The Log-Volume of Optimal Codes for Memoryless Channels, Up to a Few Nats

Pierre Moulin
Dept of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 61801

***Abstract*—This paper derives a tight asymptotic upper bound on the maximum volume $M^*(n,\epsilon)$ of length-n codes for memoryless channels subject to an average decoding error probability ϵ: $M^*(n,\epsilon) = \exp\{nC - \sqrt{nV}\,\Phi^{-1}(\epsilon) + \frac{1}{2}\log n + A_{n,\epsilon} + o(1)\}$ where C is Shannon capacity, V is channel dispersion, Φ is the tail probability of the normal distribution, and $A_{n,\epsilon}$ is a bounded sequence that can be explicitly identified and reduces to a constant in the nonlattice case. A matching lower bound is presented, differing from the upper bound by a small multiplying constant. These expressions hold under certain regularity assumptions on the channel.**

I. Introduction

Shannon's seminal paper [1] introduced the fundamental capacity limits for memoryless communication channels. For any channel code of length n and tolerable decoding error probability ϵ, the maximum volume of the code is given by $M^*(n,\epsilon) = e^{nC+o(n)}$. The $o(n)$ is significant for practical values of n, hence much effort went into characterizing it in the early 1960's [2], [3], [4], [5]. In particular, Strassen [5] discovered that, under regularity assumptions, the $o(n)$ term is of the form $-\sqrt{nV}\,\Phi^{-1}(\epsilon) + O(\log n)$ where Φ is the tail probability of the normal distribution and V is the channel dispersion, or second-order coding rate. This line of research seemed forgotten until new ideas revived it, almost half a century later [6], [7]. The sharpest general result to date is

$$M^*(n,\epsilon) = \exp\{nC - \sqrt{nV}\,\Phi^{-1}(\epsilon) + O(\log n)\} \quad (1)$$

subject to some regularity conditions on the channel law.

The appeal of asymptotic expansions such as (1) is that (i) they convey significant insights into the essence of the problem and (ii) they are practically useful as the remainder $O(\cdot)$ term can be bounded and often neglected for moderate values of n, as demonstrated in [6].

Still there remains much mystery regarding the third term in (1), which has been characterized only for the binary symmetric channel (BSC) where it is $\frac{1}{2}\log n$ [6]. For the additive white Gaussian channel, the third term is sandwiched between $O(1)$ and $\frac{3}{2}\log n + O(1)$. For the binary erasure channel (BEC), the third term is $O(1)$. For discrete memoryless channels with finite input alphabet $\mathcal{X}$, under regularity assumptions on the channel law, the third term can be upper-bounded by $\frac{1}{2}(|\mathcal{X}| - 1)\log n$. The third term is often significant for moderate values of n.

This motivates a more refined analysis of the problem, in which asymptotic *equalities* for the relevant error probabilities are obtained using strong large-deviations analysis, which are closely related to Laplace's method for asymptotic expansion of integrals. A strong large-deviations analysis provides an asymptotic expansion for the probability of rare events such as $\sum_{i=1}^n U_i \geq na$ where $\{U_i,\, 1 \leq i \leq n\}$ are independent and identically distributed (iid) random variables, and a is strictly larger than the mean of U_1 [9], [10]. In contrast, the classical ("weak") large-deviations analysis merely states that the aforementioned probability vanishes as $\exp\{-n\Lambda(a) + o(n)\}$ where $\Lambda(\cdot)$ denotes the large-deviations function for U_1.

Using this approach, we establish that the third term in the asymptotic expansion of (1) is $\frac{1}{2}\log n$ in a very general setting. To do so, we derive a new result on conditional strong large-deviations analysis. We even obtain the fourth term in the expansion, which is a constant when the underlying loglikelihood random variables are of the nonlattice type and a bounded oscillating function of n otherwise. The quest for that term requires a precise characterization of the asymptotics of channel fluctuations. To this end we use two-term Edgeworth expansions [13], [14], [15], [16] in this regime and more specifically in work by Cramér [17] and Esseen [12] during the 1930's and 1940's. The Berry-Esseen theorem [14] was used in [5], [6] and provides a bound for deviation from Gaussianity, but that bound is not sharp enough for our purposes.

All the analysis and results in this paper are based on asymptotics. Only two inequalities are used. The first is the classical union-of-events bound, which is used for analysis of our random-coding scheme and turns out to be remarkably tight. The second inequality is one introduced in [6] for proving converse theorems: it provides an upper bound for $M^*(n,\epsilon)$ in terms of a maxmin optimization problem involving the power of a Neyman-Pearson test at significance level $1-\epsilon$. This is a remarkably powerful idea which can be traced back to Strassen [5, pp. 711, 712].

In order to keep the presentation focused, we assume broad regularity conditions on the channels of interest and exclude among others channels with zero dispersion and channels for which the capacity-achieving distribution is not (essentially) unique. Due to space limitations, only sketches of the proofs are given here; complete derivations are given in the full paper [8].

Notation. We use uppercase letters for random variables,

lowercase letters for their individual values, calligraphic letters for alphabets, and boldface letters for sequences. The set of all probability distributions over a finite set $\mathcal{X}$ is denoted by $\mathscr{P}(\mathcal{X})$. Mathematical expectation with respect to probability distribution P is denoted by the symbol $\mathbb{E}_P$. Given a distribution P on the random variable X and a conditional distribution W on another random variable Y given X, we denote by $P \times W$ the joint distribution on (X, Y) and by (PW) the marginal distribution on Y. The indicator function of a set $\mathcal{A}$ is denoted by $\mathbb{1}\{x \in \mathcal{A}\}$. All logarithms are natural logarithms. The probability density function (pdf) of the normal random variable is denoted by $\phi(x)$, $x \in \mathbb{R}$ and its cdf by $\Phi(x) = \int_{-\infty}^{x} \phi$.

The symbol $f(n) \sim g(n)$ denotes asymptotic equality: $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1$. The notations $f(n) = o(g(n))$ (small oh) and $f(n) = O(g(n))$ (big oh) indicate that $\lim_{n\to\infty} \frac{f(n)}{g(n)}$ is zero and finite, respectively.

A. Definitions

Let $\mathcal{X}$ and $\mathcal{Y}$ be two finite alphabets. Consider a memoryless channel $(W, \mathcal{X}, \mathcal{Y})$ characterized by input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$ and by a conditional probability density function $\{W(y|x), x \in \mathcal{X}, \ y \in \mathcal{Y}\}$. Given an input probability distribution P on $\mathcal{X}$, denote by $(PW)(y) = \sum_{x\in\mathcal{X}} P(x)W(y|x)$, $\forall y \in \mathcal{Y}$ the output probability distribution. Given $X = x$, the conditional pdf above will often be denoted by $W_x \in \mathscr{P}(\mathcal{Y})$. Kullback-Leibler divergence between two distributions P and Q on a common alphabet is denoted by $D(P\|Q) \triangleq \mathbb{E}_P[\ln \frac{P(X)}{Q(X)}]$, divergence variance by $V(P\|Q) \triangleq \mathbb{E}_P[\ln \frac{P(X)}{Q(X)}]^2 - D^2(P\|Q)$, and divergence third moment by $T(P\|Q) \triangleq \mathbb{E}_P[\ln \frac{P(X)}{Q(X)} - D(P\|Q)]^3$. Given two alphabets $\mathcal{X}$ and $\mathcal{Y}$, a $\mathcal{X}$-valued random variable X distributed as P, and two conditional distributions W and Q on a $\mathcal{Y}$-valued random variable Y given X, we denote by $D(W\|Q|P) = \mathbb{E}_{P\times W} \ln \frac{W(Y|X)}{Q(Y|X)}$ the conditional KL divergence between W and Q given P, and likewise by $V(W\|Q|P)$ and $T(W\|Q|P)$ the conditional divergence variance and the conditional divergence third moment.

A real random variable L is said to be of the lattice type if there exists numbers d and l_0 such that L belongs to the lattice $\{l_0 + kd, \ k \in \mathbb{Z}\}$ with probability 1. The largest d for which this holds is called the *span* of the lattice, and l_0 is the *offset*.

The empirical distribution (n-type) on $\mathcal{X}$ of a sequence $\mathbf{x} \in \mathcal{X}^n$ is defined by $\hat{P}_{\mathbf{x}}(x) \triangleq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{x_i = x\}$, $x \in \mathcal{X}$. We denote by $T[P]$ the set of all sequences of type P (type class), by $U_{\mathbf{X}|P}$ the uniform distribution over type class $T[P]$, and by $\mathscr{P}_n(\mathcal{X}) \subset \mathscr{P}(\mathcal{X})$ the set of n-types over $\mathcal{X}$.

Following Gallager [20, p. 17], define the mutual information between the events $X = x$ and $Y = y$ as

$$l(x, y) = \log \frac{W(y|x)}{(PW)(y)}, \quad x \in \mathcal{X}, y \in \mathcal{Y}. \tag{2}$$

For some DMCs with capacity-achieving distribution, the random variable $l(X, Y)$ is of the lattice type. For instance $l(X; Y) \in \{\log(2\lambda), \log(2-2\lambda)\}$ for the BSC with crossover probability $\lambda \neq \{0, \frac{1}{2}, 1\}$ and uniform input distribution; the span of the lattice is $\log|\frac{\lambda}{1-\lambda}|$. However for almost every *asymmetric* binary channel, as well as for almost every *nonbinary* channel (symmetric or not), $l(X; Y)$ is not of the lattice type. We refer to capacity problems where $l(X, Y)$ is of the lattice type as the *lattice case*. The lattice case is not considered in this paper.

Let $W_x \triangleq \{W(\cdot|x)\} \in \mathscr{P}(\mathcal{Y})$ for each $x \in \mathcal{X}$. We define the following moments of the random variable $l(X, Y)$ with respect to the joint distribution $P\times W$: the *unconditional* mean (= mutual information)

$$\begin{aligned} I(P, W) &= \mathbb{E}_{P\times W}[l(X, Y)] \\ &= D(P \times W\|P \times (PW)), \end{aligned} \tag{3}$$

the *conditional* mean (given X)

$$D(W_x\|PW) = \mathbb{E}_{W_x}[l(x, Y)], \quad x \in \mathcal{X},$$

the *unconditional* information variance

$$\begin{aligned} V_{\mathrm{u}}(P, W) &= \mathrm{Var}_{P\times W}[l(X, Y)] \\ &= V(P \times W\|P \times (PW)), \end{aligned} \tag{4}$$

the *conditional* information variance (given X)

$$\begin{aligned} V(P, W) &= \mathrm{Var}_{P\times W}[l(X, Y)|X] \\ &= \sum_x P(x)V(W_x\|PW), \end{aligned} \tag{5}$$

the *conditional* third central moment (given X)

$$T(P, W) = \sum_x P(x)T(W_x\|PW), \tag{6}$$

and the *conditional* skewness

$$S(P, W) = \frac{T(P, W)}{[V(P, W)]^{3/2}}. \tag{7}$$

We also define the nonnegative definite *Fisher information matrix* J whose components are

$$J_{xx'}(P, W) \triangleq -\frac{\partial^2 I(P, W)}{\partial P(x)\partial P(x')}, \quad \forall x, x' \in \mathcal{X}. \tag{8}$$

Its rank is at most $|\mathcal{X}| - 1$ but will be equal to $|\mathcal{X}| - 1$ at the capacity-achieving P under the assumptions of our theorems. Also define the weighted quadratic norm for zero-mean vectors $h \in \mathbb{R}^{\mathcal{X}}$

$$\|h\|_J \triangleq \sqrt{\sum_{x,x'\in\mathcal{X}} J_{xx'}(P, W)h(x)h(x')}. \tag{9}$$

We also use denote by $J^\dagger$ the pseudo-inverse of J and use the nonnegative quantity

$$A(P, W) \triangleq \frac{1}{V(P, W)} \sum_{x,x'} \frac{\partial V(P, W)}{\partial P(x)} \frac{\partial V(P, W)}{\partial P(x')} J^\dagger_{xx'}(P, W). \tag{10}$$

We denote by $\nabla V(P, W) \in \mathbb{R}^{|\mathcal{X}|}$ the gradient of $V(P, W)$ with respect to P.

Next, let the triple $(X', X, Y) \in \mathcal{X}^2 \times \mathcal{Y}$ be distributed according to the joint probability law $P_{X'XY}(x', x, y) = P(x')P(x)W(y|x)$. Define the *tilted joint distribution*

$$\tilde{P}_{X'XY}(x', x, y) \triangleq \frac{W(y|x')W(y|x)P(x)P(x')}{(PW)(y)}, \qquad (11)$$

which has the same X' and (X, Y) marginals as $P_{X'XY}$ but is symmetric in X' and X. The random variables $A \triangleq \ln \frac{W(Y|X')}{(PW)(Y)}$ and $B \triangleq \ln \frac{W(Y|X)}{(PW)(Y)}$ are generally dependent but have the same marginal owing to the symmetry property above. Denote by

$$\rho(P; W) = \frac{\mathrm{Cov}_{\tilde{P}}(A, B)}{\sqrt{\mathrm{Var}_{\tilde{P}}(A)\mathrm{Var}_{\tilde{P}}(B)}} \quad \in [-1, 1] \qquad (12)$$

the normalized correlation coefficient between A and B under $\tilde{P}_{X'XY}$. For an additive-noise channel, $\rho(P^*, W) = 0$ for the (uniform) capacity-achieving distribution P^*. For the Binary Erasure Channel (BEC), $\rho(P^*, W) = 1$ [8].

B. Shannon Capacity

The message m to be transmitted is drawn uniformly from the message set $\mathcal{M}_n = \{1, 2, \cdots, M_n\}$. A code is a pair of encoder mapping $f_n : \mathcal{M}_n \to \mathsf{F} \subset \mathcal{X}^n$, $\mathbf{x}(m) = f_n(m)$, and decoder mapping $g_n : \mathcal{Y}^n \to \mathcal{M}_n$, $\hat{m} = g_n(\mathbf{y})$. The code has volume (or size) M_n and rate $R_n = \frac{1}{n} \log M_n$. We denote by $M^*(n, \epsilon) \triangleq \max\{M_n : \exists (f_n, g_n) : P_e(f_n, g_n, W) \le \epsilon\}$ the maximum possible value of M_n for (n, ϵ) codes under the *average error probability* criterion

$$P_e(f_n, g_n, W) \triangleq \frac{1}{M_n} \sum_{m \in \mathcal{M}_n} \sum_{\mathbf{y}} W^n(\mathbf{y}|f_n(m)) \mathbb{1}\{g_n(\mathbf{y}) \neq m\}.$$

Shannon capacity is given by

$$C = \max_{P \in \mathscr{P}(\mathcal{X})} I(P; W) \qquad (13)$$

Problems involving cost constraints on the codewords require a different treatment and are not considered in this paper.

C. Main Result

Assume the following:

- (A1) The maximzer P^* of the mutual information in (13) is unique and $\mathcal{X}$ is its support set.
- (A2) $0 < V(P^*; W) < \infty$
- (A3) $|S(P^*; W)| < \infty$
- (A4) $|\rho(P^*; W)| < 1$.

Let

$$t_\epsilon \triangleq \Phi^{-1}(\epsilon), \quad V = V(P^*, W), \quad S = S(P^*; W),$$
$$\rho = \rho(P^*, W), \quad A = A(P^*, W), \qquad (14)$$

(hence $t_\epsilon > 0$ for $\epsilon < 1/2$) and

$$A_\epsilon = \frac{t_\epsilon^2}{8} A - \frac{S\sqrt{V}}{6}(t_\epsilon^2 - 1) + \frac{1}{2} t_\epsilon^2 + \frac{1}{2} \log(2\pi V). \qquad (15)$$

Theorem 1.1: Assume (A1)—(A4) hold. Then $\log M^*(n, \epsilon)$ satisfies

$$\log \overline{M}(n, \epsilon) + \log \sqrt{1 - \rho^2} - 1 \le \log M^*(n, \epsilon) \le \log \overline{M}(n, \epsilon)$$

where in the nonlattice case

$$\overline{M}(n, \epsilon) = nC - \sqrt{nV} t_\epsilon + \frac{1}{2} \log n + A_\epsilon + o(1). \qquad (16)$$

The lower bound is achieved by iid random codes drawn from the distribution

$$P_n^* = P^* - n^{-1/2} \frac{t_\epsilon}{2\sqrt{V}} J(P^*, W)^\dagger \cdot \nabla V(P^*, W) + O(1/n). \qquad (17)$$

Here P_n^* achieves the maximum of the functional

$$\zeta_{n,\epsilon}(P, W) \triangleq nI(P; W) - \sqrt{nV(P; W)} t_\epsilon \qquad (18)$$

over $P \in \mathscr{P}(\mathcal{X})$.

The result applies to a broad variety of channels but not to all. For instance, we have mentioned below (12) that $\rho = 1$ for the binary erasure channel (BEC), hence Assumption (A4) is not satisfied.

A sketch of the derivation of the lower and upper bounds is presented in Secs. III and IV, respectively. It is unclear whether the gap $1 - \log(1 - \rho^2)$ between lower and upper bounds should be attributed to the use of random codes, or to the union bound, or both. Either way, the gap can easily be computed explicitly via (12) and may be as small as one nat. Hence random codes perform well, and the union bound is quite tight.

D. Special Cases

Given a finite input alphabet $\mathcal{X}$, a channel is said to be cyclic-symmetric if for any input distribution P, the mutual information $I(P; W)$ is invariant to permutations of $\{P(x), x \in \mathcal{X}\}$. This is the case when for instance, $\mathcal{X} = \mathcal{Y}$ and the matrix $W(y|x)$ is circulant Toeplitz [18].

The first two properties below is well known, and the next three are immediate.

Proposition 1.2: For a cyclic-symmetric channel, the following hold:

- (i) The capacity-achieving distribution P^* is uniform on $\mathcal{X}$.
- (ii) If $|\mathcal{X}| = |\mathcal{Y}|$, the output distribution (P^*W) is also uniform.
- (iii) The variance $V(W_x \| PW)$ and third central moment $T(W_x \| PW)$ are independent of x.
- (iv) The partial derivatives $\frac{\partial V(P)}{\partial p(x)}\Big|_{P=P^*}$ are independent of x. Equivalently, the gradient vector $\nabla V(P^*)$ has identical components.
- (v) $P_n^* = P^* + O(1/n)$ and $\Delta = 0$ in (17) and (**??**), respectively.

II. Refined Asymptotics

In this section we present three results that are used to prove the direct coding theorem and the converse. The first is a known refinement on the Central Limit Theorem (CLT) [14]. The second is a new strong large-deviations result for Neyman-Pearson tests. The third is a new conditional strong large-deviations result.

A. Central Limit Asymptotics

Under some conditions, a normalized sum of independent random variables converges in distribution to a normal pdf. Consider first iid random variables $U_i, 1 \le i \le n$ with common cdf F_U, finite mean μ, variance $\sigma^2 > 0$, and skewness $S \triangleq \mathbb{E}[(U-\mu)^3]/\sqrt{V^3}$. The normalized random variable

$$T_n = \frac{\sum_{i=1}^n U_i - n\mu}{\sqrt{n\sigma^2}}$$

has zero mean and unit variance and converges in distribution to $\mathcal{N}(0,1)$. Denote by F_n the cdf of Z_n. The Cramér-Esséen theorem [12] [14, p. 538] for non-lattice random variables states that

$$F_n(t) = \Phi(t) - \frac{S}{6\sqrt{n}}(1-t^2)\,\phi(t) + o(1/\sqrt{n}) \tag{19}$$

uniformly in t. Higher-order expansions in terms of successive powers of $n^{-1/2}$ can also be derived (Edgeworth expansions [13], [14], [15], [16]) but a two-term expansion suffices for our purposes. The Berry-Esseen formula $|F_n(t) - \Phi(t)| \le \frac{\zeta/V^{3/2}}{6\sqrt{n}}$ (where $\zeta = \mathbb{E}[|U-\mu|^3]$ is the *absolute* third central moment of U) which was used in [5], [6] is not sufficiently refined to yield the sharper asymptotics of interest here.

Let t_ϵ and $t_{\epsilon,n}$ be the ϵ-quantiles of Φ and F_n respectively, i.e.,

$$\Phi(t_\epsilon) = F_n(t_{\epsilon,n}) = \epsilon. \tag{20}$$

Then the Cornish-Fisher inversion formula [13], [16], [15] yields

$$t_{\epsilon,n} = t_\epsilon + \frac{S}{6\sqrt{n}}(t_\epsilon^2 - 1) + o(n^{-1/2}). \tag{21}$$

In case U_i, $1 \le i \le n$ are independent but have different distributions, with respective means μ_i and variances σ_i^2 such that $V_n \triangleq \frac{1}{n}\sum_{i=1}^n \sigma_i^2$ is bounded away from 0 and ∞ as $n \to \infty$, let

$$\overline{S}_n = \frac{\frac{1}{n}\sum_{i=1}^n \mathbb{E}[(U_i-\mu_i)^3]}{V_n^{3/2}}. \tag{22}$$

Then (19) holds with S replaced by $\overline{S}_n$ [14, pp. 546, 547]. [1]

B. Strong Large Deviations for Binary Hypothesis Testing

Lemma 2.1: (*Second-order Taylor series expansion of large-deviations function for binary hypothesis testing*.) Consider two probability measures P and Q over a common space and assume that P is dominated by Q ($P \ll Q$). Assume $D = D(P\|Q)$ and $V = V(P\|Q)$ are positive and finite, and $T = T(P\|Q)$ is finite. Let $\Lambda(a) = \sup_s[as - \kappa(s)]$ where the cumulant generating function (cgf) $\kappa(s) = \ln \mathbb{E}_Q[(\frac{dP}{dQ})^s]$. Then

$$\Lambda(a) = a + \frac{(a-D)^2}{2V} + O((a-D)^3) \quad \text{as } a \uparrow D. \tag{23}$$

Proof: see [8].

[1] Note a typo in [14, Eqn (6.1)], where s_n^2 should be replaced with s_n^3.

Now let Y_i, $1 \le i \le n$ be independent random variables with respective distributions P_i and Q_i under hypotheses H_1 and H_0, respectively. Write $\mathbb{P}_n \triangleq \prod_{i=1}^n P_i$ and $\mathbb{Q}_n \triangleq \prod_{i=1}^n Q_i$. Assume the following holds:

(A1) For each $i \ge 1$: $P_i \ll Q_i$ (P_i is dominated by Q_i), and P_i and Q_i have respective densities p_i and q_i with respect to a dominating measure ν.

(A2) Under P_i, the loglikelihood ratio $L_i \triangleq \ln p_i(Y_i)/q_i(Y_i)$ has finite mean $D_i = D(P_i\|Q_i)$, positive and finite variance $V_i = V(P_i\|Q_i)$, and finite third central moment $T_i = T(P_i\|Q_i)$. Let

$$\overline{D}_n = \frac{1}{n}\sum_{i=1}^n D_i, \quad \overline{V}_n = \frac{1}{n}\sum_{i=1}^n V_i,$$
$$\overline{T}_n = \frac{1}{n}\sum_{i=1}^n T_i, \quad \overline{S}_n = \frac{\overline{T}_n}{[\overline{V}_n]^{3/2}}.$$

Proposition 2.2: Let $t > 0$ and $B \in \mathbb{R}$ be arbitrary constants. Assume (A1)—(A2) hold.

(i) If $\sum_{i=1}^n L_i$ is is not a lattice random variable, then

$$\mathbb{Q}_n\left[\sum_{i=1}^n \ln\frac{p_i(Y_i)}{q_i(Y_i)} \ge n\overline{D}_n - \sqrt{n\overline{V}_n}\,t + B\right]$$
$$= \frac{\exp\{-n\overline{D}_n + \sqrt{n\overline{V}_n}\,t - (\frac{t^2}{2} + B) + o(1)\}}{\sqrt{2\pi n\overline{V}_n}} \text{ as } n \to \infty. \tag{24}$$

(ii) If $\sum_{i=1}^n L_i$ is is a lattice random variable, denote by d_n its span and by Ω_n its range. Then (24) holds if the right side is multiplied by a sequence γ_n that can be explicitly identified, is bounded from above and below, and takes the value $d_n/(1-e^{-d_n})$ for $n\overline{D}_n - \sqrt{n\overline{V}_n}\,t + B \in \Omega_n$.

Proof. For each $i \ge 1$, the random variable $L_i = \ln\frac{p_i(Y_i)}{q_i(Y_i)}$ has cgf $\kappa_i(s) = \ln\int q_i^{1-s}p_i^s\,d\nu$ (negative Chernoff distance) under Q_i. Since $\{L_i\}$ are mutually independent, the cgf for $\sum_{i=1}^n L_i$ is

$$n\overline{\kappa}_n(s) = \sum_{i=1}^n \kappa_i(s).$$

Since

$$\kappa_i(1) = 0, \quad \kappa_i'(1) = D_i, \quad \kappa_i''(1) = V_i,$$

Averaging over $i = 1, 2, \cdots, n$ yields

$$\overline{\kappa}_n(1) = 0, \quad \overline{\kappa}_n'(1) = \overline{D}_n, \quad \overline{\kappa}_n''(1) = \overline{V}_n.$$

Denote by

$$\overline{\Lambda}_n(a) = \sup_{s\in\mathbb{R}}[as - \overline{\kappa}_n(s)] \tag{25}$$

the large-deviations function for $\sum_{i=1}^n L_i$. Assume the supremum defining $\overline{\Lambda}_n(a_n)$ is achieved at s_n, hence $a_n = \overline{\kappa}_n'(s_n)$.

Applying Lemma 2.1 with $\mathbb{P} = \prod_{i=1}^n P_i$, $\mathbb{Q} = \prod_{i=1}^n Q_i$, evaluating (23) at a equal to

$$a_n = \overline{D}_n - t\sqrt{\overline{V}_n/n} + B/n \tag{26}$$

and multiplying by n, we obtain

$$n\overline{\Lambda}_n(a_n) = n\overline{D}_n - \sqrt{n\overline{V}_n}\,t + \left(B + \frac{t^2}{2}\right) + O(n^{-1/2}).$$

Moreover $s_n \to 1$ as $a_n \to \overline{D_n}$.

(i) Nonlattice case: From [11, Theorem 3.3] (with a_n, s_n, $\overline{\kappa}_n$ and $\overline{\Lambda}_n$ respectively playing the roles of m_n, τ_n, ψ_n and γ_n in [11]), we have

$$\mathbb{Q}_n\left[\sum_{i=1}^n L_i \ge na_n\right] \sim \frac{e^{-n\overline{\Lambda}_n(a_n)}}{s_n\sqrt{2\pi n\overline{\kappa}_n''(s_n)}} \quad \text{as } n\to\infty$$

where $L_i = \ln\frac{p_i(Y_i)}{q_i(Y_i)}$ for $1 \le i \le n$. Since $s_n \to 1$ as $n \to \infty$, this proves (24).

(ii) Lattice case: From [11, Theorem 3.5] we have, for $na_n \in \Omega_n$,

$$\mathbb{Q}_n\left[\sum_{i=1}^n L_i \ge na_n\right] \sim \frac{d_n}{1-e^{-s_n d_n}}\,\frac{e^{-n\overline{\Lambda}_n(a_n)}}{s_n\sqrt{2\pi n\overline{\kappa}_n''(s_n)}} \quad \text{as } n\to\infty.$$

This proves the second part of the claim. □

C. Conditional Strong Large Deviations

Fix P and define $P_{X'XY}(x',x,y) = P_X(x')P_X(x)W(y|x))$, thus X' is independent of (X,Y). The joint distribution $P_{X'XY}$ has the same X' and (X,Y) marginals as $P_{X'XY}$ but is symmetric in X' and X. The random variables $L' \triangleq \ln\frac{W(Y|X')}{(PW)(Y)}$ and $L \triangleq \ln\frac{W(Y|X)}{(PW)(Y)}$ are generally dependent but have the same marginal owing to the symmetry property above. Denote by $\rho = \rho(P;W)$ the normalized correlation coefficient of (12) between L and L' under $\tilde{P}_{X'XY}$. Let $D = I(P;W)$ and $V = V(P;W)$.

Analogously to Lemma 2.1, we have

Lemma 2.3: (*Second-order Taylor series expansion of large-deviations function*.) Assume $D = I(P;W)$ and $V_{\rm u} = V_{\rm u}(P;W)$ are positive and finite, and $T = T(P;W)$ is finite. Let

$$\Lambda(\alpha) = \sup_{s,t}[\alpha s + Dt - \kappa(s,t)] \tag{27}$$

where

$$\kappa(s,t) = \ln \mathbb{E}_P\left[\left(\frac{W(Y|X')}{(PW)(Y)}\right)^s\left(\frac{W(Y|X)}{(PW)(Y)}\right)^t\right].$$

Then

$$\Lambda(D-\eta) = D - \eta + \frac{\eta^2}{2(1-\rho^2)V_{\rm u}} + O(\eta^3) \quad \text{as } \eta\downarrow 0. \tag{28}$$

The supremum is achieved by

$$\begin{aligned} s(\eta) &= 1 - \frac{\eta}{(1-\rho^2)V_{\rm u}} + O(\eta^2) \\ t(\eta) &= \frac{\eta\rho}{(1-\rho^2)V_{\rm u}} + O(\eta^2). \end{aligned} \tag{29}$$

Now let

$$Z_n = \sum_{i=1}^n \ln\frac{W(Y_i|X_i)}{(PW)(Y_i)}, \quad T_n \triangleq \frac{-Z_n + nI(P;W)}{nV_{\rm u}(P;W)}, \tag{30}$$

$$Z_n' = \sum_{i=1}^n \ln\frac{W(Y_i|X_i')}{(PW)(Y_i)}, \quad T_n' \triangleq \frac{-Z_n' + nI(P;W)}{nV_{\rm u}(P;W)}. \tag{31}$$

Proposition 2.4: If $|\rho| \ne 1$ then for any $t \in \mathbb{R}$,

$$\begin{aligned} &P_{X'XY}^n\left[Z_n' \ge nD - \sqrt{nV_{\rm u}}t \;\middle|\; \frac{Z_n - nD}{\sqrt{nV_{\rm u}}} \in [t, t+dt)\right] \\ &= \frac{\exp\{-nD + \sqrt{nV_{\rm u}}t - \frac{t^2}{2} + o(1)\}}{\sqrt{2\pi(1-\rho^2)nV_{\rm u}}} \text{ as } n\to\infty. \end{aligned} \tag{32}$$

Notes.

(i) The event on the left of (32) is a rare event but the conditioning event is in the central regime.

(ii) If L' and L are independent, the conditioning can be removed, $\rho = 0$, and thus the expression (32) reduces to that given by Prop. 2.2 (with iid $\{L_i\}$, and $B = 0$).

(iii) When L' and L are dependent, that dependency only affects the multiplying constant $(1-\rho^2)^{-1/2} \ge 1$ in the asymptotic expression (32).

Sketch of the Proof: Let $\alpha_n = D - \sqrt{V_{\rm u}/n}\,t$. By application of Lemma 2.3 with $\eta = \sqrt{V_{\rm u}/n}\,t$, we obtain from (28)

$$\begin{aligned} n\Lambda(\alpha_n) &= s_n\alpha_n + t_nD - \kappa(s_n,t_n) \\ &= nD - \sqrt{nV_{\rm u}}t + \frac{t^2}{2(1-\rho^2)} \end{aligned} \tag{33}$$

where from (29)

$$\begin{aligned} s_n &= 1 - \frac{t}{(1-\rho^2)\sqrt{nV_{\rm u}}} + O(1/n), \\ t_n &= \frac{\rho t}{(1-\rho^2)\sqrt{nV_{\rm u}}} + O(1/n). \end{aligned} \tag{34}$$

Denote by $\nabla\kappa(s,t) \in \mathbb{R}^2$ and $\nabla^2\kappa(s,t) \in \mathbb{R}^{2\times 2}$ the gradient and the Hessian of κ at (s,t). By our assumption, for α_n in a neighborhood of the limit point α, the supremum defining $\Lambda(\alpha_n) = \sup_{s,t}[s\alpha_n + tD - \kappa(s,t)]$ is achieved at (s_n,t_n) satisfying $\nabla\kappa(s_n,t_n) = (\alpha_n, D)$. Since $\kappa(\cdot,\cdot)$ is twice continuously differentiable, (s_n,t_n) converges to $(1,0)$ and $\nabla^2\kappa(s_n,t_n)$ converges to R as $\alpha_n \to D$. Define the exponentially tilted distribution

$$\tilde{P}_{L'L}(dl',dl) \triangleq e^{s_nl' + t_nl - \kappa(s_n,t_n)}P_{LL'}(dl',dl) \tag{35}$$

on $\mathscr{B}(\mathbb{R}^2)$. We have

- $\mathbb{E}_{\tilde{P}}(L',L) = \nabla\kappa(s_n,t_n) = (\alpha_n, D)$.
- $\mathrm{Cov}_{\tilde{P}}(L',L) = \nabla^2\kappa(s_n,t_n)$. Denote by ρ_n the normalized correlation coefficient for L' and L under $\tilde{P}$.

As $n\to\infty$, (s_n,t_n) converges to $(1,0)$ and thus $\nabla^2\kappa(s_n,t_n)$ converges to $\nabla^2\kappa(1,0) = V_{\rm u}\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}$ and ρ_n to ρ given in the statement of the proposition.

Define the normalized random variables

$$T_n' = \frac{\sum_{i=1}^n L_i' - n\alpha_n}{\sqrt{n(\nabla^2\kappa(s_n,t_n))_{11}}} \quad \text{and} \quad T_n = \frac{\sum_{i=1}^n L_i - nD}{\sqrt{nV_{\rm u}}}. \tag{36}$$

Under $\tilde{P}$, both T_n' and T_n have zero mean and unit variance, and their correlation coefficient is ρ_n. By the Central Limit Theorem, their joint distribution $P_{T_n'T_n}$ converges to a normal distribution, and the conditional distribution $P_{T_n'|T_n=w}$ to $\mathcal{N}(\rho w, 1-\rho^2)$, for any w in the range of T_n. The value of

that Gaussian pdf at zero is $\frac{1}{\sqrt{2\pi(1-\rho^2)}}\exp\{-\frac{w^2\rho^2}{2(1-\rho^2)}\}$. Then we have

$$P_L^n\left[T_n\in[\beta,\beta+d\beta)\right]\to\int_{w=\beta}^{\beta+d\beta}P_{T_n}(dt) \tag{37}$$

and

$$\begin{aligned}
&P_{L'L}^n\left[\sum_{i=1}^n L_i'\ge n\alpha_n,\frac{\sum_{i=1}^n L_i-nD}{\sqrt{nV_{\mathrm{u}}}}\in[\beta,\beta+d\beta)\right]\\
&=\ \mathbb{E}_{P_{X'XY}}\mathbb{1}\left\{\sum_{i=1}^n L_i'\ge n\alpha_n,\frac{\sum_{i=1}^n L_i-nD}{\sqrt{nV_{\mathrm{u}}}}\in[\beta,\beta+d\beta)\right\}\\
&\overset{(a)}{=}\ \exp\{-n[s_n\alpha_n+t_nD-\kappa(s_n,t_n)]\}\\
&\quad\mathbb{E}_{\tilde{P}_{X'XY}}\exp\left\{-s_n\left[\sum_{i=1}^n L_i'-n\alpha_n\right]-t_n\left[\sum_{i=1}^n L_i-nD\right]\right\}\\
&\quad\times\mathbb{1}\left\{\sum_{i=1}^n L_i'\ge n\alpha_n,\frac{\sum_{i=1}^n L_i-nD}{\sqrt{nV_{\mathrm{u}}}}\in[\beta,\beta+d\beta)\right\}\\
&\overset{(b)}{=}\ e^{-n\Lambda(\alpha_n)}\int_{v=0}^{\infty}\int_{w=\beta}^{\beta+d\beta}P_{T_n'T_n}(dv,dw)\\
&\quad\exp\{-s_n\sqrt{n(\nabla^2\kappa(s_n,t_n))_{11}}\,v-t_n\sqrt{nV_{\mathrm{u}}}\,w\}\\
&\overset{(c)}{=}\ \frac{\exp\{-nD+\sqrt{nV_{\mathrm{u}}}t-\frac{t^2}{2}+o(1)\}}{\sqrt{2\pi nV_{\mathrm{u}}(1-\rho^2)}}\left(\int_{w=\beta}^{\beta+d\beta}P_{T_n}(dw)\right).
\end{aligned}\tag{38}$$

where (a) follows from (35), (b) from (36), and (c) after some manipulations. Combining (37) and (38) establishes the claim. □

III. Achievability: Sketch of the Proof

The number of codewords is $M_n=\lfloor e^{nR_n}\rfloor$ where $R_n=\log\overline{M}(n,\epsilon)+\log\sqrt{1-\rho^2}-1$.

Random coding scheme. The codewords $\{\mathbf{x}(m),\,1\le m\le M_n\}$ are drawn iid P_n^* where P_n^* is given in (17). Define the random variables

$$Z_{m,n}\triangleq\ln\frac{W^n(\mathbf{Y}|\mathbf{X}(m))}{(P_n^*W)^n(\mathbf{Y})},\quad 1\le m\le M_n. \tag{39}$$

Hence $Z_{m,n}$ is a loglikelihood score minus the constant $\ln(P_n^*W)^n(\mathbf{Y})$. The ML decoding rule can be written as

$$\hat{m}=\arg\max_{1\le m\le M_n}Z_{m,n}. \tag{40}$$

In case of a tie, an error is declared.

Overview of error probability analysis. By symmetry of the codebook construction and the decoding rule, the error probability for message m is independent of m. For the calculation below, we assume without loss of generality that $m=1$ was sent. We use the bound

$$\begin{aligned}
\Pr[\text{Error}]&=\Pr\left[\max_{m\ge2}Z_{m,n}\ge Z_{1,n}\right]\\
&=\int P_{Z_{1,n}}(dz)\Pr\left[\max_{m\ge2}Z_{m,n}\ge z|Z_{1,n}=z\right]\\
&\le\int P_{Z_{1,n}}(dz)\,\min\{1,(M_n-1)\Pr[Z_{2,n}\ge z|Z_{1,n}=z]\}\\
&=P_{Z_{1,n}}[Z_{1,n}\le z_n^*]+(M_n-1)\\
&\quad\times\int_{z>z_n^*}P_{Z_{1,n}}(dz)P_{Z_{2,n}|Z_{1,n}}[Z_{2,n}\ge z|Z_{1,n}=z]
\end{aligned}\tag{41}$$

where the inequality follows from the union bound. In the last line, z_n^* is derived so that $(M_n-1)\Pr[Z_{2,n}\ge z_n^*|Z_{1,n}=z_n^*]=1$. We then use precise asymptotics for $P_{Z_{1,n}}$ and $P_{Z_{2,n}|Z_{1,n}}$ to derive the desired result.

The statistics of $Z_{1,n}$ are obtained from (20) (21). Defining $T_n=(Z_{1,n}-\mathbb{E}[Z_{1,n}])/\sqrt{\mathrm{Var}(Z_{1,n})}$ as in (30), we obtain $F_{T_n}(t_{\epsilon,n})=1-\epsilon+o(n^{-1/2})$. The conditional probability $P_{Z_{2,n}|Z_{1,n}}[Z_{2,n}\ge z|Z_{1,n}=z]$ in (41) is evaluated using Prop. 2.4. The threshold z_n^* is selected as $z_n^*=\mathbb{E}[Z_{1,n}]-\sqrt{\mathrm{Var}[Z_{1,n}]}t_n^*$ where $t_n^*=t_{\epsilon,n}+\frac{1}{\sqrt{nV}}$. Then the two terms in the right side of (41) are respectively given by $\epsilon-\frac{\phi(t_\epsilon)}{\sqrt{nV}}+o(n^{-1/2})$ and $\frac{\phi(t_\epsilon)}{\sqrt{nV}}(1+o(1))$. The $O(n^{-1/2})$ terms cancel out and the desired result follows.

IV. Converse: Sketch of the Proof

A. Background

Some background from [6] is presented here.

Theorem 4.1: [6, Theorem 27 p. 2318] Every (M,ϵ) code with codewords in $\mathsf{F}\subseteq\mathcal{X}^n$ satisfies

$$M\le\sup_{P_{\mathbf{X}}}\inf_{Q_{\mathbf{Y}}}\frac{1}{\beta_{1-\epsilon}(P_{\mathbf{XY}},P_{\mathbf{X}}\times Q_{\mathbf{Y}})}$$

where the supremum is over all probability distributions over F, and the infimum is over all probability distributions over $\mathcal{Y}^n$.

In some cases $\beta_\alpha(P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}},Q_{\mathbf{Y}})$ is constant for all $\mathbf{x}\in\mathsf{F}$, e.g., when $Q_{\mathbf{Y}}=Q_Y^n$ is the n-fold product of a distribution Q_Y over $\mathcal{Y}$, and when the empirical distribution of the codewords over the alphabet $\mathcal{X}$ is the same for all $\mathbf{x}\in\mathsf{F}$. With some abuse of notation, we write $\beta_\alpha(\hat{P}_{\mathbf{x}},Q_Y^n)=\beta_\alpha(P_{\mathbf{Y}|\mathbf{X}=\mathbf{x}},Q_Y^n)$. Then the following result holds.

Theorem 4.2: Fix a distribution Q_Y over $\mathcal{Y}$ and a type $P\in\mathscr{P}(\mathcal{X})$. Then every (M,ϵ) code with codewords in $T[P]$ satisfies

$$M(n,\epsilon)\le\frac{1}{\beta_{1-\epsilon}(P,Q_Y^n)}.$$

This result is used in [6] to derive a converse theorem for constant-composition codes. Denote by $M_{cc}^*(n,\epsilon)$ the maximal number of codewords for any constant-composition code over the DMC W, with *maximal error probability* ϵ. The following result holds.

Theorem 4.3: [6, Theorem 48 p. 2331]. Fix a DMC W. If $0 < \epsilon \leq 1/2$, there exists a constant $F > 0$ such that

$$\log M_{cc}^*(n,\epsilon) \leq nC - \sqrt{nV}\Phi^{-1}(\epsilon) + \frac{1}{2}\ln n + F.$$

We have combined Theorem 4.2 and strong large-deviations analysis to refine Theorem 4.3 as follows:

$$\log M_{cc}^*(n,\epsilon) \leq nC - \sqrt{nV}t_\epsilon + \frac{1}{2}\log n + A_\epsilon - \Delta_{cc} + o(1) \quad (42)$$

where Δ_{cc} is a positive constant.

B. General Codes

Each codeword $\mathbf{x} \in \mathcal{X}^n$ has a type $\hat{P}_\mathbf{x} \in \mathscr{P}(\mathcal{X})$. For the constant-composition codes of the previous section, $\hat{P}_\mathbf{x}$ is the same for all codewords. For a more general code, $\hat{P}_\mathbf{x}$ is not fixed but has a (nondegenerate) empirical distribution π_n over $\mathscr{P}(\mathcal{X})$. That is,

$$\pi_n(\mathcal{A}) = \frac{1}{M_n}\sum_{1\leq m\leq M_n} \mathbb{1}\{\hat{P}_{\mathbf{x}(m)} \in \mathcal{A}\}$$

for all collections $\mathcal{A}$ of types. We refer to π_n as the type distribution of the code.

Denote by $\mathscr{P}^{p.i.}(\mathcal{X}^n)$ the set of all permutation-invariant distributions over $\mathcal{X}^n$. Define $U_{\mathbf{X}|\theta}$ as the uniform distribution over the type class $T[\theta]$. Clearly $U_{\mathbf{X}|\theta}$ is permutation-invariant for each θ, and so is any convex combination $P_\mathbf{X} = \int \pi_n(d\theta)U_{\mathbf{X}|\theta}$. With some abuse of notation, define the error probability

$$\beta_{1-\epsilon}(\pi) = \beta_{1-\epsilon}(P_\mathbf{X}W^n, P_\mathbf{X}Q^\pi)$$
$$\text{with } P_\mathbf{X} = \int \pi(d\theta)U_{\mathbf{X}|\theta} \in \mathscr{P}^{p.i.}(\mathcal{X}^n).$$

Here Q^π denotes a strategy function, i.e., a choice of $Q \in \mathscr{P}(\mathcal{Y}^n)$ that may depend on π.

The NP test is a randomized likelihood ratio test. For $P_\mathbf{X} = (P_n^*)^n$ and $Q = (P_n^*W)^n$, where P_n^* is given by (17), application of Prop. 2.2 and refined CLT asymptotics of Sec. II-A yield the optimal threshold and the asymptotic type-II error probability

$$\beta_{1-\epsilon}(\pi) = 1/\exp\{nC - \sqrt{nV}t_\epsilon + \frac{1}{2}\ln n + A_\epsilon + o(1)\}. \quad (43)$$

One difficulty with Theorem 4.1 [6] is that the minimization over all probability distributions $P_\mathbf{X}$ over $\mathcal{X}^n$ is apparently intractable. However the minimization problem can be considerably simplified, as stated in the lemma below.

Proposition 4.4: Fix a permutation-invariant strategy Q^π. Then

$$\begin{aligned}&\inf_{P_\mathbf{X}\in\mathscr{P}(\mathcal{X}^n)} \beta_{1-\epsilon}(P_\mathbf{X}W^n, P_\mathbf{X}\times Q^{\pi_n})\\ &= \inf_{P_\mathbf{X}\in\mathscr{P}^{p.i.}(\mathcal{X}^n)} \beta_{1-\epsilon}(P_\mathbf{X}W^n, P_\mathbf{X}\times Q^{\pi_n})\\ &\geq \inf_{\pi\in\mathscr{P}(\Theta)} \beta_{1-\epsilon}(\pi)\end{aligned}$$

Proof. Consider a random variable Ω that is uniformly distributed over the set of all $n!$ permutations of the set $\{1, 2, \cdots, n\}$. Denote by $\omega\mathbf{x}$ the sequence obtained by applying permutation ω to a sequence $\mathbf{x} \in \mathcal{X}^n$. Given any distribution $P_\mathbf{X}$ on $\mathcal{X}^n$, the permutation-averaged distribution $P_{\Omega\mathbf{X}}$ is permutation-invariant.

Since Q is permutation-invariant, the error probability $\beta_{1-\epsilon}(P_{\omega\mathbf{X}}W^n, P_{\omega\mathbf{X}}\times Q)$ is independent of ω. Hence, by the same arguments as in [6, Lemma 29], we have

$$\forall\omega: \quad \beta_{1-\epsilon}(P_{\omega\mathbf{X}}W^n, P_{\omega\mathbf{X}}\times Q) = \beta_{1-\epsilon}(P_{\Omega\mathbf{X}}W^n, P_{\Omega\mathbf{X}}\times Q)$$

The claim follows by taking the infimum over $P_\mathbf{X} \in \mathscr{P}(\mathcal{X}^n)$. □

Proposition 4.5: Every (M,ϵ) code with type distribution π satisfies

$$M(n,\epsilon) \leq \frac{1}{\beta_{1-\epsilon}(\pi)}. \quad (44)$$

Proof. By Theorem 4.1,

$$M \leq \sup_{P_\mathbf{X}} \frac{1}{\beta_{1-\epsilon}(P_\mathbf{X}W^n, P_\mathbf{X}\times Q^{\pi_n})}$$

By Prop. 4.4, the supremum may be taken over the set $\mathscr{P}^{p.i.}(\mathcal{X}^n)$ of permutation-invariant distributions. Then

$$\begin{aligned}\beta_{1-\epsilon}(P_\mathbf{X}W^n, P_\mathbf{X}\times Q^{\pi_n}) &= \beta_{1-\epsilon}(\pi_n)\\ &\geq \inf_\pi \beta_{1-\epsilon}(\pi).\end{aligned}$$

□

Given some π_n, the derivations of the asymptotics of the NP test is not straighforward because the loglikelihood ratio test statistic is generally not the sum of independent random variables, an exception being the product case $P_\mathbf{X} = P^n$, which lead to (43). To prove the converse for general codes, we separate the proof into different cases, where the type distribution π_n does not concentrate near the capacity-achieving distribution P^* (Class I), where π_n concentrates near P^* but slowly (Classes II, III, IV) and where π_n concentrates near P_n^* at the $O(n^{-1/2})$ scale (Class V). A different strategy Q^π is chosen in each case, and strong large-deviation analysis is used to prove the claims. Then

Class I (First-order suboptimal codes). There exists $\delta > 0$ such that

$$\limsup_{n\to\infty} \pi_n\left[\|\theta - P^*\|_J \geq \delta\right] > 0.$$

Class II (Second-order suboptimal codes). There exists $c > 0$ and a sequence δ_n such that $n^{-1/4} \leq \delta_n \ll 1$ and

$$\limsup_{n\to\infty} \pi_n[\|\theta - P^*\|_J \geq \delta_n] = c > 0.$$

Class III (Third-order suboptimal codes). There exists $c > 0$ and a sequence δ_n such that $\sqrt{\frac{\log n}{n}} \leq \delta_n \ll n^{-1/4}$ and

$$\limsup_{n\to\infty} \pi_n[\|\theta - P^*\|_J \geq \delta_n] = c > 0.$$

Class IV (Fourth-order suboptimal codes). There exists $c > 0$ and a sequence δ_n such that $\sqrt{\frac{1}{n}} \ll \delta_n \leq \sqrt{\frac{\log n}{n}}$ and

$$\limsup_{n\to\infty} \pi_n[\|\theta - P^*\|_J \geq \delta_n] = c > 0.$$

Class V (Fourth-order suboptimal codes). $P_{\mathbf{X}} \neq (P_n^*)^n$ and

$$\lim_{c\to\infty} \limsup_{n\to\infty} \pi_n[\|\theta - P_n^*\|_J \geq cn^{-1/2}] = 0$$

where P_n^* is given by (17).

Theorem. Assume P^* is unique. The following upper bounds on $M(n,\epsilon)$ hold.

$$\begin{aligned}
\text{Class I}: \ M(n,\epsilon) &\leq \exp\{n(C - \delta^2/2 + o(\delta)) - O(\sqrt{n})\} \\
\text{Class II}: \ M(n,\epsilon) &\leq \exp\{nC - \frac{1-c}{2}\, n\delta_n^2 - \sqrt{nV}t_\epsilon \\
&\quad + o(\sqrt{n})\} \\
\text{Class III}: \ M(n,\epsilon) &\leq \exp\{nC - \sqrt{nV}t_\epsilon - \frac{1-c}{2}\, n\delta_n^2 \\
&\quad + \frac{1}{2}\ln n + O(1)\} \\
\text{Class IV}: \ M(n,\epsilon) &\leq \exp\{nC - \sqrt{nV}t_\epsilon + \frac{1}{2}\ln n \\
&\quad - \frac{1-c}{2}\, n\delta_n^2 + O(1)\} \\
\text{Class V}: \ M(n,\epsilon) &\leq \exp\{nC - \sqrt{nV}t_\epsilon + \frac{1}{2}\ln n \\
&\quad + A_\epsilon + B\}, \quad B \leq 0.
\end{aligned}$$

Proof: see [8].

ACKNOWLEDGMENT

This work was supported by DARPA under the ITMANET program and was initiated during the author's sabbatical at Chinese University of Hong Kong. The author would like to thank C. Nair for stimulating discussions during this sabbatical stay and T. Riedl for informative discussions about Strassen's paper [5].

REFERENCES

[1] C. E. Shannon, *A Mathemetical Theory of Communication*, 1948.
[2] L. Weiss, "On the Strong Converse of the Coding Theorem for Symmetric Channels Without Memory," *Quarterly of Applied Mathematics*, Vol. 8, No. 3, pp. 209—214, 1960.
[3] J. Wolfowitz, *Coding Theorems of Information Theory*, 1961.
[4] R. L. Dobrushin, "Mathematical Problems in the Shannon Theory of Optimal Coding of Information," *Proc. 4th Berkeley Symp.*, 1961.
[5] V. Strassen, "Asymptotische Abschätzungen in Shannon's Informationstheorie," *Trans. 3rd Prague Conf. Info. Theory*, pp. 689—723, 1962.
[6] Y. Polyanskiy, H. V. Poor and S. Verdú, "Channel Coding Rate in the Finite Blocklength Regime," *IEEE Trans. Information Theory*, Vol. 56, No. 5, pp. 2307—2359, May 2010. Short version in *Proc. ISIT*, Toronto, CA, July 2008.
[7] M. Hayashi, "Information Spectrum Approach to Second-Order Coding Rate in Channel Coding," *IEEE Trans. Information Theory*, Vol. 55, No. 11, pp. 4947—4966, Nov. 2009.
[8] P. Moulin, "The Log-Volume of Optimal Codes for Memoryless Channels, Up to a Few Nats," *preprint*, to be posted on arxiv, Feb. 2012.
[9] D. Blackwell and J. L. Hodges, "The Probability in the Extreme Tail of a Convolution," *Ann. Math. Stat.*, Vol. 30, pp. 1113—1120, 1959.
[10] R. Bahadur and R. Ranga Rao, "On Deviations of the Sample Mean," *Ann. Math. Stat.*, Vol. 31, pp. 1015—1027, 1960.
[11] N. R. Chaganty and J. Sethuraman, "Strong Large Deviation and Local Limit Theorems," *Ann. Prob.* Vol. 21, No. 3, pp. 1671—1690, 1993.
[12] C.-G. Esseen, "Fourier analysis of distribution functions," *Acta Mathematica*, Vol. 77, pp. 1—125, 1945
[13] D. L. Wallace, "Asymptotic Approximations to Distributions," *Ann. Math. Stat.*, Vol. 29, No. 3, pp. 635—654, Sep. 1958.
[14] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. II, Wiley, NY, 1971.
[15] O.E. Barndorff-Nielsen and D. R. Cox, *Asymptotic Techniques for Use in Statistics*, Chapman and Hill, London, UK, 1991.
[16] A. DasGupta, *Asymptotic Theory of Statistics and Probability*, Springer, 2008.
[17] H. Cramér, "Sur un nouveau théorème-limite de la théorie des probabilités," *Actualités Sci. et Industrielles*, No. 736, Hermann, Paris, 1938.
[18] B. Xie and R. Wesel, "A Mutual Information Invariance Approach to Symmetry in Discrete Memoryless Channels," *Proc. ITA*, San Diego, CA, 2008.
[19] I. Csiszár and J. Körner, *Information Theory: Coding Theory for Discrete Memoryless Systems*, Academic Press, NY, 1981.
[20] R. G. Gallager, *Information Theory and Reliable Communication*, Wiley, New York, 1968.