Chapter 22
Special Topic: Applications of Large Deviation Theory

This chapter includes two applications of the large deviation theory presented in Chapter 21. One concerns an application to a problem in cryptography in which, among other motivations, hackers attempt to break a password by guessing. The other is an application to the efficiency of large sample statistical tests of hypothesis.

Example 1 *(Encrypted Security Systems*[1]*).* The problem to be considered here is of interest to cryptographers analyzing, for example, attempts by a hacker to enter a password protected system by robotically guessing it. The problem can be abstractly stated as follows: For a given finite set $S = \{1, 2, \ldots k\}$, say, Alice randomly generates a cipher $X^{(n)} = \mathbf{x} \in S^n$ of length n, where $X^{(n)} = (X_1, \ldots, X_n) \in S^n$ has a joint probability mass function $p_{X^{(n)}}(x_1, \ldots, x_n)$, $x_j \in S$, $1 \le j \le n$. According to some guessing strategy, Bob systematically steps through the messages $\mathbf{y} \in S^n$ in some specified order, and Alice responds $X^{(n)} = \mathbf{y}$ with "yes" or "no," according to whether $\mathbf{y} = \mathbf{x}$ or not. The goal is to quantify the effort required by guessing. Throughout it will be assumed without further mention that the *message source* $X_1, X_2, \ldots$ is a stationary process.

Mathematically, guessing is given by a bijection $G : S^n \to \{1, 2, \ldots, |S|^n\}$ prescribing the orders in which guesses $\mathbf{y} \in S^n$ are made in the guessing strategy. $G(\mathbf{x})$ is then the number of guesses to reach the given cipher $\mathbf{x}$.

[1] This example is based on Hanawal and Sundaresan (2011).

R. Bhattacharya, E. C. Waymire, *Stationary Processes and Discrete Parameter Markov Processes*, Graduate Texts in Mathematics 293,
https://doi.org/10.1007/978-3-031-00943-3_22

To minimize the expected number of guesses, an optimal choice is a guessing function G^* that would therefore make the order of selections according to decreasing probabilities $f(\mathbf{y}), \mathbf{y} \in S^n$. Note that if $f(\mathbf{y}) = f(\mathbf{z})$, then the order in which $\mathbf{y}$ and $\mathbf{z}$ are guessed will not affect the number of guesses to unlock the password. In particular, an optimal G is not unique with regard to minimizing the expected number of guesses.

As a measure of the attackers effort, cryptologists consider an optimal G^* to define an optimal guessing exponent by

$$g(\rho) = \lim_{n\to\infty} \frac{1}{n} \ln \mathbb{E}G^*(X^{(n)})^\rho, \tag{22.1}$$

when the limit exists. The primary focus of this chapter is on the computation of $g(\rho)$ in some generality via large deviation theory. This is achieved by systematically establishing a succession of equivalent computations: Proposition 22.2 recasts the problem in terms of an equivalent computation for word lengths, Proposition 22.3 recasts this in terms of a Rényi entropy computation, and finally Theorem 22.4, Corollary 22.5 in terms of a large deviation computation for the so-called information spectrum.

Remark 22.1. Calculations have been made for $g(\rho)$ in the case of i.i.d. encodings $X_1, \ldots, X_n$ by Arikan (1996), and irreducible Markov chain encodings by Malone and Sullivan (2004). These will appear as applications of the large deviation results of Hanawal and Sundaresan (2011) at the end of this example.

It will be helpful to introduce the *guessing length function* $L_G : S^n \to \mathbf{N}$ associated with G defined by

$$L_G(\mathbf{x}) = \lceil -\ln \frac{1}{CG(\mathbf{x})} \rceil, \quad \mathbf{x} \in S^n, \tag{22.2}$$

where $\lceil\cdot\rceil$ is the ceiling function, i.e., $\lceil x \rceil$ is the smallest integer not smaller than x, and $C = \sum_{\mathbf{x}\in S^n} \frac{1}{G(\mathbf{x})}$ is a normalization constant. In particular,

$$Q_G(x) = \frac{1}{CG(x)}, \quad x \in S^n, \tag{22.3}$$

defines a probability mass function on S^n. Note that since $C \geq 1$,

$$G(x) = \frac{1}{CQ_G(x)} \leq \frac{1}{Q_G(x)}. \tag{22.4}$$

Clearly, $\ln G(x) \leq L_G(x), x \in S^n$ by definition, and

$$\ln G(x) = -\ln Q_G(x) - \ln C \geq \lceil -\ln Q_G(x) \rceil - 1 - \ln C, \tag{22.5}$$

so that, in summary,

$$L_G(x) - 1 - \ln C \leq \ln G(x) \leq L_G(x). \tag{22.6}$$

To denote the dependence of G and L on the message length n, we write G_n, L_n, C_n, respectively, when necessary. Note that L_n satisfies the so-called *Kraft inequality* (Exercise 4)

$$\sum_{x \in S^n} \exp\{-L_n(x)\} \leq 1. \tag{22.7}$$

In general any function $L : S^n \to \mathbb{N}$ satisfying the Kraft inequality will be referred to as a *length function*. We let $\mathcal{L}_n$ denote the set of all such functions on S^n. L^* will denote a length function that minimizes $\mathbb{E}\exp\{\rho L(X^{(n)})\}$.

Suppose that $X_1, X_2, \ldots$ is a stationary process and let $Q \in \mathcal{P}_n$ denote the distribution of $(X_{1+m}, \ldots, X_{n+m})$ (m=1,2,...). The *Shannon entropy*[2] expressed in nats, i.e., using natural logarithms, is defined by

$$H(X_1, \ldots, X_n) \equiv H(Q) = -\sum_{x \in S^n} Q(\{x\}) \ln Q(\{x\}).$$

Shannon's entropy of the stationary process is defined by

$$H = \lim_{n \to \infty} \frac{H(X_1, X_2, \ldots, X_n)}{n},$$

for which existence is a direct consequence of subadditivity using Fekete's lemma from Chapter 5. Specifically, letting Q_n denote the distribution of $(X_1, X_2, \ldots, X_n)$, one has

$$\begin{aligned} H(Q_{n+m}) &\equiv H(X_1, \ldots, X_{n+m}) \\ &\leq H(X_1, \ldots, X_n) + H(X_{n+1}, \ldots, X_{n+m}) \\ &= H(X_1, \ldots, X_n) + H(X_1, \ldots, X_m) = H(Q_n) + H(Q_m), \end{aligned} \tag{22.8}$$

where the essential second line is left as Exercise 2.

Remark 22.2. Note the existence of a length function L for which the (approximate) expected lengths are minimal, i.e., the problem

$$\min_{L \in \mathcal{L}_n : \sum_x e^{-L(x)} \leq 1} \sum_{x \in S^n} p_{X^{(n)}}(x) L(x)$$

[2] See Bhattacharya and Waymire (1990, 2009), pp.184–189 for a related treatment of Shannon entropy.

can be shown to have a solution by the method of Lagrange multipliers (Exercise 5) providing one permits non-integer solutions.

Theorem 22.1 *(Shannon).* For a length function L one has

$$H(X_1,\dots,X_n) \le \mathbb{E}L(X_1,\dots,X_n),$$

with equality if and only if $p_{X^{(n)}}(x) = e^{-L(x)}$. Moreover, letting $L^*(X_1,\dots,X_n)$ denote the lengths having smallest expected value possible for the word $(X_1,\dots,X_n)$, one has

$$H(X_1,\dots,X_n) \le \mathbb{E}L^*(X_1,\dots,X_n) \le H(X_1,\dots,X_n) + 1.$$

In particular,

$$\frac{H(X_1,\dots,X_n)}{n} \le \frac{\mathbb{E}L^*(X_1,\dots,X_n)}{n} \le \frac{H(X_1,\dots,X_n)}{n} + \frac{1}{n},$$

$$\lim_{n\to\infty} \mathbb{E}\frac{L^*(X_1,\dots,X_n)}{n} = H.$$

Proof. To prove the lower bound let $q(x) = \frac{e^{-L(x)}}{\sum_{y\in S^n} e^{-L(y)}}$, and $K = \sum_{y\in S^n} e^{-L(y)} \le 1$, by the Kraft inequality. Then,

$$\begin{aligned}
&\mathbb{E}L(X_1,\dots,X_n) - H(X_1,\dots,X_n)\\
&= \sum_{x\in S^n} p_{X^{(n)}}(x)L(x) - \sum_{x\in S^n} p_{X^{(n)}}(x)\ln\frac{1}{p_{X^{(n)}}(x)}\\
&= -\sum_{x\in S^n} p_{X^{(n)}}(x)\ln e^{-L(x)} + \sum_{x\in S^n} p_{X^{(n)}}(x)\ln p_{X^{(n)}}(x)\\
&= \sum_{x\in S^n} p_{X^{(n)}}(x)\ln\frac{p_{X^{(n)}}(x)}{q(x)} - \ln K\\
&= D(p_{X^{(n)}}||q) + \ln\frac{1}{K} \ge 0. \qquad (22.9)
\end{aligned}$$

Note that approximately if $L(x) = \ln\frac{1}{p_{X(n)}(x)}$, then $H = L$. However, such a choice for L is not an integer. Taking $L(x) = \lceil \ln\frac{1}{p_{X(n)}(x)}\rceil$, the Kraft inequality is preserved by this choice Now, for this choice of lengths, a simple calculation yields,

$$H(X_1,\dots,X_n) \le \mathbb{E}L(X_1,\dots,X_n) \le H(X_1,\dots,X_n) + 1.$$

Since $\mathbb{E}L^*(X_1, \ldots, X_n) \le \mathbb{E}L(X_1, \ldots, X_n)$ both the lower and upper bounds are satisfied by $\mathbb{E}L^*(X_1, \ldots, X_n)$. ■

Lemma 1. Let G be a guessing function and L_G its associated length function. Then,

$$\left| \frac{1}{\rho} \ln \mathbb{E}G^*(X^{(n)})^\rho - \frac{1}{\rho} \ln \mathbb{E} \exp \left\{ \rho L^*(X^{(n)}) \right\} \right| \le 1 + \ln C, \tag{22.10}$$

where $C = \sum_{\mathbf{x} \in S^n} \frac{1}{G(\mathbf{x})}$.

Proof. For a length function $L \in \mathcal{L}_n$, let G_L be the guessing function that guesses in the increasing order of L-lengths. Messages of the same L-length are ordered according to an arbitrary fixed rule, say lexicographical order on S^n. Define a probability mass function on S^n by

$$Q_L(x) = \frac{\exp\{-L(x)\}}{\sum_{y \in S^n} \exp\{-L(y)\}}, \quad x \in S^n. \tag{22.11}$$

Note that G_L guesses in the decreasing order of Q_L probabilities. In particular, $G_L(x) \le \sum_{y \in S^n} \mathbf{1}[Q_L(y) \ge Q_L(x)] \le \sum_{y \in S^n} \frac{Q_L(y)}{Q_L(x)} = \frac{1}{Q_L(x)}$, so that

$$\ln G_L(x) \le -\ln Q_L(x) \quad x \in S^n. \tag{22.12}$$

Also, by definition of Q_L and using Kraft's inequality (22.7),

$$\frac{1}{Q_L(x)} = \exp\{L(x)\} \sum_{y \in S^n} \exp\{-L(y)\} \le \exp\{L(x)\},$$

so that

$$-\ln Q_L(x) \le L(x), \quad x \in S^n. \tag{22.13}$$

From these inequalities one deduces that for any $B \ge 1$,

$$\{x : L_G(x) \ge B + 1 + \ln C\} \subset \{x : G(x) \ge e^B\} \subset \{x : L_G(x) \ge B\}, \tag{22.14}$$

and

$$\{x : G_L(x) \ge e^B\} \subset \{x : L(x) \ge B\}. \tag{22.15}$$

Now, by (22.12) followed by (22.6),

$$\begin{aligned} \mathbb{E}\exp\{\rho L(X^{(n)})\} &\ge \mathbb{E}G_L(X^{(n)})^\rho \ge \mathbb{E}G^*(X^{(n)})^\rho \\ &\ge \mathbb{E}\exp\{\rho L_{G^*}(X^{(n)})\} \exp\{-\rho(1 + \ln C)\} \end{aligned}$$

$$\geq \mathbb{E}\exp\{\rho L^*(X^{(n)})\}\exp\{-\rho(1+\ln C)\}. \tag{22.16}$$

Thus,

$$\frac{\mathbb{E}G_L(X^{(n)})^\rho}{\mathbb{E}G^*(X^{(n)})^\rho} \leq \frac{\mathbb{E}\exp\{\rho L(X^{(n)})\}}{\mathbb{E}\exp\{\rho L^*(X^{(n)})\}}\exp\{\rho(1+\ln C)\}, \tag{22.17}$$

and, in terms of the length function L_G associated with G, one similarly has

$$\frac{\mathbb{E}G(X^{(n)})^\rho}{\mathbb{E}G^*(X^{(n)})^\rho} \geq \frac{\mathbb{E}\exp\{\rho L_G(X^{(n)})\}}{\mathbb{E}\exp\{\rho L^*(X^{(n)})\}}\exp\{-\rho(1+\ln C)\}. \tag{22.18}$$

The lemma now follows from these bounds upon taking logarithms with $L = L^*$ in (22.16). That is

$$1 \geq \frac{\mathbb{E}G^*(X^{(n)})^\rho}{\mathbb{E}\exp\{\rho L^*(X^{(n)})\}} \geq \exp\{-\rho(1+\ln C)\}, \tag{22.19}$$

so that $0 \geq \ln\mathbb{E}G^*(X^{(n)})^\rho - \ln\mathbb{E}\{\rho L^*(X^{(n)})\} \geq -\rho(1+\ln C)$. ■

The existence and determination of $g(\rho)$ will ultimately follow from an application of Varadhan's integral formula applied to a related function of $X_1, \ldots, X_n$ obtained from the next three propositions and their lemmas.

Proposition 22.2. The guessing exponent $g(\rho)$ exists if and only if

$$\ell(\rho) = \lim_{n\to\infty}\inf_{L\in\mathcal{L}_n}\frac{1}{n}\ln\mathbb{E}\exp\{\rho L(X^{(n)})\} \tag{22.20}$$

exists. Moreover $g(\rho) = \ell(\rho)$ when either exists.

Proof. Note that $C_n \leq 1+n\ln|S|$. Dividing both sides of the inequality in Lemma 1 by n, one has

$$\left|\frac{1}{n\rho}\ln\mathbb{E}G^{*\rho}(X^{(n)}) - \frac{1}{n\rho}\ln\mathbb{E}(\exp\{\rho L^*(X^{(n)})\})\right| \leq \frac{1}{n}(1+\ln C_n) = O\left(\frac{\ln n}{n}\right). \tag{22.21}$$

Thus the sequences differ by $o(1)$ as $n \to \infty$. ■

The next proposition requires the *Rényi entropy rate* of order $\alpha \neq 1$ defined by

$$H_\alpha(p_{X^{(n)}}) = \frac{1}{1-\alpha}\ln\sum_{\mathbf{x}\in S^n}p^\alpha_{X^{(n)}}(\mathbf{x})) \equiv \frac{1}{1-\alpha}\ln\mathbb{E}p^{\alpha-1}_{X^{(n)}}(X^{(n)}). \tag{22.22}$$

Proposition 22.3. $\lim_{n\to\infty} \inf_{L\in\mathcal{L}_n} \frac{1}{n} \ln \mathbb{E} \exp\{\rho L(X^{(n)})\}$, or equivalently $\lim_n \ln \mathbb{E}G^*(X^{(n)})^\rho$, exists if and only if $\lim_{n\to\infty} \frac{1}{n} H_\alpha(p_{X^{(n)}})$ exists for $\alpha = \frac{1}{1+\rho}$. Moreover, if the latter limit exists, then it is given by $\frac{g(\rho)}{\rho}$.

Proof. The equivalence is the content of Proposition 22.2. We focus on the former limit. For each n the Donsker–Varadhan variational formula of Corollary 21.17 yields, upon replacing g by $L(X^{(n)})$, λ by ρ, Q by $p_{X^{(n)}}$, and P by Q, that

$$\ln \mathbb{E} \exp\{\rho L(X^{(n)})\} = \sup_{Q\in\mathcal{P}_n} \{\rho \mathbb{E}_Q L(X^{(n)}) - D(Q||p_{X^{(n)}})\}. \tag{22.23}$$

Taking the infimum on both sides over all length functions $L \in \mathcal{L}_n$ and applying Fan's minimax exchange of supremum and infimum, one has

$$\begin{aligned}\inf_{L\in\mathcal{L}_n} \ln \mathbb{E} \exp\{\rho L_n(X^{(n)})\} &= \inf_{L\in\mathcal{L}_n} \sup_{Q\in\mathcal{P}_n} \{\rho \mathbb{E}_Q L_n(X^{(n)}) - D(Q||p_{X^{(n)}})\}\\ &= \sup_{Q\in\mathcal{P}_n} \inf_{L\in\mathcal{L}_n} \{\rho \mathbb{E}_Q L_n(X^{(n)}) - D(Q||p_{X^{(n)}})\}\\ &= \sup_{Q\in\mathcal{P}_n} \{\rho H(Q) - D(Q||p_{X^{(n)}})\} + O(1)\\ &= \rho H_{\frac{1}{1+\rho}}(p_{X^{(n)}}) + O(1),\end{aligned} \tag{22.24}$$

where to justify the use of Fan's minimax formula one notes firstly convexity of the map $(Q, L) \in \mathcal{P}_n \times \mathcal{L}_n \to \mathbb{E}_Q\{\rho L(X^{(n)}) - D(Q||p_{X^{(n)}}) = \sum_{x\in S^n}\{\rho L(x) + \ln Q(x) - \ln p_{X^{(n)}}(x)\}Q(x)$, as a function of $Q \in \mathcal{P}_n$, and the linearity as a function of L. The next equation follows from Theorem 22.1, namely $\inf_{L\in\mathcal{L}_n} \mathbb{E}_{Q\in\mathcal{P}_n}\{L(X^{(n)})\} = H(Q) + O(1)$. Finally, the last equation follows by writing

$$\sup_{Q\in\mathcal{P}_n} \left\{\rho H(Q) - D(Q||p_{X^{(n)}})\right\} = (1+\rho) \sup_{Q\in\mathcal{P}_n} \left\{\mathbb{E}_Q\left[-\frac{\rho}{1+\rho} \ln p_{X^{(n)}}(X^{(n)})\right] - D(Q||p_{X^{(n)}})\right\},$$

and then applying the Donsker–Varadhan variational formula of Corollary 21.17, as in the first equation, with g replaced by $\ln p_{X^{(n)}}(X^{(n)})$, λ replaced by $\frac{1}{1+\rho}$, P replaced by Q to get the scaled Rényi entropy. That is,

$$\begin{aligned}&\sup_{Q\in\mathcal{P}_n} \{\rho H(Q) - D(Q||p_{X^{(n)}})\} + O(1)\\ &= \sup_{Q\in\mathcal{P}_n} \left\{-\rho \sum_x Q(x) \ln Q(x) - \sum_x Q(x) \ln Q(x) + \sum_x Q(x) \ln p_{X^{(n)}}(x)\right\}\\ &= \sup_{Q\in\mathcal{P}_n} \left\{\sum_x Q(x) \ln p_{X^{(n)}}(x) - (1+\rho) \sum_x Q(x) \ln Q(x)\right\}\end{aligned}$$

$$
\begin{aligned}
&= (1+\rho) \sup_{Q \in \mathcal{P}_n} \left\{ \mathbb{E}_Q \frac{1}{1+\rho} \ln p_{X^{(n)}}(X^{(n)}) - D(Q \| p_{X^{(n)}}) \right\} + O(1) \\
&= (1+\rho) \ln \mathbb{E} p_{X^{(n)}}^{\frac{1}{1+\rho}-1}(X^{(n)}) + O(1) \\
&= \rho H_{\frac{1}{1+\rho}}(p_{X^{(n)}}) + O(1).
\end{aligned}
$$

Scale by $\frac{1}{n}$ and let $n \to \infty$ to complete the proof. ■

The *information spectrum* is defined by $-\frac{1}{n} \ln p_{X^{(n)}}(X^{(n)})$. The next step is to show that the Rényi entropy rate can be computed from the distributions of the information spectra.

Theorem 22.4 *(Hanawal and Sundaresan (2011)).* Let ν_n be the distribution of the information spectrum $-\frac{1}{n} \ln p_{X^{(n)}}(X^{(n)})$. If $\nu_n, n \geq 1$, satisfy a LDP with rate function I, then the limiting Rényi entropy rate of order $\alpha = \frac{1}{1+\rho}$ exists and is given by $\beta^{-1} \sup_{t \in \mathbb{R}}\{\beta t - I(t)\}$, where $\beta = \frac{\rho}{1+\rho}$.

Proof. Let ν_n denote the distribution of the information spectrum $\frac{1}{n} \ln p_{X^{(n)}}(X^{(n)})$. Then, with $A_n = \{-\frac{1}{n} \ln p_{X^{(n)}}(x) : x \in S^n\}$, one has

$$
\begin{aligned}
\int_{\mathbb{R}} \exp(n\beta t) \nu_n(dt) &= \sum_{t \in A_n} \exp(n\beta t) \sum_{\{x : p_{X^{(n)}}(x) = \exp(-nt)\}} p_{X^{(n)}}(x) \\
&= \sum_{x \in S^n} p_{X^{(n)}}(x)^{1-\beta} \\
&= \sum_{x \in S^n} p_{X^{(n)}}(x)^{\frac{1}{1+\rho}} \\
&= \exp\{\beta H_{\frac{1}{1+\rho}}(p_{X^{(n)}})\}.
\end{aligned} \tag{22.25}
$$

Now, scaling by $\frac{1}{n}$ and taking logarithms, one may apply the Varadhan integral formula to the left side to obtain in the limit $\beta^{-1} \sup_{t \in \mathbb{R}}\{\beta t - I(t)\}$, while one has on the right side $\beta \lim_n \frac{1}{n} H_{\frac{1}{1+\rho}}(p_{X^{(n)}})$. ■

Corollary 22.5. If the distributions of the information spectra satisfies a LDP with rate I, then the guessing exponent exists and is given by

$$
g(\rho) = (1+\rho) \sup_{t \in \mathbb{R}} \left\{ \frac{\rho}{1+\rho} t - I(t) \right\}.
$$

Proof. By Proposition 22.3 the limiting Rényi entropy is $\frac{g(\rho)}{\rho}$. Thus, one has $g(\rho) = \rho\beta^{-1} \sup_{t \in \mathbb{R}}\{\beta t - I(t)\} = (1+\rho) \sup_{t \in \mathbb{R}}\{\beta t - I(t)\}$. ■

First let us apply this theory to the case of i.i.d. message sources.

Theorem 22.6 *(I.I.D. Case).* Assume that $X_1, X_2, \ldots$ is i.i.d. with common probability mass function p on the finite alphabet S. Then, the limit defining the guessing exponent $g(\rho)$ exists and is given by $g(\rho) = (1+\rho)H_{\frac{1}{1+\rho}}(p)$, where $H_\alpha(p)$ denotes the Rényi entropy rate of order α of the probability mass function p.

Proof. From Proposition 22.3 one can compute $g(\rho)$ from the Rényi entropy rate which, in turn, is given by $\frac{1+\rho}{\rho}I^*(\frac{\rho}{1+\rho})$, where $I(\cdot)$ is the large deviation rate for the energy spectrum

$$-\frac{1}{n}\ln p_{X^{(n)}}(X^{(n)}) = -\frac{1}{n}\ln\prod_{j=1}^{n} p(X_j) = -\frac{1}{n}\sum_{j=1}^{n}\ln p(X_j).$$

In particular, $I(h) = c^*(h)$ is the Legendre transform of the cumulant generating function of $-\ln p(X_1)$, namely

$$c(h) = \ln \mathbb{E}e^{h(-\ln p(X_1))} = \ln \mathbb{E}p^{-h}(X_1) = hH_{1-h}(p).$$

Since the Legendre transform operation $*$ is idempotent (see Exercise 6), it follows that

$$I^*\left(\frac{\rho}{1+\rho}\right) = (c^*)^*\left(\frac{\rho}{1+\rho}\right) = c\left(\frac{\rho}{1+\rho}\right).$$

In particular, $g(\rho) = (1+\rho)\frac{\rho}{1+\rho}H_{\frac{1}{1+\rho}} = \rho H(\frac{1}{1+\rho})$, as asserted. ■

Theorem 22.7 *(Irreducible Markov Case).* Let $X_1, X_2, \ldots$ be an irreducible Markov chain on S with homogeneous transition probability matrix $p = ((p(y|x)))_{x,y\in S}$. Then the guessing exponent $g(\rho)$ exists and is given by

$$g(\rho) = (1+\rho)\lambda^+\left(\frac{\rho}{1+\rho}\right),$$

where $\lambda^+(h)$ is the largest eigenvalue of the matrix $((\pi^{1-h}(y|x)))_{x,y\in S}$.

Proof. As in the i.i.d. case, from Proposition 22.3 one can compute $g(\rho)$ from the Rényi entropy rate which, in turn, is given by $\frac{1+\rho}{\rho}I^*(\frac{\rho}{1+\rho})$, where $I(\cdot)$ is the large deviation rate for the energy spectrum

$$-\frac{1}{n}\ln p_{X^{(n)}}(X^{(n)}) = -\frac{1}{n}\left\{\ln p(X_1) + \sum_{j=1}^{n-1}\ln p(X_{j+1}|X_j)\right\}.$$

Note that $Y_j = (X_j, X_{j+1}), j = 1, 2, \ldots$ is also a stationary Markov chain with one-step transition probabilities

$$\tilde{p}((w,z)|(x,y)) = \begin{cases} p(z|y), & y = w, \\ 0, & y \neq w. \end{cases}$$

To compute $I(\cdot)$ it suffices to compute the large deviation rate for $\sum_{j=1}^{n} \varphi(Y_j)$, where $g(Y_j) = -\ln p(X_{j+1}|X_j)$. Let $v(x,y) = -\ln p(y|x)$, $(x,y) \in S \times S$. Then,

$$\begin{aligned} T_h f(x,y) &= \sum_{(w,z)\in S\times S} f(w,z)e^{-h\ln p(z|w)}\tilde{p}((w,z)|(x,y)) \\ &= \sum_{z\in S} f(y,z)p^{-h}(z|y)p(z|y) = \sum_{z\in S} f(y,z)p^{1-h}(z|y). \quad (22.26) \end{aligned}$$

Observe that $T_h g(x,y) = \lambda g(x,y), (x,y) \in S \times S$ implies $g(x,y) = g(y)$, i.e., is constant in x. In particular, $\lambda^+(h) = \lambda(1-h)$, where $\lambda(a)$ is the largest eigenvalue of the matrix $((p^a(y|x))_{(x,y)\in S\times S}$. In particular, $I(h) = \lambda^*(h)$. Again using idempotency, of the Legendre transform, $I^*(t) = \lambda(t)$. It follows that the entropy rate is given by $\frac{1+\rho}{\rho}\ln\lambda(\frac{\rho}{1+\rho})$, and therefore the guessing exponent is $g(\rho) = \rho\frac{1+\rho}{\rho}\ln\lambda(\frac{\rho}{1+\rho})$ where $\lambda(\frac{\rho}{1+\rho})$ is the largest eigenvalue of the matrix $((p^{\frac{1}{1+\rho}}(y|x)))_{(x,y)\in S\times S}$. ■

Remark 22.3. Alternative representations of the guessing exponent in both of these cases can be obtained by consideration of level-2 large deviations as given in Hanawal and Sundaresan (2011). Moreover, the computation of the guessing exponent by these methods for other general classes of message sources can be found there.

The Kraft inequality for lengths plays an essential role in this application, specifically in Theorem 22.1 and its application in the proof of Proposition 22.3. In the classic monograph of Shannon (1948) messages are defined as sequences from a finite alphabet S, referred to as ciphers.[3] In the context of message compression, for a positive integer b one often defines a *b-ary coding function* as an injective map $c : S^n \to \cup_{m=1}^{\infty}\{0,1,\ldots,b-1\}^m$ that renders a message $x \in S^n$ of length n, as a b-ary sequence $c(x)$ of length m for some m. One seeks codes c for which the average length $\mathbb{E}L(X^{(n)})$, of a message $X^{(n)}$, is minimal. A b-ary coding function is said to be *prefix-free* (or instantaneous) if for $x \neq y$ $c(x)$ is not a prefix of $c(y)$. A prefix-free code may be represented as leaves on a rooted b-ary tree obtained by coding the path from the root to the leaves (terminal vertices) with labels $\{0,1,2,\ldots,b-1\}$ from left to right at each level of the tree. Therefore, a prefix-free codeword can be instantaneously decoded without reference to future codewords since the end of a codeword is immediately recognizable as a leaf.

[3] The textbook by Cover and Thomas (2006) provides a good foundation for the general concepts and results encountered in information theory.

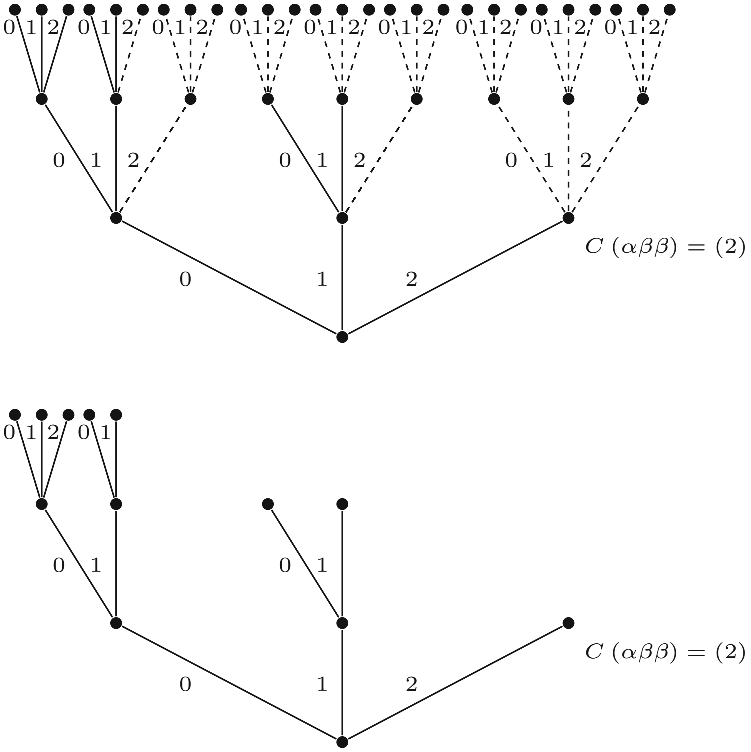

Fig. 22.1 Prefix-free code: $S = \{\alpha, \beta\}, n = 3, b = 3$; $L(\alpha, \beta, \beta) = 1$, $L(\alpha, \alpha, \alpha) = L(\alpha, \alpha, \beta) = 2$; $L(\beta, \beta, \beta) = \cdots = L(\beta, \alpha, \alpha) = 3$.

Proposition 22.8. Given any positive integers $L_1, \ldots, L_{|S|^n}$, satisfying Kraft inequality, there is a prefix-free b-ary code on S^n, $b \geq 3$, whose code words have lengths $L_1, \ldots, L_{|S|^n}$.

Proof. Observe that for positive integers $L(x), x \in S^n$, $b \geq 3$, since $2 < e < 3$, $\sum_{x \in S^n} b^{-L(x)} \leq 1$ if $\sum_{x \in S^n} \exp\{-L(x)\} \leq 1$. Let $m = |S|^n$, $L_{\max} = \max\{L_1, \ldots, L_m\}$ and construct a full rooted b-ary tree of height $L_{\max}$ for a $b \geq 3$. Then the total number of leaves available is $b^{L_{\max}}$, at vertices of height $L_{\max}$ having height one label from $\{0, 1, \ldots, b-2\}$ (see Figure 22.1). This uses $(b-1)b^{L_{\max}-1}$ of the leaves, with $b^{L_{\max}} - (b-1)b^{L_{\max}-1} = b^{L_{\max}-1}$ remaining for coding words having lengths at most $L_{\max} - 1$. Proceed inductively. ■

Remark 22.4. The prefix-free b-ary code constructed in the proof of Proposition 22.8 is referred to as the *Shannon code*. The units for message compression are referred to as "bits" when the logarithm is base 2, and "nats" for the natural logarithm. Natural logarithms are mathematically more convenient to the problem at hand and can be used without loss of generality.

The significance of Proposition 22.8 for the present chapter is that one may assume any given lengths, i.e., subject to the Kraft inequality, to be those of a prefix-free code.

Remark 22.5. The approximate code lengths $L(x) = \ln \frac{1}{p_{X(n)}}$ can be obtained as the solution to minimizing expected code lengths subject to Kraft inequality by the method of Lagrange multipliers (Exercise 5). However, as noted, these are not necessarily positive integers. The existence of an optimal code is a consequence of Theorem 22.1 and Proposition 22.8 by consideration of the prefix-free code associated with $L(x) = \lceil \ln \frac{1}{p_{X(n)}} \rceil$.

The next example illustrates a role for large deviation theory in large sample statistical inference.

Example 2 *(Efficiency of Statistical Tests of Hypothesis in Large Samples).* A common statistical test of hypothesis about an unknown parameter θ based on a random sample of size n from some distribution may be stated as follows:

The null hypothesis $H_0 : \theta \leq \theta_0$ is to be tested against the alternative hypothesis $H_1 : \theta > \theta_0$. The test is of the form: Reject H_0 (in favor of H_1) if $\overline{X} > a$, where $\overline{X}$ is the (sample) mean of i.i.d. variables $(X_1, \ldots, X_n)$ based on the random sample, and a is an appropriate number. The objective is to have small error probabilities $\alpha_n = P(\overline{X} > a|H_0)$, and $\beta_n = P(\overline{X} \leq a|H_1)$.

There are several competing notions for the *Asymptotic Relative Efficiency* (ARE) of such tests. For example, in the so-called *location problem*, the distribution function of X is of the form $F(x - \theta)$, $\theta \in \mathbb{R}$. In particular, F may be the normal distribution $N(\theta, 1)$. The *Normal test M* is of the form: Reject H_0 iff $\overline{X} > a$. The *t-test T* is of the form: Reject H_0 iff $\overline{X}/s > a$, where s is the sample standard deviation. The *Sign test S* is of the form: Reject H_0 iff $\frac{1}{n}\sum_{1 \leq i \leq n}[X_i - \theta_0 \geq 0] > a$. The a-values of these tests are not necessarily the same.

The most commonly used test ARE is the *Pitman ARE E_P* test,[4] which fixes a "small" level $\alpha_n = \alpha$, and compares two tests A, B, say, based on the smallness of their β_n. Specifically, the *Pitman ARE* of B with respect to A is $E_P(A, B) = n/h(n)$, where $h(n)$ is the sample size needed for B to attain the same level β_n as attained by A based on a sample size n. The asymptotics here are generally based on weak convergence, especially the CLT (central limit theorem).

The two other important AREs we discuss in detail here are mainly based on large deviations. **Chernoff**-ARE:[5] Based on large deviation estimates for each test A, B, Chernoff's (modified) test picks the value of a that minimizes $\alpha_n + \lambda\beta_n$ over all a for some fixed $\lambda > 0$. (It turns out the ARE does not depend on λ). The ratio of the large deviation rates $I(A), I(B)$ of decay of this minimum value δ_n, say, of $\alpha_n + \lambda\beta_n$ is compared for the tests A and B, and the *Chernoff ARE* of B with respect to A is $E_C(A, B) = I(B)/I(A)$.

[4] Serfling (1980), Chapter 10, Bhattacharya et al. (2016), Chapter 8.

[5] Serfling (1980), Chapter 10; Chernoff (1952).

Proposition 22.9. Assume $m_i(h) = \mathbb{E}(\exp\{hX_1\}|H_i) < \infty$ for all $-\infty < h < \infty (i = 0, 1)$. Let

$$c_i(a) = \sup\{ah - \ln m_i(h) : h \in \mathbb{R}\}(i = 0, 1), d(a) = \min\{c_0(a), c_1(a)\},$$
$$I = \max\{d(a) : \theta_0 \le a \le \theta_1\},$$

and $\rho = \exp\{-I\}$. Then

$$\lim_{n\to\infty} \frac{1}{n} \ln \delta_n = -I. \tag{22.27}$$

Proof. By the upper bound in the Cramér–Chernoff theorem (i.e., Chernoff's Inequality), $\alpha_n + \lambda\beta_n \le \exp\{-nc_0(a)\} + \lambda \exp\{-nc_1(a)\} \le (1 + \lambda) \exp\{-nd(a)\}$. Minimizing over a, one arrives at the inequality $\delta_n \le (1 + \lambda)\rho^n$, or $\frac{1}{n} \ln \delta_n \le -I + \frac{1}{n} \ln(1 + \lambda)$, and $\limsup_n \frac{1}{n} \ln \delta_n \le -I$. For the lower bound for δ_n, note that, by the Crameér–Chernoff theorem, $\liminf \frac{1}{n} \ln \alpha_n \ge -c_0(a)$, $\liminf_n \frac{1}{n} \ln \beta_n \ge -c_1(a)$. That is, given $\eta > 0$, for all sufficiently large n, $\min\{\alpha_n, \beta_n\} \ge \exp\{-n(d(a) + \eta)\}$, or $\alpha_n + \lambda\beta_n \ge (1 + \lambda) \exp\{-n(d(a) + \eta)\}$. Hence, taking the minimum over a, $\delta_n \ge (1 + \lambda) \exp\{-n(I + \eta)\}$, or $\frac{1}{n} \ln \delta_n \ge -(I + \eta) + \frac{1}{n} \ln(1 + \lambda)$; so that $\liminf \frac{1}{n} \ln \delta_n \ge -(I + \eta)$ for all $\eta > 0$. Hence $\liminf_n \frac{1}{n} \ln \delta_n \ge -I$. ■

The Location Problem Consider the tests M, T, S for the location problem for $F(x-\theta)$ described in the first paragraph. Assume that F has a density f, continuous at $\theta = 0$, and a finite variance σ_f^2. Then for the test $H_0 : \theta \le 0$, to be tested against the alternative hypothesis $H_1 : \theta \ge \theta_1 > 0$, one can show[6] that $E_P(S, M) = 4\sigma_f^2 f^2(0)$. In particular, (i) if F is $N(\theta, 1)$, then $E_P(S, M) = 2/\pi < 1$, (ii) if F is Double exponential (i.e., $f(x - \theta) = \frac{1}{2} \exp\{-|x - \theta|\}$), then $E_P(S, M) = 2$, and (iii) if f is uniform on $[-\frac{1}{2} - \theta, \frac{1}{2} - \theta]$, then $E_P(S, M) = 1/3$. In all these cases (and more broadly) $E_P(T, M) = 1$, where T is the t-test.

More interesting are Pitman comparisons among nonparametric tests for the so-called *two-sample problems*. Here two independent samples $(X_1, \ldots, X_m)$, $(Y_1, \ldots, Y_n)$ of sizes m and n are drawn from an unknown distribution whose density is of the form $f((x - \theta)/\sigma)$, $\theta \in \mathbb{R}$, $\sigma > 0$. One wishes to test $H_0 : \theta = 0$, against $H_1 : \theta > 0$. More generally, one wishes to test if the Y-distribution is stochastically larger than the X-distribution (i.e., $P(Y > z) \ge P(X > z)$ for all z, with strict inequality for at least some z). The most commonly used test for this uses the (nonparametric) statistic $T = \overline{Y} - \overline{X}$, which rejects H_0 if T exceeds a critical value. (The critical value is determined approximately by the CLT to meet the requirement $\alpha = P(Reject H_0|H_0)$). It turns out that appropriate nonparametric

[6] See Serfling (1980), Chapter 10; Bhattacharya and Waymire (2016), Chapter 8.

tests based on ranks of the combined observations X_i's and Y_j's, mostly outperform T.[7]

Chernoff Index Computation The Chernoff indices I are generally difficult to compute, since the indices try to minimize a linear combination of both error probabilities α_n and β_n. We consider the simple case where $F = N(\theta, 1)$, and $H_0 : \theta = 0$, $H_1 : \theta = \theta_1$, $\theta_1 > 0$. Again consider the test M: Reject H_0 if $\overline{X} > a$, otherwise Reject H_1. We leave it as an exercise to show that $I(M) = \theta_1^2/8$ (Exercise 8). For the sign test S: Reject H_0 if $\frac{1}{n}\sum_{1\leq i\leq n} \mathbf{1}[X_i > 0]) > a$ for some appropriate a, as considered in the discussion of the Chernoff-ARE above, one may compute the Chernoff index $I(S)$ from the distribution $B(n, p)$ with $p = \Phi(\theta_1)$, Φ being the distribution function of $N(0, 1)$. Namely,

$$I(S) = \ln\{2(b(\theta_1))^{b(\theta_1)}(1 - b(\theta_1))^{1-b(\theta_1)}\}, \tag{22.28}$$

where $b(\theta) = \ln[1 - \Phi(\theta)]/[(\ln\{(1 - \Phi(\theta))/\Phi(\theta)\}]$ (Exercise 10). The ratio $I(S)/I(M)$ provides the Chernoff-ARE $E_C(S, M)$. One may check that $E_C(S, M) \to 2/\pi = E_P(S, M)$ as $\theta_1 \downarrow 0$ (Exercise 11).

Bahadur-*ARE* As mentioned above, the Chernoff-ARE is generally difficult to compute. In addition, the threshold of the test itself is modified by the requirement of this notion of efficiency. The most popular ARE for tests based on large deviations is due to Bahadur (1960). Here is a brief description following Serfling (1980). The Bahadur-ARE is based on a large deviation rate comparison of the p-values of the tests. Consider a test of hypothesis $H_0 : \theta \in \Theta_0$, with a real-valued test statistic T_n based on observations $X_1, \ldots, X_n$, rejecting H_0 if T_n is large. The p-value of the *test* is $L_n = \sup[1 - F_{\theta_n}(T_n) : \theta \in \Theta_0] = 1 - F_{\theta_n^0}(T_n)$, say, where F_{θ_n} is the distribution function of the statistic under the parameter value θ. Thus L_n is the random quantity which is the probability (under H_0) of the statistic being larger than what is observed, i.e., of showing a discrepancy from the null hypothesis as large or larger than what is observed. Statisticians routinely use L_n to decide whether to reject H_0: smaller the p-value, stronger is the evidence against H_0. Under H_0, assuming that the distribution function $F_{\theta_n^{(0}}$ of T_n is continuous, $F_{\theta_n^{(0)}}(T_n)$ has the uniform distribution on $[0, 1]$, and so is the distribution of $L_n = 1-F_{\theta_n^{(0)}}(T_n)$. Under fairly general conditions, $-2n^{-1}\ln L_n$ converges almost surely to a constant $c(\theta)$, which is referred to as *Bahadur's (exact) slope* for T_n, for $\theta \in \Theta_1$. The Bahadur relative efficiency of a test I with respect to test II is defined by the ratio of their corresponding slopes (a.s. large deviation rates) $e_B(I, II) = c_I(\Theta)/c_{II}(\Theta)$.

The following is a basic result which may be used to compute the slope of tests such as $H_0 : \theta \leq \theta_0$, against $H_1 : \theta > \theta_0$.[8] Write $\Theta_1 = \Theta\backslash\Theta_0$.

[7] Bhattacharya et al. (2016), Chapter 8.

[8] We follow Serfling (1980), Chapter 10, for the proof of the following result of Bahadur (1960; 1971).

Theorem 22.10 *(Bahadur (1960)).* For a test sequence T_n which rejects H_0 for large T_n, assume (i) $n^{-\frac{1}{2}}T_n$ converges a.s. (under θ) to a finite $b(\theta)$, for all $\theta \in \Theta_1$, and (ii) one has

$$\lim_{n\to\infty} -2n^{-1}\ln\sup[1 - F_{\theta_n}(n^{\frac{1}{2}}t) : \theta \in \Theta_0] = g(t), \tag{22.29}$$

where g is continuous on an open interval I containing $\{b(\theta) : \theta \in \Theta_1\}$. Then $\forall\, \theta \in \Theta_1$, with P_θ-probability one,

$$\lim_{n\to\infty} -2n^{-1}\ln L_n = g(b(\theta)) = c(\theta). \tag{22.30}$$

Proof. Fix a $\theta \in \Theta_1$, and let ω be any point in the sample space of P_θ for which the limit (i) holds. Fix $\epsilon > 0$ sufficiently small that $(b(\theta)-\epsilon, b(\theta)+\epsilon)$ is contained in I. By (i), there exists $n = n(\omega)$ such that $b(\theta)-\epsilon \le n^{-\frac{1}{2}}T_n(\omega) \le b(\theta)+\epsilon$ for all $n \ge n(\omega)$, i.e., $n^{\frac{1}{2}}(b(\theta)-\epsilon) \le T_{n(\omega)} \le n^{\frac{1}{2}}(b(\theta)+\epsilon)$ for all $n \ge n(\omega)$. Plugging these in $-2n^{-1}\ln\sup[1 - F_{\theta_n}(n^{\frac{1}{2}}t) : \theta \in \Theta_0]$, one then has

$$\begin{aligned} &-2n^{-1}\ln\sup[1 - F_{\theta_n}(b(\theta)-\epsilon)) : \theta \in \Theta_0]) \\ &\quad\le -2n^{-1}\ln L_n(\omega) \\ &\quad\le -2n^{-1}\ln\sup[1 - F_{\theta_n}(b(\theta)+\epsilon)) : \theta \in \Theta_0])\forall\, n \ge n(\omega). \end{aligned} \tag{22.31}$$

The limits as $n \to \infty$ of the two extreme sides are $g(b(\theta)-\epsilon)$ and $g(b(\theta)+\epsilon)$. Therefore, the limit points of the middle term in (22.31) all lie in this interval. By continuity of g, it follows that the middle term converges to $g(b(\theta))$. ■

The exact Bahadur slopes for the mean test M and the t-test T may be computed for testing $H_0 : \theta \le 0$, versus the alternative $H_1 : \Theta > 0$ in the model $N(\theta, 1)$, using the upper tail of the standard normal $N(0, 1)$, and that of the (Student's) t-statistic with $n-1$ degrees of freedom. Using Bahadur's theorem, one finds $c_M(\theta) = \theta^2$, $c_T(\theta) = \ln(1+\theta^2)$ (Exercise 12). Thus $e_B(T, M) < 1$ for all $\theta \in \Theta_1$. This is in contrast with both Pitman's ARE and Chernoff's ARE, for each of which the ARE is one.

Remark 22.6. Bahadur's ARE also distinguishes between the*frequency chi-square* and the *likelihood ratio test* in the multinomial model, showing the latter is asymptotically more efficient than the former. Again the Pitman ARE is one between the two tests.[9]

[9] Abrahamson (1965).

Exercises

1. (Shannon & Renyi Entropies) Show that the Shannon entropy H may be expressed in terms of the Renyi entropy as

$$H(X_1, \dots, X_n) = \lim_{\alpha \to 1} H_\alpha(X_1, \dots, X_n).$$

2. Complete the proof of (22.8) by showing that for random vectors $H(X, Y) \leq H(X) + H(Y)$. [*Hint*: Show how to express $H(X) + H(Y) - H(X, Y)$ as $D(p_{(X,Y)} || p_X \odot p_Y) \geq 0$, where $p_{(X,Y)}, p_X, p_Y$ are the joint and marginal distributions, respectively.
3. Let P, Q be probability measures on S^n. For convenience relabel $S^n = \{1, \dots, k\}, k = |S|^n$, $q_j = Q(\{j\}), p_j = P(\{j\})$. Prove Gibbs inequality: $\sum_j p_j \ln p_j \geq \sum_j p_j \ln q_j$. [*Hint*: Consider $\sum_j p_j \ln \frac{q_j}{p_j}$ and bound $\ln x \leq x - 1, x > 0$.
4. Give a proof of the Kraft inequality for the message length associated with G. [*Hint*: $\sum_x e^{\lceil \ln Q_G(x) \rceil} \leq \sum_x e^{\ln Q_G(x)}$.
5. Show that the problem $\min_{L \in \mathcal{L}_n : \sum_x e^{-L(x)} \leq 1} \sum_{x \in S^n} p_{X^{(n)}} L(x)$ has a solution. [*Hint*: Use Lagrange multipliers to minimize $J = \sum_{x \in S^n} p_{X^{(n)}} L(x) + \lambda \sum_{x \in S^n} e^{-L(x)}$. Derivatives with respect to each $L(x)$ are zero and λ can be determined from the constraint $\sum_x e^{-L(x)} \leq 1$.
6. Let u, v be twice-continuously differentiable functions on $\mathbb{R}$ with Legendre transforms u^*, v^*, respectively, where $f^*(x) = \sup_{h \in \mathbb{R}}\{xh - f(h)\}$, $x \in \mathbb{R}$. Show that (a) u^* is convex. (b) (Idempotency) $u^{**} = u$[*Hint*: Write $u^*(x) = xh(x) - u(h(x))$ and use the smoothness hypothesis on u to optimize.
7. Give a proof for (22.6).
8. Show that $I(M) = \theta_1^2/8$.
9. Give a proof of (22.6).
10. Verify the hypotheses (i),(ii) in Theorem 22.10 for the tests M, T, S. [Hint: For M, let $T_n = n^{\frac{1}{2}}(\overline{X} - \theta_0)$, then (i) is satisfied, since for $\theta > \theta_0$, $n^{-\frac{1}{2}} T_n \to \theta - \theta_0$. For assumption (ii) assume that X_j has a finite moment generating function, and use the Cramér-Chernoff large deviation rate. A similar, but little longer, proof applies to the statistic T, using independence of the sample mean and sample variance. For S one uses the moment generating function of Bernoulli variables.]
11. Show that $E_C(S, M) \to 2/\pi = E_P(S, M)$ as $\theta_1 \downarrow 0$.
12. Show using Bahadur's theorem, that $c_M(\theta) = \theta^2$, $c_T(\theta) = \ln(1 + \theta^2)$.