A Tight Upper Bound for the Third-Order Asymptotics of Discrete Memoryless Channels

Marco Tomamichel
Centre for Quantum Technologies,
National University of Singapore
Email: cqtmarco@nus.edu.sg

Vincent Y. F. Tan
Institute for Infocomm Research, A*STAR and
ECE Dept., National University of Singapore
Email: vtan@nus.edu.sg

***Abstract*—This paper shows that the logarithm of the ε-error capacity (average error probability) for n uses of a discrete memoryless channel with positive conditional information variance at every capacity-achieving input distribution is upper bounded by the normal approximation plus a term that does not exceed $\frac{1}{2}\log n + O(1)$.**

I. Introduction

The primary information-theoretic task in channel coding is the characterization of the maximum rate of communication over n independent uses of a noisy channel W. We are concerned in this paper with *discrete memoryless channels* (DMCs), i.e., $W : \mathcal{X} \to \mathcal{Y}$ and $\mathcal{X}$ and $\mathcal{Y}$ are finite. Let $M^*(W^n, \varepsilon)$ denote the maximum size of a length-n block code for W having average error probability no larger than $\varepsilon \in (0, 1)$. Shannon's *noisy-channel coding theorem* [1] and the strong converse [2] state that for every $\varepsilon \in (0, 1)$,

$$\lim_{n\to\infty} \frac{1}{n} \log M^*(W^n, \varepsilon) = C \quad \text{bits/channel use},$$

where $C = \max_P I(P, W)$ is the *channel capacity*. Since the mid-1960s, there has been interest in determining finer asymptotic characterizations of the coding theorem. This is useful because such an analysis provides key insights into the amount of backoff from channel capacity for block codes of finite length n. In particular, Strassen in 1964 [3] showed using normal approximations that, under mild regularity conditions, the asymptotic expansion of $\log M^*(W^n, \varepsilon)$ satisfies

$$\log M^*(W^n, \varepsilon) = nC + \sqrt{nV_\varepsilon}\Phi^{-1}(\varepsilon) + \rho_n, \qquad (1)$$

where $\rho_n = O(\log n)$, V_ε is known as the *ε-channel dispersion* [4], [5] and Φ is the Gaussian distribution function. There have been recent extensions of Strassen's normal approximation, most prominently by Hayashi [6] and Polyanskiy-Poor-Verdú (PPV) [4]. Strassen's normal approximation has been shown to hold for many other classes of channels such as the additive white Gaussian noise channel [4]–[6].

Despite these impressive advances in channel coding, the third-order term ρ_n in (1) is not well understood. Indeed, Hayashi in the conclusion of his paper [6] mentions that

> *"... the third-order coding rate is expected but appears difficult. The second order is the order $\sqrt{n}$, and it is not clear whether the third order is a constant order or the order $\log n$"*

It is known for the binary symmetric channel (BSC) that $\rho_n = \frac{1}{2}\log n + O(1)$ [4, Th. 52] and for the binary erasure channel (BEC), $\rho_n = O(1)$ [4, Th. 53]. More generally, there are classes of channels for which we have bounds on ρ_n [5, Sec. 3.4.5]. For lower bounds (achievability), if we restrict ourselves to DMCs W with positive capacity and all elements of the stochastic matrix W are positive, $\rho_n \geq \frac{1}{2}\log n + O(1)$ [5, Cor. 54]. For upper bounds (converse), if we restrict our attention to so-called *weakly input-symmetric* DMCs [5, Def. 9], $\rho_n \leq \frac{1}{2}\log n + O(1)$ [5, Th. 55]. This symmetric case was also observed by Strassen [3, footnote 1]. It was shown [7] that, under some regularity assumptions, *constant-composition codes* satisfy $\rho_n = \frac{1}{2}\log n + O(1)$. It is also claimed that the same holds for a more general class of DMCs in [8].

This paper strengthens the upper (converse) bound on the third-order term ρ_n. To state our upper bound succinctly, define $\Pi := \{P \in \mathcal{P}(\mathcal{X}) \,|\, I(P, W) = C\}$ to be the set of *capacity-achieving input distributions* (CAIDs). Let $V(P, W)$ be the *conditional information variance* [4, Eqs. (242)-(244)]. If $V(P, W)$ evaluated at every CAID is positive (i.e., $V_{\min} := \min_{P\in\Pi} V(P, W) > 0$), our main result states that

$$\log M^*(W^n, \varepsilon) \leq nC + \sqrt{nV_\varepsilon}\Phi^{-1}(\varepsilon) + \frac{1}{2}\log n + O(1), \quad (2)$$

for every $\varepsilon \in (0, 1)$. Hence, for this rather general class of DMCs, $\rho_n \leq \frac{1}{2}\log n + O(1)$. We may thus dispense with the assumption that W is weakly input-symmetric [5, Def. 9].

The usual way [3]–[6] to prove an upper bound (converse) on $M^*(W^n, \varepsilon)$ is to first prove an upper bound on the maximum number of codewords in a constant-composition code [7] under the *maximum* error probability formulation $M^*_{\max}(W^n, \varepsilon)$. This upper bound can be proved using either the meta-converse [4, Th. 28] or tight bounds on the type-II error probability in a simple binary hypothesis test [3, Th. 1.1]. By the type-counting lemma [9], every length-n block code can be partitioned into no more than $(n + 1)^{|\mathcal{X}|-1}$ constant-composition subcodes. This leads to the rather conservative bound [3, Eq. (4.29)] [5, Eq. (3.259)]

$$\log M^*_{\max}(W^n, \varepsilon) \leq nC + \sqrt{nV_\varepsilon}\Phi^{-1}(\varepsilon) + \left(|\mathcal{X}| - \frac{1}{2}\right)\log n + O(1). \quad (3)$$

Subsequently, using expurgation (see [5, Eq. (3.260)]), we can conclude that the same upper bound holds for $M^*(W^n, \varepsilon)$. We adopt a different approach for the proof of our main result in (2). In a nutshell, we generalize the converse technique in Wang-Colbeck-Renner [10] and Wang-Renner [11], exploit the link [12, Lem. 12] between the *ε-hypothesis testing relative entropy* [13] and the *relative entropy information spectrum* [14, Ch. 4] and carefully weigh the contributions of each input type for a general (non-constant-composition) code by constructing an appropriate ϵ-net for the output probability simplex. The last step, which replaces the use of the type-counting lemma, allows us to bound the effect of different input types with the $O(1)$ term in (2).

II. Notation and Preliminaries

A. Discrete Memoryless Channels

As mentioned in the Introduction, we consider *discrete memoryless channels* (DMCs), which are characterized by two finite sets, the input alphabet $\mathcal{X}$ and the output alphabet $\mathcal{Y}$, and a stochastic matrix W, where $W(y|x)$ denotes the probability that the output $y \in \mathcal{Y}$ occurs given input $x \in \mathcal{X}$. The set of probability distributions on $\mathcal{X}$ is denoted $\mathcal{P}(\mathcal{X})$. For any probability distribution $P \in \mathcal{P}(\mathcal{X})$, we denote by $P \times W : (x, y) \mapsto P(x)W(y|x)$ the joint distribution of inputs and outputs of the channel, and by $PW : y \mapsto \sum_x P(x)W(y|x)$ its marginal on $\mathcal{Y}$. Finally, $W(\cdot|x)$ denotes the distribution on $\mathcal{Y}$ if the input is fixed to x.

Given two probability distributions $P, Q \in \mathcal{P}(\mathcal{X})$, the random variable $\log \frac{P(X)}{Q(X)}$ where X has distribution P the *log-likelihood ratio* of P and Q. Its mean is the *relative entropy*

$$D(P\|Q) := \mathop{\mathrm{E}}_{P}\left[\log \frac{P}{Q}\right] = \sum_{x\in\mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}.$$

The *mutual information* is $I(P, W) := D(P\times W\|P\times PW) = \sum_x P(x)D(W(\cdot|x)\|PW)$. Moreover,

$$C(W) := \max_{P\in\mathcal{P}(\mathcal{X})} I(P, W), \quad \text{and}$$
$$\Pi(W) := \{P \in \mathcal{P}(\mathcal{X}) \,|\, I(P, W) = C(W)\}$$

are the *capacity* and the set of *capacity achieving input distributions* (CAIDs), respectively.[1] The set of CAIDs is convex and compact in $\mathcal{P}(\mathcal{X})$. The unique [15, Cor. 2 to Th. 4.5.1] *capacity achieving output distribution* (CAOD) is denoted as Q^* and $Q^* = PW$ for all $P \in \Pi$. Furthermore, it satisfies $Q^*(y) > 0$ for all $y \in \mathcal{Y}$ [15, Cor. 1 to Th. 4.5.1], where we assume that all outputs are accessible.

The variance of the log-likelihood ratio of P and Q is the *divergence variance*

$$V(P\|Q) := \mathop{\mathrm{E}}_{P}\left[\Big(\log \frac{P}{Q} - D(P\|Q)\Big)^2\right].$$

We also define the *conditional divergence variance* $V(W\|Q|P) := \sum_x P(x)V(W(\cdot|x)\|Q)$ and the *conditional information variance* $V(P, W) := V(W\|PW|P)$. Note that $V(P, W) = V(P\times W\|P\times PW)$ for all $P \in \Pi$ [4, Lem. 62]. The *ε-channel dispersion* [4, Def. 2] is an operational quantity that was shown [4, Eq. (223)] to be equal to

$$V_\varepsilon(W) := \begin{cases} V_{\min} := \min_{P\in\Pi} V(P, W) & \text{if } \varepsilon \le \frac{1}{2} \\ V_{\max} := \max_{P\in\Pi} V(P, W) & \text{if } \varepsilon > \frac{1}{2} \end{cases}. \quad (4)$$

We employ the cumulative distribution function of the standard normal distribution

$$\Phi(a) := \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}x^2\Big)\,\mathrm{d}x$$

and define its inverse as $\Phi^{-1}(\varepsilon) := \sup\{a \in \mathbb{R} \,|\, \Phi(a) \le \varepsilon\}$, which evaluates to the usual inverse for $0 < \varepsilon < 1$ and continuously extended to take values $\pm\infty$ outside that range.

For a sequence $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^{\times n}$, we denote by $P_{\mathbf{x}} \in \mathcal{P}(\mathcal{X})$ the probability distribution given by the relative frequencies of $\mathbf{x}$, i.e. $P_{\mathbf{x}}(x) = \frac{1}{n}\sum_{i=1}^n 1_{\{x_i=x\}}$. This probability distribution $P_{\mathbf{x}}$ is also known as the *empirical distribution* or the *type* [9] of $\mathbf{x}$. The set of all such distributions is denoted as $\mathcal{P}_n(\mathcal{X}) = \bigcup_{\mathbf{x}} \{P_{\mathbf{x}}\}$ and satisfies $|\mathcal{P}_n(\mathcal{X})| \le (n+1)^{|\mathcal{X}|-1}$.

B. Codes and ε-Error Capacity

A *code* $\mathcal{C}$ for a channel is defined by the triple $\{\mathcal{M}, e, d\}$, where $\mathcal{M}$ is a set of messages, $e : \mathcal{M} \to \mathcal{X}$ an encoder and $d : \mathcal{Y} \to \mathcal{M}$ a decoder. We write $|\mathcal{C}| = |\mathcal{M}|$ for the cardinality of the message set. We define the *average error probability* of a code $\mathcal{C}$ for the channel W as

$$p_{\text{err}}(\mathcal{C}, W) := \Pr[M \ne M']$$

where P_M is assumed to be uniform on $\mathcal{M}$,

$$M \xrightarrow{e} X \xrightarrow{W} Y \xrightarrow{d} M'$$

forms a Markov chain, and M' thus denotes output of the decoder. The *one-shot ε-error capacity* of W is defined as

$$M^*(W, \varepsilon) := \max\big\{m \in \mathbb{N} \,\big|\, \exists\, \mathcal{C} :\ |\mathcal{C}| = m \ \wedge\ p_{\text{err}}(\mathcal{C}, W) \le \varepsilon\big\}.$$

We are also interested in the ε-error capacity for $n \ge 1$ uses of a memoryless channel. For this purpose, we consider the channel W^n, defined by the stochastic matrix $W^n(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^n W(y_i|x_i)$, where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ are strings of length n of symbols $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$, respectively. Then, the *blocklength n ε-error capacity* of the channel W is denoted as $M^*(W^n, \varepsilon)$.

III. Main Result and Proof Sketch

Theorem 1. *For every DMC W for which $V_{\min} > 0$, the blocklength n ε-error capacity satisfies*

$$\log M^*(W^n, \varepsilon) \le nC + \sqrt{nV_\varepsilon}\Phi^{-1}(\varepsilon) + \frac{1}{2}\log n + O(1).$$

In light of the existing results on ρ_n (in the Introduction and [5, Sec. 3.4.5]), the third order term is the best possible unless we impose further assumptions on W. In fact, the assumption that $V_{\min} > 0$ can be dispensed with. See [16] for details.

The proof consists of five parts, each detailed in one of the following subsections. In the first subsection, we introduce

[1] We often drop the dependence on W if it is clear from context.

two entropic quantities, the hypothesis testing divergence [10], [11], [13] and a quantity related to the information (or divergence) spectrum [14, Ch. 4]. We state some useful properties we need later. In the second subsection, we present a converse bound, valid for general channels, that involves a minimization over output distributions and maximization over input symbols. In the third subsection, we choose an appropriate output distribution for use in the general converse bound. In the fourth subsection, we state some continuity properties of information measures around the CAIDs and the unique CAOD. Finally, the fifth subsection contains a proof sketch of our main result. For the complete details of the proof and illustrations, please see the full version of this paper on the arXiv repository [16].

A. Hypothesis Testing and the Information Spectrum

We use the following divergence [10]–[13], which is closely related to binary hypothesis testing. Let $\varepsilon \in (0,1)$ and let $P, Q \in \mathcal{P}(\mathcal{Z})$, where $\mathcal{Z}$ is finite. We consider binary (probabilistic) hypothesis tests $\xi : \mathcal{Z} \to [0,1]$ and define the *ε-hypothesis testing divergence*

$$D_h^\varepsilon(P\|Q) := \sup\Big\{R\in\mathbb{R}\,\Big|\,\exists\ \xi:\ \mathop{\mathrm{E}}_{Q}\big[\xi(Z)\big] \le (1-\varepsilon)\exp(-R) \wedge \mathop{\mathrm{E}}_{P}\big[\xi(Z)\big] \ge 1-\varepsilon\Big\}.$$

Note that $D_h^\varepsilon(P\|Q) = -\log\frac{\beta_{1-\varepsilon}(P,Q)}{1-\varepsilon}$ where β_α is defined in PPV [4, Eq. (100)]. It is easy to see that $D_h^\varepsilon(P\|Q) \ge 0$, where the lower bound is achieved if and only if $P = Q$ and $D_h^\varepsilon(P\|Q)$ diverges if P and Q are orthogonal. It satisfies a data-processing inequality [10], i.e., $D_h^\varepsilon(P\|Q) \ge D_h^\varepsilon(PW\|QW)$ for all $W : \mathcal{Z} \to \mathcal{Z}'$. When evaluated for i.i.d. random variables, its asymptotic expansion in the first order is determined by the Chernoff-Stein Lemma [9, Cor. 1.2], yielding $D_h^\varepsilon(P^{\times n}\|Q^{\times n}) = nD(P\|Q)+o(n)$ for any $\varepsilon \in (0,1)$. Strassen [3, Th. 3.1] tightened this analysis and showed that $D_h^\varepsilon(P^{\times n}\|Q^{\times n}) = nD(P\|Q) + \sqrt{nV(P\|Q)}\Phi^{-1}(\varepsilon) + \frac{1}{2}\log n + O(1)$.

The following quantity, which characterizes the distribution of the log-likelihood ratio and is known as the *relative entropy information spectrum* or the *divergence spectrum* [14, Ch. 4], is sometimes easier to manipulate and evaluate.

$$D_s^\varepsilon(P\|Q) := \sup\left\{R\in\mathbb{R}\,\middle|\,\Pr_P\left[\log\frac{P}{Q}\le R\right]\le\varepsilon\right\}.$$

It is intimately related to the ε-hypothesis testing divergence, as the following lemma shows.

Lemma 2. *For any $\delta \in (0, 1-\varepsilon)$, we have*

$$D_h^\varepsilon(P\|Q) \le D_s^{\varepsilon+\delta}(P\|Q) + \log\frac{1-\varepsilon}{\delta}. \tag{5}$$

This relation follows from standard arguments relating binary hypothesis testing and the log-likelihood test to the relative entropy information spectrum. In [12, Lem. 12], an analogue of the above lemma is shown for the non-commutative case.

We can give an upper bound on $D_s^\varepsilon(P\|Q)$ if Q is a convex combination of distributions.

Lemma 3. *Let $P \in \mathcal{P}(\mathcal{Z})$ and $Q = \sum_{i\in\mathcal{I}} q(i)Q^i$ with $Q^i \in \mathcal{P}(\mathcal{Z})$ and $q \in \mathcal{P}(\mathcal{I})$ and $\mathcal{I}$ is some countable index set. Then,*

$$D_s^\varepsilon(P\|Q) \le \inf\big\{D_s^\varepsilon(P\|Q^i) - \log q(i)\big\}_{i\in\mathcal{I}}$$

The following property bounds the log-likelihood ratio of the input-output behavior of two channels in terms of the log-likelihood ratio evaluated for a single input symbol.

Lemma 4. *Let $P \in \mathcal{P}(\mathcal{X})$ and let $V, W : \mathcal{X} \to \mathcal{Y}$. Then,*

$$D_s^\varepsilon(P\times W\|P\times V) \le \sup_{x:\,P(x)>0} D_s^\varepsilon(W(\cdot|x)\|V(\cdot|x)).$$

The distribution of the log-likelihood ratio has the following asymptotic expansion.

Lemma 5. *Let $P_i, Q \in \mathcal{P}(\mathcal{Z})$ be such that $P_i \ll Q$ for all i in some set $\mathcal{I}$. Consider a sequence of distributions P_{i_k} indexed by $(i_1,\dots,i_n)$ where $i_k \in \mathcal{I}$ for each $1 \le k \le n$. Define*

$$D_n := \frac{1}{n}\sum_{k=1}^n D(P_{i_k}\|Q), \quad V_n := \frac{1}{n}\sum_{k=1}^n V(P_{i_k}\|Q), \quad \text{and}$$
$$T_n := \frac{1}{n}\sum_{k=1}^n \mathop{\mathrm{E}}_{P_{i_k}}\left[\left|\log\frac{P_{i_k}}{Q} - D(P_{i_k}\|Q)\right|^3\right].$$

If $V_n \ge V_- > 0$, then we have

$$D_s^\varepsilon\big(P_{i_1}\times\dots P_{i_n}\big\|Q^{\times n}\big) \le nD_n + \sqrt{nV_n}\Phi^{-1}\left(\varepsilon + \frac{6\,T_n}{\sqrt{nV_-^3}}\right).$$

In any case, we have

$$D_s^\varepsilon\big(P_{i_1}\times\dots P_{i_n}\big\|Q^{\times n}\big) \le nD_n + \sqrt{\frac{nV_n}{1-\varepsilon}}. \tag{6}$$

B. A Converse Bound for General Channels

Here, we give a new converse bound on the size of arbitrary codes for general channels, for average probability of error.

Proposition 6. *Let $\varepsilon \in (0,1)$ and let W be any channel. Then, for any $\delta \in (0, 1-\varepsilon)$, we have*

$$\log M^*(W,\varepsilon) \le \inf_{Q\in\mathcal{P}(\mathcal{Y})}\ \max_{x\in\mathcal{X}} D_s^{\varepsilon+\delta}\big(W(\cdot|x)\big\|Q\big) + \log\frac{1}{\delta}.$$

The first part of the proof of Prop. 6 is equivalent to the meta-converse in [4, Th. 27] (see also [10] and [11]). Via relaxation to the divergence spectrum (Lemma 2), we then replace the maximization over input distributions with a maximization over input symbols (Lemma 4), which yields a result in the spirit of [4, Th. 28 and Th. 31] but our result applies to *average* error probability. The maximization over symbols allows us to apply our converse bound on non-constant composition codes directly. See [16].

C. A Suitable Choice of Output Distribution Q

For n-fold repetitions of a DMC, the bound in Proposition 6 evaluates to

$$\log M^*(W^n,\varepsilon) \le \min_{Q^{(n)}\in\mathcal{P}(\mathcal{Y}^{\times n})}\ \max_{\mathbf{x}\in\mathcal{X}^{\times n}} D_s^{\varepsilon+\delta}\big(W^n(\cdot|\mathbf{x})\big\|Q^{(n)}\big) + \log\frac{1}{\delta}, \tag{7}$$

and it thus important to find a suitable choice of $Q^{(n)} \in \mathcal{P}(\mathcal{Y}^{\times n})$ to further upper bound the above. Symmetry considerations allow us to restrict the search to distributions that are invariant under permutations of the n channel uses. Let $\zeta := |\mathcal{Y}|(|\mathcal{Y}|-1)$ and let $\gamma > 0$ be a constant which is to be chosen later. Consider the following convex combination of product distributions:

$$Q^{(n)}(\mathbf{y}) := \frac{1}{2}\sum_{\mathbf{k}\in\mathcal{K}} \frac{\exp\big(-\gamma\|\mathbf{k}\|_2^2\big)}{F} \prod_{i=1}^{n} Q_{\mathbf{k}}(y_i) + \frac{1}{2}\sum_{P_{\mathbf{x}}\in\mathcal{P}_n(\mathcal{X})} \frac{1}{|\mathcal{P}_n(\mathcal{X})|}\prod_{i=1}^{n} P_{\mathbf{x}}W(y_i), \quad (8)$$

where $\mathbf{y} := (y_1, y_2, \ldots, y_n)$ and

$$Q_{\mathbf{k}}(y) := Q^*(y) + \frac{k_y}{\sqrt{n\zeta}},$$
$$\mathcal{K} := \Big\{\mathbf{k}\in\mathbb{Z}^{|\mathcal{Y}|}\,\Big|\,\sum_y k_y = 0 \wedge k_y \geq -Q^*(y)\sqrt{n\zeta}\Big\}.$$

In (8), F is a normalization constant that ensures $\sum_{\mathbf{y}} Q^{(n)}(\mathbf{y}) = 1$. What we have done in our choice of $Q_{\mathbf{k}}$ is to uniformly quantize the simplex $\mathcal{P}(\mathcal{Y})$ along axis-parallel directions. The constraint that each $\mathbf{k}\in\mathcal{K}$ ensures that each $Q_{\mathbf{k}}$ is a valid probability mass function. We find that

$$F \leq \sum_{\mathbf{k}\in\mathbb{Z}^{|\mathcal{Y}|}} \exp\big(-\gamma\|\mathbf{k}\|_2^2\big) \leq \left(1+\sqrt{\frac{\pi}{\gamma}}\right)^{|\mathcal{Y}|}$$

is a finite constant. Furthermore, by construction, the representation points $\{Q_{\mathbf{k}}\}_{\mathbf{k}}$ form an ϵ-*net* with $\epsilon = n^{-\frac{1}{2}}$ for $\mathcal{P}(\mathcal{Y})$. Namely, for every $Q\in\mathcal{P}(\mathcal{Y})$, there exists a $\mathbf{k}\in\mathcal{K}$ such that $\|Q - Q_{\mathbf{k}}\|_2 \leq n^{-\frac{1}{2}}$.

D. Continuity around the CAIDs and the unique CAOD

We will often be concerned with probability distributions close to the set of CAIDs Π in Euclidean distance, i.e., those distributions belonging to

$$\Pi_\mu := \Big\{P\in\mathcal{P}(\mathcal{X})\Big|\min_{P^*\in\Pi}\|P-P^*\|_2 \leq \mu\Big\}$$

for some small $\mu > 0$. The image of this set under W is denoted as $\Pi_\mu W$. We also consider a larger, "η-blown-up" version, of $\Pi_\mu W$, namely

$$\Gamma_\mu^\eta := \Big\{Q\in\mathcal{P}(\mathcal{Y})\Big|\,\exists\, P\in\Pi_\mu \text{ s.t. } \|PW-Q\|_2\leq\eta\Big\}.$$

Note that $\Gamma_\mu^0 = \Pi_\mu W$ if the matrix W has full rank. The following Lemma summarizes known results about these sets.

Lemma 7. *Let W be a DMC such that $V_{\min} > 0$. Then there exists $\mu > 0$ and $\eta > 0$, as well as finite constants $q_{\min} > 0$, $\alpha > 0$ and $\beta > 0$ such that the following holds. For all $P\in\Pi_\mu$, their projections $P^* = \arg\min_{P'\in\Pi}\|P-P'\|_2$ and for all $Q\in\Gamma_\mu^\eta$ we have*

1) $Q(y) > q_{\min}$ *for all* $y\in\mathcal{Y}$,
2) $V(W\|Q|P) > \frac{V_{\min}}{2} > 0$,
3) $I(P,W) \leq C(W) - \alpha\|P-P^*\|_2^2$,
4) $D(P\times W\|P\times Q) \leq I(P,W) + \frac{\|Q-PW\|_2^2}{q_{\min}}$,
5) $\big|\sqrt{V(P,W)} - \sqrt{V(P^*,W)}\,\big| \leq \beta\|P-P^*\|_2$,
6) $\big|\sqrt{V(W\|Q|P)} - \sqrt{V(P,W)}\,\big| \leq \beta\|Q-PW\|_2$.

E. Asymptotics for DMCs

We are now ready to prove our main result.

Proof of Theorem 1: Firstly, we employ Proposition 6 to provide a bound on $\log M^*(W^n,\varepsilon)$. We choose $\delta = n^{-\frac{1}{2}}$, which satisfies $0<\delta<1-\varepsilon$ for sufficiently large n. Substitute the output distribution $Q^{(n)}$ in (8) to get

$$\log M^*(W^n,\varepsilon) \leq \max_{\mathbf{x}\in\mathcal{X}^{\times n}} \underbrace{D_s^{\varepsilon+\delta}\big(W^n(\cdot|\mathbf{x})\big\|Q^{(n)}\big)}_{=:\ \mathrm{cv}(\mathbf{x})} + \frac{1}{2}\log n.$$

It remains to show that each term $\mathrm{cv}(\mathbf{x})$ in the maximization is upper bounded by $nC + \sqrt{nV_\varepsilon}\Phi^{-1}(\varepsilon) + G$ for a suitable constant G for all sufficiently large n.

We apply Lemma 7 that supplies us with constants μ, η, $q_{\min}$, α and β and distinguish between two cases for the following; either a) $\mathbf{x}$ satisfies $P_{\mathbf{x}}\notin\Pi_\mu$ or b) $\mathbf{x}$ satisfies $P_{\mathbf{x}}\in\Pi_\mu$. This strategy in which we partition input types into two classes was proposed by Strassen [3, Sec. 4]. See also PPV [4, Appendix I].

Case a): $P_{\mathbf{x}}\notin\Pi_\mu$: The mutual information outside Π_μ is bounded away from the capacity, i.e., $I(P_{\mathbf{x}},W)\leq C' < C$ for all $P_{\mathbf{x}}\notin\Pi_\mu$.

We first apply Lemma 3 and then Lemma 5 to bound

$$\mathrm{cv}(\mathbf{x}) \leq D_s^{\varepsilon+\delta}\big(W^n(\cdot|\mathbf{x})\big\|(P_{\mathbf{x}}W)^{\times n}\big) + \log\big(2\,|\mathcal{P}_n(\mathcal{X})|\big) \leq nI(P_{\mathbf{x}},W) + \sqrt{\frac{nV(P_{\mathbf{x}},W)}{1-\varepsilon-\delta}} + \log\big(2\,|\mathcal{P}_n(\mathcal{X})|\big).$$

Invoking [4, Lem. 62] and [14, Rmk. 3.1.1] yields the uniform bound $V(P_{\mathbf{x}},W)\leq\frac{8\log^2 e}{e^2}|\mathcal{Y}|\leq 2.3\,|\mathcal{Y}|$. Hence,

$$\mathrm{cv}(\mathbf{x}) \leq nC' + \sqrt{n}\sqrt{\frac{2.3\,|\mathcal{Y}|}{1-\varepsilon-\delta}} + \big(|\mathcal{X}|-1\big)\log\big(n+1\big) + \log 2.$$

Since $C' < C$, the linear term dominates the term growing with the square root of n and the term growing logarithmically in n asymptotically. Hence, it is evident that $\mathrm{cv}(\mathbf{x}) \leq nC + \sqrt{nV_\varepsilon}\Phi^{-1}(\varepsilon)$ for sufficiently large n.

Case b): $P_{\mathbf{x}}\in\Pi_\mu$: Before we commence, define the *third absolute moment of the log-likelihood ratio* between P and Q to be $T(P\|Q) := \mathrm{E}_P\big[\big|\log\frac{P}{Q} - D(P\|Q)\big|^3\big]$. Also define

$$V_+ := \max_{P\in\mathcal{P}(\mathcal{X})}\max_{Q\in\Gamma_\mu^\eta} V(W\|Q|P), \quad \text{and}$$
$$T_+ := \max_{P\in\mathcal{P}(\mathcal{X})}\max_{Q\in\Gamma_\mu^\eta} T(W\|Q|P),$$

where $T(W\|Q|P) := \sum_x P(x)T(W(\cdot|x)||Q)$. Note that $0 < V_+ < \infty$ and $T_+ < \infty$ by Lemma 7.

For each $\mathbf{x}$, we denote by $Q_{\mathbf{k}(\mathbf{x})}$ the element of the ϵ-net (constructed in Section III-C) closest to $P_{\mathbf{x}}W$. We note that since $\|Q_{\mathbf{k}(\mathbf{x})} - P_{\mathbf{x}}W\|_2 \leq \epsilon = n^{-\frac{1}{2}}$, we have $Q_{\mathbf{k}(\mathbf{x})}\in\Gamma_\mu^\eta$ for sufficiently large n, which enables us to apply the properties described in Lemma 7 extensively below.

We first use Lemma 3 to bound

$$\mathrm{cv}(\mathbf{x}) \le D_s^{\varepsilon+\delta}\big(W^n(\cdot|\mathbf{x})\big\|(Q_{\mathbf{k}(\mathbf{x})})^{\times n}\big) + \gamma\|\mathbf{k}(\mathbf{x})\|_2^2 + \log\big(2F\big).$$

We now employ Lemma 5, where we choose $P_i = W(\cdot|x_i)$ and may set $V_- = \frac{V_{\min}}{2}$ due to Lemma 7. The bound on the third moment, T_n in Lemma 5, still depends on $Q_{\mathbf{k}(\mathbf{x})}$. However, we can upper-bound $T(Q_{\mathbf{k}(\mathbf{x})})$ by T_+ which is finite. We then introduce the finite constant $B := 1 + 6T_+/V_-^{3/2}$, while substituting for $\delta = n^{-\frac{1}{2}}$, to get

$$\mathrm{cv}(\mathbf{x}) \le nD(P_{\mathbf{x}}\times W\|P_{\mathbf{x}}\times Q_{\mathbf{k}(\mathbf{x})}) + \sqrt{nV(W\|Q_{\mathbf{k}(\mathbf{x})}|P_{\mathbf{x}})}\,\Phi^{-1}\left(\varepsilon+\frac{B}{\sqrt{n}}\right)+\gamma\|\mathbf{k}(\mathbf{x})\|_2^2+\log\big(2F\big).$$

We now require that $n \ge N$, where N is chosen large enough such that $\varepsilon + \frac{B}{\sqrt{N}} < 1$. This ensures that the coefficient of the term growing as $\sqrt{n}$ in the above expression is finite. Next, we use the fact that Φ^{-1} is infinitely differentiable and $V(W\|Q_{\mathbf{k}(\mathbf{x})}|P_{\mathbf{x}}) \le V_+$ is finite to bound

$$\sqrt{nV(W\|Q_{\mathbf{k}(\mathbf{x})}|P_{\mathbf{x}})}\,\Phi^{-1}\left(\varepsilon+\frac{B}{\sqrt{n}}\right) \le \sqrt{nV(W\|Q_{\mathbf{k}(\mathbf{x})}|P_{\mathbf{x}})}\,\Phi^{-1}(\varepsilon) + G_1.$$

for some finite constant G_1 and all $n \ge N$. Thus, defining $G_2 := G_1 + \log(2F)$, we get

$$\mathrm{cv}(\mathbf{x}) \le nD(P_{\mathbf{x}}\times W\|P_{\mathbf{x}}\times Q_{\mathbf{k}(\mathbf{x})}) + \sqrt{nV(W\|Q_{\mathbf{k}(\mathbf{x})}|P_{\mathbf{x}})}\,\Phi^{-1}(\varepsilon) + \gamma\|\mathbf{k}(\mathbf{x})\|_2^2 + G_2,$$

Next, we would like to replace $Q_{\mathbf{k}(\mathbf{x})}$ with $P_{\mathbf{x}}W$ in the above bound. This can be done without too much loss due to Lemma 7 and $\|Q_{\mathbf{k}(\mathbf{x})} - P_{\mathbf{x}}W\|_2 \le n^{-\frac{1}{2}}$. We have,

$$D(P_{\mathbf{x}}\times W\|P_{\mathbf{x}}\times Q_{\mathbf{k}(\mathbf{x})}) - I(P_{\mathbf{x}},W) \le \frac{1}{n\,q_{\min}}, \quad \text{and}$$
$$\Big|\sqrt{V(W\|Q_{\mathbf{k}(\mathbf{x})}|P_{\mathbf{x}})} - \sqrt{V(P_{\mathbf{x}},W)}\Big| \le \frac{\beta}{\sqrt{n}}.$$

Hence, choosing $G_3 := \frac{1}{q_{\min}} + \beta\big|\Phi^{-1}(\varepsilon)\big| + G_2$, we find that

$$\mathrm{cv}(\mathbf{x}) \le nI(P_{\mathbf{x}},W) + \sqrt{nV(P_{\mathbf{x}},W)}\,\Phi^{-1}(\varepsilon) + \gamma\|\mathbf{k}(\mathbf{x})\|_2^2 + G_3.$$

In the following, we use the fact that all distributions $P_{\mathbf{x}}$ in Π_μ satisfy $I(P_{\mathbf{x}},W) \le C - \alpha\xi^2$ and $|\sqrt{V(P_{\mathbf{x}},W)} - \sqrt{V(P^*,W)}| \le \beta\xi$, where $P^* := \arg\min_{P'\in\Pi}\|P_{\mathbf{x}} - P'\|_2$ (which is unique) and $\xi := \|P_{\mathbf{x}} - P^*\|_2$. Hence,

$$\mathrm{cv}(\mathbf{x}) \le nC + \sqrt{nV(P^*,W)}\Phi^{-1}(\varepsilon) + \Big(-\alpha\xi^2 n + \beta|\Phi^{-1}(\varepsilon)|\xi\sqrt{n} + \gamma\|\mathbf{k}(\mathbf{x})\|_2^2\Big) + G_3. \quad (9)$$

It thus remains to show that the term in the bracket is upper bounded by a constant, for an appropriate choice of γ. Let $\|W\|_2$ be the spectral norm of the matrix W. From the construction of the ϵ-net in Section III-C,

$$\|\mathbf{k}(\mathbf{x})\|_2 = \sqrt{n\zeta}\,\|Q_{\mathbf{k}(\mathbf{x})} - Q^*\|_2 \le \sqrt{n\zeta}\Big(\|Q_{\mathbf{k}(\mathbf{x})} - P_{\mathbf{x}}W\|_2 + \|P_{\mathbf{x}}W - Q^*\|_2\Big) \le \sqrt{n\zeta}\left(\frac{1}{\sqrt{n}} + \|W\|_2\,\xi\right).$$

Substituting this bound into (9), we find that the term in the bracket evaluates to

$$\big(\gamma\zeta\|W\|_2^2 - \alpha\big)\xi^2 n + \big(\beta|\Phi^{-1}(\varepsilon)| + 2\gamma\zeta\|W\|_2\big)\xi\sqrt{n} + \gamma\zeta.$$

The expression is a quadratic in $\xi\sqrt{n}$ and has a finite maximum if we choose γ such that $\gamma\zeta\|W\|_2^2 < \alpha$. Hence,

$$\mathrm{cv}(\mathbf{x}) \le nC + \sqrt{nV(P^*,W)}\Phi^{-1}(\varepsilon) + G_4$$

for an appropriate constant G_4 and $n \ge N$. Maximizing over all $P^* \in \Pi$ completes the proof of Theorem 1. ∎

IV. Concluding Remarks and Extensions

Our general converse bound in Prop. 6 can be specialized to channels with cost constraints. As such, it can be applied to the AWGN channel with maximal power constraints and the evaluation of Prop. 6 using a product output distribution yields the $\frac{1}{2}\log n + O(1)$ upper bound on the third-order term [4, Th. 54]. It would be interesting to check if the evaluation of Prop. 6 yields the same upper bound for the finite-dimensional infinite constellations problem [17, Th. 13].

Acknowledgements: MT thanks Ligong Wang for helpful explanations. VYFT thanks Yanina Shkel for insightful discussions and Pierre Moulin for sharing his ITA paper [8]. MT is supported by the National Research Foundation and the Ministry of Education of Singapore. VYFT would like to acknowledge funding support from A*STAR, Singapore.

References

[1] C. E. Shannon. A mathematical theory of communication. *Bell System Tech. Journal*, 27:379–423, 1948.
[2] J. Wolfowitz. *Coding Theorems of Information Theory*. Springer-Verlag, New York, 3rd edition, 1978.
[3] V. Strassen. Asymptotische Abschätzungen in Shannons Informationstheorie. In *Trans. Third Prague Conf. Inf. Theory*, pages 689–723, Prague, 1962.
[4] Y. Polyanskiy, H. V. Poor, and S. Verdú. Channel coding in the finite blocklength regime. *IEEE Trans. on Inf. Th.*, 56:2307–59, May 2010.
[5] Y. Polyanskiy. *Channel coding: Non-asymptotic fundamental limits*. PhD thesis, Princeton University, 2010.
[6] M. Hayashi. Information spectrum approach to second-order coding rate in channel coding. *IEEE Trans. on Inf. Th.*, 55:4947–66, Nov 2009.
[7] P. Moulin. The log-volume of optimal constant-composition codes for memoryless channels, within O(1) bits. In *Int. Symp. Inf. Th.*, Cambridge, MA, 2012.
[8] P. Moulin. The log-volume of optimal codes for memoryless channels, up to a few nats. In *Info. Th. and Appl. Workshop*, San Diego, 2012.
[9] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011.
[10] L. Wang, R. Colbeck, and R. Renner. Simple channel coding bounds. In *Intl. Symp. Inf. Th.*, Seoul, South Korea, 2009.
[11] L. Wang and R. Renner. One-shot classical-quantum capacity and hypothesis testing. *Physical Review Letters*, 108:200501, May 2012.
[12] M. Tomamichel and M. Hayashi. A hierarchy of information quantities for finite block length analysis of quantum tasks. `arXiv:1208.1478 [quant-ph]`, Sep 2012.
[13] F. Dupuis, L. Krämer, P. Faist, J. M. Renes, and R. Renner. Generalized entropies. `arXiv:1211.3141 [quant-ph]`, Nov 2012.
[14] T. S. Han. *Information-Spectrum Methods in Information Theory*. Springer Berlin Heidelberg, Feb 2003.
[15] R. G. Gallager. *Information Theory and Reliable Communication*. Wiley, New York, 1968.
[16] M. Tomamichel and V. Y. F. Tan. A tight upper bound for the third-order asymptotics of discrete memoryless channels. `arXiv:1212.3689 [cs.IT]`, Nov 2012.
[17] A. Ingber, R. Zamir, and M. Feder. Finite-dimensional infinite constellations. *IEEE Trans. on Inf. Th.*, 59(3):1630–56, 2013.