PROBABILITY DISTRIBUTIONS AND MAXIMUM ENTROPY

KEITH CONRAD

1. Introduction

If we want to assign probabilities to an event, and see no reason for one outcome to occur more often than any other, then the events are assigned equal probabilities. This is called the principle of insufficient reason, or principle of indifference, and goes back to Laplace. If we happen to know (or learn) something about the non-uniformity of the outcomes, how should the assignment of probabilities be changed? There is an extension of the principle of insufficient reason that suggests what to do. It is called the principle of maximum entropy. After defining entropy and computing it in some examples, we will describe this principle and see how it provides a natural conceptual role for many standard probability distributions (normal, exponential, Laplace, Bernoulli). In particular, the normal distribution will be seen to have a distinguishing property among all continuous probability distributions on $\mathbf{R}$ that may be simpler for students in an undergraduate probability course to appreciate than the special role of the normal distribution in the central limit theorem.

This paper is organized as follows. In Sections 2 and 3, we describe the principle of maximum entropy in three basic examples. The explanation of these examples is given in Section 4 as a consequence of a general result (Theorem 4.3). Section 5 provides further illustrations of the maximum entropy principle, with details largely left to the reader as practice. In Section 6 we state Shannon's theorem, which characterizes the entropy function on finite sample spaces. Finally, in Section 7, we prove a theorem about positivity of maximum entropy distributions in the presence of suitable constraints, and derive a uniqueness theorem. In that final section entropy is considered on abstract measure spaces, while the earlier part is written in the language of discrete and continuous probability distributions in order to be more accessible to a motivated undergraduate studying probability.

2. Entropy: definitions and calculations

For a discrete probability distribution p on the countable set $\{x_1, x_2, \dots\}$, with $p_i = p(x_i)$, the *entropy* of p is defined as

$$h(p) = -\sum_{i \geq 1} p_i \log p_i.$$

For a continuous probability density function $p(x)$ on an interval I, its *entropy* is defined as

$$h(p) = -\int_I p(x) \log p(x)\, \mathrm{d}x.$$

We set $0 \log 0 = 0$, which is natural if you look at the graph of $x \log x$. See Figure 1.

This definition of entropy, introduced by Shannon [13], resembles a formula for a thermodynamic notion of entropy. Physically, systems are expected to evolve into states with higher entropy as they approach equilibrium. In our probabilistic context, $h(p)$ is viewed as a measure of the information carried by p, with higher entropy corresponding to less information (more uncertainty, or more of a lack of information).

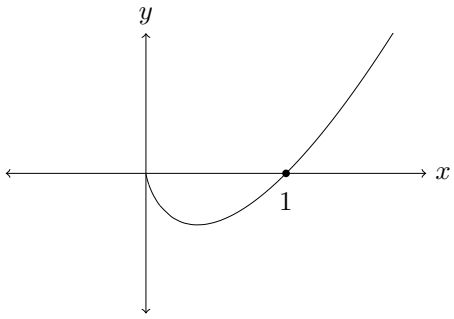

FIGURE 1. Graph of $y = x \log x$.

Example 2.1. Consider a finite set $\{x_1, x_2, \ldots, x_n\}$. If $p(x_1) = 1$ while $p(x_j) = 0$ for $j > 1$, then $h(p) = -1 \log 1 = 0$. In this case, statistics governed by p almost surely produce only one possible outcome, x_1. We have complete knowledge of what will happen. On the other hand, if p is the uniform density function, where $p(x_j) = 1/n$ for all j, then $h(p) = \log n$. We will see later (Theorem 3.1) that every probability density function on $\{x_1, \ldots, x_n\}$ has entropy $\leq \log n$, and the entropy $\log n$ occurs only for the uniform distribution. Heuristically, the probability density function on $\{x_1, x_2, \ldots, x_n\}$ with maximum entropy turns out to be the one that corresponds to the least amount of knowledge of $\{x_1, x_2, \ldots, x_n\}$.

Example 2.2. The entropy of the Gaussian density on $\mathbf{R}$ with mean μ and variance σ^2 is

$$-\int_{\mathbf{R}} \frac{1}{\sqrt{2\pi}\sigma} e^{-(1/2)((x-\mu)/\sigma)^2} \left(-\log(\sqrt{2\pi}\sigma) - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right) dx = \frac{1}{2}(1 + \log(2\pi\sigma^2)).$$

The mean μ does not enter the final formula, so all Gaussians with a common σ (Figure 2) have the same entropy.

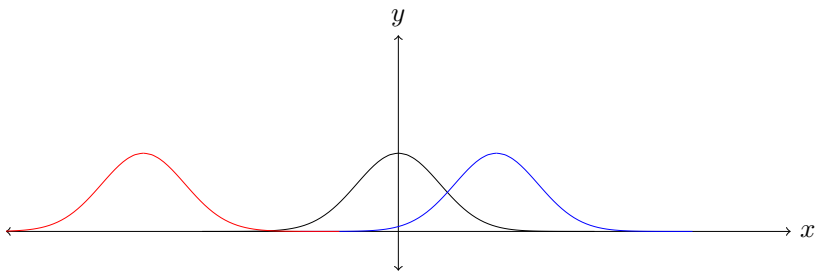

FIGURE 2. Gaussians with the same σ: same entropy.

For σ near 0, the entropy of a Gaussian is negative. Graphically, when σ is small, a substantial piece of the probability density function has values greater than 1, and there $-p \log p < 0$. For discrete distributions, on the other hand, entropy is always ≥ 0, since values of a discrete probability density function never exceed 1. That entropy can be negative in the continuous case simply reflects the fact that probability distributions in the continuous case can be more concentrated than a uniform distribution on $[0, 1]$.

Example 2.3. The entropy of the exponential density on $(0, \infty)$ with mean λ is

$$-\int_0^\infty \frac{1}{\lambda} e^{-x/\lambda} \left(-\log \lambda - \frac{x}{\lambda}\right) \mathrm{d}x = 1 + \log \lambda.$$

As in the previous example, this entropy becomes negative for small λ.

We now state the principle of maximum entropy: if we are seeking a probability density function subject to certain constraints (*e.g.*, a given mean or variance), *use the density satisfying those constraints that has entropy as large as possible.* Laplace's principle of indifference turns out to correspond to the case of a finite sample space with no constraints (Example 2.1).

The principle of maximum entropy, as a method of statistical inference, is due to Jaynes [6, 7, 8]. His idea is that this principle leads to the selection of a probability density function that is consistent with our knowledge and introduces no unwarranted information. Any probability density function satisfying the constraints that has smaller entropy will contain more information (less uncertainty), and thus says something stronger than what we are assuming. The probability density function with maximum entropy, satisfying whatever constraints we impose, is the one that should be least surprising in terms of the predictions it makes.

It is important to clear up an easy misconception: the principle of maximum entropy does not give us something for nothing. For example, a coin is not fair just because we don't know anything about it. In fact, to the contrary, the principle of maximum entropy guides us to the best probability distribution that reflects our current knowledge *and* it tells us what to do if experimental data does not agree with predictions coming from our chosen distribution: understand why the phenomenon being studied behaves in an unexpected way (find a previously unseen constraint) and maximize entropy over the distributions that satisfy all constraints we are now aware of, including the new one.

A proper appreciation of the principle of maximum entropy goes hand in hand with a certain attitude about the interpretation of probability distributions. A probability distribution can be viewed as: (1) a predictor of frequencies of outcomes over repeated trials, or (2) a numerical measure of plausibility that some *individual* situation develops in certain ways. Sometimes the first (frequency) viewpoint is meaningless, and only the second (subjective) interpretation of probability makes sense. For instance, we can ask about the probability that civilization will be wiped out by an asteroid in the next 10,000 years, or the probability that the Red Sox will win the World Series again.

Example 2.4. For an example that mixes the two interpretations of probability (frequency in repeated trials versus plausibility of a single event), consider a six-sided die rolled 1000 times. The average number of dots that are rolled is 4.7. (A fair die would be expected to have average $7/2 = 3.5$.) What is the probability distribution for the faces on this die? This is clearly an underdetermined problem. (There are infinitely many 6-tuples $(p_1, \ldots, p_6)$ with $p_i \geq 0$, $\sum_i p_i = 1$, and $\sum_i i p_i = 4.7$.) We will return to this question in Section 5.

3. Three examples of maximum entropy

We illustrate the principle of maximum entropy in the following three theorems. Proofs of these theorems are in the next section.

Theorem 3.1. *For a probability density function p on a finite set $\{x_1, \ldots, x_n\}$,*

$$h(p) \leq \log n,$$

with equality if and only if p is uniform, i.e., $p(x_i) = 1/n$ for all i.

Concretely, if $p_1, \ldots, p_n$ are nonnegative numbers with $\sum p_i = 1$ then Theorem 3.1 says $-\sum p_i \log p_i \leq \log n$, with equality if and only if every p_i is $1/n$. We are reminded now of the arithmetic-geometric mean equality, but that is really a different inequality: for positive $p_1, \ldots, p_n$ that sum to 1, the arithmetic-geometric mean inequality says $\sum \log p_i \leq -n \log n$ with equality if and only if every p_i is $1/n$. We will see in Section 7 that the arithmetic-geometric mean inequality is a special case of an inequality for entropies of multivariate Gaussians.

Theorem 3.2. *For a continuous probability density function p on $\mathbf{R}$ with variance σ^2,*

$$h(p) \leq \frac{1}{2}(1 + \log(2\pi\sigma^2)),$$

with equality if and only if p is Gaussian with variance σ^2, i.e., for some μ we have $p(x) = (1/\sqrt{2\pi}\sigma)e^{-(1/2)((x-\mu)/\sigma)^2}$.

This describes a conceptual role for Gaussians that is simpler than the Central Limit Theorem.

Theorem 3.3. *For any continuous probability density function p on $(0, \infty)$ with mean λ,*

$$h(p) \leq 1 + \log \lambda,$$

with equality if and only if p is exponential with mean λ, i.e., $p(x) = (1/\lambda)e^{-x/\lambda}$.

Theorem 3.3 suggests that for an experiment with positive outcomes whose mean value is known, the most conservative probabilistic model consistent with that mean value is an exponential distribution.

Example 3.4. To illustrate the inequality in Theorem 3.3, consider another kind of probability density function on $(0, \infty)$, say $p(x) = ae^{-bx^2}$ for $x > 0$. (Here the exponent involves $-x^2$ rather than $-x$.) For p to have total integral 1 over $(0, \infty)$ requires $b = (\pi/4)a^2$, and the mean is $a/2b$. Set $\lambda = a/2b$, so p has the same mean as the exponential distribution $(1/\lambda)e^{-x/\lambda}$. The conditions $b = (\pi/4)a^2$ and $\lambda = a/2b$ let us solve for a and b in terms of λ: $a = 2/(\pi\lambda)$ and $b = 1/(\pi\lambda^2)$. In Figure 3 we plot ae^{-bx^2} and $(1/\lambda)e^{-x/\lambda}$ with the same mean. The inequality $h(p) < 1 + \log\lambda$ from Theorem 3.3 is equivalent (after some algebra) to $1/2 + \log 2 - \log \pi$ being positive. This number is approximately .0484.

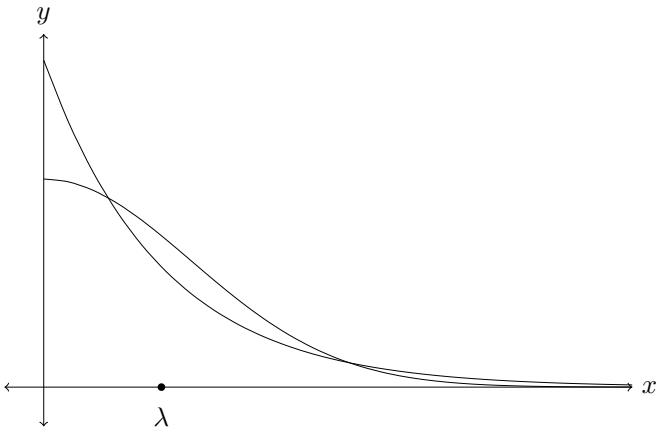

FIGURE 3. Exponential and other distribution with same mean.

While we will be concerned with the principle of maximum entropy insofar as it explains a natural role for various classical probability distributions, the principle is also widely used for practical purposes in the applied sciences [2, 9, 10].

4. Explanation of the three examples

To prove Theorems 3.1, 3.2, and 3.3, we want to maximize $h(p)$ over all probability density functions p satisfying certain constraints. This can be done using Lagrange multipliers.[1] We will instead prove the theorems by taking advantage of special properties of the functional expression $-p \log p$. The application of Lagrange multipliers to such problems is discussed in Appendix A. It is worth noting that Lagrange multipliers would would not, on its own, lead to a logically complete solution of the problem, since the maximum entropy might *a priori* occur at a probability distribution not accessible to that method (analogous to $x = 0$ as a minimum of $|x|$, which is inaccessible to methods of calculus).

Lemma 4.1. *For $x > 0$ and $y \geq 0$,*

$$y - y \log y \leq x - y \log x,$$

with equality if and only if $x = y$.

Note the right side is *not* $x - x \log x$.

Proof. The result is clear if $y = 0$, by our convention that $0 \log 0 = 0$. Thus, let $y > 0$. The inequality is equivalent to $\log(x/y) \leq x/y - 1$, and it is easy to check, for $t > 0$, that $\log t \leq t - 1$ with equality if and only if $t = 1$ (see Figure 4). □

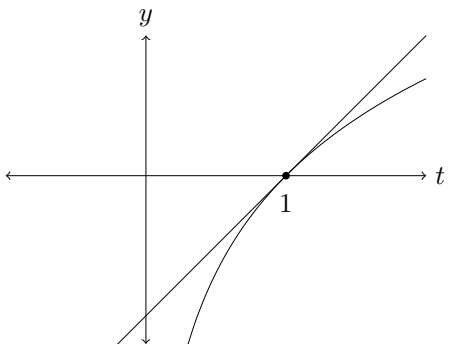

Figure 4. Graph of $y = \log t$ lies below $y = t - 1$ except at $t = 1$.

Lemma 4.2. *Let $p(x)$ and $q(x)$ be continuous probability density functions on an interval I in the real numbers, with $p \geq 0$ and $q > 0$ on I. We have*

$$-\int_I p \log p \, \mathrm{d}x \leq -\int_I p \log q \, \mathrm{d}x$$

if both integrals exist. Moreover, there is equality if and only if $p(x) = q(x)$ for all x.

[1] This is a more interesting application of Lagrange multipliers with *multiple constraints* than the type usually met in multivariable calculus books.

For discrete probability density functions p and q on a set $\{x_1, x_2, \dots\}$, with $p(x_i) \geq 0$ and $q(x_i) > 0$ for all i,

$$-\sum_{i\geq 1} p(x_i)\log p(x_i) \leq -\sum_{i\geq 1} p(x_i)\log q(x_i)$$

if both sums converge. Moreover, there is equality if and only if $p(x_i) = q(x_i)$ for all i.

Proof. We carry out the proof in the continuous case. The discrete case is identical, using sums in place of integrals. By Lemma 4.1, for any $x \in I$

$$p(x) - p(x)\log p(x) \leq q(x) - p(x)\log q(x), \tag{4.1}$$

so integrating both sides over the interval I gives

$$\int_I p\,\mathrm{d}x - \int_I p\log p\,\mathrm{d}x \leq \int_I q\,\mathrm{d}x - \int_I p\log q\,\mathrm{d}x \Longrightarrow -\int_I p\log p\,\mathrm{d}x \leq -\int_I p\log q\,\mathrm{d}x$$

because $\int_I p\,\mathrm{d}x$ and $\int_I q\,\mathrm{d}x$ are both 1. If this last inequality of integrals is an equality, then the continuous function

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

has integral 0 over I, so this nonnegative function is 0. Thus (4.1) is an equality for all $x \in I$, so $p(x) = q(x)$ for all $x \in I$ by Lemma 4.1. □

Theorem 4.3. *Let p and q be continuous probability density functions on an interval I, with finite entropy. Assume $q(x) > 0$ for all $x \in I$. If*

$$-\int_I p\log q\,\mathrm{d}x = h(q), \tag{4.2}$$

then $h(p) \leq h(q)$, with equality if and only if $p = q$.

For discrete probability density functions p and q on $\{x_1, x_2, \dots\}$, with finite entropy, assume $q(x_i) > 0$ for all i. If

$$-\sum_{i\geq 1} p(x_i)\log q(x_i) = h(q), \tag{4.3}$$

then $h(p) \leq h(q)$, with equality if and only if $p(x_i) = q(x_i)$ for all i.

Proof. By Lemma 4.2, $h(p)$ is bounded above by $-\int_I p\log q\,\mathrm{d}x$ in the continuous case and by $-\sum_i p(x_i)\log q(x_i)$ in the discrete case. This bound is being assumed to equal $h(q)$. If $h(p) = h(q)$, then $h(p)$ equals the bound, in which case Lemma 4.2 tells us p equals q. □

Remark 4.4. The conclusions of Theorem 4.3 are still true, by the same proof, if we use $\leq$ in place of the equalities in (4.2) and (4.3). However, we will see in all our applications that equality for (4.2) and (4.3) is what occurs when we want to use Theorem 4.3.

The integral $-\int_I p\log q\,\mathrm{d}x$ and sum $-\sum_{i\geq 1} p(x_i)\log q(x_i)$ that occur in Theorem 4.3 are not entropies, but merely play a technical role to get the desired maximum entropy conclusion in the theorems.

Although Theorem 4.3 allows p to vanish somewhere, it avoids the possibility that $q = 0$ anywhere. Intuitively, this is because the expression $p\log q$ becomes unwieldy around points where q vanishes and p does not. The relation between maximum entropy and zero sets of probability density functions will be explored in Section 7.

The proofs of Theorems 3.1, 3.2, and 3.3 are quite simple. We will show in each case that the constraints (if any) imposed in these theorems imply the constraint (4.2) or (4.3) from Theorem 4.3 for suitable q.

First we prove Theorem 3.1.

Proof. We give two proofs. The first proof will not use Theorem 4.3. The second will.

A probability density function on $\{x_1, \ldots, x_n\}$ is a set of nonnegative real numbers $p_1, \ldots, p_n$ that add up to 1. Entropy is a continuous function of the n-tuples $(p_1, \ldots, p_n)$, and these points lie in a compact subset of $\mathbf{R}^n$, so there is an n-tuple where entropy is maximized. We want to show this occurs at $(1/n, \ldots, 1/n)$ and nowhere else.

Suppose the p_j are not all equal, say $p_1 < p_2$. (Clearly $n \neq 1$.) We will find a new probability density with higher entropy. It then follows, since entropy is maximized at some n-tuple, that entropy is uniquely maximized at the n-tuple with $p_i = 1/n$ for all i.

Since $p_1 < p_2$, for small positive ε we have $p_1 + \varepsilon < p_2 - \varepsilon$. The entropy of $\{p_1 + \varepsilon, p_2 - \varepsilon, p_3, \ldots, p_n\}$ minus the entropy of $\{p_1, p_2, p_3, \ldots, p_n\}$ equals

$$-p_1 \log\left(\frac{p_1+\varepsilon}{p_1}\right) - \varepsilon\log(p_1+\varepsilon) - p_2\log\left(\frac{p_2-\varepsilon}{p_2}\right) + \varepsilon\log(p_2-\varepsilon). \tag{4.4}$$

We want to show this is positive for small enough ε. Rewrite (4.4) as

$$-p_1\log\left(1+\frac{\varepsilon}{p_1}\right) - \varepsilon\left(\log p_1 + \log\left(1+\frac{\varepsilon}{p_1}\right)\right) - p_2\log\left(1-\frac{\varepsilon}{p_2}\right) + \varepsilon\left(\log p_2 + \log\left(1-\frac{\varepsilon}{p_2}\right)\right). \tag{4.5}$$

Since $\log(1+x) = x + O(x^2)$ for small x, (4.5) is

$$-\varepsilon - \varepsilon\log p_1 + \varepsilon + \varepsilon\log p_2 + O(\varepsilon^2) = \varepsilon\log(p_2/p_1) + O(\varepsilon^2),$$

which is positive when ε is small enough since $p_1 < p_2$.

For the second proof, let p be any probability density function on $\{x_1, \ldots, x_n\}$, with $p_i = p(x_i)$. Letting $q_i = 1/n$ for all i,

$$-\sum_{i=1}^n p_i \log q_i = \sum_{i=1}^n p_i \log n = \log n,$$

which is the entropy of q. Therefore Lemma 4.2 says $h(p) \leq h(q)$, with equality if and only if p is uniform. □

Next, we prove Theorem 3.2.

Proof. Let p be a probability density function on $\mathbf{R}$ with variance σ^2. Let μ be its mean. (The mean exists by definition of variance.) Letting q be the Gaussian with mean μ and variance σ^2,

$$-\int_{\mathbf{R}} p(x)\log q(x)\,\mathrm{d}x = \int_{\mathbf{R}} p(x)\left(\frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)\mathrm{d}x.$$

Splitting up the integral into two integrals, the first is $(1/2)\log(2\pi\sigma^2)$ since $\int_{\mathbf{R}} p(x)\,\mathrm{d}x = 1$, and the second is $1/2$ since $\int_{\mathbf{R}}(x-\mu)^2 p(x)\,\mathrm{d}x = \sigma^2$ by definition. Thus the total integral is $(1/2)\log(2\pi\sigma^2) + 1/2$, which is the entropy of q. □

Finally, we prove Theorem 3.3.

Proof. Let p be a probability density function on the interval $(0, \infty)$ with mean λ. Letting $q(x) = (1/\lambda)e^{-x/\lambda}$,

$$-\int_0^\infty p(x)\log q(x)\,\mathrm{d}x = \int_0^\infty p(x)\left(\log\lambda + \frac{x}{\lambda}\right)\mathrm{d}x.$$

Since p has mean λ, this integral is $\log\lambda + 1$, which is the entropy of q. □

In each of Theorems 3.1, 3.2, and 3.3, entropy is maximized over distributions on a fixed domain satisfying certain constraints. Table 1 summarizes these extra constraints, which in each case amounts to fixing the value of some integral.

Distribution	Domain	Fixed Value
Uniform	Finite Set	None
Normal with mean μ	$\mathbf{R}$	$\int_{\mathbf{R}}(x-\mu)^2p(x)\,dx$
Exponential	$(0,\infty)$	$\int_0^\infty xp(x)\,dx$

TABLE 1. Extra constraints

How does one *discover* these extra constraints? They come from asking, for a given distribution q (which we aim to characterize via maximum entropy), what extra information about distributions p on the same domain is needed so that the equality in (4.2) or (4.3) takes place. For instance, in the setting of Theorem 3.3, we want to realize an exponential distribution $q(x) = (1/\lambda)e^{-x/\lambda}$ on $(0,\infty)$ as a maximum entropy distribution. For any distribution $p(x)$ on $(0,\infty)$,

$$\begin{aligned} -\int_0^\infty p(x)\log q(x)\,dx &= \int_0^\infty p(x)\left(\log\lambda + \frac{x}{\lambda}\right)dx \\ &= (\log\lambda)\int_0^\infty p(x)\,dx + \frac{1}{\lambda}\int_0^\infty xp(x)\,dx \\ &= \log\lambda + \frac{1}{\lambda}\int_0^\infty xp(x)\,dx. \end{aligned}$$

To complete this calculation, we need to know the mean of p. This is why, in Theorem 3.3, the exponential distribution is the one on $(0,\infty)$ with maximum entropy having a given mean. The reader should consider Theorems 3.1 and 3.2, as well as later characterizations of distributions in terms of maximum entropy, in this light.

Remark 4.5. For any real-valued random variable X on $\mathbf{R}$ with a corresponding probability density function $p(x)$, the entropy of X is defined to be the entropy of p: $h(X) = -\int_{\mathbf{R}} p(x)\log p(x)\,dx$. If $X_1, X_2, \ldots,$ is a sequence of independent identically distributed real-valued random variables with mean 0 and variance 1, the central limit theorem says that the sums $S_n = (X_1+\cdots+X_n)/\sqrt{n}$, which all have mean 0 and variance 1, converge in a suitable sense to the normal distribution $N(0,1)$ with mean 0 and variance 1. Since $N(0,1)$ is the unique maximum entropy distribution among continuous probability distributions on $\mathbf{R}$ with mean 0 and variance 1, each S_n has entropy less than that of $N(0,1)$. It's natural to ask if, following thermodynamic intuition, the entropies $h(S_n)$ are monotonically increasing (up to the entropy of $N(0,1)$). The answer is yes, which provides an entropic viewpoint on the central limit theorem. See [1].

5. More examples of maximum entropy

Other probability density functions can also be characterized via maximum entropy using Theorem 4.3.

Theorem 5.1. *Fix real numbers $a < b$ and $\mu \in (a,b)$. The continuous probability density function on the interval $[a,b]$ with mean μ that maximizes entropy among all such densities (on $[a,b]$ with mean μ) is a truncated exponential density*

$$q_\alpha(x) = \begin{cases} C_\alpha e^{\alpha x}, & \text{if } x \in [a,b], \\ 0, & \text{otherwise}, \end{cases}$$

where C_α is chosen so that $\int_a^b C_\alpha e^{\alpha x}\,\mathrm{d}x = 1$, and α is the unique real number such that $\int_a^b C_\alpha x e^{\alpha x}\,\mathrm{d}x = \mu$.

This answers a question posted on Math Overflow [11].

Proof. The normalization condition $\int_a^b C_\alpha e^{\alpha x}\,\mathrm{d}x = 1$ tells us $C_\alpha = \alpha/(e^{\alpha b} - e^{\alpha a})$. At $\alpha = 0$ this is $1/(b-a)$, so $q_0(x)$ is the uniform density on $[a,b]$, whose mean is $(a+b)/2$.

To show, for each $\mu \in (a,b)$, that $q_\alpha(x)$ has mean μ for exactly one real number α, we work out a general formula for the mean of $q_\alpha(x)$. When $\alpha \neq 0$, the mean μ_α of $q_\alpha(x)$ is

$$\mu_\alpha = \int_a^b x q_\alpha(x)\,\mathrm{d}x = C_\alpha \left(\frac{xe^{\alpha x}}{\alpha} - \frac{e^{\alpha x}}{\alpha^2}\right)\Bigg|_a^b = \frac{be^{\alpha b} - ae^{\alpha a}}{e^{\alpha b} - e^{\alpha a}} - \frac{1}{\alpha},$$

and we set $\mu_0 = (a+b)/2$, which is the mean of q_0. In Figure 5 is a graph of μ_α as α varies.

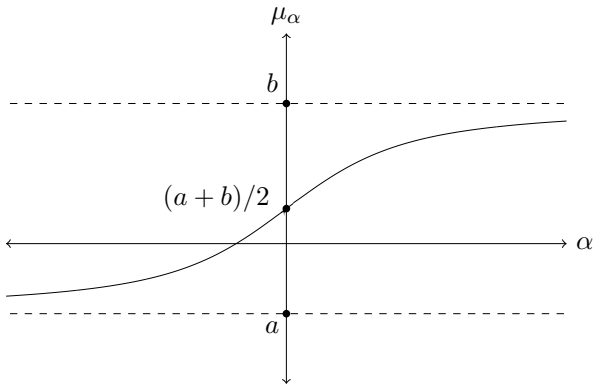

FIGURE 5. A graph of μ_α as a function of α.

Let's check μ_α has the features suggested by Figure 5:

- It is a monotonically increasing function of α,
- $\lim_{\alpha\to-\infty} \mu_\alpha = a$, $\lim_{\alpha\to\infty} \mu_\alpha = b$.

For $\alpha \neq 0$, write

$$\mu_\alpha = \frac{b(e^{\alpha b} - e^{a\alpha}) + (b-a)e^{\alpha a}}{e^{\alpha b} - e^{a\alpha}} - \frac{1}{\alpha} = b + \frac{b-a}{e^{\alpha(b-a)} - 1} - \frac{1}{\alpha}, \tag{5.1}$$

and differentiate this last formula with respect to α:

$$\frac{d\mu_\alpha}{d\alpha} = -\frac{(b-a)^2 e^{\alpha(b-a)}}{(e^{\alpha(b-a)} - 1)^2} + \frac{1}{\alpha^2} \stackrel{?}{>} 0 \Longleftrightarrow (e^{\alpha(b-a)} - 1)^2 \stackrel{?}{>} (\alpha(b-a))^2 e^{\alpha(b-a)}.$$

Setting $t = \alpha(b-a)$, the right inequality is $(e^t - 1)^2 \stackrel{?}{>} t^2 e^t$, or equivalently $e^t - 2 + e^{-t} \stackrel{?}{>} t^2$, and that inequality holds for all $t \neq 0$ since the power series of the left side is a series in even powers of t whose lowest-order term is t^2. This proves μ_α is monotonically increasing in α for both $\alpha < 0$ and $\alpha > 0$. To put these together, we want to check $\mu_\alpha < (a+b)/2$ for $\alpha < 0$ and $\mu_\alpha > (a+b)/2$ for $\alpha > 0$: using (5.1),

$$\mu_\alpha - \frac{a+b}{2} = \frac{b-a}{2} + \frac{b-a}{e^{\alpha(b-a)} - 1} - \frac{b-a}{(b-a)\alpha} = (b-a)\left(\frac{1}{2} + \frac{1}{e^t - 1} - \frac{1}{t}\right)$$

for $t = (b-a)\alpha$. As a power series in t, this is $(b-a)(t/12 - t^3/720 + \cdots)$, so for α near 0, $\mu_\alpha - (a+b)/2$ is positive for $\alpha > 0$ and negative for $\alpha < 0$.

To compute $\lim_{\alpha\to-\infty}\mu_\alpha$ and $\lim_{\alpha\to\infty}\mu_\alpha$, the second formula in (5.1) tells us that $\lim_{\alpha\to\infty}\mu_\alpha = b$ and $\lim_{\alpha\to-\infty}\mu_\alpha = b-(b-a) = a$.

Thus for each $\mu \in (a,b)$, there is a unique $\alpha \in \mathbf{R}$ such that $q_\alpha(x)$ has mean μ.

With α chosen so that $q_\alpha(x)$ has mean μ, we want to show the probability density function $q_\alpha(x)$ has maximum entropy among all continuous probability density functions $p(x)$ on $[a,b]$ with mean μ. Since

$$-\int_a^b p(x)\log q_\alpha(x)\,dx = -\int_a^b p(x)(\log C_\alpha + \alpha x)\,dx = -\log C_\alpha - \alpha\mu = h(q_\alpha),$$

Theorem 4.3 tells us that $h(p) \le h(q_\alpha)$, with equality if and only if $p = q_\alpha$ on $[a,b]$. □

Taking $[a,b] = [0,1]$, the plot of $q_\alpha(x)$ for several values of α is in Figure 6, with $\alpha \le 0$ on the left and $\alpha \ge 0$ on the right. At $\alpha = 0$ we have $\mu = 1/2$ (uniform). As $\mu \to 0$ (that is, $\alpha \to -\infty$) we have $q_\alpha(x) \to 0$ for $0 < x \le 1$, and as $\mu \to 1$ (that is, $\alpha \to \infty$) we have $q_\alpha(x) \to 0$ for $0 \le x < 1$. In the limit, $q_{-\infty}(x)$ is a Dirac mass at 0 while $q_\infty(x)$ is a Dirac mass at 1. These limiting distributions are *not* continuous.

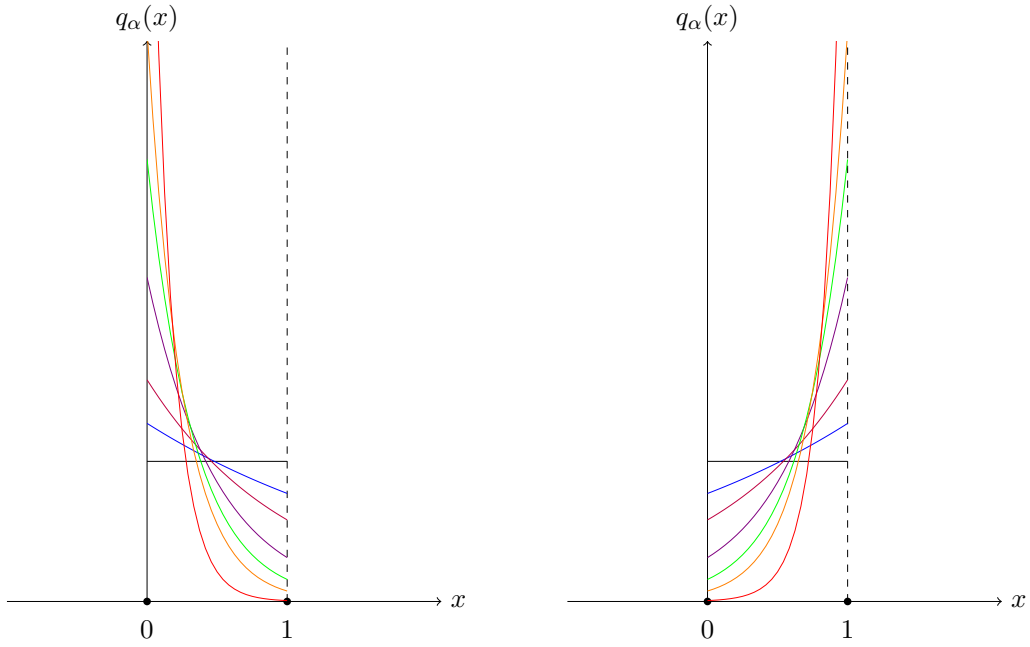

FIGURE 6. Plot of $q_\alpha(x)$ on $[0,1]$ for $\alpha = 0$, $\pm.5$, ±1, ±2, ±3, ±4, and ±7.

Theorem 5.2. *Fix $\lambda > 0$ and $\mu \in \mathbf{R}$. For any continuous probability density function p on $\mathbf{R}$ with mean μ such that*

$$\lambda = \int_{\mathbf{R}} |x-\mu|p(x)\,dx$$

we have

$$h(p) \le 1 + \log(2\lambda),$$

with equality if and only if p is the Laplace distribution with mean μ and variance $2\lambda^2$, i.e., $p(x) = (1/2\lambda)e^{-|x-\mu|/\lambda}$ almost everywhere.

Proof. Left to the reader. □

Note the constraint on p in Theorem 5.2 involves $\int_{\mathbf{R}} |x-\mu|p(x)\,dx$, not $\int_{\mathbf{R}}(x-\mu)p(x)\,dx$.

Theorem 5.3. *Fix $\lambda > 0$ and $\mu \in \mathbf{R}$. For any continuous probability density function p on $\mathbf{R}$ with mean μ such that*

$$\int_{\mathbf{R}} p(x)\log(e^{(x-\mu)/(2\lambda)} + e^{-(x-\mu)/(2\lambda)})\,dx = 1$$

we have

$$h(p) \leq 2 + \log\lambda,$$

with equality if and only if p is the logistic distribution $p(x) = \dfrac{1}{\lambda(e^{(x-\mu)/(2\lambda)} + e^{-(x-\mu)/(2\lambda)})^2}$.

Proof. Left to the reader. Use the change of variables $y = e^{(x-\mu)/(2\lambda)}$ and the integral formulas $\int_0^\infty ((\log y)/(1+y)^2)\,dy = 0$ and $\int_1^\infty ((\log y)/y^2)\,dy = 1$. □

Remark 5.4. As $|x| \to \infty$, the logistic distribution is in practice the simplest smooth probability distribution whose tail decays like $e^{-|x-\mu|/\lambda}$, which resembles the non-smooth Laplace distribution.

Now we turn to n-dimensional distributions, generalizing Theorem 3.2. Entropy is defined in terms of integrals over $\mathbf{R}^n$, and there is an obvious n-dimensional analogue of Lemma 4.2 and Theorem 4.3, which the reader can check.

Theorem 5.5. *For a continuous probability density function p on $\mathbf{R}^n$ with fixed covariances σ_{ij},*

$$h(p) \leq \frac{1}{2}(n + \log((2\pi)^n \det\Sigma)),$$

where $\Sigma = (\sigma_{ij})$ is the covariance matrix for p. There is equality if and only if p is an n-dimensional Gaussian density with covariances σ_{ij}.

We recall the definition of the covariances σ_{ij}. For an n-dimensional probability density function p, its means are $\mu_i = \int_{\mathbf{R}^n} x_i p(\mathbf{x})\,d\mathbf{x}$ and its covariances are

$$\sigma_{ij} = \int_{\mathbf{R}^n} (x_i - \mu_i)(x_j - \mu_j)p(\mathbf{x})\,d\mathbf{x}. \tag{5.2}$$

In particular, $\sigma_{ii} > 0$. When $n = 1$, $\sigma_{11} = \sigma^2$ in the usual notation. The symmetric matrix $\Sigma = (\sigma_{ij})$ is positive-definite, since the matrix $(\langle v_i, v_j\rangle)$ is positive-definite for any finite set of linearly independent v_i in a real inner product space $(V, \langle\cdot,\cdot\rangle)$.

Proof. The Gaussian densities on $\mathbf{R}^n$ are those probability density functions of the form

$$G(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det\Sigma}} e^{-(1/2)(\mathbf{x}-\boldsymbol{\mu})\cdot\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})},$$

where $\Sigma := (\sigma_{ij})$ is a positive-definite symmetric matrix and $\boldsymbol{\mu} \in \mathbf{R}^n$. Calculations show the i-th mean of G, $\int_{\mathbf{R}^n} x_i G(\mathbf{x})\,d\mathbf{x}$, is the i-th coordinate of $\boldsymbol{\mu}$, and the covariances of G are the entries σ_{ij} in Σ. The entropy of G is

$$\frac{1}{2}(n + \log((2\pi)^n \det\Sigma)),$$

by a calculation left to the reader. (Hint: it helps in the calculation to write Σ as the square of a symmetric matrix.)

Now assume p is any n-dimensional probability density function with means and covariances. Define μ_i to be its i-th mean of p and define (σ_{ij}) to be the covariance matrix of p.

Let G be the n-dimensional Gaussian with the means and covariances of p. The theorem follows from Theorem 4.3 and the equation

$$-\int_{\mathbf{R}^n} p(\mathbf{x}) \log G(\mathbf{x})\, d\mathbf{x} = \frac{1}{2}(n + \log((2\pi)^n \det \Sigma)),$$

whose verification boils down to checking that

$$\int_{\mathbf{R}^n} (\mathbf{x} - \boldsymbol{\mu}) \cdot \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) p(\mathbf{x})\, d\mathbf{x} = n, \tag{5.3}$$

which is left to the reader. (Hint: Diagonalize the quadratic form corresponding to Σ.) □

The next example, which can be omitted by the reader unfamiliar with the context, involves entropy for probability density functions on local fields.

Theorem 5.6. *Let K be a nonarchimedean local field, with dx the Haar measure, normalized so $\mathcal{O}_K$ has measure 1. Fix a positive number r. As p varies over all probability density functions on (K, dx) such that*

$$b_r = \int_K |x|^r p(x)\, dx$$

is fixed, the maximum value of $h(p)$ occurs precisely at those p such that $p(x) = ce^{-t|x|^r}$ almost everywhere, where c and t are determined by the conditions

$$\int_K ce^{-t|x|^r}\, dx = 1, \qquad \int_K |x|^r (ce^{-t|x|^r})\, dx = b_r.$$

Proof. Left to the reader. □

Theorem 5.6 applies to the real and complex fields, using Lebesgue measure. For example, the cases $r = 1$ and $r = 2$ of Theorem 5.6 with $K = \mathbf{R}$ were already met, in Theorems 5.2 and 3.2 respectively.

Our remaining examples are discrete distributions. On $\mathbf{N} = \{0, 1, 2, \dots\}$, the geometric distributions are given by $p(n) = (1 - r)r^n$, where $0 \le r < 1$, and the mean is $r/(1 - r)$. More generally, for a fixed integer k, the geometric distributions on $\{k, k+1, k+2, \dots\}$ are given by $p(n) = (1 - r)r^{n-k}$ $(0 \le r < 1)$, with mean $k + r/(1 - r)$. In particular, the mean and r in a geometric distribution on $\{k, k + 1, k + 2, \dots\}$ determine each other (k is fixed).

Any probability distribution p on $\{k, k + 1, \dots\}$ has mean at least k (if it has a mean), so there is a geometric distribution with the same mean as p.

Theorem 5.7. *On $\{k, k + 1, k + 2, \dots\}$, the unique probability distribution with a given mean and maximum entropy is the geometric distribution with that mean.*

Proof. When the mean is greater than k, this is an application of Theorem 4.3, and is left to the reader. When the mean is k, the geometric distribution with $r = 0$ is the only such distribution at all. □

Example 5.8. During a storm, 100 windows in a factory break into a total of 1000 pieces. What is (the best estimate for) the probability that a specific window broke into 4 pieces? 20 pieces?

Letting m be the number of pieces into which a window breaks, our model will let m run over $\mathbf{Z}^+$. (There is an upper bound on m due to atoms, but taking $m \in \mathbf{Z}^+$ seems most reasonable.) The statement of the problem suggests the mean number of pieces a window broke into is 10. Using Theorem 5.7 with $k = 1$, the most reasonable distribution on m is geometric with $k + r/(1 - r) = 10$, *i.e.*, $r = 9/10$. Then $p(n) = (1/10)(9/10)^{n-1}$. In particular, $p(4) \approx .0729$. That is, the (most reasonable) probability a window broke into 4 pieces is around 7.3%. The probability that a window broke into 20 pieces is $p(20) \approx .0135$, or a little over 1.3%. The probability a window did not break is $p(1) = 1/10$, or 10%.

Theorem 5.9. *For a fixed $y > 0$, the unique probability distribution on $\{0, y, 2y, 3y, \dots\}$ with mean λ and maximum entropy is $p(ny) = y\lambda^n/(\lambda+y)^{n+1}$, which has the form $(1-r)r^n$ with $r = 1/(1 + y/\lambda)$.*

Proof. Left to the reader. □

Now we consider distributions corresponding to a finite set of (say) n independent trials, each trial having just two outcomes (such as coin-flipping). The standard example of such a distribution is a Bernoulli distribution, where each trial has the same probability of success. The Bernoulli distribution for n trials, with mean $\lambda \in [0, n]$, corresponds to a probability of success λ/n in each trial.

Theorem 5.10. *Let p be a probability distribution on $\{0, 1\}^n$ corresponding to some sequence of n independent trials having two outcomes. Let $p_i \in [0, 1]$ be the probability the i-th outcome is 1, and let $\lambda = \sum p_i$ be the mean of p. Then*

$$h(p) \leq -\lambda \log(\lambda/n) - (n - \lambda) \log(1 - \lambda/n),$$

with equality if and only if p is the Bernoulli distribution with mean λ.

Proof. Left to the reader. Note the independence condition means such p are precisely the set of probability measures on $\{0, 1\}^n$ that break up into products: $p(x_1, \dots, x_n) = p_1(x_1)p_2(x_2) \dots p_n(x_n)$, where $\{p_i(0), p_i(1)\}$ is a probability distribution on $\{0, 1\}$. □

Our next and final example of the principle of maximum entropy will refine the principle of indifference on finite sample spaces and lead to a standard probability distribution from thermodynamics. Let $S = \{s_1, \dots, s_n\}$ be a finite sample space and $E\colon S \to \mathbf{R}$, with $E_j = E(s_j)$. Physically, the elements of S are states of a system, E_j is the *energy* a particle has when it's in state s_j, and p_j is the probability of a particle being in state s_j, or equivalently the probability that a particle in the system has energy E_j. For each probability distribution p on S, with $p_j = p(s_j)$, thought of as a probability distribution for the particles to be in the different states, the corresponding expected value of E is $\sum p_j E_j$. This number is between $\min E_j$ and $\max E_j$. Choose a number $\overline{E}$ such that $\min E_j \leq \overline{E} \leq \max E_j$, Our basic question is: what distribution p on S satisfies

$$\sum p_j E_j = \overline{E} \tag{5.4}$$

and has maximum entropy? When all E_j are equal, then the expected value condition (5.4) is vacuous (it follows from $\sum p_j = 1$) and we are simply seeking a distribution on S with maximum entropy and no constraints. Theorem 3.1 tells us the answer when there are no constraints: the uniform distribution. What if the E_j are not all equal? (We allow some of the E_j to coincide, just not all.)

If $n = 2$ and $E_1 \neq E_2$, then elementary algebra gives the answer: if $p_1 + p_2 = 1$ and $\overline{E} = p_1E_1 + p_2E_2$, where E_1, E_2, and $\overline{E}$ are known, then

$$p_1 = \frac{\overline{E} - E_2}{E_1 - E_2}, \quad p_2 = \frac{\overline{E} - E_1}{E_2 - E_1}.$$

If $n > 2$, then the conditions $\sum p_j = 1$ and $\sum p_j E_j = \overline{E}$ are two linear equations in n unknowns (the p_j), so there is not a unique solution. (Recall Example 2.4.)

Theorem 5.11. *If the E_j's are not all equal, then for each $\overline{E}$ between $\min E_j$ and $\max E_j$, there is a unique probability distribution q on $\{s_1, \dots, s_n\}$ satisfying the condition $\sum q_j E_j = \overline{E}$ and having maximum entropy. It is given by the formula*

$$q_j = \frac{e^{-\beta E_j}}{\sum_{i=1}^n e^{-\beta E_i}} \tag{5.5}$$

for a unique extended real number β in $[-\infty,\infty]$ that depends on $\overline{E}$. In particular, $\beta = -\infty$ corresponds to $\overline{E} = \max E_j$, $\beta = \infty$ corresponds to $\overline{E} = \min E_j$, and $\beta = 0$ (the uniform distribution) corresponds to the arithmetic mean $\overline{E} = (\sum E_j)/n$, so $\beta > 0$ when $\overline{E} < (\sum E_j)/n$ and $\beta < 0$ when $\overline{E} > (\sum E_j)/n$.

When all E_j are equal, $q_j = 1/n$ for every β, so Theorem 5.11 could incorporate Theorem 3.1 as a special case. However, we will be using Theorem 3.1 in the proof of certain degenerate cases of Theorem 5.11.

Proof. For each $\beta \in \mathbf{R}$ (we'll deal with $\beta = \pm\infty$ later), let $q_j(\beta)$ be given by the formula (5.5). Note $q_j(\beta) > 0$. We view the numbers $q_j(\beta)$ (for $j = 1, \ldots, n$) as a probability distribution $q(\beta)$ on $\{s_1, \ldots, s_n\}$, with $q_j(\beta)$ being the probability that particles from the system are in state s_j. The expected value of the energy function E for this probability distribution is

$$f(\beta) = \sum_{j=1}^{n} q_j(\beta)E_j = \frac{\sum_{j=1}^{n} E_j e^{-\beta E_j}}{\sum_{i=1}^{n} e^{-\beta E_i}}. \tag{5.6}$$

Writing $Z(\beta) = \sum_{i=1}^{n} e^{-\beta E_i}$ for the denominator of the $q_j(\beta)$'s, $f(\beta) = -Z'(\beta)/Z(\beta)$.

For each $\beta \in \mathbf{R}$, $f(\beta)$ is the expected value of E relative to the distribution $q(\beta)$. As β varies over $\mathbf{R}$, $f(\beta)$ has values strictly between $\min E_j$ and $\max E_j$. (Since we are assuming the E_j are not all equal, $n > 1$.) A calculation shows

$$f'(\beta) = \frac{\sum_{i,j}(E_iE_j - E_j^2)e^{-\beta(E_i+E_j)}}{\left(\sum_{i=1}^{n} e^{-\beta E_i}\right)^2} = -\frac{\sum_{i<j}(E_i - E_j)^2 e^{-\beta(E_i+E_j)}}{\left(\sum_{i=1}^{n} e^{-\beta E_i}\right)^2},$$

so $f'(\beta) < 0$ for all $\beta \in \mathbf{R}$. Therefore f takes each value in its range only once. Since

$$\lim_{\beta\to-\infty} f(\beta) = \max E_j, \quad \lim_{\beta\to\infty} f(\beta) = \min E_j, \tag{5.7}$$

and

$$\lim_{\beta\to-\infty} q_i(\beta) = \begin{cases} 1/c_{\max}, & \text{if } E_i = \max E_j, \\ 0, & \text{if } E_i < \max E_j, \end{cases} \quad \lim_{\beta\to\infty} q_j(\beta) = \begin{cases} 1/c_{\min}, & \text{if } E_i = \min E_j, \\ 0, & \text{if } E_i > \min E_j, \end{cases} \tag{5.8}$$

where $c_{\max} = \#\{i : E_i = \max E_j\}$ and $c_{\min} = \#\{i : E_i = \min E_j\}$, we can use (5.7) and (5.8) to extend the functions $q_j(\beta)$ and $f(\beta)$ to the cases $\beta = \pm\infty$. Therefore, for each $\overline{E}$ in $[\min E_j, \max E_j]$, there is a unique $\beta \in [-\infty,\infty]$ (depending on $\overline{E}$) such that $\sum q_j(\beta)E_j = \overline{E}$.

Now we consider all probability distributions p_j on $\{s_1, \ldots, s_n\}$ with $\sum p_jE_j = \overline{E}$. We want to show the unique distribution p with maximum entropy is given by $p_j = q_j(\beta)$, where $\beta = \beta_{\overline{E}}$ is selected from $[-\infty,\infty]$ to satisfy $\sum q_j(\beta)E_j = \overline{E}$.

Case 1: $\min E_j < \overline{E} < \max E_j$. Here $\beta \in \mathbf{R}$ and $q_j(\beta) > 0$ for all j. We will show

$$-\sum_{j=1}^{n} p_j \log q_j(\beta) = -\sum_{j=1}^{n} q_j(\beta) \log q_j(\beta), \tag{5.9}$$

and then Theorem 4.3 applies.

Set $Z(\beta) = \sum_{i=1}^n e^{-\beta E_i}$, the denominator of each $q_j(\beta)$, as before. Then (5.9) follows from

$$\begin{aligned} -\sum_{j=1}^n p_j \log q_j(\beta) &= -\sum_{j=1}^n p_j(-\beta E_j - \log Z(\beta)) \\ &= \beta \overline{E} + \log Z(\beta) \\ &= -\sum_{j=1}^n q_j(\beta) \log q_j(\beta). \end{aligned}$$

Case 2: $\overline{E} = \min E_j$. In this case, $\beta = \infty$. Let c be the number of values of i such that $E_i = \min E_j$, so

$$q_i(\infty) = \begin{cases} 1/c, & \text{if } E_i = \min E_j, \\ 0, & \text{otherwise.} \end{cases}$$

We cannot directly use Theorem 4.3 to prove this distribution has maximum entropy, since the probability distribution $q(\infty)$ vanishes at some s_i.

We will show that every distribution p such that $\sum p_j E_j = \min E_j$ vanishes at each s_i where $E_i > \min E_j$. That will imply this class of distributions is supported on the set $\{s_i : E_i = \min E_j\}$, where E_i restricts to a constant function and $q(\infty)$ restricts to the uniform distribution. Therefore $q(\infty)$ will have the maximum entropy by Theorem 3.1.

For ease of notation, suppose the E_j's are indexed so that $E_1 \le E_2 \le \cdots \le E_n$. Then

$$E_1 = \sum_{i=1}^n p_i E_1 \ge \sum_{i=1}^c p_i E_1 + \sum_{i=c+1}^n p_i E_i.$$

Subtracting the first term on the right from both sides,

$$\left(\sum_{i=c+1}^n p_i\right) E_1 \ge \sum_{i=c+1}^n p_i E_i \ge \left(\sum_{i=c+1}^n p_i\right) E_{c+1}.$$

Since $E_1 < E_{c+1}$, we must have $p_i = 0$ for $i > c$.

Case 3: $\overline{E} = \max E_j$. This is similar to Case 2 and is left to the reader. □

Example 5.12. We return to the weighted die problem from Example 2.4. When rolled 1000 times, a six-sided die comes up with an average of 4.7 dots. We want to estimate, as best we can, the probability distribution of the faces.

Here the space of 6 outcomes is $\{1, 2, \ldots, 6\}$ We do not know the probability distribution of the occurrence of the faces, although we expect it is not uniform. (In a uniform distribution, the average number of dots that occur is 3.5, not 4.7.) What is the best guess for the probability distribution? By the principle of maximum entropy and Theorem 5.11, our best guess is $q(\beta_0)$, where β_0 is chosen so that $\sum_{j=1}^6 j q_j(\beta_0) = 4.7$, *i.e.*,

$$\frac{\sum_{j=1}^6 j e^{-\beta_0 j}}{\sum_{i=1}^6 e^{-\beta_0 i}} = 4.7. \tag{5.10}$$

(That we use a numerical average over 1000 rolls as a theoretical mean value is admittedly a judgment call.) The left side of (5.10) is a monotonically decreasing function of β_0, so it is easy with a computer to find the approximate solution $\beta_0 \approx -.4632823$. The full maximum entropy distribution is, approximately,

$$q_1 \approx .039, \quad q_2 \approx .062, \quad q_3 \approx .098, \quad q_4 \approx .157, \quad q_5 \approx .249, \quad q_6 \approx .395.$$

For example, this suggests there should be about a 25% chance of getting a 5 on each independent roll of the die.

Example 5.13. If after 1000 rolls the six-sided die comes up with an average of 2.8 dots, then the maximum entropy distribution of the faces is determined by the parameter β_0 satisfying

$$\frac{\sum_{j=1}^{6} je^{-\beta_0 j}}{\sum_{i=1}^{6} e^{-\beta_0 i}} = 2.8, \tag{5.11}$$

and by a computer $\beta_0 \approx .2490454$. The maximum entropy distribution for this β_0 is

$$q_1 \approx .284, \quad q_2 \approx .221, \quad q_3 \approx .172, \quad q_4 \approx .134, \quad q_5 \approx .104, \quad q_6 \approx .081.$$

This method of solution, in extreme cases, leads to results that at first seem absurd. For instance, if we flip a coin twice and get heads both times, the principle of maximum entropy would suggest the coin has probability 0 of coming up tails! This setting is far removed from usual statistical practice, where two trials have no significance. Obviously further experimentation would be likely to lead us to revise our estimated probabilities (using maximum entropy in light of new information).

Theorem 5.14. *Let $S = \{s_1, s_2, \dots, \}$ be a countable set and $E\colon S \to \mathbf{R}$ be a nonnegative function that takes on any value a uniformly bounded number of times, Write $E(s_j)$ as E_j and assume $\lim_{i\to\infty} E_i = \infty$. For each number $\overline{E} \geq \min E_j$, there is a unique probability distribution q on S that satisfies the condition $\sum_{j\geq 1} q_j E_j = \overline{E}$ and has maximum entropy. It is given by the formula*

$$q_j = \frac{e^{-\beta E_j}}{\sum_{i\geq 1} e^{-\beta E_i}} \tag{5.12}$$

for a unique number $\beta \in (0, \infty]$. In particular, $\beta = \infty$ corresponds to $\overline{E} = \min E_j$.

Proof. This is similar to the proof of Theorem 5.11. □

That we only encounter $\beta > 0$ in the countably infinite case, while β could be positive or negative or 0 in the finite case, is related to the fact that the denominator of (5.12) doesn't converge for $\beta \leq 0$ when $E_i > 0$ and there are infinitely many E_i's.

Remark 5.15. Theorems 5.7 and 5.9 are special cases of Theorem 5.14: $S = \{k, k+1, k+2, \dots\}$ and $E_n = n$ for $n \geq k$ in the case of Theorem 5.7, and $S = \{0, y, 2y, \dots\}$ and $E_n = ny$ for $n \geq 0$ in the case of Theorem 5.9. Setting $E_n = n$ for $n \geq k$ in (5.12), for some $\beta > 0$ the maximum entropy distribution in Theorem 5.7 will have

$$q_n = \frac{e^{-\beta n}}{\sum_{i\geq k} e^{-\beta i}} = \frac{e^{-\beta n}}{e^{-\beta k}/(1-e^{-\beta})} = (1-e^{-\beta})e^{-\beta(n-k)},$$

which is the geometric distribution $(1-r)r^{n-k}$ for $r = e^{-\beta} \in [0, 1)$.

Setting $E_n = ny$ for $n \geq 0$ in (5.12), for some $\beta > 0$ the maximum entropy distribution in Theorem 5.9 will have

$$q_n = \frac{e^{-\beta ny}}{\sum_{i\geq 0} e^{-\beta iy}} = (1-e^{-\beta y})e^{-\beta ny} = (1-r)r^n$$

for $r = e^{-\beta y}$. The condition that $\sum_{n\geq 0} q_n E_n = \overline{E}$ becomes $ry/(1-r) = \overline{E}$, so $r = \overline{E}/(y+\overline{E}) = 1/(1+y/\overline{E})$.

In thermodynamics, the distribution arising in Theorem 5.11 is the Maxwell–Boltzmann energy distribution of a system of non-interacting particles in thermodynamic equilibrium having a given mean energy $\overline{E}$. The second law of thermodynamics says, roughly, that as a physical system evolves towards equilibrium its entropy is increasing, which provides the physical intuition for the principle of maximum entropy.

Jaynes [6, 8] stressed the derivation of the Maxwell–Boltzmann distribution in Theorem 5.11, which does not rely on equations of motion of particles. For Jaynes, the Maxwell–Boltzmann distribution in thermodynamics does not arise from laws of physics, but rather from "logic": we want an energy distribution with a given average value and without any unwarranted extra properties. Constraining the energy distribution only by a mean value, the distribution with maximum entropy is the unique Maxwell-Boltzmann distribution with the chosen mean value. Jaynes said the reason this distribution fits experimental measurements so well is that macroscopic systems have a tremendous number of particles, so the distribution of a macroscopic random variable (*e.g.*, pressure) is very highly concentrated.

6. A characterization of the entropy function

Shannon's basic paper [13] gives an axiomatic description of the entropy function on finite sample spaces. A version of Shannon's theorem with weaker hypotheses, due to Faddeev [3], is as follows.

Theorem 6.1. *For $n \geq 2$, let $\Delta_n = \{(p_1, \ldots, p_n) : 0 \leq p_i \leq 1, \sum p_i = 1\}$. Suppose on each Δ_n a function $H_n \colon \Delta_n \to \mathbf{R}$ is given with the following properties:*

- *Each H_n is continuous.*
- *Each H_n is a symmetric function of its arguments.*
- *For $n \geq 2$ and all $(p_1, \ldots, p_n) \in H_n$ with $p_n > 0$, and $t \in [0, 1]$,*

$$H_{n+1}(p_1, \ldots, p_{n-1}, tp_n, (1-t)p_n) = H_n(p_1, \ldots, p_{n-1}, p_n) + p_n H_2(t, 1-t).$$

Then, for some constant k, $H_n(p_1, \ldots, p_n) = -k \sum_{i=1}^n p_i \log p_i$ for all n.

Proof. See [3] or [4, Chapter 1]. The proof has three steps. First, verify the function $F(n) = H_n(1/n, 1/n, \ldots, 1/n)$ has the form $-k \log n$. (This uses the infinitude of the primes!) Second, reduce the case of rational probabilities to the equiprobable case using the third property. Finally, handle irrational probabilities by contintuity from the rational probability case. Only in the last step is the continuity hypothesis used.

Unlike Faddeev's proof, Shannon's proof of Theorem 6.1 included the condition that $H_n(1/n, 1/n, \ldots, 1/n)$ is monotonically increasing with n (and the symmetry hypothesis was not included). □

The choice of normalization constant k in the conclusion of Theorem 6.1 can be considered as a choice of base for the logarithms. If we ask for $H_n(1/n, \ldots, 1/n)$ to be an increasing function of n, which seems plausible for a measurement of uncertainty on probability distributions, then $k > 0$. In fact, the positivity of k is forced just by asking that $H_2(1/2, 1/2) > 0$.

In Theorem 6.1, only H_2 has to be assumed continuous, since the other H_n's can be written in terms of H_2 and therefore are forced to be continuous.

The third condition in Theorem 6.1 is a consistency condition. If one event can be viewed as two separate events with their own probabilities (t and $1-t$), then the total entropy is the entropy with the two events combined plus a weighted entropy for the two events separately. It implies, quite generally, that any two ways of viewing a sequence of events as a composite of smaller events will lead to the same calculation of entropy. That is, if $(q_1, \ldots, q_n) \in \Delta_n$ and each q_r is a sum of (say) m_r numbers $p_j^{(r)} \geq 0$, then

$$H(p_1^{(1)}, p_2^{(1)}, \ldots, p_{m_1}^{(1)}, \ldots, p_1^{(n)}, \ldots, p_{m_n}^{(n)}) = H(q_1, \ldots, q_n) + \sum_{r=1}^n q_r H\left(\frac{p_1^{(r)}}{q_r}, \ldots, \frac{p_{m_r}^{(r)}}{q_r}\right). \tag{6.1}$$

(The subscripts on H have been dropped to avoid cluttering the notation.)

Although $\sum p_i^2$ is a conceivable first attempt to measure the uncertainty in a probability distribution on finite sample spaces, Theorem 6.1 implies this formula has to lead to inconsistent calculations of uncertainty, since it is not the unique possible type of formula satisfying the hypotheses of Theorem 6.1.

7. Positivity and uniqueness of the maximum entropy distribution

In each of the three examples from Section 3, the probability density function with maximum entropy is positive everywhere. Positivity on the whole space played a crucial technical role in our discussion, through its appearance as a hypothesis in Theorem 4.3.

This raises a natural question: is it always true that a probability density with maximum entropy is positive? No. Indeed, we have seen examples where the density with maximum entropy vanishes in some places (*e.g.*, Theorem 5.11 when $\overline{E}$ is $\min E_j$ or $\max E_j$), or we can interpret some earlier examples in this context (*e.g.*, consider Theorem 3.3 for probability density functions on $\mathbf{R}$ rather than $(0,\infty)$, but with the extra constraint $\int_{-\infty}^0 p\,\mathrm{d}x = 0$.) In such examples, however, every probability density function satisfying the given constraints also vanishes where the maximum entropy density vanishes.

Let us refine our question: is the zero set of a probability density with maximum entropy *not a surprise*, in the sense that all other probability densities satisfying the constraints also vanish on this set? (Zero sets, unlike other level sets, are distinguished through their probabilistic interpretation as "impossible events.") Jaynes believed so, as he wrote [6, p. 623] "Mathematically, the maximum-entropy distribution has the important property that no possibility is ignored; it assigns positive weight to every situation that is not absolutely excluded by the given information." However, Jaynes did not prove his remark. We will give a proof when the constraints are convex-linear (that is, the probability distributions satisfying the constraints are closed under convex-linear combinations). This fits the setting of most of our examples. (Exceptions are Theorems 3.2, 5.5, and 5.10, where the constraints are not convex-linear.)

At this point, to state the basic result quite broadly, we will work in the general context of entropy on measure spaces. Let (S,ν) be a measure space. A probability density function p on (S,ν) is a ν-measurable function from S to $[0,\infty)$ such that $p\,\mathrm{d}\nu$ is a probability measure on S. Define the entropy of p to be

$$h(p) = -\int_S p\log p\,\mathrm{d}\nu,$$

assuming this converges. That is, we assume $p\log p \in L^1(S,\nu)$.

(Note the entropy of p depends on the choice of ν, so $h_\nu(p)$ would be a better notation than $h(p)$. For any probability measure μ on S that is absolutely continuous with respect to ν, we could define the entropy of μ with respect to ν as $-\int_S \log(\mathrm{d}\mu/\mathrm{d}\nu)\,\mathrm{d}\mu$, where $\mathrm{d}\mu/\mathrm{d}\nu$ is a Radon-Nikodym derivative. This recovers the above formulation by writing μ as $p\,\mathrm{d}\nu$.)

Theorem 7.1. *Let p and q be probability density functions on (S,ν) with finite entropy. Assume $q > 0$ on S. If*

$$-\int_S p\log p\,\mathrm{d}\nu = h(q),$$

then $h(p) \le h(q)$, with equality if and only if $p = q$ almost everywhere on S.

Proof. Mimic the proof of Lemma 4.2 and Theorem 4.3. □

Example 7.2. For $a < b$, partition the interval $[a,b]$ into n parts using $a = a_0 < a_1 < \cdots < a_n = b$. Pick n positive numbers $\lambda_1,\dots,\lambda_n$ such that $\lambda_1 + \cdots + \lambda_n = 1$. Let q be the (discontinuous) probability density function on $[a,b]$ with constant value $\lambda_i/(a_i - a_{i-1})$

on each open interval (a_{i-1}, a_i) and having arbitrary positive values at the a_i's. For any probability density function p on $[a, b]$ such that $\int_{a_{i-1}}^{a_i} p(x)\,\mathrm{d}x = \lambda_i$ for all i,

$$-\int_a^b p \log q \,\mathrm{d}x = -\sum_{i=1}^n \lambda_i \log\left(\frac{\lambda_i}{a_i - a_{i-1}}\right) = h(q).$$

Therefore q has maximum entropy among all probability density functions on $[a, b]$ satisfying $\int_{a_{i-1}}^{a_i} p(x)\,\mathrm{d}x = \lambda_i$ for all i.

Lemma 7.3. *The probability density functions on (S, ν) having finite entropy are closed under convex-linear combinations: if $h(p_1)$ and $h(p_2)$ are finite, so is $h((1-\varepsilon)p_1 + \varepsilon p_2)$ for any $\varepsilon \in [0, 1]$.*

Proof. We may take $0 < \varepsilon < 1$.

Let $f(t) = -t \log t$ for $t \geq 0$, so $h(p) = \int_S f(p(x))\,\mathrm{d}\nu(x)$. From the concavity of the graph of s,

$$f((1-\varepsilon)t_1 + \varepsilon t_2) \geq (1-\varepsilon)f(t_1) + \varepsilon f(t_2)$$

when $t_1, t_2 \geq 0$. Therefore

$$f((1-\varepsilon)p_1(x) + \varepsilon p_2(x)) \geq (1-\varepsilon)f(p_1(x)) + \varepsilon f(p_2(x)) \tag{7.1}$$

For an upper bound on $f((1-\varepsilon)t_1 + \varepsilon t_2)$, we note that

$$f(t_1 + t_2) \leq f(t_1) + f(t_2)$$

for all $t_1, t_2 \geq 0$. This is clear if either t_1 or t_2 is 0. When t_1 and t_2 are both positive,

$$\begin{aligned} f(t_1 + t_2) &= -(t_1 + t_2)\log(t_1 + t_2) \\ &= -t_1 \log(t_1 + t_2) - t_2 \log(t_1 + t_2) \\ &\leq -t_1 \log(t_1) - t_2 \log(t_2) \\ &= f(t_1) + f(t_2). \end{aligned}$$

Therefore

$$(1-\varepsilon)f(p_1(x)) + \varepsilon f(p_2(x)) \leq f((1-\varepsilon)p_1(x) + \varepsilon p_2(x)) \leq f((1-\varepsilon)p_1(x)) + f(\varepsilon p_2(x)). \tag{7.2}$$

Integrate (7.2) over S. Since $\int_S f(\delta p(x))\,\mathrm{d}\nu = -\delta \log \delta + \delta h(p)$ for $\delta \geq 0$ and p a probability distribution on S, we see $(1-\varepsilon)p_1 + \varepsilon p_2$ has finite entropy:

$$(1-\varepsilon)h(p_1) + \varepsilon h(p_2) \leq h((1-\varepsilon)p_1 + \varepsilon p_2) \leq -(1-\varepsilon)\log(1-\varepsilon) + (1-\varepsilon)h(p_1) - \varepsilon \log \varepsilon + \varepsilon h(p_2).$$

□

Let Π be any set of probability density functions p on (S, ν) that is closed under convex-linear combinations. Let Π' be the subset of Π consisting of those $p \in \Pi$ with finite entropy. By Lemma 7.3, Π' is closed under convex-linear combinations.

Example 7.4. Fix a finite set of (real-valued) random variables $X_1, \ldots, X_r$ on (S, ν) and a finite set of constants $c_i \in \mathbf{R}$. Let Π be those probability density functions p on (S, ν) that give X_i the expected value c_i:

$$\int_S X_i p \,\mathrm{d}\nu = c_i. \tag{7.3}$$

The property (7.3) is preserved for convex-linear combinations of p's.

From a physical standpoint, S is a state space (say, all possible positions of particles in a box), the X_i's are macroscopic physical observables on the states in the state space (energy, speed, *etc.*), and a choice of p amounts to a weighting of which states are more or less likely to arise. The choice of Π using constraints of the type (7.3) amounts to the consideration

of only those weightings p that give the X_i's certain fixed mean values c_i. (While (7.3) is the way constraints usually arise in practice, from a purely mathematical point of view we can simply take $c_i = 0$ by replacing X_i with $X_i - c_i$ for all i.)

I am grateful to Jon Tyson for the formulation and proof of the next lemma.

Lemma 7.5. *If there are p_1 and p_2 in Π' such that, on a subset $A \subset S$ with positive measure, $p_1 = 0$ and $p_2 > 0$, then*

$$h((1-\varepsilon)p_1 + \varepsilon p_2) > h(p_1)$$

for small positive ε.

We consider $(1-\varepsilon)p_1 + \varepsilon p_2$ to be a slight perturbation of p_1 when ε is small.

Proof. Let $B = S - A$, so $h(p_1) = -\int_B p_1 \log p_1 \, d\nu$. Since p_1 and p_2 are in Π', $(1-\varepsilon)p_1 + \varepsilon p_2 \in \Pi'$ for any $\varepsilon \in [0,1]$.

As in the proof of Lemma 7.3, let $f(t) = -t \log t$ for $t \geq 0$. Since $p_1 = 0$ on A,

$$h(p_1) = \int_S f(p_1)\, d\nu = \int_B f(p_1)\, d\nu. \tag{7.4}$$

Then

$$\begin{aligned}
h((1-\varepsilon)p_1 + \varepsilon p_2) &= \int_A f((1-\varepsilon)p_1 + \varepsilon p_2)\, d\nu + \int_B f((1-\varepsilon)p_1 + \varepsilon p_2)\, d\nu \\
&= \int_A f(\varepsilon p_2)\, d\nu + \int_B f((1-\varepsilon)p_1 + \varepsilon p_2)\, d\nu \\
&\geq \int_A f(\varepsilon p_2)\, d\nu + \int_B ((1-\varepsilon)f(p_1) + \varepsilon f(p_2))\, d\nu \quad \text{by (7.1)} \\
&= \varepsilon h(p_2) - (\varepsilon \log \varepsilon) \int_A p_2 \, d\nu + (1-\varepsilon)h(p_1) \text{ by (7.4)} \\
&= h(p_1) + \varepsilon \left(-(\log \varepsilon) \int_A p_2 \, d\nu + h(p_2) - h(p_1) \right).
\end{aligned}$$

Since $\nu(A) > 0$ and $p_2 > 0$ on A, $\int_A p_2 \, d\nu > 0$. Therefore the expression inside the parentheses is positive when ε is close enough to 0, no matter the size of $h(p_2) - h(p_1)$. Thus, the overall entropy is greater than $h(p_1)$ for small ε. □

Theorem 7.6. *If Π' contains a probability density function q with maximum entropy, then every $p \in \Pi'$ vanishes almost everywhere q vanishes.*

Proof. If the conclusion is false, there is some $p \in \Pi'$ and some $A \subset S$ with $\nu(A) > 0$ such that, on A, $q = 0$ and $p > 0$. However, for sufficiently small $\varepsilon > 0$, the probability density function $(1-\varepsilon)q + \varepsilon p$ lies in Π' and has greater entropy than q by Lemma 7.5. This is a contradiction. □

A consequence of Theorem 7.6 is the essential uniqueness of the maximum entropy distribution, if it exists.

Theorem 7.7. *If q_1 and q_2 have maximum entropy in Π', then $q_1 = q_2$ almost everywhere.*

Proof. By Theorem 7.6, we can change q_1 and q_2 on a set of measure 0 in order to assume they have the same zero set, say Z. Let $Y = S - Z$, so q_1 and q_2 are positive probability density functions on Y.

Let $q = (1/2)(q_1 + q_2)$. (Any other convex-linear combination $\varepsilon q_1 + (1-\varepsilon)q_2$, for an $\varepsilon \in (0,1)$, would do just as well for this proof.) Then $q > 0$ on Y. By Lemma 7.3, $q \in \Pi'$. By Lemma 4.2,

$$h(q_1) = -\int_Y q_1 \log q_1 \, \mathrm{d}\nu \leq -\int_Y q_1 \log q \, \mathrm{d}\nu, \tag{7.5}$$

and

$$h(q_2) = -\int_Y q_1 \log q_1 \, \mathrm{d}\nu \leq -\int_Y q_1 \log q \, \mathrm{d}\nu. \tag{7.6}$$

Since

$$\begin{aligned} h(q) &= -\int_S q \log q \, \mathrm{d}\nu \\ &= -\int_Y q \log q \, \mathrm{d}\nu \\ &= -\frac{1}{2}\int_Y q_1 \log q \, \mathrm{d}\nu - \frac{1}{2}\int_Y q_2 \log q \, \mathrm{d}\nu \\ &\geq \frac{1}{2}h(q_1) + \frac{1}{2}h(q_2) \\ &= h(q_1), \end{aligned}$$

maximality implies $h(q) = h(q_1) = h(q_2)$. This forces the inequalities in (7.5) and (7.6) to be equalities, which by Lemma 4.2 forces $q_1 = q$ and $q_2 = q$ almost everywhere on Y. Therefore $q_1 = q_2$ almost everywhere on S. □

This uniqueness confirms Jaynes' remark [6, p. 623] that "the maximum-entropy distribution [...] is *uniquely* determined as the one which is maximally noncommittal with regard to missing information." (italics mine)

The physical interpretation of Theorem 7.7 is interesting. Consider a system constrained by a finite set of macroscopic mean values. This corresponds to a Π constrained according to Example 7.4. Theorem 7.7 "proves" the system has at most one equilibrium state.

In contrast to Theorem 7.7, Theorem 3.2 is a context where there are infinitely many maximum entropy distributions: all Gaussians on $\mathbf{R}$ with a fixed variance have the same entropy. This is not a counterexample to Theorem 7.7, since the property in Theorem 3.2 of having a fixed variance is *not* closed under convex-linear combinations: consider two Gaussians on $\mathbf{R}$ with the same variance and different means. (Of course "fixed variance" alone is an artificial kind of constraint to consider, but it indicates an extent to which the hypotheses of Theorem 7.7 can't be relaxed.) On the other hand, the simultaneous conditions "fixed mean" and "fixed variance" for distributions on $\mathbf{R}$ are closed under convex-linear combinations, and we already know there is only one maximum entropy distribution on $\mathbf{R}$ satisfying those two constraints, namely the Gaussian having the chosen fixed mean and variance.

Theorem 7.7 does not tell us Π has a unique maximum entropy distribution, but rather that if it has one then it is unique. A maximum entropy distribution need not exist. For instance, the set of probability density functions on $\mathbf{R}$ constrained to have a fixed mean μ is closed under convex-linear combinations, but it has no maximum entropy distribution since the one-dimensional Gaussians with mean μ and increasing variance have no maximum entropy. As another example, which answers a question of Mark Fisher, the set of probability density functions p on $\mathbf{R}$ that satisfy $\int_a^b p(x)\,\mathrm{d}x = \lambda$, where a, b, and λ are fixed with $\lambda < 1$, has no maximum entropy distribution (even though it is closed under convex linear

combinations): for any $c > b$, the probability density function

$$p(x) = \begin{cases} \lambda/(b-a), & \text{if } a \leq x \leq b, \\ (1-\lambda)/(c-b), & \text{if } b < x \leq c, \\ 0, & \text{otherwise} \end{cases}$$

has entropy

$$\begin{aligned} h(p) &= -\int_a^b p(x)\log p(x)\,dx - \int_b^c p(x)\log p(x)\,dx \\ &= -\lambda\log\lambda - (1-\lambda)\log(1-\lambda) + \lambda\log(b-a) + (1-\lambda)\log(c-b), \end{aligned}$$

which becomes arbitrarily large as $c \to \infty$. Tweaking the graph of p near a, b, and c produces continuous distributions satisfying $\int_a^b p(x)\,dx = \lambda$ that have arbitrarily large entropy. (If $\lambda = 1$ then $p = 0$ almost everywhere outside $[a, b]$ and such probability density functions have a maximum entropy distribution, namely the uniform distribution on $[a, b]$.)

The following corollary of Theorem 7.7 answers a question of Claude Girard.

Corollary 7.8. *Suppose Π is a set of probability distributions on $\mathbf{R}^n$ that is closed under convex-linear combinations and contains a maximum entropy distribution. If Π is closed under orthogonal transformations (i.e., when $p(\mathbf{x}) \in \Pi$, so is $p(A\mathbf{x})$ for any orthogonal matrix A), then the maximum entropy distribution in Π is orthogonally invariant.*

Proof. First, note that when $p(\mathbf{x})$ is a probability distribution on $\mathbf{R}^n$, so is $p(A\mathbf{x})$ since Lebesgue measure is invariant under orthogonal transformations. Therefore it makes sense to consider the possibility that Π is closed under orthogonal transformations. Moreover, when $p(\mathbf{x})$ has finite entropy and A is orthogonal, $p(A\mathbf{x})$ has the same entropy (with respect to Lebesgue measure):

$$-\int_{\mathbf{R}^n} p(A\mathbf{x})\log p(A\mathbf{x})\,d\mathbf{x} = -\int_{\mathbf{R}^n} p(\mathbf{x})\log p(\mathbf{x})\,d\mathbf{x}.$$

Thus, if Π is closed under convex-linear combinations (so Theorem 7.7 applies to Π') and any orthogonal change of variables, and Π' contains a maximum entropy distribution, the uniqueness of this distribution forces it to be an orthogonally invariant function: $p(\mathbf{x}) = p(A\mathbf{x})$ for every orthogonal A and almost all $\mathbf{x}$ in $\mathbf{R}^n$. □

Example 7.9. For a real number $t > 0$, let Π_t be the set of probability distributions $p(\mathbf{x})$ on $\mathbf{R}^n$ satisfying the conditions

$$\int_{\mathbf{R}^n} (\mathbf{x}\cdot\mathbf{v})p(\mathbf{x})\,d\mathbf{x} = 0, \quad \int_{\mathbf{R}^n} (\mathbf{x}\cdot\mathbf{x})p(\mathbf{x})\,d\mathbf{x} = t, \tag{7.7}$$

where $\mathbf{v}$ runs over $\mathbf{R}^n$. The first condition (over all $\mathbf{v}$) is equivalent to the coordinate conditions

$$\int_{\mathbf{R}^n} x_i p(\mathbf{x})\,d\mathbf{x} = 0 \tag{7.8}$$

for $i = 1, 2, \ldots, n$, so it is saying (in a more geometric way than (7.8)) that any coordinate of a vector chosen randomly according to p has expected value 0. The second condition says the expected squared length of vectors chosen randomly according to p is t.

The conditions in (7.7) are trivially closed under convex-linear combinations on p. They are also closed under any orthogonal transformation on p, since for any orthogonal matrix

A and $\mathbf{v} \in \mathbf{R}^n$,

$$\begin{aligned}\int_{\mathbf{R}^n} (\mathbf{x}\cdot\mathbf{v})p(A\mathbf{x})\,d\mathbf{x} &= \int_{\mathbf{R}^n} (A\mathbf{x}\cdot A\mathbf{v})p(A\mathbf{x})\,d\mathbf{x}\\ &= \int_{\mathbf{R}^n} (\mathbf{x}\cdot A\mathbf{v})p(\mathbf{x})\,d\mathbf{x}\\ &= 0\end{aligned}$$

and

$$\begin{aligned}\int_{\mathbf{R}^n} (\mathbf{x}\cdot\mathbf{x})p(A\mathbf{x})\,d\mathbf{x} &= \int_{\mathbf{R}^n} (A\mathbf{x}\cdot A\mathbf{x})p(A\mathbf{x})\,d\mathbf{x}\\ &= \int_{\mathbf{R}^n} (\mathbf{x}\cdot\mathbf{x})p(\mathbf{x})\,d\mathbf{x}\\ &= t.\end{aligned}$$

Thus, if Π_t has a maximum entropy distribution, Theorem 7.7 tells us it is unique and Corollary 7.8 tells us it is an orthogonally-invariant function, *i.e.*, it must be a function of $|\mathbf{x}|$. But neither Theorem 7.7 nor Corollary 7.8 guarantees that there is a maximum entropy distribution. Is there one? If so, what is it?

We will show Π_t has for its maximum entropy distribution the n-dimensional Gaussian with mean $\mathbf{0}$ and scalar covariance matrix $\Sigma = (t/n)I_n$. Our argument will not use the non-constructive Theorem 7.7 or Corollary 7.8.

By Theorem 5.5, any probability distribution on $\mathbf{R}^n$ with finite means, finite covariances (recall (5.2)), and finite entropy has its entropy bounded above by the entropy of the n-dimensional Gaussian with the same means and covariances. The conditions (7.7) defining Π_t imply any $p \in \Pi_t$ has coordinate means $\mu_i = 0$ and finite covariances (note $(x_i + x_j)^2 \leq 2x_i^2 + 2x_j^2$), with the diagonal covariance sum $\sum_{i=1}^n \sigma_{ii}$ equal to t. Therefore a maximum entropy distribution in Π_t, if one exists, lies among the n-dimensional Gaussians in Π_t, which are the distributions of the form

$$G_\Sigma(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} e^{-(1/2)\mathbf{x}\cdot\Sigma^{-1}\mathbf{x}},$$

where Σ is a positive-definite symmetric matrix with trace t. The entropy of G_Σ is

$$h(G_\Sigma) = \frac{1}{2}(n + \log((2\pi)^n \det \Sigma)).$$

The arithmetic-geometric mean inequality on the (positive) eigenvalues of Σ says

$$\frac{1}{n}\operatorname{Tr}(\Sigma) \geq \sqrt[n]{\det \Sigma}$$

with equality if and only if all the eigenvalues are equal. Therefore

$$h(G_\Sigma) \leq \frac{n}{2}\left(1 + \log\left(\frac{2\pi t}{n}\right)\right). \tag{7.9}$$

This upper bound is achieved exactly when $\sqrt[n]{\det \Sigma} = t/n$, which occurs when the eigenvalues of Σ are all equal. Since Σ has trace t, the common eigenvalue is t/n, so $\Sigma = (t/n)I_n$. The Gaussian $G_{(t/n)I_n}$ has maximum entropy in Π_t.

Remark 7.10. In equation (7.9), the right side can be rewritten as $h(G_{(t/n)I_n})$, which shows the arithmetic-geometric mean inequality (both the inequality and the condition describing when equality occurs) is equivalent to a special case of entropy maximization.

Replacing the standard inner product on $\mathbf{R}^n$ with an inner product attached to any positive-definite quadratic form, we can extend Example 7.9 to an entropic characterization of Gaussian distributions that is more geometric than the entropic characterization in Theorem 5.5.

Theorem 7.11. *Let Q be a positive-definite quadratic form on $\mathbf{R}^n$, and $t > 0$. Consider the set of probability distributions p on $\mathbf{R}^n$ such that*

$$\int_{\mathbf{R}^n} \langle \mathbf{x}, \mathbf{v}\rangle_Q p(\mathbf{x})\, d\mathbf{x} = 0, \quad \int_{\mathbf{R}^n} Q(\mathbf{x})p(\mathbf{x})\, d\mathbf{x} = t, \tag{7.10}$$

where $\mathbf{v}$ runs over $\mathbf{R}^n$ and $\langle \cdot, \cdot \rangle_Q$ is the bilinear form attached to Q.

Entropy in this set of distributions is maximized at the Gaussian with $\boldsymbol{\mu} = \mathbf{0}$ and inverse covariance matrix Σ^{-1} attached to the quadratic form $(n/t)Q$: $\mathbf{x} \cdot \Sigma^{-1}\mathbf{x} = (n/t)Q(\mathbf{x})$.

Proof. Relative to the usual inner product on $\mathbf{R}^n$, $Q(\mathbf{x}) = \mathbf{x} \cdot C\mathbf{x}$ and $\langle \mathbf{x}, \mathbf{y}\rangle_Q = \mathbf{x} \cdot C\mathbf{y}$ for some positive-definite symmetric matrix C. We can write $C = D^2$ for a symmetric matrix D with positive eigenvalues. The conditions (7.10) become

$$\int_{\mathbf{R}^n} (\mathbf{x} \cdot \mathbf{v}) \frac{p(D^{-1}\mathbf{x})}{\det D}\, d\mathbf{x} = 0, \quad \int_{\mathbf{R}^n} (\mathbf{x} \cdot \mathbf{x}) \frac{p(D^{-1}\mathbf{x})}{\det D}\, d\mathbf{x} = t,$$

as $\mathbf{v}$ runs over $\mathbf{R}^n$. Let $p_D(\mathbf{x}) = p(D^{-1}\mathbf{x})/\det D$, which is also a probability distribution. Thus, p satisfies (7.10) exactly when $p_D(\mathbf{x})$ satisfies (7.7). Moreover, a calculation shows $h(p_D) = h(p) + \log(\det D)$, and the constant $\log(\det D)$ is independent of p. It depends only on the initial choice of quadratic form Q. Therefore, our calculations in Example 7.9 tell us that the unique maximum-entropy distribution for the constraints (7.10) is the p such that

$$p_D(\mathbf{x}) = G_{(t/n)I_n}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n (t/n)^n}} e^{-(1/2)(n/t)\mathbf{x}\cdot\mathbf{x}},$$

Unwinding the algebra, this says

$$\begin{aligned} p(\mathbf{x}) &= \frac{\det D}{\sqrt{(2\pi)^n (t/n)^n}} e^{-(1/2)(n/t)D\mathbf{x}\cdot D\mathbf{x}} \\ &= \frac{1}{\sqrt{(2\pi)^n (t/n)^n (\det D^{-1})^2}} e^{-(1/2)\mathbf{x}\cdot(n/t)D^2\mathbf{x}} \\ &= \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} e^{-(1/2)\mathbf{x}\cdot\Sigma^{-1}\mathbf{x}}, \end{aligned}$$

where $\Sigma = (t/n)C^{-1}$. This is a Gaussian, with $\mathbf{x} \cdot \Sigma^{-1}\mathbf{x} = \mathbf{x} \cdot (n/t)C\mathbf{x} = (n/t)Q(\mathbf{x})$. □

In (7.10), the second condition can be described in terms of traces. Letting $\sigma_{ij} = \int_{\mathbf{R}^n} x_i x_j p(\mathbf{x})\, d\mathbf{x}$ be the covariances of p, and $\Sigma_p = (\sigma_{ij})$ be the corresponding covariance matrix for p, the second equation in (7.10) says $\mathrm{Tr}(\Sigma_p C) = t$, where C is the symmetric matrix attached to Q.

The conclusion of Theorem 7.11 shows it is $(n/t)Q$, rather than the initial Q, which is preferred. Rescaling Q does not affect the vanishing in the first condition in (7.10), but changes both sides of the second condition in (7.10). Rescaling by n/t corresponds to making the right side equal to n (which reminds us of (5.3)). In other words, once we choose a particular geometry for $\mathbf{R}^n$ (a choice of positive-definite quadratic form, up to scaling), scale the choice so random vectors have expected squared length equal to n. With this constraint, the maxmium entropy distribution with mean vector $\mathbf{0}$ is the Gaussian with mean vector $\mathbf{0}$ and inverted covariance matrix equal to the matrix defined by Q: for $\mathbf{x} \in \mathbf{R}^n$, $\mathbf{x} \cdot \Sigma^{-1}\mathbf{x} = Q(\mathbf{x})$.

Appendix A. Lagrange Multipliers

To prove a probability distribution is a maximum entropy distribution, an argument often used in place of Theorem 4.3 is the method of Lagrange multipliers. We will revisit several examples and show how this idea is carried out. In the spirit of how this is done in practice, we will totally ignore the need to show Lagrange multipliers is giving us a *maximum*.

Example A.1. (Theorem 3.1) Among all probability distributions $\{p_1, \ldots, p_n\}$ on a finite set, to maximize the entropy we want to maximize $-\sum_{i=1}^n p_i \log p_i$ with the constraint $p_i \geq 0$ and $\sum_{i=1}^n p_i = 1$. Set

$$F(p_1, \ldots, p_n, \lambda) = -\sum_{i=1}^n p_i \log p_i + \lambda \left(\sum_{i=1}^n p_i - 1 \right),$$

where λ is a Lagrange multiplier and the p_i's are (positive) real variables. Since $\partial F/\partial p_j = -1 - \log p_j + \lambda$, at a maximum entropy distribution we have $-1 - \log p_j + \lambda = 0$, so $p_j = e^{\lambda - 1}$ for all j. This is a constant value, and the condition $p_1 + \cdots + p_n = 1$ implies $e^{\lambda-1} = 1/n$. Thus $p_i = 1/n$ for $i = 1, \ldots, n$.

Example A.2. (Theorem 3.2) Let $p(x)$ be a probability distribution on $\mathbf{R}$ with mean μ and variance σ^2. To maximize $-\int_{\mathbf{R}} p(x) \log p(x)\, dx$ subject to these constraints, we consider

$$\begin{aligned} F(p, \lambda_1, \lambda_2, \lambda_3) &= -\int_{\mathbf{R}} p(x) \log p(x)\, dx + \lambda_1 \left(\int_{\mathbf{R}} p(x)\, dx - 1 \right) + \lambda_2 \left(\int_{\mathbf{R}} x p(x)\, dx - \mu \right) \\ &\quad + \lambda_3 \left(\int_{\mathbf{R}} (x-\mu)^2 p(x)\, dx - \sigma^2 \right) \\ &= \int_{\mathbf{R}} (-p(x) \log p(x) + \lambda_1 p(x) + \lambda_2 x p(x) + \lambda_3 (x-\mu)^2 p(x))\, dx - \lambda_1 \\ &\quad - \mu\lambda_2 - \sigma^2 \lambda_3 \\ &= \int_{\mathbf{R}} \mathcal{L}(x, p(x), \lambda_1, \lambda_2, \lambda_3)\, dx - \lambda_1 - \mu\lambda_2 - \sigma^2\lambda_3, \end{aligned}$$

where $\mathcal{L}(x, p, \lambda_1, \lambda_2, \lambda_3) = -p \log p + \lambda_1 p + \lambda_2 x p + \lambda_3 (x-\mu)^2 p$. Since[2]

$$\frac{\partial \mathcal{L}}{\partial p} = -1 - \log p + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2,$$

at a maximum entropy distribution we have $-1 - \log p(x) + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2 = 0$, so $p(x) = e^{\lambda_1 - 1 + \lambda_2 x + \lambda_3 (x-\mu)^2}$. For $\int_{\mathbf{R}} p(x)\, dx$ to be finite requires $\lambda_2 = 0$ and $\lambda_3 < 0$. Thus $p(x) = e^a e^{-b(x-\mu)^2}$, where $a = \lambda_1 - 1$ and $b = -\lambda_3 > 0$.

The integral $\int_{\mathbf{R}} e^a e^{-b(x-\mu)^2}\, dx$ is $e^a \sqrt{\pi/b}$, so $p(x)$ being a probability distribution makes it $\sqrt{b/\pi} e^{-b(x-\mu)^2}$. Then $\int_{\mathbf{R}} x p(x)\, dx$ is automatically μ, and $\int_{\mathbf{R}} (x-\mu)^2 p(x)\, dx$ is $1/(2b)$, so $b = 1/(2\sigma^2)$. Thus $p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(1/2)(x-\mu)^2/\sigma^2}$, a normal distribution.

[2] In $\partial \mathcal{L}/\partial p$, p is treated as an indeterminate, not as $p(x)$.

Example A.3. (Theorem 3.3) Let $p(x)$ be a probability distribution on $(0,\infty)$ with mean μ. To maximize $-\int_{\mathbf{R}} p(x)\log p(x)\,dx$ subject to this constraint, set

$$\begin{aligned} F(p,\lambda_1,\lambda_2) &= -\int_0^\infty p(x)\log p(x)\,dx + \lambda_1\left(\int_0^\infty p(x)\,dx - 1\right) + \lambda_2\left(\int_0^\infty xp(x)\,dx - \mu\right) \\ &= \int_0^\infty (-p(x)\log p(x) + \lambda_1 p(x) + \lambda_2 xp(x))\,dx - \lambda_1 - \lambda_2\mu \\ &= \int_0^\infty \mathcal{L}(x,p(x),\lambda_1,\lambda_2)\,dx - \lambda_1 - \lambda_2\mu, \end{aligned}$$

where $\mathcal{L}(x,p,\lambda_1,\lambda_2) = -p\log p + \lambda_1 p + \lambda_2 xp$. Then $\partial\mathcal{L}/\partial p = -1 - \log p + \lambda_1 + \lambda_2 x$, so at a maximum entropy distribution we have $-1 - \log p(x) + \lambda_1 + \lambda_2 x = 0$, and thus $p(x) = e^{\lambda_1 - 1 + \lambda_2 x}$ for $x \geq 0$. The condition $\int_0^\infty p(x)\,dx < \infty$ requires $\lambda_2 < 0$. Then $\int_0^\infty e^{\lambda_1-1+\lambda_2 x}\,dx = e^{\lambda_1-1}\int_0^\infty e^{\lambda_2 x}\,dx = e^{\lambda_1-1}/|\lambda_2|$, so $e^{\lambda_1-1} = |\lambda_2|$. Thus $p(x) = |\lambda_2|e^{\lambda_2 x}$. Since $\int_0^\infty xe^{\lambda_2 x}\,dx = 1/\lambda_2^2$, the condition $\int_0^\infty xp(x)\,dx = \mu$ implies $\lambda_2 = -1/\mu$, so $p(x) = (1/\mu)e^{-x/\mu}$, which is an exponential distribution.

Example A.4. (Theorem 5.1) When $p(x)$ is a probability distribution on $[a,b]$ with mean $\mu \in (a,b)$, we maximize $-\int_a^b p(x)\log p(x)\,dx$ by looking at

$$F(p,\lambda_1,\lambda_2) = -\int_a^b p(x)\log p(x)\,dx + \lambda_1\left(\int_a^b p(x)\,dx - 1\right) + \lambda_2\left(\int_a^b xp(x)\,dx - \mu\right),$$

exactly as in the previous example. The same ideas from the previous example imply $p(x) = e^{\lambda_1-1}e^{\lambda_2 x}$ for $x \in [a,b]$, which has the shape $Ce^{\alpha x}$ from Theorem 5.1. The two constraints $\int_a^b p(x)\,dx = 1$ and $\int_a^b xp(x)\,dx = \mu$ pin down λ_1 and λ_2.

Example A.5. (Theorem 5.7) When $\{p_k, p_{k+1}, p_{k+2}, \dots\}$ is a probability distribution on $\{k, k+1, k+2, \dots\}$ with mean μ, it has maximum entropy among such distributions when

$$F(p_k,p_{k+1},\dots,\lambda_1,\lambda_2) = -\sum_{i\geq k} p_i\log p_i + \lambda_1\left(\sum_{i\geq k} p_i - 1\right) + \lambda_2\left(\sum_{i\geq k} ip_i - \mu\right)$$

satisfies $\partial F/\partial p_j = 0$ for all j. The derivative is $-1 - \log p_j + \lambda_1 + \lambda_2 j$, so $p_j = e^{\lambda_1-1}e^{\lambda_2 j}$ for all j. To have $\sum_{j\geq k} p_j < \infty$ requires $\lambda_2 < 0$. Set $r = e^{\lambda_2} \in (0,1)$, so $p_j = Cr^j$, where $C = e^{\lambda_1-1}$.

The condition $\sum_{i\geq k} p_i = 1$ implies $C = (1-r)/r^k$, so $p_i = (1-r)r^{i-k}$. The condition that the mean is μ determines the value of r, as in the proof of Theorem 5.7.

Example A.6. (Theorem 5.11) Let $E_1,\dots,E_n$ be in $\mathbf{R}$ with $\overline{E}$ lying between $\min E_j$ and $\max E_j$. To find the maximum entropy distribution $\{p_1,\dots,p_n\}$ where $\sum_{i=1}^n p_i E_i = \overline{E}$, consider

$$F(p_1,\dots,p_n,\lambda_1,\lambda_2) = -\sum_{i=1}^n p_i\log p_i + \lambda_1\left(\sum_{i=1}^n p_i - 1\right) + \lambda_2\left(\sum_{i=1}^n p_iE_i - \overline{E}\right).$$

Since $\partial F/\partial p_j = -1 - \log p_j + \lambda_1 + \lambda_2 E_j$, at a maximum entropy distribution we have $-1 - \log p_j + \lambda_1 + \lambda_2 E_j = 0$, so $p_j = e^{\lambda_1-1+\lambda_2 E_j}$. The condition $\sum_{j=1}^n p_j = 1$ becomes $e^{\lambda_1-1}\sum_{i=1}^n e^{\lambda_2 E_j} = 1$, so $e^{\lambda_1-1} = 1/\sum_{j=1}^n e^{\lambda_2 E_j}$. Thus

$$p_i = e^{\lambda_1-1}e^{\lambda_2 E_i} = \frac{e^{\lambda_2 E_i}}{\sum_{j=1}^n e^{\lambda_2 E_j}}.$$

This agrees with (5.5), where λ_2 needs to be chosen so that $\sum_{i=1}^n p_i E_i = \overline{E}$, which is the same as λ_2 satisfying

$$\frac{\sum_{i=1}^n E_i e^{\lambda_2 E_i}}{\sum_{j=1}^n e^{\lambda_2 E_i}} = \overline{E}.$$

To align the notation with Theorem 5.11, we should write λ_2 as $-\beta$.

It is left to the reader to work out other maximum entropy distributions using Lagrange multipliers, such as Theorems 5.2, 5.5, and 5.9.

References

[1] S. Artstein, K. M. Ball, F. Barthe, and A. Naor, *Solution of Shannon's Problem on the Monotonicity of Entropy*, J. Amer. Math. Soc. **17** (2004), 975–982.

[2] B. Buck and V. A. Macaulay (eds.), Maximum Entropy in Action, Clarendon Press, Oxford, 1991.

[3] D. K. Faddeev, The notion of entropy of finite probabilistic schemes (Russian), *Uspekhi Mat. Nauk* **11** (1956), 15–19.

[4] A. Feinstein, "The Foundations of Information Theory," McGraw-Hill, New York, 1958.

[5] S. Goldman, "Information Theory," Prentice-Hall, New York, 1955.

[6] E. T. Jaynes, "Information theory and statistical mechanics," Phys. Rev. **106** (1957), 620–630.

[7] E. T. Jaynes, "Information theory and statistical mechanics, II" Phys. Rev. **108** (1957), 171–190.

[8] E. T. Jaynes, Information theory and statistical mechanics, pp. 181–218 of "Statistical Physics," Brandeis Summer Insitute 1962, W. A. Benjamin, New York, 1963.

[9] J. Justice (ed.), Maximum Entropy and Bayesian Methods in Applied Statistics, Cambridge Univ. Press, Cambridge, 1986.

[10] R. D. Levine and M. Tribus (eds.), The Maximum Entropy Formalism, The MIT Press, Cambridge, 1979.

[11] Math Overflow, http://mathoverflow.net/questions/116667/whats-the-maximum-entropy-probability-distribution-given-bounds-a-b-and-mean.

[12] G. P. Patil, S. Kotz, and J. K. Ord (eds.), A Modern Course on Statistical Distributions in Scientific Work Vol. 3: Characterizations and Applications D. Reidel Publ. Company, Dordrecht, 1975.

[13] C. E. Shannon and W. Weaver, "The Mathematical Theory of Communication," Univ. of Illinois Press, Urbana, IL, 1949.