1. Questions branching off *Information Theory and Statistical Mechanics I*, Jaynes 1957, as printed in ?

Halfway through the paper, on page 11 of ?, Jaynes describes the following problem, to illustrate how, with a large number of degrees of freedom, mean-value constraints can yield incredibly tight maximum entropy distributions.

Suppose you define the following series of state values x, for small $\epsilon > 0$, and natural n.

$$\begin{aligned} x_1^{n+1} &= \epsilon \\ x_{i+1} - x_i &= \frac{\epsilon}{x_i^n} \end{aligned}$$

or, put another way,

$$x_{i+1} = (1+ \frac{\epsilon}{x_i^{n+1}})x_i$$

I claim that this series grows as follows (representative ratios are given for $n = 1$)

$$\begin{aligned} \frac{x_1}{x_1} &= 1 \\ \frac{x_2}{x_1} &= 2 \\ \frac{x_3}{x_1} &= 2 + \frac{1}{2^n} = 2.5 \\ \frac{x_4}{x_1} &= 2 + \frac{1}{2^n} + \frac{1}{(2+\frac{1}{2^n})^n} = 2.9 \\ \frac{x_5}{x_1} &= 2 + \frac{1}{2^n} + \frac{1}{(2+\frac{1}{2^n})^n} + \frac{1}{(2+\frac{1}{2^n}+\frac{1}{(2+\frac{1}{2^n})^n})^n} = 3.24 \end{aligned}$$

and so on. So intuitively you see that the terms are growing, but that they will get closer together as n increases. This also means that if you pick ϵ small enough, you can make the point density in x_i as dense as you want, wherever you want, for a fixed n. The easiest way to see this is to just plot the terms x_i, which you calculate iteratively.

Jaynes mentions that the series is unbounded above. I had a hard time seeing this; yes, the terms are growing, but the rate of growth slows down dramatically, especially for larger n. Shouldn't this plateau? Well, let's suppose it does.

Suppose $\forall \epsilon \in \mathbb{R}^+ \; \forall n \in \mathbb{N} \; \exists M \in \mathbb{R}^+ : \forall i \in \mathbb{N} \; x_i < M$. That is, suppose the series x_i has an upper bound M for any choice of parameters ϵ and n. If this is the case, we know $\forall i$, $\frac{x_{i+1}}{x_i} = 1 + \frac{\epsilon}{x_i^{n+1}} > 1 + \frac{\epsilon}{M^{n+1}} > 1$. Now, we can write

$$x_{i+1} = \left(\frac{x_{i+1}}{x_i}\right)\left(\frac{x_i}{x_{i-1}}\right)\cdots\left(\frac{x_2}{x_1}\right)(x_1) > \left(1 + \frac{\epsilon}{M^{n+1}}\right)^i x_1$$

and since we know $1 + \frac{\epsilon}{M^{n+1}} > 1$, we know that the series x_i diverges with i. This contradicts the assumption that M was an upper bound for the series, and so we conclude that the series is unbounded above.

This derivation was unsatisfying to me. I think the best way to see why the series has to be unbounded is to say "give me a bound M, and I can give you an upper bound i on how long it will take the series to pass M." This series warrants more thought.

Anyways, at this point we have a series defined on $(0, \infty)$ that can get as close to zero as you wish, since you can choose ϵ, and which densely samples the domain in a way you can control, with n. So Jaynes asks us to consider the summation

$$\sum_i f(x_i) = \sum_i f(x_i) \frac{x_{i+1} - x_i}{x_{i+1} - x_i} = \sum_i f(x_i) \frac{x_i^n}{\epsilon}(x_{i+1} - x_i) \approx \int_0^\infty f(x) \frac{x^n}{\epsilon} dx$$

which he approximates by an integral, though he says this is not necessary.

Now we get to the point of this setup. Suppose the x are actually the allowed states of your system. I'm still not sure what n is. Let's say we are given $\langle x \rangle = \sum_i x_i p_i$, the average x, presumably associated with a macroscopic measurement. Our job is to infer the probabilities p_i and estimate $\langle x^2 \rangle$. Jaynes asserts that the best inference we can make is to choose a maximum entropy distribution for p_i, under the mean-value constraint on x and the normalization constraint on the distribution. He defines a Lagrangian

$$\mathcal{L} = -\sum_i p_i \ln p_i + \mu \left(1 - \sum_i p_i \right) + \lambda \left(\langle x \rangle - \sum_i x_i p_i \right)$$

with constraints given by $\frac{\partial \mathcal{L}}{\partial \mu} = 0$ and $\frac{\partial \mathcal{L}}{\partial \lambda} = 0$. He solves $\forall i$, $\frac{\partial \mathcal{L}}{\partial p_i} = 0$ to find

$$p_i = e^{-1-\mu} e^{-\lambda x_i}$$

and applies the normalization constraint $\frac{\partial \mathcal{L}}{\partial \mu} = 1 - \sum_i p_i = 0$.

$$\begin{aligned} \sum_i e^{-1-\mu} e^{-\lambda x_i} &= 1 \\ e^{-1-\mu} &= \tfrac{1}{\sum_i e^{-\lambda x_i}} = \frac{1}{Z(\lambda)} \end{aligned}$$

We now have a properly normalized distribution

$$p_i = Z(\lambda)^{-1} e^{-\lambda x_i}$$

and our job is find λ. We notice that

$$-\frac{\partial \ln Z(\lambda)}{\partial \lambda} = \sum_i x_i p_i = \langle x \rangle$$

so if we could just calculate a closed form for $Z(\lambda)$ we'd be set. Here's where the integral approximation above comes in.

$$Z(\lambda) = \sum_i e^{-\lambda x_i} \approx \frac{1}{\epsilon} \int_0^\infty e^{-\lambda x} x^n dx$$

which can be straightforwardly evaluted as follows:

$$\int_0^\infty e^{-\lambda x} x^n dx = \int_0^\infty dx \frac{d^n}{d(-\lambda)^n} e^{-\lambda x} = \frac{d^n}{d(-\lambda)^n} \int_0^\infty dx e^{-\lambda x} = \frac{d^n}{d(-\lambda)^n} \left(\frac{1}{\lambda} \right) = \frac{n!}{\lambda^{n+1}}$$

where we are assuming uniform continuity of the integral for reasons I don't yet understand. Anyways, now we know that

$$Z(\lambda) \approx \frac{n!}{\epsilon \lambda^{n+1}}$$

and we immediately calculate

$$-\frac{\partial \ln Z(\lambda)}{\partial \lambda} = \langle x \rangle = \frac{n+1}{\lambda}$$

to solve

$$\lambda = \frac{n+1}{\langle x \rangle}$$

finally specifying our estimated distribution over x in full. At this point we have noticed that our final distribution looks a lot like the Boltzmann distribution. It would be worth trying to interpret this λ in the context of statistical mechanical/thermodynamic temperature, i.e.

$$\beta \propto \frac{1}{T} = \frac{\partial S}{\partial E}$$

versus our

$$\lambda = \frac{n+1}{\langle x \rangle} = \frac{effectived.o.f.?}{averagestate}$$

Needless to say, I'm not clear on this point. Anyways, now that we have the distribution over x inferred from the constraint on $\langle x \rangle$ we can easily estimate things like

$$\langle x^2 \rangle \{\langle x \rangle\} \approx \frac{Z^{-1}}{\epsilon} \int_0^\infty x^{n+2} e^{-\lambda x} dx = \frac{n+2}{n+1} \langle x \rangle^2$$

so we can also inductively deduce, for example, how much jitter we expect in future measurments of $\langle x \rangle$

$$Var_{Estimated}(x) = \langle x^2 \rangle_{MaxEnt} - \langle x \rangle^2_{Observed} = \frac{\langle x \rangle^2}{n+1}$$

Here, we simply note that the estimated standard deviation drops as $O(\frac{1}{\sqrt{n}})$, so if n is appreciably large, our estimated state distribution is going to be incredibly tight. I'm still not sure why n is playing the role of 'effective degrees of freedom' so well.

Also, it's interesting that if we calculate the actual maximized value of the entropy, we find that

$$\begin{aligned} S_{max} &= -\sum_i p_i \ln p_i \\ &= \frac{1}{Z} \sum_i e^{-\lambda x_i} (\lambda x_i + \ln Z) \\ S_{max} &= \lambda \langle x \rangle + \ln Z \end{aligned}$$

which could be more simply written as

$$S_{max} = n + 1 + \ln Z(\lambda(\langle x \rangle))$$

Calculating derivates via either representation should give us the same result.

$$\frac{\partial S}{\partial \lambda} = \langle x \rangle - \frac{n+1}{\lambda} - \langle x \rangle = -\frac{n+1}{\lambda} = -\langle x \rangle$$

If you don't consider λ and $\langle x \rangle$ to be functions of each other, you'd get 0 as the result. Not too sure when things are allowed to pretend to be independent; this is like the difference between Hamiltonian and Lagrangian mechanics all over again.

Similarly, we can calculate

$$\begin{aligned} \frac{\partial S}{\partial \langle x \rangle} &= \frac{\partial}{\partial \langle x \rangle} [\langle x \rangle \lambda(\langle x \rangle)] + \frac{\partial \ln Z}{\partial \lambda} \frac{\partial \lambda}{\partial \langle x \rangle} \\ &= \lambda - \frac{n+1}{\langle x \rangle} + \frac{n+1}{\langle x \rangle} = \lambda - \lambda + \lambda = \lambda \end{aligned}$$

It'd sure be worth trying to understand why the dependence on $\langle x \rangle$ cancels out so nicely, and why the two derivatives have opposite signs. Still, this is really neat! If this were the Boltzmann distribution, and $x \sim E$ you'd see straight off that $\frac{\partial S}{\partial \langle E \rangle} = \lambda$ implied $\lambda \sim g(T)$, where g is invertible (since thermodynamic temperature is the qualitative measure of equilibrium, whereas $\frac{\partial S}{\partial \langle E \rangle}$ can be shown to be the statistical indicator of equilbrium for energy-exchanging systems).

I got that invertible g thing from a text-in-progress that I still need to cite. Also, you should note that what we've done here is just as much for a toy system as any of the other examples out there. Whether you use an ideal gas to show why λ must be related to T, or you use this x, n system, it's no different. We're drawing the connection only because we were able to find a functional relationship between λ and the average constraint that was given. Is it possible to do it more generally? That is, show $\langle x \rangle$ and λ are conjugate (is that right?) without explicitly knowing how they are related? Think about Rob's polynomial approach to solving for λin the case with finite, discrete energies.

Our next task is to try the inference the other way around. That is, given $\left\langle x^2 \right\rangle$ estimate $\langle x \rangle$. Also, we would like to have a feeling for how sure we are about these estimates.