

Maximum Entropy: Clearing up Mysteries

Marian Grendár, jr^{1,*} and Marian Grendár²

¹Institute of Measurement Science, Slovak Academy of Sciences, Dúbravská cesta 9,
842 19 Bratislava, Slovakia

Tel: +(421 7) 54773269 2407, Fax: +(421 7) 54775943, E-mail: umergren@savba.sk

²E-mail: marian.grendar@bb.telecom.sk

* Author to whom correspondence should be addressed.

Received: 29 November 2000 / Accepted: 31 March 2001 / Published: 8 April 2001

Abstract:

There are several mystifications and a couple of mysteries pertinent to MaxEnt. The mystifications, pitfalls and traps are set up mainly by an unfortunate formulation of Jaynes' die problem, the cause célèbre of MaxEnt. After discussing the mystifications a new formulation of the problem is proposed. Then we turn to the mysteries. An answer to the recurring question 'Just what are we accomplishing when we maximize entropy?' [8], based on MaxProb rationale of MaxEnt [6], is recalled. A brief view on the other mystery: 'What is the relation between MaxEnt and the Bayesian method?' [9], in light of the MaxProb rationale of MaxEnt suggests that there is not and cannot be a conflict between MaxEnt and Bayes Theorem.

Keywords: Maximum entropy, Maximum Probability, Jaynes' die problem, Maximum Likelihood, MiniMax Entropy.

1 Introduction

Shannon's entropy maximization is the place where several (parallel:) roads cross: statistical physics, information theory, mathematical statistics, inverse problems, Bayesian probability theory (BPT). Though it was noted that the entropy should be named also after Boltzmann, Planck, Gibbs, Von Neumann, Jeffreys, Pauli, Kullback, Leibler, Ingarden, Good, the fundamental importance of its *maximization* was stressed and elaborated mainly by Jaynes (visit [1]), in the realm of statistical physics, where the idea had precursors in Boltzmann and Gibbs. Jaynes raised the Shannon's entropy maximization (MaxEnt) to the principle status, in the realm of (equilibrium) statistical physics, and promoted MaxEnt also in the area of mathematical statistics, inverse problems and BPT.

In this brief note we shall attempt to offer several critical comments on the corpus of Jaynes' interpretative thoughts on MaxEnt, scattered through his rich intellectual heritage. We shall concentrate to the famous Jaynes' die problem, the celebrated example of capabilities of MaxEnt method.

2 MaxEnt ex machina? (Jaynes' die problem)

The die problem in 1978 formulation appeared in the following form (see [7]):

Suppose a die has been tossed N times and we are told only that the average number of spots up was not 3.5 as one might expect for an "honest" die but 4.5. Given this information, *and nothing else*, what probability should we assign to i spots in the next toss?

The probability distribution maximizing Shannon's entropy in the class of all pmf's satisfying a constraint equating expectation to the average value is not demanding to find out, mathematically (for instance by the Lagrange multiplier method). More difficult are questions which arise when one starts to approach the set-up from different sides.

2.1 The first reformulation: there is no die

Let us begin with the following first note: it is obvious, and Jaynes did explicitly mention it, that the moment consistency constraint $\sum_{i=1}^6 p_i x_i = 4.5$ can be satisfied by an infinite number of pmf's; in other words: there is a plenty of probability mass functions with the mean value of 4.5. The problem (from the category of inverse problems) then lies in the fact that we want to pick up one pmf of the class, in the best way (in a sense) we can. But, the *unfortunate* mentioning of the die

in the problem formulation suggests that we have to assume multinomial distribution – and that the problem is in fact to choose six parameters of the multinomial pmf.

Many had fallen into this trap, which inevitably leads to claim (presented also by Jaynes) that the die problem could be solved by the maximum likelihood (ML) method, if observed data (or frequencies) were known to us. However, since they are not given to us, the ML method cannot be called upon, but – MaxEnt can, and it can produce an acceptable result. This way of reasoning is for decades repeated as the illustration of abilities and power of the MaxEnt method.

To clarify the issue about what is in fact the feasible set of pmf's in the die problem, and not to obscure the main point of the MaxEnt method we suggest to take out the dice from the die problem formulation completely. Thus, consider the first reformulation of the problem:

Let there be an object of experimentation, revealing 6 different outcomes, which can be enumerated by $\{1, 2, \dots, 6\}$. The experiment was performed N times, but instead of the entire sample we are told only the average of the results, to be 4.5. Given this information ...

2.2 The second reformulation: potential

Given this information, ... what is the question we would like to ask? Do we look for assignment of probabilities for the outcomes? After a criticism, Jaynes re-interpreted the task as laying in estimation of frequencies of the outcomes. Neither future probabilities nor past frequencies are the task of MaxEnt, as we shall argue in the next Section.

By the way: just what are we accomplishing when we maximize entropy? (see [8]). This is the question! Answer to this question could also resolve the problem of what to ask in the 'die' problem. Several verbal interpretations of the entropy maximization were offered: it was seen to produce the most uniform (in what sense? definitely not in the geometrical sense!) pmf subject to the constraint; the smoothest pmf, the flattest, the least prejudiced pmf etc.

One more fact, which usually remains uncommented both by followers and opponents of MaxEnt. The pmf recovered by MaxEnt is the same regardless of N – the sample size that the average 4.5 came from. This fact is not that much disturbing in context of problems of statistical physics where N is indeed large, but it is definitely a weak point of the MaxEnt method in statistics.

The above list of issues is already long enough, but yet there remains one more important thing: the relationship between MaxEnt and ML methods. Jaynes ([8]) had noted that 'almost every conceivable opinion about the relation between maximum entropy and maximum likelihood can be found expressed in the current literature'. MaxEnt, constrained by a moment consistency constraint recovers a pmf in exponential form. ML estimator of parameter of the exponential form, should be found out of just the moment consistency constraint. We dubbed this well-known fact

as 'complementarity' of ML and MaxEnt tasks. Surprisingly, the complementarity holds also for a more general *parametric* form of the constraints – just MaxEnt should be replaced by MiniMax Ent (see [5]).

The ML-MaxEnt relationship investigations also revealed important, hidden ingredient of the 'die' problem, and the entire MaxEnt method; which we have dubbed *potential*, due to an analogy between Boltzmann's deduction of equilibrium distribution of an ideal gas placed in an external potential field and probability density function (see [6]). Recognition of the key role of the notion of potential leads to the following reformulation of the 'die' problem:

Let there be an object of experimentation, revealing 6 different outcomes, which can be enumerated by $x = \{1, 2, \dots, 6\}$. The experiment was performed N times. A potential function $u(x) = x$ was chosen, and its average value $\sum_{i=1}^6 f_i u(x_i)$, where f_i represents frequency of the i -th outcome in the sample, was calculated. Given the average value of the potential function, and nothing else ...

So, in the Jaynes' die problem we were in fact also told that $u(x) = x$.

Now we can return to the above mentioned complementary relationship of ML and MaxEnt. Appreciation of the idea of potential has an immediate implication for the 'die' problem. The problem can be easily solved by ML method, despite the fact that the sample (or frequencies) are not given to us: first, note that the known potential function defines a unique parametric family of pmf's in exponential form (see [5]). Second, pick up any sample (frequencies) such that $\sum_{i=1}^6 f_i u(x_i)$ is equal to the the prescribed number. Then the ML estimation will produce exactly the same result as MaxEnt. The second step relies upon what is in statistics known as sufficient statistic.

2.3 The third reformulation: what is the question?

Finally we can turn to the last task of completing the 'die'-problem set-up. All the above mentioned pitfalls and traps can be viewed as mystifications rather than mysteries around MaxEnt. Now, let us address the mysteries.

Why should we maximize just Shannon's entropy to pick up a pmf out of the feasible set? Why not some other objective function? A healthy dozen of axiomatizations were developed, to show that the Shannon's entropy is the only criterion function to be used for 'judging under uncertainty'. Yet, they do not answer the more elementary question: What do we do when we maximize entropy? It is the first mystery. Building upon Boltzmann's multiplicity argument it was shown (see [6]), that MaxEnt is just an asymptotic form of elementary, self-evident method, dubbed *MaxProb*, of looking for such an empirical pmf (occurrence vector) which has maximal probability of being generated by a uniform generator. As a by-product this result also solves the

above mentioned inability of MaxEnt to take into account an information on sample size and also suggests a class of constraints which are allowed to bind MaxProb/MaxEnt.

Taking the things together, we can propose the final formulation of the 'die' problem:

Let there be an object of experimentation, revealing 6 different outcomes, which can be enumerated by $x = \{1, 2, \dots, 6\}$. The experiment was performed N times. A potential function $u(x)$ was chosen, and its average value $\sum_{i=1}^6 f_i u(x_i)$, where f_i represents frequency of the i -th outcome in the sample, was calculated. This information $\{N, u(x), \sum_{i=1}^6 f_i u(x_i)\}$ forms a feasible set \mathcal{H} of absolute frequency vectors (occurrence vectors). Now we ask: what is the most probable occurrence vector which can be generated by a uniform generator?

This simple probabilistic question leads for $N \rightarrow \infty$ to the same answer as MaxEnt (see [6] for the exact formulation, the proof, and an illustrative example).

3 MaxEnt vs. Bayes?

MaxProb rationale of MaxEnt seems to offer also a clear way through the fog of views on the relation between MaxEnt and Bayesianism.

The unclear understanding of the aim of MaxEnt and consequent suggestive analogies with Bayes Theorem have produced rather huge literature on the relation between MaxEnt and BPT. Bayesian Probability Theory is viewed either as a special case of MaxEnt, or vice versa MaxEnt as a special form of Bayes Theorem, or they are seen as unrelated at all. Moreover, Jaynes favored MaxEnt as a method for assigning priors at BPT.

Instead of going to a detailed account of the long list of different views on the relation of MaxEnt and the Bayes method we would like to highlight that MaxProb rationale of MaxEnt reveals that the Jaynes' guess of what is the task of MaxEnt (recall the question from the Jaynes' die problem formulation) was wrong. MaxEnt does not solve the problem of assigning probabilities to the *next* event. This makes clear why those of investigators of the relation between MaxEnt and the Bayes' Theorem who answered Jaynes' die problem question by means of the Bayes' formula (see for instance [10], [3], [2]) came to a result different than the MaxEnt one. They were not solving the same problem as MaxEnt solves.

MaxProb rationale of MaxEnt shows that MaxEnt and Bayes Theorem *cannot be in conflict*, for the plain reason: MaxEnt solves a specific though very general probabilistic problem, Bayes Theorem is a fundamental statement of probability theory. How can they be at odds with each other? See also [4] for another view supporting this claim.

Acknowledgements. It is our pleasure to thank Peter Cheeseman, Aleš Gottvald, George Judge, Gejza Wimmer and Viktor Witkovský for valuable discussions on MaxEnt.

In memory of Edwin T. Jaynes.

References

- [1] <http://bayes.wustl.edu> - a web page dedicated to Edwin T. Jaynes, maintained by L. Bretthorst
- [2] Cheeseman, P.; Lored, T.; Skilling, J.; Stutz, J. and Smelyanskiy V. On Bayesian inference and Maximum Entropy, NASA Ames Research Center, working paper, 2000.
- [3] Frieden, B. R. Dice, entropy and likelihood. *Proc. IEEE* **1985**, *73*, 1764-1770.
- [4] Gottvald, A. Beyond the MaxEnt principle: Bayesian, spectral and convolutionary metamorphoses. In *Proceedings of the 4th Japan-Central Europe joint workshop on energy and information in nonlinear systems*, Gottvald, A.(ed.), to appear.
- [5] Grendár, M. jr. and Grendár, M. MiniMax Entropy and Maximum Likelihood: complementarity of tasks, identity of solutions. In *Maximum Entropy and Bayesian methods*, Mohammad-Djafari, A. (ed.), AIP, NY, to appear. Proceedings of the 20th MaxEnt Workshop held at CNRS, Gif sur Yvette, France, July 8-13, 2000. Also, available on-line at LANL arXiv <http://xxx.lanl.gov/abs/math.PR/0009129>
- [6] Grendár, M. jr. and Grendár, M. What is the question that MaxEnt answers? A probabilistic interpretation. In *Maximum Entropy and Bayesian methods*, Mohammad-Djafari, A. (ed.), AIP, NY, to appear. Also, available on-line at <http://xxx.lanl.gov/abs/math-ph/0009020>
- [7] Jaynes, E. T. Where do we stand on Maximum Entropy?. In *The Maximum Entropy Formalism*; Levine, R. D. and Tribus, M., Eds.; M. I. T. Press: Cambridge, MA, 1978.
- [8] Jaynes, E. T. On The Rationale of Maximum Entropy Methods. *Proc. IEEE* **1982**, *70*, 939-952.
- [9] Jaynes, E. T. The relation of bayesian and maximum entropy methods. In *Maximum Entropy and Bayesian Methods in Science and Engineering (Vol 1)*; Erickson, G. J. and Smith, C. R., Eds.; Kluwer Academic Publishers: the Netherlands, 25-29, 1988.
- [10] Uffink, J. The Constraint Rule of the Maximum Entropy Principle. *Studies in History and Philosophy of Modern Physics* **1996**, *27*, 47-79.