ONE-PARAMETER EXTENDED FISHER INFORMATION

WORACHET BUKAEW

A Thesis Submitted to Graduate School of Naresuan University
in Partial Fulfillment of the Requirements
for the Master of Science Degree in Theoretical Physics
May 2022

Thesis entitled “One-parameter extended Fisher information”
by Worachet Bukaew
has been approved by the Graduate School as partial fulfillment of the requirements for the Master of Science in Theoretical Physics of
Naresuan University.

Oral Defense Committee

. . . . . Chair
(Assistant Professor Monsit Tanasittikosol, Ph.D.)

. . . . . Advisor
(Assistant Professor Sikarin Yoo-Kong, Ph.D.)

. . . . . Co-Advisor
(Assistant Professor Pichet Vanichchapongjaroen, Ph.D.)

. . . . . Internal Examiner
(Assistant Professor Seckson Sukhasena, Ph.D.)

Approved

. . . . . .
(Associate Professor Krongkarn Chootip, Ph.D.)
Dean of the Graduate School

ACKNOWLEDGMENTS

I would like to thank my advisor, Assistant Professor Sikarin Yoo-Kong, for giving me a good research topic. Also thank for his suggestions, discussions, and motivations till I finished my thesis. Thank for his explanation of difficult concept and training. And thank you for all knowledge which he gave to me.

I would like to thank all people at IF, they all give me a training of using programs, for example, Latex, Mathematica, etc., which are in out of my familiar using.

Furthermore, I would like to thank Center of Excellence in Theoretical and Computational Science (TaCs-CoE), KMUTT and The Prof. Dr. Sujin Jinahyon Foundation for the partial financial support.

Finally, I would like to thank my parent, my father, mother and sister, who support everything to me till I graduate my Master of Master of Science Degree.

Worachet Bukaew

LIST OF CONTENTS

Chapter **Page**

LIST OF CONTENTS (CONT.)

Chapter **Page**

LIST OF TABLES

LIST OF FIGURES

Title	ONE-PARAMETER EXTENDED FISHER INFORMATION
Author	Worachet Bukaew
Advisor	Assistant Professor Sikarin Yoo-Kong, Ph.D.
Co-Advisor	Assistant Professor Pichet Vanichchapongjaroen, Ph.D.
Academic Paper	Thesis M.S. in Theoretical Physics, Naresuan University, 2022
Keywords	Fisher information, Non-extensive statistics, Kullback-Leibler divergence, Non-standard Lagrangian, Variation principle

ABSTRACT

We introduce the generalised Fisher information or the one-parameter extended class of the Fisher information. This new form of the Fisher information is obtained from the intriguing connection between the standard Fisher information and the variational principle together with the non-uniqueness property of the Lagrangian. Furthermore, one could treat this one-parameter Fisher information as a generating function for obtaining what is called Fisher information hierarchy. The generalised Cramér-Rao inequality is also derived. The interesting point is the fact that the whole Fisher information hierarchy, except for the standard Fisher information, does not follow the additive rule. This could suggest that there is an indirect connection between the Tsallis entropy and the one-parameter Fisher information. Furthermore, the whole Fisher information hierarchy is also obtained from the two-parameter Kullback-Leibler divergence.

CHAPTER I

INTRODUCTION

1.1 Background and motivation

There is no doubt that we are now in the "information era". The information is physical [1] and plays an essential role in modern physics ranging from thermodynamics, statistical mechanics, quantum mechanics to relativity. The birth of information theory can be traced back to the seminal paper of Shannon [2] on communication. The key quantity in this context is the entropy or more precisely "Shannon entropy" as the mean value of information or uncertainty inherent in the possible outcomes. The interesting point is that the Shannon entropy is in the same form as the Gibbs-Boltzmann entropy in the context of statistical mechanics if one ignores the Boltzmann's constant, which measures the configuration of the microscopic states. However, the notion of entropy was first introduced in the context of thermodynamics the second law of thermodynamics, which is a bit more abstract relating to the heat flow in or out and the temperature of the system. However, if we trace back long before the breakthrough work of Shannon, Fisher purposed another information quantity, later known as Fisher information [3], as a measurement uncertainty on estimating unknown parameters in the system. This means that the Fisher information allows us to probe into the internal structure of the system. At this point, Shannon entropy and Fisher information provides a complete description of the system in the sense that the Fisher information can give an insight of what the system is made of and the Shannon entropy gives the system behaviour in the big picture. Moreover, the Shannon differential entropy and Fisher information are connected which was first observed by Kullback [4]. With the Kullback insight, the Fisher information matrix can be obtained from the second derivative of the Kullback-Leibler divergence(or the relative entropy).

The generalised version of the Shannon entropy was first introduced by Renyi [5]. The Renyi entropy comes with a parameter and with a suitable limit, the Shannon entropy can be recovered. Then one could think that the Renyi entropy is a one-parameter extended class of the Shannon entropy. On the statistical mechanical side, the generalised version of the Gibbs-Boltzmann entropy was proposed by Tsallis [6]. Again, this Tsallis entropy comes with a parameter and the Gibbs-Boltzmann entropy can be recovered with a suitable limit. One main feature of the Tsallis entropy is the non-additive property directly related to the non-extensivity of the system. Consequently, this leads to a new kind of research area known as the Tsallis statistics with a wide range of applications in statistics, physics, chemistry, economics, and biosciences [7]. On the other hand, several extensions of the Fisher information have been proposed with different aspects [8–13] to serve different uses in statistics.

1.2 Objectives

The aim of this work is to derive the one-parameter generalisation of the Fisher information.

1.3 Frameworks

In this contribution, we propose another one-parameter extended class of the Fisher information. The key motivation and derivation come from the intriguing connection between Fisher information and variational principle observed by Frieden [14–16] together with one-parameter extended class of the Lagrangian [21].

1.4 Structure of the thesis

The body of this thesis is the following. In chapter 2, we will give a brief review of basic statistical notation, entropies, generalized entropies, and Fisher information. In chapter 3, one parameter extended class of the Fisher information is derived by

employing the connection with the variational principle. The Fisher information hierarchy will be so obtained, the extended Cramér-Rao inequality and non-additive property are given as well. Moreover, The connection between two-parameter Kullback-Leibler divergence and Fisher information hierarchy will be established and we also show that Fisher information hierarchy has a connection to Shannon entropy throught out matric tensor in Kullback-Leibler divergence. The last chapter will be about the conclusion and discussion.

CHAPTER II

THEORIES

In this chapter, we will first provide all necessary basic ingredients. The basic of the statistics such as the notation of random variables, probabilities, the expectation of random variables and statistical moments will be given in the first section . In the next section, the concept of entropies will be discussed including thermodynamic entropy, statistical entropy, Shannon entropy, Tsallis entropy and Kullback-Leibler divergence. The definition of the generalised entropy will be mentioned as well. The standard Fisher information will be given at the end of the chapter together with the connection with the geometric metric tensor on the probability manifold.

2.1 Basic statistics

2.1.1 Random variable and probability

In statistics, we usually deal with a random process which is an event or experiment that has random outcomes, i.e., tossing a coin, rolling a die, choosing a card. For these kinds of experiments, we cannot exactly predict an outcome. Then there will be a range of possibilities that we can calculate the probability of a particular outcome. Random variables give numerical value to outcomes of random events. Normally, the random variables, defined on the sample space Ω which is a collection of all possible outcomes of a random event, are denoted by capital letters, i.e., $X =$ number of aces in a card hand or $Y =$ total of lotto numbers. The random variables can be divided into two classes which are discrete and continuous cases.

2.1.1.1 Discrete random variable.

Definition: A discrete random variable X is a measurable and countable value from a set of possible outcomes Ω to a measurable space. That is

$$X \subseteq \Omega \ , \tag{2.1}$$

where X is a set of possible values, $X = \{x_1, x_2, ..., x_N\}$.

Let us give an example. Suppose that we would like to predict the outcome of the rolling unbiased dice. Of course there are six possible outcomes and therefore $\Omega = \{1, 2, 3, 4, 5, 6\}$. If we are interested in only even number outcomes then $X = \{2, 4, 6\}$ is our a set of random number. We now further introduce a quantity, associated with a chance of getting a particular outcome, called **probability mass function (PMF)**.

Definition: Let X be a discrete random variable (finite or countably infinite). The function

$$P_X(x_i) = P(X = x_i), \ \text{ for } \ i = 1, 2, 3, ... \ , \tag{2.2}$$

is called the probability mass function (PMF) of X. The subscript X indicates that this is the PMF of random variable X. These PMFs must satisfy

$$\sum_{i=1}^{N} P_X(x_i) = 1, \tag{2.3}$$

which is called the normalisation condition.

If we again consider the unbiased dice, we can define probabilities of each value as $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$ and these all should satisfy normalization condition $\sum_{i=1}^{6} P_X(x_i) = 1$.

2.1.1.2 Continuous random variable.

Definition: A random variable X is said to be continuous, when it is measurable and uncountable value from a set of possible outcomes Ω to a measurable space. Therefore, this means that X can be any, uncertain, possible value

$$X = [a, b] \,, \tag{2.4}$$

where a and b are boundary value of X.

Let us give an example. Here we would like to measure the temperature of water and the range of possible temperature is $X = [25, 30]$ degree celsius from the sample space $\Omega \in$ R. We can also define a chance of getting random variables called **probability density function (PDF)**.

Definition: The probability that a random variable X takes a value in the (open or closed) interval $[a, b]$ is given by the integral of a function called the probability density function $f_X(x)$:

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx \,, \tag{2.5}$$

where $P(a \leq X \leq b)$ indicates probability of X in interval $[a, b]$.

If we consider random variable being any real numbers and the PDF is normalised so that

$$\int_{-\infty}^{\infty} f_X(x)dx = 1. \tag{2.6}$$

Normally, we might write it in a simple way as

$$\int_{\Omega} f(x)dx = 1 \,, \tag{2.7}$$

where random variable is finite value.

2.1.2 Moment generating function

Here, we will introduce and discuss moment generating functions which will be recalled later. The moment generating functions of a random variable X is a function $M_X(\alpha)$ defined as

$$M_X(\alpha) = \begin{cases} \sum_i^n e^{\alpha x_i} P(x_i) & \text{for discrete random variable} \\ \int_\Omega e^{\alpha x} f(x) dx & \text{for continuous random variable} \end{cases} \tag{2.8}$$

Then we can consider Taylor series of exponential

$$e^{\alpha X} = 1 + \alpha X + \frac{(\alpha X)^2}{2!} + \frac{(\alpha X)^3}{3!} + ... + \frac{(\alpha X)^n}{n!} + ... \,, \tag{2.9}$$

and expected values of Equation (2.9) can be written as

$$\begin{aligned} M_X(\alpha) = \left\langle e^{\alpha X} \right\rangle &= 1 + \alpha \left\langle X \right\rangle + \alpha^2 \left\langle \frac{X^2}{2!} \right\rangle + \alpha^3 \left\langle \frac{X^3}{3!} \right\rangle + ... + \alpha^n \left\langle \frac{X^n}{n!} \right\rangle + ... \\ &= 1 + \alpha\, m_1 + \alpha^2 \frac{m_2}{2!} + \alpha^3 \frac{m_3}{3!} + ... + \alpha^n \frac{m_n}{n!} + ... \,, \end{aligned} \tag{2.10}$$

where m_n is called the n^{th} moment. Next, we can also consider logarithm function of moment generating functions as follow,

$$\begin{aligned} K_X(\alpha) &\equiv \log M_X(\alpha) = \log \left\langle e^{\alpha X} \right\rangle \\ &= \log\Big(1 + \alpha\, m_1 + \alpha^2 \frac{m_2}{2!} + \alpha^3 \frac{m_3}{3!} + ... + \alpha^n \frac{m_n}{n!} + ...\Big). \end{aligned} \tag{2.11}$$

Normally there are 2 type of moments. The first one is **the moment about the origin (raw moment)** of a random variable X, denoted by m_n (as pervious)

$$m_n = \begin{cases} \sum_i^n x_i^n P(x_i) & \text{for discrete random variable} \\ \int_\Omega x^n f(x) dx & \text{for continuous random variable} \end{cases} . \tag{2.12}$$

The second one is **the central moment** is moment about the mean (μ) of a random variable X, denoted by m'_n ,

$$m'_n = \begin{cases} \sum_i^n (x_i - \mu)^n P(x_i) & \text{for discrete random variable} \\ \int_\Omega (x - \mu)^n f(x) dx & \text{for continuous random variable} \end{cases} . \tag{2.13}$$

Actually the meaning of the raw moments is just the expected value of x^{n} about origin. But the central moments refer to the behaviour of distribution for example,

1) the first Moment is a measure of the central location.

2) the second Moment is a measure of dispersion/spread.

3) the third Moment is a measure of asymmetry.

4) the fourth Moment is a measure of outliers/tailedness.

Moreover, there are the relations between raw and central moments as well. For example, the 2^{nd} central moment m_2' can be expressed as follows

$$\begin{aligned} m_2' &= \left\langle (x-\mu)^2 \right\rangle \\ &= \left\langle x^2 - 2\mu x + \mu^2 \right\rangle \\ &= m_2 - m_1^2, \end{aligned} \tag{2.14}$$

where $m_1 = \mu$ is the mean value. In the same fashion, some other order of central moments are related with raw moments as follows

$$m_3' = m_3 - m_3 m_2 + 2m_1^3 \tag{2.15}$$

$$m_4' = m_4 - 4m_1 m_3 + 6m_1^2 m_2 - 3m_1^4. \tag{2.16}$$

2.1.3 Cumulant generating function

If we define $Y \equiv 1 + \alpha\, m_1 + \alpha^2 \frac{m_2}{2!} + \alpha^3 \frac{m_3}{3!} + ... + \alpha^n \frac{m_n}{n!} + ...$, we can expand logarithm function as

$$\log(1+Y) = Y - \frac{Y^2}{2} + \frac{Y^3}{3} - \frac{Y^4}{4} + \dots . \tag{2.17}$$

Then Equation (2.11) is rewritten as

$$\begin{aligned} K_X(\alpha) &= \left(\alpha\, m_1 + \alpha^2 \frac{m_2}{2!} + \alpha^3 \frac{m_3}{3!} + ...\right) \\ &\quad -\frac{1}{2}\left(\alpha^2 m_1^2 + \alpha^3 \frac{m_1 m_2}{2!} + \alpha^4 \frac{m_1 m_3}{3!} + ...\right) \\ &\quad +\frac{1}{3}\left(\alpha^3 m_1^3 + \alpha^4 \frac{m_1^2 m_2}{2!} + \alpha^5 \frac{m_1^2 m_3}{3!}\right) \\ &\quad -\frac{1}{4}\left(\alpha^4 m_1^4 + \alpha^5 \frac{m_1^3 m_2}{2!} + \alpha^6 \frac{m_1^3 m_3}{3!}\right) + \dots . \end{aligned} \tag{2.18}$$

Rearranging and grouping the common terms, we will get

$$\begin{aligned} K_X(\alpha) &= m_1\alpha + \left[m_2 - m_1^2\right]\frac{\alpha^2}{2!} + \left[m_3 - 3m_1m_2 + 2m_1^3\right]\frac{\alpha^3}{3!} + \\ &\quad \left[m_4 - 4m_3m_1 - 3m_2^2 + 12m_2m_1^2 - 6m_1^4\right]\frac{\alpha^4}{4!} \\ &= k_1\alpha + k_2\frac{\alpha^2}{2!} + k_3\frac{\alpha^3}{3!} + k_4\frac{\alpha^4}{4!} + ... = \sum_{n=1}^{\infty}\frac{k_n(X)}{n!}(\alpha)^n\,, \end{aligned} \tag{2.19}$$

where k_n is n^{th} cumulants given by

$$k_1 = m_1' = m_1\,, \tag{2.20}$$

$$k_2 = m_2' = m_2 - m_1^2\,, \tag{2.21}$$

$$k_2 = m_3' = m_3 - m_3m_2 + 2m_1^3\,, \tag{2.22}$$

$$k_4 = m_4' - 3m_2 = m_4 - 4m_1m_3 - 3m_2^2 + 12m_1m_2^2 - 6m_1^4\,, \tag{2.23}$$

This means that cumulants can be used to describe the behaviour of the distribution as well. Recalling the definitions of moment and cumulant generating function (2.10) and Equation. (2.11), we then introduce **the effective values** of a random variable [18].

$$e^{K_X[\alpha-1]} = \left\langle e^{(\alpha-1)\cdot X}\right\rangle, \tag{2.24}$$

where $[\cdot]$ is a notion to emphasise the difference between the terms $\alpha - 1$ for two sides of the Equation. To match them, we define X_α such that

$$(\alpha - 1)\cdot X_\alpha = K_X[\alpha - 1]. \tag{2.25}$$

Then we will have

$$e^{(\alpha-1)\cdot X_\alpha} = e^{K_X[\alpha-1]} = \left\langle e^{(\alpha-1)\cdot X}\right\rangle\,. \tag{2.26}$$

Expanding the right hand side of Equation(2.26), one can obtain

$$X_\alpha = \sum_{n=1}^{\infty}\frac{k_n(X)}{n!}(\alpha-1)^{n-1}\,, \tag{2.27}$$

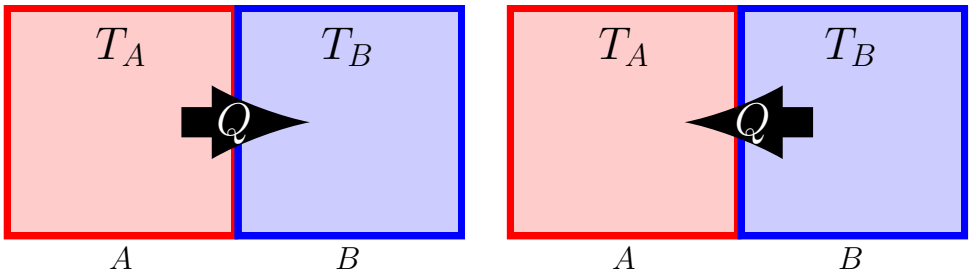
where X_α is called the effective values of order α of the random variable X.

2.2 Entropy and Fisher information

In this section, the notions of various entropies will be given together with their physical meaning. In the last subsection, the Fisher information will be discussed.

2.2.1 Thermodynamic entropy

Here we shall first discuss the notion of the entropy in the context of the thermodynamics. The notion of the entropy was introduced to capture the statement of the second law of thermodynamics which is concerned with the direction of the natural process (irreversible process). A common example is that the heat always spontaneously flows from a hot body to the cold body. We never encounter the situation that the heat spontaneously flows from a cold body to a hot body as shown in Figure 1.



(a) Natural process **(b) Impossible process**

Figure 1 When we contact system A and B together with temperature of A is greater than B ($T_A > T_b$), heat will be naturally transferred from A to B (a). While the opposite case (b) should be impossible without external conditions.

Then, we can conclude that entropy is the quantity that tell us what the processes in thermodynamic are possible or impossible. The mathematical interpretation of entropy was introduced by Rudolf Clasius. For a closed system, which evolves along the reversible path from the initial state to the final state, an infinitesimal increment of the entropy dS is given by

$$dS = \frac{đQ}{T} \ , \tag{2.28}$$

where $đQ$ is an infinitesimal transfer of heat to the system and T is a common temperature between the system and the environment which supplies heat. The symbols d and $đ$ are employed to denote exact differential and inexact differential, respectively. And, the concept of an exact differential refers to concept of path independence, while inexact differential refers to concept of path dependence [19].

Another point that we need to introduce, before we start to consider entropy change of system, is Clausius inequality. Basically, The Carnot's theorem, for the Carnot cycle, gives

$$\eta = 1 - \frac{đQ_c}{đQ_h} = 1 - \frac{T_c}{T_h} \,. \tag{2.29}$$

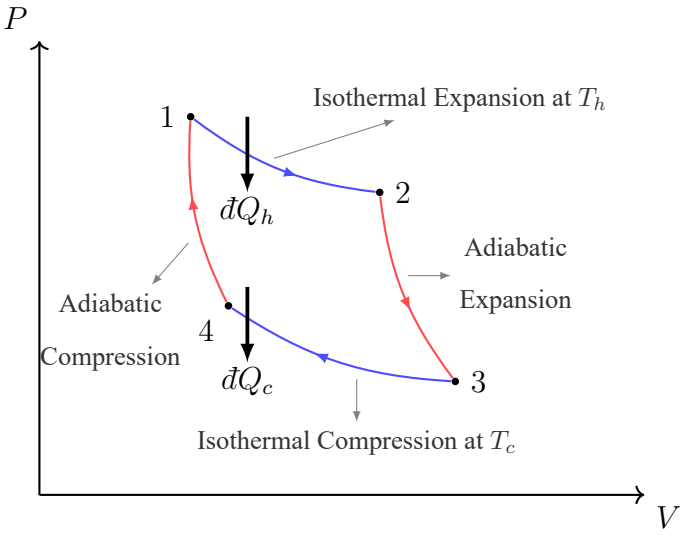


Figure 2 The Carnot cycle contains with isothermal and adiabatic process.

Where η is efficiency of Carnot heat engine and while system receives heat $đQ_h$ from high temperature T_h reservoir and rejects heat $đQ_c$ to lower temperature T_c reservoir, see Figure 2. Since, $đQ_c$ is negative, it reduces to

$$\frac{đQ_h}{T_h} + \frac{đQ_c}{T_c} = 0 \,. \tag{2.30}$$

Next, for arbitrary reversible process (closed loop), one can approximate it with many Carnot cycle as shown in Figure 3.

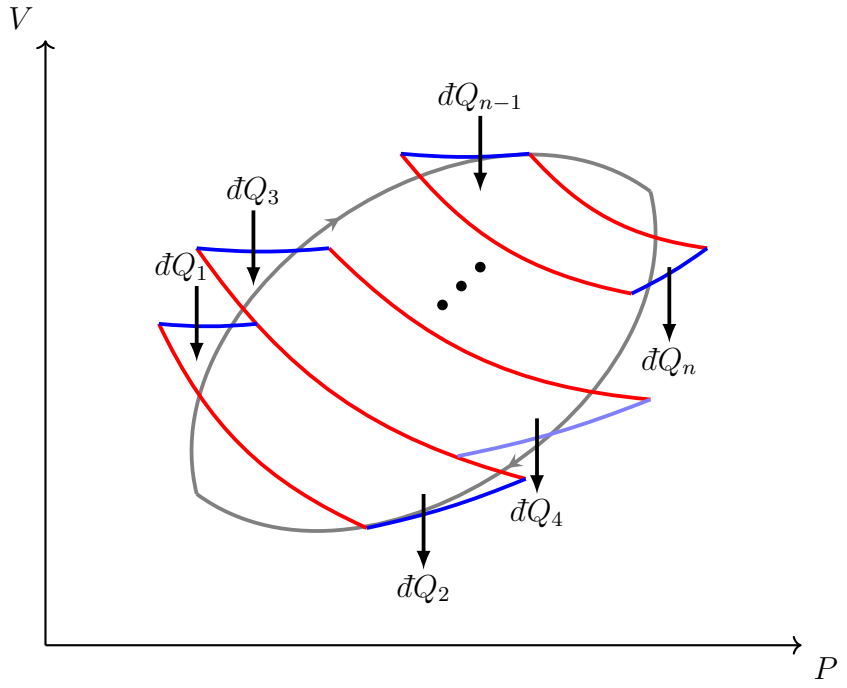


Figure 3 The reversible cycle which can be sub-divided by drawing a family of Carnot cycles.

Therefore, Equation (2.30) can be considered for every Carnot cycle in this process and then we obtain

$$\sum_{i=1}^{n} \frac{đQ_i}{T_i} = 0 \quad \text{or} \quad \oint_{\text{reversible loop}} \frac{đQ}{T} = 0 \,, \tag{2.31}$$

where $đQ_i$ is heat flow for whole process at a temperature T_i. Next, we will consider arbitrary irreversible process.

By The Carnot principle on the second law of thermodynamics, which is "efficiency of an all irreversible heat engine is always less than the efficiency of a reversible one operating between same two thermal reservoirs" [20], what we have now is

$$\begin{aligned} \eta_{ir} &< \eta_r \\ 1 - \frac{đQ'_c}{đQ'_h} &< 1 - \frac{T_c}{T_h} \,, \end{aligned} \tag{2.32}$$

or

$$\frac{đQ'_h}{T_h} + \frac{đQ'_c}{T_c} < 0 \,. \tag{2.33}$$

Please note that, we just use $đQ'$ to indicate different of heats for irreversible and reversible process but these are the same sort of quantity. With the same reason to get Equation (2.31), we lastly get

$$\sum_{i=1}^{n} \frac{đQ_i}{T_i} < 0 \quad \text{or} \quad \oint_{\text{irreversible loop}} \frac{đQ}{T} < 0 \,. \tag{2.34}$$

Actually, if we combine Equation (2.31) and (2.34) together,

$$\oint \frac{đQ}{T} \leq 0 \begin{cases} \oint \frac{đQ}{T} = 0 & \text{for reversible closed loop} \\ \oint \frac{đQ}{T} < 0 & \text{for irreversible closed loop} \end{cases} \tag{2.35}$$

which is known as Clausius inequality. Here, recalling entropy (2.28), the total change of entropy will be given by

$$\Delta S = \int_{\text{initial state}}^{\text{final state}} \frac{đQ}{T} \,. \tag{2.36}$$

Therefore, the entropy change of any closed reversible process is zero

$$\Delta S = 0 = \oint_{\text{reversible loop}} \frac{đQ}{T} \tag{2.37}$$

as a consequence of the Clausius inequality. Equation (2.37) gives an important feature called the path independent property between the initial state and the final state.

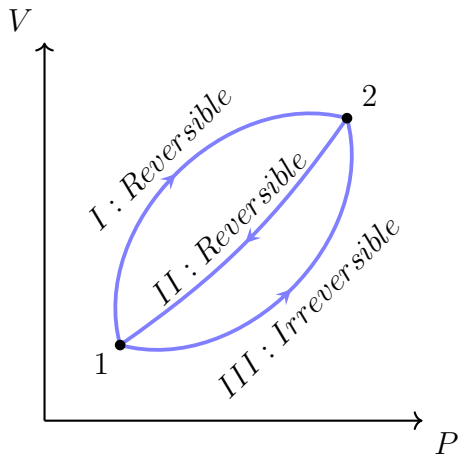


Figure 4 here are three possible paths from initial state (1) to the final state (2).

For the irreversible process, we can always connect the initial and final states with a fictitious reversible process and integrate along that path to calculate the difference in entropy. Considering an irreversible cycle in Figure 4, included path II and III, and applying Clausius inequality, we have

$$\begin{aligned} \oint_{\text{irreversible loop}} \frac{đQ}{T} &< 0 \\ \int_{\text{III}} \frac{đQ}{T} + \int_{\text{II}} \frac{đQ}{T} &< 0 \,. \end{aligned} \tag{2.38}$$

Next, if we consider the reversible cycle, included path I and II, we have

$$\begin{aligned} \oint_{\text{reversible loop}} \frac{đQ}{T} &= 0 \\ \int_{\text{I}} \frac{đQ}{T} + \int_{\text{II}} \frac{đQ}{T} &= 0 \\ \int_{\text{I}} \frac{đQ}{T} &= -\int_{\text{II}} \frac{đQ}{T} \,. \end{aligned} \tag{2.39}$$

Here, replacing the second term in (2.38) with (2.39), we obtains

$$\int_{\text{II}} \frac{đQ}{T} < \int_{\text{I}} \frac{đQ}{T} = \Delta S \,.$$

Finally, we can write entropy change for irreversible process as

$$\Delta S > \int_{\text{initial state}}^{\text{final state}} \frac{đQ}{T} \,. \tag{2.40}$$

In the case that the process is adiabatic $đQ = 0$, together with (2.37), we obtain

$$\Delta S \geq 0 \,. \tag{2.41}$$

Now we can draw a conclusion that the process of heat transferring from a hot body to a cold body is allowed because the entropy change is greater than zero. The reverse process is forbidden because the entropy change is contradict with (2.41).

2.2.2 Boltzmann-Gibbs entropy

In the previous subsection, the notion of thermodynamic entropy is discussed. Here, another notion of entropy known as the statistical entropy will derived. What we know is that thermodynamics is concerned with the macroscopic behaviour of the system. However, the system is constituted of tiny parts, i.e., a box of N atoms. The interesting fact is that the macroscopic properties of the system actually is the statistical emergence of one configuration from possible many configurations of the microscopic states. Let us now define an ensemble W which is a collection of all possible configuration of the microstates and suppose the system is composed of two subsystems at thermal equilibrium, see Figure 5. Then we could define number of microstates of the whole system $W_{1+2}(E_{1+2})$, the first subsystem $W_1(E_1)$ and the second subsystem $W_2(E_2 = E_{1+2} - E_1)$. When E_1, E_2 and E_{1+2} are energies of subsystem 1, subsystem 2 and the whole system, respectively.

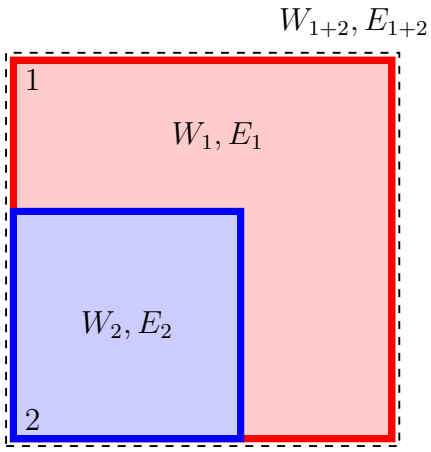


Figure 5 The composite system, which contain with subsystem 1 and 2.

Of course, relation between these three ensembles could be written as

$$W_{1+2}(E_{1+2}) = W_1(E_1)W_2(E_{1+2} - E_1) \ . \tag{2.42}$$

Therefore, probability of microstates is expressed as

$$P(E_{1+2}) = CW_1(E_1)W_2(E_{1+2} - E_1) \ , \tag{2.43}$$

where C is normalised constant of probability. Since the logarithm function is concave, the extremum point is not altered, we obtain

$$\ln P(E_{1+2}) = \ln C + \ln W_1(E_1) + \ln W_2(E_{1+2} - E_1). \tag{2.44}$$

Here, we need to know that which system provides the extremum probability respect to energy E_1. What we have now is

$$\frac{\partial}{\partial E_1} \ln P(E_{1+2}) = \frac{\partial}{\partial E_1} \ln W_1(E_1) + \frac{\partial}{\partial E_1} \ln W_2(E_{1+2} - E_1) \ . \tag{2.45}$$

Using the fact that $\frac{\partial}{\partial E_1} P(E_{1+2}) = 0$, Equation (2.45) becomes

$$\begin{aligned} 0 &= \frac{\partial}{\partial E_1} \ln W_1(E_1) + \frac{\partial}{\partial E_2} \ln W_2(E_2) \frac{\partial (E_{1+2} - E_1)}{\partial E_1} \\ &= \frac{\partial}{\partial E_1} \ln W_1(E_1) - \frac{\partial}{\partial E_2} \ln W_2(E_2) \\ &= \beta(E_1) - \beta(E_2) \ , \end{aligned} \tag{2.46}$$

where β is a new function of energy and now the relation between two subsystems is

$$\beta(E_1) = \beta(E_2) \ . \tag{2.47}$$

At thermal equilibrium there is only temperature between subsystems that will be the same (for canonical ensemble).

Therefore, we assume that $\beta = \frac{1}{k_B T}$, where k_B is Boltzmann constant. Then we obtain

$$\begin{aligned} \frac{1}{k_B T} &= \frac{\partial}{\partial E_1} \ln W_1(E_1) \\ \frac{1}{T} &= \frac{\partial}{\partial E_1} k_B \ln W_1(E_1) \ . \end{aligned} \tag{2.48}$$

There exists a function $S(E, V)$ such that

$$\frac{1}{T} = \frac{\partial S}{\partial E}. \tag{2.49}$$

Hence, from Equation (2.48) and (2.49), we now see that

$$S = k_B \ln W_1(E_1) \ . \tag{2.50}$$

We note here that one can start with $\beta(E_2)$ and will obtain Equation (2.50) as well. The quantity is called **Boltzmann entropy** (statistic entropy)

$$S_B \;=\; k_B \ln W \;, \tag{2.51}$$

where W is number of possible microstates of the system that we are interested in. To see the behaviour of the system through the Boltzmann entropy, let us consider the situation in Figure 6(a). Initially, the subsystem 1 contains three energy bunches (indistinguishable objects) and the subsystem 2 contains one energy bunch. For the system 1, we can say that we have three balls and four boxes. The number of ways to choose the balls with repetition is given by

$$\begin{aligned} \bar{C}_{n,k} &= \left(\binom{n}{k}\right) \\ &= \binom{n+k-1}{k} \;. \end{aligned} \tag{2.52}$$

This is the number of k-element combinations of n objects. Surely, each possible of combination is microstate of system and k is now defined to be number of balls and n is defined to be number of boxes. Then we can compute number of microstates for system 1 as

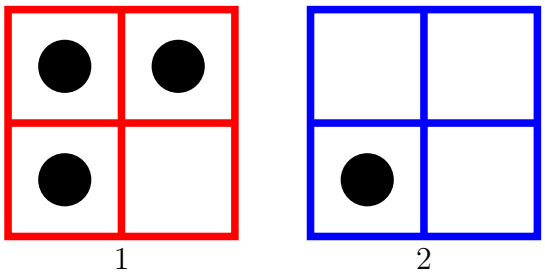
$$\begin{aligned} W_1 \;\equiv\; \bar{C}_{4,3} &= \binom{4+3-1}{3} \\ &= \binom{6}{3} \\ &= \frac{6!}{3!(6-3)!} \\ &= 20 \;. \end{aligned} \tag{2.53}$$

Also, we can compute number of microstates for the system 2 with the same process, where we have $n = 4$ and $k = 1$. This gives us $W_2 = 4$. Then Boltzmann entropy of whole system at the initial configuration is

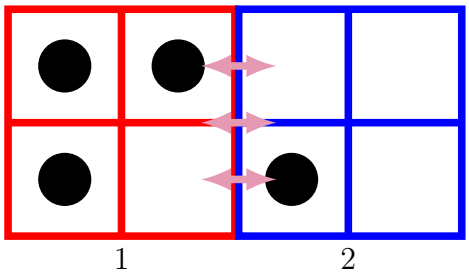
$$S_B^i \;=\; k_B \ln W_{12} \;. \tag{2.54}$$

Actually, we also know that the system which contains subsystem 1 and 2 should has number of mirostates as in Equation (2.42) (multiple form). Here, Equation (2.54) becomes

$$\begin{aligned} S_B^i &= k_B \ln W_1 W_2 \\ &= k_B \ln(20)(4) \\ &= k_B \ln 80 \, . \end{aligned} \tag{2.55}$$



(a)



(b)

Figure 6 The subsystem A and B are containing three energy bunches and one energy bunch, respectively, (a). After, they are come to together, it will be the system that contain four energy bunches (b).

Later, let 1 and 2 contact each other allowing energy bunches to move across the subsystems, see Figure 6(b). We find that the maximum entropy of the whole system is attained if each subsystem contains two energy bunch, see Figure 7.

Then the final entropy of the whole system is

$$\begin{aligned} S_B^f &= k_B \ln W_1 W_2 \\ &= k_B \ln(10)(10) \\ &= k_B \ln 100 \ . \end{aligned} \tag{2.56}$$

Therefore, the entropy change is

$$S_B^f - S_B^i = k_B \ln(10)(10) - k_B \ln(20)(4) > 0 \ , \tag{2.57}$$

which agrees with the second law of thermodynamics.

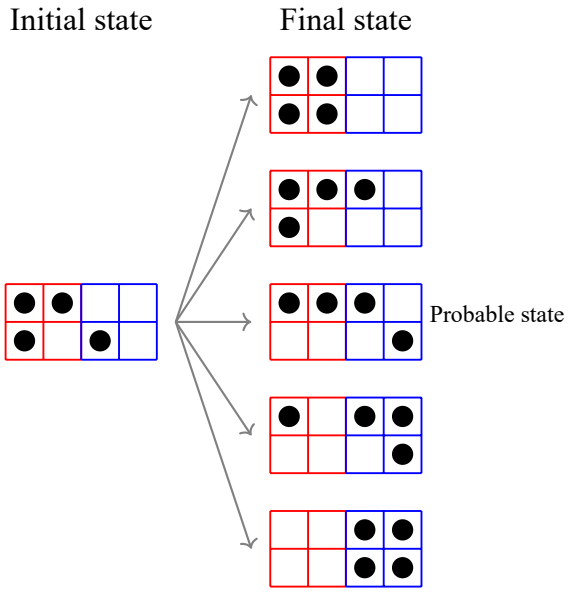


Figure 7 There are five configuration of energy bunch at final state.

Here we conclude that the statistical entropy probabilistically describes the system in the same way as the thermodynamic entropy which is the entropy of an isolated system cannot decrease with time.

So far, we treat all possible microstates on the same equal footing. This means that the probability distribution is uniform for the whole ensemble. Then if the probability distribution is not uniform, the Boltzmann entropy is no longer applicable.

However, one could employ another notion of entropy

$$S_G = -k_B \sum_{i=1}^{W} p_i \ln p_i \tag{2.58}$$

which was introduced by Josiah W Gibbs. In the case that if all microstates are equally likely $p_i = 1/W$, we obtain

$$S_G = -k_B \sum_{i=1}^{W} \frac{1}{W} \ln \frac{1}{W} = k_B W \left(\frac{1}{W} \ln \frac{1}{W} \right) = k_B \ln W, \tag{2.59}$$

which is actually the Boltzmann entropy. Then we can state that the Boltzmann entropy is the upper bound of the Gibbs entropy $S_B \geq S_G$.

2.2.3 Shannon entropy

In this section, we will introduce another type of entropy known as Shannon entropy. The origin of this entropy is nothing to do with what we have mentioned previously, namely thermodynamics and statistical mechanics, but rather from the communication.

According to Shannon, the communication is composed of 3 fundamental parts, a sender, a communication channel and a receiver. The quest is that how the receiver can identify what data is sent by the sender over the channel. Shannon demonstrated that the entropy represents an absolute mathematical limit on how well the data from the sender can be compressed onto a perfectly noiseless channel.

Recall a random variable $X = (x_1, x_2, x_3, ..., x_N)$ and a set of associated probabilities of each outcome $P = (p_1, p_2, p_3, ..., p_N)$ [1]. We now define a quantity that implies amount of surprise for i^{th}-outcome x_i as $1/p_i$. The meaning of this quantity can be interpreted as follows. We consider a coin flip experiment. There are two outcomes resulting in the random variable $X = \{x_H, x_T\}$: head and tail. If the head outcome x_H turns up every single time of flipping coin implying 100% chance associated with this particular

[1] Normally, it should be $P_X(x_i)$, but we neglect subscription X and x_i as p_i to be convenient

outcome known as maximally biased coin, the amount of surprise is 1, which is the minimum value, implying that there is nothing to be surprised with the head outcome since it turns up every single time of the experiment. In the case that there are 50% chance for outcome x_H and 50% chance for outcome x_T known as unbiased coin, the amount of surprise is 2 for each outcome, which is higher than the previous case. This means that you have to guess the outcome of the experiment for every single time. You could pick the head outcome x_H to be the one to turn up, but it might actually be x_T to turn up, with 50% chance. Of course, you are surprised with the outcome because it is not what you expected. From these two situations, we find that the more improbable a particular outcome is, the more surprised we are to observe it. A question is now how can we measure the amount of surprise properly? Here we find that if we choose logarithms to the base 2 (binary system) then the amount of surprise for each outcome is quantified in bits

$$\text{Shannon information} \equiv \mathfrak{I}_i = \log_2 \frac{1}{p_i} \,, \quad \text{[bits]} \tag{2.60}$$

which is called the Shannon information and also called surprisal. In Figure 8, the Shannon information increases logarithmically with decreasing the probability. This means that more improbable of the outcome is, the more Shannon information.

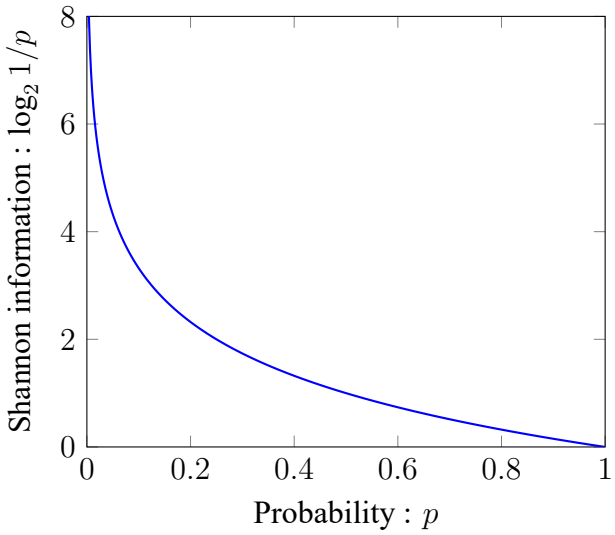


Figure 8 Shannon information related with probability of x.

Of course, in principle, the system would possess an ensemble of many possible outcomes with a particular probability distribution. This also means that we will have a corresponding ensemble of surprisals: $\{\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_N\}$. We then could look for an (linear) average of the surprisal

$$< \mathcal{I} >= \sum_{i=1}^{N} p_i \mathcal{I}_i = \sum_{i=1}^{N} p_i \log_2 \frac{1}{p_i} \ . \tag{2.61}$$

Rearranging form a bit, we obtain

$$H(p_1, p_2, ..., p_N) \equiv < \mathcal{I}_X >= -\sum_{i=1}^{N} p_i \log_2 p_i \ , \tag{2.62}$$

which is known as the Shannon entropy[2]. To see what we could say about the Shannon entropy, we again recall the flipping coin experiment. In the case of unbiased coin, we have $\{p_H = 0.5, p_T = 0.5\}$ and

$$H(p_H, p_T) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = 1 \ \text{ bit} \ . \tag{2.63}$$

This is of course the average of the Shannon information or the amount of information to be extracted from the experiment is 1 bit.

In the case of biased coin, the Shannon entropy is always less than 1 bit and the minimum value of the Shannon entropy is 0 for the maximally biased coin: either $p_H = 1$ or $p_T = 1$, see Figure 9.

In general, the maximum entropy is attained when the probability distribution of all N outcomes is fair: $\{p_i = 1/N \ , \forall i\}$

$$H_{max} = H(1/N, 1/N, ..., 1/N) = \log_2 N \quad \text{bits} \ . \tag{2.64}$$

Here is an interesting feature. For the Shannon entropy H, the variable X can be represented by $m = 2^H$ equiprobable values. In the case of fair coin, we find that $m = 2^1 = 2$. This means that we can assign two different digits in binary system to each outcomes such that $\{x_H = 0, x_T = 1\}$.

[2]The Shannon entropy is the same form with Equation (2.58), but with out Boltzmann constant K_B

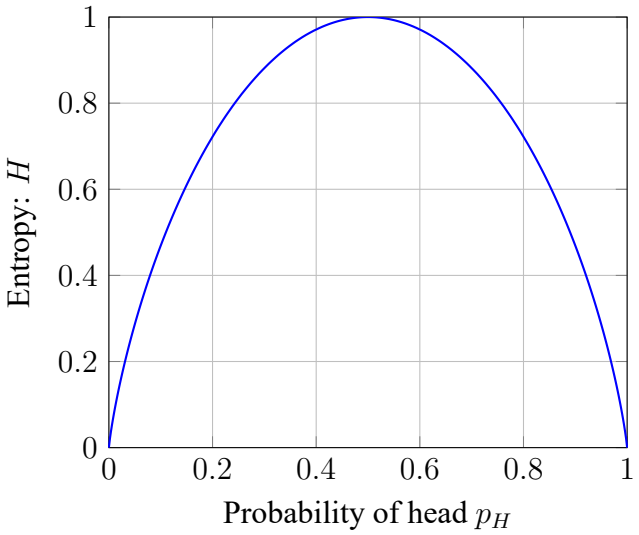


Figure 9 Shannon entropy versus the probability of getting head p_H, which is calculated by using Equation (2.62), and we automatically know that $p_T = 1 - p_H$.

In the case of unfair coin with $H = 0.469$, the number of equiprobable values is $2^{0.469} = 1.38$. This number does not look so natural comparing with the previous case. However, this way of transforming Shannon entropy to the number of equiprobable values is quite natural to associate amount of information with the variable X.

In case of continuous probability distribution, we could also define the entropy. Let X be a random variable with probability density function $f(x)$ whose domain is a set Ω. We define

$$H = -\int_{\Omega} f(x) \log f(x) dx \ , \tag{2.65}$$

which is called the differential entropy. However, the differential entropy is unlike discrete entropy becuase it can be negative. For example, we consider a normal distribution given as,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ \left(-\frac{1}{2} \left(\frac{x}{\sigma} \right)^2 \right) \right\} \ , \tag{2.66}$$

where a random variable X with $f(x)$ has zero mean ($\mu = 0$). The differential entropy

becomes

$$\begin{aligned} H &= -\int f(x)\left(\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{x^2}{2\sigma^2}\ln(e)\right)dx \\ &= \frac{1}{2}\ln\big(2\pi\sigma^2\big) + \frac{\ln e}{2\sigma^2} < X^2 > . \end{aligned} \tag{2.67}$$

Using relation $\sigma^2 =< X^2 > -\cancel{< X >^2}^{0}$ then we obtain

$$\begin{aligned} H &= \frac{1}{2}\ln\big(2\pi\sigma^2\big) + \frac{1}{2} \\ &= \frac{1}{2}\ln\big(2\pi\sigma^2\big) + \ln(e)\frac{1}{2} \\ &= \frac{1}{2}\ln\big(2e\pi\sigma^2\big). \end{aligned} \tag{2.68}$$

Equation (2.68) implies that differential entropy can be negative when $2e\pi\sigma^2$ takes the value less than 1.

2.2.4 Kullback-Leibler divergence (relative entropy)

Normally, Kullback-Leibler divergence is just a modification of the Shannon entropy. This divergence is given by

$$D(p \parallel q) = \sum_{i=1}^{N} p(x_i)\left(\log p(x_i) - \log q(x_i)\right). \tag{2.69}$$

What we see from Equation (2.69) is that it is just the expectation of the logarithm difference between the probability of data in the original distribution $p(x)$ and the approximating distribution $q(x)$. If we consider for bit unit of information, we can interpret (2.69) as "how many bits of information we expect to lose", because there will be some lost of information when we badly choose $q(x)$ to approximate the true distribution $p(x)$. For example, suppose we obtain experimental data from original distribution $p(x)$ and we expect that it might be either uniform or binomial distribution being the original distribution. We assume that Equation (2.69) yield

$$D(p \parallel \text{Binomial}) = 0.3 \;\text{ and }\; D(p \parallel \text{Uniform}) = 0.5\,. \tag{2.70}$$

We can see that the information is lost when we use the uniform approximation is greater than the binomial approximation. This means that if we have to choose some functions to

represent our observations, it is better to deal with the binomial approximation. Basically, Kullback-Leibler divergence can be written as

$$D(p \parallel q) = \sum_{i=1}^{n} p(x_i) \log \frac{p(x_i)}{q(x_i)}, \tag{2.71}$$

for the discrete case and for the continuous case, it is

$$D(f \parallel g) = \int_{\Omega} f(x) \log \frac{f(x)}{g(x)} dx. \tag{2.72}$$

We can see that if $p(X)$(or $f_X(x)$) and $q(X)$(or $g_X(x)$) are the same, D will be zero. This means that there are no difference or no distance between them. One important fact is that the KL-divergence is not a true measure of distance

$$D(p \parallel q) \neq D(q \parallel p), \tag{2.73}$$

since it is not symmetric under the commutation of the argument.

However, KL-divergence can be applied to study the probability distribution in the geometry context.

Here, the Riemann manifolds will be replaced by the statistical manifolds whose points correspond to probability distributions. To see this, let's first consider the point $P = (p^1, p^2, ..., p^n)$ and $P + dP = (p^1 + dp^1, p^2 + dp^2, ..., p^n + dp^n)$ for discrete random variable on the n-dimensional statistical manifold, see figure 10. These two points can be treated as two different probability distributions and we can use the KL-divergence to quantify the difference

$$D(P \parallel P + dP) = \sum_{i=1}^{n} p^i \ln \left(\frac{p^i}{p^i + dp^i} \right) .$$

If dp^i are infinitesimal, we find that

$$D(P \parallel P + dP) \cong \frac{1}{2} \sum_{i=1}^{n} \frac{dp^i dp^i}{p^i} .$$

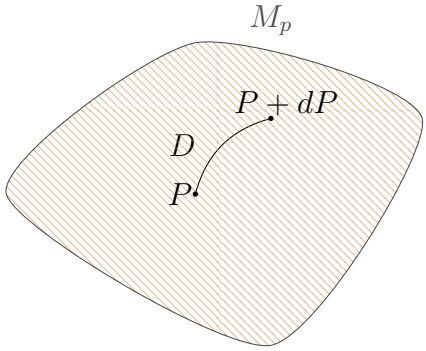


Figure 10 Two different points P and $P + dP$ on manifold M_p.

If we define $g_{ij} = \frac{1}{2}\frac{\delta_{ij}}{p^i}$, we then have

$$D(P \parallel P + dP) = \sum_{i=1}^{n}\sum_{j=1}^{n} g_{ij} dp^i dp^j \ . \tag{2.74}$$

With the present form of the Equation (2.74), one can treat the KL-divergence as the square of an infinitesimal line element ds^2 or "interval" between point P and $P + dP$. Then, of course, g_{ij} is the metric tensor, known as Fisher-Rao matrix, associated with the statistical manifold. We note here that the asymmetric property of the KL-divergence does not affect the lowest term in the expansion of Equation (2.74). Under the coordinates transformation $p^i \Rightarrow \theta^i = \theta^i(p^1, p^2, ..., p^n)$, of course, there exists an inverse transformation such that $p^i(\theta^1, \theta^2, ..., \theta^n)$. With this transformation, a new statistical manifold whose points correspond to different probability distributions $\Theta = (\theta^1, \theta^2, ..., \theta^n)$.

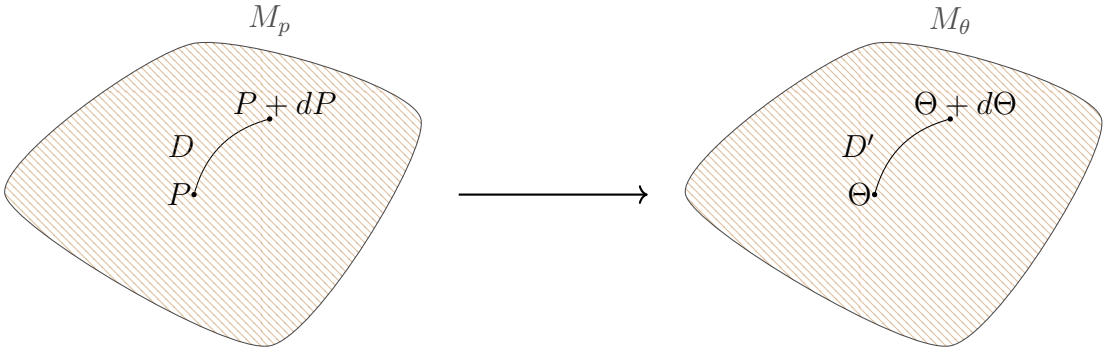


Figure 11 The probability manifold under the coordinate transformation.

The metric is transformed as follows

$$\begin{aligned} g_{ab} &= \sum_{i}^{n}\sum_{j}^{N}\frac{\partial p^i}{\partial\theta^a}\frac{\partial p^j}{\partial\theta^b}g_{ij} \\ &= \frac{1}{4}\sum_{i}^{N}\sum_{j}^{N}\frac{\partial_a p^i\partial_b p^j}{p^i}\delta_{ij}\ , \end{aligned} \tag{2.75}$$

where $\frac{\partial}{\partial\theta^a}\equiv\partial_a$ and $\frac{\partial}{\partial\theta^b}\equiv\partial_b$. A new metric can be further simplifed

$$\begin{aligned} g_{ab} &= \frac{1}{4}\sum_{i}^{N}\frac{\partial_a p^i\partial_b p^i}{p^i} \\ &= \frac{1}{4}\sum_{i}^{N}\frac{\partial_a p^i}{p^i}\frac{\partial_b p^i}{p^i}p^i \\ &= \frac{1}{4}\sum_{i}^{N}p^i\partial_a\ln p^i\partial_b\ln p^i \\ &= \frac{1}{4}\langle\partial_a\ln p^i\partial_b\ln p^i\rangle. \end{aligned} \tag{2.76}$$

We would point out that the term with the bracket is the Fisher information[3].

In addition, if we consider

$$-\frac{1}{2}\frac{\partial^2 H}{\partial p^i\partial p^j}=\frac{1}{2}\frac{\delta_{ij}}{p^i}=g_{ij}, \tag{2.77}$$

we may treat this as the connection between the Fisher information and the Shannon entropy.

[3]The derivation of the Fisher information will come up later.

2.2.5 Generalised entropies

Fadeev [23] proposed the postulates which can be used to characterize entropy (the discrete case is interested here) as follows

(a) $H(p_1, p_2, ..., p_n)$ is a symmetric function

(b) $H(p_1, p-1)$ is a continuous function of p for $0 \leqslant p \leqslant 1$.

(c) $H(1/2, 1/2) = 1$.

(d) $H(tp_1, (1-t)p_1, p_2, ..., p_n) = H(p_1, p_2, ..., p_n) + p_1 H(t, 1-t)$ for any distributions $P = (p_1, p_2, ..., p_n)$ and for $0 \leqslant t \leqslant 1$.

Here in this section, we will pay attention to Tsallis entropy and Rényi entropy since they are relevant for our context on generalising the Fisher information.

2.2.5.1 Rényi entropy.

Alfréd Rényi introduced a new quantity which is called Rényi entropy [5], through generalised probability distributions. To see what he did, let us consider $[\Omega, \mathfrak{B}, \mathcal{P}]$ be a probability space, where Ω is the elements of events (sample space), $\mathfrak{B}$ is a σ- algebra of subsets of Ω, elements of $\mathfrak{B}$ being events and $\mathcal{P}$ is probability which $\mathcal{P}(\Omega) = 1$. Then he considered function $\xi = \xi(\omega)$ for $\omega \in \Omega_1$ where $\Omega_1 \in \mathfrak{B}$ and $\mathcal{P}(\Omega_1) > 0$. Now, if $\mathcal{P}(\Omega_1) = 0$, ξ is called a complete random variable. While in the case $0 < \mathcal{P}(\Omega_1) < 1$, ξ is called incomplete random variable. What we are interested is a particular case such that ξ takes on a finite different values $x_1, x_2, ..., x_n$, the existence of distribution can be written as $p_k = \mathcal{P}(\xi = x_k)$ for $k = 1, 2, ..., n$.

Here, the generalised probability can written as

$$\mathcal{W}(P) \;=\; \sum_{k=1}^{n} p_k \;, \tag{2.78}$$

where $P \equiv (p_1, p_2, ..., p_n)$. We shall call $\mathcal{W}(P)$ as the weight of distribution with

$$0 < \mathcal{W}(P) \leqslant 1 \;. \tag{2.79}$$

The weight is called a complete distribution if $\mathcal{W}(P)$ is equal to 1, while the weight is

less than 1 will be called incomplete distribution. Next, Rényi characterised the entropy with a generalised probability distribution by given five postulates

(1) $H[P]$ is a symmetric function of the elements of P.

(2) $H[\{p\}]$ is a continuous function of p in the interval $0 < p < 1$, if p is defined as the single probability.

(3) $H[1/2] = 1$.

(4) Let Δ denote the set of all finite discrete generalised probability distributions. For $P \in \Delta$ and $Q \in \Delta$ we have

$$H[(P * Q)] = H[P] + H[Q] \ , \tag{2.80}$$

where $Q = (q_1, q_2, ..., q_m)$ is other set of probability. If we denote P and Q as two generalised distributions such that $\mathcal{W}(P) + \mathcal{W}(Q) \leqslant 1$, where the union of set being as

$$P \cup Q = (p_1, p_2, ..., p_n, q_1, q_2, ..., q_m) \ , \tag{2.81}$$

and, of course, $P \cup Q$ is not defined when $\mathcal{W}(P) + \mathcal{W}(Q) > 1$.
The last one may be called the mean-value property of entropy.

(5) If P and $Q \in \Delta$, one obtains

$$H[(P \cup Q)] = \frac{\mathcal{W}(P)H[P] + \mathcal{W}(Q)H[Q]}{\mathcal{W}(P) + \mathcal{W}(Q)} \ , \tag{2.82}$$

for $\mathcal{W}(P) + \mathcal{W}(Q) \leqslant 1$. An advantage of this way on defining the entropy from the generalised distributions is that this mean-value property is much simpler in the general case. Next, let's define

$$H[\{p\}] = h(p) \ , \tag{2.83}$$

where the probability p exists on $0 < p \leqslant 1$ and should be continous in this interval by postulate 2. The postulate 3 also gives us that $h(1/2) = 1$. Moreover, if we have another probability q with $0 < q \leqslant 1$, the postulate 4 gives

$$h(pq) = h(p) + h(q) \ . \tag{2.84}$$

Thus, it follows that

$$H[\{p\}] = h(p) = \log 1/p \tag{2.85}$$

the $h(p)$ must be the logarithm function of p.

Rényi asked a question that what he will get if he replaced postulate 5 by some other mean value. In general, weighted mean value of the numbers $x_1, x_2, ..., x_n$ with the weight probability $w_1, w_2, ..., w_n$ for $w_k > 0$ and $\sum_{k=1}^{n} w_k = 1$, can be written in the form

$$g^{-1}\left[\sum_{k=1}^{n} w_k g(x_k)\right] \tag{2.86}$$

where g^{-1} is an inverse function of $g(x)$ which is strictly an arbitrary monotonic and continuous function.

(5′) With this reason, the new postulate is introduced

$$H[(P \cup Q)] = g^{-1}\left[\frac{w(P)g(H[P]) + w(Q)g(H[Q])}{w(P) + w(Q)}\right]. \tag{2.87}$$

It can be seen clearly that, if $g(x) = ax + b$ with $a \neq 0$, then postulate 5′ reduces to 5. Another choice is

$$g_\alpha(x) = 2^{(\alpha-1)x}\ , \tag{2.88}$$

where $\alpha > 0$. Then postulate 1,2,3 and 4 characterise the entropy of order α. Obviously, postulate 5′ gives a new entropy by choosing $g(x) \equiv g_\alpha(x)$ for $P = (p_1, p_2, ..., p_n)$ as

$$H_\alpha[P] = \frac{1}{1-\alpha}\log_2\left[\frac{\sum_{k=1}^{n} p_k^\alpha}{\sum_{k=1}^{n} p_k}\right]\ . \tag{2.89}$$

This will be called the entropy of order α of generalised distribution P. For the complete distribution $\sum_{k=1}^{n} p_k = 1$, one obtains

$$H_\alpha[P] = \frac{1}{1-\alpha}\log_2\left[\sum_{k=1}^{n} p_k^\alpha\right], \tag{2.90}$$

which is known as the Rényi entropy. It is not difficult to see that under the limit $\alpha \to 1$, the Shannon entropy $H_1[P] = H[P] = -\sum_{k=1}^{n} p_k \log_2 p_k$ can be recovered.

In fact, Renyi entropy of different orders, with different values of them, are needed to describe the uncertainty [18]. To show this claim, we recall the definition of the surprisal $I_i = -\log p_i$ and rewrite as follows

$$\begin{aligned} H_\alpha[P] &= -\frac{1}{\alpha-1}\log\left[\sum_{i=1}^{n} p_i \exp[(\alpha-1)\log p_i)]\right] \\ &= -\frac{1}{\alpha-1}\log\left[\sum_{i=1}^{n} p_i \exp[(\alpha-1)(-I_i)]\right] \\ &= -\frac{1}{\alpha-1}\log\langle\exp[(\alpha-1)(-I)]\rangle\ . \end{aligned} \tag{2.91}$$

Equation (2.91) is identical to the effective values

$$X_\alpha = \sum_{n=1}^{\infty}\frac{\kappa_n(X)}{n!}(\alpha-1)^{n-1} = \frac{1}{\alpha-1}\log\langle\exp[(\alpha-1)\cdot X]\rangle \tag{2.92}$$

where the coefficients $\kappa_n(X)$ are called the cumulant. This implies that the negative Renyi entropies are the effective values of the negative surprisal (X being replaced by $-I$). Effectively, Equation (2.91) becomes

$$-H_\alpha = \sum_{n=1}^{\infty}\frac{\kappa_n(-s)}{n!}(\alpha-1)^{n-1} = \frac{1}{\alpha-1}\log\langle\exp[(\alpha-1)\cdot(-s)]\rangle\ . \tag{2.93}$$

For $n = 1$, the cumulant $\kappa_1 = \langle -s\rangle = \langle\log p_i\rangle = \sum_i p_i \log p_i$ is nothing but the Shannon entropy. Then we get

$$-H_\alpha = \sum_i p_i \log p_i + \sum_{n=2}^{\infty}\frac{\kappa_n(-s)}{n!}(\alpha-1)^{n-1}\ . \tag{2.94}$$

The second term represents the fluctuation in the uncertainty. Again, by taking limit $\alpha \to 1$, the Shannon entropy is trivially recovered from Equation (2.94).

2.2.5.2 Tsallis entropy.

Coming from different perspective, namely on the statistical mechanics, on generalising the entropy, Tsallis gave a one-parameter generalisation of the Boltzmann-Gibbs entropy [7]. To illustrate the method, let us consider a differential equation

$$\frac{dy}{dx} = a + b(y), \tag{2.95}$$

where a and b are parameters and condition $y(0) = 1$. From Equation (2.95), we consider three different possible situations.

The first equation is

$$\frac{dy}{dx} = 0 \ , \tag{2.96}$$

where a and b are set to be zero. Then its solution is trivially constant $y = 1$ and whose symmetric curve with regard to the bisector axis is $x = 1$.

The second simplest differential equation would be

$$\frac{dy}{dx} = 1. \tag{2.97}$$

Its solution is clearly

$$y = 1 + x, \tag{2.98}$$

and we know that whose inverse function is

$$y = x - 1. \tag{2.99}$$

The last one which we will consider is

$$\frac{dy}{dx} = y \ . \tag{2.100}$$

It can be explicitly seen that its solution is the exponential

$$y = e^x, \tag{2.101}$$

and its inverse function is

$$y = \ln x \ . \tag{2.102}$$

Equation (2.102) satisfies condition

$$\ln(x_A x_B) = \ln(x_A) + \ln(x_B). \tag{2.103}$$

Here, if we multiply constant K_B on both side of Equation (2.102) and replace x as a number of microstate W. We might obtain thermodynamics quantity

$$S_B = K_B \ln W \ , \tag{2.104}$$

which is Boltzmann entropy. This, of course, satisfy additive condition (2.103) as well. In addition, it is possible to unify the same three differential equations with only one parameter by considering

$$\frac{dy}{dx} = y^q, \tag{2.105}$$

with condition that $y(0) = 1$ and $q \in \mathbb{R}$. Here q is a index. Then we can find that its solution is

$$y = [1 + (1-q)\, x]^{1/(1-q)} \equiv e_q^x, \tag{2.106}$$

which is called that the Tsallis's q-exponential and there exists inverse function as

$$y = \frac{x^{1-q} - 1}{1-q} \equiv \ln_q x, \tag{2.107}$$

which is called the Tsallis's q-logarithm. Of course, if we take limit of $q \to 1$ these two q-analogues will become the original ones. It is not difficult to see that

$$\ln_q(x_A x_B) = \ln_q(x_A) + \ln_q(x_B) + (1-q) \ln_q(x_A) \ln_q(x_B). \tag{2.108}$$

Next, we can introduce the generalised Boltzmann entropy by using definition of q-logarithm

$$S_q = K_B \ln_q W = K_B \frac{W^{1-q} - 1}{1-q} \ . \tag{2.109}$$

Equation (2.109) can be rewritten as

$$\begin{aligned} S_q &= K_B \frac{\sum_{i=1}^{W} (\frac{1}{W})^q - 1}{1-q} \\ &= K_B \frac{1 - \sum_{i=1}^{W} p_i^q}{q-1}. \end{aligned} \tag{2.110}$$

Again, the Boltzmann-Gibbs entropy can be recovered by considering the limit $q \to 1$.

2.2.6 Fisher information

The new quantity can also be used to measure disorder of system such that Fisher information which is defined from maximum likelihood estimation (MLE). With random variable $X(= x_1, x_2, x_3, ..., x_N)$ and independent and identically distribution (i.i.d) of each outcome $f(x_i \mid \theta)$, we can consider likelihood as

$$L(\theta \mid X) = \prod_{i=1}^{N} f(x_i \mid \theta) \, , \tag{2.111}$$

where θ is arbitrary parameter in probability models. In principle, the likelihood measures how good the statistical model is comparing to the sample of data X for given the values of the unknown parameter θ such that at maximum value of likelihood data X are existing at the true probability models, defined through parameter θ, itself. Then we need to find maximum value of likelihood which means that

$$\frac{\partial}{\partial \theta} L(\theta \mid X) = 0 \, . \tag{2.112}$$

But, likelihood is often fussy on calculation, we introduce log-likelihood to solve the problem becauses log-likelihood is monotonic increasing function which means they have the same maximum point

$$\frac{\partial}{\partial \theta} \log L(\theta \mid X) = 0 \, . \tag{2.113}$$

We also know that derivative of log-likelihood function in Equation (2.113) is called Score function. It is obviously that the curvature of the log-likelihood function around its maximum can be used as indicator for how good it is for the estimated value: If the log-likelihood is quite narrow around the maximum we are fairly certain on the estimated value, otherwise if the log-likelihood is broad we are uncertain about the estimate. Therefore, we can consider second derivative of log-likelihood to determine curvature of log-likelihood which is averaged for all possible random variable

$$I(\theta) = \left\langle \frac{\partial^2}{\partial \theta^2} \log L(\theta \mid X) \right\rangle \, , \tag{2.114}$$

because this quantity implies concavity and convexity of function and if absolute of value is very hight its mean that function is quite sharp and easy to estimated. Thus, we define this quatity as the measurement of accuration from data which is called Fisher information. By the fact that

$$\frac{\partial}{\partial\theta}\log L(\theta \mid X) = \frac{1}{L(\theta \mid X)}\frac{\partial L(\theta \mid X)}{\partial\theta}, \tag{2.115}$$

and the first derivative of log-likelihood will be alway zero

$$\left\langle \frac{\partial}{\partial\theta}\log L(\theta \mid X)\right\rangle = 0\,, \tag{2.116}$$

Equation (2.114) can be treated as

$$\begin{aligned}\left\langle \frac{\partial^2}{\partial\theta^2}\log L(\theta \mid X)\right\rangle &= \left\langle \frac{\partial}{\partial\theta}\left(\frac{\frac{\partial}{\partial\theta}L(\theta \mid X)}{L(\theta \mid X)}\right)\right\rangle \\ &= \left\langle \frac{L(\theta \mid X)\frac{\partial^2}{\partial\theta^2}L(\theta \mid X) - \left(\frac{\partial}{\partial\theta}L(\theta \mid X)\right)^2}{L^2(\theta \mid X)}\right\rangle \\ &= \left\langle \frac{\frac{\partial^2}{\partial\theta^2}L(\theta \mid X)}{L(\theta \mid X)}\right\rangle - \left\langle \left(\frac{\partial}{\partial\theta}\log L(\theta \mid X)\right)^2\right\rangle . \end{aligned} \tag{2.117}$$

For now on, we might neglect subscription Ω on integrating to be convenient. The first term of Equation (2.117) can be obviously seen that it will be zero,

$$\begin{aligned}\left\langle \frac{\frac{\partial^2}{\partial\theta^2}L(\theta \mid X)}{L(\theta \mid X)}\right\rangle &= \int ... \int \frac{\frac{\partial^2}{\partial\theta^2}L(\theta \mid X)}{L(\theta \mid X)}L(\theta \mid X)\prod_{i=1}^{n} dx_i \\ &= \int ... \int \frac{\partial^2}{\partial\theta^2}L(\theta \mid X)\prod_{i=1}^{n} dx_i \\ &= \frac{\partial^2}{\partial\theta^2}\int ... \int L(\theta \mid X)\prod_{i=1}^{n} dx_i \\ &= \frac{\partial^2}{\partial\theta^2}(1) \\ &= 0\,. \end{aligned} \tag{2.118}$$

Therefore, Equation (2.117) can be rewritten as

$$\begin{aligned}\left\langle \frac{\partial^2}{\partial\theta^2}\log L(\theta \mid X)\right\rangle &= -\left\langle \left(\frac{\partial}{\partial\theta}\log L(\theta \mid X)\right)^2\right\rangle \\ &= -Var\left[\frac{\partial}{\partial\theta}\log L(\theta \mid X)\right] - \left\langle \frac{\partial}{\partial\theta}\log L(\theta \mid X)\right\rangle^2 . \end{aligned}$$

Using Equation (2.116), the last term vanishes, then we obtain

$$\left\langle \frac{\partial^2}{\partial\theta^2} \log L(\theta \mid X) \right\rangle \;\; = \;\; -Var\left[\frac{\partial}{\partial\theta} \log L(\theta \mid X) \right] .$$

Here, we can conclude that Fisher information tells us how much we know about the internal structure from data. With a given space of outcomes Ω, Fisher information is often defined by

$$I(\theta) \equiv \left\langle \left(\frac{\partial}{\partial\theta} \log L(\theta \mid X) \right)^2 \right\rangle = Var\left[\frac{\partial}{\partial\theta} \log L(\theta \mid X) \right] = -\left\langle \frac{\partial^2}{\partial\theta^2} \log L(\theta \mid X) \right\rangle .$$

In general, the estimated parameters could come in a set i.e., $\theta = (\theta_1, \theta_2, ..., \theta_n)$. Then the Fisher information becomes $I(\theta) = [I_{ij}(\theta)]$, where

$$I_{ij}(\theta) = -\left\langle \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log L(\theta|X) \right\rangle \; , \tag{2.119}$$

which is known as the Fisher information matrix.

To explain more clearly about Fisher information, we will show a famous example which is tossing coin. We are interested to toss the coin 10 times: $N = 10$. We observe a sequence of heads and tails which is $H, H, H, T, H, T, T, H, T, H$. If we denote heads by 1 and tails by 0, the **data** can be coded as

$$X = \{1, 1, 1, 0, 1, 0, 0, 1, 0, 1\} \; . \tag{2.120}$$

So in this experiment, head turns up 6 times. Before considering $N = 10$, we actual know that what kind of this probability model for n time tossing.

Model: We see that outcomes are independent to each other. Then the Bernoulli distribution is an appropriate model to describe the probability of observing heads for any single flip and parameter for this case is probability of getting head, that is $\theta \equiv p_H$

$$p(x_i \mid p_H) \;\; = \;\; p_H^{x_i}(1 - p_H)^{1-x_i}, \qquad x_i = \{0, 1\}. \tag{2.121}$$

Criterion: Of course, the criterion that will be used to estimate the probability associated with the heads. The likelihood function is given by

$$\begin{aligned} L(p_H \mid X) &= \prod_{i=1}^{N} p(x_i \mid p_H) \\ &= p_H^{\sum_{i=1}^{N} x_i} (1-p_H)^{N-\sum_{i=1}^{N} x_i} , \end{aligned} \tag{2.122}$$

What we need is to look at the condition for maximal likelihood function

$$\frac{d}{dp_H} L(p_H \mid X) = 0 . \tag{2.123}$$

At this point, it is useful of we consider the logarithm function of the likelihood. Then we work with the addition rather than multiplication as we mention earlier. So the log-likelihood is given by

$$\log L(p_H \mid X) = \left(\sum_{i=1}^{N} x_i \right) \log p_H + \left(n - \sum_{i=1}^{N} x_i \right) \log(1-p_H) . \tag{2.124}$$

We find that

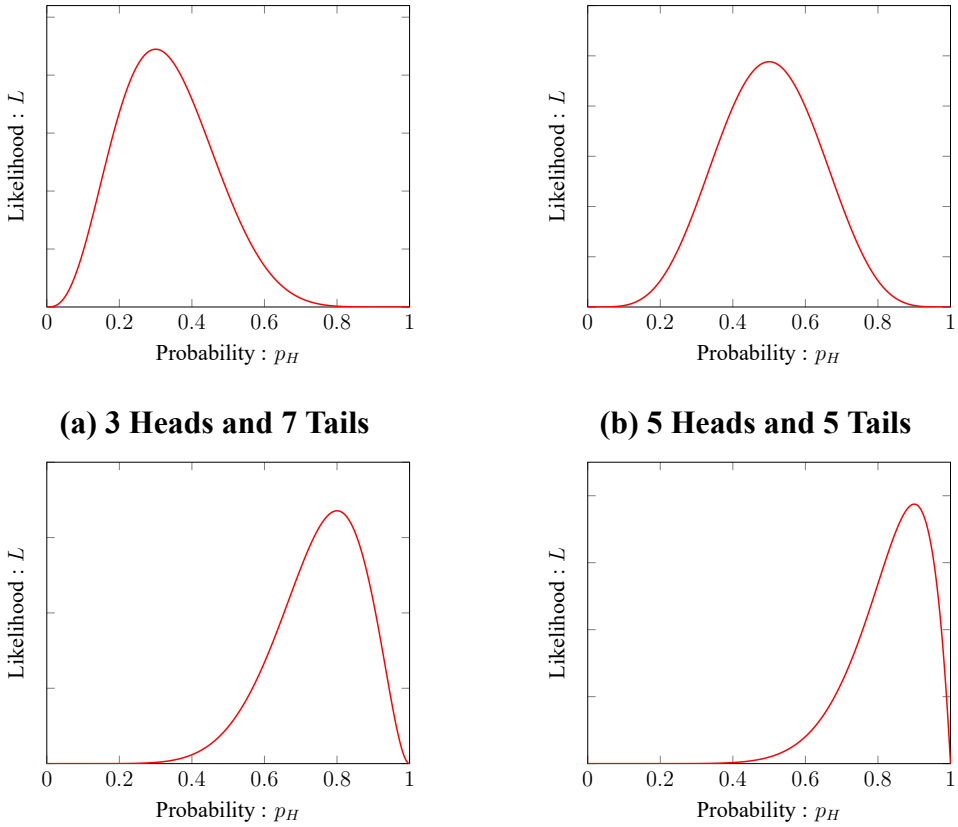
$$\frac{d}{dp_H} \log L(p_H \mid X) = 0 , \tag{2.125}$$

resulting in

$$p_H^* = \frac{\sum_{i=1}^{N} x_i}{N} . \tag{2.126}$$

Which suggests that the probability p_H is just proportion of number of head outcomes in the experiment. The Figure 12 shows the various outcomes for $N = 10$ tossing coin. We observe that for 3 heads and 7 tails were the outcome, the likelihood function reaches the maximum at $p_H = 0.3 = 3/10$.

The question is now how accurate the estimate is. According to law of large number, we would prefer a large number of data. Figure 13 shows that increase in number of flips gives decrease in the width of the distribution resulting in a better estimation. Then it suggests the curvature of the likelihood function around its maximum can be used as indicator for how good it is for the estimated value.



(a) 3 Heads and 7 Tails

(b) 5 Heads and 5 Tails

(c) 8 Heads and 2 Tails

(d) 9 Heads and 1 Tail

Figure 12 The likelihood function for several different possible outcomes for $n = 10$ flips of a coin.

If the likelihood is quite narrow around the maximum we are fairly certain on the estimated value, otherwise if the likelihood is broad we are uncertain about the estimate. We now can compute the score function

$$\text{Score} \equiv \frac{\partial}{\partial p_H} \log L(p_H \mid X) = \frac{x}{p_H} - \frac{N - x}{1 - p_H} \,, \tag{2.127}$$

where $x = \sum_{i=1}^{N} x_i$.

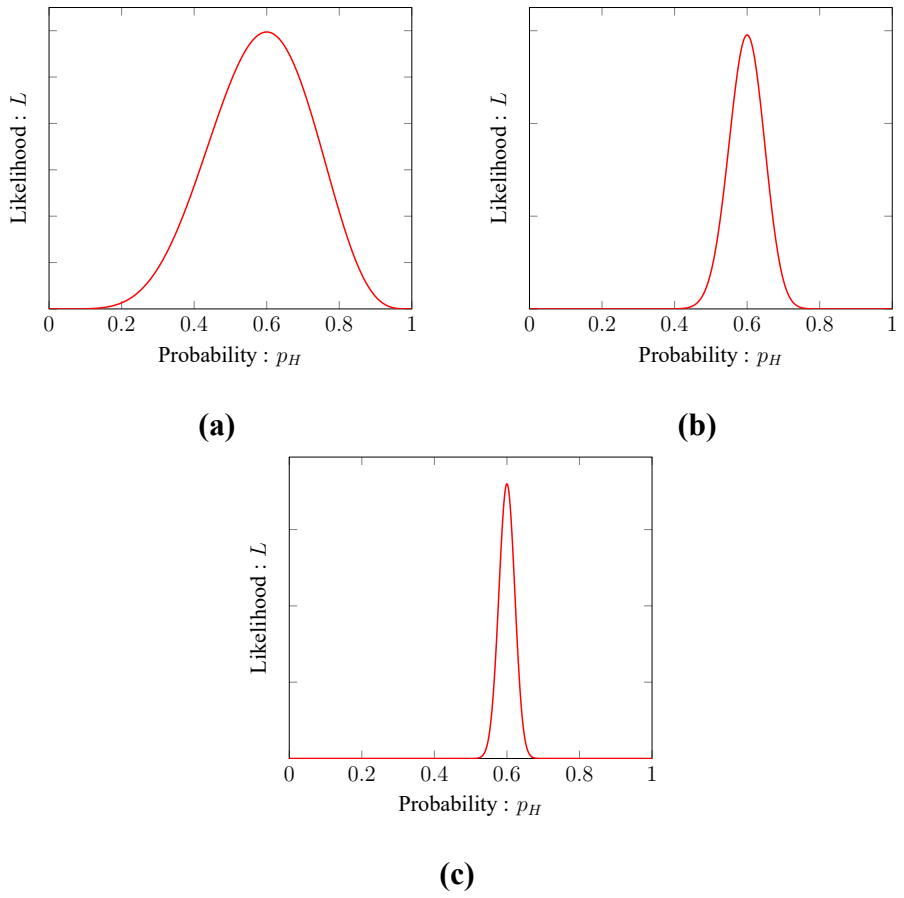


Figure 13 (a): The likelihood function for the case if 6 heads in 10 flips. (b): The likelihood function for 60 heads in 100 flips. (c): The likelihood function for 300 heads in 500 flips.

It is not difficult to see evident that, on average the score (first moment) is zero

$$\begin{aligned}\left\langle \frac{\partial}{\partial p_H} \log L(p_H \mid X) \right\rangle &= \sum_{x=0}^{N} \frac{\partial}{\partial p_H} \log L \binom{N}{x} p_H^x (1-p_H)^{1-x} \\ &= \sum_{x=0}^{N} \left(\frac{x}{p_H} - \frac{N-x}{1-p_H} \right) \binom{N}{x} p_H^x (1-p_H)^{1-x} \\ &= \frac{N p_H}{p_H} - \frac{N(1-p_H)}{1-p_H} = 0 \,. \end{aligned} \tag{2.128}$$

Next, we consider the average of the square score (the second moment)

$$\begin{aligned}
\left\langle \left(\frac{\partial}{\partial p_H} \log L(p_H \mid X) \right)^2 \right\rangle &= \sum_{x=0}^{N} \left(\frac{\partial}{\partial p_H} \log L \right)^2 \binom{N}{x} p_H^x (1-p_H)^{1-x} \\
&= \sum_{x=0}^{N} \left(\frac{x}{p_H} - \frac{N-x}{1-p_H} \right)^2 \binom{N}{x} p_H^x (1-p_H)^{1-x} \\
&= \frac{N}{p_H(1-p_H)} = -\left\langle \frac{\partial^2}{\partial p_H^2} \log L(p_H \mid X) \right\rangle \\
&= \mathrm{Var}\left(\frac{\partial}{\partial p_H} \log L(p_H \mid X) \right) \ , \qquad (2.129)
\end{aligned}$$

which gives the variance of the score. Then for a single flip, the variance is $1/p_H(1-p_H)$ which can be visualised in Figure 14. Therefore, the variance is proportional to the number n of trials, large n implying large variance as well as large negative expected second derivative of the log-likelihood function.

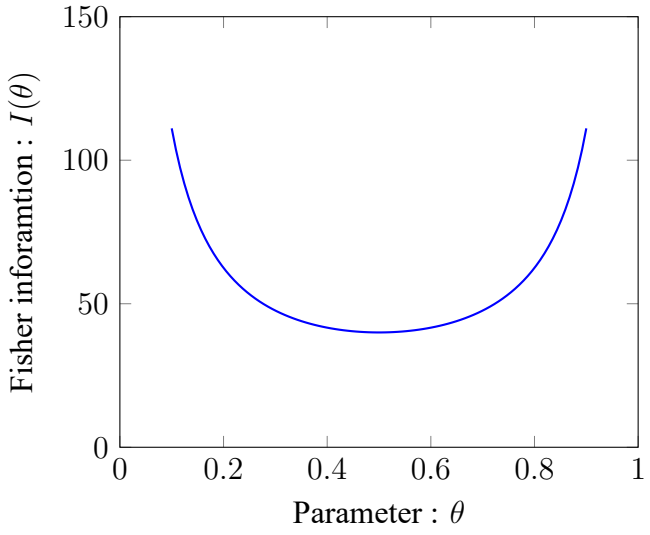


Figure 14 The variance as a function $\theta = p_H$ within the Bernoulli model. As θ reaches zero or one the variance goes to infinity. If $p_H = 1$ the outcomes will always be 1, therefore clearly conveying this information within the data .

Fisher information also provides a inforamtion lower bound on the variance of an unbiased estimator for a parameter which is called Cramér-Rao inequality. Noramlly, it is obtianed from considering unbiases estimator

$$B(\hat{\Theta}) = \left\langle \hat{\Theta} - \theta \right\rangle = \int_{\Omega} ... \int_{\Omega} (\hat{\Theta} - \theta) L(\theta \mid X) \prod_{i=1}^{N} dx_i = 0, \tag{2.130}$$

where we already know that θ is unknow parameter and define $\hat{\Theta} = h(x_1, x_2, ..., x_n)$ as point estimator.

Next, if we now consider its derivative respect to parameter θ, we will have

$$\begin{aligned} \frac{\partial}{\partial \theta} \left\langle \hat{\Theta} - \theta \right\rangle &= -\int ... \int L(\theta \mid X) \prod_{i=1}^{N} dx_i \\ &+ \int ... \int (\hat{\Theta} - \theta) \frac{\partial}{\partial \theta} L(\theta \mid X) \prod_{i=1}^{N} dx_i \ , \end{aligned} \tag{2.131}$$

using the fact that $\int ... \int L(\theta \mid X) \prod_{i=1}^{N} dx_i = 1$ and we obtain

$$\begin{aligned} \int ... \int (\hat{\Theta} - \theta) \left(\frac{\partial}{\partial \theta} \log L(\theta \mid X) \right) L(\theta \mid X) \prod_{i=1}^{N} dx_i &= 1 \\ \int ... \int \left[(\hat{\Theta} - \theta) \cdot L^{1/2}(\theta \mid X) \right] \left[\left(\frac{\partial}{\partial \theta} \log L(\theta \mid X) \right) \cdot L^{1/2}(\theta \mid X) \right] \prod_{i=1}^{N} dx_i &= 1 \ . \end{aligned}$$

Applying Cauchy–Schwarz to above equation, we get

$$\begin{aligned} \frac{1}{\int ... \int (\hat{\Theta} - \theta)^2 L(\theta \mid X) \prod_{i=1}^{N} dx_i} &\leq \int ... \int \left(\frac{\partial}{\partial \theta} \log L(\theta \mid X) \right)^2 L(\theta \mid X) \prod_{i=1}^{N} dx_i \\ \frac{1}{Var(\hat{\Theta})} &\leq I(\theta) \ , \end{aligned} \tag{2.132}$$

which is called Cramér-Rao inequality.

Furthermore, there is also one more important feature, as same as Boltzmann-Gibbs (2.103), Shannon and Rényi entropy (2.80), of the Fisher information known as the additive property. From the right hand side of the inequality (2.132), we will see that

$$\begin{aligned}
I(\theta) =& \int ... \int \left(\frac{\partial}{\partial \theta} \log \prod_{i=1}^{N} f(x_i \mid \theta) \right)^2 \prod_{i=1}^{N} f(x_i \mid \theta) \prod_{i=1}^{N} dx_i \\
=& \int ... \int \left(\sum_{i=1}^{N} \frac{1}{f(x_i \mid \theta)} \frac{\partial f(x_i \mid \theta)}{\partial \theta} \right)^2 \prod_{i=1}^{N} f(x_i \mid \theta) \prod_{i=1}^{N} dx_i \\
=& \int ... \int \Bigg(\sum_{\substack{i,j=1 \\ i \neq j}}^{N} \frac{1}{f(x_i \mid \theta) f(x_j \mid \theta)} \frac{\partial f(x_i \mid \theta)}{\partial \theta} \frac{\partial f(x_j \mid \theta)}{\partial \theta} \\
&+ \sum_{j=1}^{N} \frac{1}{f^2(x_j \mid \theta)} \left(\frac{\partial f(x_j \mid \theta)}{\partial \theta} \right)^2 \Bigg) \prod_{k=1}^{N} f(x_k \mid \theta) \prod_{k=1}^{N} dx_k \\
=& \int ... \int \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \frac{1}{f(x_i \mid \theta) f(x_j \mid \theta)} \frac{\partial f(x_i \mid \theta)}{\partial \theta} \frac{\partial f(x_j \mid \theta)}{\partial \theta} \prod_{i=k}^{N} f(x_k \mid \theta) \prod_{i=k}^{N} dx_k \\
&+ \int ... \int \sum_{j=1}^{N} \frac{1}{f^2(x_j \mid \theta)} \left(\frac{\partial f(x_j \mid \theta)}{\partial \theta} \right)^2 \prod_{k=1}^{N} f(x_k \mid \theta) \prod_{k=1}^{N} dx_k \\
=& F_1 + F_2. \qquad (2.133)
\end{aligned}$$

Next, let us first consider F_1 term. The probabilities $f(x_k \mid \theta)$ for $k \neq i$ or j integrate through as simply factors 1, by normalization. The remaining probabilities in $\prod_{k=1}^{N} f(x_k \mid \theta)$ and term $\prod_{k=1}^{N} dx_k$ will be only $f(x_i \mid \theta)$ and $f(x_j \mid \theta)$ and just $f(x_j \mid \theta)$ for F_2.

Then result is

$$I(\theta) = \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \int \int \frac{\partial f(x_i \mid \theta)}{\partial \theta} \frac{\partial f(x_j \mid \theta)}{\partial \theta} dx_i dx_j + \sum_{j=1}^{N} \int \frac{1}{f(x_j \mid \theta)} \left(\frac{\partial f(x_j \mid \theta)}{\partial \theta} \right)^2 dx_j$$

With the reason that we are dealing with identical and independent random variable,

this will be simplified as

$$
\begin{aligned}
I(\theta) &= \sum_{\substack{i,j=1 \\ i\neq j}}^{N} \int \frac{\partial f(x_i \mid \theta)}{\partial \theta} dx_i \int \frac{\partial f(x_j \mid \theta)}{\partial \theta} dx_j + \sum_{j=1}^{N} \int \frac{1}{f(x_j \mid \theta)} \left(\frac{\partial f(x_j \mid \theta)}{\partial \theta} \right)^2 dx_j \\
&= \sum_{\substack{i,j=1 \\ i\neq j}}^{N} \frac{\partial}{\partial \theta} \int f(x_i \mid \theta) dx_i \frac{\partial}{\partial \theta} \int f(x_j \mid \theta) dx_j + \sum_{j=1}^{N} \int \frac{1}{f(x_j \mid \theta)} \left(\frac{\partial f(x_j \mid \theta)}{\partial \theta} \right)^2 dx_j \\
&= \sum_{j=1}^{N} \int \frac{1}{f(x_j \mid \theta)} \left(\frac{\partial f(x_j \mid \theta)}{\partial \theta} \right)^2 dx_j \\
&= \sum_{j=1}^{N} \int \left(\frac{\partial}{\partial \theta} \log f(x_j \mid \theta) \right)^2 f(x_j \mid \theta) dx_j \\
&= \sum_{j=1}^{N} I_i(\theta). \qquad (2.134)
\end{aligned}
$$

Particularly, the Fisher matrix (2.119) can also be obtained by considering the relative entropy or Kullback-Leibler divergence between two distribution $P = (f_1, f_2, ..., f_N)$ and $Q = (q_1, q_2, ..., q_N)$ on the probability manifold. Then the Kullback-Leibler divergence between two probability distributions $L(\theta|X)$ and $L(\theta'|X)$, parametised by θ, is given by

$$
\begin{aligned}
D(\theta, \theta') &\equiv KL(L(\theta|X)||L(\theta'|X)) \\
&= \int ... \int L(\theta|X) \log \left(\frac{L(\theta|X)}{L(\theta'|X)} \right) \prod_{i=1}^{N} dx_i \ , \qquad (2.135)
\end{aligned}
$$

where likelihood is what we was already defined in Equation (2.111). For θ being fixed, the Kullback-Leibler divergence can be expanded around θ as

$$
D(\theta, \theta') = \frac{1}{2} (\theta' - \theta)^{\mathrm{T}} \left(\frac{\partial^2}{\partial \theta'_i \partial \theta'_j} D(\theta, \theta') \right) \Big|_{\theta=\theta'} (\theta' - \theta) + \mathcal{O}\left((\theta' - \theta)^2 \right) \ , \qquad (2.136)
$$

where the second order derivative is

$$
\begin{aligned}
\left(\frac{\partial^2}{\partial \theta'_i \partial \theta'_j} D(\theta, \theta') \right) \Big|_{\theta=\theta'} &= -\int ... \int \left(\frac{\partial^2}{\partial \theta'_i \partial \theta'_j} L(\theta'|X) \right) \Big|_{\theta'=\theta} L(\theta|X) \prod_{i=1}^{N} dx_i \\
&= [I_{ij}(\theta)] \ . \qquad (2.137)
\end{aligned}
$$

With this connection, one may intuitively interpret the Fisher information as the metric between two point on the probability manifold.

However, the Kullback-Leibler divergenece is not symmetric and does not follow the triangle inequality [22]. Then the Fisher information cannot be treated as a true metric.

CHAPTER III

ONE-PARAMETER EXTENDED FISHER INFORMATION

In this chapter, we will derive a new type of the Fisher information called one parameter generalised Fisher information. Then, we will construct the generalised one of Cramér-Rao inequality and show that Fisher information hierarchy is non-additive quantity. Furthermore, we will give the relation between Fisher information hierarchy and the two-parameters Kullback–Leibler divergence. We also find that the standard discrete Kullback–Leibler divergence gives a relation between Fisher information hierarchy and Shannon entropy by considering higher rank tensor metric.

3.1 Least action principle and Fisher information

We first would like to give a short review on the least action principle. Let $S[q]$ given by

$$S[q] = \int_a^b \mathscr{L}(q'(t), q(t), t)dt \ , \text{ where } \ q'(t) = \frac{dq(t)}{dt} \ , \tag{3.1}$$

be an action functional defined on the configuration space $\Sigma : q = (q_1, q_2, .., q_n)$ with dimension n. Here $\mathscr{L}$ is a Lagrangian defined

$$\mathscr{L}(q'(t), q(t), t) = T(q'(t)) - V(q(t)) \ ,$$

where T is the kinetic energy and V is the potential energy. The action $S[q]$ will take its extremal value for particular function $q_0(t)$. This means that under an infinitesimal variation $q(t, \epsilon) = q_0(t) + \epsilon\eta(t)$, where $\eta(a) = \eta(b) = 0$, see Figure 15, the action remains the same $\delta S[q] = 0$.

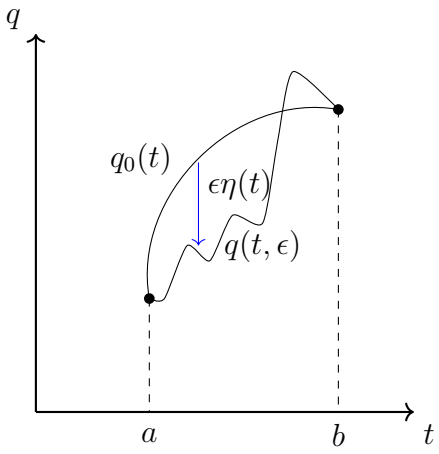


Figure 15 Small variation $q(t,\epsilon)$ of $q_0(t)$ between the endpoint a, b.

Now the new action is given by

$$S[q] \to S(\epsilon) = \int_a^b \mathscr{L}(q'(t,\epsilon), q(t,\epsilon), t)dt \ , \tag{3.2}$$

which depends on the parameter ϵ. It takes the extremal value

$$\begin{aligned} 0 = \frac{\partial S}{\partial \epsilon} &= \frac{\partial}{\partial \epsilon}\int_a^b \mathscr{L}(q'(t,\epsilon), q(t,\epsilon), t)dt \\ &= \int_a^b \left[\frac{\partial \mathscr{L}}{\partial q}\frac{\partial q}{\partial \epsilon} + \frac{\partial \mathscr{L}}{\partial q'}\frac{\partial q'}{\partial \epsilon}\right]dt. \end{aligned} \tag{3.3}$$

Integrating by parts the second term in the bracket, we obtain

$$\begin{aligned} 0 &= \int_a^b \left[\frac{\partial \mathscr{L}}{\partial q}\frac{\partial q}{\partial \epsilon} + \frac{\partial \mathscr{L}}{\partial q'}\frac{\partial q'}{\partial \epsilon}\right]dt \\ &= \int_a^b \left[\frac{\partial \mathscr{L}}{\partial q}\frac{\partial q}{\partial \epsilon} - \frac{d}{dt}\left(\frac{\partial \mathscr{L}}{\partial q'}\right)\frac{\partial q}{\partial \epsilon}\right]dt + \cancel{\frac{\partial \mathscr{L}}{\partial q'}\frac{\partial q}{\partial \epsilon}}\Big|_a^{\cancel{0}\,b} \\ &= \int_a^b \left[\frac{\partial \mathscr{L}}{\partial q} - \frac{d}{dt}\frac{\partial \mathscr{L}}{\partial q'}\right]\eta(t)dt. \end{aligned} \tag{3.4}$$

Since $\eta(t)$ is arbitrary, this means that

$$\frac{\partial \mathscr{L}}{\partial q} - \frac{d}{dt}\frac{\partial \mathscr{L}}{\partial q'} = 0 \ , \tag{3.5}$$

which is known as the Euler-Lagrange equation.

The intriguing connection between the variational principle and Fisher information was proposed by Frieden [16]. Let us now define θ as actual value of a measurement quantity, $X = (x_1, x_2, .., x_N)$ as N outcomes of the quantity and $Y = (y_1, y_2, ..., y_N)$ as random errors associated with each measurement. Then we have

$$x_i = \theta + y_i \ . \tag{3.6}$$

Next, let $f(x_i|\theta)$ be a probability distribution, whose support is a set Ω, over the x's with respect to θ. We recall the Fisher information [4]

$$I(\theta) = \sum_{i=1}^{N} \int \left(\frac{\partial}{\partial \theta} \ln f(x_i \mid \theta) \right)^2 f(x_i \mid \theta) dx_i \tag{3.7}$$

In the case $N = 1$, we do have $f(x \mid \theta) = f(x - \theta) = f(y)$. The Fisher information is simply reduced to

$$I[f(y)] = \int f(y) \left(\frac{d}{dy} \ln f(y) \right)^2 dy = \int \frac{(f'(y))^2}{f(y)} dy \quad \text{when} \quad f'(y) = \frac{df(y)}{dy}. \tag{3.8}$$

Next, we do a transformation such that $q(y) = \sqrt{f(y)}$ resulting in

$$I[q(y)] = 4 \int q'^2(y) dy, \ \ \text{when} \ \ q'(y) = \frac{dq(y)}{dy}. \tag{3.9}$$

We find that the Fisher information is now a functional with the input function $q(y)$. Here comes to an interesting point. If we define $I[q] \equiv S[q]$ as an action functional and $\mathscr{L}(q', q; y) \equiv 4q'^2(y)$ as the Lagrangian, the variational principle would give the Euler-Lagrange equation

$$\frac{\partial \mathscr{L}(q', q; y)}{\partial q(y)} - \frac{d}{dy} \left(\frac{\partial \mathscr{L}(q', q; y)}{\partial q'(y)} \right) = 0, \tag{3.10}$$

resulting in $-8q''(y) = 0$. This second order differential equation describes how a position q change with time y for a free particle. This would mean that the Fisher information (3.9) could be remarkably treated as the action functional for the free particle and of course, in the absence of the interaction, the equation of motion in physics is a direct result of extramising the Fisher information: $\delta I[q] = 0$.

[4] Here we prefer the natural logarithm function.

3.2 One-parameter extended Fisher information

Here in this section, we will employ the connection between the Fisher information and the action functional together with the non-uniqueness property of the Lagrangian to construct a one-parameter generalisation of the Fisher information. Commonly, two Lagrangians differing by the total derivative with respect to time of some function $F(q, y)$ would give the identical equation of motion on extreamising the action. However, one could ask an inverse question as follows. Imposing the equation of motion, could we solve all possible Lagrangians directly from the Euler-Lagrange equation? The answer is definitely yes and this problem is known as the inverse problem of the calculus of variations [17]. Recently, Sarawuttinack et al [21] proposed a new type of Lagrangian called the multiplicative form for the case of one degree of freedom. Here we shall employ the same technique and propose an alternative Lagrangian for $\mathscr{L}(q', q; y) = 4q'^2(y)$ as

$$\mathscr{L}_\lambda(q', q; y) = \frac{4}{\lambda}\left(e^{\lambda q'^2(y)} - 1\right) \ . \tag{3.11}$$

Of course, the Lagrangian $\mathscr{L}_\lambda(q', q; y)$ can be treated as one-parameter extended class of the standard Lagrangian $\mathscr{L}(q', q; y)$. It is not difficult to see that these two Lagrangians give exactly the same equation of motion. By considering the limit $\lambda \to 0$, one find that $\lim_{\lambda\to 0}\mathscr{L}_\lambda = \mathscr{L}(q', q; y) = 4q'^2(y)$. Then what we have is the action functional in the form

$$I_\lambda[q(y)] = \frac{4}{\lambda}\int\left(e^{\lambda q'^2(y)} - 1\right)dy \ . \tag{3.12}$$

We shall call Equation (3.12) as a one parameter generalised Fisher information. The reason can be seen as follows. If we expand the functional (3.12) with respect to the parameter λ, we obtain

$$\begin{aligned} I_\lambda[q(y)] &= 4\int q'^2(y)dy + 4\frac{\lambda}{2!}\int q'^4(y)dy + 4\frac{\lambda^2}{3!}\int q'^6(y)dy + ... \\ &= I_1[q(y)] + \frac{\lambda}{2!}I_2[q(y)] + \frac{\lambda^2}{3!}I_3[q(y)] + ... \ . \end{aligned} \tag{3.13}$$

What we have in Equation (3.13) is a hierarchy $\{I_1, I_2, I_3, ...\}$, where the first three are

$$
\begin{aligned}
I_1 &= 4\int\left(\frac{f'(y)}{2q(y)}\right)^2 dy = \frac{4}{2^2}\int\frac{f'^2(y)}{f(y)}dy = \frac{4}{2^2}\int\left(\frac{\partial}{\partial\theta}\ln f(x|\theta)\right)^2 f(x|\theta)dx\ , \\
I_2 &= 4\int\left(\frac{f'(y)}{2q(y)}\right)^4 dy = \frac{4}{2^4}\int\frac{f'^4(y)}{f^2(y)}dy = \frac{4}{2^4}\int\left(\frac{\partial}{\partial\theta}\ln f(x|\theta)\right)^4 f^2(x|\theta)dx\ , \\
I_3 &= 4\int\left(\frac{f'(y)}{2q(y)}\right)^6 dy = \frac{4}{2^6}\int\frac{f'^6(y)}{f^3(y)}dy = \frac{4}{2^6}\int\left(\frac{\partial}{\partial\theta}\ln f(x|\theta)\right)^6 f^3(x|\theta)dx\ .
\end{aligned}
$$

The first term is nothing but the standard Fisher information $I_1[q] = I[q]$ coinciding with the limit $\lim_{\lambda\to 0} I_\lambda[q(y)] = I[q(y)]$. With the above structure, one could easily write $I_n[f(y)]$ as

$$
I_n(\theta) = \frac{4}{2^{2n}}\int\left(\frac{\partial}{\partial\theta}\ln f(x|\theta)\right)^{2n} f^n(x|\theta)dx\ ,\quad n = 1, 2, 3, ... \tag{3.15}
$$

The next point is that the generalised Fisher information is in the average of the score function but the rest in the hierarchy is not. Then we shall seek a transformation to express the higher order Fisher information in the statistical average. We shall first consider the second function I_2 and introduce a new variable $\phi_1 \equiv f^2$ such that $f'(y) = \phi_1'(y)/2f(y)$, resulting in

$$
I_2[\phi_1] = \frac{4}{4^4}\int\frac{\phi_1'^4(y)}{\phi_1^4(y)}\phi_1(y)dy\ ,
$$

or

$$
I_2[\theta] = \frac{4}{4^4}\int\left[\frac{\partial}{\partial\theta}\ln\phi_1(x|\theta)\right]^4\phi_1(x|\theta)dx\ . \tag{3.16}
$$

We shall call Equation (3.16) as the 2nd order Fisher information. We can proceed the same technique of transformation and obtain the n^{th} order Fisher information as

$$
I_n(\theta) = \frac{4}{(2n)^{2n}}\int\left[\frac{\partial}{\partial\theta}\ln\phi_{n-1}(x|\theta)\right]^{2n}\phi_{n-1}(x|\theta)dx, \tag{3.17}
$$

where $\phi_{n-1}(y) = f^n(y)$ and the generalised Fisher information (3.13) can be expressed in terms of infinite series as

$$
I_\lambda(\theta) = \sum_{n=1}^{\infty}\frac{\lambda^{n-1}}{n!}I_n(\theta)\ . \tag{3.18}
$$

At this point, we may treat Equation (3.13) as the generating function for the entire hierarchy of the Fisher information by expanding with respect to the parameter λ.

3.3 Generalised Cramér-Rao inequality and non-additive property

Here in this section, we will construct the Cramér-Rao inequality associated with the Fisher information hierarchy given in the previous section. To achieve the goal, we start with

$$\left\langle \hat{\Theta} - \theta \right\rangle = \int (\hat{\Theta} - \theta) f^q(x \mid \theta) dx = 0 \ , \tag{3.19}$$

which is known as the q-expectation value [24–26] . Taking the 1st derivative, we obtain

$$\begin{aligned} \frac{\partial}{\partial\theta}\left\langle \hat{\Theta} - \theta \right\rangle &= \int \frac{\partial}{\partial\theta}(\hat{\Theta} - \theta) f^q(x \mid \theta) dx + \int (\hat{\Theta} - \theta) \frac{\partial}{\partial\theta} f^q(x \mid \theta) dx \\ &= -\int f^q(x \mid \theta) dx + q \int (\hat{\Theta} - \theta) f^{q-1}(x \mid \theta) \frac{\partial f(x \mid \theta)}{\partial\theta} dx \\ &= -\int f^q(x \mid \theta) dx + q \int (\hat{\Theta} - \theta) f^{q-1}(x \mid \theta) f(x \mid \theta) \frac{\partial \ln f(x \mid \theta)}{\partial\theta} dx \\ &= -Q_q + qJ = 0 \ , \end{aligned} \tag{3.20}$$

where

$$Q_q \equiv \int f^q(x \mid \theta) dx \ , \tag{3.21}$$

$$J \equiv \int (\hat{\Theta} - \theta) f^{q-1}(x \mid \theta) f(x \mid \theta) \frac{\partial \ln f(x \mid \theta)}{\partial\theta} dx \ . \tag{3.22}$$

Conventionally, the term Q_q is well know as information generating function [26] with Tsallis index q [6](we are now interested in case $q \geq 1$). It is also called incomplete normalization and f^q is called effective probability [25]. Next, we rewrite the J in the form

$$J = \int \left[(\hat{\Theta} - \theta) \right] \left[\frac{\partial \ln f(x \mid \theta)}{\partial\theta} f^{q-1}(x \mid \theta) \right] f(x \mid \theta) dx \ , \tag{3.23}$$

and applying the Hölder's inequality [27] to Equation (3.23), we obtain

$$\begin{aligned} J &\leq \left[\int (\hat{\Theta} - \theta)^\beta f(x \mid \theta) dx \right]^{1/\beta} \left[\int \left(\frac{\partial \ln f(x \mid \theta)}{\partial\theta} \right)^\alpha (f^{q-1}(x \mid \theta))^\alpha f(x \mid \theta) dx \right]^{1/\alpha} \\ &= \left[\int (\hat{\Theta} - \theta)^\beta f(x \mid \theta) dx \right]^{1/\beta} \left[\int \left(\frac{\partial \ln f(x \mid \theta)}{\partial\theta} \right)^\alpha f^{\alpha(q-1)+1}(x \mid \theta) dx \right]^{1/\alpha} , \end{aligned} \tag{3.24}$$

where Hölder conjugates α and β are related with the condition $1/\alpha + 1/\beta = 1$ for $\alpha, \beta = [1, \infty]$. Finally, employing Equation (3.20), the inequality (3.24) becomes

$$\begin{aligned} \frac{Q_q}{q} &= \frac{\int f^q(x \mid \theta) dx}{q} \leq \left[\int (\hat{\Theta} - \theta)^\beta f(x \mid \theta) dx \right]^{1/\beta} \\ &\times \left[\int \left(\frac{\partial \ln f(x \mid \theta)}{\partial \theta} \right)^\alpha f^{\alpha(q-1)+1}(x \mid \theta) dx \right]^{1/\alpha} , \end{aligned} \tag{3.25}$$

which is our generalised Carmer-Rao inequality. It is not difficult to see that if one takes $q = 1$, $\beta = 2$ and $\alpha = 2$, the standard Carmer-Rao inequality can be recovered. For $\alpha = 4$, $\beta = 4/3$ and $q = 5/4$, we obtain

$$\frac{4\, Q_{5/4}}{5^4 \left\langle (\hat{\Theta} - \theta)^{4/3} \right\rangle^3} \leq I_2 , \tag{3.26}$$

which is the Cramér-Rao inequality for the 2nd extended Fisher information. Basically, the inequality (3.25) provides the Cramér-Rao bound for the whole Fisher information hierarchy as shown in table 1.

Table 1 The n^{nd} Carmer-Rao inequalities and their associated three parameters.

n^{nd} Carmer-Rao inequality	Parameters		
	q	β	α
1st order	1	2	2
2nd order	5/4	4/3	4
3rd order	4/3	6/5	6
4th order	11/8	8/7	8

Next, we will investigate the additive property of the higher order Fisher information. For simplicity, we shall start with the 2nd order Fisher information. Suppose a system composed of two independent identically subsystems that are defined its random

variable $X = (x_1, x_2)$, where superscription denote for subsystems. The joint probability of the two subsystems is given by $f_{12} \equiv f(x_1, x_2|\theta) = f(x_1|\theta)f(x_2|\theta) \equiv f_1 f_2$. What we have for the 2nd order Fisher information is

$$\begin{aligned}
I_2[f_{12}] &= \frac{4}{2^4}\int\int\left(\frac{\partial}{\partial\theta}\ln(f_1 f_2)\right)^4 f_1^2 f_2^2 dx_1 dx_2 \\
&= \frac{4}{2^4}\Bigg[\int\left(\frac{\partial}{\partial\theta}\ln f_1\right)^4 f_1^2 dx_1 \int f_2^2 dx_2 + 4\int\left(\frac{\partial}{\partial\theta}\ln f_1\right)^3 f_1^2 dx_1 \\
&\times \int\left(\frac{\partial}{\partial\theta}\ln f_2\right) f_2^2 dx_2 + 6\int\left(\frac{\partial}{\partial\theta}\ln f_1\right)^2 f_1^2 dx_1 \int\left(\frac{\partial}{\partial\theta}\ln f_2\right)^2 f_2^2 dx_2 \\
&+ 4\int\left(\frac{\partial}{\partial\theta}\ln f_1\right) f_1^2 dx_1 \int\left(\frac{\partial}{\partial\theta}\ln f_2\right)^3 f_2^2 dx_2 \\
&+ \int f_1^2 dx_1 \int\left(\frac{\partial}{\partial\theta}\ln f_2\right)^4 f_2^2 dx_2\Bigg] \\
&= \frac{4}{2^4}\Bigg[Q_2(f_2)\int\left(\frac{\partial}{\partial\theta}\ln f_1\right)^4 f_1^2 dx_1 + 6\int\left(\frac{\partial}{\partial\theta}\ln f_1\right)^2 f_1^2 dx_1 \\
&\times \int\left(\frac{\partial}{\partial\theta}\ln f_2\right)^2 f_1^2 dx_2 + Q_2(f_1)\int\left(\frac{\partial}{\partial\theta}\ln f_2\right)^4 f_2^2 dx_2\Bigg] \\
&= \frac{1}{4}\Big[Q_2(f_2)I_2(f_1) + Q_2(f_1)I_2(f_2) + 6I(f_1)I(f_2)\Big]\,. \qquad (3.27)
\end{aligned}$$

Here see that the 2nd order Fisher information does not follow the additive rule.

With the result in equation (3.27), it is not difficult now to see that the nth order Fisher information could give

$$\begin{aligned}
I_n[f_{12}] =& \frac{4}{2^{2n}}\Bigg[\binom{2n}{0}Q_n(f(x_2|\theta))\int\left(\frac{\partial}{\partial\theta}\ln f(x_1|\theta)\right)^{2n} f^n(x_1|\theta)dx_1 \\
&+ \sum_{k=2}^{2n-2}\binom{2n}{k}\int\left(\frac{\partial}{\partial\theta}\ln f(x_1|\theta)\right)^{2n-k} f^n(x_1|\theta)dx_1 \\
&\times \int\left(\frac{\partial}{\partial\theta}\ln f(x_2|\theta)\right)^k f^n(x_2|\theta)dx_2 \\
&+ \binom{2n}{2n}Q_n(f(x_1|\theta))\int\left(\frac{\partial}{\partial\theta}\ln f(x_2|\theta)\right)^{2n} f^n(x_2|\theta)dx_2\Bigg]\,, \qquad (3.28)
\end{aligned}$$

where the first and last terms refer to the Fisher information for each subsystem and the middle one is the crossing-term. Therefore, our Fisher information hierarchy does not follow the additive property, except for $n = 1$ the standard Fisher information.

3.4 The Kullback–Leibler divergence revisited

Here in this section, we shall investigate on the connection between our Fisher information hierarchy and the Kullback–Leibler divergence. We shall begin with the Kullback–Leibler divergence

$$D(f \parallel q) = \int f(y) \ln \left(\frac{f(y)}{q(y)} \right) dy \ , \tag{3.29}$$

where f and q are two different points on the probability manifold. If $q(y) = f(y+\Delta) = f(y) + \Delta f'(y)$, we could have

$$D(f(y) \parallel f(y) + \Delta f'(y)) = \int f(y) \ln \left(\frac{f(y)}{f(y) + \Delta f'(y)} \right) dy \ , \tag{3.30}$$

where $f'(y) = df/dy$. We then shall expand Equation (3.30) with respect to f'. Keeping only the first dominate term, we obtain

$$D(f(y) \parallel f(y) + \Delta f'(y)) \approx \int \frac{1}{2} \frac{\left(f'(y)\right)^2}{f(y)} dy = \frac{1}{2} I[f(y)] \ . \tag{3.31}$$

We could see that the right hand side of Equation (3.31) is nothing but the standard Fisher information.

Now we introduce two-parameter Kullback-Leibler divergence

$$D_{q,q'} \left(f(y) \parallel f(y) + \Delta f'(y)\right) = \int f^q(y) \left(\ln \frac{f(y)}{f(y) + \Delta f'(y)} \right)^{q'} dy \ . \tag{3.32}$$

Here we do again the expansion with respect to p' and we obtain the two-parameter generalisation of the Fisher information from the first dominant term [8]

$$I_{a,b}[f] = \int f^a(y) \left(\frac{df(y)}{dy} \right)^b dy \ , \tag{3.33}$$

where $a = q - q' - 1$ and $b = q' + 1$ with the requirements $q > 0$ and $q' > 0$. We find that, with a suitable choice of parameters, our whole hierarchy of Fisher information can be identified as shown in the table 2.

Table 2 Comparison our one-parameter Fisher information with two-parameter Fisher information.

n^{th} Fisher information	Parameters	
	a	b
1^{st} order: I_1	1	2
2^{nd} order: I_2	2	4
3^{rd} order: I_3	3	6
4^{th} order: I_4	4	8

We note here that the quantities in Equation (3.33), of course directly connected with our Fisher information hierarchy as we already mentioned, can be possibly viewed as the generalised Fisher matrices. However, there exist also other generalised Fisher matrices for different purposes and motivations [13, 28].

3.5 Connection with the higher rank tensors

We also find that the standard discrete Kullback–Leibler divergence have a relation to Fisher information hierarchy. What we want to look is $D(p_i \parallel p_i + dp_i)$, where dp_i is small. Expanding $D(p_i \parallel p_i + dp_i)$ with respect to dp_i, we obtain

$$D(P \parallel P + dP) = \sum_{i=1}^{n} \left[\frac{1}{2!} \frac{dp^i dp^i}{p^i} - \frac{1}{3!} \frac{dp^i dp^i dp^i}{{p^i}^2} + \frac{1}{4!} \frac{dp^i dp^i dp^i dp^i}{{p^i}^3} + ... \right] . \quad (3.34)$$

We now define

$$g_{ij} = \frac{1}{2!} \frac{\delta_{ij}}{p^i} \quad \text{as the metric tensor rank 2} \quad (3.35)$$

$$u_{ijk} = \frac{1}{3!} \frac{\delta_{ijk}}{{p^i}^2} \quad \text{as the tensor rank 3} \quad (3.36)$$

$$v_{ijkl} = \frac{1}{4!} \frac{\delta_{ijkl}}{{p^i}^3} \quad \text{as the tensor rank 4 .} \quad (3.37)$$

The higher rank tensor can be generated by the same fashion. Then we will see that relative entropy becomes

$$\begin{aligned} D(P \parallel P+dP) &= \sum_{i=1}^{n}\sum_{j=1}^{n} g_{ij}dp^i dp^j + \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} u_{ijk}dp^i dp^j dp^k \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} v_{ijkl}dp^i dp^j dp^k dp^l + \dots . \end{aligned} \tag{3.38}$$

As we already mentioned, the relation between the Shannon entropy and the Fisher-Rao matrix can be obtained through the second derivative of Shannon entropy respect to all related probability

$$-\frac{1}{2}\frac{\partial^2 H}{\partial p^i \partial p^j} = \frac{1}{2!}\frac{\delta_{ij}}{p^i} = g_{ij} \ , \tag{3.39}$$

Here, we push further on order of the derivative and we obtain

$$\frac{1}{3!}\frac{\partial^3 H}{\partial p^i \partial p^j \partial p^k} = \frac{1}{3!}\frac{\delta_{ijk}}{{p^i}^2} = u_{ijk} \quad , \quad -\frac{1}{4!}\frac{\partial^4 H}{\partial p^i \partial p^j \partial p^k \partial p^l} = \frac{1}{4!}\frac{\delta_{ijkl}}{{p^i}^3} = v_{ijkl} \ . \tag{3.40}$$

Now we apply the same trick, as we did in chapter 2, on transforming the coordinates for u_{ijk} and v_{ijkl}

$$\begin{aligned} u_{abc} &= \sum_{i}^{n}\sum_{j}^{n}\sum_{k}^{n} \frac{\partial p^i}{\partial \theta^a}\frac{\partial p^j}{\partial \theta^b}\frac{\partial p^k}{\partial \theta^c} u_{ijk} \\ &= \frac{1}{3!}\sum_{i}^{n}\sum_{j}^{n}\sum_{k}^{n} \frac{\partial p^i}{\partial \theta^a}\frac{\partial p^j}{\partial \theta^b}\frac{\partial p^k}{\partial \theta^c}\frac{1}{{p^i}^2}\delta_{ijk} \\ &= \frac{1}{3!}\sum_{i}^{n} \frac{\partial_a p^i}{p^i}\frac{\partial_b p^i}{p^i}\frac{\partial_c p^i}{p^i} p^i \\ &= \frac{1}{3!}\sum_{i}^{n} p^i \partial_a \ln p^i \partial_b \ln p^i \partial_c \ln p^i \\ &= \frac{1}{3!}\langle \partial_a \ln p^i \partial_b \ln p^i \partial_c \ln p^i \rangle \ , \end{aligned} \tag{3.41}$$

and

$$\begin{aligned} v_{abcd} &= \sum_{i}^{n}\sum_{j}^{n}\sum_{k}^{n}\sum_{l}^{n}\frac{\partial p^i}{\partial\theta^a}\frac{\partial p^j}{\partial\theta^b}\frac{\partial p^k}{\partial\theta^c}\frac{\partial p^l}{\partial\theta^d}v_{ijkl} \\ &= \frac{1}{4!}\sum_{i}^{n}\sum_{j}^{n}\sum_{k}^{n}\sum_{l}^{n}\frac{\partial p^i}{\partial\theta^a}\frac{\partial p^j}{\partial\theta^b}\frac{\partial p^k}{\partial\theta^c}\frac{\partial p^l}{\partial\theta^d}\frac{1}{{p^i}^3}\delta_{ijkl} \\ &= \frac{1}{4!}\sum_{i}^{n}\frac{\partial_a p^i}{p^i}\frac{\partial_b p^i}{p^i}\frac{\partial_c p^i}{p^i}\frac{\partial_d p^i}{p^i}p^i \\ &= \frac{1}{4!}\sum_{i}^{n}p^i\partial_a\ln p^i\partial_b\ln p^i\partial_c\ln p^i\partial_d\ln p^i \\ &= \frac{1}{4!}\langle\partial_a\ln p^i\partial_b\ln p^i\partial_c\ln p^i\partial_d\ln p^i\rangle\ . \end{aligned} \tag{3.42}$$

These two transformed matrix are actually Fisher information in order of skewness and kurtosis. If we consider matric tensor (3.41) and (3.42) for only one parameter θ, we obtain $u = \frac{1}{3!}\langle(\frac{\partial}{\partial\theta}\ln p^i)^3\rangle$ and $v = \frac{1}{4!}\langle(\frac{\partial}{\partial\theta}\ln p^i)^4\rangle$, respectively. What we can see is that $\langle(\frac{\partial}{\partial\theta}\ln p^i)^4\rangle$ is nothing but the 2^{nd} order Fisher information if we defined p as ϕ_1. Of course, metric tensor rank 6 with one parameter case will give us the 4^{th} order Fisher information. Therefore, we can say that each i^{th} ($i = 1, 2, 3, ...$) extended term of one-parameter extended Fisher information (3.13) have relation with metric tensor rank j^{th} which can be obtained from j^{th}, where $j = 2, 4, 6, ...$, derivative of Shannon entropy.

CHAPTER IV

SUMMARY

We succeed to construct the one-parameter generalised Fisher information. The main method used to derive thFe one-parameter generalised Fisher information is the variational principle. We consider here the Fisher information as the action functional of free particle Lagrangian. With the new insight of the one-parameter generalised Lagrangian [21], one can naturally obtain the one-parameter generalised Fisher information which is

$$I_\lambda[q(y)] = \frac{4}{\lambda}\int\left[e^{\lambda q'^2(y)} - 1\right]dy \ ,$$

where $q(y) = \sqrt{f(y)}$ and λ is parameter. Furthermore, we can treat our one-parameter generalised Fisher information as the generator yielding

$$I_\lambda[q(y)] = I_1[q(y)] + \frac{\lambda}{2!}I_2[q(y)] + \frac{\lambda^2}{3!}I_3[q(y)] + ... \ .$$

Here $\{I_1, I_2, ...\}$ is called the Fisher information hierarchy. The first one $I_1[q(y)]$ is nothing but the standard Fisher information. By introducing $\phi_{n-1}(y) = f^n(y)$, we can rearrange Fisher information hierarchy to be in the form of ith moments such that

$$I'_n(\theta) = \frac{4}{(2n)^{2n}}\int\left[\frac{\partial}{\partial\theta}\ln\phi_{n-1}(x|\theta)\right]^{2n}\phi_{n-1}(x|\theta)dx \ .$$

Normally, Fisher information provides information lower bound on the variance of an unbiased estimator for a parameter through the relation called Cramér-Rao inequality. Here, in this present work, the generalised Cramér-Rao inequality is also obtained with the help of the Hölder's inequality and the q-expectation value of estimator

$$\begin{aligned}\frac{\int f^q(x \mid \theta)dx}{q} \ &\leq \ \left[\int(\hat{\Theta} - \theta)^\beta f(x \mid \theta)dx\right]^{1/\beta} \\ &\times\left[\int\left(\frac{\partial\ln f(x \mid \theta)}{\partial\theta}\right)^\alpha f^{\alpha(q-1)+1}(x \mid \theta)dx\right]^{1/\alpha} \ .\end{aligned}$$

Moreover, we find that our Fisher information hierarchy does not follow the addition property, except for the standard Fisher information. The interesting point is that this

non-additive property pops up when the Tsallis index q is not equal to 1 as intriguing built in the Tsallis entropy

$$S_q[f_{12}] = S_{\mathrm{q}}[f_1] + S_q[f_2] + (1-\mathrm{q})S_q[f_1]S_q[f_2] \ .$$

Of course this point is quite interesting since this non-additive property has been widely discussed in the Tsallis statistic [6]. Let us point out possibly indirect connection with the Tsallis entropy by recalling our one-parameter generalised Fisher information

$$I_\lambda[f] = \frac{4}{\lambda}\int \left[e^{\frac{\lambda}{4}\frac{f'^2(y)}{f(y)}} - 1\right] dy \ ,$$

and Tsallis entropy

$$S_q[f] = \frac{1}{1-q}\left[\int f^q(y) - 1\right] dy \ .$$

It might seem a bit strange but these two quantities more or less similar in the sense that they both contain a parameter and under the suitable limit the standard quantities can be recovered. However, more direct connection with the entropy might be the relative entropy or the Kullback–Leibler divergence, more specifically two-parameter Kullback–Leibler divergence and our whole hierarchy Fisher information can be identified with the two-parameter Fisher information with the appropriate choice of parameters.

REFERENCES

REFERENCES

[1] Landauer R. Information is physical. Physics Today. 1991;44(5):23-9.

[2] Shannon CE. A mathematical theory of communication. ACM SIGMOBILE mobile computing and communications review. 2001;5(1):3-55.

[3] Fisher RA. Theory of statistical estimation. In Mathematical proceedings of the Cambridge philosophical society. Cambridge University Press. 1925;22(5):700-725.

[4] Kullback S. Information theory and statistics. Courier Corporation; 1997.

[5] Rényi A. On measures of entropy and information. InProceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics. University of California Press. 1961;4:547-562.

[6] Tsallis C. Possible generalization of Boltzmann-Gibbs statistics. Journal of statistical physics. 1988;52(1):479-87.

[7] Tsallis C, Baldovin F, Cerbino R, Pierobon P. Introduction to nonextensive statistical mechanics and thermodynamics. In The Physics of Complex Systems (New Advances and Perspectives). IOS Press, 2003.

[8] Radhakrishnan C, Chinnarasu R, Jambulingam S. A fractional entropy in fractal phase space: Properties and characterization. International Journal of Statistical Mechanics. 2014.

[9] Boekee DE. Generalized Fisher information with application to estimation problems. IFAC Proceedings Volumes. 1977;10(12):75-82.

[10] Pennini F, Plastino A. Fisher's information measure in a Tsallis' nonextensive setting and its application to diffusive process. Physica A: Statistical Mechanics and its Applications. 1997;247(1-4):559-69.

[11] Plastino A, Plastino AR, Miller HG. Tsallis nonextensive thermostatistics and Fisher's information measure. Physica A: Statistical Mechanics and its Applications. 1997;235(3-4):577-88.

[12] Bercher JF. Some properties of generalized Fisher information in the context of nonextensive thermostatistics. Physica A: Statistical Mechanics and its Applications. 2013;392(15):3140-54.

[13] Bercher JF. On generalized Cramér–Rao inequalities, generalized Fisher information and characterizations of generalized q-Gaussian distributions. Journal of Physics A: Mathematical and Theoretical. 2012;45(25):255303.

[14] Frieden BR, Soffer BH. Lagrangians of physics and the game of Fisher-information transfer. Physical Review E. 1995;52(3):2274.

[15] Frieden BR. Fisher information, disorder, and the equilibrium distributions of physics. Physical Review A. 1990;41(8):4265.

[16] Frieden BR. Physics from Fisher information: a unification. Cambridge University Press; 2000.

[17] Zenkov DV (Ed.). The Inverse Problem of the Calculus of Variations: Local and Global Theory, Springer; 2015.

[18] Deesuwan T. Towards thermodynamics of quantum systems away from equilibrium (Doctoral dissertation, Imperial College London).

[19] Cortez LA, de Oliveira EC. On exact and inexact differentials and applications. International Journal of Mathematical Education in Science and Technology. 2017;48(4):630-41.

[20] Adkins CJ, Adkins CJ. An introduction to thermal physics. Cambridge University Press; 1987.

[21] Surawuttinack K, Yoo-Kong S, Tanasittikosol M. Multiplicative form of the Lagrangian. Theoretical and Mathematical Physics. 2016;189(3):1693-711.

[22] Stone JV. Information theory: a tutorial introduction; 2015.

[23] Fadeev DK. Zum Begriff der Entropie einer endlichen Wahrscheinlichkeitsschemas. Arbeiten zur Informationstheorie I. Deutscher Verlag der Wissenschaften. 1957:85-90.

[24] Wang QA. Extensive generalization of statistical mechanics based on incomplete information theory. Entropy. 2003;5(2):220-32.

[25] Wang QA. Incomplete statistics: nonextensive generalizations of statistical mechanics. Chaos, Solitons & Fractals. 2001;12(8):1431-7.

[26] Golomb S. The information generating function of a probability distribution (corresp.). IEEE Transactions on Information Theory. 1966;12(1):75-7.

[27] Cvetkovski Z. Inequalities: theorems, techniques and selected problems. Springer Science & Business Media; 2012.

[28] Heavens AF, Seikel M, Nord BD, Aich M, Bouffanais Y, Bassett BA, Hobson MP. Generalized fisher matrices. Monthly Notices of the Royal Astronomical Society. 2014;445(2):1687-93.

BIOGRAPHY

BIOGRAPHY

Name-Surname	Worachet Bukaew
Date of Birth	August 21, 1996
Place of Birth	Ubon Ratchathani , Thailand
Address	207 Moo 6 Tambon Pho Sai , Amphoe Phibun Mangsahan , Ubon Ratchathani Province, Thailand 34110

Education Background

2019	B.S. (Physics), Ubon Ratchathani University, Ubon Ratchathani, Thailand

APPENDIX