Asymptotic Estimates in Shannon's Information Theory

V. Strassen*

August 26, 2009

1 Introduction

Let X be a finite set and let p a probability distribution on X. Let

$$X_n = \left\{x_n = (x^1 \ldots, x^n) : x^k \in X\right\}$$

be the usual cardinal product of n copies of X and let p_n be the independent product on X_n of n copies of p. One calls the sequence $(p_n)_{n\geq 1}$ an (Information) Source with stationary, independent letters from the alphabet X. For a fixed ϵ, $0 < \epsilon < 1$, one is interested in the smallest cardinality $\beta(n, \epsilon)$ of subsets $E \subseteq X_n$ with $p_n(E) \geq 1 - \epsilon$. Let

$$h(x) = -\log p(x).$$

One calls the quantity

$$H = -\int hdp = -\sum_{p(x)>0} p(x) \log p(x)$$

the entropy of the source $(p_n)_{n\geq 1}$.

It is well-known that

$$\log \beta(n, \epsilon) = nH + O(\sqrt{n}). \tag{1.1}$$

(One may find a summary of the literature about this result and further generalizations in [2]). Recall that e^{nH} is an approximation of $\beta(n, \epsilon)$ with relative error $e^{O(\sqrt{n})}$, and note that this error could potentially go very rapidly to infinity.

Let S be the variance of h. We will write Φ to denote the Gaussian distribution. Let λ be the unique positive real number determined by

$$\Phi(\lambda) = 1 - \epsilon. \tag{1.2}$$

*This English translation was made from the original German text by Peter Luthy. This translation would not have been possible without support from Aaron Wagner.

Juschkewitsch [3] proved the following tighter version of (1.1) (not only for sources with stationary, independent letters, but rather for the more general case for stationary Markovian sources):

$$\log \beta(n, \epsilon) = nH + \lambda S\sqrt{n} + o(\sqrt{n}).$$

We will exclude the trivial case $S = 0$. Define the real polynomial Q by

$$Q(t) = \frac{\int (h - H)^3 dp}{6S^2}(t^2 - 1). \tag{1.3}$$

For a given $\varrho > 0$, let $d(t)$ be the smallest remainder of t mod ϱ so that

$$d(t) = t - k\varrho \qquad \left(1 - \tfrac{\varrho}{2} < t - k\varrho \leq \tfrac{\varrho}{2}\right),$$

and define w, with help from d, by

$$w = -d + \log\left(d + \frac{\varrho}{2}\frac{e^\varrho + 1}{e^\varrho - 1}\right). \tag{1.4}$$

The quantity w is a continuous, periodic (hence bounded) function with period ϱ.

Theorem 1.1 *Provided that the distribution of h is not lattice-like (see [4]):*

$$\log \beta(n, \epsilon) = nH + \lambda S\sqrt{n} - \frac{1}{2}\log n + Q(\lambda) - \frac{1}{2}\lambda^2 - \log(\sqrt{2\pi}S) + o(1). \tag{1.5}$$

If the distribution of h is lattice-like with lattice parameter (span) ϱ, then

$$\begin{aligned}\log \beta(n, \epsilon) = nH + \lambda S\sqrt{n} - \frac{1}{2}\log n + Q(\lambda) - \frac{1}{2}\lambda^2 - \log(\sqrt{2\pi}S)\\ + w\left(-na + \lambda S\sqrt{n} + Q(\lambda)\right) + o(1)\end{aligned}$$

and where a is a real number so that $p(h - H = a) > 0$.

From Theorem 1, it follows that the relative error of both

$$\frac{1}{S\sqrt{2\pi n}}\exp\{nH + \lambda S\sqrt{n} + Q(\lambda) - \frac{1}{2}\lambda^2\}$$

and

$$\frac{1}{S\sqrt{2\pi n}}\exp\{nH + \lambda S\sqrt{n} + Q(\lambda) - \frac{1}{2}\lambda^2 + w\left(-na + \lambda S\sqrt{n} + Q(\lambda)\right)\}$$

as approximations of $\log \beta(n, \epsilon)$ go to zero as n tends to infinity. Now, let Y be second finite set, $Y_n = \left\{y_n = (y^1 ..., y^n) : y^k \in Y\right\}$, and P a Markov kernel (transition probability function) from Y to X. For every $y \in Y$, the function $P(\cdot\,, y)$ is therefore a probability distribution in X. Rather than writing $P(\{x\}, y)$, we will write $P(x, y)$. The independent product P_n of n copies of P is then defined by

$$P_n(x_n, y_n) = P(x_1, y_1) \dots P(x_n, y_n).$$

One calls the sequence $(P_n)_{n \geq 1}$ a stationary, memoryless channel from Y to X.

An ϵ-code for P_n is a mapping f of a subset of X_n to Y_n such that the following holds for every $y_n \in f(X_n)$:

$$P_n(f^{-1}\{y_n\}, y_n) \geq 1 - \epsilon. \tag{1.6}$$

The number of points in $f(X_n)$ is called the length of the code. One is interested in the maximal length $N(n, \epsilon)$ of ϵ-codes for P_n. For any probability distribution α in Y we define the probability distribution $P\alpha$ in X and $P \times \alpha$ in $X \times Y$ by

$$(P\alpha)\{x\} = \int P(x, y)\alpha(dy) = \sum_{y \in Y} P(x, y)\alpha\{y\} \tag{1.7}$$

and

$$(P \times \alpha)\{(x, y)\} = P(x, y)\alpha\{y\}. \tag{1.8}$$

Let

$$i_\alpha(x, y) = \log \frac{P(x, y)}{(P\alpha)\{x\}} \tag{1.9}$$

and

$$I_\alpha = \int i_\alpha d(P \times \alpha) = \sum_{x,y: P(x,y)\alpha\{y\} > 0} P(x, y)\alpha\{y\} \log \frac{P(x, y)}{(P\alpha)\{x\}}, \tag{1.10}$$

and let

$$C = \sup_\alpha I_\alpha. \tag{1.11}$$

One calls C the capacity of the channel $(P_n)_{n \geq 1}$. It is well known that

$$nC + O(\sqrt{n}) < \log N(n, \epsilon) < nC + \epsilon O(n).$$

(One can find an overview of the literature concerning the Coding Theorem with weak converse in [2],[5],[6],[8], and [16].)

Wolfowitz [5] observed and proved that if, in the above estimate, one replaced the $\epsilon O(n)$ with $O(\sqrt{n})$ (strong converse) so that

$$\log N(n, \epsilon) = nC + O(\sqrt{n})$$

then for $\epsilon > \frac{1}{2}$ the term $O(\sqrt{n})$ is eventually positive.

For a special class of stationary, memoryless channels (symmetric channels with two-point input alphabet), Weiss [7] showed that for $\epsilon < \frac{1}{2}$, the term $O(\sqrt{n})$ will eventually be negative; more precisely,

$$\log N(n,\epsilon) \leq nC - T\lambda\sqrt{n} + o(\sqrt{n}) \tag{1.12}$$

holds, where $T > 0$ is the variance of i_α relative to $P \times \alpha$ where we require the condition $I_\alpha = C$, holding for a uniquely defined α (aside from some trivial, degenerate cases which we disregard).

For stationary, memoryless channels, the set

$$\bar{A} = \{\alpha | \alpha \text{ is a probability distribution in } Y, I_\alpha = C\} \tag{1.13}$$

is generally not a single point. Let G_α be the variance of i_α relative to $P \times \alpha$ and

$$T_1 = \min_{\alpha \in \bar{A}} G_\alpha \tag{1.14}$$

$$T_{-1} = \max_{\alpha \in \bar{A}} G_\alpha$$

Theorem 1.2 *It holds that*

$$\log N(n,\epsilon) = nC - \lambda T_{\text{sign}\lambda}\sqrt{n} + O(\log n),$$

more precisely

$$\log N(n,\epsilon) \leq nC - \lambda T_{\text{sign}\lambda}\sqrt{n} + |Y|\log n \text{ for sufficiently large } n, \tag{1.15}$$

$$\log N(n,\epsilon) \geq nC - \lambda T_{\text{sign}\lambda}\sqrt{n} + O(1) \text{ for } T_{\text{sign}\lambda} > 0, \tag{1.16}$$

$$\log N(n,\epsilon) \geq nC - \frac{1}{2}\log n \text{ for sufficiently large } n, T_{\text{sign}\lambda} = 0, \tag{1.17}$$

where $|Y|$ is the cardinality of Y and sign $\lambda = 1$ *($\lambda \geq 0$) and* sign $\lambda = -1$ *($\lambda < 0$).*[1]

Theorem 1.2 shows among other things that the upper bound for $\log N(n,\epsilon)$ in (1.12) is the best possible within the context of the given error term.

From Theorem 1.1 and 1.2, together, it follows that (Section 5) a stationary, memoryless Channel with capacity C from a signal with stationary source and independent letters and with entropy $H = C$ can only be transmitted with large probability of error (but we assume that one of the numbers S, T_1 is positive).

One can generalize Theorem 1.1 to non-stationary sources with independent letters by relaxing the $o(1)$ to $O(1)$ (Section 3). The corresponding statement for Theorem 1.2 also follows (but this is not proven here, cf. also [8]).

[1]For symmetric Channels, one can replace $|Y|\log n$ in (1.15) with $O(1)$ if $T = 0$ or $\frac{1}{2}\log n + O(1)$ if $T > 0$. The details in this case are greatly simplified.

For channels that are stationary and memoryless and satisfy a slightly stronger symmetry condition than in [7], Dobruschin [6] had given, without proof, the estimate

$$\log N'(n,\epsilon) = nC - \lambda T\sqrt{n} - \frac{1}{2}\log n + O(1), \tag{1.18}$$

where $N'(n,\epsilon)$ is defined as in Section 5 (iv). From (5.5) and (1.16) it follows that (1.18) is not correct in general.[1]

The details of the proofs of Theorems 1.1 and 3.1 employ essential ideas of Cramér [18] (see also Feller [19]). The proof of Theorem 1.2 is supported by methods from Feinstein and Wolfowitz. A series of work (Feinstein [20], Elias [21], Shannon, Fano [2]) concerned with the estimate of ϵ_n when $N(n,\epsilon_n)$ is given (for a stationary, memoryless channel). Theorem 1.2 can be adapted to this problem, provided $N(n,\epsilon_n)$ grows like e^{nC}. Namely, it follows that

$$N(n,\epsilon_n) = e^{nC - K_n\sqrt{n}},$$

where the K_n are bounded (not necessarily ≥ 0) and we suppose that $T_1 > 0$, so it follows easily from Theorem 1.2 that

$$\epsilon_n = \frac{1}{\sqrt{2\pi}}\int_{K_n/T_{\text{sign } K_n}}^{\infty} e^{-t^2/2}dt + O(\frac{\log n}{\sqrt{n}})$$

(One remarks that Theorem 1.2 applies uniformly for ϵ from any compact set in the open unit interval). [20] to [22] provide in this case less exact asymptotic estimates, whereupon e.g. ϵ_n could go continuously to zero (the stated results are, however, broad for finite n and can apply to an arbitrary sequence $N(n,\epsilon_n)$).

2 Formulation and Proof of A Generalization of Theorem 1.1

Let q be a finite measure in X different from p, and assume q is absolutely continuous with respect to p. Let q_n be the product measure in X_n of n copies of q, and let

$$\beta(n,\epsilon) = \min_{E\subseteq X_n, p_n(E)\geq 1-\epsilon} q_n(E) \tag{2.1}$$

We define the function h on X by

$$h(x) = -\log\frac{p(x)}{q(x)} = -\log\frac{dp}{dq}(x) \tag{2.2}$$

[1] after a communication from Mr. Prof. Dobruschin, one must put $+\frac{1}{2}\log n$ in lieu of $-\frac{1}{2}\log n$. From this and from (5.5) and (5.7) one gets for $T > 0$ exactly the result in the previous addendum and (1.16) (under Dobruschin's symmetry condition). Provided Dobruschin's result is true in the case where $T = 0$ (which appears doubtful), one obtains through comparison to the previous addendum and (5.7) even that

$$\log N(n,\epsilon) = nC + O(1)$$

for such channels.

and set

$$H = \int h dp = -\sum_{x:p(x)>0} p(x) \log \frac{p(x)}{q(x)} \tag{2.3}$$

Let S be the variance of h relative to p (in what follows, for a distribution or the moments of a function defined on X or X_n, this quantity will always be with respect to p or p_n) and let Q be defined formally as in (1.3), but with the present meaning of h, H, and S. We prove Theorem 1.1 in this new interpretation, where the other relevant quantities are defined by (2.1) and (2.2). One can then choose the measure q to be the counting measure so that one obtains Theorem 1.1 as a special case.

In the case that q is a probability measure, $\beta(n, \epsilon)$ is the probability of type II error of an optimal non-random test with probability of type I error ϵ, null hypothesis p_n, and alternate hypothesis q_n.

Let

$$h_n = -\log \frac{dp_n}{dq_n}$$

and

$$h_n(x_n) = \sum_{i=1}^{n} h(x^i) \tag{2.4}$$

p_n–almost everywhere, and let $\mu_n > 0$ chosen so that

$$p_n \left\{ \frac{dp_n}{dq_n} \geq \mu_n \right\} \leq 1 - \epsilon \tag{2.5}$$

$$p_n \left\{ \frac{dp_n}{dq_n} > \mu_n \right\} < 1 - \epsilon$$

Let

$$\bar{\beta}(n, \epsilon) = \min_{\int f dp_n \geq 1-\epsilon} \int f dq_n, \tag{2.6}$$

where f is a variable of functions on X_n taking values in the closed unit interval (in the case where q is a probability distribution, $\bar{\beta}(n, \epsilon)$ is the probability of type II errors of an optimal randomized test). From the Lemma of Neyman and Pearson [9] it follows [2]

$$\begin{aligned} q_n \left\{ \frac{dp_n}{dq_n} > \mu_n \right\} \leq \bar{\beta}(n, \epsilon) = q_n \left\{ \frac{dp_n}{dq_n} > \mu_n \right\} + \\ + \frac{1}{\mu_n} \left(1 - \epsilon - p_n \left\{ \frac{dp_n}{dq_n} < \mu_n \right\} \right) \leq q_n \left\{ \frac{dp_n}{dq_n} \geq \mu_n \right\} \end{aligned} \tag{2.7}$$

[2]For this relation, I am very thankful to Honorable Dr. W. Fieger for his critical (in the helpful sense) remarks

and hence

$$\bar{\beta}(n,\epsilon) \le \beta(n,\epsilon) \le \bar{\beta}(n,\epsilon) + \frac{1}{\mu_n} \max_{X_n} p_n(x_n).$$

One remarks that since $S > 0$, $\max_{x_n} p_n(x_n)$ goes exponentially to zero and one uses the asymptotic approximation for μ_n, so that by comparison with Theorem 1.1, it suffices to prove this theorem for $\bar{\beta}(n,\epsilon)$ in lieu of $\beta(n,\epsilon)$.

Now, it is true that

$$q_n\left\{\frac{dp_n}{dq_n} > \mu_n\right\} = \int_{dp_n/dq_n > \mu_n} \frac{1}{dp_n/dq_n} dp_n = \int_{h_n < -\log\mu_n} \exp\{h_n\} dp_n =$$
$$= \exp\{nH\} \int_{t<\lambda_n} \exp\{tS\sqrt{n}\} dF_n(t)$$

with

$$F_n = p_n\left\{\frac{h_n - nH}{S\sqrt{n}} \le t\right\} \tag{2.8}$$

and

$$\lambda_= \frac{-\log\mu_n - nH}{S\sqrt{n}} \tag{2.9}$$

Therefore,

$$q_n\left\{\frac{dp_n}{dq_n} > \mu_n\right\} = \exp\{nH\} \int_{t<\lambda_n} \exp\{tS\sqrt{n}\} dF_n(t) = \tag{2.10}$$
$$= \exp\{nH + \lambda_n S\sqrt{n}\} \int_{t<\lambda_n} \exp\{(t-\lambda_n)S\sqrt{n}\} dF_n(t) =$$
$$= \exp\{nH + \lambda_n S\sqrt{n}\} \int_{z<0} e^z dF_n\left(\frac{z}{S\sqrt{n}} + \lambda_n\right).$$

Likewise, one obtains

$$q_n\left\{\frac{dp_n}{dq_n} \ge \mu_n\right\} = \exp\{nH + \lambda_n S\sqrt{n}\} \int_{z\le 0} e^z dF_n\left(\frac{z}{S\sqrt{n}} + \lambda_n\right), \tag{2.11}$$

and therefore by (2.7)

$$\int_{z>0} e^z dF_n\left(\frac{z}{S\sqrt{n}} + \lambda_n\right) \le \bar{\beta}(n,\epsilon) \exp\{-nH - \lambda_n S\sqrt{n}\} \le \tag{2.12}$$
$$\le \int_{z\le 0} e^z dF_n\left(\frac{z}{S\sqrt{n}} + \lambda_n\right).$$

From (2.8) and (2.9) it follows that

$$F_n(\lambda_n) = p_n\left\{\frac{dp_n}{dq_n} \ge \mu_n\right\} \tag{2.13}$$
$$F_n(\lambda_n - 0) = p_n\left\{\frac{dp_n}{dq_n} > \mu_n\right\},$$

therefore as a consequence of (2.5),

$$F_n(\lambda_n) \geq 1-\epsilon \qquad (2.14)$$
$$F_n(\lambda_n - 0) < 1-\epsilon.$$

Due to (2.4), F_n is the distribution function of the normalized sum of n independent and identically distributed random variables (mean 0 and variance 1).

Let h have a non-lattice-like distribution. By a well-known theorem of Cramér and Esseen ([11],[4]), one has

$$F_n(t) = \Phi(t) + \frac{1}{\sqrt{2\pi n}} e^{-t^2/2} Q_1(t) + o\left(\frac{1}{\sqrt{n}}\right), \qquad (2.15)$$

uniformly in t, where

$$Q_1(t) = \frac{1}{6S^3}\int (h-H)^3 dp(t-t^2) = \frac{1}{S}Q(t).$$

We set

$$B(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2} Q_1(t) = \Phi'(t) Q_1(t) \qquad (2.16)$$

and obtain

$$\begin{aligned} \int_{z>0} e^z dF_n\left(\frac{z}{S\sqrt{n}}+\lambda_n\right) &= \int_{z<0} e^z d\Phi\left(\frac{z}{S\sqrt{n}}+\lambda_n\right) + \\ &+ \frac{1}{\sqrt{n}}\int_{z<0} e^z dB\left(\frac{z}{\sqrt{n}}+\lambda_n\right) + o\left(\frac{1}{\sqrt{n}}\right) = \\ = \frac{1}{\sqrt{n}}\Bigg(\frac{1}{S\sqrt{2\pi}}\int_{z<0} &\exp\left\{z - \frac{1}{2}\left(\frac{z}{S\sqrt{n}}+\lambda_n\right)^2\right\} dz + B(\lambda_n) - \\ &- \int_{z<0} e^z B\left(\frac{z}{S\sqrt{n}}+\lambda_n\right) dz + o(1)\Bigg) = \\ &= \frac{1}{\sqrt{n}}\left(\frac{1}{S\sqrt{n}} e^{-t^2/2} + o(1)\right) = \frac{1}{S\sqrt{2\pi n}} e^{-\lambda^2/2 + o(1)} \end{aligned} \qquad (2.17)$$

then by (1.2), (2.14) and the central limit theorem one has

$$\lim_{n\to\infty} \lambda_n = \lambda. \qquad (2.18)$$

Likewise, one deduces

$$\int_{z\leq 0} e^z dF_n\left(\frac{z}{S\sqrt{n}}+\lambda_n\right) = \frac{1}{S\sqrt{2\pi n}} \exp\left\{-\frac{\lambda^2}{2} + o(1)\right\}. \qquad (2.19)$$

From (2.17), (2.19), and (2.12) it follows that

$$\bar{\beta}(n,\epsilon) = \frac{1}{S\sqrt{2\pi n}} \exp\left\{nH + \lambda_n S\sqrt{n} - \frac{\lambda^2}{2} + o(1)\right\} \qquad (2.20)$$

We still estimate λ_n. Let

$$\Delta\lambda_n = \lambda_n - \lambda$$

$$\Delta\Phi_n = \Phi(\lambda_n) - \Phi(\lambda) = \Phi(\lambda_n) - (1-\epsilon).$$

From (2.15), (2.16), (2.14), and (1.2) it follows that

$$\Phi(\lambda_n) + \frac{1}{\sqrt{n}}B(\lambda_n) + o\left(\frac{1}{\sqrt{n}}\right) = F_n(\lambda_n - 0) < \Phi(\lambda) \leq F_n(\lambda_n) =$$

$$= \Phi(\lambda_n) + \frac{1}{\sqrt{n}}B(\lambda_n) + o\left(\frac{1}{\sqrt{n}}\right)$$

therefore

$$\Delta\Phi_n = -\frac{1}{\sqrt{n}}B(\lambda_n) + o\left(\frac{1}{\sqrt{n}}\right) = -\frac{1}{\sqrt{n}}B(\lambda) + o\left(\frac{1}{\sqrt{n}}\right) = \tag{2.21}$$

$$= -\frac{1}{\sqrt{n}}\Phi'(\lambda)Q_1(\lambda) + o\left(\frac{1}{\sqrt{n}}\right)$$

owing to (2.18), the continuity of B, and (2.16). On the other hand, one has

$$\Delta\Phi_n = \Phi'(\lambda)\Delta\lambda_n + o(\Delta\lambda_n) = \Phi'(\lambda)\Delta\lambda_n + o(\Delta\Phi_n)$$

and in any case by (2.21)

$$\Delta\Phi_n = O\left(\frac{1}{\sqrt{n}}\right),$$

and

$$\Delta\Phi_n = \Phi'(\lambda)\Delta\lambda_n + o\left(\frac{1}{\sqrt{n}}\right).$$

From this and from (2.21) it follows that

$$\Delta\lambda_n = -\frac{1}{\sqrt{n}}Q_1(\lambda) + o\left(\frac{1}{\sqrt{n}}\right),$$

therefore

$$\lambda_n = \lambda + \frac{1}{S\sqrt{n}}Q(\lambda) + o\left(\frac{1}{\sqrt{n}}\right). \tag{2.22}$$

Together with (2.20) this finally gives

$$\bar{\beta}(n,\epsilon) = \frac{1}{S\sqrt{2\pi n}} \exp\left\{nH + \lambda S\sqrt{n} + Q(\lambda) - \frac{\lambda^2}{2} + o(1)\right\}.$$

This proves the statements of the theorem for non-lattice-like distributions h.

Let h now be lattice-like distributed with maximal step-size ϱ and let a a real number with $p(h - H = a) > 0$. By a well-known theorem of Esseen ([11], [4]) we have

$$F_n(t) = \Phi(t) + \frac{1}{\sqrt{n}} B(t) + \frac{1}{\sqrt{n}} G_n(t) + o\left(\frac{1}{\sqrt{n}}\right) \tag{2.23}$$

uniformly in t, where B has the same meaning as earlier and

$$G_n(t) = \frac{\varrho}{S}\Phi'(t)\left(\left[\left(t - \frac{a\sqrt{n}}{S}\right)\frac{S\sqrt{n}}{\varrho}\right] - \left(t - \frac{a\sqrt{n}}{S}\right)\frac{S\sqrt{n}}{\varrho} + \frac{1}{2}\right)$$

($[t]$ means the largest whole number $\leq t$).

We tie on some estimates to the right side of (2.10). From (2.23) and (2.18) it follows that

$$\int_{z<0} e^z dF_n\left(\frac{z}{S\sqrt{n}} + \lambda_n\right) = \frac{1}{\sqrt{n}}\left(\frac{1}{S}\Phi'(\lambda) + o(1) + \right. \tag{2.24}$$

$$\left. + \int_{z<0} e^z dG_n\left(\frac{z}{S\sqrt{n}} + \lambda_n\right)\right).$$

Further,

$$\int_{z<0} e^z dG_n\left(\frac{z}{S\sqrt{n}} + \lambda_n\right) = G_n(\lambda_n - 0) - \int_{z<0} G_n\left(\frac{z}{S\sqrt{n}} + \lambda_n\right) e^z dz. \tag{2.25}$$

By assumption the random variable h takes with positive probability a value no greater than

$$H + a + k\varrho \text{ for } k \text{ a whole number}$$

and the random variable h_n therefore the quantity

$$n(H + a) + k\varrho \text{ for } k \text{ a whole number}$$

By (2.5), $-\log \mu_n$ is one of these numbers, so that

$$\frac{1}{\varrho}(S\lambda_n\sqrt{n} - na) = \frac{1}{\varrho}(-\log \mu_n - n(H + a)) \tag{2.26}$$

is a whole number. We obtain

$$G_n\left(\frac{z}{S\sqrt{n}} + \lambda_n\right) = \frac{\varrho}{S}\Phi'\left(\frac{z}{S\sqrt{n}} + \lambda_n\right)\left(\left[\frac{z}{\varrho} + \frac{1}{\varrho}(S\lambda_n\sqrt{n} - an)\right] - \right. \tag{2.27}$$

$$\left. - \left(-\frac{z}{\varrho} + \frac{1}{\varrho}(S\lambda_n\sqrt{n} - an)\right) + \frac{1}{2}\right) = \frac{\varrho}{S}\Phi'\left(\frac{z}{S\sqrt{n}} + \lambda_n\right)\left(\left[\frac{z}{\varrho}\right] - \frac{z}{\varrho} + \frac{1}{2}\right),$$

specifically,

$$G_n(\lambda_n) = \frac{\varrho}{2S}\Phi'(\lambda_n) \tag{2.28}$$

and

$$G_n(\lambda_n - 0) = -\frac{\varrho}{2S}\Phi'(\lambda_n), \tag{2.29}$$

therefore by (2.25) and the theorem of Lebesgue,

$$\int_{z<0} e^z dG_n\left(\frac{z}{S\sqrt{n}} + \lambda_n\right) = -\frac{\varrho}{2S}\Phi'(\lambda) - \frac{\varrho}{S}\Phi'(\lambda)\int_{z<0} e^z \left(\left[\frac{z}{\varrho}\right] - \frac{z}{\varrho} + \frac{1}{2}\right) dz +$$
$$+ o(1) = \frac{1}{S}\Phi'(\lambda)\left(-\varrho - \int_{z<0} e^z \left[\frac{z}{\varrho}\right]\varrho dz + \int_{z<0} e^z z dz + o(1)\right) =$$
$$= \frac{1}{S}\Phi'(\lambda)\left(-\varrho - \varrho\sum_{-\infty}^{-1}\int_{k\varrho}^{(k+1)\varrho} e^z dz - 1 + o(1)\right) =$$
$$= \frac{1}{S}\Phi'(\lambda)\left(-\varrho + \frac{\varrho}{1-e^{-\varrho}} - 1 + o(1)\right) =$$
$$= \frac{1}{S}\Phi'(\lambda)\left(\frac{\varrho}{e^{\varrho}-1} - 1 + o(1)\right).$$

Together with (2.24) it follows from this that

$$\int_{z<0} e^z dF_n\left(\frac{z}{S\sqrt{n}} + \lambda_n\right) = \frac{1}{S\sqrt{n}}\Phi'(\lambda)\left(\frac{\varrho}{e^{\varrho}-1} + o(1)\right)$$

so, owing to (2.10),

$$q_n\left\{\frac{dp_n}{dq_n} > \mu_n\right\} = \frac{1}{S\sqrt{n}}\Phi'(\lambda)\left(\frac{\varrho}{e^{\varrho}-1} + o(1)\right)\exp\{nH + S\lambda_n\sqrt{n}\}. \tag{2.30}$$

Let now $0 < \Theta_n \leq 1$, chosen so that

$$\Theta_n F_n(\lambda_n) + (1-\Theta_n)F_n(\lambda_n - 0) = 1 - \epsilon$$

(by way of (2.14) this is possible). If one sets

$$M_n = \Phi + \frac{1}{\sqrt{n}}B, \tag{2.31}$$

then for sufficiently large n, the equalities

$$M_n(t) = 1 - \epsilon$$

clearly produce the particular solutions λ_n^0. Analogously to (2.22) one proves

$$\lambda_n^0 = \lambda + \frac{1}{S\sqrt{n}}Q(\lambda) + o\left(\frac{1}{\sqrt{n}}\right). \tag{2.32}$$

Further we have

$$0 = \Theta_n F_n(\lambda_n) + (1-\Theta_n)F_n(\lambda_n - 0) - M_n(\lambda_n^0) =$$

$$\Theta_n M_n(\lambda_n) + \Theta_n \frac{\varrho}{2S\sqrt{n}}\Phi'(\lambda_n) + o\left(\frac{1}{\sqrt{n}}\right) + (1-\Theta_n)M_n(\lambda_n) +$$

$$+(\Theta_n - 1)\frac{\varrho}{2S\sqrt{n}}\Phi'(\lambda_n) + o\left(\frac{1}{\sqrt{n}}\right) - M_n(\lambda_n^0)$$

(owing to (2.23),(2.31),(2.28), and (2.29))

$$= \Theta_n M_n(\lambda_n) + \Theta_n \frac{\varrho}{2S\sqrt{n}}\Phi'(\lambda) + (1-\Theta_n)M_n(\lambda) +$$

$$= (\Theta_n - 1)\frac{\varrho}{2S\sqrt{n}}\Phi'(\lambda) - M_n(\lambda_n) - (\lambda_n^0 - \lambda_n)M_n'(\lambda_n) + o(\lambda_n^0 - \lambda_n) + o\left(\frac{1}{\sqrt{n}}\right)$$

(owing to (2.18))

$$= (2\Theta_n - 1)\frac{\varrho}{2S\sqrt{n}}\Phi'(\lambda) - (\lambda_n^0 - \lambda_n)\Phi'(\lambda) + o(\lambda_n^0 - \lambda_n) + o\left(\frac{1}{\sqrt{n}}\right),$$

therefore

$$\lambda_n^0 - \lambda_n = (2\Theta_n - 1)\frac{\varrho}{2S\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right),$$

meaning

$$-\frac{\varrho}{2S\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right) < \lambda_n^0 - \lambda_n \le \frac{\varrho}{2S\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right)$$

or

$$-\frac{\varrho}{2} < (-na + S\lambda_n^0\sqrt{n} + o(1)) - (S\lambda_n\sqrt{n} - na) \le \frac{\varrho}{2}.$$

From this, it follows by (1.4) and (2.26) that

$$d(-naS\lambda_n^0\sqrt{n} + o(1)) = S(\lambda_n^0 - \lambda_n)\sqrt{n} + o(1),$$

therefore together with (2.32)

$$\lambda_n = \lambda + \frac{1}{S\sqrt{n}}Q(\lambda) - \frac{1}{S\sqrt{n}}d(-na + S\lambda\sqrt{n} + Q(\lambda) + o(1)) + o\left(\frac{1}{\sqrt{n}}\right). \tag{2.33}$$

We apply these to (2.30) and deduce that

$$q_n\left\{\frac{dp_n}{dq_n} > \mu_n\right\} = \frac{1}{S\sqrt{n}}\Phi'(\lambda)\exp\{nH + S\lambda\sqrt{n} + Q(\lambda) - \tag{2.34}$$

$$-d(-na + S\lambda\sqrt{n} + Q(\lambda) + o(1)) + o(1)\}\frac{\varrho}{e^{\varrho} - 1}.$$

To determine $\bar{\beta}(n,\epsilon)$ using equation (2.7) we must still calculate μ_n as well as

$$\left(1 - \epsilon - p_n\left\{\frac{dp_n}{dq_n} > \mu_n\right\}\right).$$

From (2.9) and (2.33),

$$\mu_n = \exp\{-nH - S\lambda_n\sqrt{n}\} = \exp\{-nH - S\lambda\sqrt{n}+ \tag{2.35}$$

$$+d(-na + S\lambda\sqrt{n} + Q(\lambda) + o(1)) + o(1)\}.$$

According to (2.13), it holds that

$$1 - \epsilon - p_n\left\{\frac{dp_n}{dq_n} > \mu_n\right\} = 1 - \epsilon - F_n(\lambda_n - 0) = \tag{2.36}$$

$$= \Phi(\lambda) - \Phi(\lambda_n) - \frac{1}{\sqrt{n}}B(\lambda_n) - \frac{1}{\sqrt{n}}G_n(\lambda_n - 0) + o\left(\frac{1}{\sqrt{n}}\right)$$

(according to (1.2) and (2.23))

$$= \Phi(\lambda) - \Phi(\lambda) - \Phi'(\lambda)\left(\frac{1}{S\sqrt{n}}Q(\lambda) - \frac{1}{S\sqrt{n}}d(-na + S\lambda\sqrt{n}+\right.$$

$$\left.+Q(\lambda) + o(1))\right) - \frac{1}{\sqrt{n}}B(\lambda) + \frac{1}{\sqrt{n}}\frac{\varrho}{2S}\Phi'(\lambda) + o\left(\frac{1}{\sqrt{n}}\right)$$

(according to (2.33) and (2.29))

$$= \frac{S\sqrt{n}'}{\Phi}(\lambda)\left(d(-na + S\lambda\sqrt{n}Q(\lambda) + o(1)) + \frac{\varrho}{2} + o(1)\right)$$

(because of (2.16)).

Putting things together: from (2.7), (2.34), (2.35), and (2.36) it follows that,

$$\bar{\beta}(n,\epsilon) = \frac{1}{S\sqrt{n}}\Phi'(\lambda)\exp\{nH + S\lambda\sqrt{n}Q(\lambda) - d(-na + S\lambda\sqrt{n} + Q(\lambda)+$$

$$+o(1))\}\left(\frac{\varrho}{e^{\varrho}-1}e^{o(1)} + \left(d(-na + S\lambda\sqrt{n} + Q(\lambda) + o(1)) + \frac{\varrho}{2} + o(1)\right)e^{o(1)}\right)$$

and since $0 \le d + \frac{1}{2}\varrho \le \varrho$,

$$= \frac{1}{S\sqrt{n}}\Phi'(\lambda)\exp\{nH + S\lambda\sqrt{n} + Q(\lambda) - d(-na + S\lambda\sqrt{n}+$$

$$+Q(\lambda) + o(1))\}e^{o(1)}d(-na + S\lambda\sqrt{n} + Q(\lambda) + o(1)) + \frac{\varrho}{2}\frac{e^{\varrho}+1}{e^{\varrho}-1}$$

and by (1.5)

$$= \frac{1}{S\sqrt{n}}\Phi'(\lambda)\exp\{nH + S\lambda\sqrt{n} + Q(\lambda) + w(-na + S\lambda\sqrt{n}+$$

$$+Q(\lambda) + o(1)) + o(1)\}.$$

In contrast to d, w is a continuous function so that we can extract a factor $o(1)$ from the argument of w. Owing to (2.1), we finally obtain

$$\bar{\beta}(n,\epsilon) = \frac{1}{S\sqrt{n}}\Phi'(\lambda)\exp\{nH + S\lambda\sqrt{n} + Q(\lambda) +$$

$$+ w(-na + S\lambda\sqrt{n} + Q(\lambda)) + o(1)\}$$

and thereby the assertions of the theorems hold for functions h which are lattice-like distributed.

3 Non-stationary Sources with Independent Letters

Let $n \geq 1$, X^k $(1 \leq l \leq n)$ a finite set, p^k a probability distribution on X^k, q^k a finite measure on X^k so that p^k is absolutely continuous with respect to q^k. Further, let

$$X_n = X^1 \times ... \times X^n$$

$$p_n = p^1 \times ... \times p^n$$

$$q_n = q^1 \times ... \times q^n.$$

We set

(3.1) $$H^k = \int h^k dp^k \ , \ h^k = -\log \frac{dp^k}{dq^k},$$

(3.2) $$H_n = \frac{1}{n}\sum_{k=1}^{n} H^k$$

(3.3) $$S_n = \left(\frac{1}{n}\sum_{k=1}^{n}\int |h^k - H^k|^2\right)^{1/2}$$

(3.4) $$R_n = \left(\frac{1}{n}\sum_{k=1}^{n}\int |h^k - H^k|^3\right)^{1/3}$$

and

(3.5) $$\beta(n,\epsilon) = \min_{p_n(E)\geq 1-\epsilon} q_n(E) = q_n(E_{n,\epsilon}) \text{ (approximately)}$$

with $p_n(E_{n,\epsilon}) \geq 1-\epsilon$.

Theorem 3.1 *It holds that*

$$|\log\beta(n,\epsilon)-nH_n-S_n\lambda\sqrt{n}+\frac{1}{2}\log n|<\frac{140}{\delta^8}$$

$$\left(S_n\geq\delta,\ R_n\leq\frac{1}{\delta},\ \delta\leq\epsilon\leq 1-\delta,\ \sqrt{n}\geq\frac{140}{\delta^8}\right).$$

Proof. Let $\delta>0$ be given, $S_n\geq\delta$, $R_n\leq\frac{1}{\delta}$ and $\delta\leq\epsilon\leq 1-\delta$. Without loss of generality, we take $\delta\leq\frac{1}{2}$. We again determine μ_n so that

$$p_n\left\{\frac{dp_n}{dq_n}\geq\mu_n\right\}\geq 1-\epsilon \tag{3.6}$$

$$p_n\left\{\frac{dp_n}{dq_n}>\mu_n\right\}<1-\epsilon$$

and set

$$g_n=\frac{1}{S_n\sqrt{n}}\left(-\log\frac{dp_n}{dq_n}-nH_n\right). \tag{3.7}$$

$$\lambda_n=\frac{1}{S_n\sqrt{n}}\left(-\log\mu_n-nH_n\right). \tag{3.8}$$

The distribution function of g_n relative to p_n we will call F_n. g_n is a normalized sum of n independent random variables with finite third moments such that the sum of the variances is non-vanishing. A well-known theorem of Berry and Esseen ([11] S.43) states, in our notation, that

$$|F_n(t)-\Phi(t)|\leq 7.5\frac{R_n^3}{S_n^3}\frac{1}{\sqrt{n}}\leq\frac{7.5}{\delta^6}\frac{1}{\sqrt{n}}\text{ for } t \text{ real.} \tag{3.9}$$

From (3.6) through (3.8) it follows that

$$F_n(\lambda_n)\geq 1-\epsilon \tag{3.10}$$

$$F_n(\lambda_n-0)<1-\epsilon.$$

We denote now the solution λ of the equation

$$\Phi(\lambda)=1-\eta$$

with $\lambda(\eta)$, $(0<\eta<1)$. As one can clearly see,

$$\Phi'(\lambda(\eta))\geq\eta\sqrt{2}\pi\text{ for } 0<\eta\leq\frac{1}{2}. \tag{3.11}$$

From (3.9) and (3.10) it follows that

$$|\Phi(\lambda(\epsilon))-\Phi(\lambda_n)|\leq\frac{7.5}{\delta^6}\frac{1}{\sqrt{n}},$$

and hence by the mean-value theorem and (3.11)

$$|\lambda_n - \lambda(\epsilon)| \le \frac{7.5}{\delta^6\sqrt{n}} \frac{\sqrt{2\pi}}{\delta} \le \frac{20}{\delta^7\sqrt{n}} \text{ as long as } \frac{7.5}{\delta^6\sqrt{n}} < \frac{\delta}{2}. \tag{3.12}$$

Analogous to (2.7), (2.11), and (2.10) one proves that

$$q_n\left\{\frac{dp_n}{dq_n} > \mu_n\right\} \le \beta(n,\epsilon) \le q_n\left\{\frac{dp_n}{dq_n} \ge \mu_n\right\} \tag{3.13}$$

$$q_n\left\{\frac{dp_n}{dq_n} \ge \mu_n\right\} = \exp\{nH_n + S_n\lambda_n\sqrt{n}\}\int_{z\le 0} e^z dF_n\left(\frac{z}{S_n\sqrt{n}} + \lambda_n\right) \tag{3.14}$$

$$q_n\left\{\frac{dp_n}{dq_n} \ge \mu_n\right\} = \exp\{nH_n + S_n\lambda_n'\sqrt{n}\}\int_{z\le 0} e^z dF_n\left(\frac{z}{S_n\sqrt{n}} + \lambda_n'\right), \tag{3.15}$$

where it may be that $\lambda_n' < \lambda_n$. Equation (3.9) gives in a similar way to (2.17) in the proof of Theorem 1.1

$$\int_{z\le 0} e^z dF_n\left(\frac{z}{S_n\sqrt{n}} + \lambda_n\right) \le \frac{1}{\sqrt{n}}\left(\frac{1}{S_n} + \frac{15}{\delta^6}\right) \le \frac{16}{\delta^6\sqrt{n}},$$

and, together with (3.13), (3.14), (3.12), and

$$S_n \le R_n \le \frac{1}{\delta} \tag{3.16}$$

therefore

$$\log\beta(n,\epsilon) \le nH_n + S_n\lambda(\epsilon)\sqrt{n} - \frac{1}{2}\log n + \frac{24}{\delta^8} \text{ when } \frac{7.5}{\delta^6\sqrt{n}} < \frac{\delta}{2}. \tag{3.17}$$

By (3.16) and (3.9),

$$\int_{z\le 0} e^z dF_n\left(\frac{z}{S_n\sqrt{n}} + \lambda_n'\right) \ge \int_{z\le 0} e^z dF_n\left(\frac{z\delta}{\sqrt{n}} + \lambda_n'\right) \tag{3.18}$$

$$= \frac{1}{\sqrt{n}}\left(\sqrt{n}\int_{z\le 0} e^z d\Phi\left(\frac{z\delta}{\sqrt{n}} + \lambda_n'\right) + \frac{7.5}{\delta^6}\int_{z\le 0} e^z dV_n\left(\frac{z\delta}{\sqrt{n}} + \lambda_n'\right)\right),$$

where the function V_n is of bounded variation with

$$|V_n(t)| \le 1 \text{ with } t \text{ real.} \tag{3.19}$$

We have

$$\sqrt{n}\int_{z\le 0} e^z d\Phi\left(\frac{z\delta}{\sqrt{n}} + \lambda_n'\right) = \delta\int_{z\le 0} e^z \Phi'\left(\frac{z\delta}{\sqrt{n}} + \lambda_n'\right) dz \ge \tag{3.20}$$

$$\geq \delta \int_{z\leq 0} e^z \Phi'\left(\frac{z\delta}{\sqrt{n}} - |\lambda'_n|\right) dz.$$

Now, let

$$\lambda_n - \frac{K}{\sqrt{n}} \leq \lambda'_n < \lambda_n, \tag{3.21}$$

$$K = \frac{91}{\delta^6 \Phi'(\lambda(\delta))}.$$

Through (3.12) and (3.11) it holds that

$$|\lambda'_n - \lambda(\epsilon)| \leq \frac{91}{\delta^6 \Phi'(\lambda(\delta))} + \frac{20}{\delta^7 \sqrt{n}} \leq \frac{139}{\delta^7 \sqrt{n}}, \text{ so long as } \frac{7.5}{\delta^6 \sqrt{n}} < \frac{\delta}{2}, \tag{3.22}$$

therefore

$$\delta \int_{z\leq 0} e^z \Phi'\left(\frac{z\delta}{\sqrt{n}} - |\lambda'_n|\right) dz \geq \delta \int_{z\leq 0} \Phi'\left(\frac{z\delta}{\sqrt{n}} + \frac{139}{\delta^7\sqrt{n}} + \lambda(\delta)\right)$$

$$> \delta \int_0^1 \Phi'\left(\frac{140}{\delta^7 \sqrt{n}} + \lambda(\delta)\right) e^{-z} dz, \text{ provided } \frac{7.5}{\delta^6\sqrt{n}} < \frac{\delta}{2}$$

$$> \delta \int_0^1 \Phi'(\delta + \lambda(\delta)) e^{-z} dz, \text{ provided } \sqrt{n} > \frac{140}{\delta^8}$$

$$> \delta \Phi'(\lambda(\delta)) \int_0^1 e^{-z} dz \geq \frac{\delta}{4} \Phi'(\lambda(\delta)),$$

as one easily sees. From this and from (3.20) it follows that

$$\sqrt{n} \int_{z\leq 0} e^z d\Phi\left(\frac{z\delta}{\sqrt{n}} + \lambda'_n\right) \geq \frac{\delta}{4} \Phi'(\lambda(\delta)), \text{ provided } \sqrt{n} > \frac{140}{\delta^8}. \tag{3.23}$$

We now estimate below the second summand of the right side of (3.18). First,

$$\int_{z\leq 0} e^z dV_n\left(\frac{z\delta}{\sqrt{n}} + \lambda'_n\right) = V_n(\lambda'_n) - \int_{z\leq 0} V_n\left(\frac{z\delta}{\sqrt{n}} + \lambda'_n\right) e^z dz.$$

Claim: For at least one λ'_n in the interval given in (3.21),

$$V_n(\lambda'_n) - \int_{z\leq 0} V_n\left(\frac{z\delta}{\sqrt{n}} + \lambda'_n\right) e^z dz > -\frac{\delta^7}{45} \Phi'(\lambda(\delta)) = -b, \text{ approximately.}$$

Otherwise,

$$V_n(\lambda'_n) - \int_{z\leq 0} V_n\left(\frac{z\delta}{\sqrt{n}} + \lambda'_n\right) e^z dz \leq -b, \text{ so long as } \lambda_n - \frac{K}{\sqrt{n}} \leq \lambda'_n < \lambda_n, \tag{3.24}$$

by introduction of U_n and u whereby

$$V_n(t) = U_n(\frac{\sqrt{n}}{\delta}(t - \lambda_n))$$

$$u = \frac{\sqrt{n}}{\delta}(\lambda_n' - \lambda_n)$$

so that

$$U_n(u) - \int_{z \leq 0} U_n(z+u)e^z dz \leq -b, \text{ for } \frac{K}{\delta} \leq u < 0, \tag{3.25}$$

or

$$U_n(u) - e^{-u}\int_{t \leq u} U_n(t)e^t dt \leq -b, \text{ for } \frac{K}{\delta} \leq u < 0. \tag{3.26}$$

From (3.19) it follows that

$$|U_n(t)| \leq 1 \text{ for } t \text{ real.}$$

We define for a fixed n a sequence of functions U_n^i for $i = 1, 2, ...$ through

$$U_n^1 = U_n$$

$$U_n^{i+1}(u) = \begin{cases} e^{-u}\int_{b \leq u} U_n^i(t)e^t dt - b & : -\frac{K}{\delta} \leq u < 0 \\ U_n & : \text{otherwise} \end{cases}$$

For $i > 1$ the U_n^i are continuous in the interval $(-L/\delta, 0)$ and satisfy (3.26) there. The sequence $U_n^i(t)$ is for each fixed n and t monotone non-decreasing and satisfies

$$|U_n^i(t)| \leq 1 \text{ for } t \text{ real.}$$

This gives that the sequence U_n^i converges with increasing i to a function $\bar{U}_n$ with

$$|\bar{U}_n| \leq 1 \text{ for } t \text{ real,}$$

which, in the interval $(-K/\delta, 0)$, satisfies the integral equation

$$\bar{U}_n - e^{-u}\int_{t \leq u} \bar{U}_n(t)e^t dt = -b.$$

The solutions have the form:

$$\bar{U}_n = -bt + c \text{ for } -\frac{K}{\delta} \leq t < 0.$$

From this and

$$|\bar{U}_n(t)| \leq 1 \text{ for } t \text{ real.}$$

it follows that

$$\frac{K}{\delta} \leq \frac{2}{b},$$

meaning from (3.21) and the definition of b,

$$\frac{91}{\delta^7 \Phi'(\lambda(\delta))} \leq \frac{90}{\delta^7 \Phi'(\lambda(\delta))},$$

which is impossible. Thereby assertion (3.24) is proved.

We can therefore for each n pick a λ_n' in the interval $(\lambda_n - L/\sqrt{n}, \lambda_n)$ so that

$$\int_{z\leq 0} e^z dV_n\left(\frac{z\delta}{\sqrt{n}} + \lambda_n'\right) > -\frac{\delta^7}{45}\Phi'(\lambda(\delta)). \tag{3.27}$$

Now by (3.18), (3.23), and (3.27),

$$\begin{aligned}\int_{z\leq 0} e^z dF_n\left(\frac{z}{S_n\sqrt{n}} + \lambda_n'\right) &\geq \frac{1}{\sqrt{n}}\left(\frac{\delta}{4}\Phi'(\lambda(\delta)) - \frac{\delta}{6}\Phi'(\lambda(\delta))\right), \text{ when } \sqrt{n} > \frac{140}{\delta^8} \\ &\geq \frac{1}{\sqrt{n}}\frac{\delta}{12}\Phi'(\lambda(\delta)) \\ &> \frac{1}{\sqrt{n}}\frac{\delta^2}{18}\end{aligned}$$

(owing to (3.11)).

From this and from (3.15) it follows that

$$q_n\left\{\frac{dp_n}{dq_n} > \mu_n\right\} \geq \exp\{nH_n + S_n\lambda_n'\sqrt{n} - \frac{1}{2}\log n\}\frac{\delta^2}{18}, \text{ when } \sqrt{n} > \frac{140}{\delta^8} \tag{3.28}$$

therefore owing to (3.22) and (3.16)

$$\log q_n\left\{\frac{dp_n}{dq_n} > \mu_n\right\} \geq nH_n + S_n\lambda(\epsilon)\sqrt{n} - \frac{1}{2}\log n - \frac{140}{\delta^8}, \text{ when } \sqrt{n} > \frac{140}{\delta^8} \tag{3.29}$$

and by (3.13) finally

$$\log\beta(n,\epsilon) \geq nH_n + S_n\lambda\sqrt{n} - \frac{1}{2}\log n - \frac{140}{\delta^8}, \text{ when } \sqrt{n} > \frac{140}{\delta^8}. \tag{3.30}$$

From this and from (3.17) the theorem follows.

Lemma 3.2 *Let A be the set of all probability distributions in Y. I_α (see (1.10)) is a concave function on A. For α_0, $\alpha_1 \in A$ and $0 < t < 1$, the relations*

$$I_{(1-t)\alpha_0 + t\alpha_1} = (1-t)I_{\alpha_0} + tI_{\alpha_1}$$

and

$$P\alpha_0 = P\alpha_1$$

are equivalent. It holds that

$$\sum_{x:P(x,y)>0} P(x,y)\log\frac{P(x,y)}{(P\alpha)(x)} \le C \text{ for } y \in Y, \alpha \in \bar{A} \tag{3.31}$$

and

$$\sum_{x:P(x,y)>0} P(x,y)\log\frac{P(x,y)}{(P\alpha)(x)} = C \text{ for } \alpha(y) > 0, \alpha \in \bar{A}. \tag{3.32}$$

Conversely, $\alpha \in \bar{A}$ *is a consequence of*

$$\sum_{x:P(x,y)>0} P(x,y)\log\frac{P(x,y)}{(P\alpha)(x)} \ge \tag{3.33}$$

$$\ge \sum_{x:P(x,z)>0} P(x,z)\log\frac{P(x,z)}{(P\alpha)(x)} \ge \text{ for } y,z \in Y, \alpha(y) > 0.$$

Let $\bar{\alpha} \in \bar{A}$ *be arbitrary. It holds that*

$$\bar{A} = \{\alpha | \alpha \in A, P\alpha = P\bar{\alpha}; \alpha(y) = 0 \text{ for all } y \text{ with} \tag{3.34}$$

$$\sum_{x:P(x,y)>0} P(x,y)\log\frac{P(x,y)}{(P\bar{\alpha})(x)} < C\}.$$

We omit the elementary proofs of these well-known Lemmas. (cf [12]).

4 Proof of Theorem 1.2

Since I_α is a continuous function of α, $\bar{A}$ (see (1.13)) is a compact, non-empty subset of A.

$$G_\alpha = \left(\sum_{x,y:P(x,y)\alpha(y)>0} \left(\log\frac{P(x,y)}{(P\alpha)(x)} - I_\alpha \right)^2 P(x,y)\alpha(y) \right)^{1/2}, \tag{4.1}$$

is likewise a continuous function of α so that definitions (1.14) are meaningful. We pick an

$$\bar{\alpha} \in \bar{A} \tag{4.2}$$

with

$$G_{\bar{\alpha}} = T_{\text{sign}\lambda} \tag{4.3}$$

and denote by $\bar{\alpha}_n$ the independent product of n copies of $\bar{\alpha}$. For every $y_n \in Y_n$ we assign the probability distributions for $1 \le k \le n$,

$$p^k = P(\cdot, y^k) \tag{4.4}$$

(4.5) $$q^k = P\bar{\alpha}$$

and write E_{y_n} for $E_{n,\epsilon-2/\sqrt{n}}$ for $\beta(n,\epsilon-2/\sqrt{n})$ (see (3.5)). We therefore have

(4.6) $$P_n(E_{y_n}, y_n) \geq 1-\epsilon+\frac{2}{\sqrt{n}}$$

and

(4.7) $$(P\bar{\alpha})_n(E_{y_n}) = \min_{P_n(E_{y_n})\geq 1-\epsilon+w/\sqrt{n}} (P\bar{\alpha})_n(E) = \beta_{y_n}.$$

We substitute appropriately H_n, S_n, and R_n (see (3.2) through (3.4)) to produce H_{y_n}, S_{y_n}, and R_{y_n}, so it follows by Lemma 3.2

(4.8) $$H_{y_n} = -C,\ \bar{\alpha}_n\text{–almost everywhere.}$$

The function

(4.9) $$S^2(y) = \sum_{x:P(x,y)>0} \left(\log\frac{P(x,y)}{(P\bar{\alpha})(x)} - C\right)^2 P(x,y)$$

has by (4.1) and (4.3) relative to $\bar{\alpha}$ the expected value $T^2_{\mathrm{sign}\lambda}$ and finite variance. Further,

(4.10) $$S^2_{y_n} = \frac{1}{n}\sum_{k=1}^{n} S^2(y^k),\ \bar{\alpha}_n\text{–almost everywhere,}$$

so that for sufficiently large K by the Chebyshev inequality,

(4.11) $$\bar{\alpha}_n(D_n) > \frac{1}{2},\ (n\geq 1)$$

with

(4.12) $$D_n = \left\{ y_n \middle| |S^2_{y_n} - T^2_{\mathrm{sign}\lambda}| < \frac{K}{\sqrt{n}}, \bar{\alpha}_n(y_n) > 0 \right\}$$

An ϵ-code f for P_n is called admissible if

(4.13) $$f(X_n) \subseteq D_n$$

and

(4.14) $$f^{-1}(y_n) \subseteq E_{y_n},\ \text{for } y_n \in f(X_n).$$

Let the ϵ-code f be admissible with maximal length N under all admissible ϵ-codes for P_n. Claim:

(4.15) $$(P\bar{\alpha})_n(f^{-1}Y_n) \geq \frac{1}{\sqrt{n}} \text{ for } n \text{ sufficiently large.}$$

Otherwise, we have in particular that

$$\frac{1}{\sqrt{n}} > (P\bar{\alpha})_n(f^{-1}Y_n) = (P\bar{\alpha}_n)(f^{-1}Y_n) \geq$$

$$\geq \int_{D_n} P_n(f^{-1}Y_n, y_n)\bar{\alpha}_n(dy_n) \geq \frac{1}{2}\min_{y_n\in D_n} P_n(f^{-1}Y_n, y_n).$$

This therefore gives a $y_n^0 \in D_n$ with

$$P_n(f^{-1}Y_n, y_n^0) < \frac{2}{\sqrt{n}}. \tag{4.16}$$

From this and from (4.6), it follows that

$$P_n(E_{y_n^0} - f^{-1}Y_n.y_n^0) \geq 1 - \epsilon \tag{4.17}$$

Let

$$f_0(x_n) = \begin{cases} f(x_n) & x_n \in f^{-1}Y_n \\ y_n^0 & x_n \in E_{y_n^0} - f^{-1}Y_n \end{cases}$$

By reason of (4.17) $y_n^0 \in D_n$, f_0 is an admissible ϵ-code for P_n. By way of (4.16) and $\epsilon < 1$, f_0 has for sufficiently large n the length $N + 1$. This violates the maximality of N, and so (4.15) must be true.

We obtain for sufficiently large n,

$$\begin{aligned} \frac{1}{\sqrt{n}} &\leq (P\bar{\alpha})_n(f^{-1}Y_n) = \sum_{y_n\in f(X_n)} (P\bar{\alpha})_n(f^{-1}(y_n)) \leq \\ &\leq \sum_{y_n\in f(X_n)} (P\bar{\alpha})_n(E_{y_n}) \text{ by (4.14)} \\ &= \sum_{y_n\in f(X_n)} \beta_{y_n} \text{ by (4.7)} \\ &\leq N \max_{y_n\in D_n} \beta_{y_n} \text{ by (4.13).} \end{aligned} \tag{4.18}$$

When

$$T_{\text{sign}\lambda} > 0, \tag{4.19}$$

we can by (4.12) for sufficiently large n apply Theorem 3.1 and obtain from (4.8) and (4.12)

$$\max_{y_n\in D_n} \beta_{y_n} \leq \exp\{-nC + \lambda T_{\text{sign}\lambda}\sqrt{n} - \frac{1}{2}\log n + O(1)\}. \tag{4.20}$$

From (4.18) and (4.20) it gives

$$\log N(n,\epsilon) \geq \log N \geq nC - \lambda T_{\text{sign}\lambda}\sqrt{n} + O(1). \tag{4.21}$$

When

$$T_{\text{sign}\lambda} = 0, \tag{4.22}$$

it follows from (4.1) and (4.2) for y with $\bar{\alpha}(y) > 0$

$$\frac{P(x, y}{(P\bar{\alpha})(x)} = e^C, \text{ for } P(\cdot, y)\text{-almost every } x,$$

therefore for $y_n \in D_n$, for example,

$$\frac{P_n(x_n, y_n)}{P\bar{\alpha})_n(x_n)} = e^{nC}, \text{ for } P_n(\cdot, y)\text{-almost every } x_n.$$

Hence this gives

$$\beta_{y_n} \leq e^{-nC} \text{ for } y_n \in D_n$$

therefore

$$\max_{y_n \in D_n} \beta_{y_n} \leq e^{-nC}$$

which, together with (4.18), yields

$$\log N(n, \epsilon) \geq \log N \geq nC - \frac{1}{2} \log n \text{ for } n \text{ sufficiently large.} \tag{4.23}$$

In order to prove (1.15), we first take $\epsilon \leq \frac{1}{2}$ so that $\lambda > 0$. For each $y_n \in Y_n$, we associate a probability distribution $\alpha_{y_n} \in A$:

$$\alpha_{y_n} = \frac{1}{n} \sum_{k=1}^{n} \delta_{y^k}, \tag{4.24}$$

where δ_y is the unit mass at the point y (Dirac delta function). Further we set

$$A_n = \{\alpha_{y_n} | y_n \in Y_n\}. \tag{4.25}$$

Let f be an ϵ-code of maximal length for P_n. We identify f with its graph, a subset of $X_n \times Y_n$, and it is useful to decompose f into subsets f_i. Let

$$f_\alpha = \{(x_n, y_n) | (x_n, y_n) \in f, \alpha_{y_n} = \alpha\} \text{ for } \alpha \in A_n. \tag{4.26}$$

All f_α are ϵ-codes with length N_α. As f has maximal length,

$$N(n, \epsilon) = \sum_{\alpha \in A_n} N_\alpha. \tag{4.27}$$

Let $\alpha \in A_n$. We select a $y_n \in Y_n$ with $\alpha_{y_n} = \alpha$ and set for $1 \leq k \leq n$

$$\begin{aligned} p^k &= P(\cdot, y^k) \\ q^k &= P_\alpha, \end{aligned} \tag{4.28}$$

so that we can can construct $\beta(n,\epsilon)$ pursuant to (3.5). It is easy to see that $\beta(n,\epsilon)$ depends only on α (and not on the particular $y_n \in f_\alpha$ chosen). We denote $\beta(n,\epsilon)$ more accurately as $\beta_\alpha(n,\epsilon)$.

In the case that $y_n \in f_\alpha(X_n)$ and owing to

$$p_n(f_\alpha^{-1}(y_n)) = P_n(f_\alpha^{-1}(y_n), y_n) \geq 1-\epsilon,$$

we have by (4.28)

$$(P\alpha)_n(f_\alpha^{-1}(y_n)) = q_n(f_\alpha^{-1}(y_n)) \geq \min_{p_n(E)\geq 1-\epsilon} q_n(E) = \beta_\alpha(n,\epsilon),$$

and therefore also

$$1 \geq (P\alpha)_n(f_\alpha^{-1}Y_n) = \sum_{y_n\in f_\alpha(X_n)} (P_\alpha)_n(f^{-1}(y)) \geq N_\alpha \beta_\alpha(n,\epsilon).$$

We denote the cardinality of a set M with $|M|$, so out of this and (4.27) it follows that

$$(n+1)^{|Y|-1} \geq |A_n| \geq \sum_{\alpha\in A_n} N_\alpha \beta_\alpha(n,\epsilon) \geq N(n,\epsilon) \min_{\alpha\in A_n} \beta_\alpha(n,\epsilon). \tag{4.29}$$

Now for the moment let

$$T_1 > 0. \tag{4.30}$$

We can associate every $y_n \in Y_n$ by means of (4.28) with the triple H_n, S_n, R_n. In the case that $\alpha_n = \alpha_{y_n}$ and owing to

$$\begin{aligned} H_n = -\frac{1}{n} \sum_{1\leq k\leq n} \sum_{x:P(x,y^k)>0} P(x,y^k) \log \frac{P(x,y^k)}{(P_\alpha)(x)} = \\ = -\sum_{x,y:P(x,y)\alpha(y)>0} P(x,y)\alpha(y) \log \frac{P(x,y)}{(P\alpha)(x)} = -I_\alpha, \end{aligned} \tag{4.31}$$

$$\begin{aligned} S_n^2 = \frac{1}{n} \sum_k \sum_{x:P(x,y^k)>0} \Bigg(- \log \frac{P(x,y^k)}{(P\alpha)(x)} + \\ + \sum_{u:p(u,y^k)>0} P(u,y^k) \log \frac{P(u,y^k)}{(P\alpha)(x)} \Bigg)^2 P(x,y^k) = \sum_{x,y:P(x,y)\alpha(y)>0} P(x,y)\alpha(y) \times \\ \times \left(\log \frac{P(x,y)}{(P\alpha)(x)} - \sum_{u:P(u,y)>0} P(u,y) \log \frac{P(u,y)}{(P\alpha)(x)} \right)^2 = M_\alpha^2, \text{ say} \end{aligned} \tag{4.32}$$

$$(4.33)\qquad R_n^3 = \sum_{x,y:P(x,y)\alpha(y)>0} P(x,y)\alpha(y)\left|\log\frac{P(x,y)}{(P\alpha)(x)} - \right.$$

$$\left. - \sum_{u:P(u,y)>0} P(u,y)\log\frac{P(u,y)}{(P\alpha)(x)}\right|^3 = L_\alpha^3, \text{ say.}$$

the functions H_n, S_n, R_n are also functions of α (rather than of y_n). In addition to that, I_α, M_α, and L_α form continuations of H_n, S_n, R_n throughout A. From $M_\alpha > 0$ it follows naturally that $L_\alpha > 0$. The set of the $\alpha \in A$ for which $\varphi(\alpha)$ lies in the strict positive quadrant is a neighborhood of $\bar{A}$, because for $\alpha \in \bar{A}$ it holds by (4.32), (3.32), (4.1), (1.14), and (4.30)

$$(4.34)\qquad M_\alpha = G_\alpha \geq T_1 > 0.$$

As one can easily see, every $P(\cdot, y)$ is, for each $\alpha \in \bar{A}$, absolutely continuous with respect to $P\alpha$. This holds then also for the α from a suitable neighborhood of $\bar{A}$. In this neighborhood, therefore, I_α and M_α^2 are arbitrarily differentiable (smooth) functions of α. Further, $\bar{A}$ is convex (see (3.34)). Let $\eta > 0$ and let U be the closed η-neighborhood of $\bar{A}$ (in A) in such a way that $\varphi(U)$ lies completely in the strict positive quadrant and so that I_α and M_α^2 are infinitely differentiable in U. By the previous remarks, there is such an η, and since $M_\alpha^2 > 0$ for $\alpha \in U$, $M_\alpha > 0$ is also infinitely differentiable in U. U is (convex and) compact. $\varphi(U)$ is therefore also compact, and there is a $\delta > 0$ with

$$M_\alpha \geq \delta, \quad L_\alpha \leq \tfrac{1}{\delta}, \quad \delta \leq \epsilon \leq 1-\delta \text{ for } \alpha \in U.$$

By Theorem 3.1 and (4.31) through (4.33), it follows that

$$(4.35)\qquad \beta_\alpha(n,\epsilon) = \exp\{-nI_\alpha + M_\alpha\lambda\sqrt{n} - \frac{1}{2}\log n + O(1)\}$$

uniformly for $\alpha \in U \cap A_n$. Because A is compact and the open kernel $\mathring{U}$ of U contains the set $\bar{A}$, there is a $C' < C$ with

$$I_\alpha \leq C' \text{ for } \alpha \in A - \mathring{U}.$$

Let $C'' = \frac{1}{2}(C + C')$. By way of (4.31), (4.32), and because M_α is bounded, we have uniformly for y_n with $\alpha = \alpha_{y_n} \in A_n - \mathring{U}$

$$(4.36)\qquad p_n\left\{\log\frac{dp_n}{dq_n} \leq nC''\right\} \geq$$

$$\geq 1 - p_n\left\{\left(\log\frac{dp_n}{dq_n} - nI_\alpha\right)^2 > n^2(C''-C')\right\} >$$

$$> 1 - \frac{M_\alpha^2}{n(C''-C')^2} \geq \frac{1+\epsilon}{2} \text{ for } n \text{ sufficiently large.}$$

As $E_{n,\epsilon}$ can be defined with help of (4.28) as in (3.5) ($E_{n,\epsilon}$ clearly depends on y_n), it follows from (4.36), uniformly for y_n with $\alpha = \alpha_{y_n} \in A_n - \mathring{U}$

$$\beta_\alpha(n,\epsilon) = q_n(E_{n,\epsilon}) \geq q_n\left(E_{n,\epsilon} \cap \left\{\frac{dp_n}{dq_n} \leq e^{nC''}\right\}\right) \geq$$

$$\geq e^{-nC''} p_n\left(E_{n,\epsilon} \cap \left\{\log \frac{dp_n}{dq_n} \leq nC''\right\}\right) \geq \frac{1-\epsilon}{2} e^{-nC''} \text{ for } n \text{ sufficiently large,}$$

so that

$$\beta_\alpha(N,\epsilon) \geq \exp(-nC'' + O(1))$$

uniformly for $\alpha \in A_n - \mathring{U}$. From this and from (4.35) it follows that

$$\text{(4.37)} \qquad \min_{\alpha \in A_n} \beta_\alpha(n,\epsilon) \geq \min\{\min_{\alpha \in A_n \cap \mathring{U}} \beta_\alpha(n,\epsilon), \min_{\alpha \in A_n - \mathring{U}} \beta_\alpha(n,\epsilon)\}$$

$$\geq \min\{\inf_{\alpha \in U} \exp\{-nI_\alpha + M_\alpha \lambda \sqrt{n} - \frac{1}{2}\log n + O(1)\}, \exp\{-nC'' + O(1)\}\} =$$

$$= \inf_{\alpha \in U} \exp\{-nI_\alpha + M_\alpha \lambda \sqrt{n} - \frac{1}{2}\log n + O(1)\} =$$

$$= \exp\{\inf_{\alpha \in U}(-nI_\alpha + M_\alpha \lambda \sqrt{n} - \frac{1}{2}\log n) + O(1)\}.$$

We now assign each $\alpha \in U$ to that $\bar{\alpha} \in \bar{A}$ for which

$$\inf_{\bar{\alpha} \in \bar{A}} \|\alpha - \bar{\alpha}\|$$

is attained ($\|\cdot\|$ = Euclidean distance).

Let I'_α and M'_α be the first derivatives of I_α and M_α, I''_α the second derivative of I_α with respect to α (I''_α is therefore a bilinear form over the vector space that lies in the hyperplane spanned by A, see for example [13]). After the definition of $\bar{A}$ and since $\bar{\alpha} \in \bar{A}$, we have

$$\text{(4.38)} \qquad I'_{\bar{\alpha}}(\alpha - \bar{\alpha}) \leq 0.$$

Because I_α and M_α have continuous derivatives of all orders in the compact set U, there is, by the Taylor formula ([13], page 94), a $b > 0$ so that

$$\text{(4.39)} \qquad -nI_\alpha + \lambda M_\alpha \sqrt{n} \geq -n(I_{\bar{\alpha}} + I'_{\bar{\alpha}}(\alpha - \bar{\alpha}) + \frac{1}{2} I''_{\bar{\alpha}}(\alpha - \bar{\alpha})(\alpha - \bar{\alpha}) +$$

$$+ b\|\alpha - \bar{\alpha}\|^3 + \lambda(M_{\bar{\alpha}} - b\|\alpha - \bar{\alpha}\|)\sqrt{n} \text{ for } \alpha \in U.$$

Now,

$$\text{(4.40)} \qquad \begin{aligned} I_{\bar{\alpha}} &= C \\ M_{\bar{\alpha}} &= G_{\bar{\alpha}} \geq T_1 \text{ for } \alpha \in U \end{aligned}$$

(see (1.14)). One can find an $a > 0$ with

(4.41) $$I'_{\bar\alpha}(\alpha-\bar\alpha)+\frac{1}{2}I''_{\bar\alpha}(\alpha-\bar\alpha)(\alpha-\bar\alpha)\le -a\|\alpha-\bar\alpha\|^2 \text{ for } \alpha\in U$$

because, assuming this is false, one finds a sequence $\alpha_j\in U$ $(j=1,2,...)$ with

$$\liminf_j(I'_{\bar\alpha_j}(\alpha_j-\bar\alpha_j)+\frac{1}{2}I''_{\bar\alpha_j}(\alpha_j-\bar\alpha_j)(\alpha_j-\bar\alpha_j))\ge 0$$

and by (4.38)

$$\|\alpha_j-\bar\alpha_j\|=\eta \text{ for } j=1,2,....$$

Let α be a limit point of this sequence. Then it holds that

(4.42) $$\|\alpha-\bar\alpha\|=\eta>0 \text{ for } \alpha\in U$$

(4.43) $$I'_{\bar\alpha}(\alpha-\bar\alpha)+\frac{1}{2}I''_{\bar\alpha}(\alpha-\bar\alpha)(\alpha-\bar\alpha)\ge 0.$$

Let $\alpha^t=t\alpha+(1-t)\bar\alpha$ and

$$I^t=\sum_{x,y:P(x,y)\alpha^t(y)>0}P(x,y)\alpha^t(y)\log\frac{P(x,y)}{(P\alpha^t)(x)}.$$

Then

$$I''_{\bar\alpha}(\alpha-\bar\alpha)(\alpha-\bar\alpha)=\left(\frac{d^2}{dt^2}I^t\right)_{t=0}=-\sum_{(P\bar\alpha)(x)>0}\frac{((P(\alpha-\bar\alpha)(x))^2}{(P\bar\alpha)(x)}.$$

From this and from (4.38) and (4.43) it follows that

$$I'_{\bar\alpha}(\alpha-\bar\alpha)=0$$

and

$$(P\alpha)(x)=(P\bar\alpha)(x) \text{ for } (P\alpha)(x)>0,$$

therefore

$$P\alpha=P\bar\alpha.$$

By Lemma 3.2, we therefore have

$$\alpha\in\bar A$$

in contradiction to (4.42).

Thereby, (4.41) is proven, and from (4.39) to (4.41) it follows that

$$-nI_\alpha+\lambda M_\alpha\sqrt{n}\ge -nC+n\|\alpha-\bar\alpha\|^2(a-b\|\alpha-\bar\alpha\|)+$$

$$+\lambda T_1\sqrt{n}-\lambda\|\alpha-\bar\alpha\|b\sqrt{n}.$$

By shrinking U (i.e. η) if necessary, we can assume that $\|\alpha - \bar{\alpha}\| \leq a/2b$, so that

$$(4.44) \qquad -nI_\alpha + \lambda M_\alpha \sqrt{n} \geq -nC + \lambda T_1 \sqrt{n} + n\|\alpha - \bar{\alpha}\|^2 \frac{a}{2} -$$

$$-b\lambda\sqrt{n}\|\alpha - \bar{\alpha}\| \geq -nC + \lambda T_1 \sqrt{n} - \frac{\lambda^2 b^2}{2a}$$

follows.

From (4.29), (4.37), and (4.44) one ultimately gets

$$(n+1)^{|Y|-1} \geq N(n,\epsilon) \exp\{-nC + \lambda T_1 \sqrt{n} - \frac{1}{2}\log n + O(1)\},$$

and from there

$$N(n,\epsilon) < \exp\{nC - \lambda T_1 \sqrt{n} + (|Y| - \frac{1}{2}) \log n + O(1)\}.$$

Now let

$$(4.45) \qquad T_1 = 0.$$

We select an $\bar{\alpha} \in \bar{A}$ and set

$$(4.46) \qquad Z = \left\{ y | y \in Y, \log \frac{dP(\cdot,y)}{d(P\bar{\alpha})} \text{ has positive variance with respect to } P(\cdot,y) \right\}.$$

For $y \in Y - Z$, $dP(\cdot,y)/d(P\alpha)$ is therefore $P(\cdot,y)$-almost everywhere constant and indeed it holds by the fact that $\bar{\alpha} \in \bar{A}$ and Lemma 3.2 that

$$(4.47) \qquad 0 \leq \frac{P(x,y)}{(P\bar{\alpha})(x)} \leq e^C \text{ for } P(\cdot,y)\text{-almost every } x; y \in Y - Z.$$

In the case that $y_m \in Y_m$ with

$$(4.48) \qquad y^k \in Z \text{ for } 1 \leq k \leq m,$$

we form

$$p^k = P(\cdot, y^k)$$
$$q^k = P\bar{\alpha} \text{ for } 1 \leq k \leq m.$$

By (4.46) one can find a $\delta > 0$ with $S_m \geq \delta$, $R_m \leq 1/\delta$, and $\delta \leq \epsilon \leq 1 - \delta$ (independent of y_m with (4.48)). From the proof of Theorem 3.1 ((3.6) and (3.29)) it follows that

$$(4.49) \qquad p_m \left\{ \frac{dp_m}{dq_m} > \mu_m \right\} < 1 - \epsilon,$$

and

$$\text{(4.50)}\quad \log p_m\left\{\frac{dp_m}{dq_m}>\mu_m\right\}\geq mH_m+\lambda S_m\sqrt{m}-\frac{1}{2}\log m-\frac{140}{\delta^8}\ \text{if}\ m>\frac{140}{\delta^8}.$$

But now, $H_m \geq -C$ (see (3.31)), so that by above,

$$\text{(4.51)}\qquad \log q_m^m\left\{\frac{dp_m}{dq_m}>\mu_m\right\}\geq -mC-\frac{1}{2}\log m-\frac{140}{\delta^8}\ \text{if}\ m>\frac{140}{\delta^8},$$

independent of y_m with (4.48).

Now let $y_n \in Y_n$ be arbitrary. We set as above,

$$\begin{aligned} p^k &= P(\cdot, y^k)\\ q^k &= P\bar{\alpha}\ \text{for}\ 1\leq k\leq n \end{aligned}$$

and denote the $\beta(n,\epsilon)$ defined in (3.5) with $\beta_\alpha(n,\epsilon)$, where $\alpha=\alpha_{y_n}\in A_n$ (this $\beta_\alpha(n,\epsilon)$ is different from those occurring in (4.29), but the notation β_α instead of β_{y_n} is again justified).

In lieu of (4.29), we now obtain

$$\text{(4.52)}\qquad \begin{aligned} 1\geq (P\bar{\alpha})_n(f^{-1}Y_n)&=\sum_{\alpha\in A_n}(P\bar{\alpha})_n)(f_\alpha^{-1}Y_n)\geq\sum_{\alpha\in A_n}N_\alpha\beta_\alpha(n,\epsilon)\geq\\ &\geq N(n,\epsilon)\min_{\alpha\in A_n}\beta_\alpha(n,\epsilon). \end{aligned}$$

Let $\alpha\in A_n$ be arbitrary. Then either

$$\text{(4.53)}\qquad n\alpha(Z)\leq\frac{140}{\delta^8}$$

or

$$\text{(4.54)}\qquad n\alpha(Z)>\frac{140}{\delta^8}$$

In the first case we set

$$\text{(4.55)}\qquad M=\max_{x,y:P(x,y)>0}\frac{P(x,y)}{(P\bar{\alpha})(x)}.$$

Since $\bar{\alpha}\in\bar{A}$, we have that $M<\infty$ (see (3.31)). From (4.53), (4.55), and (4.47), it follows that

$$\text{(4.56)}\qquad \begin{aligned} \beta_\alpha(n,\epsilon)&\geq\frac{1-\epsilon}{\exp\{(n-n\alpha(Z))C\}M^{n\alpha(Z)}}\geq\\ &\geq\frac{1-\epsilon}{\exp\left\{nC+\frac{140}{\delta^8}\log M\right\}}=\exp\{-nC+O(1)\} \end{aligned}$$

uniformly for α with (4.53).

In the second case, (4.54), we pick a y_n with $\alpha_{y_n} = \alpha$ so that the first $m = n\alpha(Z)$ of the y^k lie in Z with the others in $Y - Z$. Then (4.49) implies that

$$(4.57)\quad p_n\left\{x_n \middle| \frac{dp_m}{dq_m}(x_n) > \mu_m, \frac{d(p^{m+1}\times ... \times p^n)}{d(q^{m+1}\times ... \times q^n)}(x^{m+1}, ..., x^n) > 0\right\} < 1 - \epsilon$$

and by the definition of Z, one finds a γ_n with

$$\left\{x_n \middle| \frac{dp_m}{dq_m}(x_n) > \mu_m, \frac{d(p^{m+1}\times ... \times p^n)}{d(q^{m+1}\times ... \times q^n)}(x^{m+1}, ..., x^n) > 0\right\} =$$
$$= \left\{x_n \middle| \frac{dp_n}{dq_n}(x_n) > \gamma_n\right\}.$$

From this, (4.57), and the definition of $\beta_\alpha(n, \epsilon)$ it holds that

(4.58)

$$\log \beta_\alpha(n, \epsilon) \geq \log q_n \left\{x_n \middle| \frac{dp_m}{dq_m}(x_n) > \mu_m, \frac{d(p^{m+1}\times ... \times p^n)}{d(q^{m+1}\times ... \times q^n)}(x^{m+1}, ..., x^n) > 0\right\} =$$
$$= \log q_m \left\{x_n \middle| \frac{dp_m}{dq_m}(x_n) > \mu_m\right\} +$$
$$+ \log(q^{m+1}\times ... \times q^n)\left\{(x^{m+1}, ..., x^n) \middle| \frac{d(p^{m+1}\times ... \times p^n)}{d(q^{m+1}\times ... \times q^n)} > 0\right\} \geq$$

(by (4.51) and (4.47))

$$\geq -mC - \frac{1}{2}\log m - \frac{140}{\delta^8} - (n-m)C \geq$$
$$\geq -nC - \frac{1}{2}\log n - \frac{140}{\delta^8}$$

uniformly in α with (4.54). Together, (4.56), and (4.58) provide

$$\beta_\alpha(n, \epsilon) > \exp\{-nC - \frac{1}{2}\log n + O(1)\}$$

uniformly in $\alpha \in A_n$ so that from (4.52)

$$N(n, \epsilon) < \exp\{nC - \lambda T_1\sqrt{n} + (|Y| - \frac{1}{2})\log n + O(1)\}$$

holds. This inequality is thereby generally (i.e. for $T_1 = 0$ and $T_1 > 0$) proven.

In a similar way, one obtains for $\epsilon > \frac{1}{2}$

$$N(n, \epsilon) < \exp\{nC - \lambda T_{-1}\sqrt{n} + (|Y| - \frac{1}{2})\log n + O(1)\}$$

whereby the theorem holds.

5 Notes and Comments

(i) In the proof of Theorem 3.1, the assumption that the underlying spaces X^k were finite was not used. The theorem is also correct for measurable spaces X^k. The same goes for the generalized version of Theorem 1.1 if one requires the occurring moments of h to be finite and one avoids the summation on the right sides of the definitions of the integrals. In the case that h has a lattice-like distribution or the distribution satisfies the requirement C [10] (this is possible only for infinite X), one can probably further sharpen Theorem 1.1 by gathering the well-known asymptotic developments of Esseen [11] and Cramér [10].

Theorem 1.1 and Theorem 3.1 remain essentially true (see [14]) if one replaces the probability distribution p, respectively p^k, and the finite measures q, respectively q^k, with totally monotone capacities in the sense of Choquet [15]. This has applications in the Theory of Continuous Information Sources with Consideration of Observation Errors, whose statistical nature one does not know exactly (see [14]).

(ii) Example of a Channel with $T_1 \neq T_{-1}$: Let $X = \{0,1,2\}$, $Y = \{0,1,2,3,4,5\}$,

$$P(x,y) = \frac{1}{2}(1-\delta_{xy}) \text{ for } x,y = 0,1,2 \tag{5.1}$$

and

$$P(x,y) = a(x+y \mod 3) \text{ for } x \in X, y = 3,4,5, \tag{5.2}$$

where $(x+y \mod 3) \in \{0,1,2\}$ and a is a probability distribution in X. Let further,

$$\alpha = \begin{cases} \frac{1}{3} & : y = 0,1,2 \\ 0 & : \text{otherwise.} \end{cases}$$

It is then $(P\alpha)(x) = \frac{1}{3}$ for $x \in X$ and

$$\sum_{x:P(x,y)>0} P(x,y) \log \frac{P(x,y)}{(P\alpha)(x)} = \log \frac{3}{2} \text{ for } y = 0,1,2.$$

By the Intermediate Value Theorem, one finds a t with $0 < t < 1$ so that for

$$a(0) = \frac{t}{4},\ a(1) = \frac{1}{4},\ a(2) = \frac{3-t}{4}$$

one finds

$$\sum_x a(x) \log \frac{a(x)}{(P\alpha)(x)} = \log \frac{3}{2}.$$

We use this a in (5.2).

From Lemma 3.2, it follows now that $\alpha \in \bar{A}$, which together with $G_\alpha = 0$ gives $T_1 = 0$. In the case that $\alpha'(y) = \alpha(5-y)$ is α' also in $\bar{A}$ by Lemma 3.2. Then $P\alpha'$ is the uniform distribution in X, a is everywhere positive (but not the uniform distribution), so we have that $G\alpha' > 0$ and therefore $T_{-1} > 0$, which means $T_1 \neq T_{-1}$.

From the proof of Theorem 1.2, it is wise in this channel to emit for $\epsilon < \frac{1}{2}$ only the symbols 0,1,2 and for $\epsilon > \frac{1}{2}$ only the symbols 3,4,5 (for large n anyway). This correlates closely with the application of the method of Random Codes (see [1], [5], [6]): the $\alpha \in \bar{A}$ are not equivalent for the construction of a Random Code. The utility of an α depends on the error probability.

One can obtain the other estimate in Theorem 1.2 under consideration of these facts also with the help of the Method of Random Codes.

(iii) One defines an "ϵ-code in mean for P_n" as a mapping g from a subset of X_n to Y_n with

$$\frac{1}{N'} \sum_{y_n \in g(Y_n)} P_n(g^{-1}(y_n, y_n) \geq 1 - \epsilon, \tag{5.3}$$

where N' is the length of the code, and one denotes with $N'(n, \epsilon)$ the maximal length of ϵ-codes in mean for P_n, wherefore

$$\log N'(n, \epsilon) = nC - \lambda T_{\mathrm{sign}\lambda}\sqrt{n} + O(\log n). \tag{5.4}$$

Because every ϵ-code is an ϵ-code in mean for P_n, one has

$$N(n, \epsilon) \leq N'(n, \epsilon). \tag{5.5}$$

Conversely, in the case that g is an ϵ-code in mean for P_n with length N', let the $(\epsilon + 1/\sqrt{n})$-code for P_n be defined by

$$f = \left\{ (x_n, y_n) \middle| y_n = g(x_n), P_n(g^{-1}(y_n), y_n) \geq 1 - \epsilon - \frac{1}{\sqrt{n}} \right\}. \tag{5.6}$$

Let N be the length of f. From (5.3) and (5.6) it follows that

$$1 - \epsilon \leq \frac{1}{N'} \sum_{y_n \in f(X_n)} P_n(g^{-1}(y_n), y_n) + \frac{1}{N'} \sum_{y_n \in g(X_n) - f(X_n)} P_n(g^{-1}(y_n), y_n) \leq \frac{N}{N'} +$$

$$+ \frac{N' - N}{N'} \left(1 - \epsilon - \frac{1}{\sqrt{n}}\right) = 1 - \epsilon - \frac{1}{\sqrt{n}} + \frac{N}{N'}\left(\epsilon + \frac{1}{\sqrt{n}}\right),$$

hence

$$N' \leq (1 + \epsilon\sqrt{n})N$$

and therefore

$$N'(n, \epsilon) \leq (1 + \epsilon\sqrt{n}) N\left(n, \epsilon + \frac{1}{\sqrt{n}}\right). \tag{5.7}$$

Thus (5.5), (5.7), and Theorem 1.2 imply (5.4).

(iv) Suppose we are given a stationary source $(p_n)_{n \geq 1}$ with independent letters from an alphabet Z and a stationary, memoryless channel $(P_n)_{n \geq 1}$ from

Y to X. We write $(p_n \xrightarrow{P_n})_\epsilon$ whenever one can find, for sufficiently large n, a mapping χ from Z_n to Y_n and a mapping φ from X_n to Z_n such that

$$\sum_{z_n \in Z_n} P_n(\varphi^{-1}(z_n), \chi(z_n)) p_n(z_n) \geq 1 - \epsilon \tag{5.8}$$

and $(p_n \overset{P_n}{\not\rightarrow})_\epsilon$ whenever, for sufficiently large n, no such mapping exists. ($(p_n \xrightarrow{P_n})_\epsilon$ means, for example, that the Channel is capable of broadcasting the message of the source in perpetuity with probability of error $\leq \epsilon$).

In the case where H is the entropy of the source and C is the capacity of the channel, the well-known ([1], [6], [7]) relations of Shannon,

$$(p_n \xrightarrow{P_n})_\epsilon, \text{ for } 0 < \epsilon < 1 \text{ when } H < C$$

and

$$(p_n \overset{P_n}{\not\rightarrow})_\epsilon, \text{ for } 0 < \epsilon < 1 \text{ when } H > C$$

apply. We consider the case

$$H = C \tag{5.9}$$

and assume that at least one of S and T_1 is positive. Let $0 < \epsilon < 1$, $(p_n \xrightarrow{P_n})_\epsilon$, and n so large that there are χ and φ as in (5.8).

Let $\delta > \epsilon$ and < 1 and

$$E = \{z_n | z_n \in Z_n, P_n(\varphi^{-1}(z_n), \chi(n)) > 1 - \delta\}.$$

From (5.8), it follows that

$$1 - \epsilon \leq \sum_{z_n \in E} P_n(\varphi^{-1}(z_n), \chi(z_n)) p_n(z_n) +$$

$$+ \sum_{z_n \in Z_n - E} P_n(\varphi^{-1}(z_n), \chi(z_n)) p_n(z_n) \leq o_n(E) + (1 - \delta)(1 - p_n(E))$$

and from here

$$p_n(E) \geq 1 - \frac{\epsilon}{\delta}.$$

Theorem 1.1 and an easy consideration in the case that $S = 0$, produce

$$|E| \geq \beta\left(n, \frac{\epsilon}{\delta}\right) = \exp\left\{nH + S\lambda\left(\frac{\epsilon}{\delta}\right)\sqrt{n} + O(\log n)\right\}, \tag{5.10}$$

where $|E|$ is the cardinality of E.

Let the mapping f from $\varphi^{-1}E$ to Y_n be defined by

$$f(x_n) = \chi(\varphi(x_n)) \text{ for } x_n \in \varphi^{-1}E.$$

Then $f(X_n) = \chi(E)$ and by the definition of E

$$P_n(f^{-1}(y_n), y_n) = P_n(\bigcup_{\substack{z_n \in E \\ \chi(z_n) = y_n}} \varphi^{-1}(z_n), y_n) = \tag{5.11}$$

$$= \sum_{\substack{z_n \in E \\ \chi(z_n)=y_n}} P_n(\varphi^{-1}(z_n), \chi(z_n)) > (1-\delta)|\chi^{-1}(y_n) \cap E| \geq 1-\delta \text{ for } y_n \in f(X_n).$$

Hence f is a δ-code for P_n of length $|\chi(E)|$.

From (5.11), it follows that

$$|\chi^{-1}(y_n) \cap E| \leq \frac{1}{1-\delta} \text{ for } y_n \in \chi(E),$$

and we obtain

$$|E| \leq \frac{1}{1-\delta}|\chi(E)| \leq \frac{1}{1-\delta} N(n,\delta) =$$

$$= \exp\{nC - \lambda(\delta) T_{\operatorname{sign}\lambda(\delta)}\sqrt{n} + O(\log n)\}.$$

This, together with (5.10) and (5.9), gives

$$S\lambda\left(\frac{\epsilon}{\delta}\right)\sqrt{n} + O(\log n) \leq -\lambda(\delta) T_{\operatorname{sign}\lambda(\delta)}\sqrt{n} + O(\log n), \tag{5.12}$$

or

$$S\lambda\left(\frac{\epsilon}{\delta}\right) + \lambda(\delta) T_{\operatorname{sign}\lambda(\delta)} \leq 0 \tag{5.13}$$

for all δ with $\epsilon < \delta < 1$ as the necessary condition for $(p_n \overset{P_n}{\to})_\epsilon$. Indeed, we have proved something more. Namely, if, for a δ with $\epsilon < \delta < 1$, it holds that

$$\lambda\left(\frac{\epsilon}{\delta}\right) S + \lambda(\delta) T_{\operatorname{sign}\lambda(\delta)} > 0 \tag{5.14}$$

it holds even that $(p_n \overset{P_n}{\not\to})_\epsilon$, since (5.13) still follows whenever (5.12) is true for infinitely many n. From here one obtains by setting $\delta = 2\epsilon$, for example when $\delta = 1/2$,

$$(p_n \overset{P_n}{\not\to})_\epsilon \text{ when } \epsilon < \frac{1}{4}. \tag{5.15}$$

In the case where $S = 0$ or $T_{-1} = 0$, setting $\delta = \epsilon$ (if $S = 0$) or $\delta > 2\epsilon$ (if $T_{-1} = 0$) implies something sharper,

$$(p_n \overset{P_n}{\not\to})_\epsilon \text{ for } \epsilon < \frac{1}{2}.$$

Now let $0 < \gamma < 1$, $0 < \delta < 1$ so that

$$\lambda(\gamma) S + \lambda(\delta) T_{\operatorname{sign}\lambda(\delta)} < 0. \tag{5.16}$$

From Theorem 1.1 (or a simple observation in the case that $S = 0$) and Theorem 1.2, it follows that

$$\beta(n,\gamma) \leq N(n,\delta)$$

for sufficiently large n. For such n one can fins a set $E_{n,\gamma}$ with

$$p_n(E_{n,\gamma}) \geq 1 - \gamma$$

and a δ-code f for P_n so that

$$|E_{n,\gamma}| \leq |f(X_n)|.$$

Let χ_0 be a bijection from $E_{n,\gamma}$ in $f(X_n)$ and define χ and φ by

$$\chi(z_n) = \begin{cases} \chi_0(z_n) & : \quad z_n \in E_{n,\gamma} \\ y_n^0 & : \quad z_n \in Z_n - E_{n,\gamma} \end{cases}$$

$$\varphi(x_n) = \begin{cases} \chi_0^{-1}(f(x_n)) & : \quad x_n \in f^{-1}\chi_0(E_{n,\gamma}) \\ z_n^0 & : \quad x_n \in X_n - f^{-1}\chi_0(E_{n,\gamma}) \end{cases},$$

where y_n^0 and z_n^0 are fixed arbitrarily chosen elements of Y_n and Z_n, respectively.

It follows that

$$\sum_{z_n \in Z_n} P_n(\varphi^{-1}(z_n), \chi(z_n)) p_n(z_n) \geq \sum_{z_n \in E_{n,\gamma}} P_n(f^{-1}(\chi(z_n)), \chi(z_n)) p_n(z_n)$$

$$\geq (1-\delta) p_n(E_{n,\gamma}) \geq (1-\delta)(1-\gamma) = 1 - (\delta + \gamma - \delta\gamma),$$

so that

$$(p_n \overset{P_n}{\longrightarrow})_{\delta+\gamma-\delta\gamma}.$$

For $\delta = \gamma > 1/2$, (5.16) is always satisfied. Hence it holds that

$$(p_n \overset{P_n}{\longrightarrow})_\epsilon \text{ when } \epsilon > 3/4.$$

The case where either $S = 0$ or $T_1 = 0$ implies in a similar fashion as the above that

$$(p_n \overset{P_n}{\longrightarrow})_\epsilon \text{ when } \epsilon > 1/2.$$

I thank Mr. K. Jacobs for his support during the development of this work and him as well as Mr A. Renyi and Mr. R. Ahlswede for many interesting comments.

References

[1] C.E. Shannon: *A Mathematical Theory of Communication.* Bell System Tech. J., Vol. 27 (1948).

[2] A. Feinstein: *Foundations of Information Theory.* New York, McGraw Hill, 1958.

[3] Juschkewtisch: *Über einen Grenzwertsatz, der mit dem Begriff der Entropie einer Markoffschen Kette zusammenhängt.* Uspechi Mat. Nauk VIII, 5 (1953), (russian).

[4] B.V. Gnedenko, A.N. Kolmogoroff: *Grenzverteilungen für Summen von unabhängigen Zufallsvariablen.* Berlin 1959.

[5] J. Wolfowitz: *Coding Theorems of Information Theory.* Ergebnisse der Mathematik und ihrer Grenzgebiete, Berlin-Göttingen-Heidelberg 1961.

[6] R.L. Dobruschin: *Mathematical Problems in the Shannon Theory of Optimal Coding of Information.* Fourth Berkeley Symposium, 1961.

[7] L. Weiss: *On the Strong Converse of the Coding Theorem for Symmetric Channels Without Memory.* Quart. Appl. Math. 18, 3 (1960).

[8] K. Jacobs: *Almost Periodic Channels.* Colloquium on Combinatorial Methods in Probability Theory, Aarhus 1962.

[9] E.L. Lehmann: *Testing Statistical Hypotheses,* New York 1959.

[10] H. Cramér: *Random Variables and Probability Distributions.* Cambridge Tracts in Mathematics and Mathematical Physics, No. 36, London 1937.

[11] C.G. Esseen: *Fourier Analysis of Distribution Functions.* Acta Mathematica, Vol. 37.

[12] C.E. Shannon: *Some Geometric Results in Channel Capacity.* Nachr. Tech. Fachber., Vol. 6 (1956).

[13] F. and R. Nevanlinna: *Absolute Analysis, Die Grundlehren der Mathematischen Wissenschaften,* Band 102, Berlin-Göttingen-Heidelberg 1959.

[14] V. Strassen: *Messfehler und Information,* Zeitschrift für Wahrscheinlichkeitstheorie, 1964.

[15] G. Choquet: *Theory of Capacities,* Ann. Inst. Fourier, 5 (1953).

[16] K. Jacobs: *Bericht über Informationstheorie,* Jahresbericht der Deutschen Mathematiker Vereinigung, 1963.

[17] K. Jacobs: *Einführung in die Informationstheorie,* Göttingen 1959.

[18] H. Cramér: *Sur un nouveau théorème-limite de la théorie des probabilités.* Colloque consacré à la théorie probabilités, Paris 1938.

[19] W. Feller: *Generalization of a Probability Limit Theorem of Cramér.* Trans. Am. Math. Soc., 54 (1943), 3.

[20] A. Feinstein: *A New Basic Theorem of Information Theory.* I. R. E. Trans. I. T. – 4,2 (1954)

[21] P. Elias: *Coding for Two Noisy Channels,* Information Theory (C. Cherry, Editor), London 1956.

[22] R.M. Fano: *Transmission of Information.* New York, London 1961.