Chapter 3

Entropy on Statistical Models

Entropy is a notion taken form Thermodynamics, where it describes the uncertainty in the movement of gas particles. In this chapter the entropy will be considered as a measure of uncertainty of a random variable.

Maximum entropy distributions, with certain moment constraints, will play a central role in this chapter. They are distributions with a maximal ignorance degree towards unknown elements of the distribution. For instance, if nothing is known about a distribution defined on the interval $[a, b]$, it makes sense to express our ignorance by choosing the distribution to be the uniform one. Sometimes the mean is known. In this case the maximum entropy decreases and the distribution is not uniform any more. More precisely, among all distributions $p(x)$ defined on $(0, \infty)$ with a given mean μ, the one with the maximum entropy is the exponential distribution. Furthermore, if both the mean and the standard variation are given for a distribution $p(x)$ defined on $\mathbb{R}$, then the distribution with the largest entropy is the normal distribution.

Since the concept of entropy can be applied to any point of a statistical model, the entropy becomes a function defined on the statistical model. Then, likewise in Thermodynamics, we shall investigate the entropy maxima, as they have a distinguished role in the theory.

O. Calin and C. Udrişte, *Geometric Modeling in Probability and Statistics*,
DOI 10.1007/978-3-319-07779-6_3,

3.1 Introduction to Information Entropy

The notion of *entropy* comes originally from Thermodynamics. It is a quantity that describes the amount of disorder or randomness in a system bearing energy or information. In Thermodynamics the entropy is defined in terms of heat and temperature.

According to the second law of Thermodynamics, during any process the change in the entropy of a system and its surroundings is either zero or positive. The entropy of a free system tends to increase in time, towards a finite or infinite maximum. Some physicists define the arrow of time in the direction in which its entropy increases, see Hawking [43]. Most processes tend to increase their entropy in the long run. For instance, a house starts falling apart, an apple gets rotten, a person gets old, a car catches rust over time, etc.

Another application of entropy is in information theory, formulated by C. E. Shannon [73] in 1948 to explain aspects and problems of information and communication. In this theory a distinguished role is played by the *information source*, which produces a sequence of messages to be communicated to the receiver. The information is a measure of the freedom of choice with which a message can be selected from the set of all possible messages. The information can be measured numerically using the logarithm in base 2. In this case the resulting units are called binary digits, or *bits*. One bit measures a choice between two equally likely choices. For instance, if a coin is tossed but we are unable to see it as it lands, the landing information contains 1 bit of information. If there are N equally likely choices, the number of bits is equal to the digital logarithm of the number of choices, $\log_2 N$. In the case when the choices are not equally probable, the situation will be described in the following.

Shannon defined a quantity that measures how much information, and at which rate this information is produced by an information source. Suppose there are n possible elementary outcomes of the source, $A_1, \dots, A_n$, which occur with probabilities $p_1 = p(A_1)$, $\dots$, $p_n = p(A_n)$, so the source outcomes are described by the discrete probability distribution

event	A_1	A_2	$\dots$	A_n
probability	p_1	p_2	$\dots$	p_n

with p_i given. Assume there is an uncertainty function, $H(p_1, \dots, p_n)$, which "measures" how much "choice" is involved in selecting an event. It is fair to ask that H satisfies the following properties (Shannon's axioms):

(i) H is continuous in each p_i;

(ii) If $p_1 = \cdots = p_n = \dfrac{1}{n}$, then H is monotonic increasing function of n (i.e., for equally likely events there is more uncertainty when there are more possible events).

(iii) If a choice is broken down into two successive choices, then the initial H is the weighted sum of the individual values of H:

$$H(p_1, p_2, \dots, p_{n-1}, p'_n, p''_n) = H(p_1, p_2, \dots, p_{n-1}, p_n) + p_n H\Big(\frac{p'_n}{p_n}, \frac{p''_n}{p_n}\Big),$$

with $p_n = p'_n + p''_n$.

Shannon proved that the only function H satisfying the previous three assumptions is of the form

$$H = -k \sum_{i=1}^{n} p_i \log_2 p_i,$$

where k is a positive constant, which amounts to the choice of a unit of measure. The negative sign in front of the summation formula implies its non-negativity. This is the definition of the *information entropy* for discrete systems given by Shannon [73]. It is remarkable that this is the same expression seen in certain formulations of statistical mechanics.

Since the next sections involve integration and differentiation, it is more convenient to use the natural logarithm instead of the digital logarithm. The entropy defined by $H = -\sum_{i=1}^{n} p_i \ln p_i$ is measured in *natural units* instead of bits.[1] Sometimes this is also denoted by $H(p_1, \dots, p_n)$.

We make some more remarks regarding notation. We write $H(X)$ to denote the entropy of a random variable X, $H(p)$ to denote the entropy of a probability density p, and $H(\xi)$ to denote the entropy $H(p_\xi)$ on a statistical model with parameter ξ. The joint entropy of two random variables X and Y will be denoted by $H(X, Y)$, while

[1] Since $\log_2 x = \ln x / \ln 2 = 1.44 \ln x$, a natural unit is about 1.44 bits.

$H(X|Y)$ will be used for the conditional entropy of X given Y. These notations will be used interchangeably, depending on the context.

The entropy can be used to measure information in the following way. The information can be measured as a reduction in the uncertainty, i.e. entropy. If X and Y are random variables that describe an event, the initial uncertainty about the event is $H(X)$. After the random variable Y is revealed, the new uncertainty is $H(X|Y)$. The reduction in uncertainty, $H(X) - H(X|Y)$, is called the information conveyed about X by Y. Its symmetry property is left as an exercise in Problem 3.3, part (d).

In the case of a discrete random variable X, the entropy can be interpreted as the weighted average of the numbers $-\ln p_i$, where the weights are the probabilities of the values of the associated random variable X. Equivalently, this can be also interpreted as the expectation of the random variable that assumes the value $-\ln p_i$ with probability p_i

$$H(X) = -\sum_{i=1}^{n} P(X = x_i) \ln P(X = x_i) = E[-\ln P(X)].$$

Extending the situation from the discrete case, the uncertainty of a continuous random variable X defined on the interval (a, b) will be defined by an integral. If p denotes the probability density function of X, then the integral

$$H(X) = -\int_a^b p(x) \ln p(x)\, dx$$

defines the entropy of X, provided the integral is finite.

This chapter considers the entropy on statistical models as a function of its parameters. It provides examples of statistical manifolds and their associated entropies and deals with the main properties of the entropy regarding bounds, maximization and relation with the Fisher information metric.

3.2 Definition and Examples

Let $\mathcal{S} = \{p_\xi = p(x;\xi); \xi = (\xi^1, \dots, \xi^n) \in \mathbb{E}\}$ be a statistical model, where $p(\cdot, \xi) : \mathcal{X} \to [0, 1]$ is the probability density function which depends on parameter vector ξ. The entropy on the manifold $\mathcal{S}$ is a function $H : \mathbb{E} \to \mathbb{R}$, which is equal to the negative of the expectation of the log-likelihood function, $H(\xi) = -E_{p_\xi}[\ell_x(\xi)]$. More precisely,

$$H(\xi) = \begin{cases} -\displaystyle\int_{\mathcal{X}} p(x,\xi) \ln p(x,\xi)\, dx, & \text{if } \mathcal{X} \text{ is continuous;} \\ -\displaystyle\sum_{x\in\mathcal{X}} p(x,\xi) \ln p(x,\xi), & \text{if } \mathcal{X} \text{ is discrete.} \end{cases}$$

Since the entropy is associated with each distribution $p(x,\xi)$, we shall also use the alternate notation $H\big(p(x,\xi)\big)$. Sometimes, the entropy in the continuous case is called *differential entropy*, while in the discrete case is called *discrete entropy.*

It is worth noting that in the discrete case the entropy is always positive, while in the continuous case might be zero or negative. Since a simple scaling of parameters will modify a continuous distribution with positive entropy into a distribution with a negative entropy (see Problem 3.4.), in the continuous case there is no canonical entropy, but just a relative entropy. In order to address this drawback, the entropy is modified into the *relative information entropy*, as we shall see in Chap. 4.

The entropy can be defined in terms of a base measure on the space $\mathcal{X}$, but for keeping the exposition elementary we shall assume that $\mathcal{X} \subseteq \mathbb{R}^n$ with the Lebesgue-measure dx.

The entropy for a few standard distributions is computed in the next examples.

Example 3.2.1 (Normal Distribution) In this case $\mathcal{X} = \mathbb{R}$, $\xi = (\mu, \sigma) \in \mathbb{R} \times (0, \infty)$ and

$$p(x;\xi) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}.$$

The entropy is

$$\begin{aligned} H(\mu,\sigma) &= -\int_{\mathcal{X}} p(x) \ln p(x)\, dx \\ &= -\int_{\mathcal{X}} p(x)\Big(-\frac{1}{2}\ln(2\pi) - \ln\sigma - \frac{(x-\mu)^2}{2\sigma^2}\Big)\, dx \\ &= \frac{1}{2}\ln(2\pi) + \ln\sigma + \frac{1}{2\sigma^2}\int_{\mathcal{X}} (x-\mu)^2 p\, dx \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2}\ln(2\pi) + \ln\sigma + \frac{1}{2\sigma^2}\cdot\sigma^2 \\
&= \frac{1}{2}\ln(2\pi) + \ln\sigma + \frac{1}{2} \\
&= \ln(\sigma\sqrt{2\pi e}).
\end{aligned}$$

It follows that the entropy does not depend on μ, and is increasing logarithmically as a function of σ, with $\lim\limits_{\sigma\searrow 0} H = -\infty$, $\lim\limits_{\sigma\nearrow\infty} H = \infty$. Furthermore, the change of coordinates $\varphi : \mathbb{E} \to \mathbb{E}$ under which the entropy is invariant, i.e. $H(\xi) = H\big(\varphi(\xi)\big)$, are only the translations $\varphi(\mu, \sigma) = (\mu + k, \sigma)$, $k \in \mathbb{R}$.

Example 3.2.2 (Poisson Distribution) In this case the sample space is $\mathcal{X} = \mathbb{N}$, and the probability density

$$p(n;\xi) = e^{-\xi}\frac{\xi^n}{n!}, \qquad n \in \mathbb{N},\ \xi \in \mathbb{R}$$

depends only on one parameter, ξ. Using $\ln p(n,\xi) = -\xi + n\ln\xi - \ln(n!)$, we have

$$\begin{aligned}
H(\xi) &= -\sum_{n\geq 0} p(n,\xi)\ln p(n,\xi) \\
&= -\sum_{n\geq 0}\Big(-\xi e^{-\xi}\frac{\xi^n}{n!} + n\ln\xi e^{-\xi}\frac{\xi^n}{n!} - \ln(n!)e^{-\xi}\frac{\xi^n}{n!}\Big) \\
&= \xi e^{-\xi}\underbrace{\sum_{n\geq 0}\frac{\xi^n}{n!}}_{=e^{\xi}} - \ln\xi\, e^{-\xi}\sum_{n\geq 0}\frac{n\xi^n}{n!} + e^{-\xi}\sum_{n\geq 0}\frac{\xi^n\ln(n!)}{n!} \\
&= \xi - \ln\xi\, e^{-\xi}\xi e^{\xi} + e^{-\xi}\sum_{n\geq 0}\frac{\ln(n!)}{n!}\xi^n \\
&= \xi(1 - \ln\xi) + e^{-\xi}\sum_{n\geq 0}\frac{\ln(n!)}{n!}\xi^n.
\end{aligned}$$

We note that $\lim\limits_{\xi\searrow 0} H(\xi){=}0$ and $H(x){<}\infty$, since the series $\sum\limits_{n\geq 0}\dfrac{\xi^n\ln(n!)}{n!}$ has an infinite radius of convergence, see Problem 3.21.

Example 3.2.3 (Exponential Distribution) Consider the exponential distribution

$$p(x;\xi) = \xi e^{-\xi x}, \qquad x > 0,\ \xi > 0$$

with parameter ξ. The entropy is

$$\begin{aligned} H(\xi) &= -\int_0^\infty p(x)\ln p(x)\,dx = -\int_0^\infty \xi e^{-\xi x}(\ln\xi - \xi x)\,dx \\ &= -\xi\ln\xi\int_0^\infty e^{-\xi x}\,dx + \xi\int_0^\infty \xi e^{-\xi x}\,x\,dx \\ &= -\ln\xi\underbrace{\int_0^\infty p(x,\xi)\,dx}_{=1} + \xi\underbrace{\int_0^\infty xp(x,\xi)\,dx}_{=1/\xi} \\ &= 1-\ln\xi, \end{aligned}$$

which is a decreasing function of ξ, with $H(\xi) > 0$ for $\xi \in (0,e)$. Making the parameter change $\lambda = \dfrac{1}{\xi}$, the model becomes $p(x;\lambda) = \frac{1}{\lambda}e^{-x/\lambda}$, $\lambda > 0$. The entropy $H(\lambda) = 1 + \ln\lambda$ increases logarithmically in λ. We note the fact that the entropy is parametrization dependent.

Example 3.2.4 (Gamma Distribution) Consider the family of distributions

$$p_\xi(x) = p_{\alpha,\beta}(x) = \frac{1}{\beta^\alpha\Gamma(\alpha)}\,x^{\alpha-1}e^{-x/\beta},$$

with positive parameters $(\xi^1,\xi^2) = (\alpha,\beta)$ and $x > 0$. We shall start by showing that

$$\int_0^\infty \ln x\; p_{\alpha,\beta}(x)\,dx = \ln\beta + \psi(\alpha), \tag{3.2.1}$$

where

$$\psi(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \tag{3.2.2}$$

is the *digamma function*. Using that the integral of $p_{\alpha,\beta}(x)$ is unity, we have

$$\int_0^\infty x^{\alpha-1}\,e^{-\frac{x}{\beta}}\,dx = \beta^\alpha\,\Gamma(\alpha),$$

and differentiating with respect to α, it follows

$$\int_0^\infty \ln x\,x^{\alpha-1}\,e^{-\frac{x}{\beta}}\,dx = \ln\beta\,\beta^\alpha\,\Gamma(\alpha) + \beta^\alpha\,\Gamma'(\alpha). \tag{3.2.3}$$

Dividing by $\beta^\alpha\Gamma(\alpha)$ yields relation (3.2.1).

Since

$$\ln p_{\alpha,\beta}(x) = -\alpha \ln \beta - \ln \Gamma(\alpha) + (\alpha - 1)\ln x - \frac{x}{\beta},$$

using $\int_0^\infty p_{\alpha,\beta}(x)\,dx = 1$, $\int_0^\infty x\,p_{\alpha,\beta}(x)\,dx = \alpha\beta$ and (3.2.1), the entropy becomes

$$\begin{aligned} H(\alpha,\beta) &= -\int_0^\infty p_{\alpha,\beta}(x) \ln p_{\alpha,\beta}(x)\,dx \\ &= \alpha \ln \beta + \ln \Gamma(\alpha) - (\alpha - 1)\int_0^\infty \ln x\, p_{\alpha,\beta}(x)\,dx \\ &\quad + \frac{1}{\beta}\int_0^\infty x\,p_{\alpha,\beta}(x)\,dx \\ &= \ln \beta + (1-\alpha)\psi(\alpha) + \ln \Gamma(\alpha) + \alpha. \end{aligned}$$

Example 3.2.5 (Beta Distribution) The beta distribution on $\mathcal{X} = [0,1]$ is defined by the density

$$p_{a,b}(x) = \frac{1}{B(a,b)}\, x^{a-1}(1-x)^{b-1},$$

with $a, b > 0$ and beta function given by

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,dx. \tag{3.2.4}$$

Differentiating with respect to a and b in (3.2.4) yields

$$\begin{aligned} \partial_a B(a,b) &= \int_0^1 \ln x\, x^{a-1}(1-x)^{b-1}\,dx \\ \partial_b B(a,b) &= \int_0^1 \ln(1-x)\, x^{a-1}(1-x)^{b-1}\,dx. \end{aligned}$$

Using

$$\ln p_{a,b} = -\ln B(a,b) + (a-1)\ln x + (b-1)\ln(1-x),$$

we find

$$\begin{aligned} H(a,b) &= -\int_0^1 p_{a,b}(x)\,\ln p_{a,b}(x)\,dx \\ &= \ln B(a,b) - \frac{a-1}{B(a,b)}\int_0^1 \ln x\, x^{a-1}(1-x)^{b-1}\,dx \\ &\quad -\frac{b-1}{B(a,b)}\int_0^1 \ln(1-x)\,x^{a-1}(1-x)^{b-1}\,dx \\ &= \ln B(a,b) - (a-1)\frac{\partial_a B(a,b)}{B(a,b)} - (b-1)\frac{\partial_b B(a,b)}{B(a,b)} \\ &= \ln B(a,b) - (a-1)\partial_a \ln B(a,b) - (b-1)\partial_b \ln B(a,b). \end{aligned} \tag{3.2.5}$$

We shall express the entropy in terms of digamma function (3.2.2). Using the expression of the beta function in terms of gamma functions

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

we have

$$\ln B(a,b) = \ln\Gamma(a) + \ln\Gamma(b) - \ln\Gamma(a+b).$$

The partial derivatives of the function $B(a,b)$ are

$$\partial_a \ln B(a,b) = \psi(a) - \psi(a+b) \tag{3.2.6}$$

$$\partial_b \ln B(a,b) = \psi(b) - \psi(a+b). \tag{3.2.7}$$

Substituting in (3.2.5) yields

$$H(a,b) = \ln B(a,b) + (a+b-2)\psi(a+b) - (a-1)\psi(a) - (b-1)\psi(b). \tag{3.2.8}$$

For example

$$\begin{aligned} H(1/2,1/2) &= \ln\sqrt{2} + \ln\sqrt{2} - \psi(1) + \psi(1/2) \\ &= \ln 2 + \gamma - 2\ln 2 - \gamma = -\ln 2 < 0, \end{aligned}$$

where we used

$$\psi(1) = -\gamma = -0.5772\ldots, \qquad \psi(1/2) = -2\ln 2 - \gamma.$$

It can be shown that the entropy is always non-positive, see Problem 3.22. For $a = b = 1$ the entropy vanishes

$$H(1,1) = \ln\Gamma(1) + \ln\Gamma(1) - \ln\Gamma(2) = 0.$$

Example 3.2.6 (Lognormal Distribution) The lognormal distribution

$$p_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}, \ (\mu,\sigma) \in (0,\infty) \times (0,\infty)$$

defines a statistical model on the sample space $\mathcal{X} = (0,\infty)$. First, using the substitution $y = \ln x - \mu$, we have

$$\begin{aligned}\int_0^\infty \ln x \, p_{\mu,\sigma}(x)\,dx &= \int_0^\infty (\ln x - \mu)\, p_{\mu,\sigma}(x)\,dx + \mu \\ &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi} y\sigma} e^{-\frac{y^2}{2\sigma^2}}\,dy + \mu = \mu. \\ \int_0^\infty (\ln x - \mu)^2 \, p_{\mu,\sigma}(x)\,dx &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y^2}{2\sigma^2}} y^2\,dy = \sigma^2.\end{aligned}$$

Using

$$\ln p_{\mu,\sigma} = -\ln(\sqrt{2\pi}\sigma) - \ln x - (\ln x - \mu)^2 \frac{1}{2\sigma^2},$$

and the previous integrals, the entropy becomes

$$\begin{aligned}H(\mu,\sigma) &= -\int_0^\infty p_{\mu,\sigma}(x)\,\ln p_{\mu,\sigma}(x)\,dx \\ &= \ln(\sqrt{2\pi}\sigma) + \int_0^\infty \ln x\, p_{\mu,\sigma}(x)\,dx \\ &\quad + \frac{1}{2\sigma^2}\int_0^\infty (\ln x - \mu)^2 p_{\mu,\sigma}(x)\,dx \\ &= \ln(\sqrt{2\pi}) + \ln\sigma + \mu + \frac{1}{2}.\end{aligned}$$

Example 3.2.7 (Dirac Distribution) A Dirac distribution on (a,b) centered at $x_0 \in (a,b)$ represents the density of an idealized point mass x_0. This can be thought of as an infinitely high, infinitely thin spike at x_0, with total area under the spike equal to 1. The Dirac distribution centered at x_0 is customarily denoted by $p(x) = \delta(x - x_0)$, and its relation with the integral can be written informally as

(i) $\displaystyle\int_a^b p(x)\,dx = \int_a^b \delta(x - x_0)\,dx = 1;$

(ii) $\displaystyle\int_a^b g(x)p(x)\,dx = \int_a^b g(x)\delta(x - x_0)\,dx = g(x_0),$

for any continuous function $g(x)$ on (a,b).

The k-th moment is given by

$$m_k = \int_a^b x^k \delta(x - x_0)\, dx = x_0^k.$$

Then the mean of the Dirac distribution is $\mu = x_0$ and the variance is $Var = m_2 - (m_1)^2 = 0$. The underlying random variable, which is Dirac distributed, is a constant equal to x_0.

In order to compute the entropy of $\delta(x - x_0)$, we shall approximate the distribution by a sequence of distributions $\varphi_\epsilon(x)$ for which we can easily compute the entropy. For any $\epsilon > 0$, consider the distribution

$$\varphi_\epsilon(x) = \begin{cases} \dfrac{1}{\epsilon}, & \text{if } |x| < \epsilon/2 \\ \\ 0, & \text{otherwise,} \end{cases}$$

with the entropy given by

$$\begin{aligned} H_\epsilon &= -\int_a^b \varphi_\epsilon(x) \ln \varphi_\epsilon(x)\, dx \\ &= -\int_{x_0-\epsilon/2}^{x_2+\epsilon/2} \frac{1}{\epsilon} \ln \frac{1}{\epsilon}\, dx \\ &= \ln \epsilon. \end{aligned}$$

Since $\lim\limits_{\epsilon \searrow 0} \varphi_\epsilon = \delta(x - x_0)$, by the Dominated Convergence Theorem the entropy of $\delta(x - x_0)$ is given by the limit

$$H = \lim_{\epsilon \searrow 0} H_\epsilon = \lim_{\epsilon \searrow 0} \ln \epsilon = -\infty.$$

In conclusion, the Dirac distribution has the lowest possible entropy. Heuristically, this is because of the lack of disorganization of the associated random variable, which is a constant.

3.3 Entropy on Products of Statistical Models

Consider the statistical manifolds $\mathcal{S}$ and $\mathcal{U}$ and let $\mathcal{S} \times \mathcal{U}$ be their product model, see Example 1.3.9. Any density function $f \in \mathcal{S} \times \mathcal{U}$, with $f(x,y) = p(x)q(y)$, $p \in \mathcal{S}$, $q \in \mathcal{U}$, has the entropy

$$\begin{aligned} H_{\mathcal{S}\times\mathcal{U}}(f) &= -\iint_{\mathcal{X}\times\mathcal{Y}} f(x,y)\ln f(x,y)\,dxdy \\ &\quad -\iint_{\mathcal{X}\times\mathcal{Y}} p(x)q(y)[\ln p(x) + \ln q(y)]\,dxdy \\ &= -\int_{\mathcal{Y}} q(y)\,dy \int_{\mathcal{X}} p(x)\ln p(x)\,dx \\ &\quad -\int_{\mathcal{X}} p(x)\,dx \int_{\mathcal{Y}} q(y)\ln q(y)\,dy \\ &= H_{\mathcal{S}}(p) + H_{\mathcal{U}}(q), \end{aligned}$$

i.e., the entropy of an element of the product model $\mathcal{S}\times\mathcal{U}$ is the sum of the entropies of the projections on $\mathcal{S}$ and $\mathcal{U}$. This can be also stated by saying that the joint entropy of two independent random variables X and Y is the sum of individual entropies, i.e.

$$H(X,Y) + H(X) + H(Y),$$

see Problem 3.5 for details.

3.4 Concavity of Entropy

Theorem 3.4.1 *For any two densities $p, q : \mathcal{X} \to \mathbb{R}$ we have*

$$H(\alpha p + \beta q) \geq \alpha H(p) + \beta H(q), \tag{3.4.9}$$

$\forall \alpha, \beta \in [0,1]$, *with* $\alpha + \beta = 1$.

Proof: Using that $f(u) = -u\ln u$ is concave on $(0,\infty)$, we obtain

$$f(\alpha p + \beta q) \geq \alpha f(p) + \beta f(q).$$

Integrating (summing) over $\mathcal{X}$ leads to expression (3.4.9). ■

With a similar proof we can obtain the following result.

Corollary 3.4.2 *For any densities $p_1, \dots, p_n$ on $\mathcal{X}$ and $\lambda_i \in [0,1]$ with $\lambda_1 + \cdots + \lambda_n = 1$, we have*

$$H\Big(\sum_{i=1}^n \lambda_i p_i\Big) \geq \sum_{i=1}^n \lambda_i H(p_i).$$

The previous result suggests to look for the maxima of the entropy function on a statistical model.

3.5 Maxima for Entropy

Let $\mathcal{S} = \{p_\xi(x); x \in \mathcal{X}, \xi \in \mathbb{E}\}$ be a statistical model. We can regard the entropy H as a function defined on the parameters space $\mathbb{E}$. We are interested in the value of the parameter ξ for which the entropy $H(\xi)$ has a local maximum. This parameter value corresponds to a distinguished density p_ξ. Sometimes, the density p_ξ satisfies some given constraints, which are provided by the given observations, and has a maximum degree of ignorance with respect to the unknown observations. This type of optimization problem is solved by considering the maximization of the entropy with constraints. In order to study this problem we shall start with the definition and characterization of critical points of entropy.

Let f be a function defined on the statistical manifold $S = \{p_\xi\}$. If $\partial_i = \partial_{\xi^i}$ denotes the tangent vector field on S in the direction of ξ^i, then

$$\partial_i f =: \partial_{\xi^i} f := \partial_{\xi^i}(f \circ p_\xi).$$

In the following the role of the function f is played by the entropy $H(\xi) = H(p_\xi)$.

Definition 3.5.1 *A point $q \in S$ is a critical point for the entropy H if*

$$X(H) = 0, \quad \forall X \in T_q S.$$

Since $\{\partial_i\}_i$ form a basis, choosing $X = \partial_i$, we obtain that the point $q = p_\xi \in S$ is a critical point for H if and only if

$$\partial_i H(\xi) = 0, \qquad i = 1, 2, \dots, n.$$

A computation provides

$$\begin{aligned}
\partial_i H &= -\partial_i \int_{\mathcal{X}} p(x,\xi) \ln p(x,\xi)\, dx \\
&= -\int_{\mathcal{X}} \Big(\partial_i p(x,\xi)\, \ln p(x,\xi) + p(x,\xi) \frac{\partial_i p(x,\xi)}{p(x,\xi)}\Big)\, dx \\
&= -\int_{\mathcal{X}} \big(\ln p(x,\xi) + 1\big) \partial_i p(x,\xi)\, dx \\
&= -\int_{\mathcal{X}} \ln p(x,\xi)\, \partial_i p(x,\xi)\, dx,
\end{aligned}$$

where we used that

$$\int_{\mathcal{X}} p(x,\xi)\, dx = 1$$

and

$$0 = \partial_i \int_{\mathcal{X}} p(x,\xi)\,dx = \int_{\mathcal{X}} \partial_i p(x,\xi)\,dx.$$

The previous computation can be summarized as in the following.

Proposition 3.5.2 *The probability distribution p_ξ is a critical point of the entropy H if and only if*

$$\int_{\mathcal{X}} \ln p(x,\xi)\,\partial_{\xi^i} p(x,\xi)\,dx = 0, \quad \forall i = 1,\ldots,m. \tag{3.5.10}$$

In the discrete case, when $\mathcal{X} = \{x^1,\ldots,x^n\}$, the Eq. (3.5.10) is replaced by the relation

$$\sum_{k=1}^{n} \ln p(x^k,\xi)\,\partial_i p(x^k,\xi) = 0, \quad \forall i = 1,\ldots,m. \tag{3.5.11}$$

Observe that the critical points characterized by the previous result do not belong to the boundary. The entropy, which is a concave function, on a convex set (such as a mixture family) sometimes attains the local minima along the boundary. Even if these points are called critical by some authors, here we do not consider them as part of our analysis.

The first derivative of the entropy can be also expressed in terms of the log-likelihood function as in the following

$$\begin{aligned}
\partial_i H &= -\int_{\mathcal{X}} \ln p(x,\xi)\,\partial_{\xi^i} p(x,\xi)\,dx \\
&= -\int_{\mathcal{X}} p(x,\xi)\ln p(x,\xi)\,\partial_i \ln p(x,\xi)\,dx \\
&= -\int_{\mathcal{X}} p(x,\xi)\ell(\xi)\,\partial_i \ell(\xi)\,dx \\
&= -E_\xi[\ell(\xi)\,\partial_{\xi^i}\ell(\xi)].
\end{aligned} \tag{3.5.12}$$

The goal of this section is to characterize the distributions p_ξ for which the entropy is maximum. Minima and maxima are among the set of critical points, see Definition 3.5.1. In order to deal with this issue we need to compute the Hessian of the entropy H.

The second order partial derivatives of the entropy H are

$$\begin{aligned}\partial_{ji}H &= \partial_j \int_{\mathcal{X}} \ln p(x,\xi)\, \partial_i p(x,\xi)\, dx \\ &= -\int_{\mathcal{X}} \Big(\frac{\partial_j p(x)}{p(x)} \partial_i p(x) + \ln p(x)\ \partial_i \partial_j p(x)\Big)\, dx \\ &= -\int_{\mathcal{X}} \Big(\frac{1}{p(x)} \partial_i p(x)\ \partial_j p(x) + \ln p(x)\ \partial_{ji} p(x)\Big)\, dx.\end{aligned}$$

In the discrete case this becomes

$$\partial_{ji}H = -\sum_{k=1}^{n} \Big(\frac{\partial_i p(x^k,\xi)\ \partial_j p(x^k,\xi)}{p(x^k,\xi)} + \ln p(x_k,\xi)\ \partial_{ij} p(x^k,\xi)\Big). \tag{3.5.13}$$

We can also express the Hessian of the entropy in terms of the log-likelihood function only. Differentiating in (3.5.12) we have

$$\begin{aligned}\partial_{ji}H &= -\partial_j \int_{\mathcal{X}} p(x,\xi)\ell(\xi)\, \partial_i \ell(\xi)\, dx \\ &= -\int_{\mathcal{X}} \Big(\partial_j p(x,\xi)\ell(\xi)\, \partial_i \ell(\xi) + p(x,\xi)\partial_j \ell(\xi)\, \partial_i \ell(\xi) \\ &\quad + p(x,\xi)\ell(\xi)\, \partial_i \partial_j \ell(\xi)\Big)\, dx \\ &= -E_\xi[\partial_i \ell\, \partial_j \ell] - E_\xi[(\partial_j \ell(\xi)\partial_i \ell(\xi) + \partial_i \partial_j \ell(\xi))\ell(\xi)] \\ &= -g_{ij}(\xi) - h_{ij}(\xi).\end{aligned}$$

We arrived at the following result that relates the entropy and the Fisher information.

Proposition 3.5.3 *The Hessian of the entropy is given by*

$$\partial_i \partial_j H(\xi) = -g_{ij}(\xi) - h_{ij}(\xi), \tag{3.5.14}$$

where $g_{ij}(\xi)$ is the Fisher–Riemann metric and

$$h_{ij}(\xi) = E_\xi[(\partial_j \ell(\xi)\partial_i \ell(\xi) + \partial_i \partial_j \ell(\xi))\ell(\xi)].$$

Corollary 3.5.4 *In the case of the mixture family (1.5.15)*

$$p(x;\xi) = C(x) + \xi^i F_i(x) \tag{3.5.15}$$

the Fisher–Riemann metric is given by

$$g_{ij}(\xi) = -\partial_i \partial_j H(\xi). \tag{3.5.16}$$

Furthermore, any critical point of the entropy (see Definition 3.5.1) is a maximum point.

Proof: From Proposition 1.5.1, part (*iii*) we have $\partial_i \partial_j \ell_x(\xi) = -\partial_i \ell_x(\xi)\, \partial_j \ell_x(\xi)$ which implies $h_{ij}(\xi) = 0$. Substituting in (3.5.14) yields (3.5.16). Using that the Fisher–Riemann matrix $g_{ij}(\xi)$ is positive definite at any ξ, it follows that $\partial_i \partial_j H(\xi)$ is globally negative definite, and hence all critical points must be maxima. We also note that we can express the Hessian in terms of F_j as in the following

$$\partial_i \partial_j H(\xi) = -\int_{\mathcal{X}} \frac{F_i(x) F_j(x)}{p(x;\xi)}\, dx.$$

■

A Hessian $Hess(F) = (\partial_{ij} F)$ is called positive definite if and only if $\sum_{i,j} \partial_{ij} F\, v^i v^j > 0$, or, equivalently,

$$\langle Hess(F) v, v\rangle > 0, \qquad \forall v \in \mathbb{R}^m.$$

In the following we shall deal with the relationship between the Hessian and the second variation of the entropy H.

Consider a curve $\xi(s)$ in the parameter space and let $\big(\xi_u(s)\big)_{|u|<\epsilon}$ be a smooth variation of the curve with $\xi_u(s)_{|u=0} = \xi(s)$. Then $s \to p_{\xi_u(s)}$ is a variation of the curve $s \to p_{_{\xi(s)}}$ on the statistical manifold S. Consider the variation

$$\xi_u(s) = \xi(s) + u\eta(s),$$

so $\partial_u \xi_u(s) = \eta(s)$ and $\partial_u^2 \xi_u(s) = 0$. The second variation of the entropy along the curve $s \to p_{\xi_u(s)}$ is

$$\begin{aligned}
\frac{d^2}{du^2} H\big(\xi_u(s)\big) &= \frac{d}{du}\langle \partial_\xi H, \partial_u \xi_u(s)\rangle \\
&= \langle \frac{d}{du}\partial_\xi H, \partial_u \xi(s)\rangle + \langle \partial_\xi H, \underbrace{\partial_u^2 \xi_u(s)}_{=0}\rangle \\
&= \frac{d}{du}(\partial_i H)\, \partial_u \xi^i(s) \\
&= \partial_i \partial_j H(\xi_u(s)) \cdot \partial_u \xi_u^i(s) \partial_u \xi_u^j(s).
\end{aligned}$$

Taking $u = 0$, we find

$$\begin{aligned}
\frac{d^2}{du^2} H\big(\xi_u(s)\big)_{|u=0} &= \partial_{ij} H\big(\xi(s)\big)\eta^i(s)\eta^j(s) \\
&= \langle Hess\, H\big(\xi(s)\big)\eta, \eta\rangle.
\end{aligned}$$

Hence $\frac{d^2}{du^2}H\big(\xi_u(s)\big)_{|u=0} < 0(> 0)$ if and only if $Hess(H)$ is negative (positive) definite. Summarizing, we have:

Theorem 3.5.5 *If ξ is such that p_ξ satisfies the critical point condition (3.5.10) (or condition (3.5.11) in the discrete case), and the Hessian $Hess(H(\xi))$ is negative definite at ξ, then p_ξ is a local maximum point for the entropy.*

We shall use this result in the next section.

Corollary 3.5.6 *Let ξ_0 be such that*

$$E_{\xi_0}[\ell(\xi_0)\partial_i\ell(\xi_0)] = 0 \tag{3.5.17}$$

and $h_{ij}(\xi_0)$ is positive definite. Then $p(x, \xi_0)$ is a distribution for which the entropy reaches a local maximum.

Proof: In the virtue of (3.5.12) the Eq. (3.5.17) is equivalent with the critical point condition $\partial_i H(\xi)_{|\xi=\xi_0} = 0$. Since $g_{ij}(\xi_0)$ is positive definite, then (3.5.14) implies that $\partial_i\partial_j H(\xi_0)$ is negative definite. Then applying Theorem 3.5.5 ends the proof. ■

3.6 Weighted Coin

Generally, for discrete distributions we may identify the statistical space $\mathcal{S}$ with the parameter space $\mathbb{E}$. We shall present next the case of a simple example where the entropy can be maximized. Flipping a weighted coin provides either heads with probability ξ^1, or tails with probability $\xi^2 = 1 - \xi^1$. The statistical manifold obtained this way depends on only one essential parameter $\xi := \xi^1$. Since $\mathcal{X} = \{x_1 = heads, x_2 = tails\}$, the manifold is just a curve in $\mathbb{R}^2$ parameterized by $\xi \in [0, 1]$. The probability distribution of the weighted coin is given by the table

outcomes	x_1	x_2
probability	ξ	$1-\xi$

We shall find the points of maximum entropy. First we write the Eq. (3.5.11) to determine the critical points

$$\begin{aligned}
\ln p(x_1, \xi)\,\partial_\xi p(x_1, \xi) + \ln p(x_2, \xi)\,\partial_\xi p(x_2, \xi) &= 0 \Longleftrightarrow \\
\ln \xi - \ln(1 - \xi) &= 0 \Longleftrightarrow \\
\xi &= 1 - \xi
\end{aligned}$$

and hence there is only one critical point, $\xi = \frac{1}{2}$.

The Hessian has only one component, so formula (3.5.13) yields

$$\begin{aligned}
\partial_\xi^2 H &= -\Big(\frac{1}{p(x_1)}\big(\partial_\xi p(x_1)\big)^2 + \ln p(x_1)\partial_\xi^2 p(x_1)\Big) \\
&\quad -\Big(\frac{1}{p(x_2)}\big(\partial_\xi p(x_2)\big)^2 + \ln p(x_2)\partial_\xi^2 p(x_2)\Big) \\
&= -\Big(\frac{1}{\xi}\cdot 1 + \ln\xi \cdot 0\Big) \\
&\quad -\Big(\frac{1}{1-\xi}\big(\partial_\xi(1-\xi)\big)^2 + \ln(1-\xi)\ \partial_\xi^2(1-\xi)\Big) \\
&= -\Big(\frac{1}{\xi} + \frac{1}{1-\xi}\Big).
\end{aligned}$$

Evaluating at the critical point, we get

$$\partial_\xi^2 H_{|\xi=\frac{1}{2}} = -4 < 0,$$

and hence $\xi = \frac{1}{2}$ is a maximum point for the entropy. In this case $\xi^1 = \xi^2 = \frac{1}{2}$. This can be restated by saying that *the fair coin has the highest entropy among all weighted coins.*

3.7 Entropy for Finite Sample Space

Again, we underline that for discrete distributions we identify the statistical space $\mathcal{S}$ with the parameter space $\mathbb{E}$.

Consider a statistical model with a finite discrete sample space $\mathcal{X} = \{x^1, \dots, x^{n+1}\}$ and associated probabilities $p(x^i) = \xi^i$, $\xi^i \in [0,1]$, $i = 1, \dots n+1$. Since $\xi^{n+1} = 1 - \sum_{i=1}^n \xi^i$, the statistical manifold is described by n essential parameters, and hence it has n dimensions. The manifold can be also seen as a hypersurface in $\mathbb{R}^{n+1}$. The entropy function is

$$H = -\sum_{i=1}^{n+1} \xi^i \ln \xi^i. \tag{3.7.18}$$

The following result deals with the maximum entropy condition. Even if it can be derived from the concavity property of H, see Theorem 3.4.1, we prefer to deduct it here in a direct way. We note that concavity is used as a tool to derive the case of continuous distributions, see Corollary 5.9.3.

Theorem 3.7.1 *The entropy (3.7.18) is maximum if and only if*

$$\xi^1 = \cdots = \xi^{n+1} = \frac{1}{n+1}. \tag{3.7.19}$$

Proof: The critical point condition (3.5.11) becomes

$$\begin{aligned}
\sum_{k=1}^{n} \ln p(x^k,\xi)\partial_{\xi^i} p(x^k,\xi) + \ln p(x^{n+1},\xi)\ \partial_{\xi^i} p(x^{n+1},\xi) &= 0 \iff \\
\sum_{k=1}^{n} \ln \xi^k\ \delta_{ik} + \ln \xi^{n+1}\ \partial_{\xi^{n+1}}(1-\xi^1-\cdots-\xi^n) &= 0 \iff \\
\ln \xi^i - \ln \xi^{n+1} &= 0 \iff \\
\xi^i &= \xi^{n+1},
\end{aligned}$$

$\forall i = 1,\dots,n$. Hence condition (3.7.19) follows.

We shall investigate the Hessian at this critical point. Following formula (3.5.13) yields

$$\begin{aligned}
Hess(H)_{ij} &= -\sum_{k=1}^{n} \frac{\partial_i(\xi^k)\cdot\partial_j(\xi^k)}{\xi^k} - \frac{\partial_i(\xi^{n+1})\cdot\partial_j(\xi^{n+1})}{\xi^{n+1}} \\
&\quad -\sum_{k=1}^{n} \ln \xi^k\ \partial_i\partial_j(\xi^k) - \ln \xi^{n+1}\ \partial_i\partial_j(\xi^{n+1}) \\
&= -\Big(\sum_{k=1}^{n} \frac{\delta_{ik}\delta_{jk}}{\xi^k} - \frac{1}{\xi^{n+1}}\Big),
\end{aligned}$$

where we have used $\partial_i(\xi^{n+1}) = \partial_i(1-\xi^1-\cdots-\xi^n) = -1$, for $i = 1,\dots,n$.

At the critical point the Hessian is equal to

$$Hess(H)_{ij|_{\xi_k=\frac{1}{n+1}}} = -(n+1)\Big(1 + \sum_{k=1}^{n} \delta_{ik}\delta_{jk}\Big) = -2(n+1)I_n,$$

which shows that it is negative definite. Theorem 3.5.5 leads to the desired conclusion. ■

Example 3.7.2 Let ξ^i be the probability that a die lands with the face i up. This model depends on five essential parameters. According to the previous result, the fair die is the one which maximizes the entropy.

3.8 A Continuous Distribution Example

Let $p(x;\xi) = 2\xi x + 3(1-\xi)x^2$ be a continuous probability distribution function, with $x \in [0,1]$. The statistical manifold defined by the above probability distribution is one dimensional, since $\xi \in \mathbb{R}$. There is only one basic vector field equal to

$$\partial_\xi = 2x - 3x^2,$$

and which does not depend on ξ. In order to find the critical points, we follow Eq. (3.5.10)

$$\begin{aligned} \int_0^1 p(x,\xi)\,\partial_\xi p(x,\xi)\,dx &= 0 \Longleftrightarrow \\ \int_0^1 (2x-3x^2)(2\xi x + 3(1-\xi)x^2)\,dx &= 0 \Longleftrightarrow \\ \frac{2}{15}\xi - \frac{3}{10} &= 0 \Longleftrightarrow \xi = \frac{9}{4}. \end{aligned}$$

Before investigating the Hessian, we note that

$$\partial_\xi p(x;\xi) = 2x - 3x^2, \quad \partial_\xi^2 p(x;\xi) = 0, \quad p\Big(x;\frac{9}{4}\Big) = \frac{9}{4}x - \frac{15}{4}x^2,$$

so

$$\begin{aligned} \partial_\xi^2 H_{|\xi=\frac{9}{4}} &= -\int_0^1 \Big(\frac{1}{p}(\partial_\xi p)^2 + \ln p\; \partial_\xi^2 p\Big)\,dx\Big|_{\xi=\frac{9}{4}} \\ &= -\int_0^1 \frac{(2x-3x^2)^2}{\frac{9}{2}x - \frac{15}{4}x^2}\,dx < 0, \end{aligned}$$

because $\frac{9}{2}x - \frac{15}{4}x^2 < 0$ for $x \in (0,1]$.

Hence $\xi = \frac{9}{4}$ is a maximum point for the entropy. The maximum value of the entropy is

$$\begin{aligned} H\Big(\frac{9}{4}\Big) &= -\int_0^1 \big(\frac{9}{2}x - \frac{15}{4}x^2\big)\ln\big(\frac{9}{2}x - \frac{15}{4}x^2\big)\,dx \\ &= -\frac{52}{25}\ln 3 + \frac{47}{30} + \frac{23}{25}\ln 2 \\ &= -0.807514878. \end{aligned}$$

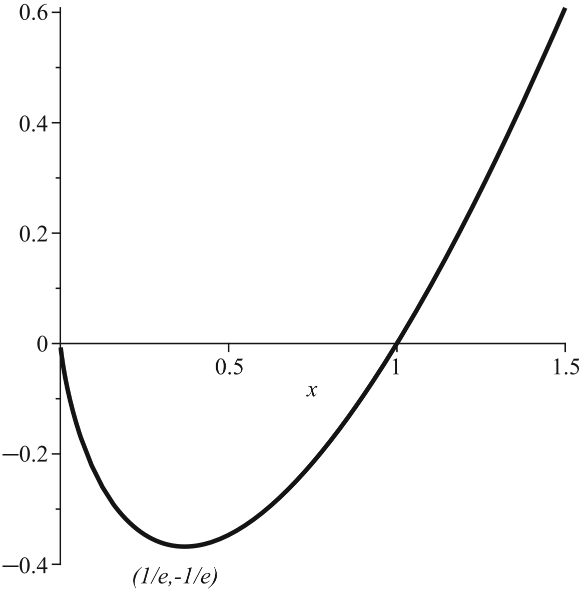

Figure 3.1: The function $x \to x \ln x$ has a global minimum value equal to $-1/e$ that is reached at $x = 1/e$

3.9 Upper Bounds for Entropy

We shall start with computing a rough upper bound for the entropy in the case when the sample space is a finite interval, $\mathcal{X} = [a, b]$. Consider the convex function

$$f : [0,1] \to \mathbb{R}, \qquad f(u) = \begin{cases} u \ln u & if \quad u \in (0,1] \\ 0 & if \quad u = 0. \end{cases}$$

Since $f'(u) = 1 + \ln u$, $u \in (0, 1)$, the function has a global minimum at $u = 1/e$, and hence $u \ln u \geq -1/e$, see Fig. 3.1.

Let $p : \mathcal{X} \to \mathbb{R}$ be a probability density. Substituting $u = p(x)$ yields $p(x) \ln p(x) \geq -1/e$. Integrating, we find

$$\int_a^b p(x) \ln p(x)\, dx \geq -\frac{b-a}{e}.$$

Using the definition of the entropy we obtain the following upper bound.

Proposition 3.9.1 *The entropy $H(p)$ of a probability distribution $p : [a, b] \to [0, \infty)$ satisfies the inequality*

$$H(p) \leq \frac{b-a}{e}. \tag{3.9.20}$$

Corollary 3.9.2 *The entropy $H(p)$ is smaller than half the length of the domain interval of the distribution p, i.e.,*

$$H(p) \leq \frac{b-a}{2}.$$

This implies that the entropy $H(p)$ is smaller than the mean of the uniform distribution.

We note that the inequality (3.9.20) becomes identity for the uniform distribution $p : [0, e] \to [0, \infty)$, $p(x) = 1/e$, see Problem 3.20. We shall present next another upper bound which is reached for all uniform distributions.

Theorem 3.9.3 *The entropy of a smooth probability distribution $p : [a, b] \to [0, \infty)$ satisfies the inequality*

$$H(p) \leq \ln(b-a). \tag{3.9.21}$$

Proof: Since the function

$$f : [0,1] \to \mathbb{R}, \qquad f(u) = \begin{cases} u \ln u & if \quad u \in (0,1] \\ 0 & if \quad u = 0 \end{cases}$$

is convex on $[0, \infty)$, an application of Jensen integral inequality yields

$$\begin{aligned} f\Big(\frac{1}{b-a}\int_a^b p(x)\,dx\Big) &\leq \frac{1}{b-a}\int_a^b f\big(p(x)\big)\,dx \Longleftrightarrow \\ f\Big(\frac{1}{b-a}\Big) &\leq \frac{1}{b-a}\int_a^b p(x)\ln p(x)\,dx \Longleftrightarrow \\ \ln\Big(\frac{1}{b-a}\Big) &\leq \int_a^b p(x)\ln p(x)\,dx \Longleftrightarrow \\ -\ln(b-a) &\leq -H(p), \end{aligned}$$

which is equivalent to (3.9.21). The identity is reached for the uniform distribution $p(x) = 1/(b-a)$. ∎

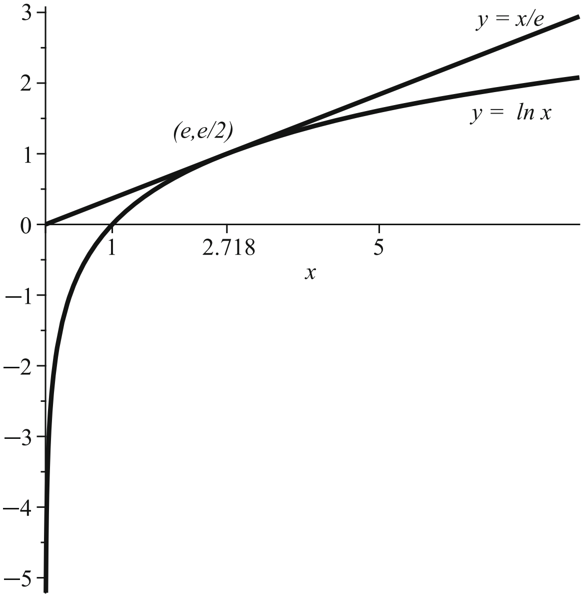

Figure 3.2: The inequality $\ln x \leq x/e$ is reached for $x = e$

The above result states that the maximum entropy is realized only for the case of the uniform distribution. In other words, the entropy measures the closeness of a distribution to the uniform distribution.

Since we have the inequality

$$\ln x \leq \frac{x}{e}, \qquad \forall x > 0$$

with equality only for $x = e$, see Fig. 3.2, it follows that the inequality (3.9.21) provides a better bound than (3.9.20).

In the following we shall present the bounds of the entropy in terms of the maxima and minima of the probability distribution. We shall use the following inequality involving the weighted average of n numbers.

Lemma 3.9.4 *If* $\lambda_1, \dots, \lambda_n > 0$ *and* $\alpha_1, \dots, \alpha_n \in \mathbb{R}$*, then*

$$\min_j \{\alpha_j\} \leq \frac{\sum_i \lambda_i \alpha_i}{\sum_i \lambda_i} \leq \max_j \{\alpha_j\}.$$

This says that if α_j are the coordinates of n points of masses λ_j, then the coordinate of the center of mass of the system is larger than the smallest coordinate and smaller than the largest coordinate.

Proposition 3.9.5 *Consider the discrete probability distribution* $p = \{p_j\}$*, with* $p_1 \le \cdots \le p_n$*. Then the entropy satisfies the double inequality*

$$-\ln p_n \le H(p) \le -\ln p_1.$$

Proof: Letting $\lambda_j = p_j$ and $\alpha_j = -\ln p_j$ in Lemma 3.9.4 and using

$$H(p) = -\sum_j p_j \ln p_j = \frac{\sum_i \lambda_i \alpha_i}{\sum_i \lambda_i},$$

we find the desired inequality. ■

Remark 3.9.6 The distribution $p = \{p_j\}$ is uniform with $p_j = \dfrac{1}{n}$ if and only if $p_1 = p_n$. In this case the entropy is given by

$$H(p) = -\ln p_1 = \ln p_n = -\ln\frac{1}{n} = \ln n.$$

The continuous analog of Proposition 3.9.5 is given below.

Proposition 3.9.7 *Consider the continuous probability distribution* $p : \mathcal{X} \to [a, b] \subset [0, \infty)$*, with* $p_m = \min_{x\in\mathcal{X}} p(x)$ *and* $p_{_M} = \max_{x\in\mathcal{X}} p(x)$*. Then the entropy satisfies the inequality*

$$-\ln p_{_M} \le H(p) \le -\ln p_m.$$

Proof: The proof is using the following continuous analog of Lemma 3.9.4,

$$\min_{x\in\mathcal{X}} \alpha(x) \le \frac{\int_{\mathcal{X}} \lambda(x)\alpha(x)\, dx}{\int_{\mathcal{X}} \lambda(x)\, dx} \le \max_{x\in\mathcal{X}} \alpha(x),$$

where we choose $\alpha(x) = -\ln p(x)$ and $\lambda(x) = p(x)$. ■

3.10 Boltzman–Gibbs Submanifolds

Let

$$\mathcal{S} = \{p_\xi : [0,1] \longrightarrow \mathbb{R}_+;\ \int_{\mathcal{X}} p_{_\xi}(x)\, dx = 1\}, \quad \xi \in \mathbb{E},$$

be a statistical model with the state space $\mathcal{X} = [0, 1]$. Let $\mu \in \mathbb{R}$ be a fixed constant and consider the set of elements of $\mathcal{S}$ with the mean μ

$$\mathcal{M}_\mu = \{p_\xi \in \mathcal{S};\ \int_{\mathcal{X}} x p_{_\xi}(x)\, dx = \mu\}.$$

and assume that $\mathcal{M}_\mu$ is a submanifold of $\mathcal{S}$.

Definition 3.10.1 *The statistical submanifold $\mathcal{M}_\mu = \{p_\xi\}$ defined above is called a Boltzman–Gibbs submanifold of $\mathcal{S}$.*

Example 3.10.1 In the case of beta distribution, the Boltzman–Gibbs submanifold $\mathcal{M}_\mu = \{p_{a,ka}; a > 0, k = (1-\mu)/\mu\}$ is just a curve. In particular, $\mathcal{M}_1 = \{p_{a,0}; a > 0\}$, with $p_{a,0}(x) = \frac{1}{B(a,0)} x^{a-1}(1-x)^{-1}$.

One of the problems arised here is to find the distribution of maximum entropy on a Boltzman–Gibbs submanifold. Since the maxima are among critical points, which are introduced by Definition 3.5.1, we shall start the study with finding the critical points of the entropy

$$H(\xi) = H(p_\xi) = -\int_{\mathcal{X}} p_{_\xi}(x)\,\ln p_{_\xi}(x)\,dx$$

on a Boltzman–Gibbs submanifold $\mathcal{M}_\mu$. Differentiating with respect to ξ^j in relations

$$\int_{\mathcal{X}} xp_{_\xi}(x)\,dx = \mu, \quad \int_{\mathcal{X}} p_{_\xi}(x)\,dx = 1 \tag{3.10.22}$$

yields

$$\int_{\mathcal{X}} x\,\partial_j p(x,\xi)\,dx = 0, \qquad \int_{\mathcal{X}} \partial_j p(x,\xi)\,dx = 0. \tag{3.10.23}$$

A computation provides

$$\begin{aligned} -\partial_j H(\xi) &= \partial_j \int_{\mathcal{X}} p_{_\xi}(x) \ln p_{_\xi}(x)\,dx \\ &= \int_{\mathcal{X}} \Big(\partial_j p_{_\xi}(x)\ \ln p_{_\xi}(x) + p_{_\xi}(x)\frac{\partial_j p_{_\xi}(x)}{p_{_\xi}(x)}\Big)\,dx \\ &= \int_{\mathcal{X}} \partial_j p(x)\ \ln p_{_\xi}(x)\,dx + \underbrace{\int_{\mathcal{X}} \partial_j p_{_\xi}(x)\,dx}_{=0 \text{ by } (3.10.23)}. \end{aligned}$$

Hence the critical points $p_{_\xi}$ satisfying $\partial_j H(\xi) = 0$ are solutions of the integral equation

$$\int \partial_j p(x,\xi)\ \ln p(x,\xi)\,dx = 0, \tag{3.10.24}$$

subject to the constraint

$$\int_{\mathcal{X}} x\partial_j p(x,\xi)\,dx = 0. \tag{3.10.25}$$

Multiplying (3.10.25) by the Lagrange multiplier $\lambda = \lambda(\xi)$ and adding it to (3.10.24) yields

$$\int_{\mathcal{X}} \partial_j p(x,\xi)\Big(\ln p(x,\xi) + \lambda(\xi)x\Big)\,dx = 0.$$

Since $\displaystyle\int \partial_j p(x,\xi)\,dx = 0$, it makes sense to consider those critical points for which the term $\ln p(x,\xi) + \lambda(\xi)x$ is a constant function in x, i.e., depends only on ξ

$$\ln p(x,\xi) + \lambda(\xi)x = \theta(\xi).$$

Then the above equation has the solution

$$p(x,\xi) = e^{\theta(\xi)-\lambda(\xi)x}, \qquad (3.10.26)$$

which is an exponential family. We still need to determine the functions θ and λ such that the constraints (3.10.22) hold. This will be done explicitly for the case when the sample space is $\mathcal{X} = [0,1]$. From the second constraint we obtain a relation between θ and λ:

$$\int_0^1 p(x,\xi)\,dx{=}1 \Longrightarrow e^{\theta(\xi)}\int_0^1 e^{-\lambda(\xi)x}\,dx = 1 \Longleftrightarrow \frac{1-e^{-\lambda(\xi)}}{\lambda(\xi)} = e^{-\theta(\xi)},$$

which leads to

$$\theta(\xi) = \ln\frac{\lambda(\xi)}{1-e^{-\lambda(\xi)}}.$$

Substituting in (3.10.26) yields

$$p(x,\xi) = \frac{\lambda(\xi)}{1-e^{-\lambda(\xi)}}e^{-\lambda(\xi)x}. \qquad (3.10.27)$$

Substituting in the constraint

$$\int_0^1 xp(x,\xi)\,dx = \mu,$$

we find

$$\begin{aligned}
\frac{\lambda(\xi)}{1-e^{-\lambda(\xi)}}\int_0^1 xe^{-\lambda(\xi)x}\,dx &= \mu \Longleftrightarrow\\
\frac{1-\big(1+\lambda(\xi)\big)e^{-\lambda(\xi)}}{\lambda(\xi)(1-e^{-\lambda(\xi)})} &= \mu \Longleftrightarrow\\
\frac{e^{\lambda(\xi)}-\lambda(\xi)-1}{\lambda(\xi)(e^{\lambda(\xi)}-1)} &= \mu \Longleftrightarrow\\
\frac{1}{\lambda(\xi)}-\frac{1}{e^{\lambda(\xi)}-1} &= \mu.
\end{aligned}$$

Given μ, we need to solve the above equation for $\lambda(\xi)$. In order to complete the computation, we need the following result.

Lemma 3.10.2 *The function*

$$f(x)=\frac{1}{x}-\frac{1}{e^x-1},\ x\in(-\infty,0)\cup(0,\infty),$$

has the following properties

i) $\lim_{x\searrow 0} f(x)=\lim_{x\nearrow 0} f(x)=\frac{1}{2}$,

ii) $\lim_{x\longrightarrow\infty} f(x)=0$, $\lim_{x\longrightarrow-\infty} f(x)=1$,

iii) $f(x)$ *is a strictly decreasing function of* x.

Proof: $i)$ Applying l'Hôspital's rule twice, we get

$$\begin{aligned}
\lim_{x\searrow 0} f(x) &= \lim_{x\searrow 0}\frac{e^x-1-x}{x(e^x-1)}=\lim_{x\searrow 0}\frac{e^x-1}{e^x-1+xe^x}\\
&= \lim_{x\searrow 0}\frac{e^x}{e^x+xe^x+e^x}=\lim_{x\searrow 0}\frac{1}{2+x}=\frac{1}{2}.
\end{aligned}$$

$ii)$ It follows easily from the properties of the exponential function. ■

Since the function f is one-to-one, the equation $f(\lambda)=\mu$ has at most one solution, see Fig. 3.3. More precisely,

- if $\mu\geq 1$, the equation has no solution;

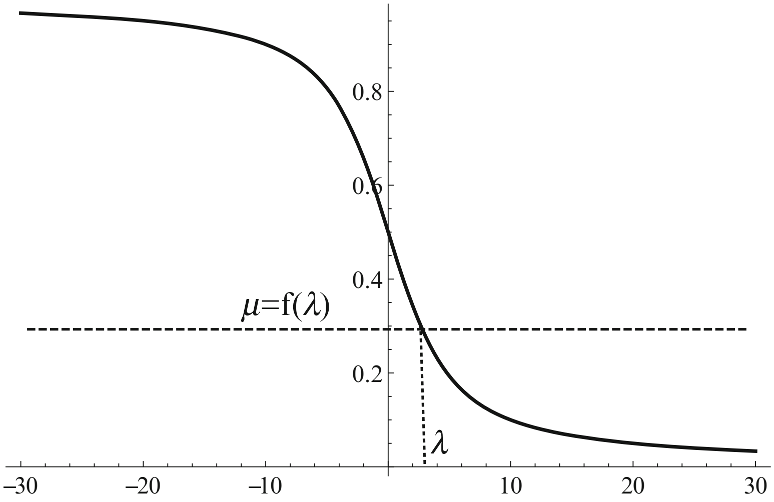

Figure 3.3: The graph of the decreasing function $f(x) = \dfrac{1}{x} - \dfrac{1}{e^x - 1}$ and the solution of the equation $f(\lambda) = \mu$ with $\mu \in (0, 1)$

- if $\mu \in (0, 1)$, the equation has a unique solution, for any ξ, i.e., λ is constant, $\lambda = f^{-1}(\mu)$. For instance, if $\mu = 1/2$, then $\lambda = 0$.

It follows that θ is also constant,

$$\theta = \ln \frac{\lambda}{1 - e^{-\lambda}}.$$

Hence the distribution becomes

$$p(x) = e^{\theta - \lambda x}, \quad x \in (0, 1).$$

3.11 Adiabatic Flows

The entropy $H(\xi)$ is a real function defined on the parameter space $\mathbb{E}$ of the statistical model $\mathcal{S} = \{p_\xi\}$. The critical points of $H(\xi)$ are solutions of the system $\partial_i H(\xi) = 0$. Suppose that the set C of critical points is void. Then the constant level sets $\sum_c := \{H(\xi) = c\}$ are hypersurfaces in $\mathbb{E}$. As usual, we accept the denomination of hypersurface for $\sum_c$ even if $\sum_c \cap C$ consists in a finite number of points.

Let $s \longrightarrow \xi(s)$, $\xi(s) \in \mathbb{E}$, be a curve situated in one of the hypersurfaces $\sum_c$. Since $H(\xi(s)) = c$, it follows

$$\frac{d}{ds} H\big(\xi(s)\big) = \partial_j H\big(\xi(s)\big)\, \dot{\xi}^j(s) = 0. \tag{3.11.28}$$

Since $\dot{\xi}^j(s)$ is an arbitrary vector tangent to $\sum_c$, the vector field $\partial_i H$ is normal to $\sum_c$. Consequently, any vector field $X = (X^i)$ on $\mathbb{E}$ that satisfies

$$\partial_i H(\xi) X^i(\xi) = 0$$

is tangent to $\sum_c$.

Let $X = (X^i)$ be a vector field tangent to $\sum_c$. The flow $\xi(s)$ defined by

$$\dot{\xi}(s) = X^i(\xi(s)),\ \ i = 1, \ldots, n = \dim\ \mathrm{S}$$

is called *adiabatic flow* on $\sum_c$. This means $H(\xi) = c$, since the entropy is unchanged along the flow, i.e., $H(\xi)$ is a first integral, or $\sum_c$ is an invariant set with respect to this flow.

Suppose now that $S = \{p_\xi\}$ refers to a continuous distribution statistical model. Then

$$\begin{aligned} \partial_j H\big(\xi(s)\big) &= \int_{\mathcal{X}} \ln p\big(x, \xi(s)\big)\ \partial_j p\big(x, \xi(s)\big)\, dx \\ &= \int_{\mathcal{X}} \ell_x(\xi(s)) \partial_j \ell_x(\xi(s))\, dx, \end{aligned}$$

and combining with (3.11.28) we arrive at the following result:

Proposition 3.11.1 *The flow $\dot{\xi}^i(s) = X^i(\xi(s))$ is adiabatic if and only if*

$$\int_{\mathcal{X}} \ell_x(\xi(s)) \frac{d}{ds} \ell_x(\xi(s))\, dx = 0.$$

Example 3.11.1 If in the case of the normal distribution the entropy along the curve $s \longrightarrow p_{\sigma(s), \mu(s)}$ is constant, i.e.,

$$H\big(\sigma(s), \mu(s)\big) = \ln\big(\sigma(s)\sqrt{2\pi e}\big) = c$$

then $\sigma(s) = \dfrac{e^c}{\sqrt{2\pi e}}$, constant. Hence the adiabatic flow in this case corresponds to the straight lines

$$\{\sigma = constant,\ \mu(s)\},$$

with $\mu(s)$ arbitrary curve.

For more information regarding flows the reader is referred to Udriste [80, 82, 83].

3.12 Problems

3.1. Use the uncertainty function axioms to show the following relations:

(a) $H\Big(\frac{1}{2},\frac{1}{3},\frac{1}{6}\Big) = H\Big(\frac{1}{2},\frac{1}{2}\Big) + \frac{1}{2}H\Big(\frac{2}{3},\frac{1}{3}\Big).$

(b) $H\Big(\frac{1}{2},\frac{1}{4},\frac{1}{8},\frac{1}{8}\Big) = H\Big(\frac{3}{4},\frac{1}{4}\Big) + \frac{3}{4}H\Big(\frac{2}{3},\frac{1}{3}\Big) + \frac{1}{4}H\Big(\frac{1}{2},\frac{1}{2}\Big).$

(c) $H(p_1,\dots,p_n,0) = H(p_1,\dots,p_n).$

3.2. Consider two events $A = \{a_1,\dots,a_m\}$ and $B = \{b_1,\dots,b_n\}$, and let $p(a_i,b_j)$ be the probability of the joint occurrence of outcomes a_i and b_j. The entropy of the joint event is defined by

$$H(A,B) = -\sum_{i,j} p(a_i,b_j)\log_2 p(a_i,b_j).$$

Prove the inequality

$$H(A,B) \le H(A) + H(B),$$

with identity if and only if the events A and B are independent (i.e., $p(a_i,b_i) = p(a_i)p(b_j)$).

3.3. If $A = \{a_1,\dots,a_m\}$ and $B = \{b_1,\dots,b_n\}$ are two events, define the *conditional entropy* of B given A by

$$H(B|A) = -\sum_{i,j} p(a_i,b_j)\log_2 p_{a_i}(b_j),$$

and the information conveyed about B by A as

$$I(B|A) = H(B) - H(B|A),$$

where $p_{a_i}(b_j) = \dfrac{p(a_i,b_j)}{\sum_j p(a_i,b_j)}$ is the conditional probability of b_j given a_i. Prove the following:

(a) $H(A,B) = H(A) + H(B|A)$;

(b) $H(B) \ge H(B|A)$. When does the equality hold?

(c) $H(B|A) - H(A|B) = H(B) - H(A)$;

(d) $I(B|A) = I(A|B)$.

3.4. Let X be a real-valued continuous random variable on $\mathbb{R}^n$, with density function $p(x)$. Define the entropy of X by

$$H(X) = -\int_{\mathbb{R}^n} p(x)\, \ln p(x)\, dx.$$

(a) Show that the entropy is translation invariant, i.e., $H(X) = H(X+c)$, for any constant $c \in \mathbb{R}$.

(b) Prove the formula $H(aX) = H(X) + \ln|a|$, for any constant $a \in \mathbb{R}$. Show that by rescaling the random variable the entropy can change from negative to positive and vice versa.

(c) Show that in the case of a vector valued random variable $Y : \mathbb{R}^n \to \mathbb{R}^n$ and an $n \times n$ matrix A we have

$$H(AY) = H(Y) + \ln|\det A|.$$

(d) Use (c) to prove that the entropy is invariant under orthogonal transformations of the random variable.

3.5. The joint and conditional entropies of two continuous random variables X and Y are given by

$$H(X,Y) = -\iint p(x,y)\, \log_2 p(x,y)\, dxdy,$$

$$H(Y|X) = -\iint p(x,y)\, \log_2 \frac{p(x,y)}{p(x)}\, dxdy,$$

where $p(x) = \displaystyle\int p(x,y)\, dy$ is the marginal probability of X. Prove the following:

(a) $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$;

(b) $H(Y|X) \le H(Y)$.

3.6. Let $\alpha(x,y)$ be a function with $\alpha(x,y) \ge 0$, $\displaystyle\int_{\mathbb{R}} \alpha(x,y)\, dx = \int_{\mathbb{R}} \alpha(x,y)\, dy = 1$. Consider the averaging operation

$$q(y) = \int_{\mathbb{R}} \alpha(x,y) p(x)\, dx.$$

Prove that the entropy of the averaged distribution $q(y)$ is equal to or greater than the entropy of $p(x)$, i.e., $H(q) \ge H(p)$.

3.7. Consider the two-dimensional statistical model defined by

$$p(x, \xi^1, \xi^2) = 2\xi^1 x + 3\xi^2 x^2 + 4(1 - \xi^1 - \xi^2)x^3, \qquad x \in (0, 1).$$

(a) Compute the Fisher metric $g_{ij}(\xi)$.

(b) Compute the entropy $H(p)$.

(c) Find ξ for which H is critical. Does it correspond to a maximum or to a minimum?

3.8. Find a generic formula for the informational entropy of the exponential family $p(\xi, x) = e^{C(x)+\xi^i F_i(x)-\phi(\xi)}$, $x \in \mathcal{X}$.

3.9. (The change of the entropy under a change of coordinates.) Consider the vector random variables X and Y, related by $Y = \phi(X)$, with $\phi : \mathbb{R}^n \to \mathbb{R}^n$ invertible transformation.

(a) Show that

$$H(Y) = H(X) - E[\ln J_{\phi^{-1}}],$$

where $J_{\phi^{-1}}$ is the Jacobian of ϕ^{-1} and $E[\,\cdot\,]$ is the expectation with respect to the probability density of X.

(b) Consider the linear transformation $Y = AX$, with $A \in \mathbb{R}^{n\times n}$ nonsingular matrix. What is the relation expressed by part (a) in this case?

3.10. Consider the Gaussian distribution

$$p(x_1, \dots, x_n) = \frac{\sqrt{\det A}}{(2\pi)^{n/2}} e^{-\frac{1}{2}\langle Ax, x\rangle},$$

where A is a symmetric $n \times n$ matrix. Show that the entropy of p is

$$H = \frac{1}{2} \ln[(2\pi e)^n \det A].$$

3.11. Let $X = (X_1, \dots, X_n)$ be a random vector in $\mathbb{R}^n$, with $E[X_j] = 0$ and denote by $A = a_{ij} = E[X_i X_j]$ the associated covariance matrix. Prove that

$$H(X) \le \frac{1}{2} \ln[(2\pi e)^n \det A].$$

When is the equality reached?

3.12. Consider the density of an exponentially distributed random variable with parameter $\lambda > 0$

$$p(x, \lambda) = \lambda e^{-\lambda x}, \qquad x \geq 0.$$

Find its entropy.

3.13. Consider the Cauchy's distribution on $\mathbb{R}$

$$p(x, \xi) = \frac{\xi}{4\pi} \frac{1}{x^2 + \xi^2}, \qquad \xi > 0.$$

Show that its entropy is

$$H(\xi) = \ln(4\pi\xi).$$

3.14. Find a generic formula for the informational energy of the mixture family $p(\xi, x) = C(x) + \xi^i F_i(x)$, $x \in \mathcal{X}$.

3.15. Let $f(x) = \dfrac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}$, $x \geq 0$, $\sigma > 0$, be the Rayleigh distribution. Prove that its entropy is given by

$$H(\sigma) = 1 + \ln \frac{\sigma}{\sqrt{2}} + \frac{\gamma}{2},$$

where γ is Euler's constant.

3.16. Show that the entropy of the Maxwell–Boltzmann distribution

$$p(x, a) = \frac{1}{a^3} \sqrt{\frac{2}{\pi}} x^2 e^{-\frac{x^2}{2a^2}}, \qquad a > 0,\ x \in \mathbb{R}$$

is $H(a) = \frac{1}{2} - \gamma - \ln(a\sqrt{2\pi})$, where γ is Euler's constant.

3.17. Consider the Laplace distribution

$$f(x, b, \mu) = \frac{1}{2b} e^{-|x-\mu|/b}, \qquad b > 0, \mu \in \mathbb{R}.$$

Show that its entropy is

$$H(b, \mu) = 1 + \ln(2b).$$

3.18. Let $\mu \in \mathbb{R}$. Construct a statistical model

$$\mathcal{S} = \{p_\xi(x);\, \xi \in \mathbb{E}, x \in \mathcal{X}\}$$

such that the functional $F : \mathcal{S} \longrightarrow \mathbb{R}$,

$$F(p(\cdot)) = \int_{\mathcal{X}} xp(x)\, dx - \mu$$

has at least one critical point. Is $\mathcal{M}_\mu = F^{-1}(0)$ a submanifold of $\mathcal{S}$?

3.19. Starting from the Euclidean space $(\mathbb{R}^n_+, \delta_{ij})$, find the Hessian metric produced by the Shannon entropy function

$$f : \mathbb{R}^n_+ \to R,\ f(x^1, \cdots, x^n) = \frac{1}{k^2} \sum_{i=1}^{n} \ln(k^2 x^i).$$

3.20. Show that the inequality (3.9.20) becomes identity for the uniform distribution $p : [0, e] \to [0, \infty)$, $p(x) = 1/e$, and this is the only distribution with this property.

3.21. (a) Let $a_n(x) = \frac{\xi^n \ln(n!)}{n!}$. Show that $\lim\limits_{n\to\infty} \left| \frac{a_{n+1}(x)}{a_n(x)} \right| = 0$ for any x;

(b) Show that the series $\sum\limits_{n\geq 0} \dfrac{\xi^n \ln(n!)}{n!}$ has an infinite radius of convergence;

(c) Deduce that the entropy for the Poisson distribution is finite.

3.22. Show that the entropy of the beta distribution

$$p_{a,b}(x) = \frac{1}{B(a,b)}\, x^{a-1}(1-x)^{b-1}, \quad 0 \leq x \leq 1$$

is always non-positive, $H(\alpha, \beta) \leq 0$, for any $a, b > 0$. For which values of a and b does the entropy vanish?