# Maximum Entropy Models

Tor Erlend Fjelde

July 4, 2018

## Contents

# 1 Issues

☐ Define entropy instead as a function of a *random variable* rather than a functional acting on probability density $p(X)$

# 2 Introduction

# 3 Background

## 3.1 Shannon / Information Entropy

Let $X$ be a *discrete* random variable, i.e. taking values from a at most countable sample space $\Omega_X$. The probability mass of $X$ is defined by $p_X : \mathscr{P}(\Omega_X) \to [0,1]$ such that $\sum_{x \in \Omega_X} p_X(x) = 1$, where we use the notation $\mathscr{P}(\Omega_X)$ to mean the *powerset of* $\Omega_X$. The probability measure $P_X$ is then

$$P_X(A) = \sum_{x \in A} p_X(x), \quad \forall A \subseteq \Omega_X$$

We will refer to the triple $\big(P, \Omega_X, \mathscr{P}(\Omega_X)\big)$ as a *probability space*.[1]

Now suppose we want to define some function which describes the *information* obtained from observing some event $A \subseteq \Omega_X$. Intuitively one would expect such a function, denoted $S_p$, to have the following properties:

1. $S_p$ is *monotonically decreasing* in $P$, i.e. if $P(A)$ is large, we gain little information by observing $A$,

2. $S_p(A) \geq 0, \quad \forall A \subseteq \Omega_X$, i.e. information is a non-negative quantity,

3. if $p_X(A) = 1$ for some $A \subseteq \Omega_X$, then $S_p(A) = 0$, i.e. events with probability 1 provides no information,

4. if $X$ and $Y$ are two *independent* random variables with probability masses $p_X$ and $p_Y$, respectively, then on the extended probability space $\big(P_X P_Y, \Omega_X \times \Omega_Y, \mathscr{P}(\Omega_X \times \Omega_Y)\big)$

   $$S_p(A \times B) = S_p(A) + S_p(B), \quad \forall A \subseteq \Omega_X, B \subseteq \Omega_Y,$$

   i.e. joint information of independent events are added together.[2]

It turns out that this is satisfied by the following definition of $S_p$.[1]

**Definition 3.1.** The **Shannon entropy** or **information entropy** $S_p$ over a discrete probability space $\big(P, \Omega, \mathscr{P}(\Omega)\big)$ is defined

$$S_p(A) = -\sum_{x \in A} p(x) \log p(x), \quad \forall A \subseteq \Omega$$

*Note that this is true for any probability space, hence entropy might also be defined for subsets of $\mathscr{P}(\Omega)$ for which we have an adapted probability measure.*

It is interesting to note that $S_p$ is *minimized* when $p(x) = 1$ and $p(y) = 0$ for all $y \neq x$, as we would be certain of the outcome, and *maximized* if $p(x) = p(y)$, i.e. $p$ is the uniform distribution, where all outcomes are equally likely. This coincides with what one would untuitively expect from a measure of information, where taking on the uniform distribution is the sensible thing to do when we have no prior information.

[1] A term often used when talking about probabilities in measure theory.

[2] Shannon information is often denoted $H$ in the informatics / mathematics literature due to its relatedness to the *entropy* of thermodynamics.

### 3.1.1 TODO Continuous random variables

Shannon also introduced entropy of probability densities of *continuous* random variables by simply substituting the summation in Definition 3.1 by an integrand over all $x$. This is often referred to as the **differential entropy**, given by

$$S_d(p) = -\int p(x)\log p(x)\ dx$$

Unfortunately this definition does not share some of the desirable properties of the discrete entropy. As we will see it fails to satisfy non-negativity, transformation invariance, and, though sometimes stated to be, does not in general correspond to taking the number of hypotheses infinity in Definition 3.1.

That $S_d$ can take on negative values is easily seen from considering the distribution $\text{Uniform}(0, 1/2)$. Then the *differential entropy* gives us

$$S_d(p) = -\int_0^{1/2} 2\log 2\ dx = -\log 2 = -1$$

which contradicts the first axiom of information entropy as defined for the discrete case.

Using the same case we can also demonstrate the scale *variance* of $S_d(p)$. Let $Y = aX$ for some $a \in \mathbb{R}$, then

$$S_d(p_Y) = -\int$$

Therefore the following definition for entropy in the case of continuous functions was proposed by Jaynes: [2]

**Definition 3.2.** Jaynes provided the following definition for the continuous case

$$S(p) = -\int p(x)\log\frac{p(x)}{m(x)}\,\mathrm{d}x$$

where $m(x)$, called the *invariant factor*, is proportional to the limiting density of discrete points.

The *invariant factor* is simply a function which enforces *non-negativity* and *scale invariance*, i.e. if $Y = aX$ where $X$ is a random random variable and $a \in \mathbb{R}$, then

$$S(p_X) = S(p_Y)$$

$m(x)$ is therefore highly dependent on the distribution in question and the transformation being made. Note that his property is most definitively something we want from a "measure of information", since $p_X = p_Y$.

## 3.2 Cross-entropy and Kullback-Leibner divergence

Similar to information entropy as described in Definition 3.1, we have **cross-entropy** which is way of comparing two distributions $p$ and $q$. Suppose we have a message following the distribution $p$, then the optimal encoding of the message has the average length given by the information entropy, using $\log_2$. Then suppose we are instead using an encoding which is optimal for some *other* distribution $q$. Then the average length of the encoded message from $p$ is given by

$$-\sum_x p(x)\log q(x)$$

which is the **cross-entropy** between $p$ and $q$, and has the property that it is always greater than or equal to $-\sum_x p(x)\log p(x)$.

Suppose we have $N$ samples $\{x_i\}$ from $p$, then observe that the negative log-likelihood for some parametrizable distribution $q$ is given by

$$-\frac{1}{N}\sum_{i=1}^{N}\log q(x_i \mid \theta)$$

By the Law of Large numbers, we have

$$\lim_{N\to\infty} -\frac{1}{N}\sum_{i=1}^{N}\log q(x_i \mid \theta) = -\mathbb{E}_p[\log q(x \mid \theta)]$$

where $\mathbb{E}_p$ denotes the expectation over the probability density $p$. But this is exactly the cross-entropy given above. This implies that the $\theta$ which minimizes the cross-entropy is also the $\theta$ which maximizes the likelihood as the number of samples goes to infinity.

This is often used in classification tasks as the objective one wants to maximize wrt. parameters $\theta$ of the model.

### 3.2.1 Kullback-Leibler divergence

A similar measure which is also often used is the **Kullback-Leibler (KL) divergence** or **Relative entropy**, of which the cross-entropy is involved, defined

$$D_{\mathrm{KL}}(p\,||\,q) = \mathbb{E}_p\left[\log\frac{p(x \mid \theta^*)}{q(x \mid \theta)}\right] = \mathbb{E}_p\big[\log p(x \mid \theta^*)\big] - \mathbb{E}_p\big[\log q(x$$

where $\theta^*$ is the optimal parameter and $\theta$ is the one we vary to approximate $p(x \mid \theta^*)$. The second term here is simply the cross-entropy above, and since $\theta^*$ is a constant, the minimization process would only depend on the second term, i.e. we have the same property of giving us the same estimate as the MLE for large $N$. Observe that in the case where $p(x \mid \theta^*) = q(x \mid \theta)$ for all $x$, the KL divergence clearly vanishes, which we would want from a similarity "measure".

In fact, the KL divergence is also well-defined for continuous distributions, unlike the information entropy, as discussed earlier in Section 3.1.1.

It is worthy to note that the KL divergence is not *symmetric*, and therefore does not define a proper metric. Nonetheless, one can easily construct a metric between probability distributions from the KL divergence, called the **Jensen-Shannon entropy**.[3]

# 4 Maximum Entropy models

## 4.1 Principle of Maximum Entropy

Suppose we have a $N$ *samples* $A = \{x_i \in \mathcal{X}, \forall i = 1, \dots, N\}$ of a *discrete* random variables $X$, drawn from a probability density $p(x)$, and we want to estimate $p$ from $A$. Let $q(x)$ denote our estimate of $p(x)$; that is, $q$ is our *model*.

In this case, the **Principle of Maximum Entropy** states that one should choose $q$ with the *largest uncertainty*; that is, choose $q$ such that

$$\max_{q'} S_{q'}(A) = S_q(A)$$

subject to constraints that the expectations of some *observed* functions $\{f_i : \mathcal{X} \to \mathbb{R}\}$ under our estimate $q$ are equal to their *observed* expectations under the $p$.[4] Formally,

$$\begin{aligned} \underset{q}{\text{minimize}} \quad & - S_q(A) \\ \text{subject to} \quad & \sum_x q(x) = 1, \quad \mathbb{E}_q[f_i] = F_i, \ i = 1, \dots, k. \end{aligned}$$

for some $F_i$, where the constrains are ensuring that $q$ is a probability distribution matching these observed expectations.

### 4.1.1 Justification

As discussed in Section 3.1, we know that the uniform distribution maximizes the entropy without any constraints. Principle of Maximum Entropy can therefore be considered as a

Now we will see an argument proposed by Graham Wallis to Jaynes, which has the benefit of being stricly combinatorial in nature, not making any references to information entropy.[5]

Suppose one wants to assign probabilities to $m$ mutually exclusive propositions, and at the same time enforce constraints corresponding to testable information, e.g. expectation of $X$ is 0.3. One could then distribute $N$ quanta of probability at random to the $m$ prepositions, enforcing the constraints by starting over again if they are not satisfied. Upon reaching a distribution of these quanta which satisfies the constraints, then the probabilities are given by

$$p_i = \frac{n_i}{N}, \quad i = 1, \dots, m$$

where $n_i$ are the number of quanta assigned to the i-th proposition. This has the same distribution as the case where one randomly places $N$ balls in $m$ different buckets, i.e. a multinomial distribution,

$$P(\mathbf{p}) = \frac{W}{m^N}, \quad W = \frac{N!}{n_1! n_2! \dots n_m!}$$

where $\mathbf{p}$ is an $N$ dimensional vector with the j-th entry corresponding to which "bucket" the j-th "ball" was assigned. To obtain the most probable $\mathbf{p}$, one simply maximizes $W$, since $m^{-N}$ is of course independent of $\mathbf{p}$. Equivalently, one can maximize any monotonically increasing function of $W$, e.g.

$$\begin{aligned} \frac{1}{N} \log W &= \frac{1}{N} \log \frac{N!}{n_1! n_2! \dots n_m!} \\ &= \frac{1}{N} \log \frac{N!}{(Np_1)!(Np_2)! \cdots (Np_m)!} \\ &= \frac{1}{N} \left( \log N! - \sum_{i=1}^{m} \log \left( (Np_i)! \right) \right) \end{aligned}$$

Letting $N \to \infty$, and making use of Stirling's approximation, we get

$$\begin{aligned} \lim_{N \to \infty} \left( \frac{1}{N} \log W \right) &= \frac{1}{N} \left( N \log N - \sum_{i=1}^{m} N p_i \log \left( N p_i \right) \right) \\ &= \log N - \sum_{i=1}^{m} p_i \log(N p_i) \\ &= \log N - \log N \sum_{i=1}^{m} p_i - \sum_{i=1}^{m} p_i \log p_i \\ &= \left( 1 - \sum_{i=1}^{m} p_i \right) \log N - \sum_{i=1}^{m} p_i \log p_i \\ &= - \sum_{i=1}^{m} p_i \log p_i \\ &= H(\mathbf{p}) \end{aligned}$$

Then, maximizing this wrt. the constraints of the testable information, we have arrived at the Principle of Maximum Entropy.

## 4.2 General expression for MaxEnt distribution

Formulating the optimization problem in Equation 4.1 using the method of Lagrange multipliers, letting $F_i = \langle f_i \rangle_A$, i.e. the expectations over $q$ constrained to be equal to the *empirical* expectation, we have

$$\mathcal{L}[q] = -S_q + \sum_{i=1}^{m} \lambda_i \left( \langle f_i \rangle_A - \sum_x f_i(x) q(x) \right) + \gamma \left( 1 - \sum_x \mathrm{d}x \, q(x) \right)$$

We solve for $q$ by taking the functional derivative and letting it equal zero

$$0 = \frac{\delta \mathcal{L}}{\delta q} = \left( \log q(x) \right) + 1 - \gamma - \sum_{i=1}^{m} \lambda_i f_i(x)$$

We observe that this is satisfied by

$$q(x) = \frac{1}{Z} \exp \left( \sum_{i=1}^{m} \lambda_i f_i(x) \right)$$

where

$$Z(\lambda_1, \ldots, \lambda_m) = \sum_x \exp\left(\sum_{i=1}^m \lambda_i f_i(x)\right)$$

is the *partition function*. This expression for $q$ is therefore the *general form* of the **maximum entropy distribution**. The Lagrange multipliers are then given by the convenient expressions

$$\langle f_i \rangle_A = \frac{\partial}{\partial \lambda_k} Z(\lambda_1, \ldots, \lambda_m)$$

It is worth noting that due to the invariant measure $m(x)$ in Definition 3.2, the procedure for continuous random variables does not necessarily give the same maximum entropy model under the same constraints.

## 4.3 Principle of Maximum Entropy (POME) and Maximum Likelihood Estimate (MLE)

As noted by Jaynes, it is important to note that there is clear distinction between POME and MLE.[6] POME is used for construction of models *a priori*, i.e. without any data. Using POME one specifies constraints on the *space of hypothesis*; one could say it is making a choice of which family of densities one wants to use for the model. Given the space of possible distributions which satisfy the stated constraints, POME provides a method of choosing which family to use, specifically it makes one choose the distribution which assumes the least amount information. What POME does *not* do, is provide us with a method for finding the "optimal" (in some specified sense) parameters *within* this family of distributions, i.e. if one of the constrains are $\mathbb{E}[f_i] = F_i$, we want the optimal $F_i$ wrt. the data. An example is in the case where we have a continuous random variable, where POME with the first and second moment constrained to some $\mu$ and $\sigma^2$, respectively. This gives us a Guassian distribution with mean $\mu$ and variance $\sigma^2$, but it does not specify which values $\mu$ and $\sigma^2$ should take.

This is exactly where the Bayesian line of thinking comes in, proposing that one ought to choose the parameters of the distribution such that one maximizes the likelihood, or equivalently the log-likelihood, of the data. That is, given a set of data $\{x_i\}$ which is believed to come from some parametrizable distribution $p_\theta$, then we want to find $\theta$ such that

$$\max_\theta p(\{x_i\} \mid \theta)$$

and since

$$p(\{x_i\} \mid \theta) = \frac{p(\theta \mid \{x_i\}) p(\theta)}{p(\{x_i\})}$$

we can equivalently work with the following objective function

$$\max_\theta p\big(\theta \mid \{x_i\}\big) p(\theta)$$

where the first factor is the *likelihood* of the data. In Section we will see how there are objectives which are heavily related to the information entropy, which gives rise to MLE estimates in the limiting case of number of samples.

### 4.3.1 MaxEnt for finding priors

In the case described above, one still needs to specify the *prior* $p(\theta)$ for the parameters. Often one does not have any prior information and would like the prior $p(\theta)$ to be as "uninformative" as possible. Therefore it is often suggested that one ought to use POME to find suitable priors, since this would supposedly give us the least informative distribution under the constraints.[2] For example, if one is working on a non-degenerate interval on the real line and do not have any prior information of $\theta$, then the uniform distribution seems intuitvely to be the best choice, which is the same as POME would give us.

MacKay suggests one should not even use POME for finding priors.[1] He brings up the following example: suppose one has an unfair die which we know has an average of 2.5. What is the probabilities $\{p_1, p_2, p_3, p_4, p_5, p_6\}$? MacKay states that such questions are not even sensible. Still, one could argue that the resulting maximum entropy model is still better than nothing.

## 4.4 "Over-constraining"

When using POME to construct MaxEnt models is important to be *reasonable* in what constraints one enforce. Consider the case where we have observations $\{x_i, i = 1, \ldots, n\}$ of a discrete variables $X$. Let each of the constrained functions $f_i$ be defined as

$$f_i(x) = \begin{cases} 1 & \text{if } x = x_i \\ 0 & \text{otherwise} \end{cases}, \quad i = 1, \ldots, n$$

Then our constraints are simply

$$\mathbb{E}[f_i] = \sum_x f_i(x) p(x) = p(x_i), \quad i = 1, \ldots, n$$

That is, we have constrained the entire distribution. In this case it does not really make any sense to talk about "uninformativeness" of a distribution when we have literally constrained it to take on exactly the values we want. Therefore one should be careful when choosing constraints.

# 5 Applications

## 5.1 Natural Language Processing

Maximum entropy models are used frequently in the natural language processing community, often in the context of large graphical models. [7] Usually this simply means that they are using exponential models, which we know are MaxEnt models with certain constraints, and using some form of "relative" entropy as the objective function, as discussed in Section 4.3.[8]

Models in the exponential family have some very nice properties for performing inference. The convexity of the exponential function makes optimization significantly easier, and computation can be simplified substantially by making use of log in many places rather than the exponentials. The optimization procedure is the same since log is monotonically increasing wrt. exp. The usage of log also has the convenient effect of being more numerically stable, which is a huge benefit when performing numerical approximations.

### 5.2 Statistical Physics

Entropy in physics was deduced using thermodynamic principles through the use of ensembles, which is very similar to the alternative justification we saw in Section 4.1.1. This was later related to Shannon entropy, and the probability densities obtained for these ensembles were argued to be nothing more than maximum entropy models with different constraints depending on which type of ensemble one was working with, e.g. isotropic system.[9]

## References

[1] D.J.C. MacKay, D.J.C.M. Kay, and Cambridge University Press. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

[2] Edwin T Jaynes. Prior probabilities. *IEEE Transactions on systems science and cybernetics*, 4(3):227–241, 1968.

[3] D.M. Endres and J.E. Schindelin. A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858–1860, Jul 2003.

[4] Pankaj Mehta, Marin Bukov, Ching-Hao Wang, Alexandre G. R. Day, Clint Richardson, Charles K. Fisher, and David J. Schwab. A high-bias, low-variance introduction to machine learning for physicists. *CoRR*, 2018.

[5] *Probability theory: the logic of science*, volume 2.

[6] Edwin T Jaynes. The relation of bayesian and maximum entropy methods. *Maximum-Entropy and Bayesian Methods in Science and Engineering (Vol. 1)*, 1988.

[7] Hai Leong Chieu and Hwee Tou Ng. Named entity recognition. *Proceedings of the 19th international conference on Computational linguistics -*, 2002.

[8] Adam L Berger, Vincent J Della Pietra, and Stephen A Della Pietra. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71, 1996.

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