Chapter 2
Information Quantities and Parameter Estimation in Classical Systems

Abstract For the study of quantum information theory, mathematical statistics, and information geometry, which are mainly examined in a nonquantum context. This chapter briefly summarizes the fundamentals of these topics from a unified viewpoint. Since these topics are usually treated individually, this chapter will be useful even for nonquantum applications.

2.1 Information Quantities in Classical Systems

When all the given density matrices $\rho_1, \ldots, \rho_n$ commute, they may be simultaneously diagonalized using a common orthonormal basis $\{u^1, \ldots, u^d\}$ according to $\rho_1 = \sum_i p_{1,i}|u^i\rangle\langle u^i|, \ldots, \rho_n = \sum_i p_{n,i}|u^i\rangle\langle u^i|$. In this case, it is sufficient to treat only the diagonal elements, i.e., we discuss only the probability distributions $p_1, \ldots, p_n$. Henceforth we will refer to such cases as **classical** because they do not exhibit any quantum properties. Let us now examine various information quantities with respect to probability distributions.

2.1.1 Entropy

Before proceeding to the definition of information quantities, we prepare the notations for basic probability theory. For a given probability distribution $p = \{p_x\}_{x\in\Omega}$ of the real-valued random variable X, we define the expectation $\mathrm{E}_p(X)$ as

$$\mathrm{E}_p(X) \stackrel{\mathrm{def}}{=} \sum_{x\in\Omega} x p_x. \tag{2.1}$$

When the number $-\log p_x$ is regarded as a real-valued random variable, the **Shannon entropy** is defined as the expectation of the real-valued random variable under the probability distribution p, i.e.,[1]

[1] In this case, we consider $0 \log 0$ to be 0 here.

M. Hayashi, *Quantum Information Theory*, Graduate Texts in Physics,
DOI 10.1007/978-3-662-49725-8_2

$$H(p) \stackrel{\text{def}}{=} \sum_{x \in \Omega} -p_x \log p_x. \tag{2.2}$$

It is often simply called **entropy**. That is, when $\mathcal{P}(\Omega)$ denotes the set of probability distributions on the probability space Ω, H is a real-valued function on $\mathcal{P}(\Omega)$. Sometimes, we denote the probability distribution of a random variable X by P_X. In this case, we write the entropy of P_X as $H(X)$. For $\Omega = \{0, 1\}$, the probability distribution is written as $(a, 1-a)$ and the entropy is called a **binary entropy**, which is given by $h(a) \stackrel{\text{def}}{=} -a \log a - (1-a)\log(1-a)$.

When the number of elements of Ω is a finite number k, it is possible to choose the distribution so that all probabilities p_i have the same value. Such a probability distribution $p = (p_i)$ is called a **uniform distribution** and is denoted by $p_{\mathrm{mix},\Omega}$. It is simplified to p_{mix} for simplicity. If it is necessary to denote the number of supports k explicitly, we write $p_{\mathrm{mix},k}$. As shown later, any distribution p on Ω satisfies the relation

$$H(p) \le \log k = H(p_{\mathrm{mix},\Omega}). \tag{2.3}$$

The entropy $H(\mathrm{P}_{X,Y}(x, y))$ of the joint distribution $\mathrm{P}_{X,Y}$ for two random variables X and Y is denoted by $H(X, Y)$. In particular, if Y can be expressed as $f(X)$, where f is a function, then[Exe. 2.1]

$$H(X, Y) = H(X, f(X)) = H(X). \tag{2.4}$$

Given a conditional probability $\mathrm{P}_{X|Y=y} = \{\mathrm{P}_{X|Y}(x|y)\}_x$, the entropy of X is given by $H(X|Y=y) \stackrel{\text{def}}{=} H(\mathrm{P}_{X|Y=y})$ when the random variable Y is known to be y. The expectation of this entropy with respect to the probability distribution of Y is called the **conditional entropy** denoted by $H(X|Y)$. We may write it as

$$\begin{aligned} H(X|Y) &\stackrel{\text{def}}{=} \sum_y \sum_x -\mathrm{P}_Y(y)\mathrm{P}_{X|Y}(x|y) \log \mathrm{P}_{X|Y}(x|y) \\ &= -\sum_{x,y} \mathrm{P}_{X,Y}(x, y) \log \frac{\mathrm{P}_{X,Y}(x, y)}{\mathrm{P}_Y(y)} \\ &= -\sum \mathrm{P}_{X,Y}(x, y) \log \mathrm{P}_{X,Y}(x, y) + \sum_y \mathrm{P}_Y(y) \log \mathrm{P}_Y(y) \\ &= H(X, Y) - H(Y). \end{aligned} \tag{2.5}$$

The final equation in (2.5) is called **chain rule**. Using chain rule (2.5) and (2.4), we have

$$H(X) = H(f(X)) + H(X|f(X)) \ge H(f(X)), \tag{2.6}$$

which is called **monotonicity**.

Applying (2.4) to the distribution $P_{X|Y=y}$, we have

$$H(X, f(X,Y)|Y) = \sum_y P_Y(y) H(X, f(X,y)|Y = y)$$
$$= \sum_y P_Y(y) H(X|Y = y) = H(X|Y). \tag{2.7}$$

Since (as will be shown later)

$$H(X) + H(Y) - H(X,Y) \geq 0\,, \tag{2.8}$$

we have

$$H(X) \geq H(X|Y)\,. \tag{2.9}$$

If Y takes values in $\{0, 1\}$, (2.9) is equivalent to the concavity of the entropy[Exe. 2.2]:

$$\lambda H(p) + (1-\lambda) H(p') \leq H(\lambda p + (1-\lambda)p'), \quad 0 < \forall\lambda < 1\,. \tag{2.10}$$

Exercises

2.1 Verify (2.4) if the variable Y can be written $f(X)$ for a function f.

2.2 Verify that (2.9) and (2.10) are equivalent.

2.3 Given a distribution $p = \{p_x\}$ on $\{1, \ldots, k\}$. Assume that the maximum probability p_x is larger than a. Verify that $H(p) \leq h(a) + (1-a)\log(k-1)$.

2.4 Define $p_A \times p_B(\omega_A, \omega_B) = p_A(\omega_A) p_B(\omega_B)$ in $\Omega_A \times \Omega_B$ for probability distributions p_A in Ω_A, p_B in Ω_B. Show that

$$H(p_A) + H(p_B) = H(p_A \times p_B). \tag{2.11}$$

2.1.2 Relative Entropy

We now consider a quantity that expresses the closeness between two probability distributions $p = \{p_i\}_{i\in\Omega}$ and $q = \{q_i\}_{i\in\Omega}$. It is called an information quantity because our access to information is closely related to the difference between the distributions reflecting the information of our interest. A typical example is the **relative entropy**[2] $D(p\|q)$, which is defined as

[2]The term relative entropy is commonly used in statistical physics. In information theory, it is generally known as the **Kullback–Leibler divergence**, while in statistics it is known as the **Kullback–Leibler information**.

$$D(p\|q) \stackrel{\text{def}}{=} \sum_{i\in\Omega} p_i \log \frac{p_i}{q_i}\,. \tag{2.12}$$

This quantity is always no less than 0, and it is equal to 0 if and only if $p = q$. This can be shown by applying the **logarithmic inequality**[Exe. 2.6] "$\log x \le x - 1$ for $x > 0$" to (2.12):

$$0 - D(p\|q) = \sum_{i=1}^{k} p_i \left(-\frac{q_i}{p_i} + 1 + \log \frac{q_i}{p_i}\right) \le \sum_{i=1}^{k} p_i\, 0 = 0\,.$$

Note that the equality of $\log x \le x - 1$ holds only when $x = 1$. We may obtain (2.3) by using the positivity of the relative entropy for the case $q = \{1/k\}$.

Let us now consider possible information processes. For simplicity, we assume that the probability space Ω is given as the set $\mathbb{N}_k \stackrel{\text{def}}{=} \{1, \ldots, k\}$. When an information process converts a set $\mathbb{N}_k \stackrel{\text{def}}{=} \{1, \ldots, k\}$ to another set $\mathbb{N}_l$ deterministically, we may denote the information processing by a function from $\mathbb{N}_k$ to $\mathbb{N}_l$. If it converts probabilistically, it is denoted by a real-valued matrix $\{Q^i_j\}$ in which every element Q^i_j represents the probability of the output data $j \in \mathbb{N}_l$ when the input data are $i \in \mathbb{N}_k$. This matrix $Q = (Q^i_j)$ satisfies $\sum_{j=1}^{l} Q^i_j = 1$ for each i. Such a matrix Q is called a **stochastic transition matrix**. In this notation, Q^i expresses the distribution $(Q^i_1, \ldots, Q^i_k)$ on the output system with the input i. When the input signal is generated according to the probability distribution p, the output signal is generated according to the probability distribution $Q(p)_j \stackrel{\text{def}}{=} \sum_{i=1}^{k} Q^i_j p_i$. The stochastic transition matrix Q represents not only such probabilistic information processes but also probabilistic fluctuations in the data due to noise. Furthermore, since it expresses the probability distribution of the output system for each input signal, we can also use it to model a channel transmitting information.

A fundamental property of a stochastic transition matrix Q is the inequality

$$D(p\|q) \ge D(Q(p)\|Q(q))\,, \tag{2.13}$$

which is called an **information-processing inequality**. This property is often called **monotonicity**.[3] The inequality implies that the amount of information should not increase via any information processing. This inequality will be proved for the general case in Theorem 2.1. It may also be shown using a logarithmic inequality.

For example, consider the stochastic transition matrix $Q = (Q^i_j)$ from $\mathbb{N}_{2k}$ to $\mathbb{N}_k$, where Q^i_j is 1 when $i = j$, $j + k$ and 0 otherwise. Given two probability distributions p, p' in $\mathbb{N}_k$, we define the probability distribution $\tilde{p}$ for $\mathbb{N}_{2k}$ as

$$\tilde{p}_i = \lambda p_i,\ \ \tilde{p}_{i+k} = (1-\lambda) p'_i, \quad 1 \le \forall i \le k$$

[3] In this book, monotonicity refers to only the monotonicity regarding the change in probability distributions or density matrices.

with a real number $\lambda \in (0, 1)$. Similarly, we define $\tilde{q}$ for two probability distributions q, q' in $\mathbb{N}_k$. Then,

$$D(\tilde{p}\|\tilde{q}) = \lambda D(p\|q) + (1-\lambda)D(p'\|q').$$

Since $Q(\tilde{p}) = \lambda p + (1-\lambda)p'$ and $Q(\tilde{q}) = \lambda q + (1-\lambda)q'$, the information-processing inequality (2.13) yields the **joint convexity** of the relative entropy

$$\lambda D(p\|q) + (1-\lambda)D(p'\|q') \geq D(\lambda p + (1-\lambda)p'\|\lambda q + (1-\lambda)q'). \quad (2.14)$$

Next, let us consider other information quantities that express the difference between the two probability distributions p and q. In order to express the amount of information, these quantities should satisfy the property given by (2.13). This property can be satisfied by constructing the information quantity in the following manner. First, we define convex functions. When a function f satisfies

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2), \quad 0 \leq \forall\lambda \leq 1, \forall x_1, x_2 \in \mathbb{R},$$

it is called a **convex function**. For a probability distribution $p = \{p_i\}$, a convex function f satisfies **Jensen's inequality**:

$$\sum_i p_i f(x_i) \geq f\left(\sum_i p_i x_i\right). \quad (2.15)$$

Theorem 2.1 (Csiszár [1]) *Let f be a convex function. The information quantity $D_f(p\|q) \stackrel{\text{def}}{=} \sum_i q_i f\left(\frac{p_i}{q_i}\right)$ then satisfies the* **monotonicity** *condition*

$$D_f(p\|q) \geq D_f(Q(p)\|Q(q)). \quad (2.16)$$

*Henceforth, $D_f(p\|q)$ will be called an f-***relative entropy**.[4]

For example, for $f(x) = x\log x$ we obtain the relative entropy. For $f(x) = 1 - \sqrt{x}$,

$$D_f(p\|q) = 1 - \sum_i \sqrt{p_i}\sqrt{q_i} = \frac{1}{2}\sum_i \left(\sqrt{p_i} - \sqrt{q_i}\right)^2. \quad (2.17)$$

Its square root is called the **Hellinger distance** and is denoted by $d_2(p, q)$. This satisfies the axioms of a distance[Exe. 2.14]. When $f(x) = \frac{4}{1-\alpha^2}(1 - x^{(1+\alpha)/2})(-1 < \alpha < 1)$, $D_f(p\|q)$ is equal to the α-divergence $\frac{4}{1-\alpha^2}\left(1 - \sum_i p_i^{(1+\alpha)/2} q_i^{(1-\alpha)/2}\right)$ according

[4]This quantity is more commonly used in information theory, where it is called f-divergence [1]. In this text, we prefer to use the term "relative entropy" for all relative-entropy-like quantities.

to Amari and Nagaoka [2]. By applying inequality (2.16) to the concave function $x \to x^s$ $(0 \le s \le 1)$ and the convex function $x \to x^s$ $(s \le 0)$, we obtain the inequalities

$$\sum_i p_i^{1-s} q_i^s \le \sum_j Q(p)_j^{1-s} Q(q)_j^s \text{ for } 0 \le s \le 1\,,$$
$$\sum_i p_i^{1-s} q_i^s \ge \sum_j Q(p)_j^{1-s} Q(q)_j^s \text{ for } s \le 0\,.$$

Hence, the quantity $\phi(s|p\|q) \stackrel{\text{def}}{=} \log(\sum_i p_i^{1-s} q_i^s)$ satisfies the **monotonicity**

$$\phi(s|p\|q) \le \phi(s|Q(p)\|Q(q)) \text{ for } 0 \le s \le 1\,,$$
$$\phi(s|p\|q) \ge \phi(s|Q(p)\|Q(q)) \text{ for } s \le 0\,.$$

The relative entropy can be expressed as

$$\phi'(0|p\|q) = -D(p\|q), \quad \phi'(1|p\|q) = D(q\|p). \tag{2.18}$$

Since $\phi(s|p\|q)$ is a convex function of s[Exe. 2.16], the **relative Rényi entropy** [3]

$$D_{1-s}(p\|q) \stackrel{\text{def}}{=} -\frac{\phi(s|p\|q)}{s} = -\frac{\phi(s|p\|q) - \phi(0|p\|q)}{s} = -\frac{1}{s} \log \sum_i p_i^{1-s} q_i^s \tag{2.19}$$

is monotone decreasing for s[Exe. 2.17]. More precise analyses for these quantities are given in Exercises 3.45, 3.52, and 3.53.

We will abbreviate it to $\phi(s)$ if it is not necessary to specify p and q explicitly. Hence, we define the minimum and the maximum relative entropies as

$$D_{\max}(p\|q) \stackrel{\text{def}}{=} -\log \max_i \frac{p_i}{q_i}, \quad D_{\min}(p\|q) \stackrel{\text{def}}{=} -\log \sum_{i:p_i>0} q_i\,. \tag{2.20}$$

Hence, we obtain the relations[Exe. 2.18, 2.19]

$$\lim_{s\to-\infty} D_{1-s}(p\|q) = D_{\max}(p\|q), \quad \lim_{s\to 1} D_{1-s}(p\|q) = D_{\min}(p\|q), \tag{2.21}$$
$$\lim_{s\to 0} D_{1-s}(p\|q) = D(p\|q). \tag{2.22}$$

That is, $D_{\max}(p\|q)$ and $D_{\min}(p\|q)$ give the maximum and the minimum values of $D_{1-s}(p\|q)$, respectively.

Proof of Theorem 2.1 Since f is a convex function, Jensen's inequality ensures that

$$\sum_i \frac{Q_j^i q_i}{\sum_{i'} Q_j^{i'} q_{i'}} f\left(\frac{p_i}{q_i}\right) \geq f\left(\sum_i \frac{Q_j^i q_i}{\sum_{i'} Q_j^{i'} q_{i'}} \frac{p_i}{q_i}\right) = f\left(\frac{\sum_i Q_j^i p_i}{\sum_{i'} Q_j^{i'} q_{i'}}\right).$$

Therefore,

$$\begin{aligned} D_f(Q(p)\|Q(q)) &= \sum_j \sum_{i''} Q_j^{i''} q_{i''} f\left(\frac{\left(\sum_i Q_j^i p_i\right)}{\left(\sum_{i'} Q_j^{i'} q_{i'}\right)}\right) \\ &\leq \sum_j \sum_{i''} Q_j^{i''} q_{i''} \sum_i \frac{Q_j^i q_i}{\sum_{i'} Q_j^{i'} q_{i'}} f\left(\frac{p_i}{q_i}\right) \\ &= \sum_j \sum_i Q_j^i q_i f\left(\frac{p_i}{q_i}\right) = \sum_i q_i f\left(\frac{p_i}{q_i}\right) = D_f(p\|q). \end{aligned}$$

■

We consider the **variational distance** as another information quantity. It is defined as

$$d_1(p,q) \stackrel{\text{def}}{=} \frac{1}{2} \sum_i |p_i - q_i| . \tag{2.23}$$

It is the f-relative entropy when $f(x)$ is chosen to be $\frac{1}{2}|1-x|$. However, it satisfies the **monotonicity** property[Exe. 2.9]

$$d_1(Q(p), Q(q)) \leq d_1(p,q) . \tag{2.24}$$

The variational distance, Hellinger distance, and relative entropy are related by the following formulas:

$$d_1(p,q) \geq d_2^2(p,q) \geq \frac{1}{2} d_1^2(p,q) , \tag{2.25}$$

$$D(p\|q) \geq -2 \log\left(\sum_i \sqrt{p_i}\sqrt{q_i}\right) \geq 2 d_2^2(p,q) . \tag{2.26}$$

The last inequality may be deduced from the logarithmic inequality. The combination of (2.25) and (2.26) is called **Pinsker inequality**.

When a stochastic transition matrix $Q = (Q_j^i)$ satisfies $\sum_i Q_j^i = 1$, i.e., its transpose is also a stochastic transition matrix, the stochastic transition matrix $Q = (Q_j^i)$ is called a **double stochastic transition matrix**. Now, we assume that the input symbol i and the output symbol j take the values in $1, \ldots, k_1$ and $1, \ldots, k_2$, respectively. When the stochastic transition matrix $Q = (Q_j^i)$ is double stochastic, we have $k_2 = \sum_{j=1}^{k_2} 1 = \sum_{j=1}^{k_2} \sum_{i=1}^{k_1} Q_j^i = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} Q_j^i = \sum_{i=1}^{k_1} 1 = k_1$. That is, any double stochastic matrix is a square matrix.

A stochastic transition square matrix Q is a double stochastic transition matrix if and only if the output distribution $Q(p_{\mathrm{mix}})$ is a uniform distribution because $Q(p_{\mathrm{mix}})_j = \sum_i Q^i_j \frac{1}{k} = \frac{1}{k}$. The double stochastic transition matrix Q and the probability distribution p satisfy

$$\log k - H(Q(p)) = D(Q(p)\|p_{\mathrm{mix},k}) \geq D(p\|p_{\mathrm{mix},k}) = \log k - H(p),$$

which implies that

$$H(Q(p)) \geq H(p)\,. \tag{2.27}$$

Exercises

2.5 Show that

$$D(p_A\|q_A) + D(p_B\|q_B) = D(p_A \times p_B\|q_A \times q_B) \tag{2.28}$$

for probability distributions p_A, q_A in Ω_A and p_B, q_B in Ω_B.

2.6 Show the logarithmic inequality, i.e., the inequality $\log x \leq x - 1$, holds for $x > 0$ and the equality holds only for $x = 1$.

2.7 Show that the f-relative entropy $D_f(p\|q)$ of a convex function f satisfies $D_f(p\|q) \geq f(1)$.

2.8 Prove (2.17).

2.9 Show that the variational distance satisfies the monotonicity condition (2.24).

2.10 Show that $d_1(p,q) \geq d_2^2(p,q)$ by first proving the inequality $|x-y| \geq (\sqrt{x} - \sqrt{y})^2$.

2.11 Show that $d_2^2(p,q) \geq \frac{1}{2} d_1^2(p,q)$ following the steps below.
(**a**) Prove

$$\left(\sum_i |p_i - q_i|\right)^2 \leq \left(\sum_i |\sqrt{p_i} - \sqrt{q_i}|^2\right)\left(\sum_i |\sqrt{p_i} + \sqrt{q_i}|^2\right)$$

using the Schwarz inequality.
(**b**) Show that $\sum_i |\sqrt{p_i} + \sqrt{q_i}|^2 \leq 4$.
(**c**) Show that $d_2^2(p,q) \geq \frac{1}{2} d_1^2(p,q)$ using the above results.

2.12 Show that $d_1(p,q) \leq \sum_{x \neq x_0} |p_x - q_x|$ for any x_0.

2.13 Show that $D(p\|q) \geq -2\log\left(\sum_i \sqrt{p_i}\sqrt{q_i}\right)$.

2.14 Verify that the Hellinger distance satisfies the axioms of a distance by following the steps below.
(**a**) Prove the following for arbitrary vectors x and y

$$(\|x\| + \|y\|)^2 \geq \|x\|^2 + \langle x, y\rangle + \langle y, x\rangle + \|y\|^2 .$$

(**b**) Prove the following for arbitrary vectors x and y:

$$\|x\| + \|y\| \geq \|x + y\| .$$

(**c**) Show the following for the three probability distributions p, q, and r:

$$\sqrt{\sum_i \left(\sqrt{p_i} - \sqrt{q_i}\right)^2} \leq \sqrt{\sum_i \left(\sqrt{p_i} - \sqrt{r_i}\right)^2} + \sqrt{\sum_i \left(\sqrt{r_i} - \sqrt{q_i}\right)^2} .$$

Note that this formula is equivalent to the axiom of a distance $d_2(p, q) \leq d_2(p, r) + d_2(r, q)$ for the Hellinger distance.

2.15 Show (2.18).

2.16 Show that $\phi(s|p\|q)$ is convex for s.

2.17 Show that $\frac{f(s)}{s}$ is (strictly) monotone increasing for s when $f(0) = 0$ and $f(s)$ is (strictly) convex for s.

2.18 Show that $\lim_{s\to-\infty} D_{1-s}(p\|q) = D_{\max}(p\|q)$ by following the steps below.
(**a**) Show that $\frac{1}{t}\log(\sum_{i=1}^k a_i b_i^t) \to \log\max(b_1, \ldots, b_k)$ as $t \to \infty$ for $a_i, b_i \geq 0$.
(**b**) Show the desired equation.

2.19 Show that $\lim_{s\to 1} D_{1-s}(p\|q) = D_{\min}(p\|q)$.

2.20 Show that

$$D(p\|q) = \max_{\lambda=(\lambda_1,\ldots,\lambda_k)\in\mathbb{R}^k} \sum_{i=1}^k p_i\lambda_i - \log\sum_{i=1}^k q_i e^{\lambda_i} \tag{2.29}$$

for two probability distributions p and q on $\{1, \ldots, k\}$.

2.1.3 Mutual Information

Given the joint probability distribution $\mathrm{P}_{X,Y}$ of two random variables X and Y, the **marginal distributions** P_X and P_Y of $\mathrm{P}_{X,Y}$ are defined as

$$\mathrm{P}_X(x) \stackrel{\mathrm{def}}{=} \sum_y \mathrm{P}_{X,Y}(x, y) \quad \text{and} \quad \mathrm{P}_Y(y) \stackrel{\mathrm{def}}{=} \sum_x \mathrm{P}_{X,Y}(x, y)\,.$$

Then, the conditional distribution is calculated as

$$\mathrm{P}_{X|Y}(x|y) = \frac{\mathrm{P}_{X,Y}(x, y)}{\mathrm{P}_Y(y)}\,.$$

When $\mathrm{P}_X(x) = \mathrm{P}_{X|Y}(x|y)$, two random variables X and Y are **independent**. In this case, the joint distribution $\mathrm{P}_{X,Y}(x, y)$ is equal to the product of marginal distributions $\mathrm{P}_X \times \mathrm{P}_Y(x, y) := \mathrm{P}_X(x)\mathrm{P}_Y(y)$. That is, the relative entropy $D(\mathrm{P}_{X,Y}\|\mathrm{P}_X \times \mathrm{P}_Y)$ is equal to zero. We now introduce **mutual information** $I(X : Y)$, which expresses how different the joint distribution $\mathrm{P}_{X,Y}(x, y)$ is from the product of marginal distributions $\mathrm{P}_X(x)\mathrm{P}_Y(y)$. This quantity satisfies the following relation:

$$\begin{aligned} I(X : Y) &\stackrel{\mathrm{def}}{=} D(\mathrm{P}_{X,Y}\|\mathrm{P}_X\mathrm{P}_Y) = \sum_{x,y} \mathrm{P}_{X,Y}(x, y) \log \frac{\mathrm{P}_{X,Y}(x, y)}{\mathrm{P}_X(x)\mathrm{P}_Y(y)} \\ &= H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X, Y)\,. \end{aligned} \tag{2.30}$$

Hence, inequality (2.8) may be obtained from the above formula and the positivity of $I(X : Y)$. Further, we can define a **conditional mutual information** in a manner similar to that of the entropy. This quantity involves another random variable Z (in addition to X and Y) and is defined as

$$\begin{aligned} I(X : Y|Z) &\stackrel{\mathrm{def}}{=} \sum_z \mathrm{P}_Z(z) I(X : Y|Z = z) \\ &= \sum_{x,y,z} \mathrm{P}_{X,Y,Z}(x, y, z) \log \frac{\mathrm{P}_{XY|Z}(x, y|z)}{\mathrm{P}_{X|Z}(x|z)\mathrm{P}_{Y|Z}(y|z)} \geq 0\,, \end{aligned} \tag{2.31}$$

where $I(X : Y|Z = z)$ is the mutual information of X and Y assuming that $Z = z$ is known. By applying (2.5) and (2.30) to the case $Z = z$, we obtain

$$\begin{aligned} &I(X : Y|Z) = H(X|Z) + H(Y|Z) - H(XY|Z) = H(X|Z) - H(X|YZ) \\ &= -\left(H(X) - H(X|Z)\right) + \left(H(X) - H(X|YZ)\right) \\ &= -I(X : Z) + I(X : YZ)\,. \end{aligned}$$

This equation is called the **chain rule** of mutual information, which may also be written as

$$I(X : YZ) = I(X : Z) + I(X : Y|Z)\,. \tag{2.32}$$

Hence, it follows that

$$I(X : YZ) \geq I(X : Z)\,.$$

Note that (2.32) can be generalized as

$$I(X : YZ|U) = I(X : Z|U) + I(X : Y|ZU)\,. \tag{2.33}$$

Next, we apply the above argument to the case where the information channel is given by a stochastic transition matrix $Q = (Q_y^x)$ and the input distribution is given by p. Let X and Y be, respectively, the random variables of the input system and output system. That is, their joint distribution is given as $\mathrm{P}_{X,Y}(x, y) = Q_y^x p_x$. Then, the mutual information $I(X : Y)$ can be regarded as the amount of information transmitted via channel Q when the input signal is generated with the distribution p. This is called **transmission information**, and it is denoted by $I(p, Q)$. Therefore, we can define the transmission information by

$$I(p, Q) \stackrel{\text{def}}{=} H(Q(p)) - \sum_x p_x H(Q^x)\,. \tag{2.34}$$

We will now discuss **Fano's inequality**, which is given by the following theorem.

Theorem 2.2 (Fano [4]) *Let X and Y be random variables that take values in the same data set $\mathbb{N}_k = \{1, \ldots, k\}$. Then, the following inequality holds:*

$$\begin{aligned} H(X|Y) &\leq \mathrm{P}\{X \neq Y\} \log(k-1) + h(\mathrm{P}\{X \neq Y\}) \\ &\leq \mathrm{P}\{X \neq Y\} \log k + \log 2\,. \end{aligned} \tag{2.35}$$

Proof We define the random variable $Z \stackrel{\text{def}}{=} \begin{cases} 0 & X = Y \\ 1 & X \neq Y \end{cases}$. Applying (2.5) to X and Z under the condition $Y = y$, we obtain

$$\begin{aligned} H(X|Y = y) &= H(X, Z|Y = y) \\ &= \sum_z \mathrm{P}_{Z|Y}(z|y) H(X|Z = z, Y = y) + H(Z|Y = y). \end{aligned}$$

The first equality follows from the fact that the random variable Z can be uniquely obtained from X. Taking the expectation with respect to y, we get

$$\begin{aligned} H(X|Y) &= H(X|Z, Y) + H(Z|Y) \leq H(X|Z, Y) + H(Z) \\ &= H(X|Z, Y) + h(\mathrm{P}\{X \neq Y\}). \end{aligned} \tag{2.36}$$

Applying (2.3), we have

$$H(X|Y = y, Z = 0) = 0, \quad H(X|Y = y, Z = 1) \leq \log(k-1).$$

Therefore,

$$H(X|Y, Z) \leq \mathrm{P}\{X \neq Y\} \log(k-1). \tag{2.37}$$

Finally, combining (2.36) and (2.37), we obtain (2.35). ■

Exercise

2.21 Show the chain rule of conditional mutual information (2.33) based on (2.32).

2.1.4 The Independent and Identical Condition and Rényi Entropy

Given a probability distribution $p = \{p_i\}_{i=1}^k$, we define the **Rényi entropy** $H_{1-s}(p)$ of order $1-s$ as

$$H_{1-s}(p) \stackrel{\mathrm{def}}{=} \frac{\psi(s|p)}{s}, \quad \psi(s|p) \stackrel{\mathrm{def}}{=} \log \sum_i p_i^{1-s} \tag{2.38}$$

for a real number s in addition to the entropy $H(p)$. We will abbreviate the quantity $\psi(s|p)$ to $\psi(s)$ when there is no risk of ambiguity. When $0 < s < 1$, the quantity $\psi(s)$ is a positive quantity that is larger when the probability distribution is closer to the uniform distribution. When $s < 0$, the quantity $\psi(s)$ is a negative quantity that is smaller when the probability distribution is closer to the uniform distribution. Finally, when $s = 0$, the quantity $\psi(s)$ is equal to 0. The derivative $\psi'(0)$ of $\psi(s)$ at $s = 0$ is equal to $H(p)$.

Hence, Rényi entropy $H_{1-s}(p)$ is always positive, and the limit $\lim_{s\to 0} H_{1-s}(p)$ equals $H(p)$. Further, since $\psi(s)$ is convex, Rényi entropy $H_{1-s}(p)$ is monotone increasing for s. In particular, Rényi entropy $H_{1-s}(p_{\mathrm{mix},k})$ is equal to $\log k$. Hence, Rényi entropy $H_{1-s}(p)$ expresses the amount of the uncertainty of the distribution of p. We also define the **minimum entropy** $H_{\min}(p)$ and the **maximum entropy** $H_{\max}(p)$ as

$$H_{\min}(p) \stackrel{\mathrm{def}}{=} -\log \max_i p_i, \quad H_{\max}(p) \stackrel{\mathrm{def}}{=} \log |\{i|p_i > 0\}|. \tag{2.39}$$

Then, we obtain

$$\lim_{s\to-\infty} H_{1-s}(p) = H_{\min}(p), \quad \lim_{s\to 1} H_{1-s}(p) = H_{\max}(p). \tag{2.40}$$

These give the minimum and the maximum of Rényi entropies $H_{1-s}(p)$.

Now consider n data $i_1, \ldots, i_n$ that are generated independently with the same probability distribution $p = \{p_i\}_{i=1}^k$. The probability of obtaining a particular data sequence $i^n = (i_1, \ldots, i_n)$ is given by $p_{i_1} \cdot \cdots \cdot p_{i_n}$. This probability distribution is called an n-fold **independent and identical distribution** (abbreviated as n-i.i.d.) and denoted by p^n. Then, we have $\psi(s|p^n) = n\psi(s|p)$, i.e., $H_{1-s}(p^n) = nH_{1-s}(p)$[Exe. 2.22]. When a sufficiently large number n of data are generated according to the independent and identical condition, the behavior of the distribution may be characterized by the entropy and the Rényi entropy.

The probability of the likelihood being less than $a \ge 0$ under the probability distribution p, i.e., the probability that $\{p_i \le a\}$, is

$$p\{p_i \le a\} = \sum_{i: p_i \le a} p_i \le \sum_{i: 1 \le \frac{a}{p_i}} \left(\frac{a}{p_i}\right)^s p_i \le \sum_{i=1}^k p_i^{1-s} a^s = e^{\psi(s) + s \log a} \tag{2.41}$$

if $0 \le s \le 1$. Accordingly,

$$p^n\{p_{i^n}^n \le e^{-nR}\} \le e^{n \min_{0 \le s \le 1} (\psi(s) - sR)}. \tag{2.42}$$

Conversely, the probability of the likelihood being greater than a, i.e., the probability that $\{p_i > a\}$, is

$$p\{p_i > a\} \le \sum_{i: 1 > \frac{a}{p_i}} \left(\frac{a}{p_i}\right)^s p_i \le \sum_{i=1}^k p_i^{1-s} a^s = e^{\psi(s) + s \log a} \tag{2.43}$$

if $s \le 0$. Similarly, we obtain

$$p^n\{p_{i^n}^n > e^{-nR}\} \le e^{n \min_{s \le 0} (\psi(s) - sR)}. \tag{2.44}$$

The exponential decreasing rate (exponent) on the right-hand side (RHS) of (2.42) is negative when $R > H(p)$. Hence, the probability $p^n\{p_{i^n}^n \le e^{-nR}\}$ approaches 0 exponentially. This fact can be shown as follows. Choosing a small $s_1 > 0$, we have $H_{1-s_1}(p) - R < 0$. Hence, we have

$$\min_{0 \le s \le 1} (\psi(s) - sR) = \min_{0 \le s \le 1} s(H_{1-s}(p) - R) \le s_1 (H_{1-s_1}(p) - R) < 0. \tag{2.45}$$

Hence, we see that the exponent on the RHS of (2.42) is negative. Conversely, the exponent on the RHS of (2.44) is negative when $R < H(p)$, and the probability $p^n\{p_{i^n}^n \le e^{-nR}\}$ approaches 0 exponentially. This can be verified from (2.45) by choosing $s_2 < 0$ with a sufficiently small absolute value.

We may generalize this argument for the likelihood $q_{i^n}^n$ of a different probability distribution q as follows. Defining $\tilde{\psi}(s) \stackrel{\text{def}}{=} \log \sum_i p_i q_i^{-s}$, we can show that

$$p^n\{q_{i^n}^n \le e^{-nR}\} \le e^{n \min_{0\le s}(\tilde{\psi}(s)-sR)}, \tag{2.46}$$

$$p^n\{q_{i^n}^n > e^{-nR}\} \le e^{n \min_{s\le 0}(\tilde{\psi}(s)-sR)}. \tag{2.47}$$

The Rényi entropy $H_{1-s}(p)$ and the entropy $H(p)$ express the concentration of probability under independent and identical distributions with a sufficiently large number of data. To investigate the concentration, let us consider the probability $P(p, L)$ of the most frequent L outcomes for a given probability distribution $p = (p_i)$.[5] This can be written as

$$P(p, L) = \sum_{i=1}^{L} p_i^{\downarrow}, \tag{2.48}$$

where $p_i^{\downarrow}$ are the elements of p_i that are reordered according to size. Let us analyze this by reexamining the set $\{p_i > a\}$. The number of elements of the set $|\{p_i > a\}|$ is evaluated as

$$|\{p_i > a\}| \le \sum_{i:p_i>a} \left(\frac{p_i}{a}\right)^{1-s} \le \sum_{i=1}^{k} p_i^{1-s} a^{-1+s} = e^{\psi(s)-(1-s)\log a} \tag{2.49}$$

when $0 < s < 1$. By using (2.41) and defining $b(s, R) \stackrel{\text{def}}{=} \frac{\psi(s)-R}{1-s}$ for R and $0 \le s < 1$, we have

$$|\{p_i > e^{b(s,R)}\}| \le e^R, \quad p\{p_i \le e^{b(s,R)}\} \le e^{\frac{\psi(s)-sR}{1-s}}.$$

We choose $s_0 \stackrel{\text{def}}{=} \operatorname{argmin}_{0\le s\le 1} \frac{\psi(s)-sR}{1-s}$[6] and define $P^c(p, e^R) \stackrel{\text{def}}{=} 1 - P(p, e^R)$; hence,

$$P^c(p, e^R) \le e^{\frac{\psi(s_0)-s_0R}{1-s_0}} = e^{\min_{0\le s\le 1}\frac{\psi(s)-sR}{1-s}}. \tag{2.50}$$

Applying this argument to the n-i.i.d p^n, we have

$$P^c(p^n, e^{nR}) \le e^{n\frac{\psi(s_0)-s_0R}{1-s_0}} = e^{n\min_{0\le s\le 1}\frac{\psi(s)-sR}{1-s}}. \tag{2.51}$$

Now, we let $R > H(p)$ and choose a sufficiently small number $0 < s_1 < 1$. Then, inequality (2.45) yields

$$\min_{0\le s<1} \frac{\psi(s)-sR}{1-s} = \min_{0\le s<1} \frac{s(H_{1-s}(p)-R)}{1-s} \le \frac{s_1(H_{1-s_1}(p)-R)}{1-s_1} < 0.$$

[5] If L is not an integer, we consider the largest integer that does not exceed L.

[6] $\operatorname{argmin}_{0\le s\le 1} f(s)$ returns the value of s that yields $\min_{0\le s\le 1} f(s)$. argmax is similarly defined.

Hence, the probability $P^c(p^n, e^{nR})$ approaches 0 exponentially. That implies that the probabilities are almost concentrated on the most frequent e^{nR} elements because $1 - P^c(p^n, e^{nR})$ equals the probability on the most frequent e^{nR} elements. Since this holds when $R > H(p)$, most of the probabilities are concentrated on $e^{nH(p)}$ elements. Therefore, this can be interpreted as meaning that the entropy $H(p)$ asymptotically expresses the degree of concentration. This will play an important role in problems such as source coding, which will be discussed later.

On the other hand, when $H(p) > R$, $P(p^n, e^{nR})$ approaches 0. To prove this, let us consider the following inequality for an arbitrary subset A:

$$pA \le a|A| + p\{p_i > a\}. \tag{2.52}$$

We can prove this inequality by considering the set $A = (A \cap \{p_i \le a\}) \cup (A \cap \{p_i > a\})$. Defining $R \stackrel{\text{def}}{=} \log |A|$ and $a \stackrel{\text{def}}{=} e^{b(s,R)}$ and using (2.43), we obtain $pA \le 2e^{\frac{\psi(s)-sR}{1-s}}$. Therefore,

$$P(p, e^R) \le 2e^{\min_{s\le 0} \frac{\psi(s)-sR}{1-s}}, \tag{2.53}$$

and we obtain

$$P(p^n, e^{nR}) \le 2e^{n \min_{s\le 0} \frac{\psi(s)-sR}{1-s}}. \tag{2.54}$$

We also note that in order to avoid $P(p^n, e^{nR}) \to 0$, we require $R \ge H(p)$ according to the condition $\min_{s\le 0} \frac{\psi(s)-sR}{1-s} < 0$.

Exercises

2.22 Show that $\psi(s|p_A \times p_B) = \psi(s|p_A) + \psi(s|p_B)$.

2.23 Define the distribution $p_s(x) := p(x)^{1-s} e^{-\psi(s)}$ and assume that a distribution q satisfies $H(q) = H(p_s)$. Show that $D(p_s\|p) \le D(q\|p)$ for $s \le 1$ by following steps below.
(**a**) Show that $\frac{1}{1-s} D(q\|p_s) = \frac{1}{1-s} \sum_x q(x) \log q(x) - \sum_x q(x) \log p(x) + \frac{\psi(s)}{1-s}$.
(**b**) Show $D(q\|p) - \frac{1}{1-s} D(q\|p_s) = D(p_s\|p)$.
(**c**) Show the desired inequality.

2.24 Show the equation

$$\sup_{0\le s\le 1} \frac{sR - \psi(s)}{1-s} = \min_{q: H(q)\ge R} D(q\|p) \tag{2.55}$$

following the steps below.
(**a**) Show that $\frac{sR-\psi(s)}{1-s} \le 0$ for $R \le H(p)$ and $s \in [0, 1]$.
(**b**) Show that both side of (2.55) are zero when $R \le H(p)$.
(**c**) Show that

$$H(p_s) = (1-s)\psi'(s) + \psi(s), \tag{2.56}$$
$$D(p_s \| p) = s\psi'(s) - \psi(s). \tag{2.57}$$

(**d**) Show that

$$\frac{d}{ds}(1-s)\psi'(s) + \psi(s) = (1-s)\psi''(s) < 0, \tag{2.58}$$
$$\frac{d}{ds}s\psi'(s) - \psi(s) = s\phi''(s) > 0 \tag{2.59}$$

for $s \in (0, 1)$.
(**e**) In the following, we consider the case $R > H(p)$. Show that there uniquely exists $s_R \in (0, 1)$ such that $H(p_{s_R}) = R$.
(**f**) Show that

$$\min_{q:H(q)=R} D(q\|p) = D(p_{s_R}\|p). \tag{2.60}$$

(**g**) Show that

$$\min_{q:H(q)\geq R} D(q\|p) = D(p_{s_R}\|p). \tag{2.61}$$

(**h**) Show that

$$D(p_{s_R}\|p) = \frac{s_R R - \psi(s_R)}{1-s_R}. \tag{2.62}$$

(**i**) Show that

$$\frac{d}{ds}\frac{sR-\psi(s)}{1-s} = \frac{R + (s-1)\psi'(s) - \psi(s)}{(1-s)^2}. \tag{2.63}$$

(**j**) Show that

$$\sup_{0\leq s\leq 1} \frac{sR-\psi(s)}{1-s} = \frac{s_R R - \psi(s_R)}{1-s_R}. \tag{2.64}$$

(**k**) Show (2.55).

2.25 Show that

$$\sup_{s\leq 0} \frac{sR-\psi(s|p)}{1-s} = \min_{q:H(q)\leq R} D(q\|p). \tag{2.65}$$

(**a**) Show that there uniquely exists $s_R \leq 0$ such that $H(p_{s_R}) = R$.
(**b**) Show that

$$\min_{q:H(q)=r} D(q\|p) = D(p_{s_R}\|p). \tag{2.66}$$

(**c**) Show that

$$\min_{q:H(q)\le R} D(q\|p) = D(p_{s_R}\|p). \tag{2.67}$$

(**d**) Show that

$$D(p_{s_R}\|p) = \frac{s_R R - \psi(s_R)}{1 - s_R}. \tag{2.68}$$

(**e**) Show that

$$\sup_{s\le 0} \frac{sR - \psi(s)}{1 - s} = \frac{s_R R - \psi(s_R)}{1 - s_R}. \tag{2.69}$$

(**f**) Show (2.65).

2.26 Assume that $R \le H_{\min}(p)$. Show that

$$\sup_{s\le 0} \frac{-\psi(s)}{1 - s} = \min_{q:H(q)=0} D(q\|p) = H_{\min}(p) \tag{2.70}$$

2.27 Show that

$$-\log \max_i p_i \le H_\alpha(p) \le -\log \min_i p_i \tag{2.71}$$

for $\alpha \ge 0$.

2.1.5 Conditional Rényi Entropy

Next, we consider the conditional extension of Rényi entropy. For this purpose, we focus on the following relation between the conditional entropy and the relative entropy. For a given joint distribution P_{XY} on $\mathcal{X} \times \mathcal{Y}$, we have two characterization for the conditional entropy[Exe. 2.28]

$$H(X|Y) = \log|\mathcal{X}| - D(\mathrm{P}_{XY}\|p_{\mathrm{mix},\mathcal{X}} \times \mathrm{P}_Y) \tag{2.72}$$
$$H(X|Y) = \log|\mathcal{X}| - \min_{Q_Y} D(\mathrm{P}_{XY}\|p_{\mathrm{mix},\mathcal{X}} \times Q_Y). \tag{2.73}$$

Based on the above relations, we define two kinds of **conditional Rényi entropies** for $s \in (-1, \infty)\backslash\{0\}$ as follows.

$$H_{1+s}(X|Y) \stackrel{\text{def}}{=} \log|\mathcal{X}| - D_{1+s}(\mathrm{P}_{XY}\|p_{\mathrm{mix},\mathcal{X}} \times \mathrm{P}_Y)$$
$$= -\frac{1}{s}\log \sum_y \mathrm{P}_Y(y) \sum_x \mathrm{P}_{X|Y=y}(x)^{1+s} \tag{2.74}$$

$$H^{\uparrow}_{1+s}(X|Y) \stackrel{\text{def}}{=} \log|\mathcal{X}| - \min_{Q_Y} D_{1+s}(\mathrm{P}_{XY}\|p_{\mathrm{mix},\mathcal{X}} \times Q_Y),$$
$$= \max_{Q_Y} -\frac{1}{s}\log \sum_{x,y} \mathrm{P}_{X,Y}(x,y)^{1+s} Q_Y(y)^{-s} \tag{2.75}$$

where Q_Y is an arbitrary distribution on $\mathcal{Y}$. In the case of $s = 0$, they are defined as $H(X|Y)$ because[Exe. 2.29]

$$\lim_{s\to 0} H_{1+s}(X|Y) = \lim_{s\to 0} H^{\uparrow}_{1+s}(X|Y) = H(X|Y). \tag{2.76}$$

According to the relations (2.40), **conditional minimum entropies** $H_{\min}(X|Y)$ and $H^{\uparrow}_{\min}(X|Y)$ and **conditional maximum entropies** $H_{\max}(X|Y)$ and $H^{\uparrow}_{\max}(X|Y)$ are defined as

$$H_{\min}(X|Y) \stackrel{\text{def}}{=} \lim_{s\to\infty} H_{1+s}(X|Y), \quad H^{\uparrow}_{\min}(X|Y) \stackrel{\text{def}}{=} \lim_{s\to\infty} H^{\uparrow}_{1+s}(X|Y), \tag{2.77}$$

$$H_{\max}(X|Y) \stackrel{\text{def}}{=} \lim_{s\to -1} H_{1+s}(X|Y), \quad H^{\uparrow}_{\max}(X|Y) \stackrel{\text{def}}{=} \lim_{s\to -1} H^{\uparrow}_{1+s}(X|Y). \tag{2.78}$$

From the definition, we find the relation

$$H_{1+s}(X|Y) \le H^{\uparrow}_{1+s}(X|Y). \tag{2.79}$$

Unfortunately, these two conditional Rényi entropies are not the same in general. Thanks to the property of the relative Rényi entropy, we have the following lemma[Exe. 2.31].

Lemma 2.1 *The functions $s \mapsto sH_{1+s}(X|Y)$ and $sH^{\uparrow}_{1+s}(X|Y)$ are concave for $s \in (-1,\infty)$. The functions $s \mapsto H_{1+s}(X|Y)$ and $H^{\uparrow}_{1+s}(X|Y)$ are monotonically decreasing.*

Lemma 2.2 *The quantity $H^{\uparrow}_{1+s}(X|Y)$ has the following form.*

$$H^{\uparrow}_{1+s}(X|Y) = \log|\mathcal{X}| - D_{1+s}\left(\mathrm{P}_{XY}\|p_{\mathrm{mix},\mathcal{X}} \times \mathrm{P}_Y^{(1+s)}\right) \tag{2.80}$$

$$= -\frac{1+s}{s}\log \sum_y \mathrm{P}_Y(y)\left(\sum_x \mathrm{P}_{X|Y}(x|y)^{1+s}\right)^{\frac{1}{1+s}} \tag{2.81}$$

$$= -\frac{1+s}{s}\log \sum_y \left(\sum_x \mathrm{P}_{XY}(x,y)^{1+s}\right)^{\frac{1}{1+s}}, \tag{2.82}$$

where $\mathrm{P}_Y^{(1+s)}(y) := \frac{(\sum_x \mathrm{P}_{XY}(x,y)^{1+s})^{\frac{1}{1+s}}}{\sum_{y'}(\sum_x \mathrm{P}_{XY}(x,y')^{1+s})^{\frac{1}{1+s}}}$.

Proof Substituting $\sum_x \mathrm{P}_{XY}(x, y)^{1+s}$ and $Q_Y(y)^{-s}$ to f and g in the reverse Hölder inequality (A.27) with $p = \frac{1}{1+s}$ and $q = -\frac{1}{s}$, we obtain

$$\begin{aligned}
&e^{-s(\log|\mathcal{X}| - D_{1+s}(\mathrm{P}_{XY}\|p_{\mathrm{mix},\mathcal{X}}\times Q_Y))} \\
&= \sum_y \sum_x \mathrm{P}_{XY}(x, y)^{1+s} Q_Y(y)^{-s} \\
&\ge \left(\sum_y \left(\sum_x \mathrm{P}_{XY}(x, y)^{1+s}\right)^{1/(1+s)}\right)^{1+s} \left(\sum_y Q_Y(y)^{-s\cdot -1/s}\right)^{-s} \\
&= \left(\sum_y \left(\sum_x \mathrm{P}_{XY}(x, y)^{1+s}\right)^{\frac{1}{1+s}}\right)^{1+s}
\end{aligned}$$

for $s \in (0, \infty]$. Since the equality holds when $Q_Y(y) = \mathrm{P}_Y^{(1+s)}(y)$, we obtain

$$e^{-sH^{\uparrow}_{1+s}(X|Y)} = \left(\sum_y \left(\sum_x \mathrm{P}_{XY}(x, y)^{1+s}\right)^{\frac{1}{1+s}}\right)^{1+s},$$

which implies (2.81) with $s \in (0, \infty]$.

The same substitution to the Hölder inequality (A.25) yields

$$e^{-s(\log|\mathcal{X}| - D_{1+s}(\mathrm{P}_{XY}\|p_{\mathrm{mix},\mathcal{X}}\times Q_Y))} \le \left(\sum_y \left(\sum_x \mathrm{P}_{XY}(x, y)^{1+s}\right)^{\frac{1}{1+s}}\right)^{1+s}$$

for $s \in (-1, 0)$. Since the equality holds when $Q_Y(y) = \mathrm{P}_Y^{(1+s)}(y)$, we obtain (2.81) with $s \in (-1, 0)$.

Finally, (2.82) follows from a simple calculation. ■

Taking the limits $s \to -1$ and $s \to \infty$ in Lemma 2.2, we obtain the following lemma[Exe. 2.30].

Lemma 2.3 *The quantities* $H_{\min}(X|Y)$, $H^{\uparrow}_{\min}(X|Y)$, $H_{\max}(X|Y)$, *and* $H^{\uparrow}_{\max}(X|Y)$ *are characterized as*

$$H_{\min}(X|Y) = -\log \max_{x,y:\mathrm{P}_Y(y)>0} \mathrm{P}_{X|Y=y}(x), \tag{2.83}$$

$$H^{\uparrow}_{\min}(X|Y) = -\log \sum_y \mathrm{P}_Y(y) \max_x \mathrm{P}_{X|Y=y}(x), \tag{2.84}$$

$$H_{\max}(X|Y) = -\log \sum_y \mathrm{P}_Y(y) |\{x|\mathrm{P}_{X|Y=y}(x) > 0\}|, \tag{2.85}$$

$$H^{\uparrow}_{\max}(X|Y) = -\log \max_{y:\mathrm{P}_Y(y)>0} |\{x|\mathrm{P}_{X|Y=y}(x) > 0\}|. \tag{2.86}$$

Further, as an inequality opposite to (2.79), we have

Lemma 2.4 ([5, Lemma 5]) *For $s \in (-1, 1)\backslash\{0\}$, we have*

$$H_{1+s}(X|Y) \ge H^{\uparrow}_{\frac{1}{1-s}}(X|Y). \tag{2.87}$$

Proof Next, we consider the case with $s \in (0, 1)$. Substituting $\mathrm{P}_{XY}(x, y)$ and $(\frac{\mathrm{P}_{XY}(x,y)}{\mathrm{P}_Y(y)})^s$ to f and g in the Hölder inequality (A.25) with $p = \frac{1}{1-s}$ and $q = \frac{1}{s}$, we obtain

$$\begin{aligned} e^{-sH_{1+s}(X|Y)} &= \sum_y \sum_x \mathrm{P}_{XY}(x, y) \left(\frac{\mathrm{P}_{XY}(x, y)}{\mathrm{P}_Y(y)}\right)^s \\ &\le \sum_y \left(\sum_x \mathrm{P}_{XY}(x, y)^{1/(1-s)}\right)^{1-s} \left(\sum_{x'} \frac{\mathrm{P}_{XY}(x', y)}{\mathrm{P}_Y(y)}\right)^s \\ &= \sum_y \left(\sum_x \mathrm{P}_{XY}(x, y)^{1/(1-s)}\right)^{1-s} = e^{-sH^{\uparrow}_{\frac{1}{1-s}}(X|Y)} \end{aligned} \tag{2.88}$$

for $s \in (0, 1)$ because $\sum_x \frac{\mathrm{P}_{XY}(x,y)}{\mathrm{P}_Y(y)} = \frac{\mathrm{P}_Y(y)}{\mathrm{P}_Y(y)} = 1$.

Next, we consider the case with $s \in (-1, 0)$. The same substitution to the reverse Hölder inequality (A.27) with $p = 1/(1 - s)$ and $q = \frac{1}{s}$ yields

$$e^{-sH_{1+s}(X|Y)} \ge e^{-sH^{\uparrow}_{\frac{1}{1-s}}(X|Y)}$$

because $(\sum_x \frac{\mathrm{P}_{X,Y}(x,y)}{\mathrm{P}_Y(y)})^s = (\frac{\mathrm{P}_Y(y)}{\mathrm{P}_Y(y)})^s = 1$. ■

Now, we consider the meaning of two kinds of conditional Rényi entropies. For this purpose, we discuss the case when $\mathrm{P}_{X^nY^n}$ is the independent and identical distribution of P_{XY}. Applying (2.42) and (2.44) to the distribution $\mathrm{P}_{X^n|Y^n=y}$ and taking the average with respect to y under the distribution P_{Y^n}, we have

$$\mathrm{P}_{X^nY^n}\{(x, y)|\mathrm{P}_{X^n|Y^n}(x|y) \le e^{-nR}\} \le e^{n \min_{-1\le s\le 0} s(R-H_{1+s}(X|Y))} \tag{2.89}$$

$$\mathrm{P}_{X^nY^n}\{(x, y)|\mathrm{P}_{X^n|Y^n}(x|y) > e^{-nR}\} \le e^{n \min_{s\ge 0} s(R-H_{1+s}(X|Y))}, \tag{2.90}$$

which gives an operational meaning of the conditional Rényi entropy $H_{1+s}(X|Y)$. Similarly, applying (2.50) and (2.53) to the distribution $\mathrm{P}_{X^n|Y^n=y}$ and taking the average with respect to y under the distribution P_{Y^n}, we have

$$\sum_y \mathrm{P}_{Y^n}(y) P(\mathrm{P}_{X^n|Y^n=y}, e^{nR}) \le e^{n \min_{-1\le s\le 0} \frac{s}{1+s}(R - H^{\uparrow}_{1+s}(X|Y))} \tag{2.91}$$

$$\sum_y \mathrm{P}_{Y^n}(y) P^c(\mathrm{P}_{X^n|Y^n=y}, e^{nR}) \le e^{n \min_{s\ge 0} \frac{s}{1+s}(R - H^{\uparrow}_{1+s}(X|Y))}, \tag{2.92}$$

which gives an operational meaning of the conditional Rényi entropy $H^{\uparrow}_{1+s}(X|Y)$. These inequalities clarify the difference between two kinds of conditional Rényi entropies.

Exercises

2.28 Show (2.72) and (2.73).

2.29 Show (2.76).

2.30 Show Lemma 2.3.

2.31 Show Lemma 2.1.

2.32 Show that the equality in (2.87) holds for a real $s \in (-1, 1)\backslash\{0\}$ if and only if $\mathrm{P}_{XY}(x, y) = \frac{1}{|\mathcal{X}|}\mathrm{P}_Y(y)$.

2.2 Geometry of Probability Distribution Family

2.2.1 Inner Product for Random Variables and Fisher Information

In Sect. 2.1, we introduced the mutual information $I(X : Y)$ as a quantity that expresses the correlation between two random variables X and Y. However, for calculating this quantity, one must calculate the logarithm of each probability, which is a rather tedious calculation amount. We now introduce the **covariance** $\mathrm{Cov}_p(X, Y)$ as a quantity that expresses the correlation between two real-valued random variables X and Y. Generally, calculations involving the covariance are less tedious than those of mutual information. Given a probability distribution p in a probability space Ω, the covariance is defined as

$$\mathrm{Cov}_p(X, Y) \stackrel{\mathrm{def}}{=} \sum_{\omega\in\Omega} (X(\omega) - \mathrm{E}_p(X))(Y(\omega) - \mathrm{E}_p(Y)) p(\omega). \tag{2.93}$$

If X and Y are independent, the covariance $\mathrm{Cov}_p(X, Y)$ is equal to 0[Exe. 2.33]. Thus far it has not been necessary to specify the probability distribution, and therefore we had no difficulties in using notations such as $H(X)$ and $I(X : Y)$. However, since it is important to emphasize the probability distribution treated in our discussion, we

will use the above notation without their abbreviation. If X and Y are the same, the covariance $\mathrm{Cov}_p(X, Y)$ coincides with the **variance** $\mathrm{V}_p(X)$ of X:

$$\mathrm{Cov}_p(X, Y) \stackrel{\mathrm{def}}{=} \sum_{\omega\in\Omega} (X(\omega) - \mathrm{E}_p(X))^2 p(\omega). \tag{2.94}$$

Given real-valued random variables $X_1, \ldots, X_d$, the matrix $\mathrm{Cov}_p(X_k, X_j)$ is called a **covariance matrix**. Now, starting from a given probability distribution p, we define the inner product in the space of real-valued random variables as[7]

$$\langle A, B\rangle_p^{(e)} \stackrel{\mathrm{def}}{=} \sum_{\omega} A(\omega)B(\omega)p(\omega)\,. \tag{2.95}$$

Then, the covariance $\mathrm{Cov}_p(X, Y)$ is equal to the above inner product between the two real-valued random variables $(X(\omega) - \mathrm{E}_p(X))$ and $(Y(\omega) - \mathrm{E}_p(Y))$ with a zero expectation. That is, the inner product (2.95) implies the correlation between the two real-valued random variables with zero expectation in classical systems. This inner product is also deeply related to statistical inference in another sense, as discussed below.

When we observe n independent real-valued random variables $X_1, \ldots, X_n$ identical to real-valued random variable X, the average value

$$\boldsymbol{X}^n \stackrel{\mathrm{def}}{=} \frac{X_1 + \ldots + X_n}{n} \tag{2.96}$$

converges to the expectation $\mathrm{E}_p(X)$ in probability. That is,

$$p^n\{|\boldsymbol{X}^n - \mathrm{E}_p(X)| > \epsilon\} \to 0, \quad \forall\epsilon > 0\,, \tag{2.97}$$

which is called the **law of large numbers**. Further, the distribution of the real-valued random variable

$$\sqrt{n}(\boldsymbol{X}^n - \mathrm{E}_p(X)) \tag{2.98}$$

goes to the Gaussian distribution with the variance $V = \mathrm{V}_p(X)$:

$$P_{G,V}(x) = \frac{1}{\sqrt{2\pi V}} e^{-\frac{x^2}{2V}}\,, \tag{2.99}$$

i.e.,

$$p^n\{a \le \sqrt{n}(\boldsymbol{X}^n - \mathrm{E}_p(X)) \le b\} \to \int_a^b P_{G,V}(x)dx\,, \tag{2.100}$$

[7]The superscript (e) means "exponential." This is because A corresponds to the exponential representation, as discussed later.

which is called the **central limit theorem**. Hence, the asymptotic behavior is almost characterized by the expectation $\mathrm{E}(X)$ and the variance $\mathrm{V}(X)$.

For l real-valued random variables $X_1, \ldots, X_l$, we can similarly define the real-valued random variables $\boldsymbol{X}_1^n, \ldots, \boldsymbol{X}_l^n$. These converge to their expectation in probability. The distribution of the real-valued random variables

$$(\sqrt{n}(\boldsymbol{X}_1^n - \mathrm{E}_p(X)), \ldots, \sqrt{n}(\boldsymbol{X}_k^n - \mathrm{E}_p(X))) \tag{2.101}$$

converges the k-multirate Gaussian distribution and the covariance matrix $\boldsymbol{V} = \mathrm{Cov}_p(X_k, X_j)$:

$$P_{G,\boldsymbol{V}}(x) \stackrel{\text{def}}{=} \frac{1}{\sqrt{(2\pi)^l \det \mathrm{V}}} e^{\langle x|\boldsymbol{V}^{-1}|x\rangle}. \tag{2.102}$$

Therefore, the asymptotic behavior is almost described by the expectation and the covariance matrix.

Consider the set of probability distributions p_θ parameterized by a single real number θ. For example, we can parameterize a binomial distribution with the probability space $\{0, 1\}$ by $p_\theta(0) = \theta$, $p_\theta(1) = 1 - \theta$. When the set of probability distributions is parameterized by a single parameter, it is called a **probability distribution family** and is represented by $\{p_\theta | \theta \in \Theta \subset \mathbb{R}\}$. Based on a probability distribution family, we can define the **logarithmic derivative** as $l_{\theta_0}(\omega) \stackrel{\text{def}}{=} \frac{d \log p_\theta(\omega)}{d\theta}|_{\theta=\theta_0} = \left.\frac{dp_\theta(\omega)}{d\theta}\right|_{\theta=\theta_0} / p_{\theta_0}(\omega)$. Since it is a real-valued function of the probability space, it can be regarded as a real-valued random variable. We can consider that this quantity expresses the sensitivity of the probability distribution to the variations in the parameter θ around θ_0. The **Fisher metric** (**Fisher information**) is defined as the variance of the logarithmic derivative l_{θ_0}. Since the expectation of l_{θ_0} with respect to p_{θ_0} is 0, the Fisher information can also be defined as

$$J_\theta \stackrel{\text{def}}{=} \langle l_\theta, l_\theta \rangle_{p_\theta}^{(e)}. \tag{2.103}$$

Therefore, this quantity represents the amount of variation in the probability distribution due to the variations in the parameter. Alternatively, it can indicate how much the probability distribution family represents the information related to the parameter. As discussed later, these ideas will be further refined from the viewpoint of statistical inference. The Fisher information J_θ may also be expressed as the limits of relative entropy and Hellinger distance[Exe. 2.35, 2.36]:

$$\frac{J_\theta}{2} = 4 \lim_{\epsilon \to 0} \frac{d_2^2(p_\theta, p_{\theta+\epsilon})}{\epsilon^2} \tag{2.104}$$

$$= \lim_{\epsilon \to 0} \frac{D(p_\theta \| p_{\theta+\epsilon})}{\epsilon^2} = \lim_{\epsilon \to 0} \frac{D(p_{\theta+\epsilon} \| p_\theta)}{\epsilon^2}. \tag{2.105}$$

The Fisher information J_θ is also characterized by the limit of relative Rényi entropy[Exe. 2.37]:

$$\frac{J_\theta}{2} = \lim_{\epsilon \to 0} \frac{-\phi(s|p_\theta \| p_{\theta+\epsilon})}{\epsilon^2 s(1-s)}. \tag{2.106}$$

Next, let us consider the probability distribution family $\{p_\theta | \theta \in \Theta \subset \mathbb{R}^d\}$ with multiple parameters. For each parameter, we define the logarithmic derivative $l_{\theta:k}(\omega)$ as

$$l_{\theta:k}(\omega) \stackrel{\text{def}}{=} \frac{\partial \log p_\theta(\omega)}{\partial \theta^k} = \frac{\partial p_\theta(\omega)}{\partial \theta^k} \Big/ p_\theta(\omega).$$

We use the covariance matrix $\langle l_{\theta:k}, l_{\theta:j} \rangle_{p_\theta}^{(e)}$ for the logarithmic derivatives $l_{\theta:1}, \ldots, l_{\theta:d}$ instead of the Fisher information. This matrix is called the **Fisher information matrix** and will be denoted by $\boldsymbol{J}_\theta = (J_{\theta:k,j})$. This matrix takes the role of the Fisher information when there are multiple parameters; we discuss this in greater detail below.

This inner product is closely related to the conditional expectation as follows. Suppose that we observe only the subsystem Ω_1, although the total system is given as $\Omega_1 \times \Omega_2$. Let us consider the real-valued random variable X of the total system. We denote the random variable describing the outcome in the probability space Ω_j by Z_j for $j = 1, 2$. Then, dependently of the distribution p of the total system, the **conditional expectation** $\kappa_p(X)$ of X is defined as a function of $\omega_1 \in \Omega_1$ by

$$\kappa_p(X)(\omega_1) := \sum_{\omega_2 \in \Omega_2} p(Z_2 = \omega_2 | Z_1 = \omega_1) X(\omega_1, \omega_2). \tag{2.107}$$

Then, we define the inclusion map i from the set of real-valued random variables on Ω_1 to the set of real-valued random variables on $\Omega_1 \times \Omega_2$. That is, for a random variable Y on Ω_1, the real-valued random variable $i(Y)$ on $\Omega_1 \times \Omega_2$ is defined as

$$i(Y)(\omega_1, \omega_2) = Y(\omega_1), \quad \forall (\omega_1, \omega_2) \in \Omega_1 \times \Omega_2. \tag{2.108}$$

To see the relation with the above defined inner product, we focus on an arbitrary real-valued random variable Y on Ω_1, which given as a function of Z_1. Then, the **conditional expectation** $\kappa_p(X)$ of X satisfies

$$\begin{aligned} \langle Y, \kappa_p(X) \rangle_p^{(e)} &= \sum_{\omega_1} p(Z_1 = \omega_1) Y(\omega_1) \sum_{\omega_2 \in \Omega_2} p(Z_2 = \omega_2 | Z_1 = \omega_1) X(\omega_1, \omega_2) \\ = \sum_{\omega_1, \omega_2} Y(\omega_1) X(\omega_1, \omega_2) p(Z_1 = \omega_1, Z_2 = \omega_2) &= \langle i_p(Y), X \rangle_p^{(e)}. \end{aligned} \tag{2.109}$$

In fact, when a real-valued random variable $\kappa_p(X)$ satisfies the condition (2.109) for an arbitrary real-valued random variable Y on Ω_1, it is uniquely determined because

the condition (2.109) guarantees that $\kappa_p(X)$ is the image of X for the dual map of i with respect to the inner product $\langle Y, X\rangle_p^{(e)}$. That is, when the linear space of random variables on Ω_1 is regarded as a subspace of the linear space of random variables on $\Omega_1 \times \Omega_2$ via the inclusion map i, the map $\kappa_p(X)$ is the projection from the linear space of random variables on $\Omega_1 \times \Omega_2$ to the sub linear space of random variables on Ω_1. So, we can regard the condition (2.109) as another definition of the conditional expectation $\kappa_p(X)$ of X. That is, the conditional expectation $\kappa_p(X)$ of X is the real-valued random variable describing the behavior of the random variable X of the total system $\Omega_1 \times \Omega_2$ in the subsystem Ω_1.

Generally, when we focus on a subspace $\mathfrak{U}$ of real-valued random variables for an arbitrary random variable X, we can define the **conditional expectation** $\kappa_{\mathfrak{U},p}(X) \in \mathfrak{U}$ as

$$\langle Y, \kappa_{\mathfrak{U},p}(X)\rangle_p^{(e)} = \langle Y, X\rangle_p^{(e)}, \quad \forall Y \in \mathfrak{U}. \tag{2.110}$$

This implies that the map $\kappa_{\mathfrak{U},p}(\ )$ is the projection from the space of all real-valued random variables to the subspace $\mathfrak{U}$ with respect to the inner product $\langle , \rangle_p$.

Exercises

2.33 Show that $\mathrm{Cov}_p(X, Y) = 0$ for real-valued random variables X and Y if they are independent.

2.34 Let J_θ be the Fisher information of a probability distribution family $\{p_\theta | \theta \in \Theta\}$. Let p_θ^n be the n-fold independent and identical distribution of p_θ. Show that the Fisher information of the probability distribution family $\{p_\theta^n | \theta \in \Theta\}$ at p_θ is nJ_θ.

2.35 Prove (2.104) using the second equality in (2.17), and noting that $\sqrt{1+x} \cong 1 + \frac{1}{2}x - \frac{1}{8}x^2$ for small x.

2.36 Prove (2.105) following the steps below.
(**a**) Show the following approximation with the limit $\epsilon \to 0$.

$$\log p_{\theta+\epsilon}(\omega) - \log p_\theta(\omega) \cong \frac{d \log p_\theta(\omega)}{d\theta}\epsilon + \frac{1}{2}\frac{d^2 \log p_\theta(\omega)}{d^2\theta}\epsilon^2.$$

(**b**) Prove the first equality in (2.105) using (**a**).
(**c**) Show the following approximation with the limit $\epsilon \to 0$.

$$p_{\theta+\epsilon}(\omega) \cong p_\theta(\omega) + \frac{dp_\theta(\omega)}{d\theta}\epsilon + \frac{1}{2}\frac{d^2 p_\theta(\omega)}{d^2\theta}\epsilon^2.$$

(**d**) Prove the second equality in (2.105) using (**a**) and (**c**).

2.37 Prove (2.106) using the approximation $(1+x)^s \cong 1 + sx + \frac{s(s-1)}{2}x^2$ for small x.

2.2.2 Bregman Divergence

To discuss divergence from a more general viewpoint, we formulate Bregman divergence based on a general strictly convex function $\mu(\theta)$ on $\mathbb{R}$. Assume that the strictly convex function $\mu(\theta)$ is twice-differentiable. Then, we define the **Bregman divergence** (**canonical divergence**) of $\mu(\theta)$ as

$$\begin{aligned} D^{\mu}(\bar{\theta}\|\theta) &:= \mu'(\bar{\theta})(\bar{\theta}-\theta) - \mu(\bar{\theta}) + \mu(\theta) \\ &\overset{(a)}{=} \max_{\tilde{\theta}} \mu'(\bar{\theta})(\tilde{\theta}-\theta) - \mu(\tilde{\theta}) + \mu(\theta) = \int_{\theta}^{\bar{\theta}} \mu''(\tilde{\theta})(\tilde{\theta}-\theta) d\tilde{\theta}\,. \end{aligned} \tag{2.111}$$

Here (a) can be derived as follows. Since the inside function of the maximum is concave for $\tilde{\theta}$, the maximum is realized when the derivative is zero, which implies that $\tilde{\theta} = \theta$. Hence, we obtain (a). In this case, the convex function $\mu(\theta)$ is called the **potential** of the Bregman divergence. Further, when $\bar{\theta} > \theta$, the above maximum is replaced by $\max_{\tilde{\theta}:\tilde{\theta}\ge\bar{\theta}}$.

Since the function μ is strictly convex, the correspondence $\theta \leftrightarrow \eta = \frac{d\mu}{d\theta}$ is one-to-one. Hence, the divergence $D^{\mu}(\bar{\theta}\|\theta)$ can be expressed with the parameter η. For this purpose, we define the **Legendre transform** ν of μ

$$\nu(\eta) \overset{\text{def}}{=} \max_{\tilde{\theta}} \eta\tilde{\theta} - \mu(\tilde{\theta}). \tag{2.112}$$

Then, the function ν is a convex function[Exe. 2.38], and we can recover the functions μ and θ as

$$\mu(\theta) = \max_{\tilde{\eta}} \theta\tilde{\eta} - \nu(\tilde{\eta}), \quad \theta = \frac{d\nu}{d\eta}\,.$$

Due to the inverse function theorem, the second derivative ($\frac{d^2\nu}{d\eta^2}$ of ν is calculated to $\frac{d\theta}{d\eta} = \frac{d\eta}{d\theta}^{-1} = \frac{d^2\mu}{d\theta^2}^{-1}$.

In particular, when $\eta = \frac{d\mu}{d\theta}(\theta)$,

$$\nu(\eta) = \theta\eta - \mu(\theta) = D^{\mu}(\theta\|0) - \mu(0), \tag{2.113}$$

$$\mu(\theta) = \theta\eta - \nu(\eta) = D^{\nu}(\eta\|0) - \nu(0). \tag{2.114}$$

Using these relations, we can obtain

$$D^{\mu}(\bar{\theta}\|\theta) = D^{\nu}(\eta\|\bar{\eta}) = \theta(\eta-\bar{\eta}) - \nu(\eta) + \nu(\bar{\eta}). \tag{2.115}$$

That is, the Bregman divergence of μ can be written by the Bregman divergence of the Legendre transform of μ.

Now, we extend Bregman to the multi-parametric case. Let $\mu(\theta)$ be a twice-differentiable and strictly convex function defined on a subset Θ of the d-dimensional real vector space $\mathbb{R}^d$. The Bregman divergence concerning the convex function μ is defined by

$$D^\mu(\bar{\theta}\|\theta) \stackrel{\text{def}}{=} \sum_k \eta_k(\bar{\theta})(\bar{\theta}^k - \theta^k) - \mu(\bar{\theta}) + \mu(\theta), \quad \eta_k(\theta) \stackrel{\text{def}}{=} \frac{\partial \mu}{\partial \theta^k}(\theta)\,. \tag{2.116}$$

This quantity has the following two characterizations:

$$D^\mu(\bar{\theta}\|\theta) = \max_{\tilde{\theta}} \sum_k \frac{\partial \mu}{\partial \theta^k}(\bar{\theta})(\tilde{\theta}^k - \theta^k) - \mu(\tilde{\theta}) + \mu(\theta) \tag{2.117}$$

$$= \int_0^1 \sum_{k,j} (\bar{\theta}^k - \theta^k)(\bar{\theta}^j - \theta^j) \frac{\partial^2 \mu}{\partial \theta^k \partial \theta^j}(\theta + (\bar{\theta} - \theta)t) t dt\,. \tag{2.118}$$

Since the strict positivity of μ implies the strict positivity of inside of the above integral, $D^\mu(\bar{\theta}\|\theta)$ is strictly positive unless $\bar{\theta} = \theta$. The strict positivity of μ is also guarantees that the correspondence $\theta^k \leftrightarrow \eta_k = \frac{\partial \mu}{\partial \theta^k}$ is one-to-one. Hence, the Bregman divergence $D^\mu(\bar{\theta}\|\theta)$ can be expressed with the parameter η. For this purpose, we define the Legendre transform ν of μ

$$\nu(\eta) \stackrel{\text{def}}{=} \max_{\tilde{\theta}} \sum_k \eta_k \tilde{\theta}^k - \mu(\tilde{\theta}). \tag{2.119}$$

Then, the function ν is a convex function[Exe. 2.38], and we can recover the functions μ and θ as

$$\mu(\theta) = \max_{\tilde{\eta}} \sum_k \theta^k \tilde{\eta}_k - \nu(\tilde{\eta}), \quad \theta^k = \frac{\partial \nu}{\partial \eta_k}\,. \tag{2.120}$$

Due to the inverse function theorem, the second derivative matrix $(\frac{\partial^2 \nu}{\partial \eta_k \partial \eta_j})_{k,j}$ of ν is calculated to $(\frac{\partial \theta^k}{\partial \eta_j})_{k,j} = ((\frac{\partial \eta^k}{\partial \theta_j})_{k,j})^{-1} = ((\frac{\partial^2 \mu}{\partial \theta^k \partial \theta^j})_{k,j})^{-1}$, which is the inverse of the matrix $\frac{\partial^2 \mu}{\partial \theta^k \partial \theta^j}$.

In particular, when $\eta_k = \frac{\partial \mu}{\partial \theta^k}(\theta)$,

$$\nu(\eta) = \sum_k \eta_k \theta^k - \mu(\theta) = D^\mu(\theta\|0) - \mu(0), \tag{2.121}$$

$$\mu(\theta) = \sum_k \theta^k \eta_k - \nu(\eta) = D^\nu(\eta\|0) - \nu(0). \tag{2.122}$$

Using these relations, we can characterize the Bregman divergence concerning the convex function μ by the Bregman divergence concerning the convex function ν as

$$D^{\mu}(\bar{\theta}\|\theta) = D^{\nu}(\eta\|\bar{\eta}) = \sum_k \theta^k(\eta_k - \bar{\eta}_k) - \nu(\eta) + \nu(\bar{\eta}) \tag{2.123}$$

$$= \int_0^1 \sum_{k,j} (\eta_k(\theta) - \eta_k(\bar{\theta}))(\eta_j(\theta) - \eta_j(\bar{\theta})) \frac{\partial^2 \nu}{\partial \eta_k \partial \eta_j}(\eta(\bar{\theta}) + (\eta(\bar{\theta}) - \eta(\theta))t) t dt, \tag{2.124}$$

where (2.124) follows from (2.118) for the Bregman divergence with respect to ν.

A subset $\mathcal{E}$ of Θ is called an exponential subfamily of Θ when there exist an element $\theta' \in \Theta$ and l independent vectors $v_1, \ldots, v_l \in \mathbb{R}^d$ such that $\mathcal{E} = \{\theta \in \Theta | \theta = \theta' + \sum_{j=1}^k a^j v_j \exists (a^1, \ldots, a^l) \in \mathbb{R}^l\}$. A subset $\mathcal{M}$ of Θ is called a mixture subfamily of Θ when there exist a l-dimensional vector $(b_1, \ldots, b_l)$ and l independent vectors $v_1, \ldots, v_l \in \mathbb{R}^d$ such that $\mathcal{M} = \{\theta \in \Theta | b_k = \sum_{j=1}^d v_k^j \eta_j(\theta)\}$. In particular, the set of vectors $\{v_1, \ldots, v_l\}$ is called a generator of $\mathcal{E}$ and $\mathcal{M}$, respectively.

Now, we focus on two points $\theta' = (\theta'^1, \ldots, \theta'^d)$ and $\theta'' = (\theta''^1, \ldots, \theta''^d)$. We choose the exponential subfamily $\mathcal{E}$ of Θ whose natural parameters $\theta^{l+1}, \ldots, \theta^d$ are fixed to $\theta''^{l+1}, \ldots, \theta''^d$, and the mixture subfamily $\mathcal{M}$ of Θ whose expectation parameters $\eta^1, \ldots, \eta^l$ are fixed to $\eta(\theta')^1, \ldots, \eta(\theta')^l$. Let $\tilde{\theta} = (\tilde{\theta}^1, \ldots, \tilde{\theta}^d)$ be an element of the intersection of these two subfamily of Θ. That is, $\tilde{\theta}^j = \theta''^j$ for $j = l+1, \ldots, d$ and $\eta_j(\tilde{\theta}) = \eta_j(\theta')$ for $j = 1, \ldots, l$.

Then, since

$$\theta'^j - \theta''^j = \begin{cases} (\theta'^j - \tilde{\theta}^j) & \text{if } j \geq l+1 \\ (\theta'^j - \tilde{\theta}^j) + (\tilde{\theta}^j - \theta''^j) & \text{if } j \leq l, \end{cases} \tag{2.125}$$

the definition (2.116) implies that

$$\begin{aligned} D^{\mu}(\theta'\|\theta'') &= \sum_{j=1}^d (\theta'^j - \theta''^j)\eta_j(\theta') - \mu(\theta') + \mu(\theta'') \\ &= \sum_{j=1}^d (\theta'^j - \tilde{\theta}^j)\eta_j(\theta') - \mu(\theta') + \mu(\tilde{\theta}) + \sum_{j=1}^l (\tilde{\theta}^j - \theta''^j)\eta_j(\tilde{\theta}) - \mu(\tilde{\theta}) + \mu(\theta'') \\ &= D^{\mu}(\theta'\|\tilde{\theta}) + D^{\mu}(\tilde{\theta}\|\theta''). \end{aligned} \tag{2.126}$$

Using (2.126), we obtain the **Pythagorean theorem** [2] as follows.

Theorem 2.3 (Amari [6]) *Given an element $\theta \in \Theta$ and a mixture subfamily $\mathcal{M}$ of Θ with the generator $\{v_1, \ldots, v_l\}$, we define $\theta^* := \mathrm{argmin}_{\theta' \in \mathcal{M}} D^{\mu}(\theta'\|\theta)$. Then, we obtain the following two items as Fig. 2.1.*

(1) Any element $\theta' \in \mathcal{M}$ satisfies $D^{\mu}(\theta'\|\theta) = D^{\mu}(\theta'\|\theta^*) + D^{\mu}(\theta^*\|\theta)$.
(2) The element θ^* is the unique element of the intersection of the mixture subfamily $\mathcal{M}$ and the exponential subfamily $\mathcal{E}$ containing θ with the generator $\{v_1, \ldots, v_l\}$.

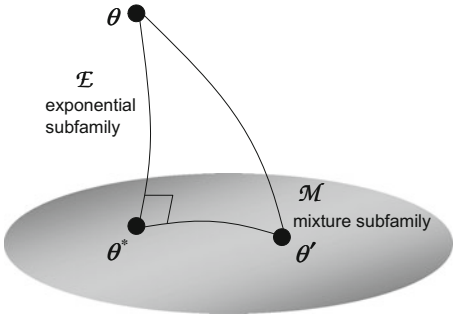

Fig. 2.1 Pythagorean theorem

Proof Choose an element $\tilde{\theta}$ in the intersection of the mixture subfamily $\mathcal{M}$ and the exponential subfamily $\mathcal{E}$ containing θ with the generator $\{v_1, \ldots, v_l\}$. Now, we choose additional vectors $\{v_{l+1}, \ldots, v_d\}$ such that the set $\{v_1, \ldots, v_d\}$ forms a basis. Then, we introduce another coordinate a^j such that $\sum_{j=1}^{d} a^j v_j^l = \theta^l$. Now, we apply the new coordinate a^j to the relation (2.126). Thus, any element $\theta' \in \mathcal{M}$ satisfies that $D^{\mu}(\theta'\|\theta) = D^{\mu}(\theta'\|\tilde{\theta}) + D^{\mu}(\tilde{\theta}\|\theta)$. Since $D^{\mu}(\theta'\|\tilde{\theta}) > 0$ except for $\theta' = \tilde{\theta}$, we have $\min_{\theta' \in \mathcal{M}} D^{\mu}(\theta'\|\theta) = D^{\mu}(\tilde{\theta}\|\theta)$, which implies that $\theta^* = \tilde{\theta}$, i.e., (2). Hence, we obtain (1). ■

We also have another version of the Pythagorean theorem as follows.

Theorem 2.4 (Amari [6]) *Given an element $\theta' \in \Theta$ and an exponential subfamily $\mathcal{E}$ of Θ with the generator $\{v_1, \ldots, v_l\}$, we define $\theta'_* := \mathrm{argmin}_{\theta \in \mathcal{E}} D^{\mu}(\theta'\|\theta)$.*

(1) Any element $\theta \in \mathcal{E}$ satisfies $D^{\mu}(\theta'\|\theta) = D^{\mu}(\theta'\|\theta'_*) + D^{\mu}(\theta'_*\|\theta)$.

(2) The element θ'_* is the unique element of the intersection of the exponential subfamily $\mathcal{E}$ and the mixture subfamily containing θ' with the generator $\{v_1, \ldots, v_l\}$.

Exercises

2.38 Show that $\nu(\eta)$ is a convex function.

2.39 Solve Exercise 2.23 by using Theorem 2.3.

2.2.3 Exponential Family and Divergence

In Sect. 2.1, relative entropy $D(p\|q)$ is defined. In this subsection, we characterize it as Bregman divergence.

Let $p(\omega)$ be a probability distribution and $X(\omega)$ be a real-valued random variable. When the family $\{p_\theta | \theta \in \Theta\}$ has the form

$$p_\theta(\omega) = p(\omega) e^{\theta X(\omega) - \mu(\theta)}, \tag{2.127}$$

$$\mu(\theta) \stackrel{\mathrm{def}}{=} \log \sum_{\omega} p(\omega) e^{\theta X(\omega)}, \tag{2.128}$$

the logarithmic derivative at respective points equals the logarithmic derivative at a fixed point with the addition of a constant. In this case, the family, X, and $\mu(\theta)$ are called an **exponential family**, the generator, and the **cumulant generating function** of X, respectively. In particular, in an exponential family, the logarithmic derivative does not depend on the point θ except for constant differences. Hence, it is often called the exponential (e) representation of the derivative. Therefore, we use the superscript (e) in the inner product $\langle\, , \,\rangle_p^{(e)}$. The function $\mu(\theta)$ is often called a **potential function** in the context of information geometry. Since the first derivative of $\mu(\theta)$ is calculated as $\mu'(\theta) = (\frac{d}{d\theta} e^{\mu(\theta)}) e^{-\mu(\theta)} = \sum_\omega p_\theta(\omega) X(\omega)$, the second derivative is as

$$\begin{aligned}\mu''(\theta) &= \left(\frac{d^2}{d\theta^2} e^{\mu(\theta)}\right) e^{-\mu(\theta)} - \left(\left(\frac{d}{d\theta} e^{\mu(\theta)}\right) e^{-\mu(\theta)}\right)^2 \\ &= \sum_\omega p_\theta(\omega) X(\omega)^2 - \left(\sum_\omega p_\theta(\omega) X(\omega)\right)^2 = J_\theta > 0,\end{aligned}$$

is the Fisher information. So, the cumulant generating function $\mu(\theta)$ is a strictly convex function. Therefore, the first derivative $\mu'(\theta) = \sum_\omega p_\theta(\omega) X(\omega)$ is monotone increasing. That is, we may regard it as another parameter identifying the distribution p_θ, and denote it by η. The original parameter θ is called a **natural parameter**, and the other parameter η is an **expectation parameter**. When the distribution is parametrized by the expectation parameter η, it is written as $\hat{p}_\eta$. Hence, we have $\hat{p}_{\eta(\theta)} = p_\theta$.

For example, in the one-trial binomial distribution, the generator X is given as $X(i) = i$, and the distribution p_0 is given as $p_0(i) = \frac{1}{2}$, for $i = 0, 1$. Then, the cumulant generating function μ is calculated to be $\mu(\theta) = \log \frac{1+e^\theta}{2}$. The distribution is written as $p_\theta(0) = 1/(1+e^\theta)$, $p_\theta(1) = e^\theta/(1+e^\theta)$ in the natural parameter θ. Hence, the binomial distribution is an exponential family. The expectation parameter is $\eta(\theta) = e^\theta/(1+e^\theta)$. That is, the distribution is written as $\hat{p}_\eta(1) = \eta$, $\hat{p}_\eta(0) = 1-\eta$ in the expectation parameter η.

Since $\mu(\theta)$ is twice-differentiable and strictly convex, we can consider the Bregman divergence of $\mu(\theta)$. Then, the divergence $D(p_{\bar{\theta}} \| p_\theta)$ can be written by using the Bregman divergence of $\mu(\theta)$ as follows.

$$\begin{aligned}&D(p_{\bar{\theta}} \| p_\theta) = D(\hat{p}_{\eta(\bar{\theta})} \| \hat{p}_{\eta(\theta)}) = (\bar{\theta} - \theta)\eta(\bar{\theta}) - \mu(\bar{\theta}) + \mu(\theta) \\ =&D^\mu(\bar{\theta} \| \theta) = \int_\theta^{\bar{\theta}} J_{\tilde{\theta}}(\tilde{\theta} - \theta) d\tilde{\theta} = \max_{\tilde{\theta}} (\tilde{\theta} - \theta)\eta(\bar{\theta}) - \mu(\tilde{\theta}) + \mu(\theta)\,. \end{aligned} \tag{2.129}$$

where equations in (2.129) follow from (2.111). When $\bar{\theta} > \theta$, the above maximum is replaced by $\max_{\tilde{\theta}: \tilde{\theta} \ge \bar{\theta}}$.

Next, we consider the multi-parameter case. Let $X_1(\omega), \ldots, X_d(\omega)$ be d real-valued random variables. We can define a d-parameter exponential family

$$p_\theta(\omega) \stackrel{\text{def}}{=} p(\omega) e^{\sum_k \theta^k X_k(\omega) - \mu(\theta)}, \quad \mu(\theta) \stackrel{\text{def}}{=} \log \sum_\omega p(\omega) e^{\sum_k \theta^k X_k(\omega)}\,. \tag{2.130}$$

The parameters θ^k are natural parameters, and the other parameters

$$\eta_k(\theta) \stackrel{\text{def}}{=} \frac{\partial \mu}{\partial \theta^k} = \sum_\omega p_\theta(\omega) X_k(\omega) \tag{2.131}$$

are expectation parameters. Since the second derivative $\frac{\partial^2 \mu(\theta)}{\partial \theta^j \partial \theta^k}$ is equal to the Fisher information matrix $J_{\theta:k,j}$, the cumulant generating function $\mu(\theta)$ is a convex function. Using (2.118), we obtain

$$D(p_{\bar{\theta}} \| p_\theta) = \int_0^1 \sum_{k,j} (\bar{\theta}^k - \theta^k)(\bar{\theta}^j - \theta^j) J_{\theta + (\bar{\theta} - \theta)t:k,j} t dt \tag{2.132}$$

similar to (2.129). Since the second derivative matrix $(\frac{\partial^2 \nu}{\partial \eta_k \partial \eta_j})_{k,j}$ of ν appearing in (2.124) is the inverse of the matrix $\frac{\partial^2 \mu}{\partial \theta^k \partial \theta^j}$, the application of (2.124) yields that

$$D(p_{\bar{\theta}} \| p_\theta) = \int_0^1 \sum_{k,j} (\eta_k(\theta) - \eta_k(\bar{\theta}))(\eta_j(\theta) - \eta_j(\bar{\theta}))(J^{-1}_{\theta(t)})^{k,j} t dt, \tag{2.133}$$

where $\theta(t)$ is defined as $\eta(\theta(t)) = \eta(\bar{\theta}) + (\eta(\bar{\theta}) - \eta(\theta))t$. Note that the inverse matrix $J^{-1}_{\theta(t)}$ is the Fisher information matrix with respect to the parameter η.

In what follows, we consider the case where p is the uniform distribution p_{mix}. Let the real-valued random variables $X_1(\omega), \ldots, X_d(\omega)$ be a basis of the space $\mathcal{R}_0(\Omega)$ of random variables that have expectation 0 under the uniform distribution p_{mix}. We also choose the dual basis $Y^1(\omega), \ldots, Y^k(\omega)$ of the space $\mathcal{R}_0(\Omega)$ satisfying $\sum_\omega Y^k(\omega) X_j(\omega) = \delta^k_j$. Then, any distribution p can be parameterized by the expectation parameter as

$$p(\omega) = \hat{p}_{\eta(\theta)}(\omega) := p_{\text{mix}}(\omega) + \sum_i \eta_k(\theta) Y^k(\omega)$$

because $p - p_{\text{mix}}$ can be regarded as an element of $\mathcal{R}_0(\Omega)$.

From (2.123) and (2.120),

$$D(\hat{p}_{\bar{\eta}} \| \hat{p}_\eta) = D^\nu(\eta \| \bar{\eta}) = \sum_k \frac{\partial \nu}{\partial \eta_k}(\eta_k - \bar{\eta}_k) - \nu(\eta) + \nu(\bar{\eta}), \tag{2.134}$$

$$\nu(\eta) = D(\hat{p}_\eta \| p_{\text{mix}}) = -H(\hat{p}_\eta) + H(p_{\text{mix}}) \tag{2.135}$$

because $\mu(0) = 0$. The second derivative matrix of ν is the inverse of the second derivative matrix of μ, i.e., the Fisher information matrix concerning the natural parameter θ. That is, the second derivative matrix of ν coincides with the Fisher information matrix concerning the expectation parameter η.

Now, for given distributions p and q, we consider the case when $Y^1(\omega) = q(\omega) - p(\omega)$. In this case, the distribution $\hat{p}_t := (1-t)p + tq$ $(0 \le t \le 1)$ depends on the

first expectation parameter η_1. Other expectation parameters η_k are constants for the distribution p_t. Hence, $\eta_1(\hat{p}_t) - \eta_1(\hat{p}_{t'}) = t - t'$ and $\eta_k(\hat{p}_t) - \eta_k(\hat{p}_{t'}) = 0$ for $k \geq 2$. Thus, as a special case of (2.133), we have

$$D(p\|q) = \int_0^1 J_t t dt \,, \tag{2.136}$$

where J_t is the Fisher information for the parameter t.

2.3 Estimation in Classical Systems

An important problem in mathematical statistics is the estimation of the parameter θ from some given data $\omega \in \Omega$ for a probability distribution that generates the data. To solve this problem, a mapping $\hat{\theta}$ called an estimator from the probability space Ω to the parameter space $\Theta \subset \mathbb{R}$ is required. The accuracy of the estimator is most commonly evaluated by the **mean square error**, which is the expectation of the square of the difference $\hat{\theta} - \theta$:

$$\mathrm{V}_\theta(\hat{\theta}) \stackrel{\mathrm{def}}{=} \mathrm{E}_{p_\theta}((\hat{\theta} - \theta)^2), \tag{2.137}$$

where θ is the true parameter. Note that sometimes the mean square error is not the same as the variance $\mathrm{V}_{p_\theta}(X)$. The estimator

$$\mathrm{E}_\theta(\hat{\theta}) \stackrel{\mathrm{def}}{=} \mathrm{E}_{p_\theta}(\hat{\theta}) = \theta, \quad \forall \theta \in \Theta \tag{2.138}$$

is called an **unbiased estimator**, and such estimators form an important class of estimators. The mean square error of the unbiased estimator $\hat{\theta}$ satisfies the **Cramér–Rao inequality**

$$\mathrm{V}_\theta(\hat{\theta}) \geq J_\theta^{-1} \,. \tag{2.139}$$

When an unbiased estimator attains the RHS of (2.139), it is called **efficient**. This inequality can be proved from the relations

$$\langle (\hat{\theta} - \theta), l_{\theta_0} \rangle_p^{(e)} = \left. \frac{d\mathrm{E}_\theta(\hat{\theta} - \theta_0)}{d\theta} \right|_{\theta=\theta_0} = 1$$

and

$$\langle (\hat{\theta} - \theta_0), (\hat{\theta} - \theta_0) \rangle_{p_{\theta_0}}^{(e)} \langle l_{\theta_0}, l_{\theta_0} \rangle_{p_{\theta_0}}^{(e)} \geq \left| \langle (\hat{\theta} - \theta), l_{\theta_0} \rangle_{p_{\theta_0}}^{(e)} \right|^2 = 1 \,, \tag{2.140}$$

which follows from Schwarz's inequality. The equality of (2.139) holds for every value of θ if and only if the probability distribution family is a one-parameter exponential family (2.127) and the expectation parameter $\eta(\theta) = \sum_\omega X(\omega) p_\theta(\omega)$ is to be estimated. In this case, the efficient estimator for the expected parameter is given as $\hat{\eta}(\omega) := X(\omega)$ (Exercise 2.40). Even in the estimation for an exponential family, there is necessarily no estimator for the natural parameter θ in (2.127) such that the equality of (2.139) holds for all θ.

Let n data $\omega^n = (\omega_1, \ldots, \omega_n) \in \Omega^n$ be generated with the n-i.i.d. of the probability distribution p_θ. The estimator may then be given by the mapping $\hat{\theta}^n$ from Ω^n to $\Theta \subset \mathbb{R}$. In this case, the Fisher information of the probability distribution family is nJ_θ, and the unbiased estimator $\hat{\theta}^n$ satisfies the Cramér–Rao inequality

$$\mathrm{V}_\theta(\hat{\theta}^n) \geq \frac{1}{n} J_\theta^{-1}.$$

However, in general, it is not necessary to restrict our estimator to unbiased estimators. In fact, rare estimators satisfy such conditions for finite n.

Therefore, in mathematical statistics, we often study problems in the asymptotic limit $n \to \infty$ rather than those with a finite number of data elements. For this purpose, let us apply the asymptotic unbiasedness conditions

$$\lim_{n\to\infty} \mathrm{E}_\theta(\hat{\theta}_n) = \theta, \quad \lim_{n\to\infty} \frac{d}{d\theta} \mathrm{E}_\theta(\hat{\theta}_n) = 1, \quad \forall \theta \in \Theta \tag{2.141}$$

to a sequence of estimators $\{\hat{\theta}^n\}$. Evaluating the accuracy with $\underline{\lim}\, n\mathrm{V}_\theta(\hat{\theta}_n)$, we have the **asymptotic Cramér–Rao inequality**[8]:

$$\underline{\lim}\, n\mathrm{V}_\theta(\hat{\theta}_n) \geq J_\theta^{-1}, \tag{2.142}$$

which is shown as follows. Based on a derivation similar to (2.139), we obtain

$$nJ_\theta \mathrm{V}_\theta(\hat{\theta}_n) \geq \left| \frac{d}{d\theta} \mathrm{E}_\theta(\hat{\theta}_n) \right|^2. \tag{2.143}$$

Combination of (2.141) and (2.143) derives Inequality (2.142).

Now, we consider what estimator attains the lower bound of (2.142). The **maximum likelihood estimator** $\hat{\theta}_{n,ML}(\omega^n)$

$$\hat{\theta}_{n,ML}(\omega^n) = \operatorname*{argmax}_{\theta\in\Theta} p_\theta^n(\omega^n) \tag{2.144}$$

[8]This inequality still holds even if the asymptotic unbiasedness condition is replaced by another weak condition. Indeed, it is a problem to choose a suitable condition to be assumed for the inequality (2.142). For details, see van der Vaart [7].

achieves this lower bound, and the limit of its mean squared error is equal to J_θ^{-1} [7]. Indeed, in an exponential family with the expectation parameter, the maximum likelihood estimator is equal to the efficient estimator[Exe. 2.41]. Hence, the maximum likelihood estimator plays an important role in statistical inference.[9]

Indeed, we choose the mean square error as the criterion of estimation error because (1) its mathematical treatment is easy and (2) in the i.i.d. case, the sample mean can be characterized by a Gaussian distribution. Hence, we can expect that a suitable estimator will also approach a Gaussian distribution asymptotically. That is, we can expect that its asymptotic behavior will be characterizable by the variance. In particular, the maximum likelihood estimator $\hat{\theta}_{n,ML}$ obeys the Gaussian distribution asymptotically:

$$p_\theta^n\{a \le \sqrt{n}(\hat{\theta}_{n,ML} - \theta) \le b\} \to \int_a^b P_{G,1/J_\theta}(x)dx, \quad \forall a, b\,.$$

Let us now consider the probability distribution family $\{p_\theta | \theta \in \Theta \subset \mathbb{R}^d\}$ with multiple parameters. We focus on the Fisher information matrix $\boldsymbol{J}_\theta = (J_{\theta:k,j})$, which was defined at the end of Sect. 2.2.1, instead of the Fisher information. The estimator is given by the map $\hat{\theta} = (\hat{\theta}^1, \ldots, \hat{\theta}^d)$ from the probability space Ω to the parameter space Θ, similar to the one-parameter case. The unbiasedness conditions are

$$\mathrm{E}_\theta^k(\hat{\theta}) \stackrel{\mathrm{def}}{=} \mathrm{E}_{p_\theta}(\hat{\theta}^k) = \theta^k, \quad \forall \theta \in \Theta, 1 \le \forall k \le d\,.$$

The error can be calculated using the **mean square error matrix** $\boldsymbol{V}_\theta(\hat{\theta}) = (\mathrm{V}_\theta^{k,j}(\hat{\theta}))$:

$$\mathrm{V}_\theta^{k,j}(\hat{\theta}) \stackrel{\mathrm{def}}{=} \mathrm{E}_{p_\theta}((\hat{\theta}^k - \theta^k)(\hat{\theta}^j - \theta^j)).$$

Then, we obtain the **multiparameter Cramér–Rao inequality**

$$\boldsymbol{V}_\theta(\hat{\theta}) \ge \boldsymbol{J}_\theta^{-1}\,. \tag{2.145}$$

Proof of (2.145) For the proof, let us assume that any vectors $|b\rangle = (b_1, \ldots, b_d)^T \in \mathbb{C}^d$ and $|a\rangle \in \mathbb{C}^d$ satisfy

$$\langle b|\boldsymbol{V}_\theta(\hat{\theta})b\rangle\langle a|\boldsymbol{J}_\theta|a\rangle \ge |\langle b|a\rangle|^2\,. \tag{2.146}$$

By substituting $a = (\boldsymbol{J}_\theta)^{-1}b$, inequality (2.146) becomes

$$\langle b|\boldsymbol{V}_\theta(\hat{\theta})|b\rangle \ge \langle b|(\boldsymbol{J}_\theta)^{-1}|b\rangle$$

[9] This is generally true for all probability distribution families, although some regularity conditions must be imposed. For example, consider the case in which Ω consists of finite elements. These regularity conditions are satisfied when the first and second derivatives with respect to θ are continuous. Generally, the central limit theorem is used in the proof [7].

since $(\boldsymbol{J}_\theta)^{-1}$ is a symmetric matrix. Therefore, we obtain (2.145) if (2.146) holds. Now, we prove (2.146) as follows. Since

$$\delta_k^j = \left.\frac{\partial \mathrm{E}_\theta^j(\hat{\theta}) - \theta_0^j}{\partial \theta^k}\right|_{\theta=\theta_0} = \left\langle l_{\theta_0:k}, (\hat{\theta}^j - \theta_0^j)\right\rangle_{\theta_0}^{(e)},$$

similarly to the proof of (2.139), the Schwarz inequality yields

$$\begin{aligned}\langle b|\boldsymbol{V}_{\theta_0}(\hat{\theta})b\rangle &= \left\langle \left(\sum_{k=1}^d (\hat{\theta}^k - \theta_0^k)b_k\right), \left(\sum_{k=1}^d (\hat{\theta}^k - \theta_0^k)b_k\right)\right\rangle_{p_{\theta_0}}^{(e)} \\ &\geq \frac{\left|\left\langle \left(\sum_{k=1}^d l_{\theta_0:k}a_k\right), \left(\sum_{k=1}^d (\hat{\theta}^k(\omega) - \theta_0^k)b_k\right)\right\rangle_{p_{\theta_0}}^{(e)}\right|^2}{\left\langle \left(\sum_{k=1}^d l_{\theta_0:k}a_k\right), \left(\sum_{k=1}^d l_{\theta_0:k}a_k\right)\right\rangle_{p_{\theta_0}}^{(e)}} = \frac{|\langle a|b\rangle|^2}{\langle a|\boldsymbol{J}_{\theta_0}|a\rangle}.\end{aligned}$$

■

Moreover, since the sequence of estimators $\{\hat{\theta}_n = (\hat{\theta}_n^1, \ldots, \hat{\theta}_n^d)\}$ satisfies the asymptotic unbiasedness condition

$$\lim_{n\to\infty} \mathrm{E}_\theta^k(\hat{\theta}_n) = \theta^k, \quad \lim_{n\to\infty} \frac{\partial}{\partial \theta^j}\mathrm{E}_\theta^k(\hat{\theta}_n) = \delta_j^k, \quad \forall \theta \in \Theta, \tag{2.147}$$

the asymptotic Cramér–Rao inequality for the multiparameter case

$$\boldsymbol{V}_\theta(\{\hat{\theta}_n\}) \geq \boldsymbol{J}_\theta^{-1} \tag{2.148}$$

holds if the limit $\boldsymbol{V}_\theta(\{\hat{\theta}_n\}) \stackrel{\mathrm{def}}{=} \lim_{n\to\infty} n\boldsymbol{V}_\theta(\hat{\theta}_n)$ exists. Next, we prove (2.148). Defining $A_{n,i}^j \stackrel{\mathrm{def}}{=} \frac{\partial}{\partial \theta^j}\mathrm{E}_\theta^k(\hat{\theta}_n)$, we have

$$n\langle a|\boldsymbol{J}_\theta|a\rangle\langle b|\boldsymbol{V}_\theta(\hat{\theta}_n)|b\rangle \geq |\langle a|\mathbf{A}_n|b\rangle|^2$$

instead of (2.146). We then obtain

$$\langle a|\boldsymbol{J}_\theta|a\rangle\langle b|\boldsymbol{V}_\theta(\{\hat{\theta}_n\})|b\rangle \geq |\langle a|b\rangle|^2,$$

from which (2.148) may be obtained in a manner similar to (2.145).

Similarly to the one-parameter case, the equality of (2.145) holds if and only if the following conditions hold: (1) The probability distribution family is a multiparameter exponential family. (2) The expectation parameter η is to be estimated. (3) The estimator for η is given by

$$\hat{\eta}_k(\omega) = X_k(\omega). \tag{2.149}$$

In this case, this estimator (2.149) equals the maximum likelihood estimator $\hat{\theta}_{n,ML} = (\hat{\theta}^1_{n,ML}, \dots \hat{\theta}^d_{n,ML})$ defined by (2.144)[Exe. 2.41], i.e.,

$$\max_{\eta} \hat{p}_\eta(\omega) = \hat{p}_{X_k(\omega)}(\omega) . \tag{2.150}$$

A probability distribution family does not necessarily have such an estimator; however, a maximum likelihood estimator $\hat{\theta}_{n,ML}$ can be defined by (2.144). This satisfies the asymptotic unbiasedness property (2.147) in a similar way to (2.144), and it satisfies the equality of (2.148). Moreover, it is known that the maximum likelihood estimator $\hat{\theta}_{n,ML}$ satisfies [7]

$$\boldsymbol{V}_\theta(\{\hat{\theta}_n\}) = \boldsymbol{J}_\theta^{-1} .$$

Note that this inequality holds independently of the choice of coordinate. Hence, for a large amount of data, it is best to use the maximum likelihood estimator. Its mean square error matrix is almost in inverse proportion to the number of observations n. This coefficient of the optimal case is given by the Fisher information matrix. Therefore, the Fisher information matrix can be considered to yield the best accuracy of an estimator.

Indeed, usually any statistical decision with the given probability distribution family $\{q_\gamma | \gamma \in \Gamma\}$ is based on the likelihood ratio $\log q_\gamma(\omega) - \log q_{\gamma'}(\omega)$. For example, the maximum likelihood estimator depends only on the likelihood ratio. A probability distribution family $\{q_\gamma | \gamma \in \Gamma\}$ is called a **curved exponential family** when it belongs to a larger multiparameter exponential family $\{p_\theta | \theta \in \Theta\}$, i.e., q_γ is given as $p_{\theta(\gamma)}$ with use of a function $\theta(\gamma)$. When $p_\theta(\omega)$ is given by (2.130), the likelihood ratio can be expressed by the relative entropy

$$\begin{aligned}
&\log q_\gamma(\omega) - \log q_{\gamma'}(\omega) = \log p_{\theta(\gamma)}(\omega) - \log p_{\theta(\gamma')}(\omega) \\
=& \sum_k (\theta(\gamma)^k - \theta(\gamma')^k) X_k(\omega) - \mu(\theta(\gamma)) + \mu(\theta(\gamma')) \\
=& \sum_k X_k(\omega)(\theta''^k - \theta(\gamma')^k) + \mu(\theta(\gamma')) - \mu(\theta'') \\
& - \left(\sum_k X_k(\omega)(\theta''^k - \theta(\gamma)^k) + \mu(\theta(\gamma)) - \mu(\theta'') \right) \\
=& D(\hat{p}_{X(\omega)} \| q_{\gamma'}) - D(\hat{p}_{X(\omega)} \| q_\gamma),
\end{aligned} \tag{2.151}$$

where θ'' is chosen as $\eta_k(\theta'') = X_k(\omega)$. That is, our estimation procedure can be treated from the viewpoint of the relative entropy geometry.

Exercises

2.40 Show that the following two conditions are equivalent for a probability distribution family $\{p_\theta | \theta \in \mathbb{R}\}$ and its estimator X by following the steps below.

① There exists a parameter η such that the estimator X is an unbiased estimator for the parameter η and the equality of (2.139) holds at all points.

② The probability distribution family $\{p_\theta | \theta \in \mathbb{R}\}$ is an exponential family, $p_\theta(\omega)$ is given by (2.127) using X, and the parameter to be estimated is the expectation parameter $\eta(\theta)$.

(**a**) Show that the estimator X is an unbiased estimator of the expectation parameter under the exponential family (2.127).
(**b**) Show that ① may be deduced from ②.
(**c**) For the exponential family (2.127), show that the natural parameter θ is given as a function of the expectation parameter η with the form $\theta = \int_0^\eta J_{\eta'}\, d\eta'$.
(**d**) Show that $\mu(\theta(\eta)) = \int_0^\eta \eta' J_{\eta'}\, d\eta'$.
(**e**) Show that $\frac{l_\eta}{J_\eta} = X - \eta$ if ① is true.
(**f**) Show that $\frac{dp_\eta}{d\eta} = J_\eta (X - \eta) p_\eta$ if ① is true.
(**g**) Show that ② is true if ① is true.

2.41 Show equation (2.150) from (2.151).

2.42 Consider the probability distribution family $\{p_\theta | \theta \in \mathbb{R}\}$ in the probability space $\{1, \ldots, l\}$ and the stochastic transition matrix $Q = (Q^i_j)$. Let the Fisher information of p_{θ_0} in the probability distribution family $\{p_\theta | \theta \in \mathbb{R}\}$ be J_{θ_0}. Let J'_{θ_0} be the Fisher information of $Q(p_{\theta_0})$ in the probability distribution family $\{Q(p_\theta) | \theta \in \mathbb{R}\}$. Show then that $J_{\theta_0} \geq J'_{\theta_0}$. This inequality is called the **monotonicity** of the Fisher information. Similarly, define $\boldsymbol{J}_{\theta_0}$, $\boldsymbol{J}'_{\theta_0}$ for the multiple variable case, and show that the matrix inequality $\boldsymbol{J}_{\theta_0} \geq \boldsymbol{J}'_{\theta_0}$ holds.

2.4 Type Method and Large Deviation Evaluation

In this section, we analyze the case of a sufficiently large number of data by using the following two methods. The first method involves an analysis based on empirical distributions, and it is called the **type method**. In the second method, we consider a particular random variable and examine its exponential behavior.

2.4.1 Type Method and Sanov's Theorem

Let n data be generated according to a probability distribution in a finite set of events $\mathbb{N}_d = \{1, \ldots, d\}$. Then, we can perform the following analysis by examining the empirical distribution of the data [8]. Let T_n be the set of empirical distributions obtained from n observations. We call each element of this set a **type**. For each type $q \in T_n$, let the subset $T^n_q \subset \mathbb{N}^n_d$ be a set of data with the empirical distribution q. Since the probability $p^n(\boldsymbol{i})$ depends only on the type q for each $\boldsymbol{i} \in T^n_q$, we can denote this probability by $p^n(q)$. Then, when the n data are generated according to the probability distribution p^n, the empirical distribution matches $q \in T_n$ with the probability $p^n(T^n_q)$ $\left(\stackrel{\text{def}}{=} \sum_{\boldsymbol{i} \in T^n_q} p^n(\boldsymbol{i})\right)$.

Theorem 2.5 *Any type $p \in T_n$ and any data $\boldsymbol{i} \in T_q^n$ satisfy the following:*

$$p^n(T_q^n) \le p^n(T_p^n)\,, \tag{2.152}$$

$$p^n(\boldsymbol{i}) = e^{-n(H(q)+D(q\|p))}. \tag{2.153}$$

Denoting the number of elements of T_n and T_q^n by $|T_n|$ and $|T_q^n|$, respectively, we obtain the relations

$$|T_n| = \frac{n!}{n_1!\cdots n_d!} \le (n+1)^{d-1}, \tag{2.154}$$

$$\frac{1}{(n+1)^d} e^{nH(q)} \le |T_q^n| \le e^{nH(q)}, \tag{2.155}$$

$$\frac{1}{(n+1)^d} e^{-nD(q\|p)} \le p^n(T_q^n) \le e^{-nD(q\|p)}. \tag{2.156}$$

Proof Let $p(i) = \frac{n_i}{n}$ and $q(i) = \frac{n_i'}{n}$. Then,

$$p^n(T_p^n) = |T_p^n| \prod_{i=1}^d p(i)^{n_i} = \frac{n!}{n_1!\cdots n_d!} \prod_{i=1}^d p(i)^{n_i},$$

$$p^n(T_q^n) = |T_q^n| \prod_{i=1}^d p(i)^{n_i'} = \frac{n!}{n_1'!\cdots n_d'!} \prod_{i=1}^d p(i)^{n_i'}.$$

Using the inequality[Exe. 2.43]

$$\frac{n!}{m!} \le n^{n-m}, \tag{2.157}$$

we have

$$\begin{aligned}\frac{p^n(T_q^n)}{p^n(T_p^n)} &= \prod_{i=1}^d \left(\frac{n_i!}{n_i'!} p(i)^{n_i'-n_i}\right) \le \prod_{i=1}^d \left(n_i^{n_i-n_i'} \left(\frac{n_i}{n}\right)^{n_i'-n_i}\right) \\ &= \prod_{i=1}^d \left(\frac{1}{n}\right)^{n_i'-n_i} = \left(\frac{1}{n}\right)^{\sum_{i=1}^d (n_i'-n_i)} = 1\,.\end{aligned}$$

Therefore, inequality (2.152) holds. For $\boldsymbol{i} \in T_q^n$, we have

$$\begin{aligned}p^n(\boldsymbol{i}) &= \prod_{i=1}^d p(i)^{n_i'} = \prod_{i=1}^d p(i)^{n\left(\frac{n_i'}{n}\right)} \\ &= \prod_{i=1}^d e^{n \log p(i) \left(\frac{n_i'}{n}\right)} = e^{n \sum_{i=1}^d q(i) \log p(i)} = e^{-n(H(q)+D(q\|p))}\,,\end{aligned}$$

which implies (2.153).

Each element q of T_n may be written as a d-dimensional vector. Each component of the vector then assumes one of the following $n+1$ values: $0, 1/n, \ldots, n/n$. Since $\sum_{i=1}^{d} q_i = 1$, the dth element is decided by the other $d-1$ elements. Therefore, inequality (2.154) follows from a combinatorial observation. Applying inequality (2.153) to the case $p = q$, we have the relation $p^n(T_q^n) = e^{-nH(p)}|T_p^n|$. Since $1 = \sum_{q \in T_n} p^n(T_q^n) \geq p^n(T_q^n)$ for $p \in T_n$, we obtain the inequality on the RHS of (2.155). Conversely, inequality (2.152) yields that $1 = \sum_{q \in T_n} p^n(T_q^n) \leq \sum_{q \in T_n} p^n(T_p^n) = e^{-nH(p)}|T_p^n||T_n|$. Combining this relation with (2.154), we obtain the inequality on the LHS of (2.155). Inequality (2.156) may be obtained by combining (2.153) and (2.155). ■

We obtain **Sanov's Theorem** using these inequalities.

Theorem 2.6 (Sanov [9]) *The following holds for a subset $\mathcal{R}$ of distributions on $\mathbb{N}_d$:*

$$\frac{1}{(n+1)^d} \exp(-n \min_{q \in \mathcal{R} \cap T_n} D(q\|p)) \leq p^n(\cup_{q \in \mathcal{R} \cap T_n} T_q^n) \leq (n+1)^d \exp(-n \inf_{q \in \mathcal{R}} D(q\|p)).$$

In particular, when the closure of the interior of $\mathcal{R}$ coincides with the closure of $\mathcal{R}$,[10]

$$\lim_{n \to \infty} -\frac{1}{n} \log p^n \left(\cup_{q \in \mathcal{R} \cap T_n} T_q^n\right) = \inf_{q \in \mathcal{R}} D(q\|p)$$

in the limit $n \to \infty$.

Based on this theorem, we can analyze how different the true distribution is from the empirical distribution. More precisely, the empirical distribution belongs to the neighborhood of the true distribution with a sufficiently large probability, i.e., the probability of its complementary event approaches 0 exponentially. This exponent is then given by the relative entropy. The discussion of this exponent is called a large deviation evaluation.

However, it is difficult to consider a quantum extension of Sanov's theorem. This is because we cannot necessarily take the common eigenvectors for plural densities. That is, this problem must be treated independently of the choice of basis. One possible way to fulfill this requirement is the group representation method. If we use this method, it is possible to treat the eigenvalues of density of the system instead of the classical probabilities [10, 11]. Since eigenvalues do not identify the density matrix, they cannot be regarded as the complete quantum extension of Sanov's theorem. Indeed, a quantum extension is available if we focus only on two densities; however, it should be regarded as the quantum extension of Stein's lemma given in

[10] The set is called the interior of a set X when it consists of the elements of X without its boundary. For example, for a one-dimensional set, the interior of $[0, 0.5] \cup \{0.7\}$ is $(0, 0.5)$ and the closure of the interior is $[0, 0.5]$. Therefore, the condition is not satisfied in this case.

Sect. 3.5. Since the data are not given without our operation in the quantum case, it is impossible to directly extend Sanov's theorem to the quantum case.

In fact, the advantage of using the type method is the universality in information theory [8]. However, if we apply the type method to quantum systems independently of the basis, the universality is not available in the quantum case. A group representation method is very effective for a treatment independent of basis [10, 12–17]. Indeed, several universal protocols have been obtained by this method.

Exercise

2.43 Prove (2.157) by considering the cases $n \geq m$ and $n < m$ separately.

2.4.2 *Cramér Theorem and Its Application to Estimation*

Next, we consider the asymptotic behavior of a random variable in the case of independent and identical trials of the probability distribution p.

For this purpose, we first introduce two fundamental inequalities[Exe. 2.44]. The **Markov inequality** states that for a real-valued random variable X where $X \geq 0$,

$$\frac{\mathrm{E}_p(X)}{c} \geq p\{X \geq c\}\,. \tag{2.158}$$

Applying the Markov inequality to the variable $|X - \mathrm{E}_p(X)|$, we obtain the **Chebyshev inequality**:

$$p\{|X - \mathrm{E}_p(X)| \geq a\} \leq \frac{\mathrm{V}_p(X)}{a^2}\,. \tag{2.159}$$

Now, consider the real-valued random variable

$$X^n \stackrel{\mathrm{def}}{=} \sum_{i=1}^{n} \frac{1}{n} X_i\,, \tag{2.160}$$

where $X_1, \ldots, X_n$ are n independent random variables that are identical to the real-valued random variable X subject to the distribution p. When the variable X^n obeys the independent and identical distribution p^n of p, the expectation of X^n coincides with the expectation $\mathrm{E}_p(X)$. Let $\mathrm{V}_p(X)$ be the variance of X. Then, its variance with n observations equals $\mathrm{V}_p(X)/n$.

Applying Chebyshev's inequality (2.159), we have

$$p^n\{|X^n - \mathrm{E}_p(X)| \geq \epsilon\} \leq \frac{\mathrm{V}_p(X)}{n\epsilon^2}$$

for arbitrary $\epsilon > 0$. This inequality yields the **(weak) law of large numbers**

$$p^n\{|X^n - \mathrm{E}_p(X)| \geq \epsilon\} \to 0, \quad \forall \epsilon > 0\,. \tag{2.161}$$

In general, if a sequence of pairs $\{(X^n, p_n)\}$ of a real-valued random variable and a probability distribution satisfies

$$p_n\{|X^n - x| \geq \epsilon\} \to 0, \quad \forall \epsilon > 0 \tag{2.162}$$

for a real number x, then the real-valued random variable X^n is said to **converge in probability** to x.

Since the left-hand side (LHS) of (2.161) converges to 0, the next focus is the speed of this convergence. Usually, this convergence is exponential. The exponent of this convergence is characterized by **Cramér's Theorem** below.

Theorem 2.7 (Cramér [18]) *Define the* **cumulant generating function** $\mu(\theta) \stackrel{\text{def}}{=} \log\left(\sum_\omega p(\omega) e^{\theta X(\omega)}\right)$. *Then*

$$\underline{\lim} -\frac{1}{n} \log p^n\{X^n \geq x\} \geq \max_{\theta \geq 0} (\theta x - \mu(\theta))\,, \tag{2.163}$$

$$\overline{\lim} -\frac{1}{n} \log p^n\{X^n \geq x\} \leq \lim_{x' \to x+0} \max_{\theta \geq 0} \left(\theta x' - \mu(\theta)\right)\,, \tag{2.164}$$

$$\underline{\lim} -\frac{1}{n} \log p^n\{X^n \leq x\} \geq \max_{\theta \leq 0} (\theta x - \mu(\theta)) \tag{2.165}$$

$$\overline{\lim} -\frac{1}{n} \log p^n\{X^n \leq x\} \leq \lim_{x' \to x-0} \max_{\theta \leq 0} \left(\theta x' - \mu(\theta)\right)\,. \tag{2.166}$$

If we replace $\{X^n \geq x\}$ *and* $\{X^n \leq x\}$ *with* $\{X^n > x\}$ *and* $\{X^n < x\}$, *respectively, the same inequalities hold.*

When the probability space consists of finite elements, the function $\max_{\theta \geq 0} (\theta x - \mu(\theta))$ is continuous, i.e., $\lim_{x' \to x+0} \max_{\theta \geq 0} \left(\theta x' - \mu(\theta)\right) = \max_{\theta \geq 0} (\theta x - \mu(\theta))$. Hence, the equality of (2.163) holds. Conversely, if the probability space contains an infinite number of elements as the set of real numbers $\mathbb{R}$, we should treat the difference between the RHS and LHS more carefully. Further, the inequality of (2.163) holds without limit, and is equivalent to (2.46) when we replace the real-valued random variable $X(\omega)$ with $-\log q(\omega)$. The same argument holds for (2.165).

Proof Inequality (2.165) is obtained by considering $-X$ in (2.163). Therefore, we prove only (2.163). Inequality (2.166) is also obtained by considering $-X$ in (2.164). Here we prove only inequality (2.163). Inequality (2.164) will be proved at the end of this section.

For a real-valued random variable X with $X(\omega)$ for each ω,

$$\mathrm{E}_{p^n}(e^{n\theta X^n}) = \mathrm{E}_{p^n}\left(\prod_{i=1}^n e^{\theta X_i}\right) = (\mathrm{E}_p e^{\theta X})^n = e^{n\mu(\theta)}\,. \tag{2.167}$$

Using the Markov inequality (2.158), we obtain

$$p^n\{X^n \geq x\} = p^n\{e^{n\theta X^n} \geq e^{n\theta x}\} \leq \frac{e^{n\mu(\theta)}}{e^{n\theta x}} = e^{n(\mu(\theta)-\theta x)} \text{ for } \theta \geq 0\,. \tag{2.168}$$

Taking the logarithm of both sides, we have

$$-\frac{1}{n}\log p^n\{X^n \geq x\} \geq \theta x - \mu(\theta).$$

Let us take the maximum on the RHS with respect to $\theta \geq 0$ and then take the limit on the LHS. We obtain inequality (2.163). ■

This theorem can be extended to the non-i.i.d. case as the **Gärtner–Ellis theorem**.

Theorem 2.8 (Gärtner [19], Ellis [20]) *Let $\{p_n\}$ be a general sequence of the probabilities with the real-valued random variables X_n. Define the* **cumulant generating functions** $\mu_n(\theta) \stackrel{\text{def}}{=} \frac{1}{n}\log\left(\sum_\omega p_n(\omega)e^{\theta n X_n(\omega)}\right)$ *and* $\mu(\theta) \stackrel{\text{def}}{=} \lim_{n\to\infty}\mu_n(\theta)$ *and the set* $G \stackrel{\text{def}}{=} \{\mu'(\theta)|\theta\}$. *Then*

$$\underline{\lim} -\frac{1}{n}\log p_n\{X_n \geq x\} \geq \max_{\theta\geq 0}\left(\theta x - \mu(\theta)\right), \tag{2.169}$$

$$\overline{\lim} -\frac{1}{n}\log p_n\{X_n \geq x\} \leq \inf_{\bar{x}\in G:\bar{x}>x}\max_{\theta\geq 0}\left(\theta \bar{x} - \mu(\theta)\right), \tag{2.170}$$

$$\underline{\lim} -\frac{1}{n}\log p_n\{X_n \leq x\} \geq \max_{\theta\leq 0}\left(\theta x - \mu(\theta)\right), \tag{2.171}$$

$$\overline{\lim} -\frac{1}{n}\log p_n\{X_n \leq x\} \leq \inf_{\bar{x}\in G:\bar{x}<x}\max_{\theta\leq 0}\left(\theta \bar{x} - \mu(\theta)\right). \tag{2.172}$$

If we replace $\{X_n \geq x\}$ and $\{X_n \leq x\}$ by $\{X_n > x\}$ and $\{X_n < x\}$, respectively, the same inequalities hold.

Inequalities (2.169) and (2.171) can be proved in a similar way to Theorem 2.7.

Next, we apply large deviation arguments to estimation theory. Our arguments will focus not on the mean square error but on the decreasing rate of the probability that the estimated parameter does not belong to the ϵ-neighborhood of the true parameter. To treat the accuracy of a sequence of estimators $\{\hat{\theta}_n\}$ with a one-parameter probability distribution family $\{p_\theta|\theta \in \mathbb{R}\}$ from the viewpoint of a large deviation, we define

$$\beta(\{\hat{\theta}_n\}, \theta, \epsilon) \stackrel{\text{def}}{=} \underline{\lim} -\frac{1}{n}\log p_\theta^n\{|\hat{\theta}_n - \theta| > \epsilon\}, \tag{2.173}$$

$$\alpha(\{\hat{\theta}_n\}, \theta) \stackrel{\text{def}}{=} \underline{\lim}_{\epsilon\to 0}\frac{\beta(\{\hat{\theta}_n\}, \theta, \epsilon)}{\epsilon^2}. \tag{2.174}$$

As an approximation, we have

$$p_\theta^n\{|\hat{\theta}_n - \theta| > \epsilon\} \cong e^{-n\epsilon^2\alpha(\{\hat{\theta}_n\},\theta)}.$$

Hence, an estimator functions better when it has larger values of $\beta(\{\hat{\theta}_n\}, \theta, \epsilon)$ and $\alpha(\{\hat{\theta}_n\}, \theta)$.

Theorem 2.9 (Bahadur [21–23]) *Let a sequence of estimators* $\{\hat{\theta}_n\}$ *satisfy the weak consistency condition*

$$p_\theta^n\{|\hat{\theta}_n - \theta| > \epsilon\} \to 0, \quad \forall \epsilon > 0, \quad \forall \theta \in \mathbb{R}\,. \tag{2.175}$$

Then, it follows that

$$\beta(\{\hat{\theta}_n\}, \theta, \epsilon) \le \inf_{\theta':|\theta'-\theta|>\epsilon} D(p_{\theta'} \| p_\theta)\,. \tag{2.176}$$

Further, if

$$D(p_{\theta'} \| p_\theta) = \lim_{\bar{\theta}\to\theta'} D(p_{\bar{\theta}} \| p_\theta), \tag{2.177}$$

the following also holds:

$$\alpha(\{\hat{\theta}_n\}, \theta) \le \frac{1}{2} J_\theta. \tag{2.178}$$

If the probability space consists of finite elements, condition (2.177) holds.

Proof of Theorem 2.9 Inequality (2.178) is obtained by combining (2.176) with (2.105). Inequality (2.176) may be derived from monotonicity (2.13) as follows. From the consistency condition (2.175), the sequence $a_n \stackrel{\text{def}}{=} p_\theta^n\{|\hat{\theta}_n - \theta| > \epsilon\}$ satisfies $a_n \to 0$. Assume that $\epsilon' \stackrel{\text{def}}{=} |\theta - \theta'| > \epsilon$. Then, when $|\hat{\theta}_n - \theta'| < \epsilon' - \epsilon$, we have $|\hat{\theta}_n - \theta| > \epsilon$. Hence, the other sequence $b_n \stackrel{\text{def}}{=} p_{\theta'}^n\{|\hat{\theta}_n - \theta| > \epsilon\} \ge p_{\theta'}^n\{|\hat{\theta}_n - \theta'| < \epsilon' - \epsilon\}$ satisfies $b_n \to 1$ because of the consistency condition (2.175). Thus, monotonicity (2.13) implies that

$$D(p_{\theta'}^n \| p_\theta^n) \ge b_n(\log b_n - \log a_n) + (1 - b_n)(\log(1 - b_n) - \log(1 - a_n)).$$

Since $nD(p_{\theta'} \| p_\theta) = D(p_{\theta'}^n \| p_\theta^n)$ follows from (2.28) and $-(1 - b_n)\log(1 - a_n) \ge 0$, we have $nD(p_{\theta'} \| p_\theta) \ge -h(b_n) - b_n \log a_n$, and therefore

$$-\frac{1}{n}\log a_n \le \frac{D(p_{\theta'} \| p_\theta)}{b_n} + \frac{h(b_n)}{nb_n}. \tag{2.179}$$

As the convergence $h(b_n) \to 0$ follows from the convergence $b_n \to 1$, we have

$$\beta(\{\hat{\theta}_n\}, \theta, \epsilon) \le D(p_{\theta'} \| p_\theta).$$

Considering $\inf_{\theta':|\theta'-\theta|>\epsilon}$, we obtain (2.176). In addition, this proof is valid even if we replace $\{|\hat{\theta}_n - \theta| > \epsilon\}$ in (2.173) by $\{|\hat{\theta}_n - \theta| \ge \epsilon\}$. ■

If no estimator satisfies the equalities in inequalities (2.176) and (2.178), these inequalities are not sufficiently useful. The following proposition gives a sufficient condition for the equalities of (2.176) and (2.178).

Proposition 2.1 *Suppose that the probability distribution family (2.127) is exponential, and the parameter to be estimated is an expectation parameter. If a sequence of estimators is given by $X^n(\omega^n)$ (see (2.160)), then the equality of (2.176) holds. The equality of (2.178) also holds.*

It is known that the maximum likelihood estimator $\hat{\theta}_{n,ML}$ satisfies (2.178) if the probability distribution family satisfies some regularity conditions [23, 24].

Proof of Proposition 2.1 *and* (2.164) *and* (2.166) *in Theorem* 2.7 Now, we prove Proposition 2.1 and its related formulas ((2.163) and (2.164) in Theorem 2.7) as follows. Because (2.129) implies $\max_{\theta' \ge \theta}(\theta' - \theta)(\eta(\theta) + \epsilon) - (\mu(\theta') - \mu(\theta)) = D(\hat{p}_{\eta(\theta)+\epsilon} \| \hat{p}_{\eta(\theta)})$, Proposition 2.1 follows from the inequalities

$$\begin{aligned} &\underline{\lim} -\frac{1}{n} \log \hat{p}^n_{\eta(\theta)}\{X^n(\omega^n) > \eta(\theta) + \epsilon\} \\ \ge &\max_{\theta' \ge \theta}(\theta' - \theta)(\eta(\theta) + \epsilon) - (\mu(\theta') - \mu(\theta)), \end{aligned} \tag{2.180}$$

$$\overline{\lim} -\frac{1}{n} \log \hat{p}^n_{\eta(\theta)}\{X^n(\omega^n) > \eta(\theta) + \epsilon\} \le \underline{\lim}_{\epsilon' \to \epsilon + 0} D(\hat{p}_{\eta(\theta)+\epsilon'} \| \hat{p}_{\eta(\theta)}) \tag{2.181}$$

for the expectation parameter η of the exponential family (2.127) and arbitrary $\epsilon > 0$. When $x = \eta(\theta) + \epsilon = \eta(\tilde{\theta}) \ge 0$ and $\theta = 0$, the formula (2.181) is the same as (2.164) in Theorem 2.7 with replacing $\ge$ by $>$ in the LHS because $D(\hat{p}_{\eta(\theta)+\epsilon} \| \hat{p}_{\eta(\theta)}) = \tilde{\theta}\eta(\tilde{\theta}) - \mu(\tilde{\theta}) = \max_\theta \theta\eta(\tilde{\theta}) - \mu(\theta)$. Since the LHS of (2.181) is not smaller than the LHS of (2.164) in this correspondence, (2.181) yields (2.164). Considering $-X$ instead of X, (2.164) implies (2.166).

To show (2.180), we choose arbitrary $\bar{\epsilon} > \epsilon$ and $\bar{\theta}$ such that $\mu'(\bar{\theta}) = \eta(\theta) + \bar{\epsilon}$. Based on the proof of (2.163) in Theorem 2.7, since the expectation of $e^{n(\theta'-\theta)X^n(\omega^n)}$ under the distribution p^n_θ is $e^{\mu(\theta')-\mu(\theta)}$, we can show that

$$\begin{aligned} &-\frac{1}{n} \log p^n_\theta\{X^n(\omega^n) > \eta(\theta) + \epsilon\} \\ \ge &\max_{\theta':\theta' \ge \theta} (\theta' - \theta)(\eta(\theta) + \epsilon) - (\mu(\theta') - \mu(\theta)), \end{aligned} \tag{2.182}$$

$$\begin{aligned} &-\frac{1}{n} \log p^n_{\bar{\theta}}\{X^n(\omega^n) \le \eta(\theta) + \epsilon\} \\ \ge &\max_{\theta':\theta' \le \bar{\theta}} (\theta' - \bar{\theta})(\eta(\theta) + \epsilon) - (\mu(\theta') - \mu(\bar{\theta})) = D(\hat{p}_{\eta(\theta)+\epsilon} \| \hat{p}_{\eta(\theta)+\bar{\epsilon}}) > 0. \end{aligned} \tag{2.183}$$

Then, (2.182) implies (2.180).

Next, using (2.183), we show (2.181) as follows. According to a discussion similar to the proof of (2.176) in Theorem 2.9, we have

$$-\frac{1}{n}\log \hat{p}^n_{\eta(\theta)}\{X^n(\omega^n) > \eta(\theta)+\epsilon\} \le \frac{D(\hat{p}_{\eta(\theta)+\epsilon'}\|\hat{p}_{\eta(\theta)})}{b_n} + \frac{h(b_n)}{nb_n} \tag{2.184}$$

for $\epsilon' > \epsilon$, where $b_n \stackrel{\rm def}{=} \hat{p}^n_{\eta(\theta)+\epsilon'}\{X^n(\omega^n) > \eta(\theta)+\epsilon\}$. From (2.183), $b_n \to 1$. Hence, we obtain the last inequality in (2.181). ■

Proof of (2.170) *and* (2.172) *in Theorem* 2.8 Finally, we will prove inequality (2.170) in Theorem 2.8, i.e., we will prove that

$$\overline{\lim} -\frac{1}{n}\log p_n\{X_n(\omega) \ge x\} \le \max_{\theta\ge 0}\left(\theta\mu'(\bar{\theta}) - \mu(\theta)\right) \tag{2.185}$$

for any $\bar{\theta}$ satisfying $\mu'(\bar{\theta}) > x$. Inequality (2.172) can be shown in the same way. Define the exponential family $p_{n,\theta}(\omega) \stackrel{\rm def}{=} p_n(\omega)e^{n\theta X_n(\omega)-n\mu_n(\theta)}$. Similarly to (2.184), we have

$$-\frac{1}{n}\log p_{n,0}\{X_n(\omega) > x\} \le \frac{D(p_{n,\bar{\theta}}\|p_{n,0})}{nb_n} + \frac{h(b_n)}{nb_n},$$

where $b_n \stackrel{\rm def}{=} p_{n,\bar{\theta}}\{X_n(\omega) > x\}$. From (2.129), $\frac{D(p_{n,\bar{\theta}}\|p_{n,0})}{n} = \max_{\theta\ge 0}\left(\theta\mu_n'(\bar{\theta}) - \mu_n(\theta)\right)$. Hence, if we show that $b_n \to 1$, we obtain (2.185). To show that $b_n \to 1$, similarly to (2.183), the inequality

$$-\frac{1}{n}\log p_{n,\bar{\theta}}\{X_n(\omega) \le x\} \ge \max_{\theta\le\bar{\theta}}(\theta-\bar{\theta})x - \mu(\theta) + \mu(\bar{\theta})$$

holds. Since the set of differentiable points of μ is open and μ' is monotone increasing and continuous in this set, there exists a point θ' in this set such that

$$\theta' < \bar{\theta}, \quad x < \mu'(\theta').$$

Since μ' is monotone increasing, we obtain

$$\begin{aligned}&\max_{\theta\le\bar{\theta}}(\theta-\bar{\theta})x - \mu(\theta) + \mu(\bar{\theta}) \ge (\theta'-\bar{\theta})x - \mu(\theta') + \mu(\bar{\theta})\\ \ge&(\mu'(\theta') - x)(\bar{\theta}-\theta') > 0\,,\end{aligned}$$

which implies that $b_n \to 1$. ■

Exercises

2.44 Prove Markov's inequality by using the inequality $\sum_{i:x_i\ge c} p_i x_i \ge c\sum_{i:x_i\ge c} p_i$.

2.45 Using Cramér's theorem and (2.42) and (2.44), show the following equations below. Show analogous formulas for (2.46), (2.47), (3.5), and (3.6).

$$\lim_{n\to\infty} -\frac{1}{n}\log p^n\{p_{i^n}^n \le e^{-nR}\} = -\min_{0\le s}(\psi(s) - sR), \tag{2.186}$$

$$\lim_{n\to\infty} -\frac{1}{n}\log p^n\{p_{i^n}^n > e^{-nR}\} = -\min_{s\le 0}(\psi(s) - sR). \tag{2.187}$$

2.46 Show that

$$\lim_{n\to\infty} -\frac{1}{n}\log P^c(p^n, e^{nR}) = -\min_{0\le s\le 1}\frac{\psi(s) - sR}{1-s} \tag{2.188}$$

by first proving (2.189) and then combining this with (2.55). The $\ge$ part may be obtained directly from (2.51)

$$\begin{aligned} P^c(p^n, e^{nR}) &\ge \max_{q\in T_n:|T_q^n|>e^{nR}} (|T_q^n| - e^{nR})e^{-n(H(p)+H(p\|q))} \\ &\ge \max_{q\in T_n:\frac{e^{nH(q)}}{(n+1)^d}>e^{nR}} \left(\frac{e^{nH(q)}}{(n+1)^d} - e^{nR}\right)e^{-n(H(p)+H(p\|q))} \\ &= \max_{q\in T_n:\frac{e^{nH(q)}}{(n+1)^d}>e^{nR}} e^{-nD(p\|q)}\left(1 - \frac{(n+1)^d e^{nR}}{e^{nH(q)}}\right). \end{aligned} \tag{2.189}$$

2.47 Show that

$$\lim_{n\to\infty} -\frac{1}{n}\log P(p^n, e^{nR}) = -\min_{s\le 0}\frac{\psi(s) - sR}{1-s} \tag{2.190}$$

by first proving (2.191) and then combining this with (2.55). The inequality $\ge$ may be obtained directly from (2.54)

$$P(p^n, e^{nR}) \ge \max_{q\in T_n:|T_q^n|\le e^{nR}} p^n(T_q^n) \ge \max_{q\in T_n:H(q)\le R} \frac{e^{-nD(q\|p)}}{(n+1)^d}. \tag{2.191}$$

2.48 Consider the case where $\Omega_n = \{0, 1\}$, $p_n(0) = e^{-na}$, $p_n(1) = 1 - e^{-na}$, $X_n(0) = a$, $X_n(1) = -b$ with $a, b > 0$. Show that $\mu(\theta) = -\min\{(1-\theta)a, \theta b\}$ and the following for $-b < x < a$:

$$\max_{\theta>0}(x\theta - \mu(\theta)) = \frac{a(x+b)}{a+b} < a, \quad \lim_{n\to\infty}\frac{1}{n}\log p_n\{X_n \ge x\} = a.$$

It gives a counterexample of Gärtner–Ellis Theorem in the nondifferentiable case.

2.5 Continuity and Axiomatic Approach

In this section, we consider how to characterize the entropy $H(p)$ by axioms. Indeed, when a real-value function S satisfies several axiomatic rules, the function S must be the entropy $H(p)$ given in (2.2). Here, we consider the following five axioms for a real-value function S for distribution, which is close to the axioms by Khinchin [25].

K1 (Normalization)

$$S(p_{\mathrm{mix},\{0,1\}}) = \log 2. \tag{2.192}$$

K2 (Continuity) S is continuous on $\mathcal{P}(\{0, 1\})$.
K3 (Nonnegativity) S is nonnegative.
K4 (Expandability) For any function f, we have

$$S(\mathrm{P}_X) = S(\mathrm{P}_{f(X)X}). \tag{2.193}$$

K5 (Chain rule) When P_{XY} is a joint distribution for X and Y, the marginal distribution P_X and the conditional distribution $\mathrm{P}_{Y|X=e}$ satisfies that

$$S(\mathrm{P}_{XY}) = S(\mathrm{P}_X) + \sum_x \mathrm{P}_X(x) S(\mathrm{P}_{Y|X=x}). \tag{2.194}$$

Here, we consider another set of axioms as follows.

A1 (Normalization)

$$S(p_{\mathrm{mix},\{0,1\}}) = \log 2. \tag{2.195}$$

A2 (Weak additivity)

$$S(p^n) = nS(p) \tag{2.196}$$

A3 (Monotonicity) For any function f, we have

$$S(\mathrm{P}_X) \geq S(\mathrm{P}_{f(X)}). \tag{2.197}$$

A4 (Asymptotic continuity) Let p_n and q_n be distributions on the set $\{0, 1\}^n$. When $d_1(p_n, q_n) \to 0$, we have

$$\frac{|S(p_n) - S(q_n)|}{n} \to 0. \tag{2.198}$$

Then, the following theorem shows the uniqueness of a function satisfying one of the above sets of axioms.

Theorem 2.10 *For a function S defined on the set of distributions, the following three conditions are equivalent.*

(1) *S satisfies Axioms* **K1-K5**.
(2) *S satisfies Axioms* **A1-A4**.
(3) $S(p) = -\sum_i p_i \log p_i$.

Before proceeding to the proof of Theorem 2.10, we consider the asymptotic convertibility for the independent and identical distribution.

Lemma 2.5 *For a distribution p on Ω and an arbitrary real number $\epsilon > 0$, there exists a sequence of maps f_n from Ω^n to $\Omega_n := \{0,1\}^{\lfloor (H(p)-\epsilon)n/\log 2 \rfloor}$ such that $d_1(p^n \circ f_n^{-1}, p_{\mathrm{mix},\Omega_n}) \to 0$.*

Lemma 2.6 *For a distribution p on Ω and an arbitrary real number $\epsilon > 0$, there exists a sequence of maps f_n from $\Omega_n' := \{0,1\}^{\lfloor (H(p)+\epsilon)n/\log 2 \rfloor}$ to Ω^n such that $d_1(p^n, p_{\mathrm{mix},\Omega_n'} \circ f_n^{-1}) \to 0$.*

These two lemmas show that the entropy $H(p)$ gives the asymptotic conversion rate between the independent and identical distribution and the uniform distribution. Rényi entropy $H_{1+s}(p)$ also satisfies Axioms **K1-K4** and **A1-A3**. However, it does not satisfies K5 (Chain rule) or A4 (Asymptotic continuity)[Exe. 2.49, 2.50]. Indeed, although the quantity $e^{-H_2(p)}$ satisfies A4 (Asymptotic continuity)[Exe. 2.51] as well as A3 (Monotonicity), it does not satisfy A2 (Weak additivity). Only the information quantity satisfying Axioms **K1-K5** or **A1-A4** gives the asymptotic conversion between the independent and identical distribution and the uniform distribution. Hence, we can conclude that K5 (Chain rule) and A4 (Asymptotic continuity) are crucial for the asymptotic conversion.

Proof of Theorem 2.10 First, we show (1) $\Rightarrow$ (2). A2 (Weak additivity) follows from K5 (Chain rule). A3 (Monotonicity) follows from K3 (Nonnegativity), K4 (Expandability), and K5 (Chain rule) by the same discussion as (2.6).

Now, we start to show A4 (Asymptotic continuity). Since the set $\mathcal{P}(\{0,1\})$ is compact, due to K2 (Continuity), S is uniformly continuous on $\mathcal{P}(\{0,1\})$. So, there exists the maximum value $R := \max_{p \in \mathcal{P}(\{0,1\})} S(p)$. For any $\epsilon > 0$, we choose $\delta > 0$ such that $|S(p) - S(q)| \le \epsilon$ for any $d_1(p,q) \le \delta$. Consider two distributions $\mathrm{P}_{\boldsymbol{X}_n}^n$ and $\overline{\mathrm{P}}_{\boldsymbol{X}_n}^n$ on the set $\{0,1\}^n$ such that $\delta_n := 2d_1(\mathrm{P}_{\boldsymbol{X}_n}^n, \overline{\mathrm{P}}_{\boldsymbol{X}_n}^n)$ goes to zero as $n \to \infty$. Then, we can choose a sufficiently large integer N such that $\delta_n \le \frac{\epsilon\delta}{2R}$ for $n \ge N$.

Here, X_i denotes the random variable on the i-th set $\{0,1\}$ in $\{0,1\}^n$ and $\boldsymbol{X}_n := (X_1, \ldots, X_n)$. For any integer $i \le n$, we have

$$\sum_{\boldsymbol{x}_{i-1}} \left| \mathrm{P}_{\boldsymbol{X}_{i-1}}^n(\boldsymbol{x}_{i-1}) - \overline{\mathrm{P}}_{\boldsymbol{X}_{i-1}}^n(\boldsymbol{x}_{i-1}) \right| \le \delta_n.$$

Also, for any value $x_i' \in \{0,1\}$, we have

$$\begin{aligned}
&\sum_{\boldsymbol{x}_{i-1}} \mathrm{P}^n_{\boldsymbol{X}_{i-1}}(\boldsymbol{x}_{i-1}) \left| \mathrm{P}^n_{X_i|\boldsymbol{X}_{i-1}=\boldsymbol{x}_{i-1}}(x_i') - \overline{\mathrm{P}}^n_{X_i|\boldsymbol{X}_{i-1}=\boldsymbol{x}_{i-1}}(x_i') \right| \\
\le & \sum_{\boldsymbol{x}_{i-1}} \left| \mathrm{P}^n_{\boldsymbol{X}_{i-1}}(\boldsymbol{x}_{i-1}) \mathrm{P}^n_{X_i|\boldsymbol{X}_{i-1}=\boldsymbol{x}_{i-1}}(x_i') - \overline{\mathrm{P}}^n_{\boldsymbol{X}_{i-1}}(\boldsymbol{x}_{i-1}) \overline{\mathrm{P}}^n_{X_i|\boldsymbol{X}_{i-1}=\boldsymbol{x}_{i-1}}(x_i') \right| \\
&+ \sum_{\boldsymbol{x}_{i-1}} \left| \mathrm{P}^n_{\boldsymbol{X}_{i-1}}(\boldsymbol{x}_{i-1}) - \overline{\mathrm{P}}^n_{\boldsymbol{X}_{i-1}}(\boldsymbol{x}_{i-1}) \right| \overline{\mathrm{P}}^n_{X_i|\boldsymbol{X}_{i-1}=\boldsymbol{x}_{i-1}}(x_i') \\
\le & \sum_{\boldsymbol{x}_i} \left| \mathrm{P}^n_{\boldsymbol{X}_i}(\boldsymbol{x}_i) - \overline{\mathrm{P}}^n_{\boldsymbol{X}_i}(\boldsymbol{x}_i) \right| + \sum_{\boldsymbol{x}_{i-1}} \left| \mathrm{P}^n_{\boldsymbol{X}_{i-1}}(\boldsymbol{x}_{i-1}) - \overline{\mathrm{P}}^n_{\boldsymbol{X}_{i-1}}(\boldsymbol{x}_{i-1}) \right| \\
\le & \delta_n + \delta_n = 2\delta_n. \qquad (2.199)
\end{aligned}$$

We define the function $Y_{x_i'}(\boldsymbol{x}_{i-1}) := |\mathrm{P}^n_{X_i|\boldsymbol{X}_{i-1}=\boldsymbol{x}_{i-1}}(x_i') - \overline{\mathrm{P}}^n_{X_i|\boldsymbol{X}_{i-1}=\boldsymbol{x}_{i-1}}(x_i')|$. Applying Markov inequality to the random variable $Y_{x_i'}(\boldsymbol{X}_{i-1})$, from (2.199), we have the inequality

$$\mathrm{P}^n_{\boldsymbol{X}_{i-1}}(\{\boldsymbol{x}_{i-1}|\ |\mathrm{P}^n_{X_i|\boldsymbol{X}_{i-1}=\boldsymbol{x}_{i-1}}(x_i') - \overline{\mathrm{P}}^n_{X_i|\boldsymbol{X}_{i-1}=\boldsymbol{x}_{i-1}}(x_i')| \le \delta\}) \ge 1 - \frac{2\delta_n}{\delta}. \qquad (2.200)$$

Let Ω_i be the set of $\boldsymbol{x}_{i-1} = (x_1, \ldots x_{i-1})$ satisfying the condition inside of the parenthesis in the LHS of (2.200). Then, K3 (Nonnegativity) implies that

$$\begin{aligned}
&\sum_{\boldsymbol{x}_{i-1}} \mathrm{P}^n_{\boldsymbol{X}_{i-1}}(\boldsymbol{x}_{i-1}) \left| S(\mathrm{P}^n_{X_i|\boldsymbol{X}_{i-1}=\boldsymbol{x}_{i-1}}) - S(\overline{\mathrm{P}}^n_{X_i|\boldsymbol{X}_{i-1}=\boldsymbol{x}_{i-1}}) \right| \\
= & \sum_{\boldsymbol{x}_{i-1} \in \Omega_i} \mathrm{P}^n_{\boldsymbol{X}_{i-1}}(\boldsymbol{x}_{i-1}) \left| S(\mathrm{P}^n_{X_i|\boldsymbol{X}_{i-1}=\boldsymbol{x}_{i-1}}) - S(\overline{\mathrm{P}}^n_{X_i|\boldsymbol{X}_{i-1}=\boldsymbol{x}_{i-1}}) \right| \\
&+ \sum_{\boldsymbol{x}_{i-1} \in \Omega_i^c} \mathrm{P}^n_{\boldsymbol{X}_{i-1}}(\boldsymbol{x}_{i-1}) \left| S(\mathrm{P}^n_{X_i|\boldsymbol{X}_{i-1}=\boldsymbol{x}_{i-1}}) - S(\overline{\mathrm{P}}^n_{X_i|\boldsymbol{X}_{i-1}=\boldsymbol{x}_{i-1}}) \right| \\
= & \sum_{\boldsymbol{x}_{i-1} \in \Omega_i} \mathrm{P}^n_{\boldsymbol{X}_{i-1}}(\boldsymbol{x}_{i-1}) \epsilon + \sum_{\boldsymbol{x}_{i-1} \in \Omega_i^c} \mathrm{P}^n_{\boldsymbol{X}_{i-1}}(\boldsymbol{x}_{i-1}) R \\
\le & \epsilon + 2\frac{\delta_n}{\delta} R \le \epsilon + \epsilon = 2\epsilon. \qquad (2.201)
\end{aligned}$$

Also, K3 (Nonnegativity) implies that

$$\begin{aligned}
&\sum_{\boldsymbol{x}_{i-1}} \left| \mathrm{P}^n_{\boldsymbol{X}_{i-1}}(\boldsymbol{x}_{i-1}) - \overline{\mathrm{P}}^n_{\boldsymbol{X}_{i-1}}(\boldsymbol{x}_{i-1}) \right| S(\overline{\mathrm{P}}_{X_i|X_1=x_1,\ldots,X_{i-1}=x_{i-1}}) \\
\le & \sum_{\boldsymbol{x}_{i-1}} \left| \mathrm{P}^n_{\boldsymbol{X}_{i-1}}(\boldsymbol{x}_{i-1}) - \overline{\mathrm{P}}^n_{\boldsymbol{X}_{i-1}}(\boldsymbol{x}_{i-1}) \right| R \le \delta_n R \le \frac{\epsilon\delta}{2}. \qquad (2.202)
\end{aligned}$$

On the other hand, K5 (Chain rule) implies that

$$S(\mathrm{P}^n_{\boldsymbol{X}_n}) = \sum_{i=1}^{n} \sum_{\boldsymbol{x}_{i-1}} \mathrm{P}^n_{\boldsymbol{X}_{i-1}}(\boldsymbol{x}_{i-1}) S(\mathrm{P}^n_{X_i|\boldsymbol{X}_{i-1}=\boldsymbol{x}_{i-1}}). \tag{2.203}$$

Thus, we have

$$\begin{aligned}
&\left|S(\mathrm{P}_{X_1,\ldots,X_n}) - S(\overline{\mathrm{P}}_{X_1,\ldots,X_n})\right| \\
\overset{(a)}{\le} &\sum_{i=1}^{n} \sum_{\boldsymbol{x}_{i-1}} \left|\mathrm{P}^n_{\boldsymbol{X}_{i-1}}(\boldsymbol{x}_{i-1}) S(\mathrm{P}^n_{X_i|\boldsymbol{X}_{i-1}=\boldsymbol{x}_{i-1}}) - \overline{\mathrm{P}}^n_{\boldsymbol{X}_{i-1}}(\boldsymbol{x}_{i-1}) S(\overline{\mathrm{P}}^n_{X_i|\boldsymbol{X}_{i-1}=\boldsymbol{x}_{i-1}})\right| \\
\le &\sum_{i=1}^{n} \sum_{\boldsymbol{x}_{i-1}} \left|\mathrm{P}^n_{\boldsymbol{X}_{i-1}}(\boldsymbol{x}_{i-1}) S(\mathrm{P}^n_{X_i|\boldsymbol{X}_{i-1}=\boldsymbol{x}_{i-1}}) - \mathrm{P}^n_{\boldsymbol{X}_{i-1}}(\boldsymbol{x}_{i-1}) S(\overline{\mathrm{P}}^n_{X_i|\boldsymbol{X}_{i-1}=\boldsymbol{x}_{i-1}})\right| \\
&+ \left|\mathrm{P}^n_{\boldsymbol{X}_{i-1}}(\boldsymbol{x}_{i-1}) S(\overline{\mathrm{P}}^n_{X_i|\boldsymbol{X}_{i-1}=\boldsymbol{x}_{i-1}}) - \overline{\mathrm{P}}^n_{\boldsymbol{X}_{i-1}}(\boldsymbol{x}_{i-1}) S(\overline{\mathrm{P}}^n_{X_i|\boldsymbol{X}_{i-1}=\boldsymbol{x}_{i-1}})\right| \\
= &\sum_{i=1}^{n} \sum_{\boldsymbol{x}_{i-1}} \mathrm{P}^n_{\boldsymbol{X}_{i-1}}(\boldsymbol{x}_{i-1}) \left|S(\mathrm{P}^n_{X_i|\boldsymbol{X}_{i-1}=\boldsymbol{x}_{i-1}}) - S(\overline{\mathrm{P}}^n_{X_i|\boldsymbol{X}_{i-1}=\boldsymbol{x}_{i-1}})\right| \\
&+ \left|\mathrm{P}^n_{\boldsymbol{X}_{i-1}}(\boldsymbol{x}_{i-1}) - \overline{\mathrm{P}}^n_{\boldsymbol{X}_{i-1}}(\boldsymbol{x}_{i-1})\right| S(\overline{\mathrm{P}}^n_{X_i|\boldsymbol{X}_{i-1}=\boldsymbol{x}_{i-1}}) \\
\overset{(b)}{\le} &\sum_{i=1}^{n} 2\epsilon + \frac{\epsilon\delta}{2} = n(2\epsilon + \frac{\epsilon\delta}{2}),
\end{aligned}$$

where (a) follows from (2.203), and (b) follows from (2.201) and (2.202). Hence, A4 (Asymptotic continuity) holds.

Next, we show (2) $\Rightarrow$ (3). For a distribution p and $\epsilon > 0$, according to Lemma 2.5, we choose a sequence of maps f_n. A1 (Normalization) and A2 (Weak additivity) imply that $S(p_{\mathrm{mix},\Omega_n}) = \lfloor (H(p) - \epsilon)n / \log 2 \rfloor \log 2$. A2 (Weak additivity) and (Monotonicity) imply that $S(p^n \circ f_n^{-1}) \le S(p^n) \le nS(p)$. By using these relations, A4 (Asymptotic continuity) implies that $H(p) - \epsilon \le S(p)$. Since ϵ is arbitrary, we have $H(p) \le S(p)$. Similarly, using Lemma 2.6, we can show that $H(p) \ge S(p)$. Thus, we obtain $H(p) = S(p)$.

Now, we show (3) $\Rightarrow$ (1). K1 (Normalization), K2 (Continuity), and K3 (Non-negativity) are oblivious from the definition (2.2). K4 (Expandability) and K5 (Chain rule) follow from (2.4) and (2.5), respectively. ■

To show Lemmas 2.5 and 2.6, we prepare another lemma as follows.

Lemma 2.7 (Han [26, Lemma 2.1.1.]) *For any two distributions* P_X *on* $\mathcal{X}$ *and* P_Y *on* $\mathcal{Y}$*, there exists a function* f *from* $\mathcal{X}$ *to* $\mathcal{Y}$ *such that*

$$d_1(\mathrm{P}_{f(X)}, \mathrm{P}_Y) \le e^{-\gamma} + \max(\mathrm{P}_X(S(a+\gamma)^c), \mathrm{P}_Y(T(a)^c)), \tag{2.204}$$

where

$$S(a) := \{x \in \mathcal{X} | \mathrm{P}_X(x) \le e^{-a}\}, \quad T(a) := \{y \in \mathcal{Y} | \mathrm{P}_Y(y) \ge e^{-a}\}.$$

Proof We define a map f from $\mathcal{X}$ to $\mathcal{Y}$ as follows. We number all of elements of $T(a)$ as $T(a) = \{y_1, \ldots, y_n\}$. So, we have

$$n = |T(a)| \le e^a. \tag{2.205}$$

For this purpose, we define n disjoint subsets $f^{-1}(y_1), \ldots, f^{-1}(y_n)$ as subsets of $\mathcal{X}$. First, we choose a subset $f^{-1}(y_1) \subset S(a+\gamma)$ such that

$$\sum_{x \in f^{-1}(y_1)} \mathrm{P}_X(x) \le \mathrm{P}_Y(y_1) < \sum_{x \in f^{-1}(y_1)} \mathrm{P}_X(x) + e^{-a-\gamma}.$$

for any $x' \in S(a+\gamma) \backslash f^{-1}(y_1)$. Next, we choose a subset $f^{-1}(y_2) \subset S(a+\gamma) \backslash f^{-1}(y_1)$ such that

$$\sum_{x \in f^{-1}(y_2)} \mathrm{P}_X(x) \le \mathrm{P}_Y(y_2) < \sum_{x \in f^{-1}(y_2)} \mathrm{P}_X(x) + e^{-a-\gamma}.$$

We repeat this selection as long as possible. Let y_l be the final element y whose inverse set $f^{-1}(y)$ can be defined in this way.

Consider the case $l = n$. We reselect $f^{-1}(y_n)$ to be $(\cup_{i=1}^{n-1} f^{-1}(y_i))^c$. Then, the set $f^{-1}(y)$ is empty for $y \in T(a)^c$. Due to Exercise 2.12, we have

$$\begin{aligned} d_1(\mathrm{P}_{f(X)}, \mathrm{P}_Y) &\le \sum_{i=1}^{n-1} |\mathrm{P}_{f(X)}(y_i) - \mathrm{P}_Y(y_i)| + \sum_{y \in T(a)^c} |\mathrm{P}_{f(X)}(y) - \mathrm{P}_Y(y)| \\ &\le \sum_{i=1}^{n-1} e^{-a-\gamma} + \sum_{y \in T(a)^c} \mathrm{P}_Y(y) \overset{(a)}{\le} e^{-\gamma} + \mathrm{P}_Y(T(a)^c), \end{aligned}$$

where (a) follows from (2.205).

Next, we consider the case $l < n$. We define $f^{-1}(y_{l+1}) := \mathcal{X} \backslash (\cup_{i=1}^{l} f^{-1}(y_i))^c$. Then, for $y \in \{y_1, \ldots y_{l+1}\}^c$, $f^{-1}(y)$ is empty. Since

$$\sum_{i=1}^{l+1} \mathrm{P}_Y(y_i) \ge \sum_{x \in S(a+\gamma)} \mathrm{P}_X(x),$$

we have

$$\sum_{y \in \{y_1, \ldots y_{l+1}\}^c} \mathrm{P}_Y(y) \le \sum_{x \in S(a+\gamma)^c} \mathrm{P}_X(x). \tag{2.206}$$

Hence, due to Exercise 2.12, we have

$$d_1(\mathrm{P}_{f(X)}, \mathrm{P}_Y) \le \sum_{i=1}^{l} |\mathrm{P}_{f(X)}(y_i) - \mathrm{P}_Y(y_i)| + \sum_{y \in \{y_1, \ldots y_{l+1}\}^c} |\mathrm{P}_{f(X)}(y) - \mathrm{P}_Y(y)|$$

$$\le \sum_{i=1}^{l} e^{-a-\gamma} + \sum_{y \in \{y_1, \ldots y_{l+1}\}^c} \mathrm{P}_Y(y)$$

$$\overset{(a)}{\le} e^{-\gamma} + \sum_{x \in S(a+\gamma)^c} \mathrm{P}_X(x) = e^{-\gamma} + \mathrm{P}_X(S(a+\gamma)^c),$$

where (a) follows from (2.205) and (2.206). ■

Now, using Lemma 2.7, we show Lemmas 2.5 and 2.6.

Proof of Lemma 2.5 We apply Lemma 2.7 to the case when $a = (H(p) - \epsilon)n$, $\gamma = n\frac{\epsilon}{2}$, and P_X and P_Y are p^n and the uniform distribution $p_{\mathrm{mix},\Omega_n}$ on the set $\Omega_n = \{0, 1\}^{\lfloor (H(p)-\epsilon)n/\log 2 \rfloor}$, respectively. Then, $\mathrm{P}_Y(T(a)^c) = 0$ and $e^{-\gamma} \to 0$. Since RHS of (2.44) goes to zero with $R < H(p)$, we have $\mathrm{P}_X(S(a+\gamma)^c) \to 0$. Therefore, we obtain the desired argument. ■

Proof of Lemma 2.6 We apply Lemma 2.7 to the case when $a + \gamma = (H(p) + \epsilon)n$, $\gamma = n\frac{\epsilon}{2}$, and P_Y and P_X are p^n and the uniform distribution $p_{\mathrm{mix},\Omega'_n}$ on the set $\Omega'_n = \{0, 1\}^{\lfloor (H(p)+\epsilon)n/\log 2 \rfloor}$, respectively. Then, $\mathrm{P}_X(S(a+\gamma)^c) = 0$ and $e^{-\gamma} \to 0$. Since RHS of (2.42) goes to zero with $R > H(p)$, we have $\mathrm{P}_Y(T(a)^c) \to 0$. Therefore, we obtain the desired argument. ■

Exercises

2.49 Show that the Rényi entropy $H_{1+s}(p)$ and the min entropy $H_{\min}(p)$ do not satisfy A4 (Asymptotic continuity) for $s > 0$ as follows.
(**a**) Define the distribution $p_{d,\epsilon}$ on $\{0, 1, \ldots, d-1\}$ by

$$p_{d,\epsilon}(i) := \begin{cases} \frac{1}{d} + \epsilon & \text{if } i = 0 \\ \frac{1}{d} - \frac{\epsilon}{d-1} & \text{if } i > 0. \end{cases} \tag{2.207}$$

Show that $d_1(p_{d,\epsilon}, p_{\mathrm{mix},d}) = \epsilon$.
(**b**) Show that $H_{\min}(p_{d,\epsilon}) = \log d - \log(1 + d\epsilon)$.
(**c**) Assume that $d\epsilon \to \infty$ as $d \to \infty$. Show that $\frac{H_{\min}(p_{\mathrm{mix},d}) - H_{\min}(p_{d,\epsilon})}{\log d} = 1 + \frac{\log \epsilon}{\log d} + O(\frac{1}{d\epsilon \log d})$ as $d \to \infty$.
(**d**) Show that $H_{1+s}(p_{d,\epsilon}) = \log d - \frac{1}{s}\log(\frac{1}{d}(1 + d\epsilon)^{1+s} + \frac{d-1}{d}(1 - \frac{d\epsilon}{d-1})^{1+s})$.
(**e**) Assume that $\frac{1}{d}(d\epsilon)^{1+s} \to \infty$ as $d \to \infty$. Show that $\frac{H_{1+s}(p_{\mathrm{mix},d}) - H_{1+s}(p_{d,\epsilon})}{\log d} = 1 + \frac{(1+s)\log \epsilon}{s \log d} + O((d\epsilon)^{-(1+s)}\frac{d}{\log d})$ as $d \to \infty$.

2.50 Show that the Rényi entropy $H_{1-s}(p)$ and the max entropy $H_{\max}(p)$ do not satisfy A4 (Asymptotic continuity) for $s \in (0, 1)$ as follows.
(**a**) Define the distribution $p'_{d,\epsilon}$ on $\{0, 1, \ldots, d-1\}$ by

$$p'_{d,\epsilon}(i) := \begin{cases} 1-\epsilon & \text{if } i = 0 \\ \frac{\epsilon}{d-1} & \text{if } i > 0. \end{cases} \tag{2.208}$$

Show that $d_1(p'_{d,\epsilon}, p'_{d,0}) = \epsilon$.
(**b**) Show that $\frac{H_{\max}(p'_{\mathrm{mix},d}) - H_{\max}(p'_{d,0})}{\log d} = 1$ for $\epsilon > 0$.
(**c**) Show that $H_{1-s}(p'_{d,\epsilon}) = -\frac{1}{s}\log((1-\epsilon)^{1-s} + (d-1)(\frac{\epsilon}{d-1})^{1-s})$.
(**d**) Show that $\frac{H_{1+s}(p_{\mathrm{mix},d}) - H_{1+s}(p_{d,\epsilon})}{\log d} = \frac{1-s}{s\log d}(d-1)^s\epsilon^{1-s} + O(\frac{\epsilon}{\log d}) + O(\frac{\epsilon^{2(1-s)}}{\log d})$ as $\epsilon \to 0$.

2.51 Show that $e^{-2H_2(p)}$ satisfies A4 (Asymptotic continuity) for $s > 0$ by showing the following inequality. That is, show that the continuity of $e^{-2H_2(p)}$ does not depend on the cardinality of the supports of p and q.

$$|e^{-H_2(p)} - e^{-H_2(q)}| \le 2d_1(p, q). \tag{2.209}$$

2.6 Large Deviation on Sphere

Next, we consider a probability distribution on the set of pure states. In quantum information, if we have no information on the given system $\mathcal{H} = \mathbb{C}^l$, it is natural to assume that the probability distribution is invariant with respect to the action of the unitary group $\mathrm{U}(l)$ on the set of pure states. Such a distribution is unique and is called the Haar measure, which is denoted by $\mu_{\mathcal{H}}$. Since the normalized vector is given as $|\phi\rangle \in \mathbb{C}^l$ satisfying $\|\phi\| = 1$, the distribution $\mu_{\mathcal{H}}$ is given as a distribution on the set of pure states satisfying that

$$\int_B \mu_{\mathcal{H}}(d\phi) = \int_B \mu_{\mathcal{H}}(dU\phi) \text{ for } U \in \mathrm{U}(l). \tag{2.210}$$

That is, the Haar measure is defined as the unique distribution satisfying (2.210). When the pure state is regarded as an element of the $2l-1$-dimensional sphere S^{2l-1}, the distribution $\mu_{\mathcal{H}}$ is given as a distribution on the $2l-1$-dimensional sphere. More generally, the Haar measure μ_{S^n} on n-dimensional sphere S^n is given as the distribution satisfying that

$$\int_B \mu_{S^n}(dx) = \int_B \mu_{S^n}(dgx) \text{ for } g \in \mathrm{O}(n+1). \tag{2.211}$$

The Haar measure has several useful properties. For example, the invariance guarantees that

$$\int |\phi\rangle\langle\phi| \mu(d\phi) = \frac{1}{l} I. \tag{2.212}$$

Further, when $\mathcal{H} = \mathbb{C}^l$ is spanned by the basis $\{|e_i\rangle\}_{i=1}^l$, for n-th permutation π, we define the unitary U_π on $\mathcal{H}^{\otimes n}$ as

$$U_\pi(|v_1, \ldots, v_n\rangle) := |v_{\pi(1)}, \ldots, v_{\pi(n)}\rangle. \tag{2.213}$$

Then, we define the n-th symmetric subspace $\mathcal{H}_{s,n} \subset \mathcal{H}^{\otimes n}$ as the space spanned by $\{\sum_\pi U_\pi(|e_1, \ldots, e_1, e_2, \ldots, e_2, \ldots, e_l, \ldots, e_l\rangle)\}$. The dimension of $\mathcal{H}_{s,n}$ is $\binom{l+n-1}{l-1}$, and the invariance implies that

$$\int |\phi\rangle\langle\phi|^{\otimes n} \mu_{\mathcal{H}}(d\phi) = \frac{1}{\binom{l+n-1}{l-1}} P_{\mathcal{H}_{s,n}}, \tag{2.214}$$

where $P_{\mathcal{H}_{s,n}}$ is the projection to $\mathcal{H}_{s,n}$. When a pure state ρ on $\mathcal{H}^{\otimes n}$ is invariant for U_π with an arbitrary n-th permutation π, the pure state ρ is a state on $\mathcal{H}_{s,n}$. Hence, we have

$$\rho \le \binom{l+n-1}{l-1} \int |\phi\rangle\langle\phi|^{\otimes n} \mu_{\mathcal{H}}(d\phi). \tag{2.215}$$

Here, $\binom{l+n-1}{l-1}$ is upper bounded by $(n+1)^{d-1}$.

In quantum information, we often consider the stochastic behavior of a function of a pure state under the Haar measure $\mu_{\mathcal{H}}$. In order to discuss this issue, we need the following preparation. First, we define the median of a real-valued random variable X as

$$\mathrm{Med}_p(X) \stackrel{\mathrm{def}}{=} \frac{\overline{\mathrm{Med}_p(X)} + \underline{\mathrm{Med}_p(X)}}{2} \tag{2.216}$$

$$\overline{\mathrm{Med}_p(X)} \stackrel{\mathrm{def}}{=} \inf\{r | p\{x | x \ge r\} < 1/2\} \tag{2.217}$$

$$\underline{\mathrm{Med}_p(X)} \stackrel{\mathrm{def}}{=} \sup\{r | p\{x | x \le r\} < 1/2\}. \tag{2.218}$$

The cumulative distribution function of the real-valued random variable X is defined as

$$F_{X,p}(a) := p\{x | x \le a\}, \tag{2.219}$$

where $p(\Omega)$ is defined for a subset $S \subset \Omega$ as

$$p(S) := \sum_{x \in S} p_x. \tag{2.220}$$

Then, we have the following lemma.

Lemma 2.8 *When given two real-valued random variables X and Y satisfies $F_{X,p} \leq F_{Y,p}$, we have $\mathrm{E}_p X \geq \mathrm{E}_p Y$.*

Then, we define the metric $d(x, y)$ between two wave functions x and y in S^{2l-1} as

$$d(x, y) := \cos^{-1} \mathbf{Re}\langle x, y\rangle \in [0, \pi]. \tag{2.221}$$

Then, for a wave function $y \in S^{2l-1}$, we define the subset $D(y, r)$ as

$$D(y, r) := \{x \in S^{2l-1} | d(x, y) \leq r\}. \tag{2.222}$$

Then, the probability $\mu_{S^{2l-1}}(D(y, r))$ depends only on r. For a given probability $p \in (0, 1)$, we define $r(p)$ as $\mu_{S^{2l-1}}(D(y, r(p))) = p$. For a given subset $\Omega \subset S^{2l-1}$, we define the subset Ω_ϵ for $\epsilon > 0$ as

$$\Omega_\epsilon := \{x \in S^{2l-1} | d(x, y) \leq \epsilon, \exists y \in \Omega\}. \tag{2.223}$$

Then, we prepare the following fundamental lemma.

Lemma 2.9 ([27, Theorem 2.1]) *For a given $p \in (0, 1)$ and $\epsilon > 0$, we have*

$$\min\{\mu_{S^{2l-1}}(\Omega_\epsilon) | \mu_{S^{2l-1}}(\Omega) = p\} = \mu_{S^{2l-1}}(D(y, r(p))_\epsilon), \tag{2.224}$$

where the set $D(y, r(p))_\epsilon$ is illustrated as Fig. 2.2.

Proof We give only an intuitive proof. First, we consider an infinitesimal $\epsilon > 0$. In this case, it is enough to consider the boundary of Ω because the size of boundary

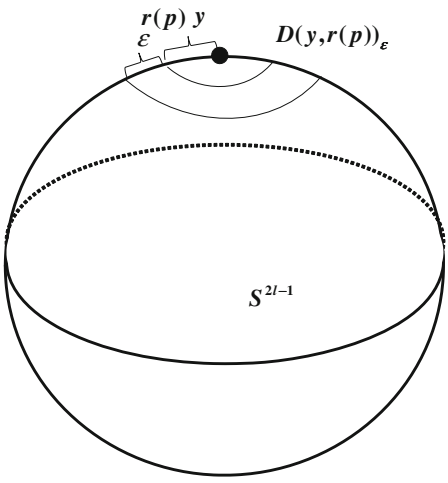

Fig. 2.2 Set $D(y, r(p))_\epsilon$

of Ω is proportional to $\frac{d\mu_{S^{2l-1}}(\Omega_\epsilon)}{d\epsilon}|_{\epsilon=0}$. We can intuitively find that the set $D(y, r(p))$ has the minimum boundary among the subsets Ω satisfying $\mu_{S^{2l-1}}(\Omega) = p$. That is, we obtain $\frac{d\mu_{S^{2l-1}}(\Omega_\epsilon)}{d\epsilon}|_{\epsilon=0} \geq \frac{d\mu_{S^{2l-1}}(D(y,r(p))_\epsilon)}{d\epsilon}|_{\epsilon=0}$.

Next, for $p' > p$ and a subset Ω satisfying $\mu_{S^{2l-1}}(\Omega) = p$, we define the function $f(p', \Omega)$ as $\mu_{S^{2l-1}}(\Omega_{f(p',\Omega)}) = p'$. Then, we have

$$\frac{df(p', \Omega)}{dp'} = \frac{1}{\frac{d\mu_{S^{2l-1}}(\Omega_{f(p',\Omega)+\epsilon})}{d\epsilon}|_{\epsilon=0}} \leq \frac{1}{\frac{d\mu_{S^{2l-1}}(D(y,r(p))_{f(p',D(y,r(p)))+\epsilon})}{d\epsilon}|_{\epsilon=0}}, \tag{2.225}$$

which implies

$$f(p', \Omega) \leq f(p', D(y, r(p))). \tag{2.226}$$

Hence, we obtain

$$\mu_{S^{2l-1}}(\Omega_{f(p',D(y,r(p)))}) \geq \mu_{S^{2l-1}}(\Omega_{f(p',\Omega)}) = \mu_{S^{2l-1}}(D(y, r(p))_{f(p',D(y,r(p)))}).$$

■

Using the above lemma, we obtain the following lemma.

Lemma 2.10 ([27, Corollary 2.2]) *When a subset $\Omega \subset S^{2l-1}$ satisfies $\mu_{S^{2l-1}}(\omega) \geq \frac{1}{2}$, we have*

$$\mu_{S^{2l-1}}(\Omega_\epsilon) \geq 1 - e^{-\epsilon^2(l-1)}/2. \tag{2.227}$$

Proof Thanks to Lemma 2.9, since $D(y, \frac{\pi}{2}) = \frac{1}{2}$, it is enough to show that $D(y, \frac{\pi}{2})_\epsilon = D(y, \frac{\pi}{2} + \epsilon) \geq 1 - e^{\epsilon^2(l-1)}/2$. The size of the boundary of $D(y, \theta)$ is proportional to $\sin^{2l-2}\theta = \cos^{2l-2}(\theta - \frac{\pi}{2})$ for $\theta \in [0, \pi]$. Hence, choosing $\theta' := \theta - \frac{\pi}{2}$, we have

$$D\left(y, \frac{\pi}{2} + \epsilon\right) = \frac{\int_{-\frac{\pi}{2}}^{\epsilon} \cos^{2l-2}\theta' d\theta'}{I_{l-1}}, \tag{2.228}$$

where

$$I_{l-1} := \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^{2l-2}\theta' d\theta' = B\left(l - \frac{1}{2}, \frac{1}{2}\right) = \frac{\Gamma(l - \frac{1}{2})\Gamma(\frac{1}{2})}{\Gamma(l)} = \frac{2l-3}{2l-2} I_{l-2}. \tag{2.229}$$

Since $\frac{2l-3}{\sqrt{2l-2}\sqrt{2l-4}} \geq 1$, we have

$$\sqrt{2l-2} I_{l-1} \geq \sqrt{2l-4} I_{l-2}, \tag{2.230}$$

which implies $\sqrt{2l-2} I_{l-1} \geq \sqrt{2} I_1 = \sqrt{2} B\left(\frac{3}{2}, \frac{1}{2}\right) = \frac{\pi}{\sqrt{2}}$.

For $t \in [0, \frac{\pi}{2}]$, the inequality $\cos t \le e^{-\frac{t^2}{2}}$ holds. Using the parameter $u := \sqrt{l-1}\theta$, we have

$$1 - D\left(y, \frac{\pi}{2} + \epsilon\right) = \frac{\int_{\epsilon}^{\frac{\pi}{2}} \cos^{2l-2}\theta' d\theta'}{I_{l-1}} = \frac{1}{\sqrt{l-1}} \frac{\int_{\epsilon\sqrt{l-1}}^{\frac{\pi}{2}\sqrt{l-1}} \cos^{2l-2}\frac{u}{\sqrt{l-1}} du}{I_{l-1}}$$
$$\le \frac{\int_{\epsilon\sqrt{l-1}}^{\frac{\pi}{2}\sqrt{l-1}} \left(e^{-\frac{u^2}{2(l-1)}}\right)^{2l-2} du}{\frac{\pi}{\sqrt{2}}} = \frac{\int_{\epsilon\sqrt{l-1}}^{\frac{\pi}{2}\sqrt{l-1}} e^{-u^2} du}{\frac{\pi}{\sqrt{2}}} \le e^{-\epsilon^2(l-1)}/2,$$

where the final inequality follows from Exercise 2.56. ■

A real-valued continuous function f of S^{2l-1} can be regarded as a real-valued random variable on S^{2l-1}. Then, we define the set Ω_f as

$$\Omega_f := \{x \in S^{2l-1} | f(x) \le \mathrm{Med}_{S^{2l-1}}(f)\}, \tag{2.231}$$

where $\mathrm{Med}_{S^{2l-1}}(f)$ is the abbreviation of the median $\mathrm{Med}_{\mu_{S^{2l-1}}}(f)$ under the Haar measure $\mu_{S^{2l-1}}$ on S^{2l-1}. Using Lemma 2.9, we obtain the inequality

$$\mu_{S^{2l-1}}((\Omega_f)_\epsilon) \ge 1 - e^{-\epsilon^2(l-1)}/2. \tag{2.232}$$

Now, we say that the function f is Lipschitz continuous with the Lipschitz constant C_0 with respect to the metric d in subset $\Omega \subset S^{2l-1}$ when

$$\frac{|f(x) - f(y)|}{d(x, y)} \le C_0, \quad \forall x, y \in \Omega. \tag{2.233}$$

In particular, when $\Omega = S^{2l-1}$, we simply say that the function f is Lipschitz continuous with the Lipschitz constant C_0 with respect to the metric d, which is assumed in the following. Since $(\Omega_f)_\epsilon \subset \{x \in S^{2l-1} | f(x) \ge \mathrm{Med}_{S^{2l-1}}(f) + C_0\epsilon\}^c$, (2.232) implies that

$$\mu_{S^{2l-1}}\{x \in S^{2l-1} | f(x) \ge \mathrm{Med}_{S^{2l-1}}(f) + C_0\epsilon\} \le \mu_{S^{2l-1}}((\Omega_f)_\epsilon^c) \le \frac{e^{-\epsilon^2(l-1)}}{2}. \tag{2.234}$$

Similarly, we can show that

$$\mu_{S^{2l-1}}\{x \in S^{2l-1} | f(x) \le \mathrm{Med}_{S^{2l-1}}(f) - C_0\epsilon\} \le \frac{e^{-\epsilon^2(l-1)}}{2}. \tag{2.235}$$

Hence, we obtain

$$\mu_{S^{2l-1}}\{x \in S^{2l-1} | \, |f(x) - \mathrm{Med}_{S^{2l-1}}(f)| \ge \epsilon\} \le e^{-\frac{\epsilon^2(l-1)}{C_0^2}}, \tag{2.236}$$

which implies that the cumulative distribution function of the real-valued random variable $|f(x) - \mathrm{Med}_{S^{2l-1}}(f)|$ is less than $F(x) := 1 - e^{-\frac{x^2(l-1)}{C_0^2}}$. Now, we simplify the expectation $\mathrm{E}_{\mu_{S^{2l-1}}}$ under the Haar measure $\mu_{S^{2l-1}}$ on S^{2l-1} to $\mathrm{E}_{S^{2l-1}}$. Thus, Lemma 2.8 guarantees that

$$\begin{aligned}&\mathrm{E}_{S^{2l-1}}|f(X) - \mathrm{Med}_{S^{2l-1}}(f)| \le \int_0^\infty x \frac{dF(x)}{dx} dx \\ =& \int_0^\infty \frac{2(l-1)}{C_0^2} x^2 e^{-\frac{x^2(l-1)}{C_0^2}} dx = \frac{C_0}{2}\sqrt{\frac{\pi}{l-1}},\end{aligned}$$

where we used the relation in Exercise 2.55. Thus, we obtain

$$|\mathrm{E}_{S^{2l-1}} f(X) - \mathrm{Med}_{S^{2l-1}}(f)| \le \mathrm{E}_{S^{2l-1}}|f(X) - \mathrm{Med}_{S^{2l-1}}(f)| \le \frac{C_0}{2}\sqrt{\frac{\pi}{l-1}}. \tag{2.237}$$

Finally, given positive numbers δ and C_1, we define the sets

$$\begin{aligned}\Omega_{\delta,C_1} :=& \left\{x \in S^{2l-1} \middle| f(x) \ge \mathrm{E}_{S^{2l-1}} f(X) + \frac{C_0}{2}\sqrt{\frac{\pi}{l-1}} + C_1\delta\right\} \\ \subset& \{x \in S^{2l-1} | f(x) \ge \mathrm{Med}_{S^{2l-1}}(f) + C_1\delta\}, \\ \tilde{\Omega}_{\delta,C_1} :=& \left\{x \in S^{2l-1} \middle| \mathrm{E}_{S^{2l-1}} f(X) - \frac{C_0}{2}\sqrt{\frac{\pi}{l-1}} < f(x) \right. \\ & \left. < \mathrm{E}_{S^{2l-1}} f(X) + \frac{C_0}{2}\sqrt{\frac{\pi}{l-1}} + C_1\delta\right\} \\ \supset& \{x \in S^{2l-1} | \mathrm{Med}_{S^{2l-1}}(f) < f(x) < \mathrm{Med}_{S^{2l-1}}(f) + C_1\delta\}.\end{aligned}$$

Then, we obtain the large deviation type bound with respect to the Haar measure on the $2l-1$-dimensional sphere as follows.

Theorem 2.11 *When the function $f(x)$ has the Lipschitz constant C_1 on the subset $\tilde{\Omega}_{\delta,C_1}$, we have*

$$\mu_{S^{2l-1}}(\Omega_{\delta,C_1}) \le e^{-\delta^2(l-1)}/2. \tag{2.238}$$

Here, C_0 is the Lipschitz constant for the whole set, and C_1 is the Lipschitz constant for the specific subset $\tilde{\Omega}_{\delta,C_1}$.

Next, we apply the Haar measure to construct a proper subset of S^{2l-1}. A subset Ω of S^{2l-1} is called an ϵ net of S^{2l-1} when for any element $x \in S^{2l-1}$, there exists an element $y \in S^{2l-1}$ such that $d(x, y) \le \epsilon$.

Lemma 2.11 *There exists an ϵ net Ω of S^{2l-1} whose cardinality is less than* $\frac{\sqrt{(2l-1)\pi}}{\sin^{2l-1}\frac{\epsilon}{2}} < \sqrt{(2l-1)}\pi(\frac{2}{\sin\epsilon})^{2l-1}$.

Proof We choose a subset Ω of S^{2l-1} satisfying the condition that $d(x, y) > \epsilon$ for any two distinct elements $x, y \in \Omega$. We choose the subset Ω so that no subset Ω' strictly larger than Ω satisfies the required condition. Here, a set Ω' is called strictly larger than Ω when Ω' contains Ω and there is at least an element of Ω' that is not included in Ω. A rigorous proof of the existence of such a subset can be given by using Zorn's lemma.

Hence, for any element $x \in S^{2l-1}$, there exists an element $y \in S^{2l-1}$ such that $d(x, y) \leq \epsilon$. That is, the set Ω is an ϵ net of S^{2l-1}. Due to the construction, $D(x, \epsilon/2) \cap D(y, \epsilon/2) = \emptyset$ for any two distinct elements $x, y \in \Omega$. Thus, $|\Omega| \mu_{S^{2l-1}}(D(x, \epsilon/2)) = \sum_{x \in \Omega} \mu_{S^{2l-1}}(D(x, \epsilon/2)) \leq 1$. That is, $|\Omega| \leq \frac{1}{\mu_{S^{2l-1}}(D(x, \epsilon/2))}$. The probability $\mu_{S^{2l-1}}(D(x, \epsilon/2))$ is evaluated by using Exercise 2.57 as

$$\begin{aligned}
&\mu_{S^{2l-1}}(D(x, \epsilon/2)) = \int_0^{\epsilon/2} \sin^{2l-2}\theta d\theta / I_{l-1} \geq \int_0^{\epsilon/2} \left(\frac{\sin\frac{\epsilon}{2}}{\epsilon/2}\theta\right)^{2l-2} d\theta / I_{l-1} \\
=&\frac{\sin^{2l-2}\frac{\epsilon}{2}}{(\epsilon/2)^{2l-2}}[\theta^{2l-1}/(2l-1)I_{l-1}]_0^{\epsilon/2} = \frac{\sin^{2l-2}\frac{\epsilon}{2}}{(\epsilon/2)^{2l-2}}(\frac{\epsilon}{2})^{2l-1}/(2l-1)I_{l-1} \\
=&\frac{\frac{\epsilon}{2}\sin^{2l-2}\frac{\epsilon}{2}}{(2l-1)I_{l-1}} \geq \frac{\sin^{2l-1}\frac{\epsilon}{2}}{\sqrt{(2l-1)\pi}}.
\end{aligned}$$

where the relation $\frac{\epsilon}{2} \geq \sin\frac{\epsilon}{2}$ is used. ■

Exercises

2.52 Show that $\||x\rangle\langle x| - |y\rangle\langle y|\|_1 \leq 2\sin\epsilon$ when $d(x, y) = \epsilon \leq \frac{\pi}{2}$ and $x, y \in S^{2l-1}$.

2.53 Show that $\||x\rangle\langle x| - |y\rangle\langle y|\|_2 \leq \sqrt{2}d(x, y)$.

2.54 Show that $\||x\rangle - |y\rangle\| \leq 2\sin\frac{d(x,y)}{2} \leq d(x, y)$.

2.55 Show that $\int_0^\infty 2cx^2 e^{-cx^2} dx = \frac{1}{2}\sqrt{\frac{\pi}{c}}$.

2.56 Show $\frac{\int_{\epsilon\sqrt{l-1}}^{\frac{\pi}{2}\sqrt{l-1}} e^{-u^2} du}{\frac{\pi}{\sqrt{2}}} \leq e^{-\epsilon^2(l-1)}/2$ when $u \geq 0$ and $\epsilon > 0$.

2.57 Show that

$$(2l-1)B\left(l - \frac{1}{2}, \frac{1}{2}\right) \leq \sqrt{(2l-1)\pi} \tag{2.239}$$

by following the steps below.

(**a**) Show the equation $B(l - \frac{1}{2}, \frac{1}{2}) = \pi \cdot \prod_{k=1}^{l-1} \frac{2k-1}{2k}$.
(**b**) Show the inequality $\sum_{k=1}^{l-1} \log\frac{2k}{2k-1} \geq \frac{1}{2}\log(2l-1)$.
(**c**) Show the inequality (2.239).

2.7 Related Books

In this chapter, we treat several important topics in information science from the probabilistic viewpoint. In Sect. 2.1, information quantities e.g., entropy, relative entropy, mutual information, Rényi entropy, and conditional Rényi entropy are discussed. Its discussion and its historical notes except for Rényi entropy and Conditional Rényi entropy appear in Chap. 2 of Cover and Thomas [28]. Conditional Rényi entropy is recently introduced and discussed by several papers [29–31] from various viewpoints. This quantity will be investigated much more deeply in future.

Section 2.2 focuses on information geometry. Amari and Nagaoka [2] is a textbook on this topic written by the pioneers in the field. Bregman divergence plays a central role in this section. Although their book [2] contains the Bregman divergence, it discusses information geometry from a more general viewpoint. Recent Amari's paper [6] focuses on the Bregman divergence and derives several important theorems only from the structure of Bregman divergence. This section follows his derivation.

Section 2.3 briefly treats the estimation theory of probability distribution families. Lehmann and Casella [32] is a good textbook covering all of estimation theory. For a more in-depth discussion of its asymptotic aspect, see van der Vaart [7].

Section 2.4.1 reviews the type method. It has been formulated by Csiszár and Köner [8]. Section 2.4.2 treats the large deviation theory including estimation theory. Its details are given in Dembo and Zeitouni [33] and Bucklew [34]. In this book, we give a proof of Cramér's theorem and the Gártner–Ellis theorem. In fact, (2.163), (2.165), (2.169), and (2.171) follow from Markov's inequality. However, its opposite parts are not simple. Many papers and books give their proof. In this book, we prove these inequalities by combining the estimation of the exponential theory and the Legendre transform. This proof seems to be the simplest of known proofs.

Section 2.5 explains how to derive the entropy from natural axioms. This section addresses two sets of axioms. One is close to the axioms proposed by Khinchin [25]. The other is related to asymptotic continuity, and has not been given in anywhere. The latter is related to the entropy measure discussed in Sect. 8.7.

Section 2.6 focuses on the Haar measure, which is a natural distribution on the set of pure states. Milman and Schechtman [27] discusses the asymptotic behavior of a function of the random variable subject to the Haar measure. Since this type discussion attracts much attention in quantum information recently and is applied in Sects. 8.13, 2.6 is devoted to this topic.

2.8 Solutions of Exercises

Exercise 2.1 When $y = f(x)$, $P_{X,Y}(x, y) = P_X(x)$. Hence, $H(X, f(X)) = -\sum_{x,y:y=f(x)} P_{X,Y}(x, y) \log P_{X,Y}(x, y) = -\sum_x P_X(x) \log P_X(x) = H(X)$.

Exercise 2.2 Consider the case $\mathrm{P}_Y(1) = \lambda, \mathrm{P}_Y(0) = 1 - \lambda, \mathrm{P}_{X|Y=1} = p, \mathrm{P}_{X|Y=0} = p'$.

Exercise 2.3 The concavity of entropy guarantees that the maximum of $H(p)$ under the above condition is realized by the distribution $(a, \frac{1-a}{k-1}, \ldots, \frac{1-a}{k-1})$, whose entropy is $h(a) + (1-a)\log(k-1)$.

Exercise 2.4 $H(p_A \times p_B) = -\sum_{\omega_A,\omega_B} p_A(\omega_A)p_B(\omega_B)\log(p_A(\omega_A)p_B(\omega_B)) = -\sum_{\omega_A} p_A(\omega_A)\log p_A(\omega_A) - \sum_{\omega_B} p_B(\omega_B)\log p_B(\omega_B) = H(p_A) + H(p_B)$.

Exercise 2.5 $D(p_A \times p_B \| q_A \times q_B) = \sum_{\omega_A,\omega_B} p_A(\omega_A)p_B(\omega_B)(\log(p_A(\omega_A)p_B(\omega_B)) - \log(q_A(\omega_A)q_B(\omega_B))) = \sum_{\omega_A} p_A(\omega_A)(\log p_A(\omega_A) - \log q_A(\omega_A)) + \sum_{\omega_B} p_B(\omega_B)(\log p_B(\omega_B) - \log q_B(\omega_B)) = D(p_A\|q_A) + D(p_B\|q_B)$.

Exercise 2.6 Define $f(x) := \log x - (x-1)$. Since $f'(x) = \frac{1}{x} - 1$, we find that the maximum of $f(x)$ is attained only when $x = 1$. That is, $f(x) < f(1) = 0$.

Exercise 2.7 Apply a stochastic transition matrix of rank 1 to Theorem 2.1.

Exercise 2.8 $D_f(p\|q) = \sum_i p_i \left(1 - \sqrt{\frac{q_i}{p_i}}\right) = 1 - \sum_i \sqrt{p_i q_i} = \frac{1}{2}\sum_i \left(\sqrt{p_i} - \sqrt{q_i}\right)^2$.

Exercise 2.9 Use the fact that $\sum_j \sum_i Q^i_j |p_i - q_i| \geq \sum_j |\sum_i Q^i_j (p_i - q_i)|$.

Exercise 2.10 Consider the $x \geq y$ and $x < y$ cases separately.

Exercise 2.11

(**a**) Use $|p_i - q_i| = |\sqrt{p_i} - \sqrt{q_i}||\sqrt{p_i} + \sqrt{q_i}|$.
(**b**) Use $p_i + q_i \geq 2\sqrt{p_i}\sqrt{q_i}$.

Exercise 2.12 We find that $p_{x_0} - q_{x_0} = -\sum_{x \neq x_0}(p_x - q_x)$. Thus, $|p_{x_0} - q_{x_0}| \leq \sum_{x\neq x_0} |p_x - q_x|$. Hence, $d_1(p,q) = \frac{1}{2}|p_{x_0} - q_{x_0}| + \frac{1}{2}\sum_{x\neq x_0}|p_x - q_x| \leq \sum_{x\neq x_0} |p_x - q_x|$.

Exercise 2.13 Assume that the datum i generates with the probability distribution p_i. Apply Jensen's inequality to the random variable $\sqrt{q_i/p_i}$ and the convex function $-\log x$.

Exercise 2.14

(**a**) Since Schwartz inequality implies that $\|x\|\|y\| \geq \langle x, y\rangle$ and $\|x\|\|y\| \geq \langle y, x\rangle$, we have

$$\begin{aligned}&(\|x\| + \|y\|)^2 - (\|x\|^2 + \langle x, y\rangle + \langle y, x\rangle + \|y\|^2)\\ =&2\|x\|\|y\| - \langle x, y\rangle - \langle y, x\rangle \geq 0.\end{aligned}$$

(**b**)

$$(\|x\| + \|y\|)^2 \geq \|x\|^2 + \langle x, y\rangle + \langle y, x\rangle + \|y\|^2 = \|x + y\|^2.$$

(**c**) Substitute $\sqrt{p_i} - \sqrt{r_i}$ and $\sqrt{r_i} - \sqrt{q_i}$ into x and y in the inequality given in (**b**).

Exercise 2.15 Check that $\phi'(s|p\|q) = \dfrac{\sum_i p_i^{1-s} q_i^s (\log q_i - \log p_i)}{\sum_i p_i^{1-s} q_i^s}$.

Exercise 2.16 Check that $\phi''(s|p\|q) =$

$$\frac{(\sum_i p_i^{1-s} q_i^s)(\sum_i p_i^{1-s} q_i^s (\log q_i - \log p_i)^2) - (\sum_i p_i^{1-s} q_i^s (\log q_i - \log p_i))^2}{(\sum_i p_i^{1-s} q_i^s)^2}.$$

Next, use Schwarz's inequality between two vectors 1 and $(-\log p_i + \log q_i)$.

Exercise 2.17 For $0 < s < s'$, we have $\frac{s}{s'} f(s') = (1 - \frac{s}{s'}) f(0) + \frac{s}{s'} f(s') \ge f((1 - \frac{s}{s'}) \cdot 0 + \frac{s}{s'} \cdot s') = f(s)$, which implies that $\frac{f(s')}{s'} \ge \frac{f(s)}{s}$. Similarly, for $0 > s > s'$, we have $\frac{f(s')}{s'} \le \frac{f(s)}{s}$. Thus, $\frac{f(s)}{s}$ is monotone increasing When $f(s)$ is strictly convex for s, the above inequalities $\le$ and $\ge$ can be replaced by $<$ and $>$. Hence, $\frac{f(s)}{s}$ is strictly monotone increasing

Exercise 2.18

(**a**) For simplicity, we denote $\max(b_1, \ldots, b_k)$ by b_M. We choose a subset $S \subset \{1, \ldots, k\}$ such that $b_M = b_i$ for $i \in S$ and $b_M > b_i$ for $i \notin S$. Thus, $\frac{1}{t} \log(\sum_{i=1}^k a_i b_i^t) = \log b_M + \frac{1}{t} \log(\sum_{i \in S} a_i + \sum_{i \notin S} a_i (\frac{b_i}{b_M})^t) \to \log b_M + \frac{1}{t} \log(\sum_{i \in S} a_i) \to \log b_M$ as $t \to \infty$.

Exercise 2.19 $\sum_i p_i^{1-s} q_i^s = \sum_{i:p_i>0} p_i^{1-s} q_i^s \to \sum_{i:p_i>0} q_i^s$ as $s \to 1$.

Exercise 2.20 Solving the equation that the partial derivative equals zero on the RHS. Then, we obtain $\lambda_i = p_i/q_i$. Substituting it into th RHS, we obtain the LHS.

Exercise 2.21 Apply the formula (2.32) to the conditional distribution $P_{XYZ|U=u}$. Then, we have

$$I(X : YZ|U = u) = I(X : Z|U = u) + \sum_z P_Z(z) I(X : Y|Z = z, U = u). \tag{2.240}$$

Taking the expectation for U, we obtain (2.33).

Exercise 2.22 $e^{\psi(s|p_A \times p_B)} = \sum_{a,b} p_A(a)^{1-s} p_B(b)^{1-s} = \sum_a p_A(a)^{1-s} \sum_b p_B(b)^{1-s} = e^{\psi(s|p_A)} e^{\psi(s|p_B)}$.

Exercise 2.23

(**b**)

$$\begin{aligned} & D(q\|p) - \frac{1}{1-s} D(q\|p_s) \\ =& \sum_x q(x)(\log q(x) - \log p(x)) - \frac{1}{1-s} \sum_x q(x) \log q(x) \end{aligned}$$

$$
\begin{aligned}
&+ \sum_x q(x) \log p(x) + \frac{\psi(s)}{1-s} \\
&= - \frac{s}{1-s} \sum_x q(x) \log q(x) + \frac{\psi(s)}{1-s} \\
&= - \frac{s}{1-s} H(q) + \frac{\psi(s)}{1-s} = \frac{s}{1-s} H(p_s) + \frac{\psi(s)}{1-s} \\
&= - \frac{s}{1-s} \sum_x p_s(x)(1-s) \log p(x) - \frac{s}{1-s} \psi(s) + \frac{\psi(s)}{1-s} \\
&= - s \sum_x p_s(x) \log p(x) + \psi(s) = D(p_s \| p).
\end{aligned}
$$

(**c**) The desired inequality follows from the inequality $\frac{1}{1-s} D(q \| p_s) \geq 0$ for $s \leq 1$.

Exercise 2.24

(**a**) It follows from $\psi(s) \geq H(p)$ for $s \in [0, 1]$.
(**b**) The left hand side is zero when $s = 0$.
(**c**) $\psi'(s) = - \sum_x p_s(x) \log p(x)$, $H(p_s) = -(1-s) \sum_x p_s(x) \log p(x) + \psi(s)$, and $D(p_s \| p) = \sum_x p_s(x) \log p(x) - \psi(s)$.
(**e**) It follows from the relations $\frac{d}{ds} H(p_s) < 0$ and $H(p_1) = H(p) < R$.
(**f**) It follows from Exercise 2.23.
(**g**) It follows from (**f**) and the continuity of $H(q)$ and $D(q \| p)$ for q.
(**h**) Since $\psi'(s_R) = (H(p_{s_R}) - \psi(s_R))/(1 - s_R) = (R - \psi(s_R))/(1 - s_R)$, we have $D(p_{s_R} \| p) = s_R \psi'(s_R) - \psi(s_R) = s_R (R - \psi(s_R))/(1 - s_R) - \psi(s_R) = \frac{s_R R - \psi(s_R)}{1 - s_R}$.
(**j**) When $s = s_R$, $\frac{R + (s-1)\psi'(s) - \psi(s)}{(1-s)^2}. = 0$. Further, since $\frac{d}{ds}(R + (s-1)\psi'(s) - \psi(s)) = (s-1)\psi''(s) > 0$, $\frac{R + (s-1)\psi'(s) - \psi(s)}{(1-s)^2} > 0$ for $s > s_R$ and $\frac{R + (s-1)\psi'(s) - \psi(s)}{(1-s)^2} < 0$ for $s < s_R$. Hence, the maximum of $\frac{s R - \psi(s)}{1-s}$ can be realized with $s = s_R$.
(**k**) Combine (**g**), (**h**), and (**j**).

Exercise 2.25

(**a**) See (**e**) of Exercise 2.24.
(**b**) It follows from Exercise 2.23.
(**c**) See (**g**) of Exercise 2.24.
(**d**) See (**h**) of Exercise 2.24.
(**e**) See (**j**) of Exercise 2.24.
(**f**) Combine (**c**), (**d**), and (**e**).

Exercise 2.26 $\frac{d}{ds} \frac{-\psi(s)}{1-s} = \frac{(s-1)\psi'(s) - \psi(s)}{(1-s)^2} = \frac{-H(p_s)}{(1-s)^2} < 0$. Hence, the supremum is attained with $s \to -\infty$.

Exercise 2.27 Since $-\log \max_i p_i \le H_\alpha(p) = H_{\min}(p) \le H_\alpha(p) \le H_{\max}(p)$, it is enough to show $H_{\max}(p) \le -\log \min_i p_i$. This inequality is equivalent with $\min_i p_i \le \frac{1}{|\{i|p_i>0\}|}$.

Exercise 2.28 Equation (2.72) can be shown by a simple calculation. Equation (2.73) is shown by the following way.

$$\log |\mathcal{X}| - \min_{Q_Y} D(\mathrm{P}_{XY} \| p_{\mathrm{mix},\mathcal{X}} \times Q_Y) = H(X|Y) - \min_{Q_Y} D(\mathrm{P}_Y \| Q_Y). = H(X|Y).$$

Exercise 2.29 Due to (2.74), we have

$$\lim_{s\to 0} H_{1+s}(X|Y) = -\frac{d}{ds} \sum_y \mathrm{P}_Y(y) \sum_x \mathrm{P}_{X|Y=y}(x)^{1+s}|_{s=0} = H(X|Y).$$

Due to (2.74), we have

$$\lim_{s\to 0} H^{\uparrow}_{1+s}(X|Y) = \max_{Q_Y} -\frac{d}{ds} \sum_{x,y} \mathrm{P}_{X,Y}(x,y)^{1+s} Q_Y(y)^{-s}|_{s=0}$$

$$= \max_{Q_Y} -\sum_{x,y} \mathrm{P}_{X,Y}(x,y)(\log \mathrm{P}_{X,Y}(x,y) - \log Q_Y(y)) = H(X|Y).$$

Exercise 2.30 The second expression in (2.74) yields (2.83) and (2.85). (2.81) yields (2.84) and (2.86).

Exercise 2.31 The concavity of $s \mapsto sH_{1+s}(X|Y)$ can be shown from the convexity of $s \mapsto D_{1+s}(p\|q)$(Exercise 2.16). Since the function $s \mapsto D_{1+s}(\mathrm{P}_{XY} \| p_{\mathrm{mix},\mathcal{X}} \times Q_Y)$ is convex, the function $s \mapsto \min_{Q_Y} D_{1+s}(\mathrm{P}_{XY} \| p_{\mathrm{mix},\mathcal{X}} \times Q_Y)$ is also convex. Hence, the function $s \mapsto sH^{\uparrow}_{1+s}(X|Y)$ is concave. Similar to Exercise 2.17, we can show that the functions $s \mapsto H_{1+s}(X|Y)$ and $H^{\uparrow}_{1+s}(X|Y)$ are monotonicallly decreasing.

Exercise 2.32 Due to the equality condition of Hölder inequality, the equality in (2.88) holds if and only if there exists a function $c(y)$ such that $\mathrm{P}_{X|Y=y}(x) = c(y)\mathrm{P}_{XY}(x,y)$, which implies that $\mathrm{P}_{XY}(x,y)^{-s/(1-s)} = c(y)\mathrm{P}_Y(y)$. Hence, we obtain $\mathrm{P}_{XY}(x,y) = c(y)^{-(1-s)/s}\mathrm{P}_Y(y)^{-(1-s)/s}$. This condition is equivalent to $\mathrm{P}_{XY}(x,y) = \frac{1}{|\mathcal{X}|}\mathrm{P}_Y(y)$.

Exercise 2.33 We denote the marginal distributions of X and Y p_X and p_Y respectively. Then, $\mathrm{Cov}_p(X,Y) = \sum_{x,y} p(x,y)(X - \mathrm{E}_p X)(Y - \mathrm{E}_p Y) = \sum_{x,y} p_X(x) P_Y(y)(X - \mathrm{E}_p X)(Y - \mathrm{E}_p Y) = \sum_x p_X(x)(X - \mathrm{E}_p X) \sum_y P_Y(y)(Y - \mathrm{E}_p Y) = 0$.

Exercise 2.34 For $i \ne j$, we have $\sum_{\omega_1,\ldots,\omega_n} p_\theta(\omega_1)\cdots p_\theta(\omega_n) \frac{d \log p_\theta(\omega_i)}{d\theta} \frac{d \log p_\theta(\omega_j)}{d\theta} = 0$. Hence, $\sum_{\omega_1,\ldots,\omega_n} p^n_\theta(\omega_1,\ldots,x_n)(\frac{d \log p^n_\theta(\omega_1,\ldots,x_n)}{d\theta})^2$
$= \sum_{\omega_1,\ldots,\omega_n} p_\theta(\omega_1)\cdots p_\theta(\omega_n)(\frac{d \log p_\theta(\omega_1)+\cdots+\log p_\theta(\omega_n)}{d\theta})^2$
$= \sum_{\omega_1,\ldots,\omega_n} p_\theta(\omega_1)\cdots p_\theta(\omega_n)(\frac{d \log p_\theta(\omega_1)}{d\theta} + \cdots + \frac{d \log p_\theta(\omega_n)}{d\theta})^2$

$= \sum_{\omega_1,\ldots,\omega_n} p_\theta(\omega_1)\cdots p_\theta(\omega_n) \sum_{i=1}^n (\frac{d\log p_\theta(\omega_i)}{d\theta})^2 + \sum_{i\neq j} \frac{d\log p_\theta(\omega_i)}{d\theta}\frac{d\log p_\theta(\omega_j)}{d\theta}$
$= \sum_{\omega_1,\ldots,\omega_n} p_\theta(\omega_1)\cdots p_\theta(\omega_n) \sum_{i=1}^n (\frac{d\log p_\theta(\omega_i)}{d\theta})^2$
$= \sum_{i=1}^n \sum_{\omega_i} p_\theta(\omega_i)(\frac{d\log p_\theta(\omega_i)}{d\theta})^2 = nJ_\theta.$

Exercise 2.35 Use the approximation
$\sqrt{p_{\theta+\epsilon}(\omega)} \cong \sqrt{p_\theta(\omega)}\sqrt{1 + l_\theta(\omega)\epsilon + \frac{1}{2}\frac{d^2 p_\theta(\omega)}{d\theta^2}\epsilon^2}.$

Exercise 2.36

(**a**) It follows from the Taylor expansion of $p_{\theta+\epsilon}(\omega)$ for ϵ.
(**b**) Since $\frac{d^2\log p_\theta(\omega)}{d^2\theta} = \frac{d^2 p_\theta(\omega)}{d^2\theta}/p_\theta(\omega) - (\frac{dp_\theta(\omega)}{d\theta}/p_\theta(\omega))^2$, we have
$\sum_\omega p_\theta(\omega)\frac{d^2\log p_\theta(\omega)}{d^2\theta}\epsilon^2 = \sum_\omega p_\theta(\omega)(-(\frac{dp_\theta(\omega)}{d\theta}/p_\theta(\omega))^2 + \frac{d^2 p_\theta(\omega)}{d^2\theta}/p_\theta(\omega))$
$= -\sum_\omega p_\theta(\omega)(\frac{dp_\theta(\omega)}{d\theta}/p_\theta(\omega))^2 = -J_\theta$. Thus, $D(p_\theta\|p_{\theta+\epsilon})$
$= \sum_\omega p_\theta(\omega)(\log p_\theta(\omega) - \log p_{\theta+\epsilon}(\omega)) \cong -\sum_\omega p_\theta(\omega)(\frac{d\log p_\theta(\omega)}{d\theta}\epsilon +$
$\frac{1}{2}\frac{d^2\log p_\theta(\omega)}{d^2\theta}\epsilon^2)$
$= -\sum_\omega p_\theta(\omega)\frac{d\log p_\theta(\omega)}{d\theta}\epsilon - \frac{1}{2}\sum_\omega p_\theta(\omega)\frac{d^2\log p_\theta(\omega)}{d^2\theta}\epsilon^2$
$= -\frac{1}{2}\sum_\omega p_\theta(\omega)\frac{d^2\log p_\theta(\omega)}{d^2\theta}\epsilon^2 = \frac{1}{2}J_\theta\epsilon^2.$
(**d**) $D(p_{\theta+\epsilon}\|p_\theta) = \sum_\omega p_{\theta+\epsilon}(\omega)(\log p_{\theta+\epsilon}(\omega) - \log p_\theta(\omega))$
$\cong \sum_\omega (p_\theta(\omega) + \frac{dp_\theta(\omega)}{d\theta}\epsilon + \frac{1}{2}\frac{d^2 p_\theta(\omega)}{d^2\theta}\epsilon^2)(\frac{d\log p_\theta(\omega)}{d\theta}\epsilon + \frac{1}{2}\frac{d^2\log p_\theta(\omega)}{d^2\theta}\epsilon^2)$
$\cong \sum_\omega p_\theta(\omega)(\frac{d\log p_\theta(\omega)}{d\theta}\epsilon + \frac{1}{2}\frac{d^2\log p_\theta(\omega)}{d^2\theta}\epsilon^2) + \sum_\omega \frac{dp_\theta(\omega)}{d\theta}\epsilon\frac{d\log p_\theta(\omega)}{d\theta}\epsilon$
$= \sum_\omega p_\theta(\omega)\frac{1}{2}\frac{d^2\log p_\theta(\omega)}{d^2\theta}\epsilon^2) + \sum_\omega \frac{dp_\theta(\omega)}{d\theta}\frac{d\log p_\theta(\omega)}{d\theta}\epsilon^2 = J_\theta\epsilon^2 - \frac{1}{2}J_\theta\epsilon^2 = \frac{1}{2}J_\theta\epsilon^2.$

Exercise 2.37 $e^{\phi(s|p_\theta\|p_{\theta+\epsilon})} = \sum_x p_\theta(\omega)^{1-s}p_{\theta+\epsilon}(\omega)^s$
$\cong \sum_x p_\theta(\omega)^{1-s}(p_\theta(\omega) + \frac{dp_\theta(\omega)}{d\theta}\epsilon + \frac{1}{2}\frac{d^2 p_\theta(\omega)}{d^2\theta}\epsilon^2)^s$
$= \sum_x p_\theta(\omega)^{1-s}p_\theta(\omega)^s(1 + \frac{dp_\theta(\omega)}{d\theta}p_\theta(\omega)^{-1}\epsilon + \frac{1}{2}\frac{d^2 p_\theta(\omega)}{d^2\theta}p_\theta(\omega)^{-1}\epsilon^2)^s$
$\cong \sum_x p_\theta(\omega)(1 + s(\frac{dp_\theta(\omega)}{d\theta}p_\theta(\omega)^{-1}\epsilon + \frac{1}{2}\frac{d^2 p_\theta(\omega)}{d^2\theta}p_\theta(\omega)^{-1}\epsilon^2) + \frac{s(s-1)}{2}(\frac{dp_\theta(\omega)}{d\theta}$
$p_\theta(\omega)^{-1}\epsilon)^2)$
$= 1 + \sum_x p_\theta(\omega)s(\frac{dp_\theta(\omega)}{d\theta}p_\theta(\omega)^{-1}\epsilon + \sum_x p_\theta(\omega)\frac{1}{2}\frac{d^2 p_\theta(\omega)}{d^2\theta}p_\theta(\omega)^{-1}\epsilon^2$
$+ \frac{s(s-1)}{2}\sum_x p_\theta(\omega)(\frac{dp_\theta(\omega)}{d\theta})^2 p_\theta(\omega)^{-2}\epsilon^2$
$= 1 + \frac{s(s-1)}{2}\sum_x p_\theta(\omega)^{-1}(\frac{dp_\theta(\omega)}{d\theta})^2\epsilon^2 = 1 + \frac{s(s-1)}{2}\epsilon^2 J_\theta$. Thus, $\phi(s|p_\theta\|p_{\theta+\epsilon}) \cong \log$
$(1 + \frac{s(s-1)}{2}\epsilon^2 J_\theta) \cong \frac{s(s-1)}{2}\epsilon^2 J_\theta.$

Exercise 2.38 For arbitrary η and η', and a real number $\lambda \in (0,1)$, we choose $\tilde{\theta}_0$ such that $\max_{\tilde{\theta}} \sum_k (\lambda\eta_k + (1-\lambda)\eta'_k)\tilde{\theta}^k - \mu(\tilde{\theta}) = \sum_k (\lambda\eta_k + (1-\lambda)\eta'_k)\tilde{\theta}_0^k - \mu(\tilde{\theta})$. Hence,

$$\nu(\lambda\eta + (1-\lambda)\eta') = \sum_k (\lambda\eta_k + (1-\lambda)\eta'_k)\tilde{\theta}_0^k - \mu(\tilde{\theta})$$
$$= \lambda \sum_k \eta_k \tilde{\theta}_0^k - \mu(\tilde{\theta}) + (1-\lambda) \sum_k \eta_k \tilde{\theta}_0^k - \mu(\tilde{\theta})$$
$$\le \lambda \max_{\tilde{\theta}} \sum_k \eta_k \tilde{\theta}^k - \mu(\tilde{\theta}) + (1-\lambda) \max_{\tilde{\theta}} \sum_k \eta_k \tilde{\theta}^k - \mu(\tilde{\theta}) = \lambda\nu(\eta) + (1-\lambda)\nu(\eta').$$

Exercise 2.39 Choose the generator $-\log p(x)$. Then, the set $\{p_s(x)\}$ is an exponential family generated by $-\log p(x)$. The set $\{q|H(q) = H(p_s)\}$ is a mixture family generated by $-\log p(x)$. So, Theorem 2.3 directly solves Exercise 2.23.

Exercise 2.40

(**a**) Since $\eta(\theta) = \sum_\omega p_\theta(\omega) X(\omega)$, X is an unbiased estimator.
(**b**) Since $\frac{d}{d\eta} \log p_\theta(\omega) = \frac{d\theta}{d\eta}\frac{d}{d\theta} \log p_\theta(\omega) = (\frac{d\eta}{d\theta})^{-1}\frac{d}{d\theta} \log p_\theta(\omega) = (J_\theta)^{-1}\frac{d}{d\theta} \log p_\theta(\omega)$, the Fisher information for η is $J_\theta (J_\theta)^{-2} = J_\theta^{-1}$. Then, the lower bound of the variance of unbiased estimator given by Crámer-Rao inequality is J_θ. The variance of X is also J_θ.
(**c**) Use $\frac{d\theta}{d\eta} = J_\eta$.
(**d**) Since $\frac{d\mu}{d\theta} = \eta$, we have $\frac{d\mu}{d\eta} = \frac{d\theta}{d\eta}\frac{d\mu}{d\theta} = J_\eta \eta$. Taking the integral, we obtain the desired equation.
(**e**) Inequality (2.140) is derived by Schwartz inequality. Since $|\langle X - \eta, l_\eta \rangle_{p_\eta}| = 1$, the equality condition is $\frac{l_\eta}{J_\eta} = X - \eta$.
(**f**) Replace l_η by $\frac{p_\eta}{d\eta}/p_\eta$. We obtain $\frac{dp_\eta}{d\eta} = J_\eta(X-\eta)p_\eta$.
(**g**) Define $\theta := \int_0^\eta J_{\eta'}\, d\eta'$, and $\mu(\theta(\eta)) := \int_0^\eta \eta' J_{\eta'}\, d\eta'$.
$\frac{d\mu(\theta(\eta))}{d\theta} = \frac{d\mu(\theta(\eta))}{d\eta}\frac{d\eta}{d\theta} = \eta J_\eta (\frac{d\theta}{d\eta})^{-1} = \eta J_\eta J_\eta^{-1} = \eta$. The function $\log \sum_\omega p_\eta(\omega) e^{\theta X(\omega)}$ also satisfies the same differential equation. Due to the uniqueness of the solution of the differential equation, we have $\mu(\theta(\eta)) = \log \sum_\omega p_\eta(\omega) e^{\theta X(\omega)}$.
Since $\frac{d \log p_\eta}{d\eta} = \frac{dp_\eta}{d\eta}/p_\eta = J_\eta(X-\eta) = J_\eta X - \eta J_\eta$, we have $\log p_\eta = \theta X - \mu(\theta(\eta))$. Hence, we have $p_\eta = e^{\theta X - \mu(\theta)}$.

Exercise 2.41 Show that $\frac{p_\theta(\omega)}{d\theta} = 0$ if and only if $\eta(\theta) = X(\omega)$.

Exercise 2.42 Combine (2.13) and (2.105).

Exercise 2.43 The case of $n \ge m$ can be obtained from $n, n-1, \ldots, m+1 \ge m$. The $n < m$ case may be obtained from $\frac{1}{m}, \frac{1}{m-1}, \ldots, \frac{1}{n+1} \le \frac{1}{n}$.

Exercise 2.44 $\mathrm{E}_p X = \sum_i p_i x_i \ge \sum_{i: x_i \ge c} p_i x_i \ge c \sum_{i: x_i \ge c} p_i$.

Exercise 2.45 Apply Cramér's theorem to the random variable $\log p_i$.

Exercise 2.46 Equation (2.189) implies that

$$\begin{aligned}
&\lim_{n\to\infty} -\frac{1}{n}\log P^c(p^n, e^{nR}) \\
&\le \lim_{n\to\infty} -\frac{1}{n}\log \max_{q\in T_n:\frac{e^{nH(q)}}{(n+1)^d}>e^{nR}} e^{-nD(p\|q)}\left(1-\frac{(n+1)^d e^{nR}}{e^{nH(q)}}\right) \\
&\le \min_{q:H(q)\ge R} D(p\|q).
\end{aligned}$$

Combing (2.55), we obtain the $\le$ part of (2.188).

Exercise 2.47

$$\begin{aligned}
\lim_{n\to\infty} -\frac{1}{n}\log P(p^n, e^{nR}) &\le \lim_{n\to\infty} -\frac{1}{n}\log \max_{q\in T_n:H(q)\le R} \frac{e^{-nD(q\|p)}}{(n+1)^d} \\
&\le \min_{q:H(q)\le R} D(p\|q).
\end{aligned}$$

Combing (2.65), we obtain the $\le$ part of (2.188).

Exercise 2.48 Since $p_n(0)e^{n\theta\cdot a} + p_n(1)e^{n\theta\cdot -b} = e^{-na}e^{n\theta a} + (1-e^{-na})e^{-n\theta b}$, we have $\mu(\theta) = \lim_{n\to\infty}\frac{1}{n}\log(e^{-na}e^{n\theta a} + (1-e^{-na})e^{-n\theta b}) = -\theta b$ for $\theta < \frac{a}{a+b}$ and $\mu(\theta) = -a(1-\theta)$ for $\theta \ge \frac{a}{a+b}$. Hence, we obtain $\mu(\theta) = -\min\{(1-\theta)a, \theta b\}$. Since $-b < x < a$, we have $\max_{\theta>0}(x\theta - \mu(\theta)) = \max(\max_{\frac{a}{a+b}>\theta>0}(x\theta + \theta b), \max_{\theta\ge\frac{a}{a+b}}(x\theta + a(1-\theta))) = \max((x+b)\frac{a}{a+b}, a + (x-a)\frac{a}{a+b}) = \max(\frac{a(x+b)}{a+b}, \frac{a(x+b)}{a+b}) = \frac{a(x+b)}{a+b} < a$.

On the other hand, since $a > x > -b$, $\lim_{n\to\infty}\frac{1}{n}\log p_n\{X_n \ge x\} = \lim_{n\to\infty}\frac{1}{n}\log p_n(0) = \lim_{n\to\infty}\frac{1}{n}\log e^{-na} = a$.

Exercise 2.49

(**b**) Since $e^{-H_{\min}(p_{d,\epsilon})} = \frac{1}{d} + \epsilon$, we have $H_{\min}(p_{d,\epsilon}) = \log d - \log(1+d\epsilon)$.
(**c**) Since $H_{\min}(p_{\mathrm{mix},d}) - H_{\min}(p_{d,\epsilon}) = \log(1+d\epsilon) = (\log d + \log\epsilon) + O(\frac{1}{d\epsilon})$, we have $\frac{H_{\min}(p_{\mathrm{mix},d}) - H_{\min}(p_{d,\epsilon})}{\log d} = 1 + \frac{\log\epsilon}{\log d} + O(\frac{1}{d\epsilon\log d})$.
(**d**) Since $e^{-sH_{1+s}(p_{d,\epsilon})} = (\frac{1}{d}+\epsilon)^{1+s} + (d-1)(\frac{1}{d} - \frac{\epsilon}{d-1})^{1+s}$, we have $H_{1+s}(p_{d,\epsilon})$ $= -\frac{1}{s}\log((\frac{1}{d}+\epsilon)^{1+s} + (d-1)(\frac{1}{d} - \frac{\epsilon}{d-1})^{1+s}) = \log d - \frac{1}{s}\log(\frac{1}{d}(1+d\epsilon)^{1+s}$ $+ \frac{d-1}{d}(1 - \frac{d\epsilon}{d-1})^{1+s})$.
(**e**) Since $H_{1+s}(p_{\mathrm{mix},d}) - H_{1+s}(p_{d,\epsilon}) = \frac{1}{s}\log(\frac{1}{d}(1+d\epsilon)^{1+s} + \frac{d-1}{d}(1 - \frac{d\epsilon}{d-1})^{1+s}) +$ $O(d(d\epsilon)^{-(1+s)}) = \frac{1}{s}\log(\frac{1}{d}(1+d\epsilon)^{1+s}) + O(d(d\epsilon)^{-(1+s)}) = -\frac{\log d}{s} + \frac{1+s}{s}\log(d\epsilon) +$ $O(d(d\epsilon)^{-(1+s)}) = \log d + \frac{1+s}{s}\log\epsilon + O(d(d\epsilon)^{-(1+s)})$, we have $\frac{H_{1+s}(p_{\mathrm{mix},d}) - H_{1+s}(p_{d,\epsilon})}{\log d}$ $= 1 + \frac{(1+s)\log\epsilon}{s\log d} + O((d\epsilon)^{-(1+s)}\frac{d}{\log d})$ as $d\to\infty$.

Exercise 2.50

(**b**) Since the cardinality of $p'_{d,\epsilon}$ is d, we have $H_{\max}(p'_{d,\epsilon}) = \log d$. Thus, $\frac{H_{\max}(p'_{\mathrm{mix},d}) - H_{\max}(p'_{d,0})}{\log d} = 1$ for $\epsilon > 0$.

(**c**) Since $e^{sH_{1-s}(p'_{d,\epsilon})} = (1-\epsilon)^{1-s} + (d-1)(\frac{\epsilon}{d-1})^{1-s} = 1-(1-s)\epsilon + (d-1)^s \epsilon^{1-s} + O(\epsilon^2) = 1-(d-1)^s\epsilon^{1-s} + O(\epsilon)$, we have $H_{1-s}(p'_{d,\epsilon}) = -\frac{1}{s}\log((1-\epsilon)^{1-s} + (d-1)(\frac{\epsilon}{d-1})^{1-s})$.
(**d**) Since $(1-\epsilon)^{1-s} + (d-1)(\frac{\epsilon}{d-1})^{1-s} = 1-(1-s)\epsilon + (d-1)^s\epsilon^{1-s} + O(\epsilon^2) = 1-(d-1)^s\epsilon^{1-s} + O(\epsilon)$, we have $H_{1-s}(p'_{d,\epsilon}) = -\frac{1}{s}\log(1-(d-1)^s\epsilon^{1-s} + O(\epsilon)) = \frac{1-s}{s}(d-1)^s\epsilon^{1-s} + O(\epsilon) + O(\epsilon^{2(1-s)})$. Thus, $\frac{H_{1+s}(p'_{d,\epsilon}) - H_{1+s}(p'_{d,0})}{\log d} = \frac{1-s}{s\log d}(d-1)^s\epsilon^{1-s} + O(\frac{\epsilon}{\log d}) + O(\frac{\epsilon^{2(1-s)}}{\log d})$ as $\epsilon \to 0$.

Exercise 2.51 We have

$$\begin{aligned}
&|e^{-H_2(p)} - e^{-H_2(q)}| = |e^{-H_2(p)} - 2c + dc^2 - e^{-H_2(q)} + 2c - dc^2| \\
=&|\sum_i (p_i - c)^2 - (q_i - c)^2| = |\sum_i (p_i - q_i)(p_i + q_i - 2c)| \\
\le& \sum_i |p_i - q_i||p_i + q_i - 2c| \le \left(\sum_i |p_i - q_i|\right) \max_i |p_i + q_i - 2c| \\
=&2d_1(p,q) \max_i |p_i + q_i - 2c|.
\end{aligned}$$

Since $\min_c \max_i |p_i + q_i - 2c| \le 1$, we obtain $|e^{-H_2(p)} - e^{-H_2(q)}| \le 2d_1(p,q)$, which implies (2.209).

Exercise 2.52 It is enough to show that $\||x\rangle\langle x| - |y\rangle\langle y|\|_1 = 2\sin\epsilon$ when $|\langle x|y\rangle| = \cos\epsilon$. When the state $|x\rangle\langle x|$ is written as $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, the other state $|y\rangle\langle y|$ is written as $\begin{pmatrix} cos^2\theta & \cos\theta\sin\theta \\ \cos\theta\sin\theta & \sin^2\theta \end{pmatrix}$. Hence,

$$|x\rangle\langle x| - |y\rangle\langle y| = \begin{pmatrix} cos^2\theta - 1 & \cos\theta\sin\theta \\ \cos\theta\sin\theta & \sin^2\theta \end{pmatrix}.$$

Solving the characteristic equation, we obtain the eigenvalues $\pm\sin\epsilon$. Thus, we have $\||x\rangle\langle x| - |y\rangle\langle y|\|_1 = 2\sin\epsilon$.

Exercise 2.53 It is enough to show the same case as Exercise 2.52. Since the eigenvalues of $|x\rangle\langle x| - |y\rangle\langle y|$ are $\pm\sin\epsilon$, we have $\||x\rangle\langle x| - |y\rangle\langle y|\|_2 = \sqrt{2\sin^2\epsilon} = \sqrt{2}\sin\epsilon$.

Exercise 2.54 It is enough to show the same case as Exercise 2.52. Choose ϵ as $d(x,y) = \epsilon$. Then, $|x\rangle - |y\rangle = \begin{pmatrix} 1-\cos\epsilon \\ \sin\epsilon \end{pmatrix}$. Thus $\||x\rangle - |y\rangle\|^2 = (1-\cos\epsilon)^2 + \sin^2\epsilon = 2(1-\cos\epsilon) = 4\sin^2\frac{\epsilon}{2}$. The second inequality follows from $\sin\frac{\epsilon}{2} \le \frac{\epsilon}{2}$.

Exercise 2.55 Use the relation $\int_0^\infty x^2 e^{-\frac{x^2}{2}} dx = \sqrt{\frac{\pi}{2}}$.

Exercise 2.56 Since $u, \epsilon \geq 0$, we have $(u + \epsilon\sqrt{l-1})^2 \geq u^2 + (\epsilon\sqrt{l-1})^2$. Thus

$$\begin{aligned}
&\frac{\int_{\epsilon\sqrt{l-1}}^{\frac{\pi}{2}\sqrt{l-1}} e^{-u^2} du}{\frac{\pi}{\sqrt{2}}} = \frac{\int_{\epsilon\sqrt{l-1}}^{\frac{\pi}{2}\sqrt{l-1}} e^{-u^2} du}{\frac{\pi}{\sqrt{2}}} = \frac{\int_0^{\frac{\pi}{2}\sqrt{l-1}-\epsilon\sqrt{l-1}} e^{-(u+\epsilon\sqrt{l-1})^2} du}{\frac{\pi}{\sqrt{2}}} \\
\leq & e^{-\epsilon^2(l-1)} \frac{\int_0^{\frac{\pi}{2}\sqrt{l-1}-\epsilon\sqrt{l-1}} e^{-u^2} du}{\frac{\pi}{\sqrt{2}}} \leq e^{-\epsilon^2(l-1)} \frac{\int_0^{\infty} e^{-u^2} du}{\frac{\pi}{\sqrt{2}}} \\
= & e^{-\epsilon^2(l-1)} \frac{\sqrt{2\pi}/4}{\frac{\pi}{\sqrt{2}}} = e^{-\epsilon^2(l-1)}/2
\end{aligned}$$

Exercise 2.57

(**a**) Use $B(k - \frac{1}{2}, \frac{1}{2}) = \frac{k-1-\frac{1}{2}}{k-1} B(k - 1 - \frac{1}{2}, \frac{1}{2})$ and $B(\frac{1}{2}, \frac{1}{2}) = \pi$.
(**b**) Since $\log(1 + x)$ is concave, we have $\log(1 + \frac{x}{2}) \geq \frac{1}{2}\log(1 + x)$. Thus $\sum_{k=1}^{l-1} \log \frac{2k}{2k-1} = \sum_{k=1}^{l-1} \log(1 + \frac{1}{2k-1}) \geq \frac{1}{2}\sum_{k=1}^{l-1} \log(1 + \frac{1}{k-1/2}) = \frac{1}{2}\sum_{k=1}^{l-1} \log \frac{k+1/2}{k-1/2} = \frac{1}{2} \log \frac{l-1/2}{1/2} = \frac{1}{2}\log(2l - 1)$.
(**c**) Due to (**b**), we have $\sum_{k=1}^{l-1} \log \frac{2k-1}{2k} \leq -\frac{1}{2}\log(2l - 1)$. Thus $(2l - 1)B(l - \frac{1}{2}, \frac{1}{2}) \leq (2l - 1)\pi(2l - 1)^{-1/2} = \sqrt{(2l - 1)}\pi$.

References

1. I. Csiszár, Information type measures of difference of probability distribution and indirect observations. Studia Scient. Math. Hungar. **2**, 299–318 (1967)
2. S. Amari, H. Nagaoka, *Methods of Information Geometry* (AMS & Oxford University Press, Oxford, 2000)
3. A. Rényi, On measures of information and entropy, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (University of California Press, Berkeley, 1961), pp. 547–561
4. R.M. Fano, *Transmission of Information: A Statistical Theory of Communication* (Wiley, New York, 1961)
5. M. Hayashi, Security analysis of ε-almost dual universal$_2$ hash functions: smoothing of min entropy vs. smoothing of Rényi entropy of order 2 (2013). arXiv:1309.1596
6. S. Amari, α-divergence Is unique, belonging to both f-divergence and Bregman divergence classes. IEEE Trans. Inform. Theory **55**(11), 4925–4931 (2009)
7. A.W. van der Vaart, *Asymptotic Statistics* (Cambridge University Press, Cambridge, 1998)
8. I. Csiszár, J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems* (Academic, 1981)
9. I.N. Sanov, On the probability of large deviations of random variables. Mat. Sbornik **42**, 11–44 (1957) (in Russian). English translation: Selected Translat. Math. Stat. **1**, 213–244 (1961)
10. M. Keyl, R.F. Werner, Estimating the spectrum of a density operator. Phys. Rev. A **64**, 052311 (2001)
11. K. Matsumoto, Seminar notes (1999)
12. M. Hayashi, Optimal sequence of POVMs in the sense of Stein's lemma in quantum hypothesis. J. Phys. A Math. Gen. **35**, 10759–10773 (2002)

13. M. Hayashi, Exponents of quantum fixed-length pure state source coding. Phys. Rev. A **66**, 032321 (2002)
14. M. Hayashi, K. Matsumoto, Variable length universal entanglement concentration by local operations and its application to teleportation and dense coding, quant-ph/0109028 (2001); K. Matsumoto, M. Hayashi, Universal entanglement concentration. Phys. Rev. A **75**, 062338 (2007)
15. M. Hayashi, K. Matsumoto, Quantum universal variable-length source coding. Phys. Rev. A **66**, 022311 (2002)
16. M. Hayashi, K. Matsumoto, Simple construction of quantum universal variable-length source coding. Quant. Inf. Comput. **2**, Special Issue, 519–529 (2002)
17. M. Hayashi, Asymptotics of quantum relative entropy from a representation theoretical viewpoint. J. Phys. A Math. Gen. **34**, 3413–3419 (2001)
18. H. Cramér, Sur un nouveaux theoorème-limite de la théorie des probabilités, in *Actualités Scientifiques et Industrielles*, no. 736, in *Colloque consacré à la thèorie des probabilités* (Hermann, Paris, 1938), pp. 5–23
19. J. Gärtner, On large deviations from the invariant measure. Theory Prob. Appl. **22**, 24–39 (1977)
20. R. Ellis, Large deviations for a general class of random vectors, Ann. Probab. **12**, 1, 1–12 (1984); *Entropy, Large Deviations and Statistical Mechanics* (Springer, Berlin, 1985)
21. R.R. Bahadur, On the asymptotic efficiency of tests and estimates. Sankhyā **22**, 229 (1960)
22. R.R. Bahadur, Rates of Convergence of Estimates and Test Statistics. Ann. Math. Stat. **38**, 303 (1967)
23. R.R. Bahadur, Some limit theorems in statistics, in *Regional Conference Series in Applied Mathematics*, no. 4 (SIAM, Philadelphia, 1971)
24. J.C. Fu, On a theorem of Bahadur on the rate of convergence of point estimators. Ann. Stat. **1**, 745 (1973)
25. A.I. Khinchin, *Mathematical Foundations of Information Theory* (Dover, New York, 1957)
26. T.S. Han, *Information-Spectrum Methods in Information Theory* (Springer, Berlin, 2002) (originally appeared in Japanese in 1998)
27. V.D. Milman, G. Schechtman, *Asymptotic theory of finite-dimensional normed spaces*, vol. 1200, Lecture Notes in Mathematics (Springer, Berlin, 1986)
28. T. Cover, J. Thomas, *Elements of Information Theory* (Wiley, New York, 1991)
29. M. Hayashi, Exponential decreasing rate of leaked information in universal random privacy amplification. IEEE Trans. Inf. Theory **57**, 3989–4001 (2011)
30. M. Iwamoto, J. Shikata, Information theoretic security for encryption based on conditional Rényi entropies. Inform. Theor. Secur. Lect. Notes Comput. Sci. **8317**(2014), 103–121 (2014)
31. M. Müller-Lennert, F. Dupuis, O. Szehr, S. Fehr, M. Tomamichel, On quantum Renyi entropies: a new generalization and some properties. J. Math. Phys. **54**, 122203 (2013)
32. E.L. Lehman, G. Casella, *Theory of Point Estimation* (Springer, Berlin Heidelberg New York, 1998)
33. A. Dembo, O. Zeitouni, *Large Deviation Techniques and Applications* (Springer, Berlin, 1997)
34. J.A. Bucklew, *Large Deviation Techniques in Decision, Simulation, and Estimation* (Wiley, New York, 1990)

http://www.springer.com/978-3-662-49723-4

Quantum Information Theory
Mathematical Foundation
Hayashi, M.
2017, XLIII, 636 p. 24 illus., 1 illus. in color., Hardcover
ISBN: 978-3-662-49723-4