# Uses of Maximum Entropy

Sami Perttu

19th April 2001

## Contents



## 1 Introduction

In this seminar paper we point our lenses to Peter Grünwald's UAI-2000 article "Maximum Entropy and the Glasses You are Looking Through" [5]. Along the way, we provide an introduction to the *Maximum Entropy (MaxEnt) Principle* and its precursor, the *principle of insufficient reason*. Both are methods for *inductive inference*, where we would like to arrive at a general conclusion (*model*) from specific examples (*data*).

Specifically, the aim of both methods is to come up with a reasonable distribution for a random variable when we have only partial knowledge concerning the random variable. After a distribution has been chosen, the full range of methods for decision-theoretic and probabilistic inference becomes available. As the name implies, however, inductive inference is not guaranteed to lead to correct predictions. Grünwald [5] reinterprets MaxEnt in a game-theoretic setting and provides a characterization of the uses of MaxEnt. showing when it can be applied in a manner that does lead to reliable predictions.

First we discuss the principle of insufficient reason in Section 2. MaxEnt is introduced in Section 3. The game-theoretic reinterpretation is explained in Section 4, and in Section 5 we attempt to give some idea of the problems of *representation dependence* that are often claimed to beset MaxEnt. Finally, some concluding remarks are offered in Section 6.

## 2 The Principle of Insufficient Reason

Laplace's *principle of insufficient reason* (PIR) can be stated: if there is no reason to believe that either of two events is more possible than the other, then both should be assigned the same probability. By transitivity, the principle extends to any finite number of events. Let $\Omega = \{x_1, ..., x_n\}$ be a finite sample space consisting of such events $x_i$, and let $P$ be a probability measure on $\Omega$. PIR ascribes the uniform probability measure

$$P(x_i) = \frac{1}{n}, \quad i = 1 \ldots n \tag{1}$$

to events in $\Omega$. At first hearing, PIR may sound like a tautology or a case of circular reasoning. After all, what else does 'possible' mean but probable? But one can argue that it refers to rational judgments, and this firmly links the principle to the subjectivist, or Bayesian, interpretation of probability.

Since its introduction, PIR has come under attack from scientists in many disciplines [11]. Part of the resistance no doubt stems from its close connection to the subjectivist interpretation, but there are other reasons.

One objection to PIR is the 'nothing from nothing' argument, as expounded by Ellis in 1850[3]: "Mere ignorance is no ground for any inference whatsoever. *Ex nihilo nihil.* It cannot be that because we are ignorant of the matter we know something about it."

One's response to the above argument depends on how one views probability and the principle of insufficient reason. Is it just a practical prescription for assigning non-informative probability values? In probability theory, one cannot formally distinguish between "informative" and "non-informative" probabilities. Therefore some care must be taken when applying the principle.

## 3 The Maximum Entropy Principle

The Maximum Entropy (MaxEnt) Principle was introduced by Jaynes in 1957 [7]. His original motivation seems to have been to clarify and justify the foundations of statistical mechanics, but later Jaynes [9] and many others have extended it to a general principle of inductive inference.

Let us again assume we would like to infer a probability distribution $P$ for a random variable $X$ over a finite sample space $\Omega = \{x_1, \ldots, x_n\}$, e.g. for prediction of data or to decide upon a course of action with the aid of a utility function. As a note, MaxEnt can be extended to the continuum, but we will discuss only finite sample spaces. The true distribution of $X$ is $P^*$, but we do not know what it is. However, we have in our possession some background knowledge in the form of a set of constraints that the distribution $P^*$ meets, and a *measure* $M : \Omega \to \mathbb{R}^+$ that can be interpreted either as a way of counting outcomes in $\Omega$ or as a prior distribution over $\Omega$. Let $\mathcal{C}$ denote the set of distributions that satisfy the constraints. Assuming we have no other knowledge about $P^*$, which distribution $P$ should we adopt as our guess?

An intuitive notion would be that the distribution should embody all of our background knowledge, but no more: it should be *maximally uninformative* with respect to the constraints. Whether the principle we are about to introduce actually meets this goal has been the center of a long-standing debate [11].

In any case, the Maximum Entropy Principle tells us that the distribution *maximizing the relative entropy with respect to M given the constraints* is the one that should be picked. Entropy can be seen as the amount of information in a random variable. Relative entropy $\mathcal{H}_M(P)$ of $P$ with respect to measure $M$ is defined as

$$\mathcal{H}_M(P) = -\sum_{x \in \Omega} \Big[ P(x) \log_2 \frac{P(x)}{M(x)} \Big], \tag{2}$$

and the maximum entropy distribution $P^{me}_M$ relative to measure $M$ is formally

$$P^{me}_M := \arg\ \max_{P \in \mathcal{C}} \mathcal{H}_M(P) = \arg\ \max_{P \in \mathcal{C}} -\sum_{x \in \Omega} \Big[ P(x) \log_2 \frac{P(X)}{M(X)} \Big]. \tag{3}$$

Since the base of the logarithm does not matter, we will drop it from now on and just assume base 2. We can see from Equation (3) that if $M$ is the uniform measure $M(x) = 1$, then relative entropy coincides with traditional Shannon entropy. In case $M$ is a probability measure, MaxEnt is, confusingly enough, also sometimes called the *minimum relative entropy* principle. If there are no constraints, that is, if $\mathcal{C}$ contains all probability distributions over $\Omega$, then the uniform probability distribution ascribed by PIR is also the maximum entropy distribution. MaxEnt can thus be seen as an extension of the principle of insufficient reason.

The most common situation for applying MaxEnt is as follows. We translate the *observed average* value of a random variable to the constraint that its *expectation* should equal the average.

Now to a justification of MaxEnt. In the case of a uniform underlying measure $M$, if we assume that the samples generated by the underlying process are

- independent and identically distributed, and that
- every sequence of samples satisfying the expectation constraint is equally likely,

we arrive at the maximum entropy distribution in the limit as the amount of data increases [4]. Furthermore, the rate of convergence to the maximum entropy distribution is exponential. This convergence is termed the *concentration phenomenon*.

If one is willing to believe that the process is such that the above two conditions are met, then one is unequivocally lead to the distribution maximizing the entropy. However, as Grünwald [5] proceeds to point out, MaxEnt distributions have other interesting and potentially useful properties. We might also like to predict the value of some other random variable $Y$ based on observed values of $X$. Assuming that $X$ obeys the constraints above, how do the two random variables need to be related for us to be able to make reliable predictions? This question is somewhat technical, and we will not discuss it in depth in this paper. One possible difficulty with applying MaxEnt could be that the sample space $\Omega$, i.e. how many outcomes there are, must be known in advance. The role of measure $M$ is also not very clear if $\Omega$ does not permit a convenient interpretation for it. Then it may be best to just forget about it and assign the uniform measure $M(x) = 1$. A simple MaxEnt example is now in order.

> *Example 1* (Brandeis dice) [8]. The expected number of spots coming up in a toss of a certain six-sided die is 4.5 instead of 3.5. The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$, the random variable we would like to obtain a distribution for is trivially $X(x) = x$, the constraint is $E(X) = 4.5$, and we select the uniform measure $M(x) = 1$. The resulting maximum entropy distribution is $P_M^{me} = \{0.05435, 0.07877, 0.11416, 0.16545, 0.23977, 0.34749\}$. The concentration phenomenon tells us that if we toss an *honest* die a large number of times and the average number of spots coming

up is 4.5, then with overwhelming probability the frequencies correspond to $P_M^{me}$.

# 4 MaxEnt as a Game

In this section we explain how MaxEnt can be interpreted as a game. The *Kullback-Leibler divergence* $D(P \mid\mid Q)$ between two distributions $P$ and $Q$ in $\Omega$ is

$$D(P \mid\mid Q) = -\sum_{x \in \Omega_X} P(x) \log \frac{Q(x)}{P(x)}. \tag{4}$$

A basic theorem in information theory, the *information inequality* [2], states that $D(P \mid\mid Q) \geq 0$ with equality if and only if $Q = P$. K-L divergence can be interpreted as the expected number of *extra* bits needed to code a random variable $P$ using a code based on $Q$, instead of a code based on $P$, which the information inequality tells us would be the optimum.

The expected number of *total* bits needed to code $P$ using a code based on $Q$ is $-\sum_{x \in \Omega} P(x) \log Q(x)$, which reaches its minimum, i.e. equals entropy $\mathcal{H}(P)$, only with the choice $Q = P$. Throwing the underlying measure $M$ in the mix, we get a characterization of entropy as

$$\mathcal{H}_M(P) = \inf_{Q \in \mathcal{P}} - \sum_{x \in \Omega} \Big[ P(x) \log \frac{Q(x)}{M(x)} \Big], \tag{5}$$

where $\mathcal{P}$ is the set of probability distributions over $\Omega$. The maximum-entropy distribution in a set $\mathcal{C}$, relative to measure $M$, is thus [5]

$$\mathcal{H}_M(P_M^{me}) = \sup_{P \in \mathcal{C}} \mathcal{H}_M(P) = \sup_{P \in \mathcal{C}} \inf_{Q \in \mathcal{P}} - \sum_{x \in \Omega} \Big[ P(x) \log \frac{Q(x)}{M(x)} \Big]. \tag{6}$$

Here is where game theory steps in, since the above is the *maximin gain* of a certain *two-player zero-sum game*. In the game, loss is measured with the *logarithmic loss* function. We will return to the loss function shortly.

The game can be interpreted as follows [5]: Player 1 - let us call her the Agent - *repeatedly* takes part in a lottery that has $n$ possible outcomes $\Omega = \{x_1, x_2, \ldots, x_n\}$. Each round, one outcome is chosen by player 2 - the Bank - as the winner according to a fixed probability distribution $P^*$. The only thing the Agent knows about the distribution is that it is chosen from the set $\mathcal{C}$. The game is played by betting money on the outcomes in $\Omega$; if outcome $x_i$ occurs, the Agent gets back a constant $b$ times the amount of money she invested in that outcome. All the money invested in the other outcomes is lost and goes to the Bank.

The Agent has some initial capital that she reinvests each round: all her money is on each round used to bet on outcomes in the lottery. We will examine this constraint shortly, but let us assume it for now.

The different choices facing the Agent, i.e. how much money to invest in each outcome, can be described as a *gambling strategy* that is formally a probability distribution $Q$ in $\Omega$: if the agent has $K$ units of money, she invests $K \cdot Q(x_i)$ units of money on outcome $x_i$ for all $i = 1 \ldots n$. Similarly, the Bank's (fixed) strategy is $P^*$.

Von Neumann's *minimax theorem* [1] states that every two-person zero-sum game has a solution in randomized (probabilistic) strategies. More formally, the game then has a definite *value* $V$:

$$V = \inf_{Q \in \mathcal{P}} \sup_{P \in \mathcal{C}} L(P, Q) = \sup_{P \in \mathcal{C}} \inf_{Q \in \mathcal{P}} L(P, Q), \tag{7}$$

where $L(P, Q)$ is the expected loss to player 1 when he is using the randomized strategy $Q$ and player 2 is using the randomized strategy $P$. Player 1 is naturally attempting to minimize her loss. Since the game is zero-sum, the goal of player 2 is just the opposite: to maximize the loss of player 1. The loss function in this case is $L(P, Q) = -\sum_{x \in \Omega_X} P(x) \log \frac{Q(x)}{M(x)}$, which, if $M(x) = 1$ for all $x$, can be interpreted as the number of bits needed to code $P$ using a code based on $Q$. But there is also another interpretation.

Equation (7) says that it does not matter whether we maximize with respect to the minimum or minimize with respect to the maximum - the result will be the same. The solution corresponds to a pair of strategies, one for each player, that neither player has incentive to alter, since they could not expect to do better by adopting a different strategy. What that means is that if player 1 changes her strategy, then the other player, player 2, has a counterstrategy that makes player 1 worse off on the average, and vice versa. The solution to the game is stable in the sense that either player can safely announce her strategy in advance, knowing that the other player cannot gain anything by the knowledge. It should be noted, though, that the solution assumes two perfectly rational players who have perfect knowledge of the mechanics of the game and also possess unlimited powers of reasoning.

In terms of the game above, both players choose a strategy: the Bank's strategy, constrained to lie in the set $\mathcal{C}$, is $P^*$, and the Agent's strategy is $Q$. Von Neumann's theorem does not cover the case where there are constraints on the set of randomized strategies the other player may choose. Grünwald [5] states that if $\mathcal{C}$ is *convex*, then a unique solution exists. Convexity is ensured e.g. in the case of expected-value constraints. Furthermore, the unique solution to the game is the MaxEnt distribution for the Agent, and it is an *equalizer strategy*: it does not matter in terms of logarithmic loss which distribution $P^*$ the Bank chooses: all distributions in $\mathcal{C}$ are equal in that they result in the same logarithmic loss for the Agent using the MaxEnt distribution. Therefore, whether or not $Q$ is correct, the Agent knows how well - in terms of logarithmic loss - she will do on the average.

Why involve logarithmic loss here? It turns out that logarithmic loss can be interpreted as the *expected growth rate* of capital in this type of games [2]. The MaxEnt distribution thus maximizes the worst-case expected growth rate of capital. Now we reexamine the constraint that the Agent has to reinvest all her money on each round of the game. There are three alternatives, depending on the constant $b$ and the set $\mathcal{C}$: the game is set up such that either the Agent can make money on the average, lose money, or make no profit nor loss. If the odds are such that the Agent cannot make money on the average, it is no longer optimal to reinvest everything; it should be clear that it is not even optimal to play the game at all if all we are interested in is the expected growth rate of capital. In the first case, on the other hand, it makes sense to put all available money in the lottery.

The game-theoretic interpretation does shed new light on MaxEnt, although the situation where it could be applied does not seem very natural: even though the Agent plays the game repeatedly, she does not learn anything about the distribution other than possibly its mean, which translates to the constraint that the strategy of the Bank must satisfy.

Also, why would the Agent wish to maximize *worst-case* loss in prediction in general, even though it is in this case a solution to a game? Presumably, one does not often apply MaxEnt in situations where the opponent is an actual adversary such as the Bank, but rather some unknown process, often called Nature, that is independent and uncaring of the choices one makes. We illustrate this with a simple example.

> *Example 2.* Let us suppose that the only knowledge we have about a coin is that its probability of coming up heads is in the range $[0.6, 0.9]$. Then we choose the uniform measure $M(heads) = M(tails) = 1$; the distribution maximizing the entropy is $P_M^{me}(\text{heads} = 0.6)$, which may seem unintuitive, since the probability is picked from the boundary instead of the interior. But it does optimize worst-case loss. Of course, situations such as above may not come up frequently in practice.

## 5 Representation Dependence in MaxEnt

An oft-cited objection to MaxEnt is its *representation dependence*. Representation here essentially means the sample space $\Omega$, the measure $M$ and the interpretation given to the two. An intuitive notion is that if the same reasoning problem is represented in a different way, any inference method should still lead to the same results. However, Halpern and Koller [6] show that no reasonable method of probabilistic inference can be absolutely independent of representation, and that any probabilistic reasoning system must make a compromise between representation independence and other desiderata.

So far, we have largely ignored the role of the underlying measure $M$ in MaxEnt. What happens if we transform our sample space $\Omega$ to $\Omega_Y$? This assumes that we can represent the same problem (the set of constraints) in $\Omega_Y$. It is known that if we transform the measure along with the sample space, then both representations give the same results.

> *Example 3.* Our original sample space is $\Omega = \{a, b\}$ with a uniform measure $M(x) = 1$ and no constraints. In this case, MaxEnt prescribes the uniform distribution. We transform the problem to a new sample space $\Omega_Y = \{a_1, a_2, b\}$, where the previous element $a$ is split into $a_1$ and $a_2$; for instance, instead of just "animal", we now distinguish between the cases "vertebrate" and "invertebrate". If we also transform the measure such that $M_Y(a_1) = M_Y(a_2) = 0.5$ and $M_Y(b) = 1$, the maximum entropy distribution for both representations remains the same: $P_Y(a_1) = P_Y(a_2) = 0.25$ and $P_Y(b) = 0.5$.

Grünwald [5] also covers the case where we would like to obtain a distribution for some random variable $Y$, but the constraints still concern random variable $X$. We will not discuss that case, even though it comprises the real contribution of the paper.

## 6 Concluding Remarks

Jaynes [9] has advocated MaxEnt aggressively as the cornerstone of an objective methodology of inductive inference. Grünwald [4, 5] takes a more cautious route with his characterization of reliable prediction. Attempts have been made [9, 10] to derive MaxEnt from some intuitive and compelling desiderata, but all of these attempts seem to have been flawed [11]. The concentration phenomenon remains the most powerful argument in favor of MaxEnt.

A pragmatic way of viewing MaxEnt is that it is simply a reasonable set of independence assumptions that has some desirable properties. The motive for assumptions is often a desire to limit the complexity of the model.

# References


[1] J. Berger. *Statistical Decision Theory and Bayesian Analysis, 2nd ed.* Springer-Verlag, New York, 1985.

[2] T. Cover and J. Thomas. *Elements of Information Theory.* Wiley-Interscience, New York, 1991.

[3] R. L. Ellis. Remarks on an alleged proof of the method of least squares, contained in a late number of the *edinburgh review.* In W. Walton, editor, *Mathematical and other Writings of R. L. Ellis*, pages 53–61. Cambridge University Press, Cambridge, 1863.

[4] Peter Grünwald. *The Minimum Description Length Principle and Reasoning Under Uncertainty.* PhD dissertation, University of Amsterdam, Institute for Logic, Language and Computation, 1998.

[5] Peter Grünwald. Maximum entropy and the glasses you are looking through. In *Proc. Uncertainty in Artificial Intelligence (UAI-00)*, 2000.

[6] J. Halpern and D. Koller. Representation dependence in probabilistic inference. In *Proc. Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95)*, pages 1853–1860, 1995.

[7] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957.

[8] E. T. Jaynes. *Papers on Probability, Statistics and Statistical Physics.* Kluwer Academic Publishers, 1989.

[9] E. T. Jaynes. *Probability Theory: the Logic of Science.* A preliminary version, 1996. Available at http://bayes.wustl.edu/.

[10] J. Shore and R. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, IT-26:26–37, 1980.

[11] Jos Uffink. Can the maximum entropy principle be explained as a consistency requirement? *Studies in History and Philosophy of Modern Physics*, 26B:223–261, 1995.