MAXIMUM ENTROPY ANALYSIS - A COMPANION

FLEMMING TOPSØE

Abstract. This is a companion to Robert Niven's block 4, 2007-lectures. I point to an alternative way of deriving the MaxEnt and the MinXEnt-distributions which is claimed to be more fundamental as well as technically a good deal simpler than the standard method via Lagrange multipliers (sic, still what you find in many (any ?) textbook you may pick up in this field). I hope you, the students in Robert's class, will realize that the method presented offers a deeper understanding with relevant interpretations which focus on what the physicist can do in terms of "describing" or "representing" the system studied, as always respecting the available knowledge. To fully benefit from the approach, you have to understand more of coding and to embark on a more thorough game-theoretical analysis.

1. The price to pay

Nothing is free. We have to learn a new concept. No way around that. In this section I give the shortest possible introduction to what we need.

Given is an *alphabet* $\mathbb{A}$, say consisting of s basic elements or *states* which we may label by an index i, running from 1 to s[1].

We know what a *probability distribution* (or just a *distribution*) over $\mathbb{A}$ is: a set $p = (p_i)_{i \le s}$ of non-negative numbers with sum 1:

$$\sum_{i \in \mathbb{A}} p_i = 1\,. \tag{1}$$

Now define an *idealized code length function*, in the following simply a *code*, as a set of numbers $\kappa = (\kappa_i)_{i \le s}$ such that *Kraft's equality* holds:

$$\sum_{i \in \mathbb{A}} \mathrm{e}^{-\kappa_i} = 1\,. \tag{2}$$

It is striking from (1) and (2) that you can go quite freely from distributions to codes and vica versa: Given p, the *adapted code* is the code

[1]in fact we could, as I think also Robert has hinted at, allow a countably infinite alphabet; this will only result in minor modifications in the following (basically infinite sums instead of finite ones). Actually, for most formulas I keep the option of an infinite alphabet open by writing $i \in \mathbb{A}$ instead of $i \le s$.

κ determined by

$$\kappa_i = -\ln p_i\,, \tag{3}$$

and, given κ, the *matching distribution* is the distribution p determined by

$$p_i = \mathrm{e}^{-\kappa_i}\ . \tag{4}$$

For a code, all values are ≥ 0. One should also allow the value ∞ which corresponds to an *impossible event*, an event with zero probability ($p_i = 0$). The other extreme, $\kappa_i = 0$, corresponds to a *certain event* ($p_i = 1$). If κ is a code and p a distribution, we call (κ, p) a *matching pair* if (3) and (4) hold.

Why bother to introduce codes? After all, the relationships (3) and (4) are extremely simple and thus the new concept can be avoided altogether. We will see! Some indications now, more later.

Often we will consider codes and distributions which need not form a matching pair. For any such pair, say (κ, p), a key quantity to consider is the *average code length*, denoted $\langle \kappa, p\rangle$ and defined – as you would expect – as the average

$$\langle \kappa, p\rangle = \sum_{i\in\mathbb{A}} p_i\kappa_i\,. \tag{5}$$

To "put you in the mood" , $(\kappa, p) \curvearrowright \langle \kappa, p\rangle$ is the *complexity function* and:

> *$\langle \kappa, p\rangle$ is to be thought of as the complexity seen from the point of view of the physicist when he is using κ as his tool for describing the system under study if the system is in fact governed by the distribution p.*

Instead of the word "describing" , pointing to codes as "descriptors" , you could talk about "representations" or "observation strategies" or "measurement strategies" and emphasize that this should be seen as reflecting the physicists ideas about the system, indeed his *knowledge* about the system. The above point of view is thus in complete conformity with the ideas brought forward by Jaynes.

2. First results

Two definitions suggest themselves:

Definitions The *entropy* of a distribution p is the minimal associated complexity:

$$\mathrm{H}(p) = \min_{\kappa} \langle \kappa, p \rangle \,. \tag{6}$$

And the *redundancy* $\mathrm{D}(p\|\kappa)$ of κ given p is the actual complexity minus the smallest achievable complexity, i.e.

$$\mathrm{D}(p\|\kappa) = \langle \kappa, p \rangle - \mathrm{H}(p) \,. \tag{7}$$

It is understood that in (6), the minimum is over all codes. We may write (7) in the form

$$\langle \kappa, p \rangle = \mathrm{H}(p) + \mathrm{D}(p\|\kappa) \tag{8}$$

which we refer to as the *linking identity* – indeed, it links together the three key quantities, *complexity, entropy* and *divergence*. As a companion to the definition (7) we define the *divergence* $\mathrm{D}(p\|q)$ between two distributions p and q as the corresponding redundancy, replacing q with the code adapted to q, i.e.

$$\mathrm{D}(p\|q) = \mathrm{D}(p\|\kappa) \text{ with } \kappa \text{ the code adapted to } q \,. \tag{9}$$

Then we may write the linking identity in the form

$$\langle \kappa, p \rangle = \mathrm{H}(p) + \mathrm{D}(p\|q) \,, \tag{10}$$

it being understood that (κ, q) is a matching pair.

Theorem 1.

$$\mathrm{H}(p) = -\sum_{i \in \mathbb{A}} p_i \ln p_i \tag{11}$$

$$\mathrm{D}(p\|q) = \sum_{i \in \mathbb{A}} p_i \ln \frac{p_i}{q_i} \,. \tag{12}$$

The simple proof below uses the elementary inequality $\ln x \leq x - 1$.

Proof. Let p and q be two distributions and κ the code adapted to q. As

$$\langle \kappa, p \rangle = -\sum_{i \in \mathbb{A}} p_i \ln p_i + \sum_{i \in \mathbb{A}} p_i \ln \frac{p_i}{q_i} \tag{13}$$

and as

$$\sum_{i \in \mathbb{A}} p_i \ln \frac{p_i}{q_i} = -\sum_{i \in \mathbb{A}} p_i \ln \frac{q_i}{p_i} \geq -\sum_{i \in \mathbb{A}} p_i \big(\frac{q_i}{p_i} - 1 \big) = 0 \,, \tag{14}$$

(with equality if $p = q$), the result follows. □

So entropy is nothing but the familiar Boltzmann-Gibbs-Shannon entropy and divergence the, likewise familiar, Kullback-Leibler divergence. Comforting!

The above proof told us that $\mathrm{D}(p\|q) \geq 0$ (with equality if and only if $p = q$). This is the most fundamental inequality of information theory. Our findings also show that entropy equals complexity for the adapted code:

$$\mathrm{H}(p) = \langle \kappa, p \rangle \text{ with } \kappa \text{ the code adapted to } p\,. \tag{15}$$

3. MaxEnt made easy

By a *model* we shall here understand any set of distributions over $\mathbb{A}$. If $\mathcal{P}$ is such a model, we denote by $H_{\max}(\mathcal{P})$ the *maximum entropy value*, defined as the suppremum $\mathrm{H}(p) = \sup_{p\in\mathcal{P}} \mathrm{H}(p)$[2]. And we say that the distribution p is the *maximum entropy distribution* (the *MaxEnt distribution*) if $p \in \mathcal{P}$ and $\mathrm{H}(p) = H_{\max}(\mathcal{P})$.

Our first result is an almost trivial but, nevertheless, extremely useful observation. It relies on the following notion: A code κ^* is *robust (for the model $\mathcal{P}$)* if $\langle \kappa^*, p \rangle$ is independent of p as long as $p \in \mathcal{P}$. The common value of $\langle \kappa^*, p \rangle$ for $p \in \mathcal{P}$ is the *constant of robustness*.

Lemma 1. *Let (κ^*, p^*) be a matching pair with $p^* \in \mathcal{P}$ and κ^* robust. Then p^* is the unique MaxEnt distribution and $H_{\max}(\mathcal{P}) = h$, the constant of robustness.*

Proof. By (15), $\mathrm{H}(p^*) = \langle \kappa^*, p^* \rangle = h$. And if $p \in \mathcal{P}$ is distinct from p^*, then by (10), $\mathrm{H}(p) < \mathrm{H}(p) + \mathrm{D}(p\|p^*) = \langle \kappa^*, p \rangle = h$. □

As an example, we conclude that if $\mathcal{P}$ consists of all distributions over $\mathbb{A}$, then the uniform distribution is the MaxEnt distribution. This follows as the code with all codeword lengths equal ($= \ln s$) is robust.

Now, to handle a more general situation, in fact the most important model for statistical physics as well as for numerous other applications, let us consider finitely many constraints, say R constraints, all given by specifying the meanvalues, denoted $\langle f_r \rangle$, for given functions f_r, $r \leq R$. Thus, the f_r's are given real-valued functions defined on $\mathbb{A}$ and the $\langle f_r \rangle$'s are given constants, the specified meanvalues. The model we have in mind is given by:

$$\mathcal{P} = \{p | \langle f_r, p \rangle = \langle f_r \rangle \text{ for } r = 1, \cdots, R\}\,. \tag{16}$$

The strategy we shall adopt in the search for the MaxEnt distribution for $\mathcal{P}$ is to search for a robust code κ^* with matching distribution in

[2]for models occurring in typical applications the supremum is achieved and may thus be replaced by a maximum.

the model. By Lemma 1, this will solve the problem. Clearly, any code of the form

$$\kappa^* = \lambda_0 + (\lambda_1 f_1 + \lambda_2 f_2 + \cdots + \lambda_R f_R) \tag{17}$$

is robust. In more detail, such a code is given by the values $\kappa_i^* = \lambda_0 + (\lambda_1 f_{1,i} + \cdots + \lambda_R f_{R,i})$ where the function values of f_r are denoted by $f_{r,i}$, $i \in \mathbb{A}$. There are many codes of this form. Indeed, for any set of constants λ_r, $r \leq R$ we can define λ_0 so that κ^* given by (17) is a code. We only have to make sure that Krafts equation (2) holds, and this is easily achieved by taking for λ_0 the value

$$\lambda_0 = \ln \sum_{i \in \mathbb{A}} \mathrm{e}^{-\lambda_1 f_{1,i} - \cdots - \lambda_R f_{R,i}} \ . \tag{18}$$

The sum in (18) plays a central role. It is the *partition function* associated with $\mathcal{P}$. We denote it by the letter "Z" (german: "Zustandssumme" – in danish the term is "tilstandssum"). The function is defined for all vectors $(\lambda_1, \cdots, \lambda_R) \in \mathbb{R}^R$ [3] by the formula

$$Z(\lambda_1, \cdots, \lambda_R) = \sum_{i \in \mathbb{A}} \mathrm{e}^{-\lambda_1 f_{1,i} - \cdots - \lambda_R f_{R,i}} \ . \tag{19}$$

Let us introduce a more streamlined notation by using boldface letters for vectors: $\boldsymbol{\lambda}$ for $(\lambda_1, \cdots, \lambda_R)$, $\mathbf{f}$ for $(f_1, \cdots, f_R)$ and, for any $i \in \mathbb{A}$, $\mathbf{f}(i)$ for $(f_{1,i}, \cdots, f_{R,i})$. Also, we use $\cdot$ to denote an *inner product* (the same as a *scalar product*) of two vectors. Thus (17), (18) and (19) can be written as follows:

$$\kappa^* = \lambda_0 + \boldsymbol{\lambda} \cdot \mathbf{f} \,, \tag{20}$$

$$\lambda_0 = \ln Z(\boldsymbol{\lambda}) \text{ with} \tag{21}$$

$$Z(\boldsymbol{\lambda}) = \sum_{i \in \mathbb{A}} \mathrm{e}^{-\boldsymbol{\lambda} \cdot \mathbf{f}(i)} \ . \tag{22}$$

Applying Lemma 1 in conjunction with the above analysis, we obtain the following key result:

Theorem 2. *The MaxEnt distribution for the model $\mathcal{P}$ given by* (16) *is the distribution p^* of the form*

$$p_i^* = \frac{\mathrm{e}^{-\boldsymbol{\lambda} \cdot \mathbf{f}(i)}}{Z(\boldsymbol{\lambda})} \,;\, i \in \mathbb{A} \tag{23}$$

for which the R parameters in $\boldsymbol{\lambda}$ are determined from the R constraints $\langle f_1, p^* \rangle = \langle f_1 \rangle, \cdots, \langle f_R, p^* \rangle = \langle f_R \rangle$. The MaxEnt value is

$$H_{\max}(\mathcal{P}) = \ln Z(\boldsymbol{\lambda}) + \boldsymbol{\lambda} \cdot \langle \mathbf{f} \rangle \,. \tag{24}$$

[3] if the alphabet is infinite, the restriction that the defining sum (19) must be convergent also has to be taken into account.

In (24), $\boldsymbol{\lambda} \cdot \langle \mathbf{f} \rangle$ is short for $\lambda_1 \langle f_1 \rangle + \cdot + \lambda_R \langle f_R \rangle$.

Proof. As stated, this really follows from the analysis above. Indeed, with λ_0 given by (18) (equivalently, by (21) and (22)), and κ^* given by (17) (equivalently, by (20)), p^* given by (23) is the distribution which matches κ^*. And if we make sure that the R constraints are satisfied, this distribution will be in $\mathcal{P}$ and then Lemma 1 applies directly. As the constant on the right hand side of (24) is the constant of robustness, (24) follows. □

There is no general formula which allows us to express the solution in closed form. However, it is pretty clear that a solution exists as we have R equations to determine the R unknowns $\lambda_1, \cdots, \lambda_R$ [4] .

4. MinXEnt made easy

In this section we study exactly the same model as before – so the model is given by (16) – but now with the change that the physicist has already – based on previous experience or for other reasons – settled for a *prior distribution*. Let us denote this prior distribution by q and let κ_0 denote the code adapted to q. Again, the physicist is searching for an optimal choice of a distribution p^*, this time thought of as a *posterior distribution* representing a suitable *updating* of the prior, taking the information available, expressed by the model (16), into account.

Instead of basing the performance on complexity as before, it is now more reasonable to look at the *saving* one can achieve. This saving can be measured by the difference between the *a priori complexity* and the new, hopefully lower, *a posteriori complexity*, i.e. by the difference $\langle \kappa_0, p \rangle - \langle \kappa^*, p \rangle$ with $p \in \mathcal{P}$.

Note that $\langle \kappa_0, p \rangle - \langle \kappa^*, p \rangle = \mathrm{D}(p\|q) - \mathrm{D}(p\|p^*)$, hence, if the physicist knew p, his response would be to update by $p^* = p$ (of course!) and his maximal saving would be $\mathrm{D}(p\|q)$. We argue with Jaynes that the proper choice of the physicist should be "least committal" , i.e. the physicist should be prepared for the least favourable eventuality – anything else would correspond to extra information which the physicist does not have. Therefore, we realize that the physicists strategy should be to choose as update, that distribution $p^* \in \mathcal{P}$ which minimizes the cross entropy $\mathrm{D}(p\|q)$ among all distributions $p \in \mathcal{P}$. This distribution is the *minimum cross entropy distribution* – for short, the *MinXEnt distribution*. Of course, it need not exist. But in natural models – as the one we are considering – it does. In the information theoretical and the statistical literature, this distribution is called the *I-projection of*

[4] mathematically, the situation is more complicated if we allow an infinite alphabet – even if we stick to finitely many constraints.

q *on* $\mathcal{P}$. We denote by $D_{\min}(\mathcal{P}\|q)$ the value to aim for which then is the minimum (strictly speaking the infimum) of $\mathrm{D}(p\|q)$ for $p \in \mathcal{P}$.

The problem then is to identify the MinXEnt-distribution and to calculate $D_{\min}(\mathcal{P}\|q)$. Really, the solution is just as easy as before. This time we rely on the following notion of robustness: A code κ^* is *robust relative to the prior* q if, for some constant d, $\langle \kappa_0, p\rangle - \langle \kappa^*, p\rangle = d$ for all $p \in \mathcal{P}$. The constant d is the *constant of relative robustness*. The relevant lemma now reads:

Lemma 2. *Let (κ^*, p^*) be a matching pair with $p^* \in \mathcal{P}$ and κ^* robust relative to q. Then p^* is the unique MinXEnt distribution and $D_{\min}(\mathcal{P}\|q) = d$, the constant of robustness relative to the prior q.*

Proof. Clearly, $\mathrm{D}(p^*\|q) = \langle \kappa_0, p^*\rangle - \langle \kappa^*, p^*\rangle = d$ and, since, for any $p \in \mathcal{P}$ which is different from p^*, we find that

$$\mathrm{D}(p\|q) = \langle \kappa_0, p\rangle - \mathrm{H}(p) > \langle \kappa_0, p\rangle - \big(\mathrm{H}(p) + \mathrm{D}(p\|p^*)\big) = \langle \kappa_0, p\rangle - \langle \kappa^*, p\rangle = d\,,$$

the result follows. □

Armed with Lemma 2, the MinXEnt-problem is as easy to solve as the MaxEnt-problem: One notes that all codes of the form $\kappa^* = \kappa_0 + \lambda_0 + \boldsymbol{\lambda}\cdot\mathbf{f}$ are robust relative to q. Indeed, for $p \in \mathcal{P}$,

$$\langle \kappa_0, p\rangle - \langle \kappa^*, p\rangle = \langle \kappa_0 - \kappa^*, p\rangle = \langle -\lambda_0 - \boldsymbol{\lambda}\cdot\mathbf{f}, p\rangle = -\lambda_0 - \boldsymbol{\lambda}\cdot\langle\mathbf{f}\rangle\,.$$

For given $\boldsymbol{\lambda}$ we must adjust λ_0 so that κ^* becomes a genuine code. This leads to the consideration of Z_q, the *partition function relative to the prior* q, which is defined by

$$Z_q(\boldsymbol{\lambda}) = \sum_{i\in\mathbb{A}} q_i\, \mathrm{e}^{-\boldsymbol{\lambda}\cdot\mathbf{f}(i)}\ . \tag{25}$$

Imitating the proof of Theorem 2, we then find the solution to our problem:

Theorem 3. *The MinXEnt distribution for the model $\mathcal{P}$ given by* (16) *and with prior q is determined as the distribution p^* of the form*

$$p^*_i = \frac{q_i\, \mathrm{e}^{-\boldsymbol{\lambda}\cdot\mathbf{f}(i)}}{Z_q(\boldsymbol{\lambda})}\,;\ i \in \mathbb{A} \tag{26}$$

for which the R parameters in $\boldsymbol{\lambda}$ satisfy the R constraints $\langle f_r, p^*\rangle = \langle f_r\rangle$, $r \le R$. The MinXEnt value is

$$D_{\min}(\mathcal{P}\|q)) = -\ln Z_q(\lambda) - \boldsymbol{\lambda}\cdot\langle\mathbf{f}\rangle\,. \tag{27}$$

5. Discussion

We have given natural intrinsic proofs of key results. To do so we had to introduce some new concepts, rooted in coding. In this way, it appears that Jaynes views can be supported more strongly than by other means. However, to reach the full benefit of the approach taken, one has to consider more explicitly the two sides involved, that of the model ("Nature") and that of the physicist. This requires more acquaintance with concepts from game theory.

The elegant and important general technique of Lagrange multipliers has been avoided. Not a deed in itself, but as it turned out, it gives more insight into the problem, e.g. the constants that appear in the solutions to the MaxEnt- and the MinXEnt problems – the same constants that would appear had we applied the standard method via Lagrange multipliers – have a clear influence on the code which the physicist is adviced to use. As a technical advantage, apart from being very expedient, our method leaves no doubt as to the nature of the solutions found (with Lagrange multipliers involved, extra effort is necessary if you want to ensure what kind of stationary point you have found – a minimum, a maximum or some kind of saddlepoint).

Our method can also be generalized considerably to involve different kinds of complexity functions (however, I cannot indicate here the wide range of further possibilities).

The technique presented was published in the seventies. One can ask why it is not universally adopted in textbooks concerned, especially in statistical physics. Perhaps, this is partly due to a resistence among scientists to adopt unfamiliar concepts (codes etc.), partly to the not-so-efficient promotional efforts. In fact, the results were first published in journals of pure mathematics and are only now finding their way to the more applied literature.

Note the great similarity between Theorems 2 and 3. It lies nearby to ask if a general result can be developed comprising both results at one stroke. Indeed, this is possible. We leave it to the readers ingenuity to suggest exactly how this can be achieved (not quite trivial but once done, it is nice!).

References

Instead of references, I point to http://isis.ku.dk/kurser/index.aspx?kursusid=26251&xslt=default regarding Robert's lectures, to my homepage http://www.math.ku.dk/~topsoe, or my book "Informationsteori" – or better still, to http://www.illc.uva.nl/HPI/ for information on codes, games etc. The latter relates to the publication later this year of "Handbook on the Philosophy of Information". That link and further links there give a wealth of information of interest for a wide audience!