Three Questions of a Linguist to a Gibbsian

Pierre Collet

Centre de Physique Theorique
CNRS UMR 7644
Ecole Polytechnique
F-91128 Palaiseau Cedex (France)
Email: collet@cpht.polytechnique.fr

This file is best viewed in full screen mode.

Special Thanks to Tony Kroch for the questions

Question 1

When you describe the Gibbs probability measures you always take the example of a gas. This has spatial extension, we need time extension.

Geographical phenomena are neglected.

Answer

The two situations are indeed very similar.

Answer

The two situations are indeed very similar.

- In space we have to consider all the possible configurations of molecules (the set of positions and velocities of all the molecules in the room).

Answer

The two situations are indeed very similar.

- In space we have to consider all the possible configurations of molecules (the set of positions and velocities of all the molecules in the room).
- In time we have to consider all the grammatically correct possible sequences.

Answer

The two situations are indeed very similar.

- In space we have to consider all the possible configurations of molecules (the set of positions and velocities of all the molecules in the room).
- In time we have to consider all the grammatically correct possible sequences.

A common goal is to put some adequate probability distribution on these different (numerous) possibilities.

Both have constraints

For example:

- Molecules do not overlap
- The order OVS does not exits in French.

Both have constraints

For example:

- Molecules do not overlap
- The order OVS does not exits in French.

Another analogy

For molecules we have the energy function of a configuration. In linguistics we can define a cost function for a sentence (OT, Sotaq).

An example of successful analogy.

In the ’70 Sinai, Ruelle and Bowen applied Gibbs formalism to time evolution (dynamical systems) and constructed what is now called SRB measures.

Apparent difference:

space is dimension 3, time is dimension 1.

This is not important for fundamental concepts. But has important consequences: you cannot boil water in dimension 1 (no phase transition for short range interactions).

Question 2

You are always worried about the thermodynamic limit (infinite volume). But this room is finite and speech signals are finite also.

Answer

It is all a matter of scales.
Molecules are very tiny, they interact only with their neighbors. Only those near the wall feel it.

Answer

It is all a matter of scales.
Molecules are very tiny, they interact only with their neighbors. Only those near the wall feel it.
In the time direction, the correlations are due to memory effects.

Answer

It is all a matter of scales.
Molecules are very tiny, they interact only with their neighbors. Only those near the wall feel it.
In the time direction, the correlations are due to memory effects.
If the duration of the signal is much longer than the forgetting time one can assume an infinite signal to compute statistics.

Answer

It is all a matter of scales.
Molecules are very tiny, they interact only with their neighbors. Only those near the wall feel it.
In the time direction, the correlations are due to memory effects.
If the duration of the signal is much longer than the forgetting time one can assume an infinite signal to compute statistics.
This is not true if the signal is too small or near the ends.

Question 3

When you write Gibbs measures the cost function does not seem to distinguish between past and future. This seems to break causality. What about the arrow of time?

Answer

The cost function associated to a Gibbs measure is not unique.

Answer

The cost function associated to a Gibbs measure is not unique.
Given any cost function in dimension one (depending on past and future), there is another cost function depending only on the past giving the same probability Gibbs measure (same statistics).
D.Ruelle

Answer

The cost function associated to a Gibbs measure is not unique.
Given any cost function in dimension one (depending on past and future), there is another cost function depending only on the past giving the same probability Gibbs measure (same statistics).
D.Ruelle

However in order to determine if a sentence is grammatically acceptable or not, near future may play a role.

The problem of time reversibility has been important since the beginning of statistical mechanics (Boltzmann, second law of thermodynamics).

Is time going backward the same as time going forward?

There is a difference with space: I can walk in both directions.
In the talk of E.Dupoux we heard samples of backward speech. They are not recognized by babies and by monkeys.

We have to distinguish two concepts:

We have to distinguish two concepts:

1) Kinematic reversibility. Is any admissible sentence also admissible when run backward? (see Kadiwéu).

We have to distinguish two concepts:

1) Kinematic reversibility. Is any admissible sentence also admissible when run backward? (see Kadiwéu).
2) Statistical reversibility. If Kinematic reversibility holds, are frequencies the same?

We have to distinguish two concepts:

1) Kinematic reversibility. Is any admissible sentence also admissible when run backward? (see Kadiwéu).

2) Statistical reversibility. If Kinematic reversibility holds, are frequencies the same?

If 1) and 2) hold, you cannot distinguish the arrow of time by looking at a typical emission.
There are many exciting recent developments when 1) holds but not 2) related to entropy creation.

1) and 2) hold for reversible Markov chains.

Question 4

Why are you always talking of Gibbs measures? is there nothing else in life ?

Question 4

Why are you always talking of Gibbs measures? is there nothing else in life ?

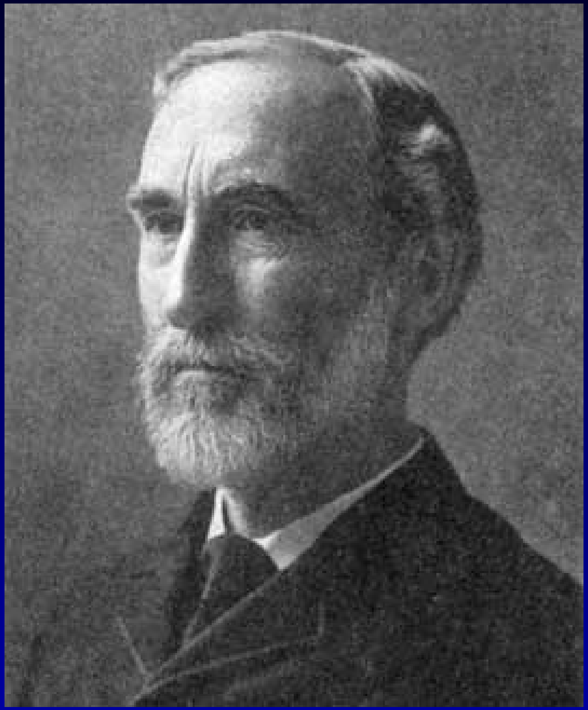
Well, let's see.

Answer

Gibbs measures are the most natural ones.

Nature prefers Gibbs measures.

The Founders



Josiah Willard Gibbs

1839-1903

Ludwig Boltzmann

1844-1906

Pictures from:

`http://www-groups.dcs.st-andrews.ac.uk/~history/Mathematicians/`

Once upon a time you were a prisoner of the trifaces, condemned to be beheaded. Their money is tricoins with three faces: head, bosom and tail, and the guards play all the day hbt (head bosom or tail). To try to save your life you propose to astonish them by making a deal. They will play a thousand times and record carefully the outcome. Head is valued 3 points, bosom two and tail one. You promised that if they give you the average number of points you will give them the number of tails. If you guess within 20 you are free else they execute you.

The average is 3.5, what is your guess for the number of tails?

The average is 3.5, what is your guess for the number of tails?

Beware that your life depends on your answer!

The situation

There are three numbers: N_1 the number of heads, N_2 the number of bosoms and N_3 the number of tails. We have

$$N_1 + N_2 + N_3 = N$$

and

$$N_1 + 2N_2 + 3N_3 = \alpha N$$

where N is the total number of throws (1000) and α the average (2.5).

THIS IS THE ONLY INFORMATION AVAILABLE.

The deficiency

Note that for a fair tricoin you have $N_1 \approx N_2 \approx N_3 \approx N/3$ and hence $\alpha = 2$. Their tricoin is heavily loaded, what do you expect from these strange people anyway?

The deficiency

Note that for a fair tricoin you have $N_1 \approx N_2 \approx N_3 \approx N/3$ and hence $\alpha = 2$. Their tricoin is heavily loaded, what do you expect from these strange people anyway?
You have to guess three numbers N_1, N_2, N_3 but have only two constraints. This is not enough, the problem has many solutions (even much more than you think as we will see below). Yet, only one of these came out!

The maxentropy argument

Only one sequence was observed $(1, 3, 3, 2, 1, 2.....)$ but you do not know it. The observed sequence has N_1 ones and if you knew this number you would immediately derive $N_2 = (3 - \alpha)N - 2N_1$ and $N_3 = (\alpha - 2)N + N_1$ using the two previous constraints. Still, for a given N_1 there are many sequences of 1, 2 and 3 which are possible under the constraints. How many?

The maxentropy argument

Only one sequence was observed $(1, 3, 3, 2, 1, 2.....)$ but you do not know it. The observed sequence has N_1 ones and if you knew this number you would immediately derive $N_2 = (3 - \alpha)N - 2N_1$ and $N_3 = (\alpha - 2)N + N_1$ using the two previous constraints. Still, for a given N_1 there are many sequences of 1, 2 and 3 which are possible under the constraints. How many?

$$W(N_1) = \frac{N!}{N_1!N_2!N_3!}$$

recall that $n! = n(n-1)(n-2)\cdots 1$, N_2 and N_3 should be replaced by their above expression.

Where there is an idea

Since we do not know N_1 all possibilities can occur,
AND SINCE WE HAVE NO MORE INFORMATION,
THEY ARE ALL EQUIVALENT!
The total number of sequences compatible with the two constraints (information) is

$$W = \sum_{N_1} W(N_1) \ .$$

Note that since all numbers must be non negative and less than N we must have $1 \leq \alpha \leq 3$ and $0 \leq N_1 \leq (3-\alpha)N/2$.
This sum has many terms (here 251), how is it organized?

Are all the terms about equal, is there a dominant term?

is it completely random?

Stirling to the rescue

For large n, $n!$ is a huge and complicated number to calculate, but we have Stirling's formula

$$\log(n!) = n \log n - n + \frac{1}{2} \log n + \text{a rest of order one} \ .$$

We will only keep the first two terms (dominant ones) and apply this to $W\left(N_1\right)$.

We get using the constraint on the sum

$$\log\left(W\left(N_1\right)\right) = N\log N - N_1\log N_1 - N_2\log N_2 - N_3\log N_3 \ .$$

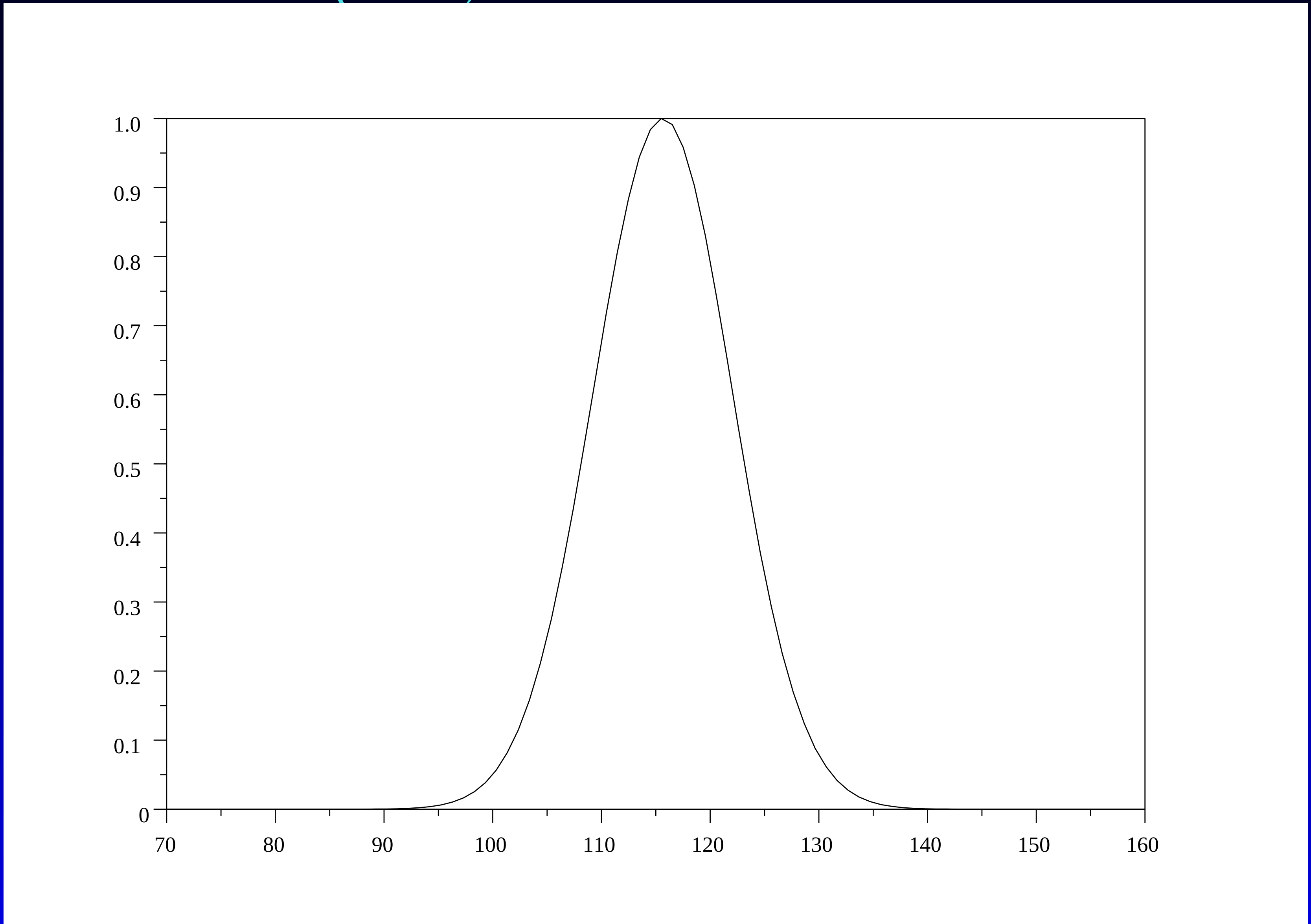
To see if there is maximum we can take the derivative with respect to N_1 after having replaced N_2 and N_3 by their value, we get

$$\frac{d\ \log\left(W\left(N_1\right)\right)}{d\,N_1} = -\log N_1 + 2\log N_2 - \log N_3 = \log\left(\frac{N_2^2}{N_1 N_3}\right) \ .$$

So we have a maximum (derivative is zero) for

$$N_1 N_3 = N_2^2 \ .$$

Plot of $W(N_1)$



Plot of $W(N_1)/W(N_1^*)$ as a function of N_1.

Solving for N_1 the equation

$$N_1 N_3 = N_2^2 \qquad \text{gives} \qquad N_1^* = 116.20406$$

This is the value of N_1 with the largest number of possibilities, i.e. the largest number $W(N_1)$. Moreover, if we depart substantially from this value, the number of sequences decreases very fast. Even more, the total number of sequences corresponding to $|N_1 - N_1^*| > 20$ is negligible with respect to the total number of possibilities: this is proportional to the surface below the curve in this domain.

Therefore,your best guess
IN THE ABSENCE OF OTHER INFORMATION
is

$$N_1 = [N_1^*] = 116$$

In the absence of any other information, 116 is the most likely choice.
It is the one realized by the most numerous equivalent possibilities.

It is the only answer which does not assume any other (subjective) information.
Any other answer assumes consciously or not another constraint.
There is NO REASON to give another answer.

Nature seems to like maximum entropy solutions because it realizes the answer in the largest possible equivalent ways.

If the measurement is different enough, it strongly suggests the existence of another constraint. Find it and redo the computation including that constraint. See the articles by Jaynes for more:

`http://bayes.wustl.edu`

Gibbs distribution

An extension of the above argument leads to the Gibbs distribution on the sequences $\mathcal{S}$

$$\mathbf{P}(\mathcal{S}) = \frac{e^{\lambda C(\mathcal{S})}}{Z}$$

where C is our constraint

$$C(\mathcal{S}) = N_1(\mathcal{S}) + 2N_2(\mathcal{S}) + 3N_3(\mathcal{S})$$

and Z is a normalization factor. λ is determined by the equation

$$N\alpha =< C > \quad .$$

This is a general situation. A probability satisfies the maximum entropy principle for a sequence of constraints $C_1, \cdots, C_p$ with averages $\alpha_1, \cdots, \alpha_p$ if and only if it is a Gibbs measure of the form

$$\mathbf{P}(X) = \frac{e^{\lambda_1 C_1(X) + \cdots + \lambda_p C_p(X)}}{Z}$$

where the numbers $\lambda_1, \cdots, \lambda_p$ are obtained by solving the equations

$$N\alpha_1 =< C_1 >, \cdots, N\alpha_p =< C_p > \quad .$$

Warning

If you have more information (for example the entire sequence) use it! You will get a more confident statistics. The maximum entropy principle applies if you only have partial information of a special kind: averages.

For a Physical approach see G.E. Uhlenbeck, G.W. Ford. *Lectures in Statistical Mechanics I.* Amer. Math. Soc. Providence 1963.

For a mathematical derivation see O.E.Lanford III. Entropy and Equilibrium States in Classical Statistical Mechanics. In *Statistical Mechanics and Mathematical Problems.* A.Lenard editor; Lecture Notes in Physics **20**. Springer-Verlag, Berlin-Heidelberg-New York 1973.

The End