Maximum Entropy and Exponential Families

Christopher Ré
(edits by Tri Dao and Anand Avati)

August 5, 2019

Abstract

The goal of this note is to derive the exponential form of probability distribution from more basic considerations, in particular Entropy. It follows a description by ET Jaynes in Chapter 11 of his book *Probability Theory: the Logic of Science* [1].[1]

1 Motivating the Exponential Model

This section will motivate the exponential model form that we've seen in lecture.

The Setup The setup for our problem is that we are given a finite set of instances $\mathcal{Y}$ and a set of m statistics (T_j, c_j) in which $T_j : \mathcal{Y} \to \mathbb{R}$ and $c_j \in \mathbb{R}$. An instance (or possible world) is just an element in a set. We can think about a statistic as a measurement of an instance, it tells us the important features of that instance that are important for our model. More precisely, the only information we have about the instances is the values of T_i on these instances. Our goal is to find a probability function p such that

$$p : \mathcal{Y} \to [0, 1] \text{ such that } \sum_{y \in \mathcal{Y}} p(y) = 1.$$

The main goal of this note is to provide a set of assumptions under which such distributions have a specific functional form, the exponential family, that we saw in generalized linear model:

$$p(y; \eta) = \exp\{\eta \cdot T(y) - a(\eta)\}$$

in which $\eta \in \mathbb{R}^m$, $T(y) \in \mathbb{R}^m$ and $T(y)_j = T_j(y)$. Notice that there is exactly one parameter for each statistic. As we'll see for discrete distributions, we are able to derive this exponential form as a consequences of a maximizing entropy subject to matching the statistics.[2]

[1] This work is available online in many places including `http://omega.albany.edu:8008/ETJ-PS/cc11g.ps`.

[2] Unfortunately, for continuous distributions, such a derivation does not work due to some technical issues with Entropy–this hasn't stopped folks from using it as justification.

1.1 The problem: Too many distributions!

We'll see the problem of defining a distribution from statistics (measurements). We'll see that often there are often many probability distributions that satisfy our constraints, and we'll be forced to pick among them.[3]

The Constraints We interpret a statistic as a constraint on p of the following form:

$$\mathbb{E}_p[T_j] = c_j \text{ i.e., } \sum_{i=1}^{N} T_j(y_i)p_i = \langle T_j, p\rangle = c_j$$

Let's get some notation to describe these constraints. Let $N = |\mathcal{Y}|$, then the probability we are after is $p \in \mathbb{R}^N$ subject to constraints.

- There are m constraints of the form

$$\langle T_j, p\rangle = c_j \text{ for } j = 1, \ldots, m.$$

- A single constraint of the form $\sum_{i=1}^{N} p_i = 1$ to ensure that p is a probability distribution. We can write this more succinctly as $\langle \mathbf{1}, p\rangle = 1$.
- We also have that $p_i \geq 0$ for $i = 1, \ldots, N$.

More compactly, we can write all constraints in a matrix G as

$$G = \begin{pmatrix} \mathbf{1} \\ T \end{pmatrix} \in \mathbb{R}^{(m+1)\times N} \text{ so that } Gp = \begin{pmatrix} 1 \\ c \end{pmatrix}.$$

If $\mathsf{N}(G) = \emptyset$, then this means that p is uniquely defined as G has an inverse. In this case, $p = G^{-1}c$. However often m is much smaller than N, so that $\mathsf{N}(G) \neq \emptyset$–and there are many solutions that satisfy the constraints.

Example 1.1 *Suppose we have three possible worlds, i.e., $\mathcal{Y} = \{y_1, y_2, y_3\}$ and one statistic $T(y_i) = i$ and $c = 2.5$. Then, we have:*

$$G = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{pmatrix} \text{ and } \mathsf{N}(G) = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}$$

Let $p^{(1)} = (1/12, 1/3, 7/12)$ then $Gp = (1, 2.5)^T$–but so do (infinitely) many others, in particular $q(\alpha) = p^{(1)} + \alpha(1, -2, 1)$ is valid so long as $\alpha \in [-1/12, 1/6]$ (due to positivity).

[3]Throughout this section, it will be convenient to view p and T_j as functions from $\mathcal{Y} \to \mathbb{R}$–and also as vectors indexed by $\mathcal{Y}$. Their use should be clear from the context.

Picking a probability distribution p In the case $\emptyset \neq \mathsf{N}(G)$, there are *many* probability distributions we can pick. All of these distributions can be written as follows:

$$p = p^{(0)} + p^{(1)} \text{ in which } p^{(0)} \in \mathsf{N}(G) \text{ and } p^{(1)} \text{ satisfies } Gp^{(1)} = \begin{pmatrix} 1 \\ c \end{pmatrix}$$

Example 1.2 *Continuing the computation above, we see* $p^{(0)} = \alpha(1, -2, 1)$ *is a vector in* $\mathsf{N}(G)$.

Which p should we pick? Well, we'll use one method called the method of maximum entropy. In turn, this will lead to the fact that our function p has a very special form–the form of exponential family distributions!

1.2 Entropy

To pick among the distributions, we'll need some scoring method.[4] We'll cut to the chase here and define the entropy, which is a function on probability distributions $p \in \mathbb{R}^N$ such that $p \geq 0$ and $\langle \mathbf{1}, p \rangle = 1$.

$$H(p) = -\sum_{i=1}^{N} p_i \log p_i$$

Effectively, the entropy rewards one for "spreading" the distribution out more. One can motivate Entropy from axioms, and either Jaynes or the Wikipedia page is pretty good on this account.[5]. The intuition should be that entropy can be used to select the *least informative* prior, it's a way of making as few additional assumptions as possible. In other words, we want to encode the prior information given by the constraints on the statistics while being as "objective" or "agnostic" as possible. This is called the maximum entropy principle.

For example, one can verify that under no constraints, $H(p)$ is maximized with $p_i = N^{-1}$– that is all alternatives have equal probability. This is what we mean by spread out.

We'll pick the distribution that maximizes entropy subject to our constraints. Mathematically, we'll examine:

$$\max_{p \in \mathbb{R}^N} H(p) \text{ s.t. } \langle \mathbf{1}, p \rangle = 1, p \geq 0, \text{ and } Tp = c$$

We will not discuss it, but under appropriate conditions there is a unique solution p.

[4] A few natural methods don't work as we might think they should (minimizing variance, etc.) See [1, Ch.11] for a description of these alternative approaches.

[5] `https://en.wikipedia.org/wiki/Entropy_(information_theory)#Rationale`

1.3 The Lagrangian

We'll create a function called the Lagrangian that has the property that any critical point of the Lagrangian is a critical point of the *constrained* problem. We will show that all critical points of the Lagrangian (and so our original problem) can be written in the exponential format we described above.

To simplify our discussion, let's imagine that $p > 0$, i.e,. there are no possible worlds y such that $p(y) = 0$. In this case, our problem reduces to:

$$\max_{p \in \mathbb{R}^N} H(p) \text{ s.t. } Tp = c \text{ and } \langle \mathbf{1}, p \rangle = 1$$

We can write the Lagrangian $\mathcal{L} : \mathbb{R}^N \times (\mathbb{R}^m \times \mathbb{R}) \to \mathbb{R}$ as follows:

$$\mathcal{L}(p; \eta, \lambda) = H(p) + \langle \eta, Tp - c \rangle + \lambda \left(\langle \mathbf{1}, p \rangle - 1 \right)$$

The special property of $\mathcal{L}$ is that any critical point of our original solution, in particular any maximum or minimum corresponds to a critical point of the Lagrangian. Thus, if we prove something about critical points of the Lagrangian, we prove something about the critical points of the original function. Later in the course, we'll see more sophisticated uses of Lagrangians but for now we include a simple derivation below to give a hint what's going on. For this section, we'll assume this special property is true.

Due to that special property, we find the critical points of $\mathcal{L}$ by differentiating with respect to p_i and setting the resulting equations to 0.

$$\begin{aligned} \frac{\partial}{\partial p_i} \mathcal{L} &= \frac{\partial}{\partial p_i} \left[H(p) + \langle \eta, Tp - c \rangle + \lambda (\langle \mathbf{1}, p \rangle - 1) \right] \\ &= -(\log p_i + 1) + \langle \eta, T(y_i) \rangle + \lambda \end{aligned}$$

Setting this expression equal to 0 and solving for p_i we learn:

$$\begin{aligned} p_i &= e^{\lambda - 1} \exp \left\{ \langle \eta, T(y_i) \rangle \right\} \\ \Rightarrow p(y) &\propto \exp\{\eta \cdot T(y)\} \end{aligned}$$

which is of the right form–except that we have one too many parameters, namely λ. Nevertheless, this is remarkable: at a critical point, it's always the case that the exponential family "pops out"!

Eliminating λ The parameter λ can be eliminated, which is the final step to match our original claimed exponential form. To do so, we sum over all the p_i which we know on one hand is equal to 1, and the other hand, we have the above expression for p_i. This gives us the following equation:

$$\sum_{i=1}^N p_i = 1 \text{ and } \sum_{i=1}^N p_i = e^{\lambda - 1} \left(\sum_{i=1}^N \exp \left\{ \eta \cdot T(y_i) \right\} \right) \text{ thus } e^{-\lambda + 1} = \left(\sum_{y \in \mathcal{Y}} \exp \left\{ \eta \cdot T(y) \right\} \right)$$

Thus, we have expressed λ as a function of η and we can eliminate it. To do so, we write:

$$\begin{aligned} Z(\eta) &= \sum_{y \in \mathcal{Y}} \exp\{\eta \cdot T(y)\} \\ \Rightarrow p(y;\eta) &= Z(\eta)^{-1} \exp\{\eta \cdot T(y)\} \\ &= \exp\{\eta \cdot T(y) - a(\eta)\} \quad \text{where } a(\eta) = \log Z(\eta) \end{aligned}$$

This function Z is called the *partition function*, and a is called the *log-partition function*. The above is the claimed exponential form we saw in lecture.

2 Why the Lagrangian? [optional]

We observe that this is a constrained optimization problem with *linear* constraints.[6]

Let r be the rank of G and so $\mathsf{dim}(\mathsf{N}(G)) = N - r$. We create a function $\phi : \mathbb{R}^{N-r} \to \mathbb{R}$ such that there is a map between any point in the domain of ϕ and a feasible solution to our constrained problem, and moreover ϕ will take the same value as H. In contrast to our original constrained problem, ϕ has an unconstrained domain (all of $\mathbb{R}^{N-r}$), and so we can apply standard calculus to find its critical points. To that end, we define a (linear) map $B \in \mathbb{R}^{N \times (N-r)}$ that has rank $N - r$. We also insist that $B^T B = I_{N-r}$. Such a B exists, as it is simply the first $N - r$ columns of a change of basis matrix from the standard basis to an orthonormal basis for $\mathsf{N}(G)$. We have

$$\phi(x) = H(Bx + p^{(1)}),$$

where $p^{(1)}$ is a fixed vector satisfying $Gp^{(1)} = \begin{pmatrix} 1 \\ c \end{pmatrix}$.

Observe that for any $x \in \mathbb{R}^{N-r}$, $Bx \in \mathsf{N}(G)$ so that $G(Bx + p^{(1)}) = Gp^{(1)} = \begin{pmatrix} 1 \\ c \end{pmatrix}$ and so $Bx + p^{(1)}$ is feasible. Moreover, B is a bijection from $\mathbb{R}^{N-r}$ to the set of feasible solutions.[7] Importantly, ϕ is now *unconstrained*, and so any saddle point (and so any maximum or minimum) must satisfy:

$$\nabla_x \phi(x) = 0$$

Gradient Decomposition Any critical point of H yields a critical point of ϕ, that is, if $p = p^{(0)} + p^{(1)}$ is a critical point of H then $x = B^T p^{(0)}$ is a critical point of ϕ. Consider any critical point p, then we can uniquely decompose the gradient as:

$$\nabla_p H(p) = g_0 + g_1 \text{ in which } g_0 \in \mathsf{N}(G) \text{ and } g_1 \in \mathsf{N}(G)^{\perp}.$$

[6] One can form the Lagrangian for non-linear constraints, but to derive it we need to use fancier math like the implicit function theorem. We only need linear constraints for our applications.

[7] For contradiction, suppose p, q are distinct feasible solutions then, $p \neq q$ but $B^T p = B^T q$ but we can write $p = p^{(0)} + p^{(1)}$ and $q = q^{(0)} + p^{(1)}$ from the above. However, $B^T p = B^T q$ implies that $B^T p^{(0)} = B^T q^{(0)}$. In turn since B is a bijection on $\mathsf{N}(G)$ this implies that $p^{(0)} = q^{(0)}$.

We claim $g_0 = B\nabla\phi(B^T p)$ or equivalently $B^T g = \nabla_x \phi(B^T p)$. From direct calculation, $\nabla_x \phi(x) = \nabla_x H(Bx + p^{(1)}) = B^T \nabla_p H(p^{(0)} + p^{(1)}) = B^T \nabla_p H(p) = B^T g_0$, where the last equality is due to $g_1 \in \mathsf{N}(G)^\perp$. A critical point of H satisfying the constraints must not change along any direction that satisfies the constraints, which is to say that we must have $g_0 = 0$. Very roughly, one can have the intuition that if p were a maximum (or minimum), then if g_0 were non-zero there would be a way to strictly increase (or decrease) the function in a neighbor around p–contradicting p being a maximum (minimum).

Lagrangian Since $g_1 \in \mathsf{N}(G)^\perp = \mathsf{R}(G^T)$ (see the fundamental theorem of linear algebra), we can find a $\eta(p)$ such that $g_1 = -G^T \eta(p)$, which motivates the following functional form:

$$\mathcal{L}(p, \eta(p)) = H(p) + \langle \eta(p), Gp - c \rangle$$

By the definition of $\eta(p)$, we have:

$$\nabla_p \mathcal{L}(p, \eta(p)) = g_0 + g_1 + G^T \eta(p) = g_0.$$

That is, for any critical point p of the original function (which corresponds to $g_0 = 0$) we can select $\eta(p)$ so that it is a critical point of $\mathcal{L}(p, \eta)$. Informally, the multipliers combines the rows of G to cancel g_1, the component of the gradient in the direction of the constraints. This establishes that any critical point of the original constrained function is also a critical point of the Lagrangian.

References

[1] Jaynes, Edwin T, *Probability theory: The logic of science*, Cambridge University Press, 2003.