Elements of Information Theory
Thomas M. Cover, Joy A. Thomas

Print ISBN 0-471-06259-6 Online ISBN 0-471-20061-1

Chapter 11

Maximum Entropy and Spectral Estimation

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 actual gas velocities). Implicit in the use of maximum entropy methods in physics is a sort of AEP that says that all microstates are equally probable.

11.1 MAXIMUM ENTROPY DISTRIBUTIONS

Consider the following problem:

Maximize the entropy $h(f)$ over all probability densities f satisfying

$$\begin{aligned} &1.\ f(x) \geq 0, \text{ with equality outside the support set } S, \\ &2.\ \textstyle\int_S f(x)\,dx = 1, \\ &3.\ \textstyle\int_S f(x) r_i(x)\,dx = \alpha_i, \quad \text{for } 1 \leq i \leq m. \end{aligned} \tag{11.1}$$

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

$$J(f) = -\int f \ln f + \lambda_0 \int f + \sum_{i=1}^{m} \lambda_i \int f r_i \tag{11.2}$$

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

$$\frac{\partial J}{\partial f(x)} = -\ln f(x) - 1 + \lambda_0 + \sum_{i=1}^{m} \lambda_i r_i(x) . \tag{11.3}$$

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

$$f(x) = e^{\lambda_0 - 1 + \sum_{i=1}^{m} \lambda_i r_i(x)}, \quad x \in S , \tag{11.4}$$

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 (11.1) and if f^* is of the form (11.4), 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 11.1.1 (*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* (*11.1*). *Then* f^* *uniquely maximizes* $h(f)$ *over all probability densities* f *satisfying constraints* (*11.1*).

Proof: Let g satisfy the constraints (11.1). Then

$$h(g) = -\int_S g \ln g \tag{11.5}$$

$$= -\int_S g \ln \frac{g}{f^*} f^* \tag{11.6}$$

$$= -D(g\|f^*) - \int_S g \ln f^* \tag{11.7}$$

$$\overset{(a)}{\leq} -\int_S g \ln f^* \tag{11.8}$$

$$\overset{(b)}{=} -\int_S g\left(\lambda_0 + \sum \lambda_i r_i\right) \tag{11.9}$$

$$\overset{(c)}{=} -\int_S f^*\left(\lambda_0 + \sum \lambda_i r_i\right) \tag{11.10}$$

$$= -\int_S f^* \ln f^* \tag{11.11}$$

$$= h(f^*), \tag{11.12}$$

where (a) follows from the non-negativity of relative entropy, (b) from the definition of f^* and (c) 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.

11.2 EXAMPLES

Example 11.2.1 (*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

$$f(x) = e^{\lambda_0 + \lambda_1 x + \lambda_2 x^2}. \tag{11.13}$$

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.

Example 11.2.2 (*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 11.2.3 (*Dice, with* $EX = \Sigma\, ip_i = \alpha$): This important example was used by Boltzmann. Suppose 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 observed constraint on the total number of spots,

$$\sum_{i=1}^{6} in_i = n\alpha. \tag{11.14}$$

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

$$\binom{n}{n_1, n_2, \ldots, n_6} \approx \frac{\left(\frac{n}{e}\right)^n}{\prod_{i=1}^{6} \left(\frac{n_i}{e}\right)^{n_i}} \tag{11.15}$$

$$= \prod_{i=1}^{6} \left(\frac{n}{n_i}\right)^{n_i} \tag{11.16}$$

$$= e^{nH\left(\frac{n_1}{n}, \frac{n_2}{n}, \ldots, \frac{n_6}{n}\right)} . \tag{11.17}$$

Thus maximizing $\binom{n}{n_1, n_2, \ldots, n_6}$ under the constraint (11.14) is almost equivalent to maximizing $H(p_1, p_2, \ldots, p_6)$ under the constraint $\Sigma\, ip_i = \alpha$. Using Theorem 11.1.1 under this constraint, we find the maximum entropy probability mass function to be

$$p_i^* = \frac{e^{\lambda i}}{\sum_{i=1}^{6} e^{\lambda i}} , \tag{11.18}$$

where λ is chosen so that $\Sigma\, ip_i^* = \alpha$. 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 12, we shall show that the reasoning and the approximations are essentially correct. In fact, we shall 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,

$$\Pr\left\{ \left| \frac{N_i}{n} - p_i^* \right| < \epsilon, i = 1, 2, \ldots, 6 \middle| \sum_{i=1}^{n} X_i = n\alpha \right\} \to 1 , \tag{11.19}$$

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

Example 11.2.4: Let $S = [a, b]$, with no other constraints. Then the maximum entropy distribution is the uniform distribution over this range.

Example 11.2.5: $S = [0, \infty)$ and $EX = \mu$. Then the entropy maximizing distribution is

$$f(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}}, \quad x \geq 0 . \tag{11.20}$$

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 11.2.6: $S = (-\infty, \infty)$, 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 11.2.7: $S = (-\infty, \infty)$, $EX = \alpha_1$ and $EX^2 = \alpha_2$. The maximum entropy distribution is $\mathcal{N}(\alpha_1, \alpha_2 - \alpha_1^2)$.

Example 11.2.8: $S = \mathcal{R}^n$, $EX_iX_j = K_{ij}$, $1 \leq i, j \leq n$. This is a multivariate example, but the same analysis holds and the maximum entropy density is of the form

$$f(\mathbf{x}) = e^{\lambda_0 + \sum_{i,j} \lambda_{ij} x_i x_j}. \tag{11.21}$$

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

$$f(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^n |K|^{\frac{1}{2}}} e^{-\frac{1}{2}\mathbf{x}^T K^{-1} \mathbf{x}}, \tag{11.22}$$

which has an entropy

$$h(\mathcal{N}_n(0, K)) = \frac{1}{2} \log(2\pi e)^n |K|, \tag{11.23}$$

as derived in Chapter 9.

11.3 AN ANOMALOUS MAXIMUM ENTROPY PROBLEM

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

$$\int_S h_i(x) f(x)\, dx = \alpha_i \tag{11.24}$$

is of the form

$$f(x) = e^{\lambda_0 + \sum \lambda_i h_i(x)} \tag{11.25}$$

if $\lambda_0, \lambda_1, \ldots, \lambda_p$ satisfying the constraints (11.24) 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

$$\int_{-\infty}^{\infty} f(x)\,dx = 1\,, \tag{11.26}$$

$$\int_{-\infty}^{\infty} xf(x)\,dx = \alpha_1\,, \tag{11.27}$$

$$\int_{-\infty}^{\infty} x^2 f(x)\,dx = \alpha_2\,, \tag{11.28}$$

$$\int_{-\infty}^{\infty} x^3 f(x)\,dx = \alpha_3\,. \tag{11.29}$$

In this case, the maximum entropy distribution, if it exists, must be of the form

$$f(x) = e^{\lambda_0 + \lambda_1 x + \lambda_2 x^2 + \lambda_3 x^3}\,. \tag{11.30}$$

But if λ_3 is non-zero, then $\int_{-\infty}^{\infty} f = \infty$ 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 an 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 11.2.1 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 the old one, with 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 significantly reducing the entropy 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

$$\sup h(f) = h(\mathcal{N}(0, \alpha_2 - \alpha_1^2)) = \frac{1}{2} \ln 2\pi e(\alpha_2 - \alpha_1^2) . \tag{11.31}$$

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

11.4 SPECTRUM ESTIMATION

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

$$R(k) = EX_iX_{i+k} . \tag{11.32}$$

The Fourier transform of the autocorrelation function for a zero mean process is the power spectral density $S(\lambda)$, i.e.,

$$S(\lambda) = \sum_{m=-\infty}^{\infty} R(m)e^{-im\lambda}, \quad -\pi < \lambda \leq \pi . \tag{11.33}$$

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,

$$\hat{R}(k) = \frac{1}{n-k} \sum_{i=1}^{n-k} X_iX_{i+k} . \tag{11.34}$$

If we use all the values of the sample correlation function $\hat{R}(\cdot)$ to calculate the spectrum, the estimate that we obtain from (11.33) does not converge to the true power spectrum for large n. Hence this method, called 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 estima-

tion for geophysical applications, Burg suggested an alternative method. Instead of setting the autocorrelations at high lags to zero, he set them to values that make the fewest assumptions about the data, i.e., values that maximize the entropy rate of the process. This is consistent with the maximum entropy principle as articulated by Jaynes [143]. Burg assumed the process to be stationary and Gaussian and found that the process which maximizes the entropy subject to the correlation constraints is an autoregressive Gaussian process of the appropriate order. In some applications where we can assume an underlying autoregressive model for the data, this method has proved useful in determining the parameters of the model (e.g., linear predictive coding for speech). This method (known as the maximum entropy method or Burg's method) is a popular method for estimation of spectral densities. We prove Burg's theorem in Section 11.6.

11.5 ENTROPY RATES OF A GAUSSIAN PROCESS

In Chapter 9, we defined the differential entropy of a continuous random variable. We can now extend the definition of entropy rates to real-valued stochastic processes.

Definition: The *differential entropy rate* of a stochastic process $\{X_i\}, X_i \in \mathcal{R}$, is defined to be

$$h(\mathcal{X}) = \lim_{n\to\infty} \frac{h(X_1, X_2, \ldots, X_n)}{n} \tag{11.35}$$

if the limit exists.

Just as in the discrete case, we can show that the limit exists for stationary processes and that the limit is given by the two expressions

$$h(\mathcal{X}) = \lim_{n\to\infty} \frac{h(X_1, X_2, \ldots, X_n)}{n} \tag{11.36}$$

$$= \lim_{n\to\infty} h(X_n | X_{n-1}, \ldots, X_1). \tag{11.37}$$

For any sample of a stationary Gaussian stochastic process, we have

$$h(X_1, X_2, \ldots, X_n) = \frac{1}{2} \log(2\pi e)^n |K^{(n)}|, \tag{11.38}$$

where the covariance matrix $K^{(n)}$ is Toeplitz with entries $R(0)$, $R(1), \ldots, R(n-1)$ along the top row. Thus $K_{ij}^{(n)} = R(|i-j|) = E(X_i - EX_i)(X_j - EX_j)$. As $n \to \infty$, the density of the eigenvalues of the

covariance matrix tends to a limit, which is the spectrum of the stochastic process. Indeed, Kolmogorov showed that the entropy rate of a stationary Gaussian stochastic process can be expressed as

$$h(\mathscr{X}) = \frac{1}{2}\log 2\pi e + \frac{1}{4\pi}\int_{-\pi}^{\pi} \log S(\lambda)\, d\lambda\,. \tag{11.39}$$

The entropy rate is also $\lim_{n\to\infty} h(X_n|X^{n-1})$. Since the stochastic process is Gaussian, the conditional distribution is also Gaussian and hence the conditional entropy is $\frac{1}{2}\log 2\pi e\sigma_\infty^2$, where σ_∞^2 is the variance of the error in the best estimate of X_n given the infinite past. Thus

$$\sigma_\infty^2 = \frac{1}{2\pi e}\, 2^{2h(\mathscr{X})}, \tag{11.40}$$

where $h(\mathscr{X})$ is given by (11.39). Hence the entropy rate corresponds to the minimum mean squared error of the best estimator of a sample of the process given the infinite past.

11.6 BURG'S MAXIMUM ENTROPY THEOREM

Theorem 11.6.1: *The maximum entropy rate stochastic process* $\{X_i\}$ *satisfying the constraints*

$$EX_iX_{i+k} = \alpha_k, \quad k = 0, 1, \ldots, p, \quad \textit{for all } i\,, \tag{11.41}$$

is the pth order Gauss-Markov process of the form

$$X_i = -\sum_{k=1}^{p} a_k X_{i-k} + Z_i\,, \tag{11.42}$$

where the Z_i *are i.i.d.* $\sim \mathcal{N}(0, \sigma^2)$ *and* $a_1, a_2, \ldots, a_p, \sigma^2$ *are chosen to satisfy (11.41).*

Remark: We do not assume that $\{X_i\}$ is (a) zero mean, (b) Gaussian, or (c) wide-sense stationary.

Proof: Let $X_1, X_2, \ldots, X_n$ be any stochastic process that satisfies the constraints (11.41). Let $Z_1, Z_2, \ldots, Z_n$ be a Gaussian process with the same covariance matrix as $X_1, X_2, \ldots, X_n$. Then since the multivariate normal distribution maximizes the entropy over all vector-valued random variables under a covariance constraint, we have

$$h(X_1, X_2, \ldots, X_n) \leq h(Z_1, Z_2, \ldots, Z_n) \tag{11.43}$$

$$= h(Z_1, \ldots, Z_p) + \sum_{i=p+1}^{n} h(Z_i | Z_{i-1}, Z_{i-2}, \ldots, Z_1) \tag{11.44}$$

$$\leq h(Z_1, \ldots, Z_p) + \sum_{i=p+1}^{n} h(Z_i | Z_{i-1}, Z_{i-2}, \ldots, Z_{i-p}) \tag{11.45}$$

by the chain rule and the fact that conditioning reduces entropy. Now define $Z_1', Z_2', \ldots, Z_n'$ as a pth order Gauss-Markov process with the same distribution as $Z_1, Z_2, \ldots, Z_n$ for all orders up to p. (Existence of such a process will be verified using the Yule-Walker equations immediately after the proof.) Then since $h(Z_i | Z_{i-1}, \ldots, Z_{i-p})$ depends only on the pth order distribution, $h(Z_i | Z_{i-1}, \ldots, Z_{i-p}) = h(Z_i' | Z'_{i-1}, \ldots, Z'_{i-p})$, and continuing the chain of inequalities, we obtain

$$h(X_1, X_2, \ldots, X_n) \leq h(Z_1, \ldots, Z_p) + \sum_{i=p+1}^{n} h(Z_i | Z_{i-1}, Z_{i-2}, \ldots, Z_{i-p}) \tag{11.46}$$

$$= h(Z_1', \ldots, Z_p') + \sum_{i=p+1}^{n} h(Z_i' | Z'_{i-1}, Z'_{i-2}, \ldots, Z'_{i-p}) \tag{11.47}$$

$$= h(Z_1', Z_2', \ldots, Z_n'), \tag{11.48}$$

where the last equality follows from the pth order Markovity of the $\{Z_i'\}$. Dividing by n and taking the limit, we obtain

$$\overline{\lim} \frac{1}{n} h(X_1, X_2, \ldots, X_n) \leq \lim \frac{1}{n} h(Z_1', Z_2', \ldots, Z_n') = h^*, \tag{11.49}$$

where

$$h^* = \frac{1}{2} \log 2\pi e \sigma^2, \tag{11.50}$$

which is the entropy rate of the Gauss-Markov process. Hence, the maximum entropy rate stochastic process satisfying the constraints is the pth order Gauss-Markov process satisfying the constraints. □

A bare bones summary of the proof is that the entropy of a finite segment of a stochastic process is bounded above by the entropy of a segment of a Gaussian random process with the same covariance structure. This entropy is in turn bounded above by the entropy of the minimal order Gauss-Markov process satisfying the given covariance

constraints. Such a process exists and has a convenient characterization by means of the Yule-Walker equations given below.

Note on the choice of $a_1, \ldots, a_p$ *and* σ^2: Given a sequence of covariances $R(0)$, $R(1), \ldots, R(p)$, does there exist a pth order Gauss-Markov process with these covariances? Given a process of the form (11.42), can we choose the a_k's to satisfy the constraints? Multiplying (11.42) by X_{i-l} and taking expectations, and noting that $R(k) = R(-k)$, we get

$$R(0) = -\sum_{k=1}^{p} a_k R(-k) + \sigma^2 \tag{11.51}$$

and

$$R(l) = -\sum_{k=1}^{p} a_k R(l-k), \quad l = 1, 2, \ldots . \tag{11.52}$$

These equations are called the *Yule-Walker* equations. There are $p+1$ equations in the $p+1$ unknowns $a_1, a_2, \ldots, a_p, \sigma^2$. Therefore, we can solve for the parameters of the process from the covariances.

Fast algorithms such as the Levinson algorithm and the Durbin algorithm [213] have been devised to use the special structure of these equations to efficiently calculate the coefficients $a_1, a_2, \ldots, a_p$ from the covariances. (We set $a_0 = 1$ for a consistent notation.) Not only do the Yule-Walker equations provide a convenient set of linear equations for calculating the a_k's and σ^2 from the $R(k)$'s, they also indicate how the autocorrelations behave for lags greater than p. The autocorrelations for high lags are an extension of the values for lags less than p. These values are called the Yule-Walker extension of the autocorrelations. The spectrum of the maximum entropy process is seen to be

$$S(l) = \frac{\sigma^2}{\left|1 + \sum_{k=1}^{p} a_k e^{-ikl}\right|^2} . \tag{11.53}$$

This is the maximum entropy spectral density subject to the constraints $R(0), R(1), \ldots, R(p)$.

In a practical problem, we are generally given a sample sequence $X_1, X_2, \ldots, X_n$, from which we calculate the autocorrelations. An important question is how many autocorrelation lags we should consider, i.e., what is the optimum value of p? A logically sound method is to choose the value of p that minimizes the total description length in a two stage description of the data. This method has been proposed by Rissanen [218, 223] and Barron [17] and is closely related to the idea of Kolmogorov complexity.

SUMMARY OF CHAPTER 11

Maximum entropy distribution: Let f be a probability density satisfying the constraints

$$\int_S f(x) r_i(x) = \alpha_i, \quad \text{for } 1 \le i \le m. \tag{11.54}$$

Let $f^*(x) = f_\lambda(x) = e^{\lambda_0 + \sum_{i=1}^m \lambda_i r_i(x)}$, $x \in S$, and let $\lambda_0, \ldots, \lambda_m$ be chosen so that f^* satisfies (11.54). Then f^* uniquely maximizes $h(f)$ over all f satisfying these constraints.

Maximum entropy spectral density estimation: The entropy rate of a stochastic process subject to autocorrelation constraints $R_0, R_1, \ldots, R_p$ is maximized by the pth order zero-mean Gauss-Markov process satisfying these constraints. The maximum entropy spectrum is

$$S(l) = \frac{\sigma^2}{\left|1 + \sum_{k=1}^p a_k e^{-ikl}\right|^2}. \tag{11.55}$$

PROBLEMS FOR CHAPTER 11

1. *Maximum entropy.* Find the maximum entropy density f defined for $x \ge 0$ satisfying $EX = \alpha_1, E \ln X = \alpha_2$. That is, maximize $-\int f \ln f$ subject to $\int x f(x)\, dx = \alpha_1$, $\int (\ln x) f(x)\, dx = \alpha_2$, where the integrals are over $0 \le x < \infty$. What family of densities is this?

2. *Min $D(P\|Q)$ under constraints on P.* We wish to find the (parametric form) of the probability mass function $P(x), x \in \{1, 2, \ldots\}$ that minimizes the relative entropy $D(P\|Q)$ over all P such that $\sum P(x) g_i(x) = \alpha_i, i = 1, 2, \ldots$.

   (a) Use Lagrange multipliers to guess that

   $$P^*(x) = Q(x) e^{\sum_{i=1}^\infty \lambda_i g_i(x) + \lambda_0} \tag{11.56}$$

   achieves this minimum if there exist λ_i's satisfying the α_i constraints. This generalizes the theorem on maximum entropy distributions subject to constraints.

   (b) Verify that P^* minimizes $D(P\|Q)$.

3. *Maximum entropy processes.* Find the maximum entropy rate stochastic process $\{X_i\}_{-\infty}^{\infty}$ subject to the constraints:

   (a) $EX_i^2 = 1, \quad i = 1, 2, \ldots,$

   (b) $EX_i^2 = 1, EX_iX_{i+1} = \frac{1}{2}, i = 1, 2, \ldots.$

4. Find the maximum entropy spectrum for the processes in parts (a) and (b) of Problem 3.

5. *Maximum entropy with marginals.* What is the maximum entropy distribution $p(x, y)$ that has the following marginals? Hint: You may wish to guess and verify a more general result.

$x \backslash y$	1	2	3	
1	p_{11}	p_{12}	p_{13}	1/2
2	p_{21}	p_{22}	p_{23}	1/4
3	p_{31}	p_{32}	p_{33}	1/4
	2/3	1/6	1/6	

6. *Processes with fixed marginals.* Consider the set of all densities with fixed pairwise marginals $f_{X_1, X_2}(x_1, x_2)$, $f_{X_2, X_3}(x_2, x_3), \ldots,$ $f_{X_{n-1}, X_n}(x_{n-1}, x_n)$. Show that the maximum entropy process with these marginals is the first-order (possibly time-varying) Markov process with these marginals. Identify the maximizing $f^*(x_1, x_2, \ldots, x_n)$.

7. *Every density is a maximum entropy density.* Let $f_0(x)$ be a given density. Given $r(x)$, consider the parametric family of densities $g_\alpha(x)$ maximizing $h(X)$ over all f satisfying $\int f(x)r(x)\,dx = \alpha$. Now let $r(x) = \ln f_0(x)$. Show that $g_\alpha(x) = f_0(x)$ for an appropriate choice $\alpha = \alpha_0$. Thus $f_0(x)$ is a maximum entropy density under the constraint $\int f \ln f_0 = \alpha_0$.

HISTORICAL NOTES

The maximum entropy principle arose in statistical mechanics in the nineteenth century and has been advocated for use in a broader context by Jaynes [143]. It was applied to spectral estimation by Burg [47]. The information theoretic proof of Burg's theorem is from Choi and Cover [56].