4 Hypothesis Testing

4.1 Hypothesis Testing

For two given general sources $\mathbf{X} = \{X^n\}_{n=1}^{\infty}$ and $\overline{\mathbf{X}} = \{\overline{X}^n\}_{n=1}^{\infty}$ we consider the *hypothesis testing problem* with the *null hypothesis* $\mathbf{X}$ and the *alternative hypothesis* $\overline{\mathbf{X}}$. This problem is also called the hypothesis testing $\mathbf{X}$ against $\overline{\mathbf{X}}$ for simplicity. Here, both X^n and $\overline{X}^n$ are supposed to be $\mathcal{X}^n$-valued random variables, where $\mathcal{X}$ denotes a source alphabet. In ordinary hypothesis testing problems we choose a subset $\mathcal{A}_n \subset \mathcal{X}^n$ as an *acceptance region.* If $\mathbf{x}$, an output from one of the two sources, belongs to $\mathcal{A}_n$, then we judge that the null hypothesis $\mathbf{X} = \{X^n\}_{n=1}^{\infty}$ is true. Otherwise, we judge that the alternative hypothesis $\overline{\mathbf{X}} = \{\overline{X}^n\}_{n=1}^{\infty}$ is true. We define the *error probability of the first kind* and the *error probability of the second kind* by

$$\mu_n \equiv \Pr\left\{X^n \notin \mathcal{A}_n\right\},$$

and

$$\lambda_n \equiv \Pr\{\overline{X}^n \in \mathcal{A}_n\},$$

respectively. From the definitions above, μ_n is the probability that we misjudge the alternative hypothesis $\overline{\mathbf{X}}$ true when the null hypothesis $\mathbf{X}$ is actually true and λ_n the probability that we misjudge $\mathbf{X}$ true when $\overline{\mathbf{X}}$ is actually true. The complement $\mathcal{C}_n \equiv \mathcal{X}^n - \mathcal{A}_n$ is called the *critical region* of the hypothesis testing. Throughout this chapter suppose that alphabet $\mathcal{X}$ is arbitrary ($\mathcal{X}$ can be countably infinite or abstract) unless stated otherwise.

The hypothesis testing is formulated as the problem of choosing an acceptance region $\mathcal{A}_n$ that makes the error probability of the second kind λ_n as small as possible subject to the constraint that the error probability of the first kind μ_n is upper bounded by a given constant. The following two simple, but powerful, lemmas are useful in order to obtain fundamental results on the hypothesis testing:

Lemma 4.1.1. *Define*

$$\mathcal{A}_n = \left\{\mathbf{x} \in \mathcal{X}^n \mid \frac{1}{n}\log\frac{P_{X^n}(\mathbf{x})}{P_{\overline{X}^n}(\mathbf{x})} \geq t\right\}$$

for an arbitrary real number * t*. Then, it holds that* $\Pr\{\overline{X}^n \in \mathcal{A}_n\} \le e^{-nt}$.

Proof. Since it follows that

$$\begin{aligned}1 \ge \Pr\{X^n \in \mathcal{A}_n\} &= \sum_{\mathbf{x}\in\mathcal{A}_n} P_{X^n}(\mathbf{x}) \\ &\ge \sum_{\mathbf{x}\in\mathcal{A}_n} P_{\overline{X}^n}(\mathbf{x})e^{nt} \\ &= \Pr\{\overline{X}^n \in \mathcal{A}_n\}e^{nt},\end{aligned}$$

we have $\Pr\{\overline{X}^n \in \mathcal{A}_n\} \le e^{-nt}$. □

Lemma 4.1.2. *For any real number* t *and* $\mathcal{A}_n \subset \mathcal{X}^n$*, it holds that*

$$\Pr\{X^n \notin \mathcal{A}_n\} + e^{nt}\Pr\{\overline{X}^n \in \mathcal{A}_n\} \ge \Pr\left\{\frac{1}{n}\log\frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)} \le t\right\}.$$

Remark 4.1.1. Lemma 4.1.2 can also be established as a consequence of Neyman-Pearson lemma [74]. However, there is no essential difference between the claims of Lemma 4.1.2 and Neyman-Pearson lemma if we consider the asymptotic situation with n being sufficiently large. □

Proof of Lemma 4.1.2.
Set

$$S_n = \left\{\mathbf{x}\in\mathcal{X}^n \mid \frac{1}{n}\log\frac{P_{X^n}(\mathbf{x})}{P_{\overline{X}^n}(\mathbf{x})} \le t\right\}.$$

Then, it follows that

$$\begin{aligned}&\Pr\left\{\frac{1}{n}\log\frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)} \le t\right\} = \Pr\{X^n \in S_n\} \\ &= \Pr\{X^n \in S_n \cap \mathcal{A}_n^c\} + \Pr\{X^n \in S_n \cap \mathcal{A}_n\} \\ &\le \Pr\{X^n \notin \mathcal{A}_n\} + \Pr\{X^n \in S_n \cap \mathcal{A}_n\}.\end{aligned}$$

By noticing that $\mathbf{x}\in S_n$ implies $P_{X^n}(\mathbf{x}) \le P_{\overline{X}^n}(\mathbf{x})e^{nt}$, we have

$$\begin{aligned}\Pr\{X^n \in S_n \cap \mathcal{A}_n\} &= \sum_{\mathbf{x}\in S_n\cap\mathcal{A}_n} P_{X^n}(\mathbf{x}) \\ &\le \sum_{\mathbf{x}\in S_n\cap\mathcal{A}_n} P_{\overline{X}^n}(\mathbf{x})e^{nt}\end{aligned}$$

* In the case where the source alphabet $\mathcal{X}$ is *abstract* in general, it is understood that $g_n(\mathbf{x}) \equiv \frac{P_{X^n}(\mathbf{x})}{P_{\overline{X}^n}(\mathbf{x})}$ $(\mathbf{x}\in\mathcal{X}^n)$ denotes the Radon-Nikodym derivative between probability measures on $\mathcal{X}^n$ with values on a singular set assumed conventionally to be $+\infty$. Then, $\frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)}$ is defined to be $g_n(X^n)$, which is obviously a random variable.

$$\leq \sum_{\mathbf{x} \in \mathcal{A}_n} P_{\overline{X}^n}(\mathbf{x}) e^{nt}$$
$$= e^{nt} \Pr\{\overline{X}^n \in \mathcal{A}_n\},$$

which completes the proof of the lemma. □

Now, we give definitions required for formulation of the hypothesis testing with the null hypothesis $\mathbf{X} = \{X^n\}_{n=1}^{\infty}$ and the alternative hypothesis $\overline{\mathbf{X}} = \{\overline{X}^n\}_{n=1}^{\infty}$. In order to define the hypothesis testing problems we need to fix a constraint that μ_n, the error probability of the first kind, must satisfy. We first consider the constraint that μ_n satisfies $\mu_n \to 0$ as $n \to \infty$. Since subject to this constraint the error probability of the second kind λ_n can usually be written as

$$\lambda_n \simeq e^{-nR} \quad (R > 0),$$

i.e., λ_n goes to zero of exponential order of block length n, it is fundamental to consider how we can make the exponent R large. The following definitions formulate such a situation.

Definition 4.1.1.

Rate R is achievable $\overset{\text{def}}{\Longleftrightarrow}$ There exists an acceptance region $\mathcal{A}_n$ satisfying

$$\lim_{n\to\infty} \mu_n = 0 \text{ and } \liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\lambda_n} \geq R.$$

Definition 4.1.2 (Supremum achievable error probability exponent).

$$B(\mathbf{X}||\overline{\mathbf{X}}) = \sup \{R \mid R \text{ is achievable}\}.$$

In the hypothesis testing problems the random variable $\frac{1}{n} \log \frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)}$ plays a crucial role. We call this the *divergence density rate* or the *likelihood-ratio density rate* and its probability distribution the *divergence-spectrum* (or, more generally, the *information-spectrum*).

Here, we define:

Definition 4.1.3.

$$\underline{D}(\mathbf{X}||\overline{\mathbf{X}}) = \text{p-}\liminf_{n\to\infty} \frac{1}{n} \log \frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)}$$

and call $\underline{D}(\mathbf{X}||\overline{\mathbf{X}})$ the *spectral inf-divergence rate* of $\mathbf{X}$ with respect to $\overline{\mathbf{X}}$.

Then, the spectral inf-divergence rate turns out to be nonnegative from its definition and Lemma 3.2.1, using X^n and $\overline{X}^n$ instead of U_n and V_n, respectively. That is, it holds that

$$\underline{D}(\mathbf{X}||\overline{\mathbf{X}}) \geq 0.$$

We have the following fundamental theorem on $B(\mathbf{X}||\overline{\mathbf{X}})$.

Theorem 4.1.1 (Verdú [90]).

$$B(\mathbf{X}||\overline{\mathbf{X}}) = \underline{D}(\mathbf{X}||\overline{\mathbf{X}}).$$

Proof.
1) *Direct part:*
Define $R = \underline{D}(\mathbf{X}||\overline{\mathbf{X}}) - \gamma$ for an arbitrary $\gamma > 0$ and consider the hypothesis testing with the acceptance region

$$\mathcal{A}_n = \left\{ \mathbf{x} \in \mathcal{X}^n \mid \frac{1}{n} \log \frac{P_{X^n}(\mathbf{x})}{P_{\overline{X}^n}(\mathbf{x})} \geq R \right\}.$$

Then, the definition of $\underline{D}(\mathbf{X}||\overline{\mathbf{X}})$ tells us that

$$\mu_n = \Pr\{X^n \notin \mathcal{A}_n\} \to 0 \quad \text{as } n \to \infty.$$

On the other hand, Lemma 4.1.1 with $t = R$ implies that

$$\Pr\{\overline{X}^n \in \mathcal{A}_n\} \leq e^{-nR},$$

i.e., $\lambda_n \leq e^{-nR}$. Accordingly, we obtain

$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\lambda_n} \geq R.$$

This establishes that $R = \underline{D}(\mathbf{X}||\overline{\mathbf{X}}) - \gamma$ is achievable for any $\gamma > 0$, which means $B(\mathbf{X}||\overline{\mathbf{X}}) \geq \underline{D}(\mathbf{X}||\overline{\mathbf{X}})$.

2) *Converse part:*
Suppose that R is achievable. Then, there exists an acceptance region $\mathcal{A}_n$ satisfying $\lim_{n\to\infty} \mu_n = 0$ and $\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\lambda_n} \geq R$. Hence, for any $\gamma > 0$ it follows that

$$\frac{1}{n} \log \frac{1}{\lambda_n} \geq R - \gamma \quad (\forall n \geq n_0),$$

which leads to

$$\lambda_n \leq e^{-n(R-\gamma)} \quad (\forall n \geq n_0).$$

On the other hand, Lemma 4.1.2 with $t = R - 2\gamma$ implies that

$$\mu_n + e^{n(R-2\gamma)}\lambda_n \geq \Pr\left\{ \frac{1}{n} \log \frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)} \leq R - 2\gamma \right\}.$$

Therefore, for all $n \geq n_0$ we have

$$\mu_n + e^{-n\gamma} \geq \Pr\left\{ \frac{1}{n} \log \frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)} \leq R - 2\gamma \right\}.$$

Since $\lim_{n\to\infty} (\mu_n + e^{-n\gamma}) = 0$, it follows that

$$\lim_{n\to\infty} \Pr\left\{ \frac{1}{n} \log \frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)} \leq R - 2\gamma \right\} = 0.$$

Thus, we obtain

$$\underline{D}(\mathbf{X}||\overline{\mathbf{X}}) \geq R - 2\gamma.$$

Since $\gamma > 0$ is arbitrary, $\underline{D}(\mathbf{X}||\overline{\mathbf{X}}) \geq R$ follows. Hence, $B(\mathbf{X}||\overline{\mathbf{X}}) \leq \underline{D}(\mathbf{X}||\overline{\mathbf{X}})$ is established. □

Here, as an application of Theorem 4.1.1, we consider the case that $\mathbf{X}$ and $\overline{\mathbf{X}}$ are the *stationary memoryless sources* subject to probability distributions P_X and $P_{\overline{X}}$, respectively. By Khintchin's theorem (Theorem 1.3.2) the divergence-spectrum of $\frac{1}{n} \log \frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)}$ converges to the one-point spectrum with a peak of probability one at $D(X||\overline{X})$ as $n \to \infty$. This fact implies

$$\underline{D}(\mathbf{X}||\overline{\mathbf{X}}) = D(X||\overline{X}),$$

where $D(X||\overline{X})$ denotes the divergence between X and $\overline{X}$. As a consequence, we obtain the following well-known result (see also Theorem 4.3.2 below):

Corollary 4.1.1.

$$B(\mathbf{X}||\overline{\mathbf{X}}) = D(X||\overline{X}). \tag{4.1.1}$$

The combination of this corollary with Corollary 4.2.1 in the following section is called Stein's lemma.

Example 4.1.1 (Hypothesis testing for the mixed source). Suppose that the null hypothesis $\mathbf{X} = \{X^n\}_{n=1}^{\infty}$ is the mixed source with probability distribution

$$P_{X^n}(\mathbf{x}) = \alpha_1 P_{X_1^n}(\mathbf{x}) + \alpha_2 P_{X_2^n}(\mathbf{x}) \quad (\alpha_1 > 0, \alpha_2 > 0, \alpha_1 + \alpha_2 = 1)$$

and the alternative hypothesis $\mathbf{X} = \{\overline{X}^n\}$ is not the mixed source. Setting $\mathbf{X}_1 = \{X_1^n\}_{n=1}^{\infty}$ and $\mathbf{X}_2 = \{X_2^n\}_{n=1}^{\infty}$, we obtain the formula

$$B(\mathbf{X}||\overline{\mathbf{X}}) = \underline{D}(\mathbf{X}||\overline{\mathbf{X}}) = \min(\underline{D}(\mathbf{X}_1||\overline{\mathbf{X}}), \underline{D}(\mathbf{X}_2||\overline{\mathbf{X}})) \tag{4.1.2}$$

(this formula can be verified by using the argument given in the proofs of Lemma 1.4.1 and Lemma 3.3.1). In particular, if $\mathbf{X}_1$, $\mathbf{X}_2$ and $\overline{\mathbf{X}}$ are the stationary memoryless sources subject to P_{X_1}, P_{X_2} and $P_{\overline{X}}$, respectively, then the divergence-spectrum of $\frac{1}{n} \log \frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)}$ converges to the two-point spectrum with two peaks of probabilities α_1 and α_2 at $D(X_1||\overline{X})$ and $D(X_2||\overline{X})$, respectively, as $n \to \infty$. Therefore, $B(\mathbf{X}||\overline{\mathbf{X}})$ is given by

$$B(\mathbf{X}||\overline{\mathbf{X}}) = \min(D(X_1||\overline{X}), D(X_2||\overline{X})) \tag{4.1.3}$$

(see also Remark 4.4.3). □

Example 4.1.2 (Hypothesis testing for a nonstationary memoryless source). Let us consider the case that both $\mathbf{X} = \{X^n\}_{n=1}^\infty$ and $\overline{\mathbf{X}} = \{\overline{X}^n\}_{n=1}^\infty$ are memoryless sources without stationarity under the assumption that $\mathcal{X}$ is a finite source alphabet. Letting $X^n = (X_1, X_2, \cdots, X_n)$ and $\overline{X}^n = (\overline{X}_1, \overline{X}_2, \cdots, \overline{X}_n)$ be the two memoryless sources, Theorem 4.1.1 and Chebyshev's inequality yield the formula

$$B(\mathbf{X}||\overline{\mathbf{X}}) = \underline{D}(\mathbf{X}||\overline{\mathbf{X}}) = \liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^n D(X_i||\overline{X}_i).$$

For example, if

$$P_{X_i} = \begin{cases} P_{X_1} & (i \text{ is odd}), \\ P_{X_2} & (i \text{ is even}), \end{cases}$$

$$P_{\overline{X}_i} = \begin{cases} P_{\overline{X}_1} & (i \text{ is odd}), \\ P_{\overline{X}_2} & (i \text{ is even}), \end{cases}$$

then it is easy to see that

$$\begin{aligned} B(\mathbf{X}||\overline{\mathbf{X}}) &= \underline{D}(\mathbf{X}||\overline{\mathbf{X}}) \\ &= \frac{1}{2}D(X_1||\overline{X}_1) + \frac{1}{2}D(X_2||\overline{X}_2). \end{aligned}$$

In addition, for the set J defined by (3.2.23) in Remark 3.2.3 in §3.2, if

$$P_{X_i} = \begin{cases} P_{X_1} & \text{for } i \in J, \\ P_{X_2} & \text{for } i \notin J, \end{cases}$$

$$P_{\overline{X}_i} = \begin{cases} P_{\overline{X}_1} & \text{for } i \in J, \\ P_{\overline{X}_2} & \text{for } i \notin J, \end{cases}$$

then we have

$$\begin{aligned} B(\mathbf{X}||\overline{\mathbf{X}}) &= \min_{\frac{1}{3}\le\lambda\le\frac{2}{3}} (\lambda D(X_1||\overline{X}_1) + (1-\lambda)D(X_2||\overline{X}_2)) \\ &= \frac{2}{3}\min(D(X_1||\overline{X}_1), D(X_2||\overline{X}_2)) \\ &\quad + \frac{1}{3}\max(D(X_1||\overline{X}_1), D(X_2||\overline{X}_2)). \end{aligned}$$

We can generalize the mixed source considered in Example 4.1.1 in the following way. For arbitrarily given infinitely many general sources $\mathbf{X}_i = \{X_i^n\}_{n=1}^\infty$ $(i = 1, 2, \cdots)$, we call the source $\mathbf{X} = \{X^n\}_{n=1}^\infty$ defined by

$$P_{X^n}(\mathbf{x}) = \sum_{i=1}^\infty \alpha_i P_{X_i^n}(\mathbf{x}) \quad (\forall n = 1, 2, \cdots; \forall \mathbf{x} \in \mathcal{X}^n) \tag{4.1.4}$$

the *mixed source* of the source family $\{\mathbf{X}_i\}_{i=1}^\infty$, where α_i $(i = 1, 2, \cdots)$ are constants satisfying

$$\sum_{i=1}^{\infty} \alpha_i = 1 \quad (\alpha_i \geq 0 : \ \forall i = 1, 2, \cdots).$$

We have the following lemma characterizing the spectral inf-divergence rate of such a mixed source $\mathbf{X}$ with respect to an arbitrarily given general source $\overline{\mathbf{X}} = \{\overline{X}^n\}_{n=1}^{\infty}$.

Lemma 4.1.3. *For the mixed source* $\mathbf{X}$ *defined in (4.1.4),*

$$\underline{D}(\mathbf{X}||\overline{\mathbf{X}}) = \inf_{i \geq 1: \alpha_i > 0} \underline{D}(\mathbf{X}_i||\overline{\mathbf{X}}). \tag{4.1.5}$$

Proof. We have only to calculate the information-spectrum similarly to the proofs of Lemma 1.4.3 (§1.4) and Lemma 3.3.2 (§3.3). □

Theorem 4.1.1 and Lemma 4.1.3 immediately yield the following theorem.

Theorem 4.1.2. *For the mixed source defined in (4.1.4),*

$$B(\mathbf{X}||\overline{\mathbf{X}}) = \inf_{i \geq 1: \alpha_i > 0} \underline{D}(\mathbf{X}_i||\overline{\mathbf{X}}). \tag{4.1.6}$$

If we consider a special case that all of $\overline{\mathbf{X}} = \{\overline{X}^n\}_{n=1}^{\infty}$ and $\mathbf{X}_i = \{X_i^n\}_{n=1}^{\infty}$ $(i = 1, 2, \cdots)$ are stationary memoryless sources, we obtain the following corollary from Theorem 4.1.2.

Corollary 4.1.2. *Let* $\mathcal{X}$ *be an arbitrary (not necessarily countable) source alphabet. If* $\overline{\mathbf{X}} = \{\overline{X}^n\}_{n=1}^{\infty}$ *and* $\mathbf{X}_i = \{X_i^n\}_{n=1}^{\infty}$ *are the stationary memoryless sources subject to probability distributions* $P_{\overline{X}}$ *and* P_{X_i} $(i = 1, 2, \cdots)$*, respectively, then*

$$B(\mathbf{X}||\overline{\mathbf{X}}) = \inf_{i \geq 1: \alpha_i > 0} D(X_i||\overline{X}) \tag{4.1.7}$$

for the mixed source $\mathbf{X} = \{X^n\}_{n=1}^{\infty}$ *defined in (4.1.4), where* $D(X_i||\overline{X})$ *denotes the divergence.*

Next, let us consider a mixed source with a more general way of mixing (see §1.4 in Chapter 1). Let Φ be an arbitrary set (probability space) and assign a general source $\mathbf{X}_\theta = \{X_\theta^n\}_{n=1}^{\infty}$ to each $\theta \in \Phi$. Here, we assume that, denoting a source alphabet by $\mathcal{X}$, $P_{X_\theta^n}(A)$, the probability of A, is a measurable function of θ for all $n = 1, 2, \cdots$ and for all measurable sets $A \subset \mathcal{X}^n$. If we fix an arbitrary probability measure w on Φ, we have a source $\mathbf{X} = \{X^n\}_{n=1}^{\infty}$ subject to the probability distribution

$$P_{X^n}(A) = \int_{\Phi} P_{X_\theta^n}(A) dw(\theta) \quad (\forall n = 1, 2, \cdots). \tag{4.1.8}$$

This source is called the *mixed source* of the source family $\{\mathbf{X}_\theta\}_{\theta \in \Phi}$. Furthermore, letting $\overline{\mathbf{X}} = \{\overline{X}^n\}_{n=1}^{\infty}$ be another general source, we define the following two functions of R instead of the spectral inf-divergence rate:

$$\underline{K}(R|\mathbf{X}||\overline{\mathbf{X}}) \equiv \liminf_{n\to\infty} \Pr\left\{\frac{1}{n}\log\frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)} \le R\right\}, \tag{4.1.9}$$

$$\overline{K}(R|\mathbf{X}||\overline{\mathbf{X}}) \equiv \limsup_{n\to\infty} \Pr\left\{\frac{1}{n}\log\frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)} \le R\right\}, \tag{4.1.10}$$

each of which is determined from the divergence-spectrum itself. We attempt to characterize these two functions by using $w(\cdot)$. However, since such characterization is difficult for general sources $\mathbf{X}_\theta$ ($\theta \in \Phi$) and $\overline{\mathbf{X}}$, we assume that $\mathcal{X}$ is a finite source alphabet and $\overline{\mathbf{X}}$ and $\mathbf{X}_\theta$ ($\theta \in \Phi$) are the stationary memoryless sources subject to probability distributions $P_{\overline{X}}$ and P_{X_θ} ($\theta \in \Phi$), respectively (we use the notations $\overline{\mathbf{X}} = \{\overline{X}\}$ and $\mathbf{X}_\theta = \{X_\theta\}$ for simplicity). Then, we have the following lemma. This lemma corresponds to Lemma 1.4.4 in §1.4 and Lemma 3.3.3 in §3.3.

Lemma 4.1.4. *Let $\mathcal{X}$ be a finite source alphabet. If each $\mathbf{X}_\theta = \{X_\theta\}$ is stationary and memoryless and so is $\overline{\mathbf{X}} = \{\overline{X}\}$, we have*

$$\int_{\{\theta|D(X_\theta||\overline{X})<R\}} dw(\theta) \le \underline{K}(R|\mathbf{X}||\overline{\mathbf{X}}) \le \overline{K}(R|\mathbf{X}||\overline{\mathbf{X}}) \le \int_{\{\theta|D(X_\theta||\overline{X})\le R\}} dw(\theta) \quad (\forall R \ge 0) \tag{4.1.11}$$

for the mixed source $\mathbf{X}$ defined in (4.1.8), where $D(X_\theta||\overline{X})$ denotes the divergence and the inequalities in (4.1.11) hold with equality except for at most countably infinite R.

Proof. We can prove the lemma by calculating the information-spectrum similarly to the proofs of Lemma 1.4.4 (§1.4) and Lemma 3.3.3 (§3.3). □

Theorem 4.1.1 and Lemma 4.1.4 immediately yield the following theorem.

Theorem 4.1.3. *For the mixed source $\mathbf{X}$ defined in Lemma 4.1.4,*

$$B(\mathbf{X}||\overline{\mathbf{X}}) = w\text{-ess.inf}\, D(X_\theta||\overline{X}). \tag{4.1.12}$$

Proof. We can prove the theorem similarly to the proof of Theorem 1.4.3 (§1.4) with using Lemma 4.1.4. □

Example 4.1.3. Let $\{P_{X_\theta}\}$ be a family of probability distributions with parameter θ over a finite source alphabet $\mathcal{X}$. For each θ denote by $\mathbf{X}_\theta$ the stationary memoryless source subject to a probability distribution P_{X_θ}. Then, $\mathbf{x} = (x_1, \cdots, x_n) \in \mathcal{X}^n$ is generated with probability

$$P_\theta(\mathbf{x}) \equiv \prod_{i=1}^{n} P_{X_\theta}(x_i).$$

Now, let $w(\theta)$ be an arbitrary probability measure and denote by $\mathbf{X} = \{X^n\}_{n=1}^{\infty}$ the mixed source obtained by mixing $\mathbf{X}_\theta$ with respect to the probability density $w(\theta)$. Then, the probability distribution of X^n is given by

$$P_{X^n}(\mathbf{x}) = \int P_\theta(\mathbf{x})dw(\theta) \quad (\forall n = 1, 2, \cdots; \forall \mathbf{x} \in \mathcal{X}^n).$$

Suppose that $\overline{\mathbf{X}} = \{\overline{X}^n\}_{n=1}^{\infty}$ is a stationary memoryless source subject to a probability distribution $P_{\overline{X}}$. Then, Lemma 4.1.4 guarantees that, in the limit of $n \to \infty$, the divergence-spectrum of the mixed source $\mathbf{X}$ against the source $\overline{\mathbf{X}}$ is distributed along the horizontal axis $D(X_\theta||\overline{X})$ with the probability density $w(\theta)$ (Fig. 4.1). Then, the *divergence rate* of $(\mathbf{X}, \overline{\mathbf{X}})$ (see

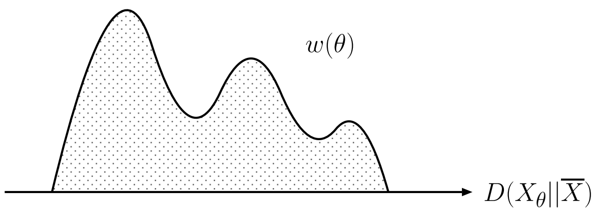

Fig. 4.1.

Remark 4.3.3 in §4.3) is computed as

$$D(\mathbf{X}||\overline{\mathbf{X}}) \equiv \lim_{n\to\infty} \frac{1}{n} D(X^n||\overline{X}^n) = \int D(X_\theta||\overline{X})dw(\theta). \tag{4.1.13}$$

In particular, the divergence-spectrum of Example 4.1.1 becomes the two-point spectrum with the two peaks of probabilities α_1 and α_2 at $D(X_1||\overline{X})$ and $D(X_2||\overline{X})$, respectively. □

We conclude this section by mentioning the hypothesis testing with a *compound source* as the null hypothesis which is deeply related to the hypothesis testing with a mixed source as the null hypothesis (called the *mixed hypothesis testing*) described above. First, suppose that infinitely many null hypotheses $\mathbf{X}_i = \{X_i^n\}_{n=1}^{\infty}$ $(i = 1, 2, \cdots)$ and an alternative hypothesis $\overline{\mathbf{X}} = \{\overline{X}^n\}_{n=1}^{\infty}$ are given. If we define an acceptance region $\mathcal{A}_n \subset \mathcal{X}^n$, the error probability of the first kind is determined by

$$\mu_n^{(i)} \equiv \Pr\{X_i^n \notin \mathcal{A}_n\} \quad (i = 1, 2, \cdots) \tag{4.1.14}$$

for each null hypothesis $\mathbf{X}_i = \{X_i^n\}_{n=1}^{\infty}$ and the error probability of the second kind is determined by

$$\lambda_n \equiv \Pr\{\overline{X}^n \in \mathcal{A}_n\}. \tag{4.1.15}$$

Here, note that the acceptance region $\mathcal{A}_n \subset \mathcal{X}^n$ above does not depend on the suffices $i = 1, 2, \cdots$ of the null hypotheses $\mathbf{X}_i = \{X_i^n\}_{n=1}^\infty$. Such a situation happens when one of the null hypotheses $\mathbf{X}_i = \{X_i^n\}_{n=1}^\infty$ $(i = 1, 2, \cdots)$ surely occurs but a hypothesis tester does not know which one occurs. We call such hypothesis testing the *compound hypothesis testing* $\{\mathbf{X}_i\}_{i=1}^\infty$ against $\overline{\mathbf{X}}$. In the compound hypothesis testing we want to keep the error probability of the first kind small for any null hypothesis that can occur. We attempt to make the error probability of the second kind as small as possible under such a requirement. We give the following definitions.

Definition 4.1.4.

Rate R is achievable $\overset{\text{def}}{\Longleftrightarrow}$ There exists an acceptance region $\mathcal{A}_n \subset \mathcal{X}^n$ satisfying $\lim_{n\to\infty} \mu_n^{(i)} = 0$ $(\forall i = 1, 2, \cdots)$ and

$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\lambda_n} \geq R.$$

Definition 4.1.5. (Supremum achievable error probability exponent in the compound hypothesis testing)

$$B(\{\mathbf{X}_i\}_{i=1}^\infty || \overline{\mathbf{X}}) = \sup \{R \mid R \text{ is achievable}\}.$$

Then, we obtain the following theorem describing a relationship between the supremum achievable error probability exponents of the mixed hypothesis testing and the compound hypothesis testing. The theorem corresponds to Theorem 3.3.5 in Chapter 3 treating channel coding.

Theorem 4.1.4. *Suppose that countably infinite null hypotheses $\mathbf{X}_i = \{X_i^n\}_{n=1}^\infty$ $(i = 1, 2, \cdots)$ and an alternative hypothesis $\overline{\mathbf{X}} = \{\overline{X}^n\}_{n=1}^\infty$ are given. Then, the supremum achievable error probability exponent $B(\{\mathbf{X}_i\}_{i=1}^\infty || \overline{\mathbf{X}})$ in the compound hypothesis testing is equal to $B(\mathbf{X} || \overline{\mathbf{X}}) \equiv B(\{\alpha_i, \mathbf{X}_i\}_{i=1}^\infty || \overline{\mathbf{X}})$, the supremum achievable error probability exponent in the mixed hypothesis testing with the mixed source $\mathbf{X}$ defined by (4.1.4). That is, we have*

$$B(\{\alpha_i, \mathbf{X}_i\}_{i=1}^\infty || \overline{\mathbf{X}}) = B(\{\mathbf{X}_i\}_{i=1}^\infty || \overline{\mathbf{X}}), \tag{4.1.16}$$

where we assume that $\alpha_i > 0$ for all $i = 1, 2, \cdots$.

Proof. Let $\mu_n^{(i)}$ $(i = 1, 2, \cdots)$ and μ_n be the error probability of the first kind for the null hypotheses $\mathbf{X}_i$ $(i = 1, 2, \cdots)$ and the mixed null hypothesis $\mathbf{X}$ with the same acceptance region $\mathcal{A}_n$. From (4.1.4), we have

$$\mu_n = \sum_{i=1}^\infty \alpha_i \mu_n^{(i)}.$$

Then, we can prove this theorem similarly to the proof of Theorem 3.3.5 in Chapter 3. □

The combination of Corollary 4.1.2 with Theorem 4.1.4 immediately yields the following corollary on the compound hypothesis testing.

Corollary 4.1.3. *Let $\mathcal{X}$ be an arbitrary (not necessarily countable) source alphabet. If $\overline{\mathbf{X}} = \{\overline{X}^n\}_{n=1}^\infty$ and $\mathbf{X}_i = \{X_i^n\}_{n=1}^\infty$ $(i = 1, 2, \cdots)$ are stationary memoryless sources subject to $P_{\overline{X}}$ and P_{X_i} $(i = 1, 2, \cdots)$, respectively, then we have*

$$B(\{\mathbf{X}_i\}_{i=1}^\infty \| \overline{\mathbf{X}}) = \inf_{i \geq 1} D(X_i \| \overline{X}) \tag{4.1.17}$$

for the compound hypothesis testing $\{\mathbf{X}_i\}_{i=1}^\infty$ against $\overline{\mathbf{X}}$, where $D(X_i \| \overline{X})$ denotes the divergence. □

4.2 ε-Hypothesis Testing

In the hypothesis testing described in the preceding section, the error probability of the first kind is required to satisfy

$$\mu_n \equiv \Pr\{X^n \notin \mathcal{A}_n\} \to 0 \quad \text{as } n \to \infty.$$

On the other hand, we can consider another requirement that the error probability of the first kind satisfies only

$$\limsup_{n \to \infty} \mu_n \leq \varepsilon$$

for an arbitrary constant $0 \leq \varepsilon < 1$. The exponent of the error probability of the second kind is expected to be large under this weakened requirement on the error probability of the first kind. This section is devoted to analysis of this problem. We first give definitions required for the analysis.

Definition 4.2.1.

$$\text{Rate } R \text{ is } \varepsilon\text{-achievable} \overset{\text{def}}{\Longleftrightarrow} \begin{array}{l} \text{There exists an acceptance region } \mathcal{A}_n \\ \text{satisfying } \limsup_{n \to \infty} \mu_n \leq \varepsilon \text{ and} \\ \liminf_{n \to \infty} \frac{1}{n} \log \frac{1}{\lambda_n} \geq R. \end{array}$$

Definition 4.2.2 (Supremum ε-achievable error probability exponent).

$$B_f(\varepsilon | \mathbf{X} \| \overline{\mathbf{X}}) = \sup\{R \mid R \text{ is } \varepsilon\text{-achievable}\}.$$

We define a function $K(R)$ by

$$K(R) = \limsup_{n\to\infty} \Pr\left\{\frac{1}{n}\log\frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)} \le R\right\} \tag{4.2.1}$$

(see Fig. 4.2). This function is nothing but $\overline{K}(R|\mathbf{X}||\overline{\mathbf{X}})$ defined in §4.1. Then, we obtain the following theorem.

Theorem 4.2.1 (Chen [14]).

$$B_f(\varepsilon|\mathbf{X}||\overline{\mathbf{X}}) = \sup\{R \mid K(R) \le \varepsilon\} \quad (0 \le \forall\varepsilon < 1). \tag{4.2.2}$$

Remark 4.2.1. The right-hand side of (4.2.2) is a right-continuous and monotone increasing function of ε. □

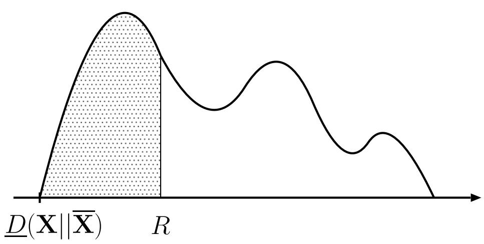

Fig. 4.2.

Proof of Theorem 4.2.1.
1) *Direct part:*

Define $R_0 = \sup\{R \mid K(R) \le \varepsilon\}$. We prove that for an arbitrary $\gamma > 0$ $R = R_0 - \gamma$ is ε-achievable. First, from the definition of $R_0 = \sup\{R \mid K(R) \le \varepsilon\}$, we have $K(R) \le \varepsilon$, i.e.,

$$\limsup_{n\to\infty} \Pr\left\{\frac{1}{n}\log\frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)} \le R\right\} \le \varepsilon. \tag{4.2.3}$$

If we consider the hypothesis testing with the acceptance region

$$\mathcal{A}_n = \left\{\mathbf{x} \in \mathcal{X}^n \mid \frac{1}{n}\log\frac{P_{X^n}(\mathbf{x})}{P_{\overline{X}^n}(\mathbf{x})} > R\right\},$$

Lemma 4.1.1 with $t = R$ implies

$$\lambda_n \le e^{-nR}. \tag{4.2.4}$$

We notice here that, since the left-hand side of (4.2.3) is equal to $\limsup_{n\to\infty} \mu_n$, we have

$$\limsup_{n\to\infty} \mu_n \le \varepsilon.$$

In addition, (4.2.4) guarantees that

$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\lambda_n} \ge R.$$

Consequently, $R = R_0 - \gamma$ is ε-achievable. Since $\gamma > 0$ can be arbitrarily small, $B_f(\varepsilon|\mathbf{X}||\overline{\mathbf{X}}) \ge R_0$ is established.

2) *Converse part:*

Suppose that R is ε-achievable. Then, there exists an acceptance region $\mathcal{A}_n$ satisfying

$$\limsup_{n\to\infty} \mu_n \le \varepsilon \quad \text{and} \quad \liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\lambda_n} \ge R.$$

The second inequality leads to

$$\lambda_n \le e^{-n(R-\gamma)} \quad (\forall n \ge n_0), \tag{4.2.5}$$

where $\gamma > 0$ is an arbitrary constant. If we set $t = R - 2\gamma$ and apply Lemma 4.1.2, it follows that

$$\mu_n + e^{n(R-2\gamma)}\lambda_n \ge \Pr\left\{ \frac{1}{n} \log \frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)} \le R - 2\gamma \right\}.$$

By substituting (4.2.5) into the left-hand side and taking $\limsup_{n\to\infty}$ of the both hand sides, we obtain

$$\varepsilon \ge \limsup_{n\to\infty} \mu_n \ge \limsup_{n\to\infty} \Pr\left\{ \frac{1}{n} \log \frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)} \le R - 2\gamma \right\},$$

i.e.,

$$K(R - 2\gamma) \le \varepsilon. \tag{4.2.6}$$

Here, define $R_0 = \sup\{R \mid K(R) \le \varepsilon\}$ and assume that $R > R_0$. We can choose a sufficiently small $\gamma > 0$ such that $R - 2\gamma > R_0$. Then, the definition of R_0 gives rise to $K(R - 2\gamma) > \varepsilon$, which contradicts (4.2.6). Therefore, $R \le R_0$ must be satisfied. The proof of $B_f(\varepsilon|\mathbf{X}||\overline{\mathbf{X}}) \le R_0$ is now completed. □

Now, consider the case that $\mathbf{X}$ and $\overline{\mathbf{X}}$ are the stationary memoryless sources subject to P_X and $P_{\overline{X}}$, respectively. Since Khintchin's law of large numbers implies that $K(R)$ can be expressed as

$$K(R) = \begin{cases} 0 \text{ for } 0 \le R < D(X||\overline{X}), \\ 1 \text{ for } R > D(X||\overline{X}), \end{cases}$$

we obtain the following corollary from Theorem 4.2.1.

Corollary 4.2.1 (Stein's lemma).

$$B_f(\varepsilon|\mathbf{X}||\overline{\mathbf{X}}) = D(X||\overline{X}) \quad (0 \le \forall\varepsilon < 1). \tag{4.2.7}$$

Example 4.2.1. Let us consider the case that $\overline{\mathbf{X}} = \{\overline{X}^n\}_{n=1}^{\infty}$ is the stationary memoryless source subject to a probability distribution $\overline{P}$ over $\mathcal{X}$ and $\mathbf{X} = \{X^n\}_{n=1}^{\infty}$ the mixed source of $\mathbf{X}_1 = \{X_1^n\}_{n=1}^{\infty}$ and $\mathbf{X}_2 = \{X_2^n\}_{n=1}^{\infty}$ with the probability distribution given by

$$P_{X^n}(\mathbf{x}) = \alpha_1 P_{X_1^n}(\mathbf{x}) + \alpha_2 P_{X_2^n}(\mathbf{x}).$$

Here, $\mathbf{X}_1$ and $\mathbf{X}_2$ are the stationary memoryless sources subject to probability distributions P_1 and P_2, respectively. If we apply the property described in Remark 1.4.1 in §1.4 to

$$\frac{1}{n}\log\frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)},$$

$K(R)$ can be expressed as

$$K(R) = \begin{cases} 0 & \text{for } 0 \le R < D(P_1||\overline{P}), \\ \alpha_1 & \text{for } D(P_1||\overline{P}) < R < D(P_2||\overline{P}). \end{cases}$$

Thus, $B_f(\varepsilon|\mathbf{X}||\overline{\mathbf{X}})$, which is illustrated in Fig. 4.3, is dependent on $0 \le \varepsilon < 1$. □

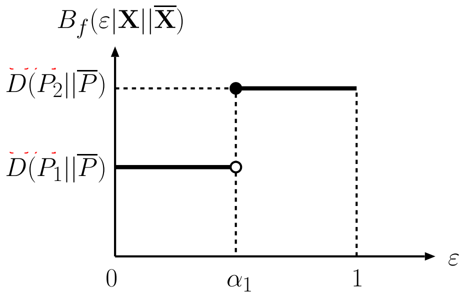

Fig. 4.3.

Example 4.2.2. In the example given in Example 4.1.3 in §4.1, Lemma 4.1.4 guarantees

$$\int_{\{\theta|D(X_\theta||\overline{X})<R\}} dw(\theta) \le K(R) \le \int_{\{\theta|D(X_\theta||\overline{X})\le R\}} dw(\theta).$$

This shows that $K(R)$ is a monotone increasing function of R. Therefore, the formula

$$B_f(\varepsilon|\mathbf{X}||\overline{\mathbf{X}}) = \sup\left\{R\left|\int_{\{\theta|D(X_\theta||\overline{X})\le R\}} dw(\theta) \le \varepsilon\right.\right\} \tag{4.2.8}$$

is obtained from Theorem 4.2.1. This function is also monotone increasing with respect to ε. □

4.3 Strong Converse Theorem for Hypothesis Testing

We also have the strong converse theorem on hypothesis testing corresponding to the strong converse theorems on source coding (§1.5), random number generation (§2.3) the channel coding (§3.5).

Definition 4.3.1. *Consider the hypothesis testing* $\mathbf{X}$ *against* $\overline{\mathbf{X}}$ *and choose a rate* R *satisfying* $R > B(\mathbf{X}||\overline{\mathbf{X}})$ *(cf. Definition 4.1.2) arbitrarily. If* $\lim_{n\to\infty} \mu_n = 1$ *holds for all acceptance regions* $\mathcal{A}_n$ *satisfying* $\liminf_{n\to\infty} \frac{1}{n}\log\frac{1}{\lambda_n} \ge R$*, the hypothesis testing* $\mathbf{X}$ *against* $\overline{\mathbf{X}}$ *is called to satisfy the* strong converse property*.* □

Here, we define:

Definition 4.3.2.

$$\overline{D}(\mathbf{X}||\overline{\mathbf{X}}) = \text{p-}\limsup_{n\to\infty} \frac{1}{n}\log\frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)}$$

and call $\overline{D}(\mathbf{X}||\overline{\mathbf{X}})$ the *spectral sup-divergence rate of* $\mathbf{X}$ against $\overline{\mathbf{X}}$.

Then, we have the following theorem.

Theorem 4.3.1 (Strong converse theorem). *The hypothesis testing* $\mathbf{X}$ *against* $\overline{\mathbf{X}}$ *satisfies the strong converse property if and only if*

$$\underline{D}(\mathbf{X}||\overline{\mathbf{X}}) = \overline{D}(\mathbf{X}||\overline{\mathbf{X}}).$$

Remark 4.3.1. This theorem means that the hypothesis testing satisfies the strong converse property if and only if the information-spectrum of the divergence density rate asymptotically becomes the one-point spectrum with a peak of probability one. □

Proof of Theorem 4.3.1.
1) *Sufficiency:*

Assume that $\underline{D}(\mathbf{X}||\overline{\mathbf{X}}) = \overline{D}(\mathbf{X}||\overline{\mathbf{X}})$. For an arbitrary constant $\gamma > 0$ define R by

$$R = B(\mathbf{X}||\overline{\mathbf{X}}) + 3\gamma = \underline{D}(\mathbf{X}||\overline{\mathbf{X}}) + 3\gamma \tag{4.3.1}$$

and consider an arbitrary hypothesis testing with an acceptance region $\mathcal{A}_n$ satisfying

$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\lambda_n} \geq R.$$

Then, it follows that

$$\frac{1}{n} \log \frac{1}{\lambda_n} \geq R - \gamma \quad (\forall n \geq n_0),$$

which can be written as

$$\lambda_n \leq e^{-n(R-\gamma)} \quad (\forall n \geq n_0). \tag{4.3.2}$$

By noticing that $\mu_n = \Pr\{X^n \notin \mathcal{A}_n\}$ and $\lambda_n = \Pr\left\{\overline{X}^n \in \mathcal{A}_n\right\}$ and substituting (4.3.2) into the inequality in Lemma 4.1.2, setting $t = R - 2\gamma$, we obtain

$$\mu_n + e^{-n\gamma} \geq \Pr\left\{\frac{1}{n} \log \frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)} \leq R - 2\gamma\right\}. \tag{4.3.3}$$

We notice here that (4.3.1) implies $R - 2\gamma = \overline{D}(\mathbf{X}||\overline{\mathbf{X}}) + \gamma$ due to the assumption of $\underline{D}(\mathbf{X}||\overline{\mathbf{X}}) = \overline{D}(\mathbf{X}||\overline{\mathbf{X}})$. Therefore, we obtain

$$\lim_{n\to\infty} \Pr\left\{\frac{1}{n} \log \frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)} \leq R - 2\gamma\right\} = 1.$$

Then, (4.3.3) guarantees $\liminf_{n\to\infty} \mu_n \geq 1$, i.e., $\lim_{n\to\infty} \mu_n = 1$.

2) *Necessity:*
Define $R = B(\mathbf{X}||\overline{\mathbf{X}}) + \gamma$ for an arbitrary constant $\gamma > 0$. If we consider a hypothesis testing with an acceptance region $\mathcal{A}_n$ defined by

$$\mathcal{A}_n = \left\{\mathbf{x} \in \mathcal{X} \mid \frac{1}{n} \log \frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)} \geq R\right\}, \tag{4.3.4}$$

Lemma 4.1.1 implies that $\lambda_n \leq e^{-nR}$, i.e.,

$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\lambda_n} \geq R > B(\mathbf{X}||\overline{\mathbf{X}}).$$

Then, it follows from the assumption of the strong converse property that

$$\lim_{n\to\infty} \Pr\{X^n \notin \mathcal{A}_n\} = \lim_{n\to\infty} \mu_n = 1.$$

By using (4.3.4), we obtain

$$\lim_{n\to\infty} \Pr\left\{\frac{1}{n} \log \frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)} \geq R\right\} = 0,$$

which leads to

$$\overline{D}(\mathbf{X}||\overline{\mathbf{X}}) \le R = B(\mathbf{X}||\overline{\mathbf{X}}) + \gamma = \underline{D}(\mathbf{X}||\overline{\mathbf{X}}) + \gamma.$$

Since $\gamma > 0$ is arbitrary,

$$\overline{D}(\mathbf{X}||\overline{\mathbf{X}}) \le \underline{D}(\mathbf{X}||\overline{\mathbf{X}})$$

is established. By noticing that $\overline{D}(\mathbf{X}||\overline{\mathbf{X}}) \ge \underline{D}(\mathbf{X}||\overline{\mathbf{X}})$, the inequality in the opposite direction, always holds, $\overline{D}(\mathbf{X}||\overline{\mathbf{X}}) = \underline{D}(\mathbf{X}||\overline{\mathbf{X}})$ is established. □

Remark 4.3.2. If the hypothesis testing satisfies the strong converse property, then $B_f(\varepsilon|\mathbf{X}||\overline{\mathbf{X}})$ becomes a constant independent of $0 \le \varepsilon < 1$ (however, the converse is not always true). In particular, if both $\mathbf{X}$ and $\overline{\mathbf{X}}$ are stationary and memoryless, it is obvious that Khintchin's law of large numbers guarantees that the strong converse property is satisfied. Therefore, Corollary 4.2.1 in the preceding section can also be obtained as a consequence of Theorem 4.3.1. □

Remark 4.3.3. For $\mathbf{X} = \{X^n\}_{n=1}^{\infty}$ and $\overline{\mathbf{X}} = \{\overline{X}^n\}_{n=1}^{\infty}$ we define $D(\mathbf{X}||\overline{\mathbf{X}})$ by

$$D(\mathbf{X}||\overline{\mathbf{X}}) = \liminf_{n\to\infty} \frac{1}{n} D(X^n||\overline{X}^n) \tag{4.3.5}$$

and call $D(\mathbf{X}||\overline{\mathbf{X}})$ the *inf-divergence rate* of $\mathbf{X}$ with respect to $\overline{\mathbf{X}}$ (in particular, $D(\mathbf{X}||\overline{\mathbf{X}})$ is simply called the *divergence rate* if the right-hand side of (4.3.5) has a limit). Then, the inequality

$$\underline{D}(\mathbf{X}||\overline{\mathbf{X}}) \le D(\mathbf{X}||\overline{\mathbf{X}}) \tag{4.3.6}$$

can be proved for an arbitrary alphabet $\mathcal{X}$ similarly to the proof of Theorem 3.5.2 using Lemma 3.2.4 with X^n and $\overline{X}^n$ instead of U_n and V_n respectively. On the other hand, the inequality

$$D(\mathbf{X}||\overline{\mathbf{X}}) \le \overline{D}(\mathbf{X}||\overline{\mathbf{X}}) \tag{4.3.7}$$

does not always hold even if $\mathcal{X}$ is a finite alphabet. Therefore, the strong converse property does not always guarantee

$$\underline{D}(\mathbf{X}||\overline{\mathbf{X}}) = D(\mathbf{X}||\overline{\mathbf{X}}) = \overline{D}(\mathbf{X}||\overline{\mathbf{X}}). \tag{4.3.8}$$

This fact means that the property corresponding to Corollary 1.7.1 on source coding and Corollary 3.5.1 on channel coding does not hold on hypothesis testing. □

Though in Remark 4.3.3 above we have seen that the strong converse property does not always imply (4.3.8), we can make (4.3.8) true under a certain condition on sources $\mathbf{X} = \{X^n\}_{n=1}^{\infty}$ and $\overline{\mathbf{X}} = \{\overline{X}^n\}_{n=1}^{\infty}$. That is, for an arbitrary source alphabet $\mathcal{X}$, if $\mathbf{X}$ and $\overline{\mathbf{X}}$ are a *stationary ergodic source* and a *stationary irreducible Markov source* of finite order, respectively, then the divergence density rate $Z_n \equiv \frac{1}{n} \log \frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)}$ converges *almost surely* to $\lim_{n\to\infty} \frac{1}{n} D(X^n||\overline{X}^n)$ (Barron [7]), which implies that $\frac{1}{n} D(X^n||\overline{X}^n)$ on the right-hand side of (4.3.5) has a limit and satisfies (4.3.8). Therefore, the hypothesis testing $\mathbf{X}$ against $\overline{\mathbf{X}}$ for such $\mathbf{X}$ and $\overline{\mathbf{X}}$ satisfies the strong converse property, and hence, we have:

Theorem 4.3.2.

$$B_f(\varepsilon|\mathbf{X}||\overline{\mathbf{X}}) = \lim_{n\to\infty} \frac{1}{n} D(X^n||\overline{X}^n) \quad (0 \le \forall\varepsilon < 1). \tag{4.3.9}$$

This theorem is regarded as a considerable generalization of the formulae (4.1.1) and (4.2.7) given in the preceding sections.

Now, consider a special case that $\mathcal{X}$ is a finite source alphabet and $\mathbf{X}$ and $\overline{\mathbf{X}}$ are *stationary irreducible Markov sources of the first order* with transition probabilities $P(\cdot|\cdot)$ and $\overline{P}(\cdot|\cdot)$, respectively. Then, the formula (4.3.9) yields

$$B_f(\varepsilon|\mathbf{X}||\overline{\mathbf{X}}) = D(P||\overline{P}|p) \quad (0 \le \forall\varepsilon < 1), \tag{4.3.10}$$

where p denotes the stationary distribution of P and the conditional divergence is defined by

$$D(P||\overline{P}|p) = \sum_{x\in\mathcal{X}} p(x) D(P(\cdot|x)||\overline{P}(\cdot|x)).$$

This is a natural generalization of Stein's lemma (Corollary 4.2.1).

Remark 4.3.4. By considering that the hypothesis testing $\mathbf{X}$ against $\overline{\mathbf{X}}$ in (4.3.10) satisfies the strong converse property, Lemma 4.1.4 and Theorem 4.1.3 in §4.1 are generalized in the following way. If $\overline{\mathbf{X}}$ and $\mathbf{X}_\theta$ are the stationary irreducible Markov sources of the first order with transition probabilities $\overline{P}(\cdot|\cdot)$ and $P_\theta(\cdot|\cdot)$, respectively, then all of (4.1.11) and (4.1.12) in §4.1 and (4.2.8) in §4.2 hold, where $D(X_\theta||\overline{X})$ is replaced by the conditional divergence $D(P_\theta||\overline{P}|p_\theta)$ (see Remark 1.4.4 in §1.4 of Chapter 1). □

4.4 Hypothesis Testing and Large Deviation Probability of Testing Error

In §4.1 and §4.2 we have studied the hypothesis testing problems with the error probability of the first kind μ_n converging to 0 or asymptotically bounded by a constant $0 \le \varepsilon < 1$. In this section we consider the hypothesis testing with the error probability of the first kind μ_n required to asymptotically satisfy

$$\mu_n \simeq e^{-nr}$$

for a given constant $r > 0$. Under this constraint we would like to find the maximum of the exponent $R > 0$ when the error probability of the second kind λ_n is expressed as $\lambda_n \simeq e^{-nR}$. Such a problem formulation means to simultaneously evaluate the large deviation behaviors of μ_n and λ_n similarly to the analysis of the large deviation behaviors on the fixed-length source coding given in §1.9. The idea of the information-spectrum slicing plays an important role in this section as well as §1.9.

First, we give two definitions. In this section as well, we denote the null hypothesis and the alternative hypothesis by $\mathbf{X} = \{X^n\}_{n=1}^{\infty}$ and $\overline{\mathbf{X}} = \{\overline{X}^n\}_{n=1}^{\infty}$, respectively.

Definition 4.4.1.

$$\text{Rate } R \text{ is } r\text{-achievable} \stackrel{\text{def}}{\Longleftrightarrow} \text{There exists an acceptance region } \mathcal{A}_n \text{ satisfying } \liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\mu_n} \ge r \text{ and } \liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\lambda_n} \ge R.$$

Definition 4.4.2 (Supremum r-achievable error probability exponents).

$$B_e(r|\mathbf{X}||\overline{\mathbf{X}}) = \sup \{R \mid R \text{ is } r\text{-achievable}\}.$$

The objective of this section is determining this $B_e(r|\mathbf{X}||\overline{\mathbf{X}})$ as a (left-continuous and monotone decreasing) function of r. To this end, we define $\eta(R)$ by

$$\eta(R) = \liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\Pr\left\{\frac{1}{n} \log \frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)} \le R\right\}}. \tag{4.4.1}$$

Though $\eta(R)$ is clearly a monotone decreasing function of R, $\eta(R)$ is not continuous in general.

Lemma 4.4.1. *If $R > \underline{D}(\mathbf{X}||\overline{\mathbf{X}})$, then $\eta(R) = 0$.*

Proof. If $R > \underline{D}(\mathbf{X}||\overline{\mathbf{X}})$, then the definition of $\underline{D}(\mathbf{X}||\overline{\mathbf{X}})$ guarantees the existence of some $0 < \varepsilon_0 < 1$ such that

$$\Pr\left\{\frac{1}{n} \log \frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)} \le R\right\} > \varepsilon_0$$

for infinitely many n. Hence,

$$\eta(R) \leq \liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\varepsilon_0} = 0.$$

□

Lemma 4.4.1 means that $R \leq \underline{D}(\mathbf{X}||\overline{\mathbf{X}})$ must be satisfied for $\eta(R) > 0$.

We have the following quite general theorem.

Theorem 4.4.1 (Han [38]). *For an arbitrary* $r \geq 0$

$$B_e(r|\mathbf{X}||\overline{\mathbf{X}}) = \inf_R \{R + \eta(R) \mid \eta(R) < r\}, \tag{4.4.2}$$

where $B_e(0|\mathbf{X}||\overline{\mathbf{X}})$ $(r = 0)$ *is defined as* $+\infty$.

Remark 4.4.1. Note that $\eta(R) < r$ on the right-hand side of (4.4.2) is not $\eta(R) \leq r$. There is an essential difference between these two as is clarified in the following proof. In addition, $R + \eta(R) \geq 0$ is satisfied for all $-\infty < R < +\infty$ since $\eta(R) \geq -R$ is guaranteed from Lemma 3.2.1 in Chapter 3. □

Remark 4.4.2. From Lemma 4.4.1, we have

$$\inf_{R > \underline{D}(\mathbf{X}||\overline{\mathbf{X}})} \{R + \eta(R) \mid \eta(R) < r\} = \inf_{R > \underline{D}(\mathbf{X}||\overline{\mathbf{X}})} R,$$

where the infimum on the right-hand side is attained at $R = \underline{D}(\mathbf{X}||\overline{\mathbf{X}})$. Therefore, $\inf_R$ on the right of (4.4.2) can be replaced with $\inf_{R \leq \underline{D}(\mathbf{X}||\overline{\mathbf{X}})}$ if $\eta(R)$ is continuous at $R = \underline{D}(\mathbf{X}||\overline{\mathbf{X}})$. □

Proof of Theorem 4.4.1.
1) *Direct part:*

We use the following notation:

$$S_n(a) = \left\{ \mathbf{x} \in \mathcal{X}^n \;\middle|\; \frac{1}{n} \log \frac{P_{X^n}(\mathbf{x})}{P_{\overline{X}^n}(\mathbf{x})} > a \right\}. \tag{4.4.3}$$

Set

$$\underline{R} = \inf \{R \mid \eta(R) < r\} \tag{4.4.4}$$

and consider the hypothesis testing with the acceptance region

$$\mathcal{A}_n = S_n(\underline{R} - \gamma),$$

where $\gamma > 0$ is an arbitrarily small constant. Then, the error probability of the first kind can be written as

$$\begin{aligned}
\mu_n &= \Pr\{X^n \notin \mathcal{A}_n\} \\
&= \Pr\left\{ \frac{1}{n} \log \frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)} \leq \underline{R} - \gamma \right\},
\end{aligned}$$

which leads to

$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\mu_n} = \eta(\underline{R} - \gamma).$$

We notice here that (4.4.4) implies $\eta(\underline{R} - \gamma) \geq r$. Therefore,

$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\mu_n} \geq r \tag{4.4.5}$$

is established. Next, we evaluate the error probability of the second kind. Set

$$\rho_0 = \inf_R \{R + \eta(R) \mid \eta(R) < r\}. \tag{4.4.6}$$

Let K be an arbitrarily large number satisfying $K > \rho_0$ and define $L = (K - \underline{R} + \gamma)/(2\gamma)$. Denote by

$$I_i = (\underline{R} - \gamma + 2(i-1)\gamma, \underline{R} - \gamma + 2i\gamma] \quad (i = 1, 2, \cdots, L) \tag{4.4.7}$$

the L subintervals of the interval $(\underline{R} - \gamma, K]$ each of which has the width 2γ. According to these subintervals we partition the set

$$T_0 = \left\{ \mathbf{x} \in \mathcal{X}^n \,\middle|\, \underline{R} - \gamma < \frac{1}{n} \log \frac{P_{X^n}(\mathbf{x})}{P_{\overline{X}^n}(\mathbf{x})} \leq K \right\}$$

into L subsets as follows:

$$S_n^{(i)} = \left\{ \mathbf{x} \in \mathcal{X}^n \,\middle|\, \frac{1}{n} \log \frac{P_{X^n}(\mathbf{x})}{P_{\overline{X}^n}(\mathbf{x})} \in I_i \right\} \quad (i = 1, 2, \cdots, L)$$

(*information-spectrum slicing*). In addition, set

$$S_n^{(0)} = \left\{ \mathbf{x} \in \mathcal{X}^n \,\middle|\, \frac{1}{n} \log \frac{P_{X^n}(\mathbf{x})}{P_{\overline{X}^n}(\mathbf{x})} > K \right\}.$$

It is clear that

$$S_n(\underline{R} - \gamma) = \bigcup_{i=0}^{L} S_n^{(i)}. \tag{4.4.8}$$

If we set $b_i = \underline{R} - \gamma + 2i\gamma$ for simplicity, (4.4.7) can be expressed as

$$I_i = (b_i - 2\gamma, b_i] \quad (i = 1, 2, \cdots, L).$$

Since for $i = 1, 2, \cdots, L$

$$\Pr\left\{ X^n \in S_n^{(i)} \right\} \leq \Pr\left\{ \frac{1}{n} \log \frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)} \leq b_i \right\},$$

it follows that

$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\Pr\left\{ X^n \in S_n^{(i)} \right\}} \geq \eta(b_i).$$

Hence, we obtain

$$\Pr\left\{X^n \in S_n^{(i)}\right\} \le e^{-n(\eta(b_i)-\gamma)} \quad (\forall n \ge n_0). \tag{4.4.9}$$

We notice here that, since $\mathbf{x} \in S_n^{(i)}$ implies that

$$\frac{1}{n}\log\frac{P_{X^n}(\mathbf{x})}{P_{\overline{X}^n}(\mathbf{x})} > b_i - 2\gamma,$$

we have the following inequality:

$$P_{\overline{X}^n}(\mathbf{x}) \le P_{X^n}(\mathbf{x})e^{-n(b_i-2\gamma)}.$$

Then, it follows from (4.4.9) that

$$\begin{aligned}\Pr\left\{\overline{X}^n \in S_n^{(i)}\right\} &\le \sum_{\mathbf{x}\in S_n^{(i)}} P_{X^n}(\mathbf{x})e^{-n(b_i-2\gamma)} \\ &\le e^{-n(b_i+\eta(b_i)-3\gamma)}. \end{aligned} \tag{4.4.10}$$

If we note here that $b_i \ge \underline{R}+\gamma$ for all $i = 1, 2, \cdots, L$, we have

$$b_i + \eta(b_i) \ge \rho_0 \quad (i = 1, 2, \cdots, L).$$

By substituting this into (4.4.10), it holds that

$$\Pr\left\{\overline{X}^n \in S_n^{(i)}\right\} \le e^{-n(\rho_0-3\gamma)} \quad (i = 1, 2, \cdots, L). \tag{4.4.11}$$

On the other hand, by taking the fact that $P_{\overline{X}^n}(\mathbf{x}) \le P_{X^n}(\mathbf{x})e^{-nK}$ for $\mathbf{x} \in S_n^{(0)}$ into consideration, we have

$$\begin{aligned}\Pr\left\{\overline{X}^n \in S_n^{(0)}\right\} &= \sum_{\mathbf{x}\in S_n^{(0)}} P_{\overline{X}^n}(\mathbf{x}) \\ &\le e^{-nK}\sum_{\mathbf{x}\in S_n^{(0)}} P_{X^n}(\mathbf{x}) \\ &\le e^{-nK}. \end{aligned} \tag{4.4.12}$$

Then, (4.4.8), (4.4.11) and (4.4.12) lead to

$$\lambda_n = \Pr\left\{\overline{X}^n \in S_n(\underline{R}-\gamma)\right\} \le Le^{-n(\rho_0-3\gamma)} + e^{-nK}.$$

We now obtain

$$\liminf_{n\to\infty}\frac{1}{n}\log\frac{1}{\lambda_n} \ge \rho_0 - 3\gamma$$

since $K > \rho_0$ guarantees that $K > \rho_0 - 3\gamma$ for $\gamma > 0$. By noticing (4.4.5), we can conclude that $\rho_0 - 3\gamma$ is r-achievable (note that $\gamma > 0$ can be arbitrarily small).

2) *Converse part:*

Let $\underline{R}$ and ρ_0 be defined as in (4.4.4) and (4.4.6), respectively. Then, since $\eta(R)$ is monotone decreasing in R, there exists an R_0 satisfying $R_0 \geq \underline{R}$ and

$$\lim_{\varepsilon \downarrow 0}(R_0 + \varepsilon + \eta(R_0 + \varepsilon)) = \rho_0. \tag{4.4.13}$$

Let us consider the set

$$S_0 = \left\{ \mathbf{x} \in \mathcal{X}^n \,\middle|\, \frac{1}{n} \log \frac{P_{X^n}(\mathbf{x})}{P_{\overline{X}^n}(\mathbf{x})} \leq R_0 + \gamma \right\},$$

where $\gamma > 0$ is an arbitrarily small constant. Then, from the definition of $\eta(R)$, there exists some divergent sequence $n_1 < n_2 < \cdots \to \infty$ of integers such that

$$\Pr\{X^{n_j} \in S_0\} \geq e^{-n_j(\eta(R_0+\gamma)+\tau)} \quad (\forall j \geq j_0), \tag{4.4.14}$$

where $\tau > 0$ is an arbitrarily small constant. We prove the converse part by the contradiction argument. To do so, assume that $R = \rho_0 + 2\delta$ ($\delta > 0$ is a fixed constant) is r-achievable, i.e., assume that there exists an acceptance region $\mathcal{A}_n$ satisfying

$$\liminf_{n \to \infty} \frac{1}{n} \log \frac{1}{\mu_n} \geq r \tag{4.4.15}$$

and

$$\liminf_{n \to \infty} \frac{1}{n} \log \frac{1}{\lambda_n} \geq R \equiv \rho_0 + 2\delta. \tag{4.4.16}$$

Since $\mathbf{x} \in S_0$ implies

$$P_{X^n}(\mathbf{x}) \leq P_{\overline{X}^n}(\mathbf{x}) e^{n(R_0+\gamma)},$$

we have

$$\begin{aligned} \Pr\{X^n \in S_0 \cap \mathcal{A}_n\} &= \sum_{\mathbf{x} \in S_0 \cap \mathcal{A}_n} P_{X^n}(\mathbf{x}) \\ &\leq \sum_{\mathbf{x} \in S_0 \cap \mathcal{A}_n} P_{\overline{X}^n}(\mathbf{x}) e^{n(R_0+\gamma)} \\ &\leq e^{n(R_0+\gamma)} \sum_{\mathbf{x} \in \mathcal{A}_n} P_{\overline{X}^n}(\mathbf{x}) \\ &= \lambda_n e^{n(R_0+\gamma)}. \end{aligned} \tag{4.4.17}$$

Furthermore, it follows from (4.4.16) that

$$\lambda_n \leq e^{-n(R-\gamma)} \quad (\forall n \geq n_0).$$

Substitution of this into (4.4.17) yields

$$\begin{aligned} \Pr\{X^n \in S_0 \cap \mathcal{A}_n\} &\leq e^{-n(R-R_0-2\gamma)} \\ &= e^{-n(\rho_0 - R_0 + 2\delta - 2\gamma)}. \end{aligned} \tag{4.4.18}$$

By virtue of (4.4.13), for any $\gamma > 0$ small enough,

$$\rho_0 \geq R_0 + \gamma + \eta(R_0 + \gamma) - \delta.$$

Therefore, by (4.4.18) we have

$$\Pr\{X^n \in S_0 \cap \mathcal{A}_n\} \leq e^{-n(\eta(R_0+\gamma)+\delta-\gamma)}.$$

If we choose $\tau > 0$ and $\gamma > 0$ so small as to satisfy $\delta > 2\tau + \gamma$, then

$$\Pr\{X^n \in S_0 \cap \mathcal{A}_n\} \leq e^{-n(\eta(R_0+\gamma)+2\tau)}, \tag{4.4.19}$$

where $\tau > 0$ is the same one as in (4.4.14). On the other hand, by using (4.4.15), we obtain

$$\begin{aligned} \Pr\{X^n \in S_0 \cap \mathcal{A}_n^c\} &\leq \Pr\{X^n \in \mathcal{A}_n^c\} \\ &= \mu_n \leq e^{-n(r-\tau)} \quad (\forall n \geq n_0). \end{aligned} \tag{4.4.20}$$

We observe here that $\eta(R_0 + \gamma) < r$ for all $\gamma > 0$, and hence, for any sufficiently small $\tau > 0$,

$$\eta(R_0 + \gamma) + 2\tau < r - \tau.$$

Then, it follows from (4.4.19) and (4.4.20) that

$$\begin{aligned} \Pr\{X^n \in S_0\} &= \Pr\{X^n \in S_0 \cap \mathcal{A}_n\} + \Pr\{X^n \in S_0 \cap \mathcal{A}_n^c\} \\ &\leq e^{-n(\eta(R_0+\gamma)+2\tau)} + e^{-n(r-\tau)} \\ &\leq 2e^{-n(\eta(R_0+\gamma)+2\tau)} \end{aligned} \tag{4.4.21}$$

for all $n \geq n_0$. However, since $\tau > 0$, (4.4.21) contradicts (4.4.14). Thus, the rate $R = \rho_0 + 2\delta$ cannot be r-achievable. Since $\delta > 0$ is arbitrary, we can conclude that any R such that $R > \rho_0$ cannot be r-achievable. □

Example 4.4.1. Let $\mathbf{X} = \{X^n\}_{n=1}^{\infty}$ and $\overline{\mathbf{X}} = \{\overline{X}^n\}_{n=1}^{\infty}$ be *stationary memoryless sources* subject to probability distributions P and $\overline{P}$ over a finite alphabet $\mathcal{X}$, respectively, and consider the corresponding hypothesis testing. We first define the plane

$$\kappa_R = \left\{ Q \in \mathcal{P}(\mathcal{X}) \,\middle|\, \sum_{x \in \mathcal{X}} Q(x) \log \frac{P(x)}{\overline{P}(x)} = R \right\} \tag{4.4.22}$$

in $\mathcal{P}(\mathcal{X})$, where $\mathcal{P}(\mathcal{X})$ denotes the set of all probability distributions over $\mathcal{X}$. Denote by P_R the *projection* of P on κ_R in the sense of the divergence (see Example 1.9.1 in §1.9). Figure 4.4 illustrates such a situation. Then, Sanov's theorem (cf. Dembo and Zeitouni [22]) tells us that $\eta(R) = 0$ if $R \geq D(P||\overline{P})$ and $\eta(R) = D(P_R||P)$ if $R \leq D(P||\overline{P})$. Here, note that $\underline{D}(\mathbf{X}||\overline{\mathbf{X}}) = D(P||\overline{P})$ from the law of large numbers. Thus, in (4.4.2) we have only to consider R satisfying $R \leq D(P||\overline{P})$ (see Remark 4.4.2). Since P_R is on κ_R, P_R satisfies the equation

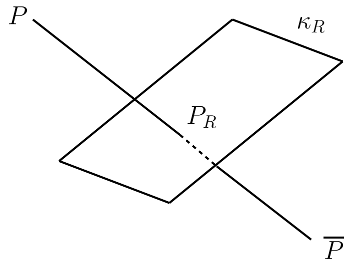

Fig. 4.4.

$$\sum_{x\in\mathcal{X}} P_R(x)\log\frac{P(x)}{\overline{P}(x)} = R,$$

which can be written as

$$D(P_R||\overline{P}) - D(P_R||P) = R. \tag{4.4.23}$$

Hence, we have

$$R + \eta(R) = D(P_R||\overline{P}).$$

Here, note that (4.4.23) and the definition of P_R imply that

$$D(P_R||\overline{P}) = \inf_{Q\in\kappa_R} D(Q||\overline{P}).$$

Then, it follows from Theorem 4.4.1 that

$$\begin{aligned} B_e(r|\mathbf{X}||\overline{\mathbf{X}}) &= \inf_R \{R + \eta(R) \mid \eta(R) < r\} \\ &= \inf_R \left\{D(P_R||\overline{P}) \mid D(P_R||P) < r\right\}, \end{aligned}$$

which can be immediately written as

$$B_e(r|\mathbf{X}||\overline{\mathbf{X}}) = \inf_{Q:D(Q||P)<r} D(Q||\overline{P}). \tag{4.4.24}$$

This is nothing but Hoeffding's theorem [50], well-known in statistics. This formula also implies that $B_e(r|\mathbf{X}||\overline{\mathbf{X}}) = 0$ for all $r \geq D(\overline{P}||P)$. □

Example 4.4.2. Let $\mathcal{X}$ be a finite alphabet. Let us consider the case that the null hypothesis $\mathbf{X} = (X_1, X_2, \cdots)$ and the alternative hypothesis $\overline{\mathbf{X}} = (\overline{X}_1, \overline{X}_2, \cdots)$ are first-order *stationary irreducible Markov sources* subject to transition probabilities $P(x_2|x_1) = \Pr\{X_2 = x_2|X_1 = x_1\}$ and $\overline{P}(x_2|x_1) = \Pr\left\{\overline{X}_2 = x_2|\overline{X}_1 = x_1\right\}$ for $x_1, x_2 \in \mathcal{X}$, respectively. Similarly to Example 1.9.2 in §1.9 of Chapter 1, we denote by $\mathcal{P}(\mathcal{X}\times\mathcal{X})$ the set of all joint probability distributions over $\mathcal{X}\times\mathcal{X}$ and define the conditional divergences by

$$D(Q||P|q) = \sum_{x_1 \in \mathcal{X}} q(x_1) D(Q(\cdot|x_1)||P(\cdot|x_1)),$$
$$D(Q||\overline{P}|q) = \sum_{x_1 \in \mathcal{X}} q(x_1) D(Q(\cdot|x_1)||\overline{P}(\cdot|x_1))$$

for any $Q \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$, where $q(\cdot)$ and $Q(\cdot|\cdot)$ denote the marginal distribution and the conditional probability distribution defined by

$$q(x_1) = \sum_{x_2 \in \mathcal{X}} Q(x_1, x_2),$$
$$Q(x_2|x_1) = \frac{Q(x_1, x_2)}{q(x_1)},$$

respectively. Then, from the argument using Sanov's theorem for stationary irreducible Markov sources similarly to Example 4.4.1, we have $\eta(R) = 0$ for $R \geq D(P||\overline{P}|p)$ and

$$\eta(R) = D(P_R||P|p_R), \tag{4.4.25}$$
$$R + \eta(R) = D(P_R||\overline{P}|p_R) \tag{4.4.26}$$

for $R \leq D(P||\overline{P}|p)$. Here, p means the stationary distribution of P, $\mathcal{P}_0$ denotes the set of all probability distributions of $Q \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$ with the stationarity (see Example 1.9.2 in §1.9 of Chapter 1), $P_R \in \mathcal{P}_0$ denotes the *projection* of P on the plane

$$\overline{\kappa}_R = \left\{ Q \in \mathcal{P}_0 \,\middle|\, \sum_{x_1, x_2 \in \mathcal{X}} Q(x_1, x_2) \log \frac{P(x_2|x_1)}{\overline{P}(x_2|x_1)} = R \right\}, \tag{4.4.27}$$

and p_R means the marginal distribution of P_R. We define the projection of P as the distribution P_R satisfying

$$\inf_{Q \in \overline{\kappa}_R} D(Q||P|q) = D(P_R||P|p_R),$$

where q denotes the marginal distribution of Q. Hence, we obtain

$$\begin{aligned} B_e(r|\mathbf{X}||\overline{\mathbf{X}}) &= \inf_R \left\{ D(P_R||\overline{P}|p_R) \mid D(P_R||P|p_R) < r \right\} \\ &= \inf_{Q \in \mathcal{P}_0 : D(Q||P|q) < r} D(Q||\overline{P}|q) \quad (\forall r > 0) \end{aligned} \tag{4.4.28}$$

from Theorem 4.4.1 (cf. Natarajan [72]). This formula tells us that $B_e(r|\mathbf{X}||\overline{\mathbf{X}}) = 0$ for all $r \geq D(\overline{P}||P|\overline{p})$, where $\overline{p}$ denotes the stationary distribution of $\overline{P}$. □

Example 4.4.3. Let us generalize Example 4.4.2 above to the hypothesis testing for *unifilar finite-state sources* (see Example 1.9.3 in §1.9 of Chapter 1). To this end, let $\mathcal{X}$ be a finite source alphabet and $\mathcal{S}$ a finite set of states. Let the null hypothesis $\mathbf{X} = \{X^n = (X_1, \cdots, X_n)\}_{n=1}^{\infty}$ be the unifilar finite-state source subject to

$$P_{X^n}(\mathbf{x}) = \prod_{i=1}^{n} P(x_i|s_i) \quad (\mathbf{x} = (x_1, x_2, \cdots, x_n) \in \mathcal{X}^n) \tag{4.4.29}$$

$$s_{i+1} = f(x_i, s_i) \quad (s_i \in \mathcal{S};\ i = 1, 2, \cdots, n) \tag{4.4.30}$$

and the alternative hypothesis $\overline{\mathbf{X}} = \{\overline{X}^n = (\overline{X}_1, \cdot, \overline{X}_n)\}_{n=1}^{\infty}$ the unifilar finite-state source subject to

$$P_{\overline{X}^n}(\mathbf{x}) = \prod_{i=1}^{n} \overline{P}(x_i|s_i) \quad (\mathbf{x} = (x_1, x_2, \cdots, x_n) \in \mathcal{X}^n) \tag{4.4.31}$$

$$s_{i+1} = f(x_i, s_i) \quad (s_i \in \mathcal{S};\ i = 1, 2, \cdots, n). \tag{4.4.32}$$

We now fix an initial state $s_1 \in \mathcal{S}$ arbitrarily and denote by $\mathcal{S}_0$ the set of all states that can be reached from s_1 with "positive probability" with respect to P_{X^n}. Next, let $XS \equiv (X, S)$ be an arbitrary random variable taking values in $\mathcal{X} \times \mathcal{S}_0$ and define

$$S' = f(X, S). \tag{4.4.33}$$

Furthermore, denote by $\mathcal{V}_0$ the set of all random variables XS satisfying both the stationary condition

$$P_{S'}(\cdot) = P_S(\cdot)$$

and the condition that the probability transition matrix $P_{S'|S}(\cdot|\cdot)$ is irreducible. Now, setting

$$\overline{\lambda}_R = \left\{ P_{XS} \in \mathcal{V}_0 \,\middle|\, \sum_{x \in \mathcal{X}, s \in \mathcal{S}_0} P_{XS}(x, s) \log \frac{P(x|s)}{\overline{P}(x|s)} = R \right\}, \tag{4.4.34}$$

we define the *projection* $P_{X_R S_R} \in \mathcal{V}_0$ of $P(\cdot|\cdot)$ on the plane $\overline{\lambda}_R$ by

$$\inf_{P_{XS} \in \overline{\lambda}_R} D(P_{XS}||P|P_S) = D(P_{X_R S_R}||P|P_{S_R}).$$

Then, similarly to Example 4.4.2 Sanov's theorem for unifilar finite-state sources (cf. Han [37]) yields

$$\eta(R) = D(P_{X_R S_R}||P|P_{S_R}) \tag{4.4.35}$$

$$R + \eta(R) = D(P_{X_R S_R}||\overline{P}|P_{S_R}). \tag{4.4.36}$$

Thus, by substituting these equalities into Theorem 4.4.1, we obtain the following formula for the hypothesis testing for unifilar finite-state sources $\mathbf{X}$ against $\overline{\mathbf{X}}$:

$$\begin{aligned} &B_e(r|\mathbf{X}||\overline{\mathbf{X}}) \\ &= \inf_{R} \left\{ D(P_{X_R S_R}||\overline{P}|P_{S_R}) \,|D(P_{X_R S_R}||P|P_{S_R}) < r \right\} \\ &= \inf_{P_{XS} \in \mathcal{V}_0 : D(P_{XS}||P|P_S) < r} D(P_{XS}||\overline{P}|P_S) \quad (\forall r > 0), \end{aligned} \tag{4.4.37}$$

where P_{XS} and P_S denote the probability distributions of random variables XS and S, respectively, and the conditional divergences are defined by

$$D(P_{XS}||P|P_S) = \sum_{s \in \mathcal{S}_0} P_S(s) D(P_{X|S}(\cdot|s)||P(\cdot|s)), \tag{4.4.38}$$

$$D(P_{XS}||\overline{P}|P_S) = \sum_{s \in \mathcal{S}_0} P_S(s) D(P_{X|S}(\cdot|s)||\overline{P}(\cdot|s)). \tag{4.4.39}$$

Recall here that, in general, every unifilar finite-state source is asymptotically a *mixed source* of stationary (or periodic) irreducible sources (see Example 1.9.3 in §1.9 of Chapter 1).

If we consider the hypothesis testing $\mathbf{X}$ against $\overline{\mathbf{X}}$ for the unifilar finite-state sources $\mathbf{X}$ and $\overline{\mathbf{X}}$, it is easy to verify that the following formula on the supremum achievable error probability exponent $B(\mathbf{X}||\overline{\mathbf{X}})$ (see Definition 4.1.2) holds:

$$B(\mathbf{X}||\overline{\mathbf{X}}) = \inf_{XS \in \overline{\mathcal{V}}_0} D(P_{XS}||\overline{P}|P_S), \tag{4.4.40}$$

where $\overline{\mathcal{V}}_0$ denotes the set of all random variables $XS \in \mathcal{V}_0$ satisfying the condition $P_{X|S}(\cdot|\cdot) = P(\cdot|\cdot)$. □

Example 4.4.4. Let $\mathcal{X}$ be a finite alphabet and consider the mixed source $\mathbf{X} = \{X^n\}_{n=1}^{\infty}$ and the stationary memoryless source $\overline{\mathbf{X}} = \{\overline{X}^n\}_{n=1}^{\infty}$ subject to the probability distribution $\overline{P}$ given in Example 4.2.1. Recall that the *mixed source* $\mathbf{X} = \{X^n\}_{n=1}^{\infty}$ is defined by

$$P_{X^n}(\mathbf{x}) = \alpha_1 P_{X_1^n}(\mathbf{x}) + \alpha_2 P_{X_2^n}(\mathbf{x}) \quad (\forall \mathbf{x} \in \mathcal{X}^n) \tag{4.4.41}$$

for the two stationary memoryless sources $\mathbf{X}_1 = \{X_1^n\}_{n=1}^{\infty}$ and $\mathbf{X}_2 = \{X_2^n\}_{n=1}^{\infty}$ subject to probability distributions P_1 and P_2, respectively. We define ν_1 and ν_2 by

$$\nu_1 = \left\{ Q \in \mathcal{P}(\mathcal{X}) \,\middle|\, \sum_{x \in \mathcal{X}} Q(x) \log \frac{P_1(x)}{P_2(x)} \geq 0 \right\}, \tag{4.4.42}$$

$$\nu_2 = \left\{ Q \in \mathcal{P}(\mathcal{X}) \,\middle|\, \sum_{x \in \mathcal{X}} Q(x) \log \frac{P_1(x)}{P_2(x)} \leq 0 \right\} \tag{4.4.43}$$

as are defined in (1.9.34) and (1.9.35) in Example 1.9.4 in §1.9 of Chapter 1 and two half-spaces in $\mathcal{P}(\mathcal{X})$ by

$$\kappa_R^{(1)} = \left\{ Q \in \mathcal{P}(\mathcal{X}) \,\middle|\, \sum_{x \in \mathcal{X}} Q(x) \log \frac{P_1(x)}{\overline{P}(x)} \leq R \right\}, \tag{4.4.44}$$

$$\kappa_R^{(2)} = \left\{ Q \in \mathcal{P}(\mathcal{X}) \,\middle|\, \sum_{x \in \mathcal{X}} Q(x) \log \frac{P_2(x)}{\overline{P}(x)} \leq R \right\}, \tag{4.4.45}$$

where $\mathcal{P}(\mathcal{X})$ denotes the space of all probability distributions over $\mathcal{X}$. By taking (1.9.39) and (1.9.40) in Example 1.9.4 into account, Sanov's theorem yields

$$\eta(R) = \min(D(P_R^{(1)}||P_1), D(P_R^{(2)}||P_2)), \tag{4.4.46}$$

where $P_R^{(1)}$ and $P_R^{(2)}$ denote the *projections* of P_1 and P_2 on $\nu_1 \cap \kappa_R^{(1)}$ and $\nu_2 \cap \kappa_R^{(2)}$, respectively. If we substitute this $\eta(R)$ into the right-hand side of (4.4.2) in Theorem 4.4.1, we can compute values of $B_e(r|\mathbf{X}||\overline{\mathbf{X}})$ for the mixed source as a function of r.

We note here that it easily follows from (4.4.46) that $\eta(R) = 0$ for R satisfying $R \geq \min(D(P_1||\overline{P}), D(P_2||\overline{P}))$ and $\eta(R)$ is a monotone decreasing and continuous function of R. Therefore,

$$B_e(r|\mathbf{X}||\overline{\mathbf{X}}) \leq \min(D(P_1||\overline{P}), D(P_2||\overline{P})) \quad (\forall r > 0). \tag{4.4.47}$$

On the other hand, since $\eta(h) > 0$ for any rate h satisfying $h < \min(D(P_1||\overline{P}), D(P_2||\overline{P}))$ is verified from (4.4.46), we have

$$\inf_R \{R + \eta(R)|\eta(R) < \eta(h)\} \geq h,$$

which implies that h is $\eta(h)$-achievable. Hence, it holds that

$$\lim_{r\downarrow 0} B_e(r|\mathbf{X}||\overline{\mathbf{X}}) = \min(D(P_1||\overline{P}), D(P_2||\overline{P})). \tag{4.4.48}$$

Example 4.4.5. Example 4.4.4 can be generalized in the following way (suppose that $\mathcal{X}$ is a finite alphabet). First, denote by $\mathbf{X}_1 = \{X_1^n\}_{n=1}^\infty$, $\mathbf{X}_2 = \{X_2^n\}_{n=1}^\infty$, $\overline{\mathbf{X}}_1 = \{\overline{X}_1^n\}_{n=1}^\infty$ and $\overline{\mathbf{X}}_2 = \{\overline{X}_2^n\}_{n=1}^\infty$ the stationary memoryless sources subject to probability distributions $P_1, P_2, \overline{P}_1$ and $\overline{P}_2$, respectively. Consider the hypothesis testing with the *mixed source* $\mathbf{X} = \{X^n\}_{n=1}^\infty$ defined by

$$P_{X^n}(\mathbf{x}) = \alpha_1 P_{X_1^n}(\mathbf{x}) + \alpha_2 P_{X_2^n}(\mathbf{x}) \quad (\alpha_1 > 0, \alpha_2 > 0, \alpha_1 + \alpha_2 = 1) \tag{4.4.49}$$

as the null hypothesis and the *mixed source* $\overline{\mathbf{X}} = \{\overline{X}^n\}_{n=1}^\infty$ defined by

$$P_{\overline{X}^n}(\mathbf{x}) = \beta_1 P_{\overline{X}_1^n}(\mathbf{x}) + \beta_2 P_{\overline{X}_2^n}(\mathbf{x}) \quad (\beta_1 > 0, \beta_2 > 0, \beta_1 + \beta_2 = 1) \tag{4.4.50}$$

as the alternative hypothesis. We define ν_1 and ν_2 by (4.4.42) and (4.4.43) in Example 4.4.4, respectively, and μ_1 and μ_2 by

$$\mu_1 = \left\{ Q \in \mathcal{P}(\mathcal{X}) \,\middle|\, \sum_{x\in\mathcal{X}} Q(x) \log \frac{\overline{P}_1(x)}{\overline{P}_2(x)} \geq 0 \right\}, \tag{4.4.51}$$

$$\mu_2 = \left\{ Q \in \mathcal{P}(\mathcal{X}) \,\middle|\, \sum_{x\in\mathcal{X}} Q(x) \log \frac{\overline{P}_1(x)}{\overline{P}_2(x)} \leq 0 \right\}, \tag{4.4.52}$$

where $\mathcal{P}(\mathcal{X})$ denotes the space of all probability distributions over $\mathcal{X}$. Furthermore, define the four half-spaces in $\mathcal{P}(\mathcal{X})$ by

$$\kappa_R^{(1)} = \left\{ Q \in \mathcal{P}(\mathcal{X}) \middle| \sum_{x \in \mathcal{X}} Q(x) \log \frac{P_1(x)}{\overline{P}_1(x)} \le R \right\}, \tag{4.4.53}$$

$$\kappa_R^{(2)} = \left\{ Q \in \mathcal{P}(\mathcal{X}) \middle| \sum_{x \in \mathcal{X}} Q(x) \log \frac{P_1(x)}{\overline{P}_2(x)} \le R \right\}, \tag{4.4.54}$$

$$\kappa_R^{(3)} = \left\{ Q \in \mathcal{P}(\mathcal{X}) \middle| \sum_{x \in \mathcal{X}} Q(x) \log \frac{P_2(x)}{\overline{P}_2(x)} \le R \right\}, \tag{4.4.55}$$

$$\kappa_R^{(4)} = \left\{ Q \in \mathcal{P}(\mathcal{X}) \middle| \sum_{x \in \mathcal{X}} Q(x) \log \frac{P_2(x)}{\overline{P}_1(x)} \le R \right\} \tag{4.4.56}$$

and denote by $P_R^{(1)}$ and $P_R^{(2)}$ the *projections* of P_1 and P_2 on

$$\nu_1 \cap \left((\mu_1 \cap \kappa_R^{(1)}) \cup (\mu_2 \cap \kappa_R^{(2)}) \right),$$

$$\nu_2 \cap \left((\mu_2 \cap \kappa_R^{(3)}) \cup (\mu_1 \cap \kappa_R^{(4)}) \right),$$

respectively. By applying Sanov's theorem similarly to Example 4.4.4, we have

$$\eta(R) = \min(D(P_R^{(1)}||P_1), D(P_R^{(2)}||P_2)). \tag{4.4.57}$$

If this $\eta(R)$ is substituted into the right-hand side of (4.4.2) in Theorem 4.4.1, we can compute values of $B_e(r|\mathbf{X}||\overline{\mathbf{X}})$ for the mixed sources as a function of r.

We note here that it easily follows from (4.4.57) that $\eta(R) = 0$ if

$$R \ge \min(D(P_1||\overline{P}_1), D(P_1||\overline{P}_2), D(P_2||\overline{P}_1), D(P_2||\overline{P}_2))$$

and $\eta(R)$ is a monotone decreasing and continuous function of R. Therefore,

$$\begin{aligned} &B_e(r|\mathbf{X}||\overline{\mathbf{X}}) \\ &\le \min(D(P_1||\overline{P}_1), D(P_1||\overline{P}_2), D(P_2||\overline{P}_1), D(P_2||\overline{P}_2)) \quad (\forall r > 0). \end{aligned} \tag{4.4.58}$$

On the other hand, since $\eta(h) > 0$ is obtained from (4.4.57) for any rate h satisfying

$$0 < h < \min(D(P_1||\overline{P}_1), D(P_1||\overline{P}_2), D(P_2||\overline{P}_1), D(P_2||\overline{P}_2)),$$

it follows that

$$\inf_R \{R + \eta(R) | \eta(R) < \eta(h)\} \ge h,$$

which implies that h is $\eta(h)$-achievable. Hence, it holds that

$$\lim_{r \downarrow 0} B_e(r|\mathbf{X}||\overline{\mathbf{X}}) = \min(D(P_1||\overline{P}_1), D(P_1||\overline{P}_2), D(P_2||\overline{P}_1), D(P_2||\overline{P}_2)). \tag{4.4.59}$$

Remark 4.4.3. In fact, we can generalize Example 4.4.4 and Example 4.4.5 in a much simpler way without computation of the information-spectrum. Let $\mathcal{X}$ be an arbitrary (not necessarily finite) alphabet and for four general sources $\mathbf{X}_1 = \{X_1^n\}_{n=1}^\infty$, $\mathbf{X}_2 = \{X_2^n\}_{n=1}^\infty$, $\overline{\mathbf{X}}_1 = \{\overline{X}_1^n\}_{n=1}^\infty$ and $\overline{\mathbf{X}}_2 = \{\overline{X}_2^n\}_{n=1}^\infty$ define $\mathbf{X} = \{X^n\}_{n=1}^\infty$ as the *mixed source* of $\mathbf{X}_1$ and $\mathbf{X}_2$ and $\overline{\mathbf{X}} = \{\overline{X}^n\}_{n=1}^\infty$ as the *mixed source* of $\overline{\mathbf{X}}_1$ and $\overline{\mathbf{X}}_2$ given by (4.4.49) and (4.4.50), respectively. Then, we have the following formula on the hypothesis testing $\mathbf{X}$ against $\overline{\mathbf{X}}$:

$$B_e(r|\mathbf{X}||\overline{\mathbf{X}}) = \min_{1\le i,j\le 2} B_e(r|\mathbf{X}_i||\overline{\mathbf{X}}_j) \quad (\forall r > 0). \tag{4.4.60}$$

This formula is established in the following way. First, we arbitrarily choose four rates R_{11}, R_{12}, R_{21} and R_{22} satisfying

$$R_{ij} < B_e(r|\mathbf{X}_i||\overline{\mathbf{X}}_j) \quad (\forall i, j = 1, 2). \tag{4.4.61}$$

Then, the definition of $B_e(r|\mathbf{X}_i||\overline{\mathbf{X}}_j)$ guarantees the existence of an acceptance region $\mathcal{A}_n^{(i,j)}$ of the hypothesis testing $\mathbf{X}_i$ against $\overline{\mathbf{X}}_j$ satisfying

$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\mu_n^{(i,j)}} \ge r, \tag{4.4.62}$$

$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\lambda_n^{(i,j)}} \ge R_{ij}, \tag{4.4.63}$$

where

$$\mu_n^{(i,j)} = P_{X_i^n}((\mathcal{A}_n^{(i,j)})^c), \quad \lambda_n^{(i,j)} = P_{\overline{X}_j^n}(\mathcal{A}_n^{(i,j)}) \quad (i, j = 1, 2). \tag{4.4.64}$$

We define the acceptance region $\mathcal{A}_n$ of the hypothesis testing $\mathbf{X}$ against $\overline{\mathbf{X}}$ by

$$\mathcal{A}_n = \left(\mathcal{A}_n^{(1,1)} \cap \mathcal{A}_n^{(1,2)}\right) \cup \left(\mathcal{A}_n^{(2,1)} \cap \mathcal{A}_n^{(2,2)}\right). \tag{4.4.65}$$

Since we have

$$\begin{aligned}
\mu_n \equiv P_{X^n}(\mathcal{A}_n^c) &= \alpha_1 P_{X_1^n}(\mathcal{A}_n^c) + \alpha_2 P_{X_2^n}(\mathcal{A}_n^c) \\
&\le \alpha_1 P_{X_1^n}((\mathcal{A}_n^{(1,1)} \cap \mathcal{A}_n^{(1,2)})^c) \\
&\quad + \alpha_2 P_{X_2^n}((\mathcal{A}_n^{(2,1)} \cap \mathcal{A}_n^{(2,2)})^c) \\
&\le \alpha_1 P_{X_1^n}((\mathcal{A}_n^{(1,1)})^c) + \alpha_1 P_{X_1^n}((\mathcal{A}_n^{(1,2)})^c) \\
&\quad + \alpha_2 P_{X_2^n}((\mathcal{A}_n^{(2,1)})^c) + \alpha_2 P_{X_2^n}((\mathcal{A}_n^{(2,2)})^c),
\end{aligned}$$

(4.4.62) and (4.4.64) guarantee

$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\mu_n} \ge r. \tag{4.4.66}$$

On the other hand, since

$$\begin{aligned}\lambda_n \equiv P_{\overline{X}^n}(\mathcal{A}_n) &= \beta_1 P_{\overline{X}_1^n}(\mathcal{A}_n) + \beta_2 P_{\overline{X}_2^n}(\mathcal{A}_n)\\ &\le \beta_1 P_{\overline{X}_1^n}(\mathcal{A}_n^{(1,1)} \cup \mathcal{A}_n^{(2,1)})\\ &\quad + \beta_2 P_{\overline{X}_2^n}(\mathcal{A}_n^{(1,2)} \cup \mathcal{A}_n^{(2,2)})\\ &\le \beta_1 P_{\overline{X}_1^n}(\mathcal{A}_n^{(1,1)}) + \beta_1 P_{\overline{X}_1^n}(\mathcal{A}_n^{(2,1)})\\ &\quad + \beta_2 P_{\overline{X}_2^n}(\mathcal{A}_n^{(1,2)}) + \beta_2 P_{\overline{X}_2^n}(\mathcal{A}_n^{(2,2)}),\end{aligned}$$

we obtain from (4.4.63) and (4.4.64) that

$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\lambda_n} \ge \min_{1\le i,j\le 2} R_{ij}. \tag{4.4.67}$$

By noticing that the rates R_{11}, R_{12}, R_{21} and R_{22} are arbitrary as far as they satisfy (4.4.61), (4.4.67) means that

$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\lambda_n} \ge \min_{1\le i,j\le 2} B_e(r|\mathbf{X}_i||\overline{\mathbf{X}}_j). \tag{4.4.68}$$

We can conclude from the combination of (4.4.66) and (4.4.68) that the right-hand side of (4.4.68) is r-achievable as a rate of the hypothesis testing $\mathbf{X}$ against $\overline{\mathbf{X}}$. That is, we have established the inequality

$$B_e(r|\mathbf{X}||\overline{\mathbf{X}}) \ge \min_{1\le i,j\le 2} B_e(r|\mathbf{X}_i||\overline{\mathbf{X}}_j) \tag{4.4.69}$$

meaning the direct part.

Next, to establish the inequality in the opposite direction, meaning the converse part, let R be an arbitrary r-achievable rate of the hypothesis testing $\mathbf{X}$ against $\overline{\mathbf{X}}$ and denote by $\mathcal{A}_n$ its corresponding acceptance region. From the definition, we have

$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\mu_n} \ge r, \tag{4.4.70}$$

$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\lambda_n} \ge R, \tag{4.4.71}$$

where

$$\mu_n = P_{X^n}(\mathcal{A}_n^c), \quad \lambda_n = P_{\overline{X}^n}(\mathcal{A}_n). \tag{4.4.72}$$

Now, consider the hypothesis testing $\mathbf{X}_i$ against $\overline{\mathbf{X}}_j$ with this $\mathcal{A}_n$ as an acceptance region and set

$$\mu_n^{(i,j)} = P_{X_i^n}(\mathcal{A}_n^c), \quad \lambda_n^{(i,j)} = P_{\overline{X}_j^n}(\mathcal{A}_n) \quad (i,j = 1,2). \tag{4.4.73}$$

Since

$$\begin{aligned}\mu_n &= \alpha_1 P_{X_1^n}(\mathcal{A}_n^c) + \alpha_2 P_{X_2^n}(\mathcal{A}_n^c)\\ &= \alpha_1 \mu_n^{(1,1)} + \alpha_2 \mu_n^{(2,1)},\end{aligned}$$

it follows that

$$\mu_n^{(1,1)} = \mu_n^{(1,2)} \le \frac{\mu_n}{\alpha_1},$$
$$\mu_n^{(2,1)} = \mu_n^{(2,2)} \le \frac{\mu_n}{\alpha_2}.$$

Therefore, (4.4.70) guarantees

$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\mu_n^{(i,j)}} \ge r \quad (\forall i, j = 1, 2). \tag{4.4.74}$$

In addition, by noticing that

$$\begin{aligned}\lambda_n &= \beta_1 P_{\overline{X}_1^n}(\mathcal{A}_n) + \beta_2 P_{\overline{X}_2^n}(\mathcal{A}_n)\\ &= \beta_1 \lambda_n^{(1,1)} + \beta_2 \lambda_n^{(1,2)},\end{aligned}$$

we have

$$\lambda_n^{(1,1)} = \lambda_n^{(2,1)} \le \frac{\lambda_n}{\beta_1},$$
$$\lambda_n^{(1,2)} = \lambda_n^{(2,2)} \le \frac{\lambda_n}{\beta_2}.$$

Hence, it holds from (4.4.71) that

$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\lambda_n^{(i,j)}} \ge R \quad (\forall i, j = 1, 2). \tag{4.4.75}$$

Equations (4.4.74) and (4.4.75) mean that R is r-achievable for all hypothesis testings $\mathbf{X}_i$ against $\overline{\mathbf{X}}_j$ $(i, j = 1, 2)$. Thus,

$$R \le \min_{1\le i,j\le 2} B_e(r|\mathbf{X}_i||\overline{\mathbf{X}}_j). \tag{4.4.76}$$

We note here that (4.4.76) implies

$$B_e(r|\mathbf{X}||\overline{\mathbf{X}}) \le \min_{1\le i,j\le 2} B_e(r|\mathbf{X}_i||\overline{\mathbf{X}}_j) \tag{4.4.77}$$

since R is an arbitrary r-achievable rate of the hypothesis testing $\mathbf{X}$ against $\overline{\mathbf{X}}$. Now (4.4.60) follows from the combination of (4.4.69) with (4.4.77).

Here, let us apply the formula (4.4.60) to Example 4.4.4 as a special case. Since $\overline{\mathbf{X}}_1 = \overline{\mathbf{X}}_2 = \overline{\mathbf{X}}$, (4.4.60) can be written as

$$B_e(r|\mathbf{X}||\overline{\mathbf{X}}) = \min(B_e(r|\mathbf{X}_1||\overline{\mathbf{X}}),\ B_e(r|\mathbf{X}_2||\overline{\mathbf{X}})).$$

By substituting the formula (4.4.24) in Example 4.4.1 into the right-hand side of the equation above, we obtain the following simple formula on the hypothesis testing for the mixed sources:

$$B_e(r|\mathbf{X}||\overline{\mathbf{X}}) = \min\left(\inf_{Q:D(Q||P_1)<r} D(Q||\overline{P}),\ \inf_{Q:D(Q||P_2)<r} D(Q||\overline{P})\right). \tag{4.4.78}$$

Similarly, if we consider the case of Example 4.4.5, (4.4.60), together with the formula (4.4.24) in Example 4.4.1, yields the following simple formula on the hypothesis testing for the mixed sources:

$$B_e(r|\mathbf{X}||\overline{\mathbf{X}}) = \min_{1\le i,j\le 2} \inf_{Q:D(Q||P_i)<r} D(Q||\overline{P}_j). \tag{4.4.79}$$

In addition, we also have the following formula of $B(\mathbf{X}||\overline{\mathbf{X}})$ (see Definition 4.1.1 and Definition 4.1.2):

$$\begin{aligned} B(\mathbf{X}||\overline{\mathbf{X}}) &= \min_{1\le i,j\le 2} B(\mathbf{X}_i||\overline{\mathbf{X}}_j) \\ &= \min_{1\le i,j\le 2} \underline{D}(\mathbf{X}_i||\overline{\mathbf{X}}_j), \end{aligned} \tag{4.4.80}$$

which can be verified by using the argument yielding the formula (4.4.60) (see Theorem 4.1.1 and Example 4.1.1). □

Example 4.4.6. In the *mixed hypothesis testing* given in Remark 4.4.3, suppose that $\mathbf{X}_1 = \{X_1^n\}_{n=1}^\infty, \mathbf{X}_2 = \{X_2^n\}_{n=1}^\infty$, $\overline{\mathbf{X}}_1 = \{\overline{X}_1^n\}_{n=1}^\infty$ and $\overline{\mathbf{X}}_2 = \{\overline{X}_2^n\}_{n=1}^\infty$ are first-order *stationary irreducible Markov sources* subject to transition probabilities $P_1(\cdot|\cdot)$, $P_2(\cdot|\cdot)$, $\overline{P}_1(\cdot|\cdot)$ and $\overline{P}_2(\cdot|\cdot)$, respectively (assume that $\mathcal{X}$ is a finite source alphabet). In this case, by substituting (4.4.28) in Example 4.4.2 into the formula (4.4.60) in Remark 4.4.3, we obtain the following formula on the mixed hypothesis testing $\mathbf{X}$ against $\overline{\mathbf{X}}$:

$$B_e(r|\mathbf{X}||\overline{\mathbf{X}}) = \min_{1\le i,j\le 2} \inf_{Q\in\mathcal{P}_0:D(Q||P_i|q)<r} D(Q||\overline{P}_j|q) \quad (\forall r > 0). \tag{4.4.81}$$

Example 4.4.7. Let us consider here the case that $\mathcal{X}$ is a *countably infinite* alphabet, say $\mathcal{X} = \{1, 2, \cdots\}$. Then, we can use *Cramér's Theorem* (cf. Dembo and Zeitouni [22]), which always holds, although Sanov's theorem used in Example 4.4.1 and Example 4.4.2 does not always hold. First, let $P = (p_1, p_2, \cdots)$ and $\overline{P} = (\overline{p}_1, \overline{p}_2, \cdots)$ be two arbitrary probability distributions over $\mathcal{X}$ and denote by X and $\overline{X}$ the random variables that are equal to k with probabilities p_k and $\overline{p}_k$ $(k = 1, 2, \cdots)$, respectively. Let $\mathbf{X} = \{X^n = (X_1, X_2, \cdots, X_n)\}_{n=1}^\infty$ and $\overline{\mathbf{X}} = \{\overline{X}^n = (\overline{X}_1, \overline{X}_2, \cdots, \overline{X}_n)\}_{n=1}^\infty$ be the stationary memoryless sources specified by X and $\overline{X}$, respectively. Since the divergence density rate can be written as

$$\frac{1}{n} \log \frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)} = \frac{1}{n} \sum_{i=1}^n \log \frac{P_{X_i}(X_i)}{P_{\overline{X}_i}(X_i)}, \tag{4.4.82}$$

$\eta(R)$ in (4.4.1) can be expressed as

$$\eta(R) = \inf_{x\le R} I(x), \tag{4.4.83}$$

where $I(x)$ denotes the large deviation rate function of (4.4.82). If we notice here that the *moment generating function* $M(\theta)$ of $\log \dfrac{P_X(X)}{P_{\overline{X}}(X)}$ is written as

$$
\begin{aligned}
M(\theta) &= \mathrm{E}e^{\theta \log \frac{P_X(X)}{P_{\overline{X}}(X)}} = \sum_{i=1}^{\infty} p_i e^{\theta \log \frac{p_i}{\overline{p}_i}} \\
&= \sum_{i=1}^{\infty} p_i^{1+\theta} \overline{p}_i^{-\theta}.
\end{aligned} \tag{4.4.84}
$$

Cramér's theorem tells that the rate function $I(x)$ is given by

$$
I(x) = \sup_{\theta} (\theta x - \Lambda(\theta)), \tag{4.4.85}
$$

where $\Lambda(\theta) = \log M(\theta)$ and $-\log M(\theta)$ is usually called *Chernoff's θ-distance* (cf. Blahut [11], Cover and Thomas [17]). Here, since the expectation of $\log \dfrac{P_X(X)}{P_{\overline{X}}(X)}$ is

$$
\mathrm{E}\left[\log \frac{P_X(X)}{P_{\overline{X}}(X)}\right] = \sum_{i=1}^{\infty} p_i \log \frac{p_i}{\overline{p}_i} \equiv D(P||\overline{P}) \quad \text{(divergence)},
$$

we notice from (4.4.83) that $\eta(R) = 0$ for $R \geq D(P||\overline{P})$ and $\eta(R) = I(R)$ for $R \leq D(P||\overline{P})$ ($I(x)$ is monotone increasing for $x \geq D(P||\overline{P})$, monotone decreasing for $x \leq D(P||\overline{P})$ and $I(x) = 0$ at $x = D(P||\overline{P})$). Therefore, we obtain the formula for computing values of $B_e(r|\mathbf{X}||\overline{\mathbf{X}})$ by substituting (4.4.83) into (4.4.2) in Theorem 4.4.1.

If (4.4.84) is substituted into (4.4.85) with $x = R$, we obtain

$$
I(R) = \sup_{\theta} (\theta R - \log \sum_{i=1}^{\infty} p_i^{1+\theta} \overline{p}_i^{-\theta}). \tag{4.4.86}
$$

Thus, we can compute $I(R)$ by using this equation. To this end, we differentiate the terms on the right-hand side with respect to θ and set it to 0. Then, we have the following equation with respect to θ:

$$
R = \frac{\displaystyle\sum_{i=1}^{\infty} p_i^{1+\theta} \overline{p}_i^{-\theta} \log \frac{p_i}{\overline{p}_i}}{\displaystyle\sum_{i=1}^{\infty} p_i^{1+\theta} \overline{p}_i^{-\theta}} \equiv \varphi(\theta). \tag{4.4.87}
$$

As far as $P \neq \overline{P}$ is satisfied, $\varphi(\theta)$ on the right-hand side turns out to be a continuous and strictly monotone increasing function of θ owing to the term by term differentiability of $M(\theta)$ (cf. Dembo and Zeitouni [22]), which is easily verified by using the Schwarz inequality (cf. Gallager [30]). If we define $\mathcal{D} = \{-\infty < \varphi(\theta) < +\infty \mid \theta\}$, then $\mathcal{D}$ forms an interval on the real line. Consequently, if $R \in \mathcal{D}$, $I(R)$ can be computed as

$$
I(R) = \theta R - \log \sum_{i=1}^{\infty} p_i^{1+\theta} \overline{p}_i^{-\theta}, \tag{4.4.88}
$$

where θ is determined by (4.4.87). In this case, denoting by $\mathcal{P}(\mathcal{X})$ the set of all probability distributions over $\mathcal{X}$ and Q_R the projection of P on the plane in $\mathcal{P}(\mathcal{X})$ defined by

$$\kappa_R = \left\{ Q \in \mathcal{P}(\mathcal{X}) \left| \sum_{i=1}^{\infty} Q(i) \log \frac{p_i}{\overline{p}_i} = R \right. \right\},$$

we can verify by direct computation that

$$I(R) = D(Q_R||P) \tag{4.4.89}$$

and

$$Q_R(i) = \frac{p_i^{1+\theta}\overline{p}_i^{-\theta}}{\sum_{i=1}^{\infty} p_i^{1+\theta}\overline{p}_i^{-\theta}} \quad (i \in \mathcal{X}) \tag{4.4.90}$$

with θ satisfying (4.4.87). That is, if $R \in \mathcal{D}$, then Cramér's theorem is reduced to Sanov's theorem as in Example 4.4.1 for a finite alphabet case. However, equality such as (4.4.89) does not hold for R satisfying $R \notin \mathcal{D}$. Therefore, it is important to know what kind of interval $\mathcal{D}$ forms. In particular, if

$$D(\overline{P}||P) < +\infty, \quad D(P||\overline{P}) < +\infty, \tag{4.4.91}$$

then

$$[-D(\overline{P}||P), D(P||\overline{P})] \subset \mathcal{D}.$$

Thus, in this case we obtain

$$B_e(r|\mathbf{X}||\overline{\mathbf{X}}) = \inf_{Q:D(Q||P)<r} D(Q||\overline{P}) \tag{4.4.92}$$

for $0 < r \leq D(\overline{P}||P)$ from Sanov's theorem in the same way as in Example 4.4.1 (see Fig. 4.5). This equation clearly holds for $r > D(\overline{P}||P)$; we

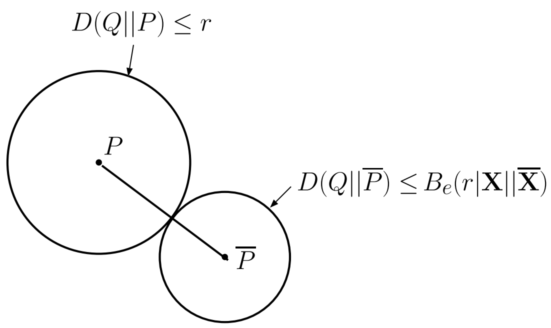

Fig. 4.5.

have $B_e(r|\mathbf{X}||\overline{\mathbf{X}}) = 0$ in this case. The formula (4.4.92) is regarded as an extension of Hoeffding's theorem in (4.4.24) for the case of a finite alphabet to the case with a countably infinite alphabet $\mathcal{X}$. In fact, the formula (4.4.92) holds in general with *any infinite alphabet* $\mathcal{X}$ that is not necessarily countably infinite under the condition (4.4.91). This fact follows from the fact that, if we rewrite $p_i^{1+\theta}\overline{p}_i^{-\theta}$, appearing in the proof above, in the equivalent form of $p_i\left(\frac{p_i}{\overline{p}_i}\right)^{\theta}$, $\frac{p_i}{\overline{p}_i}$ is well-defined as the Radon-Nikodym derivatives (cf. Billingsley [9]) for any infinite alphabet $\mathcal{X}$. We note here that the condition (4.4.91) is equivalent to the condition that the probability measure $\overline{P}$ is *absolutely continuous* with respect to the probability measure P and, conversely, P is absolutely continuous with respect to $\overline{P}$.

Cramér type equivalent of the formula (4.4.92) under the condition (4.4.91) is found in Dembo and Zeitouni [22] where the Neyman-Pearson lemma is directly invoked, while here Theorem 4.4.1 is invoked. □

Example 4.4.8. Let us consider the hypothesis testing for *autoregressive processes* given in Example 1.9.9 in §1.9. Let the null hypothesis $\mathbf{X} = \{X^n = (X_1, X_2, \ldots, X_n)\}_{n=1}^{\infty}$ and the alternative hypothesis $\overline{\mathbf{X}} = \{\overline{X}^n = (\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_n)\}$ be the autoregressive processes defined by

$$X_n = aX_{n-1} + W_n \quad (0 < a < 1;\ n = 1, 2, \cdots),$$
$$\overline{X}_n = a\overline{X}_{n-1} + \overline{W}_n \quad (0 < a < 1;\ n = 1, 2, \cdots),$$

respectively, where

$$\mathbf{W} = (W_1, W_2, \cdots) \quad (W^n = (W_1, \cdots, W_n)),$$
$$\overline{\mathbf{W}} = (\overline{W}_1, \overline{W}_2, \cdots) \quad (\overline{W}^n = (\overline{W}_1, \cdots, \overline{W}_n))$$

are the stationary memoryless sources subject to probability distributions P_W and $P_{\overline{W}}$ over the same alphabet $\mathcal{W}$, respectively, and we define $X_0 = \overline{X}_0 = 0$. The alphabet $\mathcal{W}$ can be any countably infinite set, continuous set or even any subset of real numbers. Then, since there is a one-to-one correspondence between the two divergence density rates

$$\frac{1}{n}\log\frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)}$$

and

$$\frac{1}{n}\log\frac{P_{W^n}(W^n)}{P_{\overline{W}^n}(W^n)} = \frac{1}{n}\sum_{i=1}^{n}\log\frac{P_{W_i}(W_i)}{P_{\overline{W}_i}(W_i)},$$

the information-spectrum of $\frac{1}{n}\log\frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)}$ coincides with the information-spectrum of $\frac{1}{n}\log\frac{P_{W^n}(W^n)}{P_{\overline{W}^n}(W^n)}$. Therefore, if

$$D(P_{\overline{W}}||P_W) < +\infty, \quad D(P_W||P_{\overline{W}}) < +\infty$$

are satisfied, from the formula (4.4.92) developed in Example 4.4.7 we obtain the following formula on the hypothesis testing for autoregressive sources:

$$B_e(r|\mathbf{X}||\overline{\mathbf{X}}) = \inf_{Q:D(Q||P_W)<r} D(Q||P_{\overline{W}}). \tag{4.4.93}$$

Example 4.4.9. Let us consider here the case that the null hypothesis $\mathbf{X}$ and the alternative hypothesis $\overline{\mathbf{X}}$ are the stationary memoryless sources subject to *Gaussian distributions* $N(\kappa, \sigma^2)$ and $N(\overline{\kappa}, \sigma^2)$, respectively. We denote the probability density functions by

$$P_\kappa(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\kappa)^2}{2\sigma^2}},$$
$$P_{\overline{\kappa}}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\overline{\kappa})^2}{2\sigma^2}}.$$

Denote by X the random variable subject to the probability density function P_κ. Since the moment generating function $M(\theta) = \mathrm{E}(e^{\theta Y})$ of

$$Y = \log \frac{P_\kappa(X)}{P_{\overline{\kappa}}(X)} \tag{4.4.94}$$

is computed as

$$M(\theta) = e^{\frac{(\kappa-\overline{\kappa})^2(\theta+\theta^2)}{2\sigma^2}},$$

we have

$$\theta x - \log M(\theta) = \theta x - \frac{(\kappa-\overline{\kappa})^2(\theta+\theta^2)}{2\sigma^2}.$$

Then, simple computation tells us that the large deviation rate function $I(x)$ for (4.4.94) can be written as

$$I(x) = \sup_\theta(\theta x - \log M(\theta)) = \frac{\sigma^2(x-a)^2}{2(\kappa-\overline{\kappa})^2}, \tag{4.4.95}$$

where we set $a = \dfrac{(\kappa-\overline{\kappa})^2}{2\sigma^2}$ for simplicity. In addition, we note that $D(P_\kappa||P_{\overline{\kappa}})$ $= a$. Now, Cramér's theorem implies that $\eta(R)$ in Theorem 4.4.1 can be computed as

$$\begin{aligned}\eta(R) &= \inf_{x\le R} I(x) = \min_{x\le R} \frac{\sigma^2(x-a)^2}{2(\kappa-\overline{\kappa})^2} \\ &= \min\left\{[a-R]^+, \frac{\sigma^2(R-a)^2}{2(\kappa-\overline{\kappa})^2}\right\}.\end{aligned} \tag{4.4.96}$$

Furthermore, since

$$
\begin{aligned}
R+\eta(R) &= \min\left\{R+[a-R]^+, R+\frac{\sigma^2(R-a)^2}{2(\kappa-\overline{\kappa})^2}\right\} \\
&= \min\left\{R+[a-R]^+, \frac{\sigma^2(R+a)^2}{2(\kappa-\overline{\kappa})^2}\right\},
\end{aligned}
\tag{4.4.97}
$$

the substitution of (4.4.96) and (4.4.97) into the right-hand side of (4.4.2) in Theorem 4.4.1 and a little computation lead to

$$
\begin{aligned}
B_e(r|\mathbf{X}||\overline{\mathbf{X}}) &= \min\left\{[a-r]^+, (\sqrt{r}-\sqrt{a})^2\right\} \\
&= (\sqrt{r}-\sqrt{a})^2 \mathbf{1}[r \le a],
\end{aligned}
$$

where $\mathbf{1}[\,\cdot\,]$ denotes the characteristic function (Fig 4.6). This formula tells

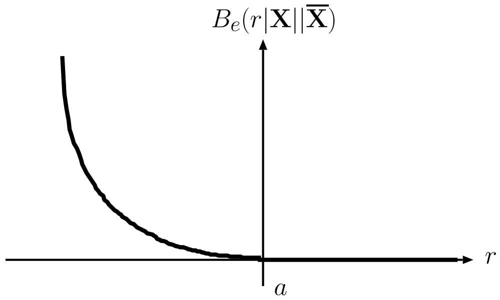

Fig. 4.6.

us that $B_e(r|\mathbf{X}||\overline{\mathbf{X}})$ is monotone decreasing for $0 < r < a$. It also tells us that $B_e(0|\mathbf{X}||\overline{\mathbf{X}}) = a = D(P_\kappa||P_{\overline{\kappa}})$ and $B_e(r|\mathbf{X}||\overline{\mathbf{X}}) = 0$ for $r \ge a$. □

Example 4.4.10. In all the examples that we have given so far, the function $\eta(R)$ is continuous in R. However, we can construct an example where $\eta(R)$ is *not* continuous in the following way. Let the source alphabet be $\mathcal{X} = \{0,1\}$ and fix a subset S_n of $\mathcal{X}^n$ satisfying $|S_n| = 2^{\alpha n}$ arbitrarily, where α is a constant satisfying $0 < \alpha < 1$. We also arbitrarily fix $\mathbf{x}_0$ and $\mathbf{x}_1$ satisfying $\mathbf{x}_0, \mathbf{x}_1 \in \mathcal{X}^n - S_n$ and $\mathbf{x}_0 \ne \mathbf{x}_1$. Now, we define the null hypothesis $\mathbf{X} = \{X^n\}_{n=1}^\infty$ as

$$
P_{X^n}(\mathbf{x}) = \begin{cases}
2^{-2\alpha n} & \text{for} \quad \mathbf{x} \in S_n, \\
2^{-3\alpha n} & \text{for} \quad \mathbf{x} = \mathbf{x}_1, \\
1-2^{-\alpha n}-2^{-3\alpha n} & \text{for} \quad \mathbf{x} = \mathbf{x}_0, \\
0 & \text{for } \mathbf{x} \notin S_n \cup \{\mathbf{x}_1, \mathbf{x}_0\},
\end{cases}
\tag{4.4.98}
$$

which clearly satisfies $P_{X^n}(S_n) = 2^{-\alpha n}$. We define the alternative hypothesis $\overline{\mathbf{X}} = \{\overline{X}^n\}_{n=1}^\infty$ as $P_{\overline{X}^n}(\mathbf{x}) = 2^{-n}$ $(\forall \mathbf{x} \in \mathcal{X}^n)$. Then, simple computation tells us that the divergence-spectrum becomes the three-point spectrum with three peaks of probabilities $1-2^{-\alpha n}-2^{-3\alpha n}$, $2^{-\alpha n}$ and $2^{-3\alpha n}$ at $1+\frac{1}{n}\log(1-2^{-\alpha n}-2^{-3\alpha n})$, $1-2\alpha$ and $1-3\alpha$, respectively. Thus, $\eta(R)$ is computed from its definition as follows:

$$\eta(R) = \begin{cases} +\infty & \text{for} \quad R < 1-3\alpha, \\ 3\alpha & \text{for } 1-3\alpha \le R < 1-2\alpha, \\ \alpha & \text{for} \quad 1-2\alpha \le R < 1, \\ 0 & \text{for} \quad 1 \le R. \end{cases} \tag{4.4.99}$$

Then, $R+\eta(R)$ is expressed as

$$R+\eta(R) = \begin{cases} +\infty & \text{for} \quad R < 1-3\alpha, \\ R+3\alpha & \text{for } 1-3\alpha \le R < 1-2\alpha, \\ R+\alpha & \text{for} \quad 1-2\alpha \le R < 1, \\ R & \text{for} \quad 1 \le R. \end{cases} \tag{4.4.100}$$

By using Theorem 4.4.1, we obtain the following formula:

$$B_e(r|\mathbf{X}||\overline{\mathbf{X}}) = \begin{cases} 1-\alpha & \text{for} \quad r > \alpha, \\ 1 & \text{for } 0 < r \le \alpha. \end{cases} \tag{4.4.101}$$

We note here that, if $r > \alpha$, $\inf_R$ on the right-hand side of (4.4.2) is attained by $R = R^\circ \equiv 1-2\alpha$ as

$$\begin{aligned} \inf_R \{R+\eta(R) \mid \eta(R) < r\} &= R^\circ + \eta(R^\circ) \quad (R^\circ \equiv 1-2\alpha) \\ &= 1-\alpha. \end{aligned}$$

In particular, if $r > 3\alpha$, $\inf_R$ is not attained by the boundary point $\underline{R} \equiv \inf\{R|\eta(R) < r\} = 1-3\alpha$ of $\{R|\eta(R) < r\}$, but attained by the internal point $R = R^\circ \equiv 1-2\alpha$. This kind of phenomenon only occurs in the hypothesis testing treating general sources that do not satisfy the consistency condition, i.e., such a phenomenon never occurs as far as we treat ordinary sources given in the preceding examples. □

4.5 Hypothesis Testing and Large Deviation: Probability of Correct Testing

In §4.4 we have considered the large deviation behavior of the error probability of the second kind λ_n subject to the constraint that the error probability of the first kind μ_n asymptotically satisfies $\mu_n \simeq e^{-nr}$ for a constant $r > 0$. However, λ_n comes to satisfy $\lambda_n \simeq 1$ if $r > 0$ is sufficiently large (see Example 4.4.1). In this kind of situation it is important to investigate the large deviation behavior of $1-\lambda_n$, the probability of correct testing against the alternative hypothesis, instead of the error probability λ_n itself. The problem to be considered is how the exponent $R > 0$ can be small when $1-\lambda_n$ is expressed as $1 - \lambda_n \simeq e^{-nR}$. This section is devoted to analysis of this problem. We begin with formulation of this problem. The null hypothesis and the alternative hypothesis are denoted by $\mathbf{X} = \{X^n\}_{n=1}^{\infty}$ and $\overline{\mathbf{X}} = \{\overline{X}^n\}_{n=1}^{\infty}$ in this section as well.

Definition 4.5.1.

$$\text{Rate } R \text{ is } r\text{-achievable} \overset{\text{def}}{\Longleftrightarrow} \begin{array}{l}\text{There exists an acceptance region } \mathcal{A}_n \\ \text{satisfying } \liminf_{n\to\infty} \dfrac{1}{n} \log \dfrac{1}{\mu_n} \geq r \text{ and} \\ \limsup_{n\to\infty} \dfrac{1}{n} \log \dfrac{1}{1-\lambda_n} \leq R.\end{array}$$

Definition 4.5.2 (infimum r-achievable correct probability exponent).

$$B_e^*(r|\mathbf{X}\|\overline{\mathbf{X}}) = \inf\left\{R \mid R \text{ is } r\text{-achievable}\right\}.$$

The objective of this section is determination of this $B_e^*(r|\mathbf{X}\|\overline{\mathbf{X}})$ as a (left-continuous and monotone increasing) function of r. To this end, let us define a function $\eta(R)$ by

$$\eta(R) = \lim_{n\to\infty} \frac{1}{n} \log \frac{1}{\Pr\left\{\dfrac{1}{n}\log\dfrac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)} \leq R\right\}}. \tag{4.5.1}$$

This function is in the same form as $\eta(R)$ in (4.4.1) in the preceding section. However, $\eta(R)$ in (4.5.1) is different from $\eta(R)$ in (4.4.1) in the sense that the existence of the limit on the right-hand side of (4.5.1) is assumed. We also note that Lemma 4.4.1 implies that $\eta(R)$ is monotone decreasing in R and $\eta(R) = 0$ for $R > \underline{D}(\mathbf{X}\|\overline{\mathbf{X}})$.

We give one assumption on the information-spectrum. That is, we assume that for any constant $M > 0$ there exists some sufficiently large constant $K > 0$ satisfying

$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\Pr\left\{\dfrac{1}{n}\log\dfrac{P_{\overline{X}^n}(\overline{X}^n)}{P_{X^n}(\overline{X}^n)} \geq K\right\}} \geq M. \tag{4.5.2}$$

Remark 4.5.1. This assumption means that the information-spectrum of $\overline{\mathbf{X}}$ with respect to $\mathbf{X}$ does not shift to the right more than some specified speed as n increases. For example, if $\mathbf{X}$ and $\overline{\mathbf{X}}$ are the stationary memoryless sources subject to probability distributions P_X and $P_{\overline{X}}$ over a finite alphabet $\mathcal{X}$, respectively, and there is no $x \in \mathcal{X}$ satisfying $P_X(x) = 0$ and $P_{\overline{X}}(x) > 0$, the assumption (4.5.2) trivially holds. The assumption (4.5.2) holds for other stationary sources that we usually treat. □

We have the following theorem that is a dual counterpart of Theorem 4.4.1.

Theorem 4.5.1 (Han [38]). *Assume that (4.5.2) holds. Then, for any $r \geq 0$*

$$B_e^*(r|\mathbf{X}||\overline{\mathbf{X}}) = \inf_R \left\{ R + \eta(R) + [r - \eta(R)]^+ \right\}, \tag{4.5.3}$$

where $B_e^*(0|\mathbf{X}||\overline{\mathbf{X}})$ $(r = 0)$ is defined as 0.

Remark 4.5.2. Lemma 4.4.1 implies that

$$\inf_{R > \underline{D}(\mathbf{X}||\overline{\mathbf{X}})} \left\{ R + \eta(R) + [r - \eta(R)]^+ \right\} = \inf_{R > \underline{D}(\mathbf{X}||\overline{\mathbf{X}})} (R + r),$$

where the infimum on the right-hand side is attained at $R = \underline{D}(\mathbf{X}||\overline{\mathbf{X}})$. Hence, $\inf_R$ on the right-hand side of (4.5.3) can be replaced with $\inf_{R \leq \underline{D}(\mathbf{X}||\overline{\mathbf{X}})}$ if $\eta(R)$ is continuous at $R = \underline{D}(\mathbf{X}||\overline{\mathbf{X}})$. Recently, another general expression for $B_e^*(r|\mathbf{X}||\overline{\mathbf{X}})$ was given by Nagaoka and Hayashi [69]. □

Proof of Theorem 4.5.1.
1) *Direct part:*
In the proof of the direct part we do not need the assumption (4.5.2). First, keep in mind that $\eta(R)$ in

$$R + \eta(R) + [r - \eta(R)]^+$$

on the right-hand side of (4.5.3) is monotone decreasing, and set

$$\rho_0^* = \inf_R \left\{ R + \eta(R) + [r - \eta(R)]^+ \right\}. \tag{4.5.4}$$

Then, there exists an R_0 such that ρ_0^* is expressed as

$$\rho_0^* = \lim_{\varepsilon \downarrow 0} (R_0 + \varepsilon + \eta(R_0 + \varepsilon) + [r - \eta(R_0 + \varepsilon)]^+), \tag{4.5.5}$$

which we rewrite as

$$\rho_0^* = R_0 + \gamma + \eta(R_0 + \gamma) + [r - \eta(R_0 + \gamma)]^+ - \nu(\gamma), \tag{4.5.6}$$

where $\gamma > 0$ is an arbitrarily small constant and $\nu(\gamma) \to 0$ as $\gamma \to 0$. We use here the notation that

$$S_n^*(a) = \left\{ \mathbf{x} \in \mathcal{X}^n \,\middle|\, \frac{1}{n} \log \frac{P_{X^n}(\mathbf{x})}{P_{\overline{X}^n}(\mathbf{x})} \leq a \right\}. \tag{4.5.7}$$

Then, since the existence of the limit in (4.5.1) is assumed, we have

$$e^{-n(\eta(R_0+\gamma)+\tau)} \leq \Pr\{X^n \in S_n^*(R_0 + \gamma)\} \leq e^{-n(\eta(R_0+\gamma)-\tau)} \quad (\forall n \geq n_0), \tag{4.5.8}$$

where $\tau > 0$ is an arbitrarily small constant. Next, define a subset $\mathcal{C}_n$ of $S_n^*(R_0+\gamma)$ as follows; if $\eta(R_0+\gamma) \geq r$ then set $\mathcal{C}_n = S_n^*(R_0+\gamma)$, otherwise set $\mathcal{C}_n = T_n$, where T_n is an arbitrary subset of $S_n^*(R_0+\gamma)$ satisfying

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{\Pr\{X^n \in T_n\}} = r. \tag{4.5.9}$$

It should be noted here that it is always possible to choose such a subset T_n, because in the case with $\eta(R_0+\gamma) < r$ we can make $\eta(R_0+\gamma)+\tau < r$ hold with a $\tau > 0$ small enough, where we may consider a *randomized* hypothesis testing if necessary. Now, consider the hypothesis testing with $\mathcal{C}_n$ as the critical region. First, we evaluate the error probability of the first kind μ_n. In the case with $\eta(R_0+\gamma) \geq r$, since $\mathcal{C}_n = S_n^*(R_0+\gamma)$, by means of (4.5.8) we have

$$\begin{aligned}\Pr\{X^n \in \mathcal{C}_n\} &\leq e^{-n(\eta(R_0+\gamma)-\tau)}\\ &\leq e^{-n(r-\tau)} \quad (\forall n \geq n_0),\end{aligned}$$

while in the case where $\eta(R_0+\gamma) < r$, by means of (4.5.9) we have

$$\Pr\{X^n \in \mathcal{C}_n\} \leq e^{-n(r-\tau)} \quad (\forall n \geq n_0).$$

Then, in either case, it holds that

$$\Pr\{X^n \in \mathcal{C}_n\} \leq e^{-n(r-\tau)}. \tag{4.5.10}$$

Therefore, the error probability of the first kind μ_n is evaluated as

$$\mu_n \equiv \Pr\{X^n \in \mathcal{C}_n\} \leq e^{-n(r-\tau)}.$$

Hence,

$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\mu_n} \geq r - \tau.$$

Since $\tau > 0$ is arbitrary, we can conclude that

$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\mu_n} \geq r. \tag{4.5.11}$$

Next, we evaluate $1-\lambda_n$, the probability of correct testing, where λ_n is the error probability of the second kind. First, we observe that if $\mathbf{x} \in S_n^*(R_0+\gamma)$ then

$$P_{\overline{X}^n}(\mathbf{x}) \geq P_{X^n}(\mathbf{x}) e^{-n(R_0+\gamma)} \tag{4.5.12}$$

holds. Then, in the case of $\eta(R_0+\gamma) \geq r$, since $\mathcal{C}_n = S_n^*(R_0+\gamma)$, it follows from (4.5.8) that

$$\begin{aligned}\Pr\left\{\overline{X}^n \in \mathcal{C}_n\right\} &= \sum_{\mathbf{x}\in\mathcal{C}_n} P_{\overline{X}^n}(\mathbf{x})\\ &\geq \sum_{\mathbf{x}\in\mathcal{C}_n} P_{X^n}(\mathbf{x}) e^{-n(R_0+\gamma)}\\ &= e^{-n(R_0+\gamma)} \Pr\{X^n \in S_n^*(R_0+\gamma)\}\\ &\geq e^{-n(R_0+\gamma+\eta(R_0+\gamma)+\tau)} \quad (\forall n \geq n_0).\end{aligned} \tag{4.5.13}$$

Similarly, in the case of $\eta(R_0+\gamma) < r$, since $\mathcal{C}_n = T_n$, it follows from (4.5.9) that

$$\Pr\left\{\overline{X}^n \in \mathcal{C}_n\right\} \geq e^{-n(R_0+\gamma+r+\tau)} \quad (\forall n \geq n_0). \tag{4.5.14}$$

Summarizing (4.5.13) and (4.5.14), in either case we have

$$\Pr\left\{\overline{X}^n \in \mathcal{C}_n\right\} \geq e^{-n(R_0+\gamma+\eta(R_0+\gamma)+[r-\eta(R_0+\gamma)]^+ +\tau)}. \tag{4.5.15}$$

Substitution of (4.5.6) into (4.5.15) yields

$$\Pr\left\{\overline{X}^n \in \mathcal{C}_n\right\} \geq e^{-n(\rho_0^*+\tau+\nu(\gamma))}.$$

Hence,

$$\begin{aligned} 1-\lambda_n &= \Pr\left\{\overline{X}^n \in \mathcal{C}_n\right\} \\ &\geq e^{-n(\rho_0^*+\tau+\nu(\gamma))}, \end{aligned}$$

from which it follows that

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{1}{1-\lambda_n} \leq \rho_0^* + \tau + \nu(\gamma). \tag{4.5.16}$$

We notice here that we can make $\tau + \nu(\gamma) \to 0$, because $\tau > 0$ and $\gamma > 0$ are both made arbitrarily small. Thus, by virtue of (4.5.11) and (4.5.16) we can conclude that any rate R satisfying $R > \rho_0^*$ is r-achievable.

2) *Converse part:*

In the proof of the converse part we need the assumption (4.5.2). First, let $K > 0$ be a large enough constant (to be specified below) and $\gamma > 0$ be an arbitrarily small constant. Putting $L = \frac{2K}{\gamma}$, we divide the interval $(-K, K]$ into L subintervals with equal width γ to have

$$I_i = (c_i - \gamma, c_i] \quad (i = 1, 2, \cdots, L),$$

where $c_i \equiv K - (i-1)\gamma$. According to this interval partition, divide the set

$$T_n^* = \left\{ \mathbf{x} \in \mathcal{X}^n \,\middle|\, -K < \frac{1}{n} \log \frac{P_{X^n}(\mathbf{x})}{P_{\overline{X}^n}(\mathbf{x})} \leq K \right\}$$

into the L subsets

$$S_n^{(i)} = \left\{ \mathbf{x} \in \mathcal{X}^n \,\middle|\, \frac{1}{n} \log \frac{P_{X^n}(\mathbf{x})}{P_{\overline{X}^n}(\mathbf{x})} \in I_i \right\} \quad (i = 1, 2, \cdots, L).$$

This operation is called the *information-spectrum slicing.* Moreover, we define

$$\begin{aligned} S_n^{(0)} &= \left\{ \mathbf{x} \in \mathcal{X}^n \,\middle|\, \frac{1}{n} \log \frac{P_{X^n}(\mathbf{x})}{P_{\overline{X}^n}(\mathbf{x})} \leq -K \right\}, \\ S_n^{(-1)} &= \left\{ \mathbf{x} \in \mathcal{X}^n \,\middle|\, \frac{1}{n} \log \frac{P_{X^n}(\mathbf{x})}{P_{\overline{X}^n}(\mathbf{x})} > K \right\}, \end{aligned}$$

where it is obvious that $\mathcal{X}^n = \bigcup_{j=-1}^{L} S_n^{(j)}$. Suppose that R is r-achievable, i.e., suppose that there exists a critical region $\mathcal{C}_n$ such that

$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\mu_n} \geq r, \tag{4.5.17}$$

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{1}{1-\lambda_n} \leq R. \tag{4.5.18}$$

Then, from (4.5.17) we have

$$\mu_n \leq e^{-n(r-\tau)} \quad (\forall n \geq n_0), \tag{4.5.19}$$

where $\tau > 0$ is an arbitrarily small constant. In order to evaluate the value of $\Pr\left\{\overline{X}^n \in \mathcal{C}_n\right\}$, let us first evaluate the value of

$$\Pr\left\{X^n \in \mathcal{C}_n^{(i)}\right\} \quad (i = 1, 2, \cdots, L),$$

where $\mathcal{C}_n^{(i)} \equiv S_n^{(i)} \cap \mathcal{C}_n$ $(i = -1, 0, 1, 2, \cdots, L)$. We now evaluate $\Pr\left\{X^n \in \mathcal{C}_n^{(i)}\right\}$ $(i = 1, 2, \cdots, L)$ in two ways as follows. First, we observe that

$$\Pr\left\{X^n \in \mathcal{C}_n^{(i)}\right\} \leq \Pr\left\{X^n \in \mathcal{C}_n\right\} = \mu_n,$$

which, together with (4.5.19), yields

$$\Pr\left\{X^n \in \mathcal{C}_n^{(i)}\right\} \leq e^{-n(r-\tau)}. \tag{4.5.20}$$

Next, by the definitions of $\eta(c_i)$ and $S_n^{(i)}$, we see that

$$\begin{aligned}\Pr\{X^n \in S_n^{(i)}\} &\leq \Pr\left\{\frac{1}{n} \log \frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)} \leq c_i\right\} \\ &\leq e^{-n(\eta(c_i)-\tau)} \quad (\forall n \geq n_0).\end{aligned}$$

Hence,

$$\begin{aligned}\Pr\left\{X^n \in \mathcal{C}_n^{(i)}\right\} &\leq \Pr\left\{X^n \in S_n^{(i)}\right\} \\ &\leq e^{-n(\eta(c_i)-\tau)}.\end{aligned} \tag{4.5.21}$$

A consequence of (4.5.20) and (4.5.21) is

$$\Pr\left\{X^n \in \mathcal{C}_n^{(i)}\right\} \leq e^{-n(\eta(c_i)+[r-\eta(c_i)]^+-\tau)} \quad (i = 1, 2, \cdots, L). \tag{4.5.22}$$

We can now evaluate $\Pr\left\{\overline{X}^n \in \mathcal{C}_n^{(i)}\right\}$ as follows. Since $\mathbf{x} \in \mathcal{C}_n^{(i)}$ implies $\mathbf{x} \in S_n^{(i)}$ $(i = 1, 2, \cdots, L)$ and hence also $P_{\overline{X}^n}(\mathbf{x}) \leq P_{X^n}(\mathbf{x})e^{-n(c_i-\gamma)}$, we have

$$\begin{aligned}
\Pr\left\{\overline{X}^n \in \mathcal{C}_n^{(i)}\right\} &= \sum_{\mathbf{x}\in\mathcal{C}_n^{(i)}} P_{\overline{X}^n}(\mathbf{x}) \\
&\le \sum_{\mathbf{x}\in\mathcal{C}_n^{(i)}} P_{X^n}(\mathbf{x}) e^{-n(c_i-\gamma)} \\
&= e^{-n(c_i-\gamma)} \Pr\left\{X^n \in \mathcal{C}_n^{(i)}\right\} \\
&\le e^{-n(c_i+\eta(c_i)+[r-\eta(c_i)]^+-\gamma-\tau)} \qquad (4.5.23)
\end{aligned}$$

for $i = 1, 2, \cdots, L$, where we have used (4.5.22) in the last inequality. Furthermore, let us evaluate $\Pr\left\{\overline{X}^n \in S_n^{(-1)}\right\}$ and $\Pr\left\{\overline{X}^n \in S_n^{(0)}\right\}$. Since $\mathbf{x} \in S_n^{(-1)}$ implies $P_{\overline{X}^n}(\mathbf{x}) \le P_{X^n}(\mathbf{x})e^{-nK}$, we obtain

$$\begin{aligned}
\Pr\left\{\overline{X}^n \in S_n^{(-1)}\right\} &= \sum_{\mathbf{x}\in S_n^{(-1)}} P_{\overline{X}^n}(\mathbf{x}) \\
&\le \sum_{\mathbf{x}\in S_n^{(-1)}} P_{X^n}(\mathbf{x}) e^{-nK} \\
&\le e^{-nK}. \qquad (4.5.24)
\end{aligned}$$

Recalling here that

$$\begin{aligned}
\Pr\left\{\overline{X}^n \in S_n^{(0)}\right\} &= \Pr\left\{\frac{1}{n}\log\frac{P_{X^n}(\overline{X}^n)}{P_{\overline{X}^n}(\overline{X}^n)} \le -K\right\} \\
&= \Pr\left\{\frac{1}{n}\log\frac{P_{\overline{X}^n}(\overline{X}^n)}{P_{X^n}(\overline{X}^n)} \ge K\right\}
\end{aligned}$$

and noting the assumption (4.5.2), we see that for any $M > 0$ there exists a $K > 0$ large enough such that

$$\Pr\left\{\overline{X}^n \in S_n^{(0)}\right\} \le e^{-n(M-\tau)} \quad (\forall n \ge n_0). \qquad (4.5.25)$$

Summarizing (4.5.23)–(4.5.25), we have

$$\begin{aligned}
&1 - \lambda_n \\
&= \Pr\left\{\overline{X}^n \in \mathcal{C}_n\right\} = \sum_{i=-1}^{L} \Pr\left\{\overline{X}^n \in \mathcal{C}_n^{(i)}\right\} \\
&\le \sum_{i=1}^{L} e^{-n(c_i+\eta(c_i)+[r-\eta(c_i)]^+-\gamma-\tau)} + e^{-nK} + e^{-n(M-\tau)}. \qquad (4.5.26)
\end{aligned}$$

On the other hand, since, by the definition (4.5.4) of ρ_0^*,

$$c_i + \eta(c_i) + [r - \eta(c_i)]^+ \ge \rho_0^* \quad (i = 1, 2, \cdots, L),$$

it follows from (4.5.26) that

$$1 - \lambda_n \le L e^{-n(\rho_0^* - \gamma - \tau)} + e^{-nK} + e^{-n(M-\tau)}.$$

Thus, if we take $M > 0$ and $K > 0$ large enough, then

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{1}{1-\lambda_n} \ge \rho_0^* - \gamma - \tau. \tag{4.5.27}$$

Therefore, $R \ge \rho_0^* - \gamma - \tau$ holds, owing to (4.5.18) and (4.5.27). Since both of $\gamma > 0$ and $\tau > 0$ are arbitrary, we can let $\gamma \to 0$ and $\tau \to 0$ to obtain $R \ge \rho_0^*$. Thus, we can conclude that any r-achievable rate R cannot be smaller than ρ_0^*. □

Example 4.5.1. Suppose that $\mathcal{X}$ is a finite source alphabet. Let the null hypothesis $\mathbf{X}$ and the alternative hypothesis $\overline{\mathbf{X}}$ be the *stationary memoryless sources* subject to probability distributions P and $\overline{P}$, respectively. Here, for simplicity, we assume that $P(x) > 0$ for all $x \in \mathcal{X}$. which is the case that the assumption (4.5.2) is satisfied. As is shown in Example 4.4.1, in this setting $\eta(R) = 0$ for $R \ge D(P||\overline{P})$ and

$$\begin{aligned} \eta(R) &= D(P_R||P), \\ R + \eta(R) &= D(P_R||\overline{P}) \end{aligned}$$

for $R \le D(P||\overline{P})$, where P_R denotes the *projection* of P on the plane κ_R defined in (4.4.22). We note here that, since $\underline{D}(\mathbf{X}||\overline{\mathbf{X}}) = D(P||\overline{P})$, it suffices to consider R satisfying $R \le D(P||\overline{P})$ in (4.5.3) of Theorem 4.5.1 (see Remark 4.5.2). Then, since we have

$$R + \eta(R) + [r - \eta(R)]^+ = D(P_R||\overline{P}) + [r - D(P_R||P)]^+,$$

Theorem 4.5.1 yields

$$B_e^*(r|\mathbf{X}||\overline{\mathbf{X}}) = \inf_R \left\{ D(P_R||\overline{P}) + [r - D(P_R||P)]^+ \right\}. \tag{4.5.28}$$

This formula indicates that $B_e^*(r|\mathbf{X}||\overline{\mathbf{X}})$ is a monotone increasing function of r. In addition, while $B_e^*(r|\mathbf{X}||\overline{\mathbf{X}}) = 0$ for the case of $r \le D(\overline{P}||P)$, $B_e^*(r|\mathbf{X}||\overline{\mathbf{X}})$ can be expressed as

$$\begin{aligned} B_e^*(r|\mathbf{X}||\overline{\mathbf{X}}) &= \inf_{R: D(P_R||P) \le r} \left\{ D(P_R||\overline{P}) + r - D(P_R||P) \right\} \\ &= \inf_{Q: D(Q||P) \le r} \left\{ D(Q||\overline{P}) + r - D(Q||P) \right\} \end{aligned} \tag{4.5.29}$$

for the case of $r \ge D(\overline{P}||P)$ because it is easily verified that $\inf_R$ on the right-hand side of (4.5.28) is achieved by R satisfying $D(P_R||P) \le r$ (Fig. 4.7). This formula is the same as the formula first developed by Han and Kobayashi [43] based on the argument of types. □

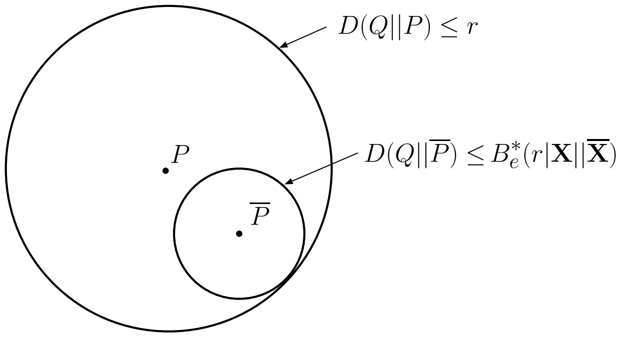

Fig. 4.7.

Example 4.5.2. Let us consider the hypothesis testing for first-order *stationary irreducible Markov sources* $\mathbf{X}$ and $\overline{\mathbf{X}}$ with a finite alphabet that are considered in Example 4.4.2. We use the notation that appeared in Example 4.4.2. From (4.4.25) and (4.4.26), we obtain $\eta(R) = 0$ for $R \geq D(P||\overline{P}|p)$ and

$$\eta(R) = D(P_R||P|p_R),$$
$$R + \eta(R) = D(P_R||\overline{P}|p_R),$$

for $R \leq D(P||\overline{P}|p)$. Then, Theorem 4.5.1 yields

$$B_e^*(r|\mathbf{X}||\overline{\mathbf{X}}) = \inf_R \left\{ D(P_R||\overline{P}|p_R) + [r - D(P_R||P|p_R)]^+ \right\}.$$

Since it is easy to verify that, if $r \geq D(\overline{P}||P|\overline{p})$, $\inf_R$ is attained by R satisfying

$$D(P_R||P|p_R) \leq r,$$

we obtain

$$\begin{aligned} B_e^*(r|\mathbf{X}||\overline{\mathbf{X}}) &= \inf_{R: D(P_R||P|p_R) \leq r} \left\{ D(P_R||\overline{P}|p_R) + r - D(P_R||P|p_R) \right\} \\ &= \inf_{Q \in \mathcal{P}_0 : D(Q||P|q) \leq r} \left\{ D(Q||\overline{P}|q) + r - D(Q||P|q) \right\}, \quad (4.5.30) \end{aligned}$$

where $\overline{p}$ denotes the stationary distribution of $\overline{P}$ (cf. Nakagawa and Kanaya [71]). In addition, we can check that $B_e^*(r|\mathbf{X}||\overline{\mathbf{X}}) = 0$ for all $r \leq D(\overline{P}||P|\overline{p})$. □

Example 4.5.3. In order to generalize Example 4.5.2 above, let us consider the hypothesis testing with *unifilar finite-state sources* $\mathbf{X}$ and $\overline{\mathbf{X}}$ given in Example 4.4.3 in §4.4. We use the same notations used in Example 4.4.3. Since (4.4.35) and (4.4.36) hold from Sanov's Theorem on unifilar finite-state sources, Theorem 4.5.1 leads to the following formula of $B_e^*(r|\mathbf{X}||\overline{\mathbf{X}})$ for the hypothesis testing $\mathbf{X}$ against $\overline{\mathbf{X}}$:

$$
\begin{aligned}
&B_e^*(r|\mathbf{X}||\overline{\mathbf{X}}) \\
&= \inf_R \left\{ D(P_{X_R S_R}||\overline{P}|P_{S_R}) + [r - D\left(P_{X_R S_R}||P|P_{S_R}\right)]^+ \right\} \\
&= \inf_{P_{XS}\in\mathcal{V}_0} \left\{ D(P_{XS}||\overline{P}|P_S) + [r - D\left(P_{XS}||P|P_S\right)]^+ \right\}. \qquad (4.5.31)
\end{aligned}
$$

Example 4.5.4. Let us consider the hypothesis testing for *autoregressive processes* with a finite alphabet $\mathcal{W}$ treated in Example 4.4.8 in §4.4. If $r \geq D(P_{\overline{W}}||P_W)$, then the result of Example 4.5.1 yields the following formula:

$$
B_e^*(r|\mathbf{X}||\overline{\mathbf{X}}) = \inf_{Q:D(Q||P_W)\leq r} \left\{ D(Q||P_{\overline{W}}) + r - D(Q||P_W) \right\}. \qquad (4.5.32)
$$

Here, $B_e^*(r|\mathbf{X}||\overline{\mathbf{X}}) = 0$ for $r \leq D(P_{\overline{W}}||P_W)$. □

Example 4.5.5. Let $\mathcal{X}$ be a finite alphabet. Consider the *mixed source* $\mathbf{X} = \{X^n\}_{n=1}^\infty$ and the stationary memoryless source $\overline{\mathbf{X}} = \{\overline{X}^n\}_{n=1}^\infty$ subject to a probability distribution $\overline{P}$ given in Example 4.2.1. We assume that $P_1(x) > 0$ and $P_2(x) > 0$ are satisfied for all $x \in \mathcal{X}$ in order to meet the assumption (4.5.2). Recall that the mixed source $\mathbf{X} = \{X^n\}_{n=1}^\infty$ is defined as

$$
P_{X^n}(\mathbf{x}) = \alpha_1 P_{X_1^n}(\mathbf{x}) + \alpha_2 P_{X_2^n}(\mathbf{x}) \quad (\forall \mathbf{x} \in \mathcal{X}^n), \qquad (4.5.33)
$$

where $\mathbf{X}_1 = \{X_1^n\}_{n=1}^\infty$ and $\mathbf{X}_2 = \{X_2^n\}_{n=1}^\infty$ denote the stationary memoryless sources subject to probability distributions P_1 and P_2, respectively. Now, we define ν_1 and ν_2 by (4.4.42) and (4.4.43) in Example 4.4.4, respectively, and the half-spaces $\kappa_R^{(1)}$ and $\kappa_R^{(2)}$ in $\mathcal{P}(\mathcal{X})$ by

$$
\kappa_R^{(1)} = \left\{ Q \in \mathcal{P}(\mathcal{X}) \,\middle|\, \sum_{x\in\mathcal{X}} Q(x) \log \frac{P_1(x)}{\overline{P}(x)} \leq R \right\}, \qquad (4.5.34)
$$

$$
\kappa_R^{(2)} = \left\{ Q \in \mathcal{P}(\mathcal{X}) \,\middle|\, \sum_{x\in\mathcal{X}} Q(x) \log \frac{P_2(x)}{\overline{P}(x)} \leq R \right\}, \qquad (4.5.35)
$$

where $\mathcal{P}(\mathcal{X})$ denotes the set of all probability distributions over $\mathcal{X}$. Denote by $P_R^{(1)}$ and $P_R^{(2)}$ the *projections* of P_1 and P_2 on $\nu_1 \cap \kappa_R^{(1)}$ and $\nu_2 \cap \kappa_R^{(2)}$, respectively. By taking (1.9.39) and (1.9.40) in Example 1.9.4 into consideration and applying Sanov's theorem, we obtain

$$
\eta(R) = \min(D(P_R^{(1)}||P_1), D(P_R^{(2)}||P_2)). \qquad (4.5.36)
$$

This formula indicates that $\eta(R) = 0$ if $R \geq \min(D(P_1||\overline{P}), D(P_2||\overline{P}))$ and $\eta(R)$ is a continuous and monotone decreasing function of R. By substituting this $\eta(R)$ into the right-hand side of (4.5.3) in Theorem 4.5.1, we can compute values of $B_e^*(r|\mathbf{X}||\overline{\mathbf{X}})$ as a function of r. □

Example 4.5.6. Let us consider the hypothesis testing for the *mixed sources* given in Example 4.4.5. First, we define ν_1, ν_2, μ_1 and μ_2 as in the same way as Example 4.4.5 and the half-spaces $\kappa_R^{(1)}$, $\kappa_R^{(2)}$, $\kappa_R^{(3)}$ and $\kappa_R^{(4)}$ by

$$\kappa_R^{(1)} = \left\{ Q \in \mathcal{P}(\mathcal{X}) \left| \sum_{x \in \mathcal{X}} Q(x) \log \frac{P_1(x)}{\overline{P}_1(x)} \leq R \right. \right\}, \tag{4.5.37}$$

$$\kappa_R^{(2)} = \left\{ Q \in \mathcal{P}(\mathcal{X}) \left| \sum_{x \in \mathcal{X}} Q(x) \log \frac{P_1(x)}{\overline{P}_2(x)} \leq R \right. \right\}, \tag{4.5.38}$$

$$\kappa_R^{(3)} = \left\{ Q \in \mathcal{P}(\mathcal{X}) \left| \sum_{x \in \mathcal{X}} Q(x) \log \frac{P_2(x)}{\overline{P}_2(x)} \leq R \right. \right\}, \tag{4.5.39}$$

$$\kappa_R^{(4)} = \left\{ Q \in \mathcal{P}(\mathcal{X}) \left| \sum_{x \in \mathcal{X}} Q(x) \log \frac{P_2(x)}{\overline{P}_1(x)} \leq R \right. \right\} \tag{4.5.40}$$

in the same way as Example 4.4.5. Denote by $P_R^{(1)}$ and $P_R^{(2)}$ the *projections* of P_1 and P_2 on

$$\nu_1 \cap \left((\mu_1 \cap \kappa_R^{(1)}) \cup (\mu_2 \cap \kappa_R^{(2)}) \right),$$

$$\nu_2 \cap \left((\mu_2 \cap \kappa_R^{(3)}) \cup (\mu_1 \cap \kappa_R^{(4)}) \right),$$

respectively. Then, if we apply Sanov's theorem in the same way as Example 4.4.5, we obtain

$$\eta(R) = \min(D(P_R^{(1)}||P_1), D(P_R^{(2)}||P_2)), \tag{4.5.41}$$

which is the same as $\eta(R)$ obtained in Example 4.4.5. This formula indicates that $\eta(R) = 0$ if

$$R \geq \min(D(P_1||\overline{P}_1), D(P_1||\overline{P}_2), D(P_2||\overline{P}_1), D(P_2||\overline{P}_2))$$

is satisfied and $\eta(R)$ is a continuous and monotone decreasing function of R. By substituting this $\eta(R)$ into the right-hand side of (4.5.3) in Theorem 4.5.1, we can compute values of $B_e^*(r|\mathbf{X}||\overline{\mathbf{X}})$ as a function of r. □

Remark 4.5.3. Unfortunately, there is no simple form of $B_e^*(r|\mathbf{X}||\overline{\mathbf{X}})$ corresponding to $B_e(r|\mathbf{X}||\overline{\mathbf{X}})$ for the mixed sources $\mathbf{X}$ and $\overline{\mathbf{X}}$ given in (4.4.60) in Remark 4.4.3. □

Example 4.5.7. So far, we have only considered the cases where $\mathcal{X}$ is a finite alphabet. If we consider general stationary memoryless sources with an alphabet $\mathcal{X}$ not restricted to a finite set, Sanov's theorem does not always hold. However, we can compute $B_e^*(r|\mathbf{X}||\overline{\mathbf{X}})$ by using *Cramér's theorem*, which always holds (suppose here that $\mathbf{X}$ and $\overline{\mathbf{X}}$ are *stationary memoryless sources*).

That is, as is mentioned in Example 4.4.7, we have only to use the rate function $I(x)$ and set

$$\eta(R) = \inf_{x \le R} I(x) \tag{4.5.42}$$

similarly to (4.4.83). Note that the right-hand side of (4.5.42) is expressed in terms of divergences (similarly to Sanov's theorem) only if $R \in \mathcal{D}$ is satisfied, where we use the notation given in Example 4.4.7. □

Example 4.5.8. Let us consider the *stationary memoryless Gaussian sources* $\mathbf{X} = \{P_\kappa\}$ and $\overline{\mathbf{X}} = \{P_{\overline{\kappa}}\}$ treated in Example 4.4.9 in the preceding section. Since $\eta(R)$ and $R + \eta(R)$ are given in (4.4.96) and (4.4.97), respectively, substitution of these into (4.5.3) in Theorem 4.5.1 and some simple calculation yield

$$B_e^*(r|\mathbf{X}||\overline{\mathbf{X}}) = (\sqrt{r} - \sqrt{a})^2 \mathbf{1}[r \ge a] \tag{4.5.43}$$

(Fig 4.8), where $a = D(P_\kappa || P_{\overline{\kappa}})$. Notice here that this function and $B_e(r|\mathbf{X}||\overline{\mathbf{X}})$ in Example 4.4.9 are symmetric with respect to the vertical axis. The formula (4.5.43) tells us that $B_e^*(r|\mathbf{X}||\overline{\mathbf{X}})$ is a monotone increasing function of r. It also tells us that $B_e^*(r|\mathbf{X}||\overline{\mathbf{X}}) = 0$ for $r \le a$. □

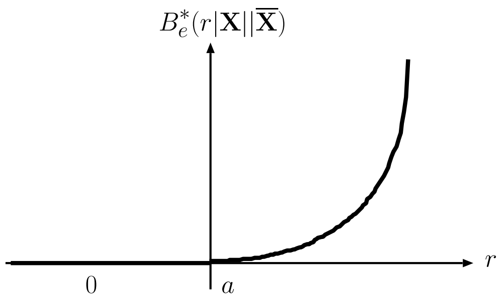

Fig. 4.8.

4.6 Generalized Hypothesis Testing

In all the hypothesis testing problems treated in this chapter, $P_{\overline{X}^n}$ in the alternative hypothesis $\overline{\mathbf{X}} = \{P_{\overline{X}^n}\}_{n=1}^\infty$ is regarded as a probability distribution (probability measure) over $\mathcal{X}^n$. However, all of the theorems, lemmas and remarks except Theorem 4.3.2 still hold if $P_{\overline{X}^n}$ is replaced with another nonnegative measure (not necessarily a probability measure) G_n satisfying $G_n(\emptyset) = 0$. Here, the error probability of the second kind $\lambda_n \equiv \Pr\{\overline{X}^n \in \mathcal{A}_n\}$ is interpreted as $\lambda_n \equiv G_n(\mathcal{A}_n)$. In addition, the inequality $\underline{D}(\mathbf{X}||\overline{\mathbf{X}}) \geq 0$ must be replaced with the inequality $\underline{D}(\mathbf{X}||\overline{\mathbf{X}}) \geq -\overline{\kappa}$, where

$$\kappa_n \equiv G_n(\mathcal{X}^n) \quad (n = 1, 2, \cdots),$$

$$\overline{\kappa} \equiv \limsup_{n\to\infty} \frac{1}{n} \log G_n(\mathcal{X}^n),$$

$$\underline{\kappa} \equiv \limsup_{n\to\infty} \frac{1}{n} \log \frac{1}{G_n(\mathcal{X}^n)}.$$

We also note that Theorem 4.5.1 holds only if $\underline{\kappa} < +\infty$ is satisfied. We must replace $B_e^*(0|\mathbf{X}||\overline{\mathbf{X}}) = 0$ in Theorem 4.5.1 with $B_e^*(0|\mathbf{X}||\overline{\mathbf{X}}) = \underline{\kappa}$ and $1 - \lambda_n$ in Definition 4.5.1 by $\kappa_n - \lambda_n$.

As an example of such nonnegative measures G_n $(n = 1, 2, \cdots)$ we may consider the measure (called the *counting measure*) satisfying $G_n(\mathbf{x}) = 1$ $(\forall \mathbf{x} \in \mathcal{X}^n;\ \forall n = 1, 2, \cdots)$ if $\mathcal{X}$ is a finite or a countably infinite alphabet. Another example may be the n-dimensional Lebesgue measure if $\mathcal{X}$ is the set of all real numbers (Theorem 4.3.2 holds if $P_{\overline{X}^n}$ is replaced with the counting measure or the Lebesgue measure). In particular, the hypothesis testing with the counting measure as G_n is nothing but the fixed-length source coding described in Chapter 1 as will be shown in the following section.

Remark 4.6.1. If the probability distribution P_{X^n} of the null hypothesis is replaced with another nonnegative measure F_n such that $F_n(\emptyset) = 0$, we can easily verify that Theorem 4.4.1 and Theorem 4.5.1 still hold. We have only to interpret probabilities in the proofs as the corresponding measures. □

4.7 Hypothesis Testing and Source Coding

So far we have described theorems on the hypothesis testing. In this section, we point out that the *hypothesis testing problems* with a *countably infinite alphabet* $\mathcal{X}$ are deeply related to the *fixed-length source coding problems* described in Chapter 1.

For example, we can see that Theorem 1.9.1 on the source coding is obtained as a special case of (the generalized version of) Theorem 4.4.1. To this end, let the null hypothesis $\mathbf{X} = \{X^n\}_{n=1}^{\infty}$ be arbitrary and let the alternative hypothesis $\overline{\mathbf{X}} = \{C_n\}_{n=1}^{\infty}$ be the counting measure

$$C_n(\mathbf{x}) = 1 \quad (\forall \mathbf{x} \in \mathcal{X}^n)$$

described in the preceding section. We denote by $\mathbf{C} = \{C_n\}_{n=1}^{\infty}$ this alternative hypothesis. For an arbitrary given acceptance region $\mathcal{A}_n \subset \mathcal{X}^n$, set $M_n = |\mathcal{A}_n|$ and define the mapping $\varphi_n : \mathcal{X}^n \to \mathcal{M}_n$ that maps each element of $\mathcal{A}_n$ to a distinct element of $\mathcal{M}_n = \{1, 2, \cdots, M_n\}$ in the order of $1, 2, \cdots,$

and all elements of $\mathcal{A}_n^c$ to 1. Define $\psi_n : \mathcal{M}_n \to \mathcal{X}^n$ as the inverse mapping of $\varphi_n|_{\mathcal{A}_n}$. If we consider the source coding with φ_n as an encoder and ψ_n as a decoder, we have $\mathcal{A}_n = \{\mathbf{x} \in \mathcal{X}^n \mid \psi_n(\varphi_n(\mathbf{x})) = \mathbf{x}\}$, which leads to the fact that the error probability of the first kind $\mu_n = \Pr\{X^n \notin \mathcal{A}_n\}$ of this hypothesis testing is equal to the decoding error probability ε_n caused by the code (φ_n, ψ_n). Such a relationship between the hypothesis testing and the source coding is one-to-one if we identify codes sharing the set $\mathcal{A}_n = \{\mathbf{x} \in \mathcal{X}^n \mid \psi_n(\varphi_n(\mathbf{x})) = \mathbf{x}\}$, the set of all correctly decodable $\mathbf{x} \in \mathcal{X}^n$, as the same code. In this case, the (generalized) error probability of the second kind λ_n can be written as

$$\begin{aligned}\lambda_n &= C_n(\mathcal{A}_n) = |\mathcal{A}_n| = M_n \\ &= e^{nr_n} \end{aligned} \tag{4.7.1}$$

under the counting measure C_n, where

$$r_n = \frac{1}{n}\log M_n$$

means the coding rate of the code (φ_n, ψ_n). Then, we obtain from (4.7.1) that

$$\liminf_{n\to\infty} \frac{1}{n}\log\frac{1}{\lambda_n} = -\limsup_{n\to\infty} r_n.$$

Hence, R being an r-achievable rate of the (generalized) hypothesis testing is equivalent to $-R$ being an r-achievable rate of the source coding. From Definition 1.9.1, Definition 1.9.2, Definition 4.4.1 and Definition 4.4.2, we can obtain

$$B_e(r|\mathbf{X}||\mathbf{C}) = -R_e(r|\mathbf{X}) \quad (\forall r > 0) \tag{4.7.2}$$

connecting $B_e(r|\mathbf{X}||\mathbf{C})$ with $R_e(r|\mathbf{X})$.

By using (4.7.2), we can obtain Theorem 1.9.1 from Theorem 4.4.1 and vice versa. For example, Theorem 4.4.1 implies Theorem 1.9.1 in the following manner. First, by recalling that the alternative hypothesis is the counting measure C_n, the probability on the right-hand side of (4.4.1) defining $\eta(R)$ can be written as

$$\begin{aligned}&\Pr\left\{\frac{1}{n}\log\frac{P_{X^n}(X^n)}{P_{\overline{X}^n}(X^n)} \le R\right\} \\ &= \Pr\left\{\frac{1}{n}\log\frac{P_{X^n}(X^n)}{C_n(X^n)} \le R\right\} \\ &= \Pr\left\{\frac{1}{n}\log P_{X^n}(X^n) \le R\right\} \\ &= \Pr\left\{\frac{1}{n}\log\frac{1}{P_{X^n}(X^n)} \ge -R\right\}.\end{aligned}$$

Taking the definition $\sigma(R)$ in (1.9.2) into consideration, we have $\eta(R) = \sigma(-R)$, which leads to

$$\sigma(R) = \eta(-R). \tag{4.7.3}$$

Then, Theorem 4.4.1 on the (generalized) hypothesis testing and (4.7.2) yield

$$\begin{aligned}
R_e(r|\mathbf{X}) &= -B_e(r|\mathbf{X}||\mathbf{C}) \\
&= -\inf_R \{R + \eta(R) \mid \eta(R) < r\} \\
&= \sup_R \{-R - \eta(R) \mid \eta(R) < r\}.
\end{aligned}$$

By replacing R by $-R$ and using (4.7.3), it follows that

$$R_e(r|\mathbf{X}) = \sup_{R \geq 0} \{R - \sigma(R) \mid \sigma(R) < r\},$$

which is exactly the same as Theorem 1.9.1 on the source coding.

By using an argument similar to the argument above, it is easy to verify that, Theorem 4.1.1, Theorem 4.2.1 and Theorem 4.3.1 with the counting measure $\mathbf{C} = \{C_n\}_{n=1}^{\infty}$ as the alternative hypothesis $\overline{\mathbf{X}}$ coincide with Theorem 1.3.1, Theorem 1.6.1 and Theorem 1.5.1 on the fixed-length source coding that are described in Chapter 1. In addition, in this case (4.1.2) coincides with Lemma 1.4.1 in Chapter 1.

Readers may feel that there may be a relationship between Theorem 4.5.1 on the hypothesis testing and Theorem 1.10.1 on the source coding similar to the relationship between Theorem 4.4.1 and 1.9.1. Nevertheless, there is no such a relationship. This is because the two definitions of the r-achievable rates are different. Recall that, while the r-achievable rate R in Definition 1.10.1 is defined under the constraint

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{1}{1-\varepsilon_n} \leq r,$$

the r-achievable rate R in Definition 4.5.1 is defined under the constraint

$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\mu_n} \geq r,$$

though we have $\mu_n = \varepsilon_n$ in the relationship between the hypothesis testing and the source coding discussed so far.

However, we can obtain a relationship between them by modifying the formulation of the source coding problem treated in §1.10. To this end, assume that a source alphabet $\mathcal{X}$ is finite and consider the "dual coding rate"

$$\rho_n = \frac{1}{n} \log(|\mathcal{X}|^n - M_n) \tag{4.7.4}$$

instead of the coding rate $r_n = \dfrac{1}{n} \log M_n$. Furthermore, let us define, instead of Definition 1.10.1 and Definition 1.10.2, respectively:

Definition 4.7.1.

$$\text{Rate } R \text{ is } r\text{-achievable} \overset{\text{def}}{\Longleftrightarrow} \text{There exists an } (n, M_n, \varepsilon_n)\text{-code satisfying } \liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\varepsilon_n} \geq r \text{ and } \liminf_{n\to\infty} \frac{1}{n} \log \rho_n \geq R.$$

Definition 4.7.2 (supremum r-achievable fixed-length dual coding rate).

$$\overline{R}_e(r|\mathbf{X}) = \sup \{R \mid R \text{ is } r\text{-achievable}\},$$

Such definitions make sense in the following case. Suppose that the decoding error probability is required to satisfy $\varepsilon_n \simeq e^{-nr}$ for a large enough $r > 0$. Since ε_n becomes quite small in such a situation, the coding rate $r_n = \frac{1}{n} \log M_n$ is nearly equal to $\log |\mathcal{X}|$. This implies that $M_n \simeq |\mathcal{X}|^n$. Therefore, it is meaningful to evaluate $|\mathcal{X}|^n - M_n$ instead of M_n itself. In this case it is a fundamental problem on the source coding to make the dual coding rate ρ_n in (4.7.4) satisfying

$$|\mathcal{X}|^n - M_n = e^{n\rho_n}$$

as large as possible. Note that $\overline{R}_e(r|\mathbf{X})$ in Definition 4.7.2 means the supremum of the dual coding rate ρ_n with respect to all codes satisfying the condition

$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\varepsilon_n} \geq r.$$

In this modified source coding problem, since $\mathbf{C} = \{C_n\}_{n=1}^{\infty}$ is defined as the counting measure, we obtain

$$|\mathcal{X}|^n - \lambda_n = e^{n\rho_n}$$

instead of (4.7.1) and

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{1}{|\mathcal{X}|^n - \lambda_n} = -\liminf_{n\to\infty} \rho_n.$$

Therefore, it turns out that R being the r-achievable rate of the (generalized) hypothesis testing is equivalent to $-R$ being the r-achievable source coding rate. From Definitions 4.5.1, 4.5.2, 4.7.1 and 4.7.2, we have the following relationship:

$$B_e^*(r|\mathbf{X}||\mathbf{C}) = -\overline{R}_e(r|\mathbf{X}). \tag{4.7.5}$$

By making use of this equality, if either $B_e^*(r|\mathbf{X}||\mathbf{C})$ or $\overline{R}_e(r|\mathbf{X})$ can be computed, we can compute the other. However, as is easily verified, the assumption (4.5.2) in Theorem 4.5.1 does not hold for the case that $\overline{\mathbf{X}}$ is equal to the counting measure $\mathbf{C}$. This means that the formula (4.5.3) for $B_e^*(r|\mathbf{X}||\mathbf{C})$ no longer holds. Thus, it is temporally impossible to obtain a formula for $\overline{R}_e(r|\mathbf{X})$ on the source coding via the formula on the hypothesis testing, though the formula for $R_e(r|\mathbf{X})$ can be obtained in such a way.