Why maximize entropy?

TSILB*

Version 1.0, 19 May 1982

It is commonly accepted that if one is asked to select a distribution satisfying a bunch of constraints, and if these constraints do not determine a unique distribution, then one is best off picking the distribution having maximum entropy. The idea is that this distribution incorporates the least possible information. Explicitly, we start off with no reason to prefer any distribution over any other, and then we are given some information about the distribution, namely that it satisfies some constraints. We want to be as conservative as possible; we want to extract as little as possible from what we have been told about the distribution; we don't want to jump to any conclusions; if we are going to come to any conclusions we want to be forced to them.

Lying behind this conservative attitude there is doubtless an Occam's razor kind of attitude: We tend to prefer, in the language of the LSAT, 'the narrowest principle covering the facts'. There is also an element of sad experience: We easily call to mind a host of memories of being burned when we jumped to conclusions.

For some time I have had the idea of making this latter feeling precise, by interpreting the process of picking a distribution satisfying a bunch of constraints as a strategy in a game we play with God. God tells us the constraints, we pick a distribution meeting those constraints, and then we have to pay according to how badly we did in guessing the distribution. The maximum entropy distribution should be our optimal strategy in this game.

*This Space Intentionally Left Blank. Contributors include: Peter Doyle.

Last night I recognized for the first time what the rules of this game with God would have to be, or rather one possible set of rules—perhaps there are other possibilities. I haven't yet convinced myself that these are the only natural rules for such a game, or even that they are all that natural. In thinking about these rules, the important question will be: Is this game just something that was rigged up to justify the maximum entropy distribution? After all, any point is the location of the maximum of some function. Does the statement that choosing the maximum entropy distribution is the optimal strategy in this game have any real mathematical content? (I purposely say 'mathematical' rather than 'philosophical', from the prejudice that once can never have the latter without the former. 'Except as a man handled an axe, he had no way of knowing a fool.') Obviously I think the answers to these questions are favorable in the case of the game I'm proposing, but I haven't taken the time to think through them carefully yet. (Added later: Now that I've finished writing this I'm much more confident.)

IDEA OF THE GAME: We are told the constraints, we pick a distribution, God gets to pick the 'real' distribution, satisfying the constraints of course, some disinterested party picks an outcome according to the 'real' distribution that God has just picked, and we have to pay according to how surprised we are to see that outcome.

Of course the big question is, how much do we have to pay? The big answer is, the log of the probability we assigned to the outcome. Actually, it is better to have us pay

$$-\log(p_i/(1/n)),$$

where p_i is the probability we assigned to the point that got picked (let's call outcomes 'points'), and n is the total number of possible outcomes. To put it more positively, we *get paid*

$$\log(p_i/(1/n)),$$

the log of the factor by when we changed the weight of the point that got picked from the value it is given by the uniform distribution. A big factor means 'I thought so'; a little factor means 'Fooled me!'. We choose this factor rather than the new weight itself so that if we start with a non-uniform a priori distribution the theory will continue to work, and so if no constraints are made at all and we stick with the a priori distribution then no money changes hands, and because it feels like the right thing to do. We take the log of the factor because we are trying to measure surprise and independent surprises should add, and because it feels like the right thing to do.

PROOF that choosing the maximum entropy distribution is the optimal strategy in this game.

Suppose for simplicity that there is only one constraint. I should have said before that the kind of constraints I am thinking about are constraints of the form: the function f having value f_i at point i has expected value $\bar{f}$, i.e.

$$\sum_i p_i f_i = \bar{f}.$$

The maximum entropy distribution is obtained via a Gibbs factor:

$$p_i = \mu_i e^{\alpha f_i}/Z,$$

where

$$Z = \sum_i \mu_i e^{\alpha f_i},$$

and μ is the a priori distribution. So we end up getting paid

$$\alpha f_i - \log Z$$

if i is picked. Our expected payment is therefore

$$\alpha \bar{f} - \log Z$$

no matter what distribution God picks. This appears significant and makes us think we're on the right track.

And indeed, let's look at the problem this way: We are supposed to pick a distribution from a collection C of distributions so as to attain

$$\max_{p \in C} \min_{q \in C} \sum q_i \log(p_i/\mu_i).$$

We want to verify that this is equivalent to maximizing entropy, i.e. that

$$\begin{aligned} & \arg\max_{p \in C} \min_{q \in C} \sum q_i \log(p_i/\mu_i) \\ = \; & \arg\max_{p \in C} \sum -p_i \log(p_i/\mu_i) \\ = \; & \arg\min_{p \in C} \sum p_i \log(p_i/\mu_i), \end{aligned}$$

where arg max means the location of the maximum.

If we call the quantity

$$\nu_\mu(q,p) = \sum q_i \log(p_i/\mu_i)$$

'the degree to which q verifies p', then to optimize our strategy we want to pick p so as to maximize the minimum verification. We want to know if this is the same as picking p so as to minimize the self-verification. (That's pessimism for you.)

But in the case we are talking about we have seen that for the maximum entropy distribution—that is, the least self-fulfilling distribution—the degree of verification doesn't depend on the q chosen. So here we have a distribution p_{me} that doesn't like itself any better than anyone else likes it. That is, this distribution is its own worst enemy. But it is a general fact of life that a distribution likes itself at least as well as it likes anyone else:

$$\begin{aligned} & \sum q_i \log(p_i/\mu_i) \\ = & \sum q_i \log(q_i/\mu_i) - \sum q_i \log(q_i/p_i) \\ \leq & \sum q_i \log(q_i/\mu_i). \end{aligned}$$

So p is the least despised distribution in the world. In symbols:

$$\min_q \nu_\mu(q, p_{\mathrm{me}}) = \nu_\mu(p_{\mathrm{me}}, p_{\mathrm{me}})$$

and

$$\nu_\mu(p_{\mathrm{me}}, p) \leq \nu_\mu(p_{\mathrm{me}}, p_{\mathrm{me}}),$$

so

$$\min_{q\in C} \nu_\mu(q,p) \leq \nu_\mu(p_{\mathrm{me}}, p) \leq \nu_\mu(p_{\mathrm{me}}, p_{\mathrm{me}}) = \min_{q\in C} \nu_\mu(q, p_{\mathrm{me}}),$$

so

$$\arg\max_{p\in C} \min_{q\in C} \nu_\mu(q,p) = p_{\mathrm{me}} = \arg\min_{p\in C} \nu_\mu(p,p),$$

q.e.d.

PROBLEM: For which sets C of distributions does the equality

$$\arg\max_{p\in C} \min_{q\in C} \nu(q,p) = \arg\min_{p\in C} \nu(p,p)$$

hold? All sets? Seems unlikely. Sets convex in a suitable sense? Check out Csiszar's paper on I-divergence geometry. These are the sets from which it makes sense to pick the maximum entropy distribution.

Addendum

21 May 1982: Talking to Roger Rosenkrantz has made me realize that these is no reason to limit my original alternatives to the set C of God's alternatives. Let me simply be given the information that God will pick from the set C. If I want to choose some distribution outside of the set C for my best guess, fine.

Then, according to Roger, the reason we take the log in deciding how much I will be paid is that this is (roughly?) the only function with the property that when D is a one-element set I am always best off choosing that element. This seems like a pretty cogent reason for using the log.

Things to think about:

- mixed strategies. What if all I tell God is a distribution of distributions, and the disinterested party picks one of my distributions independently of picking the point according to God's distribution. Does this change anything?

- contradictory information. What if I am told that the real distribution belongs to both C and D, which are disjoint. To make sense of this, let me assume that one of the two constraints is the real one, and that this will be decided by flipping a p-weighted coin at some point in the proceedings. Investigate this.