Quantities of Information: Continuous Case 10

This lecture introduces some basic facts when studying the quantities of information for continuous random variables. Continuous random variables are frequently encountered in a physical world, where source and channel are generally governed by continuous physical laws. For example, a source may be a sensor taking analog (i.e., continuous) measurements of the environment, and a channel may contain noise that is the best modeled as a continuous random variable (e.g., Gaussian). Since a mathematically rigorous treatment of quantities of information for general random variables (including continuous, continuous-discrete mixture, and even more abstract types) requires deep mathematical tools, we shall content ourselves with a largely heuristic treatment, and provide some advanced results without proof.

10.1 From Entropy to Differential Entropy

Recall that for a discrete random variable $X \sim P_X(x)$, $x \in \mathcal{X}$, its entropy is defined in Definition 2.10 as

$$H(X) = -\sum_{x \in \mathcal{X}} P_X(x) \log P_X(x). \tag{10.1}$$

Now, consider a continuous random variable X on the real line. We assume that X has a probability density function $f_X(x)$, $x \in \mathcal{S} \subseteq \mathbb{R}$. Here $\mathcal{S}$ indicates the set over which $f_X(x)$ is non-zero, and is called the support of X.

How to extend the concept of entropy to such a continuous X? A seemingly reasonable idea is to discretize X into a "quantized" discrete variable, and let the discretization go finer and finer. In light of this, fix a discretization cell size δ, and consider a discrete random variable $\tilde{X}_\delta$ with pmf

$$P_{\tilde{X}_\delta}(i) = \int_{i\delta}^{(i+1)\delta} f_X(x) \mathrm{d}x, \tag{10.2}$$

for $i = \ldots, -1, 0, 1, \ldots$. That is, we partition the real line into size-δ intervals called "cells", and discretize the continuous random variable X into these cells by assigning to each cell the probability of X occurring in that cell.

Under mild assumption on f_X, by the mean value theorem, for each i, there exists some $x_i \in [i\delta, (i+1)\delta]$, such that

$$P_{\tilde{X}_\delta}(i) = f_X(x_i)\delta. \tag{10.3}$$

With this, calculating the entropy of $\tilde{X}_\delta$, we have

$$\begin{aligned} H(\tilde{X}_\delta) &= -\sum_i P_{\tilde{X}_\delta}(i) \log P_{\tilde{X}_\delta}(i) \\ &= -\sum_i f_X(x_i)\delta \log(f_X(x_i)\delta) \\ &= -\sum_i f_X(x_i)\delta \log f_X(x_i) - \sum_i f_X(x_i)\delta \log \delta, \end{aligned} \tag{10.4}$$

where the first term converges to

$$-\int_{\mathcal{S}} f_X(x) \log f_X(x) \mathrm{d}x, \tag{10.5}$$

if this integral exists, as $\delta \to 0$, and the second term is exactly $\log \delta$.

Therefore, we have seen that

$$\lim_{\delta \to 0} \left[H(\tilde{X}_\delta) - \log \frac{1}{\delta} \right] = -\int_{\mathcal{S}} f_X(x) \log f_X(x) \mathrm{d}x. \tag{10.6}$$

That is, if we attempt to define the entropy of a continuous random variable as the entropy of its discretization in the fine discretization limit, this generally leads to infinite entropy since $\log \frac{1}{\delta}$ diverges as $\delta \to 0$. Nevertheless, after removing this $\log \frac{1}{\delta}$ dominant term, the remaining bias term, $-\int_{\mathcal{S}} f_X(x) \log f_X(x) \mathrm{d}x$, provides a relative measure of the entropy of X.

Given the preceding discussion, we introduce the definition of differential entropy.

Definition 10.1 For a continuous random variable X defined over the real line, with probability density function $f_X(x)$, $x \in \mathcal{S} \subseteq \mathbb{R}$, its differential entropy is defined as

$$h(X) = h(f_X) = -\int_{\mathcal{S}} f_X(x) \log f_X(x) \mathrm{d}x, \tag{10.7}$$

if the integral exists.

Remark 10.1 The preceding "derivation" of differential entropy is heuristic in nature. In particular, it heavily depends upon discretizing the real line using a uniform size-δ cells. It can be easily verified that, if a non-uniform discretization is applied, then the asymptotic behavior of the entropy of the discretized random variable will not yield the simple form as (10.6). Anyway, a cautious view is to treat differential entropy not as an extension of entropy, but a mathematical entity in its own right.

We can extend the concept of differential entropy to multiple random variables.

Definition 10.2 For a continuous random vector $\underline{X}$ defined over $\mathbb{R}^k$, with joint probability density function $f_{\underline{X}}(\underline{x})$, $\underline{x} \in \mathcal{S} \subseteq \mathbb{R}^k$, its differential entropy is defined as

$$h(\underline{X}) = h(f_{\underline{X}}) = -\int_{\mathcal{S}} f_{\underline{X}}(\underline{x}) \log f_{\underline{X}}(\underline{x}) \mathrm{d}\underline{x}, \tag{10.8}$$

if the integral exists.

We can also define conditional differential entropy for a pair of random variables.

Definition 10.3 For a pair of random variables X and Y, with joint probability density function $f_{X,Y}(x,y)$, $(x,y) \in \mathcal{S} \subseteq \mathbb{R}^k \times \mathbb{R}^l$, the conditional differential entropy is defined as

$$h(Y|X) = -\int_{\mathcal{S}} f_{X,Y}(x,y) \log f_{Y|X}(y|x) \mathrm{d}x\mathrm{d}y, \tag{10.9}$$

if the integral exists.

When X is a discrete random variable, as long as $f_{Y|X}$ exists for each $X = x$, we can still define conditional differential entropy $h(Y|X)$, as follows:

$$h(Y|X) = -\sum_{x \in \mathcal{X}} P_X(x) \int_{\mathcal{S}_x} f_{Y|X}(y|x) \log f_{Y|X}(y|x) \mathrm{d}y, \tag{10.10}$$

if the integrals exist, where $\mathcal{S}_x$ is the support of $f_{Y|X}(y|x)$.

Example 10.1 Let X be uniform with $f_X(x) = \frac{1}{b-a}$, $x \in [a,b]$. The differential entropy of X is

$$h(X) = \int_a^b \frac{1}{b-a} \log(b-a) \mathrm{d}x = \log(b-a) \text{ nats}. \tag{10.11}$$

Note that when $b - a < 1$, $h(X) < 0$. So differential entropy does not need to be non-negative.

Example 10.2 Let X be exponential with $f_X(x) = be^{-bx}$, $x \geq 0$. The differential entropy of X is

$$\begin{aligned} h(X) &= \int_0^\infty f_X(x)(bx - \ln b) \mathrm{d}x \\ &= b\mathbf{E}[X] - \ln b \\ &= b\frac{1}{b} - \ln b = 1 - \ln b \text{ nats}. \end{aligned} \tag{10.12}$$

Example 10.3 Let X be Gaussian with $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in \mathbb{R}$. The differential entropy of X is

$$\begin{aligned} h(X) &= \int_{-\infty}^{\infty} f_X(x) \left[\frac{(x-\mu)^2}{2\sigma^2} + \frac{1}{2}\ln(2\pi\sigma^2) \right] \mathrm{d}x \\ &= \frac{1}{2\sigma^2}\mathrm{var}[X] + \frac{1}{2}\ln(2\pi\sigma^2) \\ &= \frac{1}{2}\ln(2\pi e\sigma^2) \text{ nats.} \end{aligned} \tag{10.13}$$

Example 10.4 Let $\underline{X}$ be a k-dimensional Gaussian random vector with

$$f_{\underline{X}}(\underline{x}) = \frac{1}{(2\pi)^{k/2}\sqrt{|\mathbf{K}|}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T \mathbf{K}^{-1}(\underline{x}-\underline{\mu})}, \tag{10.14}$$

where the covariance matrix $\mathbf{K}$ is positive definite; i.e., $\underline{X} \sim \mathcal{N}(\underline{\mu}, \mathbf{K})$. The differential entropy of $\underline{X}$ is

$$\begin{aligned} h(\underline{X}) &= \int_{\mathbb{R}^k} f_{\underline{X}}(\underline{x}) \left[\frac{1}{2}(\underline{x}-\underline{\mu})^T \mathbf{K}^{-1}(\underline{x}-\underline{\mu}) + \frac{k}{2}\ln(2\pi) + \frac{1}{2}\ln|\mathbf{K}| \right] \mathrm{d}\underline{x} \\ &\overset{(a)}{=} \int_{\mathbb{R}^k} \frac{1}{(2\pi)^{k/2}\sqrt{|\mathbf{K}|}} e^{-\frac{1}{2}\underline{\tilde{x}}^T\underline{\tilde{x}}} \frac{1}{2}\underline{\tilde{x}}^T\underline{\tilde{x}} |\mathbf{K}^{1/2}| \mathrm{d}\underline{\tilde{x}} + \frac{k}{2}\ln(2\pi) + \frac{1}{2}\ln|\mathbf{K}| \\ &= \frac{1}{2}\int_{\mathbb{R}^k} \frac{1}{(2\pi)^{k/2}} e^{-\frac{1}{2}\underline{\tilde{x}}^T\underline{\tilde{x}}} \underline{\tilde{x}}^T\underline{\tilde{x}} \mathrm{d}\underline{\tilde{x}} + \frac{k}{2}\ln(2\pi) + \frac{1}{2}\ln|\mathbf{K}| \\ &= \frac{1}{2}\mathbf{E}_{\mathcal{N}(0,\mathbf{I}_{k\times k})}[\underline{\tilde{X}}^T\underline{\tilde{X}}] + \frac{k}{2}\ln(2\pi) + \frac{1}{2}\ln|\mathbf{K}| \\ &\overset{(b)}{=} \frac{1}{2}\mathrm{tr}\mathbf{E}_{\mathcal{N}(0,\mathbf{I}_{k\times k})}[\underline{\tilde{X}}\underline{\tilde{X}}^T] + \frac{k}{2}\ln(2\pi) + \frac{1}{2}\ln|\mathbf{K}| \\ &= \frac{1}{2}\mathrm{tr}\mathbf{I}_{k\times k} + \frac{k}{2}\ln(2\pi) + \frac{1}{2}\ln|\mathbf{K}| \\ &= \frac{k}{2}\ln(2\pi e) + \frac{1}{2}\ln|\mathbf{K}| \text{ nats,} \end{aligned} \tag{10.15}$$

where in (a) we have conducted a change of variable $\underline{\tilde{x}} = \mathbf{K}^{-1/2}(\underline{x} - \underline{\mu})$, and in (b) we have used an identity for trace $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.

Some important properties of entropy have analogies for differential entropy.

Proposition 10.1 Differential entropy satisfies the chain rule, i.e.,

$$h(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} h(X_i|X_{i-1}, \ldots, X_1), \tag{10.16}$$

where X_0 is understood as degenerated.

Proposition 10.2 Conditioning reduces differential entropy, i.e.,

$$h(Y|X) \leq h(Y), \tag{10.17}$$

with equality holding if and only if X and Y are independent.

Recall that for entropy, $H(X) = H(f(X))$ if and only if f is a bijection (Corollary 3.6). In general, differential entropy is no longer invariant under transform, even when the transform is bijective.

Proposition 10.3 Consider a continuous random vector $\underline{X}$ with probability density function $f_{\underline{X}}(\underline{x})$, $\underline{x} \in \mathcal{S} \subseteq \mathbb{R}^k$. We have

- For any given vector $\underline{b}$, $h(\underline{X} + \underline{b}) = h(\underline{X})$.
- For any given non-singular square matrix $\mathbf{A}$, $h(\mathbf{A}\underline{X}) = h(\underline{X}) + \log|\mathbf{A}|$.

For two probability density functions f_X and g_X, we can define their relative entropy in a similar fashion as in the discrete case.

Definition 10.4 For two probability density functions f_X and g_X, their relative entropy is defined by

$$D(f_X \| g_X) = \int_{\mathcal{S}} f_X(x) \log \frac{f_X(x)}{g_X(x)} \mathrm{d}x, \tag{10.18}$$

where $\mathcal{S}$ is the support over which either f_X or g_X is non-zero.

Note that if there exists an interval over which $g_X(x)$ is zero but $f_X(x)$ is strictly positive, then $D(f_X \| g_X) = \infty$.

It can be proved in a way analogous to the discrete case that relative entropy is always non-negative.

Proposition 10.4 For any f_X and g_X, $D(f_X \| g_X) \geq 0$, with equality holding if and only if $f_X(x) = g_X(x)$ almost everywhere over $\mathcal{S}$.

10.2 Maximum Entropy Density

Consider all probability density functions that satisfy

- Constraint [0]: $\int_{\mathcal{S}} f_X(x)\mathrm{d}x = 1$, with support $\mathcal{S}$;
- Constraints [1] to $[m]$: $\int_{\mathcal{S}} f_X(x) r_i(x)\mathrm{d}x = \alpha_i$, $i = 1, \ldots, m$.

Definition 10.5 Under Constraints [0] to $[m]$, the maximum entropy probability density function is defined as

$$f_X^*(x) = e^{\lambda_0 + \sum_{i=1}^{m} \lambda_i r_i(x)}, \quad x \in \mathcal{S}, \tag{10.19}$$

where $\lambda_0, \lambda_1, \ldots, \lambda_m$ are chosen so that $f_X^*(x)$ satisfies Constraints [0] to $[m]$ simultaneously.

Theorem 10.1 The maximum entropy probability density function $f_X^*(x)$ maximizes the differential entropy among all probability density functions satisfying Constraints [0] to [m].

Proof: The idea is to utilize the non-negativity of relative entropy. Specifically, consider an arbitrary probability density function $f_X(x)$ satisfying Constraints [0] to [m], and evaluate $D(f_X \| f_X^*)$:

$$\begin{aligned} D(f_X \| f_X^*) &= \int_{\mathcal{S}} f_X(x) \ln \frac{f_X(x)}{f_X^*(x)} \mathrm{d}x \\ &= -h(f_X) - \int_{\mathcal{S}} f_X(x) \left[\lambda_0 + \sum_{i=1}^{m} \lambda_i r_i(x) \right] \mathrm{d}x \\ &= -h(f_X) - \lambda_0 \int_{\mathcal{S}} f_X(x) \mathrm{d}x - \sum_{i=1}^{m} \lambda_i \int_{\mathcal{S}} f_X(x) r_i(x) \mathrm{d}x \\ &= -h(f_X) - \lambda_0 - \sum_{i=1}^{m} \lambda_i \alpha_i \\ &= -h(f_X) - \lambda_0 \int_{\mathcal{S}} f_X^*(x) \mathrm{d}x - \sum_{i=1}^{m} \lambda_i \int_{\mathcal{S}} f_X^*(x) r_i(x) \mathrm{d}x \\ &= -h(f_X) - \int_{\mathcal{S}} f_X^*(x) \left[\lambda_0 + \sum_{i=1}^{m} \lambda_i r_i(x) \right] \mathrm{d}x \\ &= -h(f_X) + h(f_X^*). \end{aligned} \tag{10.20}$$

Since $D(f_X \| f_X^*) \geq 0$ with equality holding if and only if $f_X(x) = f_X^*(x)$ almost everywhere over $\mathcal{S}$, we conclude that $h(f_X^*) \geq h(f_X)$. □

Example 10.5 Consider all probability density functions over $\mathcal{S} = [a, b]$, without any further constraint. So only Constraint [0] is present, and the maximum entropy probability density function is

$$f_X^*(x) = e^{\lambda_0}, \quad x \in [a, b]. \tag{10.21}$$

This means that $f_X^*(x)$ is a constant over $\mathcal{S}$, immediately suggesting that $f_X^*(x)$ should be a uniform probability density function, i.e.,

$$f_X^*(x) = \frac{1}{b - a}, \quad x \in [a, b]. \tag{10.22}$$

Example 10.6 Consider all probability density functions over $\mathcal{S} = \mathbb{R}$, satisfying $\mathbf{E}[X] = 0$ and $\mathbf{E}[X^2] = \sigma^2$. So the maximum entropy probability density function is

$$f_X^*(x) = e^{\lambda_0 + \lambda_1 x + \lambda_2 x^2}, \quad x \in \mathbb{R}. \tag{10.23}$$

The form of the probability density function is proportional to the exponential of a quadratic form, and hence it should be Gaussian. With mean zero and variance σ^2, the Gaussian probability density function is

$$f_X^*(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}. \tag{10.24}$$

This maximum entropy probability density function will be useful for calculating the capacity-cost function of Gaussian channels in the next lecture.

Example 10.7 Consider all probability density functions over $\mathcal{S} = [0, \infty)$, satisfying $\mathbf{E}[X] = \mu$. So the maximum entropy probability density function is

$$f_X^*(x) = e^{\lambda_0 + \lambda_1 x}, \quad x \in [0, \infty). \tag{10.25}$$

Such a form immediately implies that $f_X^*(x)$ should follow an exponential probability density function, i.e.,

$$f_X^*(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}}, \quad x \geq 0. \tag{10.26}$$

In this example, if we change the support $\mathcal{S}$ from $[0, \infty)$ to $\mathbb{R}$, then it can be easily verified that no maximum entropy probability density exists. In fact, letting X be a Gaussian random variable with mean μ and variance σ^2, the resulting differential entropy diverges as $\sigma^2 \longrightarrow 0$.

10.3 Entropy Power Inequality

For a continuous random variable X over $\mathbb{R}$ with differential entropy $h(X)$, we call

$$N(X) = \frac{1}{2\pi e} e^{2h(X)} \tag{10.27}$$

the entropy power of X. Rewriting

$$h(X) = \frac{1}{2} \ln(2\pi e N(X)), \tag{10.28}$$

we immediately observe that when X is Gaussian, $N(X)$ is nothing but the variance of X. So in general, $N(X)$ may be viewed as a "generalization" of the variance (i.e., "power") of X.

The idea can be extended to higher dimensions. For $\underline{X}$ over $\mathbb{R}^k$ with differential entropy $h(\underline{X})$, we call

$$N(\underline{X}) = \frac{1}{2\pi e} e^{\frac{2}{k} h(\underline{X})} \tag{10.29}$$

the entropy power of $\underline{X}$. When $\underline{X}$ is Gaussian with covariance matrix $\mathbf{K}$, $N(\underline{X}) = |\mathbf{K}|^{1/k}$.

The following fundamental inequality is called the entropy power inequality (EPI), relating the entropy powers of two independent random vectors and of their sum.

Theorem 10.2 For independent continuous random vectors $\underline{X}$ and $\underline{Y}$ over $\mathbb{R}^k$, we have

$$N(\underline{X} + \underline{Y}) \geq N(\underline{X}) + N(\underline{Y}), \tag{10.30}$$

with equality holding if and only if $\underline{X}$ and $\underline{Y}$ are Gaussian with proportional covariance matrices.

There are several ways of proving the EPI, but none of them appears to be simple enough to be taught in our lecture notes. So we simply give the result, without providing a proof.

10.4 Mutual Information

We can define mutual information for a pair of continuous random variables via relative entropy (Definition 10.4) as follows.

Definition 10.6 For a pair of random variables X and Y, with joint probability density function $f_{X,Y}(x, y)$, $(x, y) \in \mathcal{S} \subseteq \mathbb{R}^k \times \mathbb{R}^l$, the mutual information between X and Y is defined as

$$\begin{aligned} I(X;Y) &= D(f_{X,Y} \| f_X f_Y) && (10.31) \\ &= \int_{\mathcal{S}} f_{X,Y}(x, y) \log \frac{f_{X,Y}(x, y)}{f_X(x) f_Y(y)} \mathrm{d}x\mathrm{d}y, && (10.32) \end{aligned}$$

if the integral exists.

Remark 10.2 The definition (10.31) is not the most fundamental form of mutual information for general random variables. The most fundamental and also the most general definition of the mutual information between two random variables X and Y is as follows:

$$I(X;Y) = \sup_{\mathbb{Q}_X, \mathbb{Q}_Y} I([X]_{\mathbb{Q}_X}; [Y]_{\mathbb{Q}_Y}), \tag{10.33}$$

where $\mathbb{Q}_X$ (resp. $\mathbb{Q}_Y$) denotes a partitioning of $\mathcal{X}$ (resp. $\mathcal{Y}$), i.e., a finite or coutable number of subsets of $\mathcal{X}$ (resp. $\mathcal{Y}$) which are disjoint and whose union is $\mathcal{X}$ (resp. $\mathcal{Y}$). Accordingly, $[X]_{\mathbb{Q}_X}$ (resp. $[Y]_{\mathbb{Q}_Y}$) denotes the discretized random variable obtained by assigning to $[X]_{\mathbb{Q}_X} = i$ (resp. $[Y]_{\mathbb{Q}_Y} = j$) the probability that X (resp. Y) belongs to the subset indexed by i (resp. j) according to partitioning $\mathbb{Q}_X$ (resp. $\mathbb{Q}_Y$). This definition applies to very general random variables, and for continuous random variables with joint probability density functions, it is reduced into the form given by (10.31).

We can also define conditional mutual information for a triple of continuous random variables, in a similar fashion as for discrete random variables.

Definition 10.7 For a triple of random variables X, Y and Z, with joint probability density function $f_{X,Y,Z}(x, y, z)$, $(x, y, z) \in \mathcal{S} \subseteq \mathbb{R}^k \times \mathbb{R}^l \times \mathbb{R}^m$, the conditional mutual information between X and Y given Z is defined as

$$I(X; Y|Z) = \int_{\mathcal{S}} f_{X,Y,Z}(x, y, z) \log \frac{f_{X,Y|Z}(x, y|z)}{f_{X|Z}(x|z) f_{Y|Z}(y|z)} \mathrm{d}x\mathrm{d}y\mathrm{d}z, \tag{10.34}$$

if the integral exists.

The following properties hold for $I(X; Y)$ and $I(X; Y|Z)$.

Corollary 10.1 Mutual information and conditional mutual information satisfy the following properties:

- $I(X; Y) \geq 0$, with equality holding if and only if X and Y are independent.
- $I(X; Y|Z) \geq 0$, with equality holding if and only if $X \leftrightarrow Z \leftrightarrow Y$.
- (DPI) If $X \leftrightarrow Y \leftrightarrow Z$, $I(X; Y) \geq I(X; Z)$ with equality holding if and only if $X \leftrightarrow Z \leftrightarrow Y$ holds as well.
- $I(X; Y) = h(Y) - h(Y|X)$.
- $I(X; Y|Z) = h(Y|Z) - h(Y|X, Z)$.

Remark 10.3 The first three properties of Corollary 10.1 are general. The last two properties require the existence of the related differential entropies. Nevertheless, they also hold when X is discrete and Y has a conditional probability density function given X (and Z, when applicable), and for this reason they are useful when calculating mutual information of a channel with discrete input X and continuous output Y, as frequently encountered in practical communication systems.

Notes

Differential entropy already appeared in Shannon's 1948 two-part article [1], and several of its key properties including the EPI also appeared therein, without proof. The most fundamental and also the most general treatment of quantities of information for general random variables is attributed to the Russian school; see, e.g., [5].

Exercises

1. Calculate the differential entropy of the following probability density functions.
   - Cauchy: $f(x) = \frac{\lambda}{\pi}\frac{1}{\lambda^2+x^2}$, $\mathcal{S} = \mathbb{R}$.
   - Laplace: $f(x) = \frac{1}{2\lambda}e^{-|x-\theta|/\lambda}$, $\mathcal{S} = \mathbb{R}$.
   - Rayleigh: $f(x) = \alpha x e^{-\alpha x^2/2}$, $\mathcal{S} = [0, \infty)$.
2. Consider a sequence of i.i.d. random variables $\{X_i, i = 1, 2, \ldots\}$, and their sample means $\{S_n, n = 1, 2, \ldots\}$, $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.
   a) When X_i is discrete, with entropy $H(X)$, calculate $\frac{1}{n}H(S_1, S_2, \ldots, S_n)$.
   b) When X_i is continuous, with differential entropy $h(X)$, calculate $\frac{1}{n}h(S_1, S_2, \ldots, S_n)$.
3. For independent continuous random variables X and Y, prove that $h(X + Y) \geq h(X)$.
4. Consider a k-dimensional continuous random vector $\underline{X}$.
   - If $\underline{X}$ has zero mean, and has covariance matrix $\mathbf{K}$, what is the maximum differential entropy of $\underline{X}$?
   - Prove Hadamard's inequality, $|\mathbf{K}| \leq \prod_{i=1}^{k} \mathbf{K}_{ii}$.
   - Prove that the log-determinant $\ln|\mathbf{K}|$ is concave with respect to $\mathbf{K}$.
5. Prove the following generalization of the maximum entropy principle: for any given probability density function $g_X(x)$, $x \in \mathcal{S}$, the probability density function $f_X(x)$ that minimizes $D(f_X \| g_X)$ while satisfying $\int_{\mathcal{S}} f_X(x) r_i(x) \mathrm{d}x = \alpha_i$, $i = 1, 2, \ldots, m$, is given by the following form:

$$f_X(x) = g_X(x)e^{\lambda_0 + \sum_{i=1}^{m}\lambda_i r_i(x)},$$

   where $\{\lambda_i\}_{i=0,1,\ldots,m}$ are parameters.
6. Verify that in the scalar case, the EPI can be rewritten in the following equivalent form: for independent continuous random variables X and Y over $\mathbb{R}$, letting $\tilde{X}$ and $\tilde{Y}$ be independent Gaussian random variables satisfying $h(\tilde{X}) = h(X)$ and $h(\tilde{Y}) = h(Y)$, it holds that

$$h(X + Y) \geq h(\tilde{X} + \tilde{Y}). \tag{10.35}$$