The role of information in statistical mechanics

Derivations and Justifications of Jaynes' Principle of Maximum Entropy

Simone Balmelli

Contents

Jayne’s justification

Shannon:

a “measure of uncertainty” should satisfy some given axioms.

Only the entropy (up to a factor) satisfies them.

Jayne’s intuitive conclusion:

maximizing entropy with respect to our information reflects in the best possible way our uncertainty of the system.

Criticism

- An intuitive justification is insufficient.
- Why not maximize some other function?
- A formal derivation is needed.

Definitions and notation (1)

Consider a system with n possible states x_i, and define $N := \{x_i\}$.

$\mathcal{D}$ is the set of all discrete probability distributions with n components, i.e.,

$$\mathbf{p} \in \mathcal{D} \Rightarrow 0 \leq p_i \leq 1, \sum_{i=1}^{n} p_i = 1$$

$\mathbf{p}^\dagger \in \mathcal{D}$ the unknown "true" probability distribution.

For a subset $S \subseteq N$ and $\mathbf{p} \in \mathcal{D}$ let be $(\mathbf{p} * S)$ the conditional probability distribution of $\mathbf{p}$ given S, i.e.,

$$(\mathbf{p} * S)_i := \begin{cases} p_i / \sum_{x_j \in S} p_j & \text{if } x_i \in S \\ 0 & \text{if } x_i \notin S \end{cases}$$

Definitions and notation (2)

Let us assume that our information is provided in constraints:

$$\sum_{k=1}^{n} a_{ik} p_k^\dagger = 0 \qquad \sum_{k=1}^{n} c_{jk} p_k^\dagger \geq 0$$

The elements of $\mathcal{D}$ obeying these constraints define a subset $\mathcal{I} \subseteq \mathcal{D}$. We denote the information by

$$I := (\mathbf{p}^\dagger \in \mathcal{I})$$

Definitions and notation (3)

Our goal is to find a posterior distribution, that we denote by

$$\circ I \in \mathcal{D}$$

where the posterior is obtained by maximizing some function

$$H(\mathbf{p})$$

taking the constraints in to account, that is

$$H(\circ I) = \max_{\mathbf{p} \in \mathcal{I}} H(\mathbf{p})$$

Towards the axioms

What is it required from a method of inductive inference to make sense?

The answer here:

different ways of considering the problem with respect to the same information should lead to the same result.

Axiom 1 (Uniqueness)

The posterior $\circ I$ is unique.

Axiom 2 (Permutation invariance)

Let be π a permutation of the elements of $\{1, ..., n\}$, and define:

$$\pi\mathbf{p} := \left(p_{\pi^{-1}(1)}, ..., p_{\pi^{-1}(n)}\right)$$
$$\pi I := \left(\pi\mathbf{p}^{\dagger} \in \pi\mathcal{I}\right)$$

Then for any π it holds:

$$\circ\left(\pi I\right) = \pi\left(\circ I\right)$$

Axiom 3 (System independence)

Let $I_1 = \left(\mathbf{p}_1^\dagger \in \mathcal{I}_1\right)$ and $I_2 = \left(\mathbf{p}_2^\dagger \in \mathcal{I}_2\right)$ be informations about two systems.

Then it holds

$$\circ\left(I_1 \wedge I_2\right) = \left(\circ I_1\right)\left(\circ I_2\right)$$

Axiom 4 (Subset independence)

$S_1, ..., S_n$ disjoint sets whose union is N.

I_i information about the conditional distribution $\mathbf{p} * S_i$.

M information giving the probability m_i of having each subset S_i .

Then it holds

$$(\circ (I_1 \wedge ... \wedge I_n \wedge M)) * S_i = (\circ I_i)$$

Theorems

Theorem 1. Let $H(\mathbf{p})$ satisfy uniqueness, permutation invariance and subset independence. Then it is equivalent to a function of the form

$$\sum_i f(p_i).$$

Theorem 2. Let $H(\mathbf{p})$ satisfy uniqueness, permutation invariance, system independence and subset independence. Then it is equivalent to the function

$$-\sum_i p_i \log(p_i).$$

Theorem 3. The entropy $-\sum_i p_i \log(p_i)$ satisfies the four axioms.

Towards the proof

Lemma 1. Let the assumptions of Axiom 4 hold, and $\mathbf{p} = \circ(I \wedge M)$. Let $x_j \in S_i$ and $x_k \notin S_i$. Then p_j is independent of p_k and n.

Lemma 2. Let $H(\mathbf{p})$ satisfy permutation invariance. Then it is equivalent to a symmetric function of the n variables $p_1, \ldots, p_n$.

Theorems

Theorem 1. Let $H(\mathbf{p})$ satisfy uniqueness, permutation invariance and subset independence. Then it is equivalent to a function of the form

$$F(\mathbf{p}) = \sum_i f(p_i)$$

.

Theorem 2. Let $H(\mathbf{p})$ satisfy uniqueness, permutation invariance, system independence and subset independence. Then it is equivalent to the function

$$-\sum_i p_i \log(p_i).$$

Theorem 3. The entropy $-\sum_i p_i \log(p_i)$ satisfies the four axioms.

Comment on the axiomatic derivation

• General problem in statistical inference: how do we have to interpret the posterior, and how is it related with the “true” distribution?

•If we require some intuitive conditions on the treatement of information (i.e. the axioms), then Jayne’s principle is the unique consistent method of statistical inference.

• The information about the probability distributions is intended to exist *a priori*. But this can be controversial in the applications of the principle.

Information *a priori*

Suppose that Gino is telling you:

“I have chosen a probability distribution, and I tell you only its expectation value. Try to estimate the distribution!”

Then Jayne’s principle could be, on the basis of the axiomatic derivation, reasonably retained the best method.

Is this a common situation?

The constraint rule problem

Suppose now you have a dice, and you would like to estimate its probability distribution.

You make some repeated experiments and you measure then the average.

Does it make sense to set the average equal to the expectation value, in order to apply Jayne’s principle?

An example: Jaynes' principle vs Bayes' method (1)

Let be y_i the result of the i-th throw, and N the total number of throws.

Suppose that $\frac{1}{N}\sum_{i=1}^{N} y_i = 6$

Jaynes' principle yields $(p_1, ..., p_6) = (0, 0, 0, 0, 0, 1)$

This could seem plausible only if N is large.

Parenthesis: the Bayes' method (1)

Suppose that each probability distribution is equiprobable.

Let be $\phi(p_1, ..., p_n)$ the uniform probability density on the set $\mathcal{D}$.

We denote by N_i the number of tries that have given the result x_i. It must hold

$$\sum_i N_i = N$$

Parenthesis: the Bayes' method (2)

The Bayes' calculations are made as follows:

$$P(N_1, ..., N_n) = \int ... \int P(N_1, ..., N_n | p_1, ..., p_n) \phi(p_1, ..., p_n) dp_1 ... dp_n$$

$$\phi(p_1, ..., p_n | N_1, ..., N_n) = \frac{P(N_1, ..., N_n | p_1, ..., p_n) \phi(p_1, ..., p_n)}{P(N_1, ..., N_n)}$$

$$P\left(y_{N+1} = x_i | N_1, ..., N_n\right) = \int ... \int p_i \phi(p_1, ..., p_n | N_1, ..., N_n) dp_1 ... dp_n$$

$$P\left(y_{N+1} = x_i | \frac{1}{N} \sum_{j=1}^{n} x_j N_j = a\right) = \frac{\sum_{\frac{1}{N}\sum_{j=1}^{n} x_j N_j = a} P(y_{N+1} = x_i | N_1, ..., N_n) P(N_1, ..., N_n)}{P\left(\frac{1}{N}\sum_{j=1}^{n} x_j N_j = a\right)}$$

An example:
Jaynes' principle vs Bayes' method (2)

Recall: the Jaynes' solution in the case $\frac{1}{N}\sum_{i=1}^{N} y_i = 6$

is $(p_1, ..., p_6) = (0, 0, 0, 0, 0, 1)$

Bayes' method of inverse probability yields:

$$p_1 = ... = p_5 = \frac{1}{N+6}, p_6 = \frac{N+1}{N+6}$$

An example:
Jaynes' principle vs Bayes' method (3)

Suppose now: $\frac{1}{N}\sum_{i=1}^{N} y_i = 3.5$

Jaynes yields: $p_1 = \ldots = p_6 = \frac{1}{6}$

Bayes:

	p_1, p_6	p_2, p_5	p_3, p_4
$N = 2$	0.1667	0.1667	0.1667
$N = 4$	0.1500	0.1667	0.1833
$N = 20$	0.1440	0.1658	0.1901
$N = 30$	0.1432	0.1658	0.1909
$N = 60$	0.1423	0.1658	0.1919

Conclusions

Jaynes' principle is the unique method of statistical inference which is consistent with our intuition.

The problem is that it needs an information *a priori* about the "true" probability distribution in order to be applied.

Any information provided by a measurement is only an estimation of the "true" information.

This could lead to results which don't seem plausible.