CHAPTER 12

MAXIMUM ENTROPY

The temperature of a gas corresponds to the average kinetic energy of the molecules in the gas. What can we say about the distribution of velocities in the gas at a given temperature? We know from physics that this distribution is the maximum entropy distribution under the temperature constraint, otherwise known as the Maxwell–Boltzmann distribution. The maximum entropy distribution corresponds to the macrostate (as indexed by the empirical distribution) that has the most microstates (the individual gas velocities). Implicit in the use of maximum entropy methods in physics is a sort of AEP which says that all microstates are equally probable.

12.1 MAXIMUM ENTROPY DISTRIBUTIONS

Consider the following problem: Maximize the entropy $h(f)$ over all probability densities f satisfying

$$\begin{aligned} &1.\quad f(x) \ge 0, \text{ with equality outside the support set } S \\ &2.\quad \int_S f(x)\,dx = 1 \\ &3.\quad \int_S f(x) r_i(x)\,dx = \alpha_i \quad \text{for } 1 \le i \le m. \end{aligned} \tag{12.1}$$

Thus, f is a density on support set S meeting certain moment constraints $\alpha_1, \alpha_2, \ldots, \alpha_m$.

Approach 1 (Calculus) The differential entropy $h(f)$ is a concave function over a convex set. We form the functional

$$J(f) = -\int f \ln f + \lambda_0 \int f + \sum_{i=1}^{m} \lambda_i \int f r_i \tag{12.2}$$

and "differentiate" with respect to $f(x)$, the xth component of f, to obtain

$$\frac{\partial J}{\partial f(x)} = -\ln f(x) - 1 + \lambda_0 + \sum_{i=1}^{m} \lambda_i r_i(x). \tag{12.3}$$

Elements of Information Theory, Second Edition, By Thomas M. Cover and Joy A. Thomas

Setting this equal to zero, we obtain the form of the maximizing density

$$f(x) = e^{\lambda_0 - 1 + \sum_{i=1}^{m} \lambda_i r_i(x)}, \qquad x \in S, \tag{12.4}$$

where $\lambda_0, \lambda_1, \ldots, \lambda_m$ are chosen so that f satisfies the constraints.

The approach using calculus only suggests the form of the density that maximizes the entropy. To prove that this is indeed the maximum, we can take the second variation. It is simpler to use the information inequality $D(g||f) \geq 0$.

Approach 2 *(Information inequality)* If g satisfies (12.1) and if f^* is of the form (12.4), then $0 \leq D(g||f^*) = -h(g) + h(f^*)$. Thus $h(g) \leq h(f^*)$ for all g satisfying the constraints. We prove this in the following theorem.

Theorem 12.1.1 (*Maximum entropy distribution*) *Let $f^*(x) = f_\lambda(x) = e^{\lambda_0 + \sum_{i=1}^{m} \lambda_i r_i(x)}$, $x \in S$, where $\lambda_0, \ldots, \lambda_m$ are chosen so that f^* satisfies (12.1). Then f^* uniquely maximizes $h(f)$ over all probability densities f satisfying constraints (12.1).*

Proof: Let g satisfy the constraints (12.1). Then

$$\begin{aligned}
h(g) &= -\int_S g \ln g && (12.5)\\
&= -\int_S g \ln \frac{g}{f^*} f^* && (12.6)\\
&= -D(g||f^*) - \int_S g \ln f^* && (12.7)\\
&\overset{(a)}{\leq} -\int_S g \ln f^* && (12.8)\\
&\overset{(b)}{=} -\int_S g \left(\lambda_0 + \sum \lambda_i r_i\right) && (12.9)\\
&\overset{(c)}{=} -\int_S f^* \left(\lambda_0 + \sum \lambda_i r_i\right) && (12.10)\\
&= -\int_S f^* \ln f^* && (12.11)\\
&= h(f^*), && (12.12)
\end{aligned}$$

where (a) follows from the nonnegativity of relative entropy, (b) follows from the definition of f^*, and (c) follows from the fact that both f^* and g satisfy the constraints. Note that equality holds in (a) if and only if $g(x) = f^*(x)$ for all x, except for a set of measure 0, thus proving uniqueness. □

The same approach holds for discrete entropies and for multivariate distributions.

12.2 EXAMPLES

Example 12.2.1 (*One-dimensional gas with a temperature constraint*) Let the constraints be $EX = 0$ and $EX^2 = \sigma^2$. Then the form of the maximizing distribution is

$$f(x) = e^{\lambda_0 + \lambda_1 x + \lambda_2 x^2}. \tag{12.13}$$

To find the appropriate constants, we first recognize that this distribution has the same form as a normal distribution. Hence, the density that satisfies the constraints and also maximizes the entropy is the $\mathcal{N}(0, \sigma^2)$ distribution:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}. \tag{12.14}$$

Example 12.2.2 (*Dice, no constraints*) Let $S = \{1, 2, 3, 4, 5, 6\}$. The distribution that maximizes the entropy is the uniform distribution, $p(x) = \frac{1}{6}$ for $x \in S$.

Example 12.2.3 (*Dice, with $EX = \sum i p_i = \alpha$*) This important example was used by Boltzmann. Suppose that n dice are thrown on the table and we are told that the total number of spots showing is $n\alpha$. What proportion of the dice are showing face i, $i = 1, 2, \ldots, 6$?

One way of going about this is to count the number of ways that n dice can fall so that n_i dice show face i. There are $\binom{n}{n_1, n_2, \ldots, n_6}$ such ways. This is a macrostate indexed by $(n_1, n_2, \ldots, n_6)$ corresponding to $\binom{n}{n_1, n_2, \ldots, n_6}$ microstates, each having probability $\frac{1}{6^n}$. To find the most probable macrostate, we wish to maximize $\binom{n}{n_1, n_2, \ldots, n_6}$ under the constraint observed on the total number of spots,

$$\sum_{i=1}^{6} i n_i = n\alpha. \tag{12.15}$$

Using a crude Stirling's approximation, $n! \approx (\frac{n}{e})^n$, we find that

$$\binom{n}{n_1, n_2, \ldots, n_6} \approx \frac{(\frac{n}{e})^n}{\prod_{i=1}^{6} (\frac{n_i}{e})^{n_i}} \tag{12.16}$$

$$= \prod_{i=1}^{6} \left(\frac{n}{n_i}\right)^{n_i} \tag{12.17}$$

$$= e^{nH\left(\frac{n_1}{n}, \frac{n_2}{n}, \ldots, \frac{n_6}{n}\right)}. \tag{12.18}$$

Thus, maximizing $\binom{n}{n_1, n_2, \ldots, n_6}$ under the constraint (12.15) is almost equivalent to maximizing $H(p_1, p_2, \ldots, p_6)$ under the constraint $\sum ip_i = \alpha$. Using Theorem 12.1.1 under this constraint, we find the maximum entropy probability mass function to be

$$p_i^* = \frac{e^{\lambda i}}{\sum_{i=1}^{6} e^{\lambda i}}, \tag{12.19}$$

where λ is chosen so that $\sum ip_i^* = \alpha$. Thus, the most probable macrostate is $(np_1^*, np_2^* \ldots, np_6^*)$, and we expect to find $n_i^* = np_i^*$ dice showing face i.

In Chapter 11 we show that the reasoning and the approximations are essentially correct. In fact, we show that not only is the maximum entropy macrostate the most likely, but it also contains almost all of the probability. Specifically, for rational α,

$$\Pr\left\{\left|\frac{N_i}{n} - p_i^*\right| < \epsilon, i = 1, 2, \ldots, 6 \left| \sum_{i=1}^{n} X_i = n\alpha \right.\right\} \to 1, \tag{12.20}$$

as $n \to \infty$ along the subsequence such that $n\alpha$ is an integer.

Example 12.2.4 Let $S = [a, b]$, with no other constraints. Then the maximum entropy distribution is the uniform distribution over this range.

Example 12.2.5 $S = [0, \infty)$ and $EX = \mu$. Then the entropy-maximizing distribution is

$$f(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}}, \quad x \geq 0. \tag{12.21}$$

This problem has a physical interpretation. Consider the distribution of the height X of molecules in the atmosphere. The average potential energy of the molecules is fixed, and the gas tends to the distribution that has the maximum entropy subject to the constraint that $E(mgX)$ is fixed. This is the exponential distribution with density $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$. The density of the atmosphere does indeed have this distribution.

Example 12.2.6 $S = (-\infty, \infty)$, and $EX = \mu$. Here the maximum entropy is infinite, and there is no maximum entropy distribution. (Consider normal distributions with larger and larger variances.)

Example 12.2.7 $S = (-\infty, \infty)$, $EX = \alpha_1$, and $EX^2 = \alpha_2$. The maximum entropy distribution is $\mathcal{N}(\alpha_1, \alpha_2 - \alpha_1^2)$.

Example 12.2.8 $S = \mathcal{R}^n$, $EX_iX_j = K_{ij}$, $1 \le i, j \le n$. This is a multivariate example, but the same analysis holds and the maximum entropy density is of the form

$$f(\mathbf{x}) = e^{\lambda_0 + \sum_{i,j} \lambda_{ij} x_i x_j}. \tag{12.22}$$

Since the exponent is a quadratic form, it is clear by inspection that the density is a multivariate normal with zero mean. Since we have to satisfy the second moment constraints, we must have a multivariate normal with covariance K_{ij}, and hence the density is

$$f(\mathbf{x}) = \frac{1}{\left(\sqrt{2\pi}\right)^n |K|^{1/2}} e^{-\frac{1}{2}\mathbf{x}^T K^{-1} \mathbf{x}}, \tag{12.23}$$

which has an entropy

$$h(\mathcal{N}_n(0, K)) = \frac{1}{2} \log(2\pi e)^n |K|, \tag{12.24}$$

as derived in Chapter 8.

Example 12.2.9 Suppose that we have the same constraints as in Example 12.2.8, but $EX_iX_j = K_{ij}$ only for some restricted set of $(i, j) \in A$. For example, we might know only K_{ij} for $i = j \pm 2$. Then by comparing (12.22) and (12.23), we can conclude that $(K^{-1})_{ij} = 0$ for $(i, j) \in A^c$ (i.e., the entries in the inverse of the covariance matrix are 0 when (i, j) is outside the constraint set).

12.3 ANOMALOUS MAXIMUM ENTROPY PROBLEM

We have proved that the maximum entropy distribution subject to the constraints

$$\int_S h_i(x) f(x)\, dx = \alpha_i \tag{12.25}$$

is of the form

$$f(x) = e^{\lambda_0 + \sum \lambda_i h_i(x)} \tag{12.26}$$

if $\lambda_0, \lambda_1, \ldots, \lambda_p$ satisfying the constraints (12.25) exist.

We now consider a tricky problem in which the λ_i cannot be chosen to satisfy the constraints. Nonetheless, the "maximum" entropy can be found. We consider the following problem: Maximize the entropy subject to the constraints

$$\int_{-\infty}^{\infty} f(x)\,dx = 1, \tag{12.27}$$

$$\int_{-\infty}^{\infty} xf(x)\,dx = \alpha_1, \tag{12.28}$$

$$\int_{-\infty}^{\infty} x^2 f(x)\,dx = \alpha_2, \tag{12.29}$$

$$\int_{-\infty}^{\infty} x^3 f(x)\,dx = \alpha_3. \tag{12.30}$$

Here, the maximum entropy distribution, if it exists, must be of the form

$$f(x) = e^{\lambda_0+\lambda_1 x+\lambda_2 x^2+\lambda_3 x^3}. \tag{12.31}$$

But if λ_3 is nonzero, $\int_{-\infty}^{\infty} f = \infty$ and the density cannot be normalized. So λ_3 must be 0. But then we have four equations and only three variables, so that in general it is not possible to choose the appropriate constants. The method seems to have failed in this case.

The reason for the apparent failure is simple: The entropy has a least upper bound under these constraints, but it is not possible to attain it. Consider the corresponding problem with only first and second moment constraints. In this case, the results of Example 12.2.1 show that the entropy-maximizing distribution is the normal with the appropriate moments. With the additional third moment constraint, the maximum entropy cannot be higher. Is it possible to achieve this value?

We cannot achieve it, but we can come arbitrarily close. Consider a normal distribution with a small "wiggle" at a very high value of x. The moments of the new distribution are almost the same as those of the old one, the biggest change being in the third moment. We can bring the first and second moments back to their original values by adding new wiggles to balance out the changes caused by the first. By choosing the position of the wiggles, we can get any value of the third moment without reducing the entropy significantly below that of the associated normal. Using this method, we can come arbitrarily close to the upper bound for the maximum entropy distribution. We conclude that

$$\sup h(f) = h(\mathcal{N}(0, \alpha_2 - \alpha_1^2)) = \frac{1}{2}\ln 2\pi e(\alpha_2 - \alpha_1^2). \tag{12.32}$$

This example shows that the maximum entropy may only be ϵ-achievable.

12.4 SPECTRUM ESTIMATION

Given a stationary zero-mean stochastic process $\{X_i\}$, we define the autocorrelation function as

$$R(k) = EX_i X_{i+k}. \tag{12.33}$$

The Fourier transform of the autocorrelation function for a zero-mean process is the power spectral density $S(\lambda)$:

$$S(\lambda) = \sum_{m=-\infty}^{\infty} R(m)e^{-im\lambda}, \quad -\pi < \lambda \le \pi, \tag{12.34}$$

where $i = \sqrt{-1}$. Since the power spectral density is indicative of the structure of the process, it is useful to form an estimate from a sample of the process.

There are many methods to estimate the power spectrum. The simplest way is to estimate the autocorrelation function by taking sample averages for a sample of length n,

$$\hat{R}(k) = \frac{1}{n-k}\sum_{i=1}^{n-k} X_i X_{i+k}. \tag{12.35}$$

If we use all the values of the sample correlation function $\hat{R}(\cdot)$ to calculate the spectrum, the estimate that we obtain from (12.34) does not converge to the true power spectrum for large n. Hence, this method, the *periodogram method*, is rarely used. One of the reasons for the problem with the periodogram method is that the estimates of the autocorrelation function from the data have different accuracies. The estimates for low values of k (called the *lags*) are based on a large number of samples and those for high k on very few samples. So the estimates are more accurate at low k. The method can be modified so that it depends only on the autocorrelations at low k by setting the higher lag autocorrelations to 0. However, this introduces some artifacts because of the sudden transition to zero autocorrelation. Various windowing schemes have been suggested to smooth out the transition. However, windowing reduces spectral resolution and can give rise to negative power spectral estimates.

In the late 1960s, while working on the problem of spectral estimation for geophysical applications, Burg suggested an alternative method. Instead of