Entropy **2015**, *17*, 401-406; doi:10.3390/e17010401

OPEN ACCESS
entropy
ISSN 1099-4300
www.mdpi.com/journal/entropy

Letter

On an Objective Basis for the Maximum Entropy Principle

David J. Miller * and Hossein Soleimani

Department of Electrical Engineering, The Pennsylvania State University, University Park, PA 16802, USA; E-Mail: hsoleimani@psu.edu

* Author to whom correspondence should be addressed; E-Mail: djmiller@engr.psu.edu; Tel.: +1-814-865-6510.

Academic Editor: Kevin H. Knuth

Received: 26 September 2014 / Accepted: 15 January 2015 / Published: 19 January 2015

Abstract: In this letter, we elaborate on some of the issues raised by a recent paper by Neapolitan and Jiang concerning the maximum entropy (ME) principle and alternative principles for estimating probabilities consistent with known, measured constraint information. We argue that the ME solution for the "problematic" example introduced by Neapolitan and Jiang has stronger objective basis, rooted in results from information theory, than their alternative proposed solution. We also raise some technical concerns about the Bayesian analysis in their work, which was used to independently support their alternative to the ME solution. The letter concludes by noting some open problems involving maximum entropy statistical inference.

Keywords: maximum entropy; asymptotic equipartition principle

---

1. Introduction

In a recent paper, "A note of caution on maximizing entropy" [1], the authors considered the problem of estimating a probability mass function given supplied constraint information. They identified as "problematic" the maximum entropy solution for the example of a 3-sided die, where the given constraint information is that the mean die value is two. For this example, maximum entropy (ME) solves the following problem:

$$\underline{p}_{\mathrm{ME}} = \arg\max_{\underline{p}} -\sum_{i=1}^{3} p_i \log p_i$$

$$\text{subject to} \quad \sum_{i=1}^{3} ip_i = 2 \qquad (1)$$
$$\sum_{i=1}^{3} p_i = 1.$$

In this case, one can easily show that the maximum entropy solution, consistent with the given constraints, is the uniform distribution $p_i = \frac{1}{3}$, $i = 1, 2, 3$.

In their paper, the authors propose an alternative "objectively-based" approach for solving this problem. Specifically, they suppose that the probabilities are random variables, uniformly distributed over their ranges, which are prescribed by the given constraints, *i.e.*, $p_1 = p_3 \in [0, 0.5]$ and $p_2 \in [0, 1]$. Accordingly, they choose, as their estimated probabilities, the expected values of these (uniformly distributed) random variables: $\underline{p} = (p_1, p_2, p_3) = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$. The authors argue on intuitive grounds that their solution may be preferable to the ME solution, as they state: "p_2 could be as high as 1, while the other probabilities are bounded above by 0.5....[so] we may be inclined to bet on 2. Once the information gives us reason to prefer one alternative over the others, it is troublesome to claim that the probabilities...are equal." They then also consider a Bayesian learning setting and show, under particular stated assumptions, that Bayesian updating is consistent with their $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ solution. Beyond identifying what the authors call a "problematic" example for the maximum entropy principle, their paper gives historical background on the interpretation of probability, including excerpts of Jaynes' views on maximum entropy and some of the multiple senses in which, based on Jaynes' writings, one can construe that the maximum entropy principle gives "objective" probabilities.

In this letter, we do not attempt to elucidate or specifically articulate Jaynes' understanding of the maximum entropy principle. The purpose of this letter is to elaborate further on the 3-sided die problem from [1] (as well as related problems, where ME is often applied) in order to further understand and explicate several statistically objective bases for preferring one set of probability assignments over another. In so doing, we will argue that there is strong, objective support for the ME solution, as opposed to the alternative solution proposed by Neapolitan and Jiang. We also identify some open problems in maximum entropy statistical inference.

2. On Objective Bases For Preferring One Probability Assignment Over Another

2.1. "Most Probable" Interpretation of Maximum Entropy

In [2], Jaynes does provide a principled basis for preferring the maximum entropy solution over alternative probability assignments. Specifically, let N be the number of repeated trials of an experiment with K possible outcomes $\{\omega_1, \omega_2, \ldots, \omega_K\}$, and with some constraint information, such as the mean die value measured based on these N repeated trials. Note that the outcomes of the individual trials are not known. Nor is it known the number of occurrences (N_k) of each distinct outcome, ω_k. However, suppose that we *did* know $N_k, k = 1, \ldots, K$. For large N, by the weak law of large numbers, e.g., [3], we know that $\frac{N_k}{N} \to p_k$ with probability 1, where $p_k, k = 1, \ldots, K$ are the true probabilities. Thus, if $(N_1, N_2, \ldots, N_K)$ were known, a good choice for the probability assignments would be the frequency counts $(\frac{N_1}{N}, \frac{N_2}{N}, \ldots, \frac{N_K}{N})$. Accordingly, estimating $\underline{p} = (p_1, p_2, \ldots, p_K)$ amounts to estimating

$(N_1, \ldots, N_K)$. Let $(x_1, x_2, \ldots, x_N)$, $x_i \in \{\omega_1, \ldots, \omega_K\}$, be a particular N-trial realization sequence (microstate) for the experiment, with associated *macrostate* (counts) $(N'_1, N'_2, \ldots, N'_K)$ that agrees with the given constraint information. Suppose all such microstates are *a priori* equally likely. Then the probability of macrostate $(N_1, N_2, \ldots, N_K)$ is:

$$P(N_1, \ldots, N_K) = \frac{1}{K^N} \binom{N}{N_1, ..., N_K}, \tag{2}$$

where the multinomial $\binom{N}{N_1,\ldots,N_K}$ is the number of distinct microstates that are consistent with the (constraint-achieving) macrostate. Since, for any given realization sequence, with macrostate $(N_1, \ldots, N_K)$, we would form the probability estimate as $\underline{p} = (\frac{N_1}{N}, \ldots, \frac{N_K}{N})$, $P(N_1, \ldots, N_K)$ is *also* the probability that we form the probability assignment $(\frac{N_1}{N}, \ldots, \frac{N_K}{N})$. Thus, if we choose $(N_1, \ldots, N_K)$ to maximize (2), we are determining the probability mass function $(p_1, p_2, \ldots, p_K) = (\frac{N_1}{N}, \ldots, \frac{N_K}{N})$ that we are *most likely to produce as an estimate*, given the specified constraint information and the number of die tosses. To maximize (2), one maximizes the multinomial coefficient $\binom{N}{N_1,\ldots,N_K}$. Based on Stirling's approximation [4], $\binom{N}{N_1,\ldots,N_K} \sim e^{NH(\frac{N_1}{N},\ldots,\frac{N_K}{N})}$, with $H(\cdot)$ Shannon's entropy function. Accordingly, one can closely approximate maximizing (2) subject to, e.g., a mean value constraint $\sum_{i=1}^{K} iN_i = \mu$ by maximizing Shannon's entropy function $H(\frac{N_1}{N}, \ldots, \frac{N_K}{N})$ subject to the constraint. Allowing unconstrained probabilities, rather than fractions of N, this amounts to solving:

$$\max_{p_1,p_2,\ldots,p_K} -\sum_{i=1}^{K} p_i \log p_i \tag{3}$$

subject to $\sum_{i=1}^{K} ip_i = \mu$ and $\sum_{i=1}^{K} p_i = 1$.

Note, too, that since (2) or its approximate

$$P(p_1, \ldots, p_K) = \frac{e^{NH(p_1,\ldots,p_K)}}{K^N} \tag{4}$$

is the probability of forming the estimate $(p_1, \ldots, p_K)$, we can use (4) to evaluate the relative likelihoods of producing different candidate probability assignments, *i.e.*, $\frac{P(p_1,\ldots,p_K)}{P(p'_1,\ldots,p'_K)} = e^{N(H(p_1,\ldots,p_K)-H(p'_1,\ldots,p'_K))}$. Thus, the likelihood of the maximum entropy solution, relative to an alternative distribution, grows exponentially with the entropy difference.

In [1], they acknowledged this statistically-based justification for the maximum entropy distribution, as applied to the 6-sided Brandeis die problem. Moreover, they motivated the 3-sided die problem by stating that "suppose a friend later tossed the die many times...". Thus, their 3-sided die problem genuinely does consider the scenario where the constraint information was accurately measured, based on many repeated die tosses. Accordingly, the above interpretation should be applicable to their 3-sided die problem, just as it is to the Brandeis die problem. For the 3-sided die problem, we have $\frac{P(\frac{1}{3},\frac{1}{3},\frac{1}{3})}{P(\frac{1}{4},\frac{1}{2},\frac{1}{4})} = 2^{N(\log_2(3)-1.5)} \sim 2^{0.08N}$. For example, if $N = 1000$, the ME distribution is more than 10^{24} times more likely to be produced as the estimate than the proposed distribution from [1]. The only real assumption in this analysis is that realization sequences (microstates) consistent with a given macrostate are equally likely.

In [1], they also stated that "since there is no reason to believe the die is fair, we could consider all probability distributions that satisfy the constraints in the problem equally probable. That is, we apply the principle of indifference to the probability values themselves". While it is true that *a priori* there is no reason to believe the die is fair, there is also no evidence to support that all probability distributions are equally probable given that our only knowledge about the die is the mean value constraint. Applying a principle of indifference in this way, by [1], imposes a further (rather strong and unjustified) constraint which, as shown above, leads to a solution 10^{24} times less probable than the solution built on only the available evidence. The "least presumptuous" strategy should only consider the mean value constraint and not impose further constraints. Nevertheless, a different distribution than the ME solution may be achieved if the principle of indifference is applied correctly; *i.e.*, if it is supported by some evidence or by available knowledge about the distribution for a particular application.

2.2. Asymptotic Equipartition Principle (AEP) Interpretation

Another related albeit somewhat different vantage point from which to evaluate the two candidate distributions $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ is with respect to the asymptotic equipartition principle (AEP) [4]. The AEP considers N independent trials of random draws from a given distribution $(p_1, p_2, \ldots, p_K)$. It defines the *ϵ-typical* set $\mathcal{A}_\epsilon^{(N)}$ as the set of sequences $(x_1, \ldots, x_N)$ such that $2^{-N(H(X)+\epsilon)} \leq P(x_1, x_2, \ldots, x_N) \leq 2^{-N(H(X)-\epsilon)}$. One can show the following [4]:

(1) The cardinality of this set is approximately $2^{NH(p_1,p_2,\ldots,p_K)}$.

(2) Each sequence in this set is approximately equally likely.

(3) The typical set accounts for nearly all the probability, *i.e.*, the joint pmf $P(x_1, x_2, \ldots, x_N)$ is very nearly approximated as a uniform distribution on all typical sequences, with a zero probability assignment on all non-typical sequences.

Note that not all sequences that are typical will precisely achieve the mean value constraint. However, their deviation from the constraint is made arbitrarily small as N gets large. From the perspective of the AEP, according to the $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ model, there are roughly $2^{1.5N}$ equally likely realizations (each with self-information very close to $H(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$). On the other hand, according to the ME distribution $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, there are $2^{N \log_2(3)}$ equally likely realizations (typical sequences). Note that the sequences that are typical for $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ are not typical for $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, and vice versa. Thus, in choosing between the distributions, one is either rejecting $2^{1.5N}$ realizations typical of $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ or $2^{N \log_2(3)}$ realizations typical of $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Again, for $N = 1000$, choosing $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ means rejecting more than 10^{24} times the number of realizations than if one chooses $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. In the absence of additional information, it seems one should reject the hypothesis that contains (far) fewer realizations that are (approximately) consistent with the measured constraints. Again, from this vantage, the ME solution is preferred.

2.3. The Bayesian Learning Analysis from [1]

The authors in [1] sought to independently validate the $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ distribution by considering a Bayesian learning setting, starting from an uninformative Dirichlet prior on the probabilities. Here, they imposed

the mean value constraint, fixed $N_1 = N_3$, and assumed all (N_1, N_2) configurations consistent with satisfying the mean value constraint to be equally likely. They then let $N \to \infty$ and evaluated the conditional probability of die value 2, which they found to be 0.5. While this analysis does appear to support the proposed $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ distribution, a concern with this analysis is the consistency between letting $N \to \infty$ and the assumption that all (N_1, N_2) configurations meeting the constraint are equally likely. Specifically, by the weak law of large numbers, as $N \to \infty$, $\frac{N_1}{N} \to P_1$ and $\frac{N_2}{N} \to P_2$ with probability 1, *i.e.*, the assumption of *equally likely* macrostate pairs appears to be incompatible with letting $N \to \infty$.

2.4. Encoding Additional Constraints

Objectively, it is of course possible that the true distribution (used to randomly generate N realizations and achieve a mean value of 2) is the $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ distribution. This would be revealed within the ME framework if we imposed one additional constraint. Specifically, if, in addition to the mean value constraint, we imposed the second moment constraint $\sum_{i=1}^{3} i^2 p_i = 4.5$, the (unique) ME distribution is indeed the $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ distribution. Thus, ME is not incompatible with the $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ distribution. It is simply producing the "best" distribution given whatever accurate constraint information is made available.

3. Open Problems for Maximum Entropy

While we believe (as exposited here) that ME is an objectively supported method for estimating probabilities given supplied constraints, one open problem is perhaps how to properly account for inaccuracies that may exist in the measured constraints. Constraints in the ME framework are usually in the form of expectations, but since in practice the true values of expectations are not known, they are often estimated by sample averages from (observed) data. For example, in the problem discussed in this letter, $\mu = E[X]$ is replaced by its sample-based estimate $\hat{\mu} = \sum_i i \frac{N_i}{N}$. That is, expectations taken with respect to the ME distribution are constrained to agree with empirical averages, rather than with the true expectations. This may cause overfitting if the sample size is not large enough and, thus, if the expectations are poorly estimated. [5] proposed a framework to address this problem by relaxing strict constraint satisfaction. This approach may be problematic for applications where multiple constraints of different orders are imposed, with each constraint estimated based on a different sample size. For example, for a die experiment, the mean constraint is estimated based on the full set of N trials, whereas a conditional probability constraint such as $P(X = 6|X \geq 3)$ would be based on a smaller sample size. Consider also [6], where joint probability constraints, from pairwise probabilities up to fifth order, were encoded, in learning ME conditional probability models. It appears that a principled framework, properly accounting for varying degrees of constraint inaccuracy (based, e.g., on different sample sizes used to measure the constraints) is still needed.

Another open problem for maximum entropy is to determine a systematic procedure to search over all possible constraints and impose the most relevant ones for any particular problem. Especially for high-dimensional problems and domains where there is a lack of expert knowledge of the most suitable constraints, we need to objectively determine the relevant constraints to impose. There have been some efforts to address this issue. For instance, [7] described an iterative procedure to decide which constraints

to impose, applied to a natural language processing task. Also, [6] used the Kullback distance and the Bayesian Information Criterion to choose relevant constraints and applied this approach to the analysis of genome-wide association study. Nevertheless, it may be fruitful to further investigate alternative approaches for this problem.

4. Conclusions

In this letter, we elaborated on some of the issues raised by a recent paper [1] concerning the maximum entropy (ME) principle and alternative principles for estimating probabilities consistent with known, measured constraint information. We have argued that the ME solution for the "problematic" example introduced in [1] has stronger objective basis than their alternative proposed solution. We also noted some open problems involving maximum entropy statistical inference.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Neapolitan, R.E.; Jiang, X. A note of caution on maximizing entropy. *Entropy* **2014**, *16*, 4004–4014.
2. *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*; Rosenkrantz, R.D., Ed.; D. Reidel: Dordrecht, The Netherlands, 1982.
3. Rotar, V. *Probability and Stochastic Modeling*; CRC Press: Boca Raton, FL, USA, 2013.
4. Cover, T.M.; Thomas, J. *Elements of Information Theory*; Wiley: New York, NY, USA, 2006.
5. Dudik, M.; Phillips, S.J.; Schapire, R.E. Performance guarantees for regularized maximum entropy density estimation. In Proceedings of the 17th Annual Conference on Computational Learning Theory, Banff, Canada, 1–4 July 2004; pp. 472–486.
6. Miller, D.J.; Zhang, Y.; Yu, G.; Liu, Y.; Chen, L.; Langefeld, C.D.; Herrington, D.; Wang, Y. An algorithm for learning maximum entropy probability models of disease risk that efficiently searches and sparingly encodes multilocus genomic interactions. *Bioinformatics* **2009**, *25*, 2478–2485.
7. Berger, A.L.; Pietra, V.J.D.; Pietra, S.A.D. A maximum entropy approach to natural language processing. *Comput. Linguist.* **1996**, *22*, 39–71.