{-\infty}^{\infty} \mathrm{f}(x)dx = 1$$

Continuous type

- After algebraic manipulations, we conclude that

$$\mathrm{H}(\mathbf{x}_\delta) = -\ln\delta - \sum_{n=-\infty}^{\infty} \delta\,\bar{\mathrm{f}}(n\delta)\ln\bar{\mathrm{f}}(n\delta)$$

- As $\delta \to 0$, the RV $\mathbf{x}_\delta \to \mathbf{x}$, but its entropy H($\mathbf{x}_\delta$) $\to \infty$ because: – ln$\delta \to \infty$.

- For this reason, we define the entropy H($\mathbf{x}$) of $\mathbf{x}$ not as the limit of H($\mathbf{x}_\delta$) but as the limit of the sum: H($\mathbf{x}_\delta$) + lnδ, as $\delta \to 0$. This yields

$$\mathrm{H}(\mathbf{x}_\delta) + \ln\delta \xrightarrow[\delta\to 0]{} -\int_{-\infty}^{\infty} \mathrm{f}(x)\ln\mathrm{f}(x)\,dx$$

- <u>*Definition*</u>: The entropy of a continuous-type RV $\mathbf{x}$ is by definition the integral

$$\mathrm{H}(\mathbf{x}) = -\int_{-\infty}^{\infty} \mathrm{f}(x)\ln\mathrm{f}(x)\,dx$$

- <u>Example</u>: If $\mathbf{x}$ is uniform in the interval (0, a), where f(x) = 1/a, then

$$\mathrm{H}(\mathbf{x}) = -\frac{1}{a}\int_0^a \ln\frac{1}{a}\,dx = \ln a$$

Considerations

- The entropy $H(\mathbf{x}_\delta)$ of $\mathbf{x}_\delta$ is a measure of our uncertainty about the RV **x** rounded off to the nearest $n\delta$. If δ is small, the resulting uncertainty is large and it tends to ∞ as $\delta \to 0$.

- This conclusion is based on the assumption that **x** can be observed perfectly; that is, its various values can be recognized as distinct no matter how close they are.

- In a physical experiment, however, this assumption is not realistic. Values of **x** that differ slightly cannot always be treated as distinct (noise considerations or round-off errors, for example).

- Accounting for the term $\ln\delta$ in the definition of entropy of a continuous-type RV **x** is, in a sense, a recognition of this ambiguity.

Considerations

- As in the case of arbitrary partitions, the entropy of a discrete-type RV **x** is positive and it is used as a measure of uncertainty about **x**.

- This is not so, however, for continuous-type RVs. Their entropy can take any value from $-\infty$ to ∞ and it is used to measure only changes in uncertainty.

- The various properties of partitions also apply to continuous-type RVs if, as is generally the case, they involve only differences of entropies.

Entropy as Expected Value

- The entropy of a continuous-type RV $\mathbf{x}$ can be expressed as the expected value of the RV $\mathbf{y} = -\ln f(\mathbf{x})$:

$$\mathrm{H}(\mathbf{x}) = \mathrm{E}\{-\ln \mathrm{f}(\mathbf{x})\} = -\int_{-\infty}^{\infty} \mathrm{f}(x) \ln \mathrm{f}(x)\, dx$$

- Similarly, the entropy of a discrete-type RV $\mathbf{x}$ can be written as the expected value of the RV $-\ln p(\mathbf{x})$:

$$\mathrm{H}(\mathbf{x}) = \mathrm{E}\{-\ln \mathrm{p}(\mathbf{x})\} = -\sum_i p_i \ln p_i$$

where now $p(x)$ is a function defined only for $x = x_i$ and such that $p(x_i) = p_i$.

- If the RV $\mathbf{x}$ is *exponentially* distributed, then $f(x) = \lambda e^{-\lambda x} U(x)$, where $U(x)$ is the Heaviside step function. Hence:

$$\mathrm{H}(\mathbf{x}) = \mathrm{E}\{-\ln \mathrm{f}(\mathbf{x})\} = 1 - \ln \lambda = \ln \frac{e}{\lambda}$$

- If the RV $\mathbf{x}$ is *normally* distributed, then

$$\mathrm{f}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \mathrm{H}(\mathbf{x}) = \mathrm{E}\{-\ln \mathrm{f}(\mathbf{x})\} = \ln\left(\sigma\sqrt{2\pi e}\right)$$

Joint Entropy

- Suppose that **x** and **y** are two <u>discrete-type</u> RVs taking the values x_i and y_j respectively with P{**x** = x_i, **y** = y_j} = p_{ij}.

- Their joint entropy, denoted by H(**x**, **y**), is by definition the entropy of the product of their respective partitions. Clearly, the elements of $\mathbf{A_x} \cdot \mathbf{A_y}$ are the events {**x** = x_i, **y** = y_j}. Hence

$$\mathrm{H}(\mathbf{x},\mathbf{y}) = \mathrm{H}(\mathbf{A_x} \cdot \mathbf{A_y}) = -\sum_{i,j} p_{ij} \ln p_{ij}$$

- The above can be written as an expected value:

$$\mathrm{H}(\mathbf{x},\mathbf{y}) = \mathrm{E}\{-\ln \mathrm{p}(\mathbf{x},\mathbf{y})\}$$

where p(x, y) is a function defined only for $x = x_i$ and $y = y_j$ and it is such that p(x_i, y_j) = p_{ij}.

- The joint entropy H(**x**, **y**) of two <u>continuous-type</u> RVs **x** and **y** is defined as the limit of the sum: H($\mathbf{x}_\delta$, $\mathbf{y}_\delta$) + 2 lnδ, as $\delta \to 0$, where $\mathbf{x}_\delta$ and $\mathbf{y}_\delta$ are their staircase approximation. Thus we have:

$$\mathrm{H}(\mathbf{x},\mathbf{y}) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{f}(x,y)\ln \mathrm{f}(x,y)\,dx\,dy = \mathrm{E}\{-\ln \mathrm{f}(\mathbf{x},\mathbf{y})\}$$

Conditional Entropy

- Consider two <u>discrete-type</u> RVs **x** and **y** taking the values x_i and y_j respectively with $$\mathrm{P}\left(\mathbf{x}=x_i \middle| \mathbf{y}=y_j\right)=\pi_{ji}=p_{ji}/p_j$$

- The conditional entropy H(**x**|y_j) of **x** assuming **y** = y_j is by definition the conditional entropy of the partition $\mathbf{A_x}$ of **x** assuming {**y** = y_j}. From the above it follows that: $$\mathrm{H}\left(\mathbf{x}\middle|y_j\right)=-\sum_i \pi_{ji}\ln\pi_{ji}$$

- The conditional entropy H(**x**|**y**) of **x** assuming **y** is the conditional entropy of $\mathbf{A_x}$ assuming $\mathbf{A_y}$. Thus $$\mathrm{H}\left(\mathbf{x}\middle|\mathbf{y}\right)=-\sum_j p_j\,\mathrm{H}\left(\mathbf{x}\middle|y_j\right)=-\sum_{i,j} p_{ji}\ln\pi_{ji}$$

- For <u>continuous-type</u> RVs the corresponding concepts are defined similarly

$$\mathrm{H}\left(\mathbf{x}\middle|y\right)=-\int_{-\infty}^{\infty}\mathrm{f}\left(x\middle|y\right)\ln\mathrm{f}\left(x\middle|y\right)dx=\mathrm{E}\left\{-\ln\mathrm{f}\left(\mathbf{x}\middle|\mathbf{y}\right)\middle|\mathbf{y}=y\right\}$$

$$\mathrm{H}\left(\mathbf{x}\middle|\mathbf{y}\right)=-\int_{-\infty}^{\infty}\mathrm{f}(y)\mathrm{H}\left(\mathbf{x}\middle|y\right)dy=-\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathrm{f}(x,y)\ln\mathrm{f}\left(x\middle|y\right)dx\,dy=\mathrm{E}\left\{-\ln\mathrm{f}\left(\mathbf{x}\middle|\mathbf{y}\right)\right\}$$

Mutual Information

- We shall define the mutual information of the RVs **x** and **y** as follows

$$I(\mathbf{x},\mathbf{y}) = H(\mathbf{x}) + H(\mathbf{y}) - H(\mathbf{x},\mathbf{y})$$

- I(**x**, **y**) can be written as an expected value

$$I(\mathbf{x},\mathbf{y}) = E\left\{\ln \frac{f(\mathbf{x},\mathbf{y})}{f(\mathbf{x})f(\mathbf{y})}\right\}$$

- Since f(x, y) = f($x|y$)f(y) it follows from the above that

$$I(\mathbf{x},\mathbf{y}) = H(\mathbf{x}) - H(\mathbf{x}|\mathbf{y}) = H(\mathbf{y}) - H(\mathbf{y}|\mathbf{x})$$

- The properties of entropy, developed before for arbitrary partitions, are obviously true for the entropy of discrete-type RVs and can be established as appropriate limits for continuous-type RVs.

Transformations of RVs

- We shall compare the entropy of the RVs **x** and **y** = g(**x**).

- If the RV **x** is of <u>discrete type</u>, then $\mathrm{H}(\mathbf{y}) \le \mathrm{H}(\mathbf{x})$
  with equality if and only if the transformation $y = \mathrm{g}(x)$ has a unique inverse $x = \mathrm{g}^{(-1)}(y)$.

- If the transformation $y = \mathrm{g}(x)$ has not a unique inverse (it is not one-to-one), then **y** = y_i for more than one value of **x**. This results in a reduction of H(**x**).

- If the RV **x** is of <u>continuous type</u>, then $\mathrm{H}(\mathbf{y}) \le \mathrm{H}(\mathbf{x}) + \mathrm{E}\{\ln|\mathrm{g}'(\mathbf{x})|\}$
  where g'(x) is the derivative of g(x). The equality holds if and only if the transformation $y = \mathrm{g}(x)$ has a unique inverse.

- Similarly, if $\mathbf{y}_i = \mathrm{g}_i(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $i = 1, \ldots, n$, are n functions of the RVs $\mathbf{x}_i$, then

$$\mathrm{H}(\mathbf{y}_1, \ldots, \mathbf{y}_n) \le \mathrm{H}(\mathbf{x}_1, \ldots, \mathbf{x}_n) + \mathrm{E}\{\ln|\mathrm{J}(\mathbf{x}_1, \ldots, \mathbf{x}_n)|\}$$

Entropy of Stochastic Processes

- The statistics of most stochastic processes are determined in terms of the joint density f(x_1, …, x_m) of the RVs $\mathbf{x}(t_1)$, …, $\mathbf{x}(t_m)$.

- The joint entropy of these RVs is the *m*th-order entropy of the process $\mathbf{x}(t)$

$$\mathrm{H}(\mathbf{x}_1,\ldots,\mathbf{x}_m) = \mathrm{E}\{-\ln \mathrm{f}(\mathbf{x}_1,\ldots,\mathbf{x}_m)\}$$

- This function equals the uncertainty about the above RVs and it equals the information gained when they are observed.

- In general, the uncertainty about the values of $\mathbf{x}(t)$ on the entire t axis or even on a finite interval, no matter how small, is infinite.

- However, we assume $\mathbf{x}(t)$ expressed in terms of its values on a countable set of points, then a rate of uncertainty can be introduced. It suffices, therefore, to consider only discrete-time processes $\mathbf{x}_n$.

Entropy of Stochastic Processes

- The mth-order entropy of a discrete-time process $\mathbf{x}_n$ is the joint entropy $\mathrm{H}(\mathbf{x}_1, \ldots, \mathbf{x}_m)$ of the m RVs: $\mathbf{x}_n, \mathbf{x}_{n-1}, \ldots, \mathbf{x}_{n-m+1}$

- We shall assume throughout that the process $\mathbf{x}_n$ is strict-sense stationary (SSS). In this case, $\mathrm{H}(\mathbf{x}_1, \ldots, \mathbf{x}_m)$ is the uncertainty about any m consecutive values of the process $\mathbf{x}_n$.

- The first-order entropy will be denoted by $\mathrm{H}(\mathbf{x})$ and equals the uncertainty about $\mathbf{x}_n$ for a specific n.

- Recalling the properties of entropy, we have:

$$\mathrm{H}(\mathbf{x}_1, \ldots, \mathbf{x}_m) \leq \mathrm{H}(\mathbf{x}_1) + \ldots + \mathrm{H}(\mathbf{x}_m) = m\,\mathrm{H}(\mathbf{x})$$

- <u>Example</u>: If the process $\mathbf{x}_n$ is *strictly white*, that is, if the RVs $\mathbf{x}_n, \mathbf{x}_{n-1}, \ldots$ are independent, then $$\mathrm{H}(\mathbf{x}_1, \ldots, \mathbf{x}_m) = m\,\mathrm{H}(\mathbf{x})$$

Entropy of Stochastic Processes

- The conditional entropy of order *m*, $H(\mathbf{x}_n|\mathbf{x}_{n-1}, \ldots, \mathbf{x}_{n-m})$, of a process $\mathbf{x}_n$ is the uncertainty about its present under the assumption that its *m* most recent values have been observed.

- Recalling that $H(\mathbf{x}|\mathbf{y}) \leq H(\mathbf{x})$, we can readily show that:

$$H\left(\mathbf{x}_n \middle| \mathbf{x}_{n-1}, \ldots, \mathbf{x}_{n-m}\right) \leq H\left(\mathbf{x}_n \middle| \mathbf{x}_{n-1}, \ldots, \mathbf{x}_{n-m-1}\right)$$

- Thus the above conditional entropy is a decreasing function of *m*. If, therefore, it is bounded from below, it tends to a limit. This is certainly the case if the RVs $\mathbf{x}_n$ are of discrete type because then all entropies are positive.

- The limit will be denoted by $H_c(\mathbf{x})$ and will be called the <u>conditional entropy</u> of the process $\mathbf{x}_n$:

$$H_c(\mathbf{x}) = \lim_{m \to \infty} H\left(\mathbf{x}_n \middle| \mathbf{x}_{n-1}, \ldots, \mathbf{x}_{n-m}\right)$$

- The function $H_c(\mathbf{x})$ is a measure of our uncertainty about the present of $\mathbf{x}_n$ under the assumption that its entire past is observed.

4. MAXIMUM ENTROPY METHOD

Maximum Entropy Method

- The ME method is used to determine various parameters of a probability space subject to given constraints.

- The resulting problem can be solved, in general, only numerically and it involves the evaluation of the maximum of a function of several variables.

- In a number of important cases, however, the solution can be found analytically or it can be reduced to a system of algebraic equations.

- We consider herein certain special cases, concentrating on constraints in the form of expected values.

- For most problems under consideration, the following inequality is used. If f(x) and φ(x) are two arbitrary densities, then it can be proven that:

$$-\int_{-\infty}^{\infty} \varphi(x) \ln \varphi(x)\, dx \le -\int_{-\infty}^{\infty} \varphi(x) \ln \mathrm{f}(x)\, dx$$

Example

- In the coin experiment, the probability of heads is often viewed as an RV **p** (bayesian estimation).

- We shall show that if no prior information about **p** is available, then, according to the ME principle, its density f(*p*) is <u>uniform</u> in the interval (0,1).

- In this problem we must maximize H(**p**) subject to the constraint (dictated by the meaning of **p**) that f(*p*) = 0 outside the interval (0, 1).

- The corresponding entropy is, therefore, given by $\mathrm{H}(\mathbf{p}) = -\int_0^1 \mathrm{f}(p) \ln \mathrm{f}(p)\, dp$ and our problem is to find f(*p*) such as to maximize the above integral.

- We maintain that H(**p**) is maximum if f(*p*) = 1, hence H(**p**) = 0.

- Indeed, if φ(*p*) is any other density such that φ(*p*) = 0 outside (0, 1), then

$$-\int_0^1 \varphi(p) \ln \varphi(p)\, dp \le -\int_0^1 \varphi(p) \ln \mathrm{f}(p)\, dp = 0 = \mathrm{H}(\mathbf{p})$$

Constraints as Expected Values

- We shall consider now a class of problems involving constraints in the form of expected values. Such problems are common in hydrology.

- We wish to determine the density f(x) of an RV **x** subject to the condition that the expected values η_i of n known functions g$_i$(x) of **x** are given

$$\mathrm{E}\{\mathrm{g}_i(\mathbf{x})\} = \int_{-\infty}^{\infty} \mathrm{g}_i(x)\,\mathrm{f}(x)\,dx = \eta_i \qquad i = 1,\ \ldots,\ n$$

- We shall show that the ME method leads to the conclusion that f(x) must be an <u>exponential</u>

$$\mathrm{f}(x) = A\exp\{-\lambda_1\,\mathrm{g}_1(x) - \ldots - \lambda_n\,\mathrm{g}_n(x)\}$$

- Where λ_i are n constants determined from the above equations E{g$_i$(**x**)} and A is such as to satisfy the density condition

$$A\int_{-\infty}^{\infty} \exp\{\lambda_1\,\mathrm{g}_1(x) - \ldots - \lambda_n\,\mathrm{g}_n(x)\}\,dx = 1$$

Proof

- Suppose that $\mathrm{f}(x) = A\exp\{-\lambda_1\,\mathrm{g}_1(x) - \ldots - \lambda_n\,\mathrm{g}_n(x)\}$

  In this case:

$$\int_{-\infty}^{\infty} \mathrm{f}(x)\ln \mathrm{f}(x)\,dx = \int_{-\infty}^{\infty} \mathrm{f}(x)\left[\ln A - \lambda_1\,\mathrm{g}_1(x) - \ldots - \lambda_n\,\mathrm{g}_n(x)\right]dx$$

- Hence: $\mathrm{H}(\mathbf{x}) = \lambda_1\eta_1 + \ldots + \lambda_n\eta_n - \ln A$

- Now it suffices, therefore, to show that, if φ(x) is any other density satisfying the constraints E{g_i(**x**)}, then its entropy cannot exceed the right side of the above equation

$$-\int_{-\infty}^{\infty} \varphi(x)\ln\varphi(x)\,dx \le -\int_{-\infty}^{\infty} \varphi(x)\ln \mathrm{f}(x)\,dx$$

$$= \int_{-\infty}^{\infty} \varphi(x)\left[\lambda_1\,\mathrm{g}_1(x) + \ldots + \lambda_n\,\mathrm{g}_n(x) - \ln A\right]dx$$

$$= \lambda_1\eta_1 + \ldots + \lambda_n\eta_n - \ln A$$

Stochastic Processes

- The ME method can be used to determine the statistics of a stochastic process subject to given constraints.
- Suppose that $\mathbf{x}_n$ is a wide-sense stationary (WSS) process with autocorrelation $R[m] = E\{\mathbf{x}_{n+m}\, \mathbf{x}_n\}$.
- We wish to find its various densities assuming that $R[m]$ is specified either for some or for all values of m.
- The ME principle leads to the conclusion that, in both cases, $\mathbf{x}_n$ must be a normal process with zero mean. This completes the statistical description of $\mathbf{x}_n$ if $R[m]$ is known for all m.
- If, however, we know $R[m]$ only partially, then we must find its unspecified values. For finite-order densities, this involves the maximization of the corresponding entropy with respect to the unknown values of $R[m]$ and it is equivalent to the maximization of the correlation determinant Δ.

5. CONCLUSIONS

Conclusions

- Entropy is a valuable tool to provide a quantitative measure of uncertainty of stochastic modelling of natural processes.

- An important application of entropy is the determination of the statistics of a stochastic process subject to various constraints, with the maximum entropy (ME) method.

- We should emphasize, however, that as in the case of the classical definition of probability, the conclusions drawn from the ME method must be accepted with skepticism particularly when they involve elaborate constraints.

- Extremal entropy considerations may provide an important connection with statistical mechanics. Thus, the ME principle may provide a physical background in the stochastic representation of natural processes.

Thank you for your attention

"Von Neumann told me, 'You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name. In the second place, and more important, no one knows what entropy really is, so in a debate you will always have the advantage."

Claude Elwood Shannon