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

Marco Tomamichel\(^1,\ast\) and Vincent Y. F. Tan\(^2,3,\dagger\)

\(^1\)Centre for Quantum Technologies, National University of Singapore
\(^2\)Institute for Infocomm Research (I\(^2\)R), A\(^*\)STAR, Singapore
\(^3\)Department of Electrical and Computer Engineering, National University of Singapore

This paper shows that the \(\varepsilon\)-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 point-to-point 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 DMC \(W\) having average error probability no larger than \(\varepsilon \in (0, 1)\). Shannon’s noisy-channel coding theorem [1] and the corresponding 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

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

where \(\rho_n = O(\log n)\), \(V_\varepsilon\) is known as the \(\varepsilon\)-channel dispersion [4, 5] and \(\Phi\) is the Gaussian cumulative distribution function. These quantities will be defined precisely in the following section. The asymptotic expansion implies that if an error probability of \(\varepsilon\) is tolerable, the backoff from channel capacity \(C\) at finite blocklength \(n\) is roughly \(\sqrt{nV_\varepsilon} \Phi^{-1}(\varepsilon)\). There have been recent refinements to and extensions of Strassen’s normal approximation in (1), most prominently by Hayashi [6] and Polyanskiy-Poor-Verdú (PPV) [4]. Strassen’s normal approximation has also 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 the fundamental limits of channel coding, the third-order term \(\rho_n\) 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 \(\log n\)”

\(*\) cqtmarco@nus.edu.sg
\(\dagger\) vtan@nus.edu.sg
What we do know is that for the binary symmetric channel (BSC), \( \rho_n = \frac{1}{2} \log n + O(1) \) [4, Thm. 52] and for the binary erasure channel (BEC), \( \rho_n = O(1) \) [4, Thm. 53]. More generally, there are classes of channels for which we have bounds on \( \rho_n \) [5, Sec. 3.4.5]. For lower bounds (achievability), if we restrict ourselves to DMCs 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, Thm. 55]. It was shown [7] using strong large-deviation techniques that, under some regularity assumptions, constant-composition codes satisfy \( \rho_n = \frac{1}{2} \log n + O(1) \). Recall that a constant-composition code is one where all the codewords are of the same empirical distribution or type [8].

This paper strengthens the upper (converse) bound on the third-order term \( \rho_n \). To state our upper bound succinctly, define \( \Pi := \{ P \in \mathcal{P}(\mathcal{X}) \mid I(P, W) = C \} \) 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

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

for every \( \varepsilon \in (0, 1) \). Hence, for this rather general class of DMCs, we prove that the third-order term \( \rho_n \leq \frac{1}{2} \log n + O(1) \). We may thus dispense with the assumption that \( W \) is weakly input-symmetric [5, Def. 9] or the code is a constant-composition one [7]. Note that if \( \Pi \) is a singleton (i.e., there exists a unique CAID), the set of DMCs that satisfies our converse bound in (2) coincides with the set of DMCs with positive channel dispersion [4, Def. 1].

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]. This upper bound can be proved using either the meta-converse [4, Thm. 28] or tight bounds on the type-II error probability in a simple binary hypothesis test [3, Thm. 1.1]. By the type-counting lemma [8], 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)]

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

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 [9] and Wang-Renner [10], exploit the link [11, Lem. 12] between the \( \varepsilon \)-hypothesis testing relative entropy [12] and the relative entropy information spectrum [13, Ch. 4] and carefully weigh the contributions of each input type for a general (non-constant-composition) code by constructing an appropriate \( \varepsilon \)-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).

Note that unlike in (3), the third-order term in our upper bound in (2) is independent of the cardinality of the input alphabet \( |\mathcal{X}| \). This makes intuitive sense upon doing the following thought experiment. Let \( n \) be a large even integer and consider using transmitting information across \( n \) uses of a DMC \( W: \mathcal{X} \to \mathcal{Y} \). Clearly, the same amount of information can be transmitted through \( n/2 \) uses of the product channel \( W^2: \mathcal{X}^2 \to \mathcal{Y}^2 \), where \( W^2(y, y', x, x') := W(y | x)W(y' | x') \). The capacity and the dispersion of \( W^2 \) are respectively twice the capacity and the dispersion of \( W \) so the normal approximation terms for \( n \) uses of \( W \) and \( n/2 \) uses of \( W^2 \) are identical. If the coefficient of the third-order logarithmic term were dependent on the size of the input alphabet, say via some function \( g(|\mathcal{X}|) \), then for the first experiment, \( \rho_n = g(|\mathcal{X}|) \log n + O(1) \) while for the second experiment, \( \rho_n = g(|\mathcal{X}|^2) \log(n/2) + O(1) = g(|\mathcal{X}|^2) \log n + O(1) \). Thus, at an intuitive level, we expect that \( g(|\mathcal{X}|) \) is independent of \( |\mathcal{X}| \).
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 $X$ and the output alphabet $Y$, and a stochastic matrix $W$, where $W(y|x)$ denotes the probability that the output $y \in Y$ occurs given input $x \in X$. The set of probability distributions on $X$ is denoted $P(X)$. For any probability distribution $P \in P(X)$, 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 $PW : y \mapsto \sum_x P(x)W(x|y)$ its marginal on $Y$. Finally, $W(\cdot|x)$ denotes the distribution on $Y$ if the input is fixed to $x$.

Given two probability distributions $P, Q \in P(X)$, we call 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

$$D(P\|Q) := E_P \left[ \log \frac{P}{Q} \right] = \sum_x P(x) \log \frac{P(x)}{Q(x)}.$$ 

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

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

are the capacity and the set of capacity achieving input distributions (CAIDs), respectively. The set of CAIDs is convex and compact in $P(X)$. The unique [14, Cor. 2 to Thm. 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 $y \in Y$ [14, Cor. 1 to Thm. 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

$$V(P\|Q) := E_P \left[ \left( \log \frac{P}{Q} - D(P\|Q) \right)^2 \right].$$

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 $\varepsilon$-channel dispersion [4, Def. 2] is an operational quantity that was shown [4, Eq. (223)] to be equal to

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

We employ the cumulative distribution function of the standard normal distribution

$$\Phi(a) := \int_{-\infty}^a \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2} x^2 \right) \, dx$$

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

For a sequence $x = (x_1, x_2, \ldots, x_n) \in X^n$, we denote by $P_x \in P(X)$ the probability distribution given by the relative frequencies of $x$, i.e. $P_x(x) = \frac{1}{n} \sum_{i=1}^n 1\{x_i = x\}$. This probability distribution $P_x$ is also known as the empirical distribution or the type [8] of $x$. The set of all such distributions is denoted as $P_n(X) = \bigcup_x \{P_x\}$ and satisfies $|P_n(X)| \leq (n+1)^{|X|^{-1}}$.

---

1 We often drop the dependence on $W$ if it is clear from context.
B. Codes and ε-Error Capacity

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

\[
p_{\text{err}}(C, W) := \Pr[M \neq M'] = 1 - \frac{1}{|M|} \sum_{m \in M} W(d^{-1}(m)|e(m))
\]

where the distribution over messages \( P_M \) is assumed to be uniform on \( M \), forming a Markov chain, and \( M' \) thus denotes output of the decoder. The one-shot \( \varepsilon \)-error capacity of the channel \( W \) is then defined as

\[
M^*(W, \varepsilon) := \max \{ m \in \mathbb{N} \mid \exists C : |C| = m \land p_{\text{err}}(C, W) \leq \varepsilon \}.
\]

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

III. MAIN RESULT AND PROOF

Let us reiterate our main result.

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

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

In light of the existing results on \( \rho_n \) (in the Introduction and [5, Sec. 3.4.5]), the third order term is the best possible unless we impose further assumptions on \( W \).

The proof consists of five parts, each detailed in one of the following subsections. In the first subsection, we introduce two entropic quantities, the hypothesis testing divergence [9, 10, 12] and a quantity related to the information (or divergence) spectrum [13, Ch. 4]. We state and prove some useful properties we need later. In the second subsection, we derive a converse bound, valid for general DMCs, 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 and prove some continuity properties of information measures around the CAIDs and the unique CAOD. Finally, the fifth subsection contains the proof of our main result.

A. Hypothesis Testing and the Information Spectrum

We use the following divergence [9, 10, 12], 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 \( \varepsilon \)-hypothesis testing divergence

\[
D_\varepsilon^h(P\|Q) := \sup \left\{ R \in \mathbb{R} \mid \exists \xi : E_Q[\xi(\mathcal{Z})] \leq (1 - \varepsilon) \exp(-R) \land E_P[\xi(\mathcal{Z})] \geq 1 - \varepsilon \right\}.
\]
Note that $D^\xi_h(P\|Q) = -\log \frac{\beta_{\alpha}}{1-\varepsilon} (P,Q)$ where $\beta_{\alpha}$ is defined in PPV [4, Eq. (100)]. It is easy to see that $D^\xi_h(P\|Q) \geq 0$, where the lower bound is achieved if and only if $P = Q$ and $D^\xi_h(P\|Q)$ diverges if $P$ and $Q$ are orthogonal. It satisfies a data-processing inequality [9]

$$D^\xi_h(P\|Q) \geq D^\xi_h(PW\|QW) \quad \text{for all channels } W \text{ from } \mathcal{Z} \text{ to } \mathcal{Z}'.$$

When evaluated for independent and identical distributions (i.i.d.), its asymptotic expansion in the first order is determined by the Chernoff-Stein Lemma [8, Cor. 1.2], yielding $D^\xi_h(P^{\times n}\|Q^{\times n}) = nD(P\|Q) + o(n)$ for any $\varepsilon \in (0, 1)$. This analysis was tightened by Strassen [3, Thm. 3.1] and he showed that

$$D^\xi_h(P^{\times n}\|Q^{\times n}) = nD(P\|Q) + \sqrt{nV(P\|Q)}\Phi^{-1}(\varepsilon) + \frac{1}{2} \log n + O(1).$$

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 [13, Ch. 4], is sometimes easier to manipulate and evaluate.

$$D^\xi_s(P\|Q) := \sup \left\{ R \in \mathbb{R} \mid \Pr_P \left[ \log \frac{P}{Q} \leq R \right] \leq \varepsilon \right\}.$$

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

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

$$D^\xi_s(P\|Q) - \log \frac{1}{1-\varepsilon} \leq D^\xi_h(P\|Q) \leq D^\xi_s(P\|Q) + \log \frac{1-\varepsilon}{\delta}. \quad (4)$$

These relations follow from standard arguments relating binary hypothesis testing and the log-likelihood test to the relative entropy information spectrum. In [11, Lem. 12], an analogue of the above lemma is shown for the strictly more general non-commutative case. For completeness we show the second inequality, which we will employ later.

**Proof of Second Inequality in (4).** If $D^\xi_h(P\|Q)$ is infinite, $P$ is not absolutely continuous with respect to $Q$ and it is easy to see that $D^{\xi+\delta}_h(P\|Q)$ is also infinite. Hence, the second inequality in (4) trivially holds. We thus consider the case where $D^\xi_h(P\|Q)$ is finite, and fix any optimal test $\xi$ for $D^\xi_h(P\|Q)$. Set $R^* := D^\xi_h(P\|Q) + \log \frac{\delta}{1-\varepsilon}$. We find

$$\Pr_P \left[ \log \frac{P}{Q} > R^* \right] = \sum_{z \in \mathcal{Z}} P(z) 1_{\{P(z) > \exp(R^*)Q(z)\}}$$

$$\geq \sum_{z \in \mathcal{Z}} (P(z) - \exp(R^*)Q(z)) 1_{\{P(z) > \exp(R^*)Q(z)\}}$$

$$\geq \sum_{z \in \mathcal{Z}} (P(z) - \exp(R^*)Q(z)) \xi(z)$$

$$= E_{P}[\xi(Z)] - \exp(R^*) E_{Q}[\xi(Z)]$$

$$\geq 1 - \varepsilon - \delta.$$

In the last step we used the fact that $\xi$ is an optimal test, which implies that $E_P[\xi(Z)] \geq 1 - \varepsilon$ and $E_Q[\xi(Z)] \leq (1 - \varepsilon) \exp(-D^\xi_h(P\|Q))$. Thus, $D^\xi_{\alpha+\delta}(P\|Q) \geq R^*$, concluding the proof. \qed

We can give an upper bound on $D^\xi_s(P\|Q)$ if $Q$ is a convex combination of distributions.
Lemma 3. Let \( P \in \mathcal{P}(Z) \) and \( Q = \sum_{i \in \mathcal{I}} q(i)Q^i \) with \( Q^i \in \mathcal{P}(Z) \) and \( q \in \mathcal{P}(\mathcal{I}) \) and \( \mathcal{I} \) is some countable index set. Then,

\[
D_s^\epsilon(P\|Q) \leq \inf_{i \in \mathcal{I}} \{ D_s^\epsilon(P\|Q^i) - \log q(i) \}
\]

Proof. Note that for all \( z \in Z \), for all \( i \in \mathcal{I} \), we have

\[
\log \frac{P(z)}{Q(z)} = \log \frac{P(z)}{\sum_j q(j)Q^j(z)} \leq \log \frac{P(z)}{q(i)Q^i(z)} = \log \frac{P(z)}{Q^i(z)} - \log q(i).
\]

Hence,

\[
\Pr_P \left[ \log \frac{P}{Q} \leq R \right] \geq \Pr_P \left[ \log \frac{P}{Q^i} \leq R + \log q(i) \right]
\]

and, relaxing the optimization in the definition of \( D_s^\epsilon \), we get \( D_s^\epsilon(P\|Q) \leq D_s^\epsilon(P\|Q^i) - \log q(i) \) as desired. \( \square \)

The following property will be particularly useful, as it allows us to bound 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 \( P \in \mathcal{P}(X) \) and let \( V, W \) be channels from \( X \) to \( Y \). Then,

\[
D_s^\epsilon(P \times W \| P \times V) \leq \sup_{x : P(x) > 0} D_s^\epsilon(W(\cdot|x)\|V(\cdot|x)).
\]

Proof. We first note that the log-likelihood ratio takes on the form

\[
\log \frac{P \times W}{P \times V} : (x, y) \mapsto \log \frac{P(x)W(y|x)}{P(x)V(y|x)} = \log \frac{W(y|x)}{V(y|x)},
\]

and is thus independent of \( P \). Now, we may write

\[
R^* = D_s^\epsilon(P \times W \| P \times V) = \sup \left\{ R \in \mathbb{R} \left| \frac{\Pr_P \left[ \log \frac{P \times W}{P \times V} \leq R \right]}{R} \leq \epsilon \right\}
\]

\[
= \sup \left\{ R \in \mathbb{R} \left| \sum_{x \in X} P(x) \Pr_{W(\cdot|x)} \left[ \log \frac{W(\cdot|x)}{V(\cdot|x)} \leq R \right] \leq \epsilon \right\}.
\]

Inspecting this expression, for any \( \mu > 0 \), we find at least one \( x^* \in X \) such that

\[
P(x^*) > 0 \quad \text{and} \quad \Pr_{W(\cdot|x^*)} \left[ \log \frac{W(\cdot|x^*)}{V(\cdot|x^*)} \leq R^* - \mu \right] \leq \epsilon.
\]

Hence, \( D_s^\epsilon(W(\cdot|x^*)\|V(\cdot|x^*)) \geq R^* - \mu \), which implies the lemma as \( \mu \to 0 \). \( \square \)

The distribution of the log-likelihood ratio has the following asymptotic expansion for not necessarily identical product distributions.

Lemma 5. Let \( P_i, Q \in \mathcal{P}(Z) \) be such that \( P_i \ll Q \) for all \( i \) in some finite set \( \mathcal{I} \). We consider a sequence of distributions \( P_{ik} \) indexed by \( (i_1, i_2, \ldots, i_n) \) where \( i_k \in \mathcal{I} \) for each \( 1 \leq k \leq n \). Define

\[
D_n := \frac{1}{n} \sum_{k=1}^n D(P_{ik} \| Q), \quad V_n := \frac{1}{n} \sum_{k=1}^n V(P_{ik} \| Q), \quad \text{and} \quad T_n := \frac{1}{n} \sum_{k=1}^n \mathbb{E} \left[ \left( \log \frac{P_{ik}}{Q} - D(P_{ik} \| Q) \right)^3 \right].
\]
If $V_n \geq V_\varepsilon > 0$, then we have
\[
D_\varepsilon^x(P_1 \times \ldots P_n \| Q^n) \leq nD_n + \sqrt{nV_n} \Phi^{-1}(\varepsilon + \frac{6T_n}{\sqrt{nV_n^3}}).
\]
In any case, we have
\[
D_\varepsilon^x(P_1 \times \ldots P_n \| Q^n) \leq nD_n + \sqrt{nV_n} \Phi^{-1}(\varepsilon + \frac{6T_n}{\sqrt{nV_n^3}}).
\]

Proof. We consider the cumulative distribution of the random variable $S_n := \sum_k \log P_{ik}(X_{ik}) - \log Q(X_{ik})$ where each $X_{ik}$ has distribution $P_{ik}$. The random variable $S_n$ has mean $nD_n$ and variance $nV_n$. The general case, Eq. (5), is shown using Chebyshev’s inequality, which yields
\[
\varepsilon \geq \Pr\left[\sum_k \log \frac{P_{ik}}{Q} \leq R\right] \geq 1 - \frac{nV_n}{(R - nD_n)^2} \quad \text{for } R > nD_n
\]
Hence, restricting to $R > nD_n$ and relaxing the condition of the supremum, we find
\[
D_\varepsilon^x(P_1 \times \ldots P_n \| Q^n) \leq \sup\left\{R > nD_n \middle| 1 - \frac{nV_n}{(R - nD_n)^2} \leq \varepsilon \right\} = nD_n + \sqrt{nV_n} \Phi^{-1}(\varepsilon + \frac{6T_n}{\sqrt{nV_n^3}}).
\]
Furthermore, if $V_n \geq V_\varepsilon > 0$, the Berry-Essèen theorem [15, Ch. XVI.5] states that
\[
\left|\Pr\left[\sum_k \log \frac{P_{ik}}{Q} \leq R\right] - \Phi\left(\frac{R - nD_n}{\sqrt{nV_n}}\right)\right| \leq \frac{6T_n}{\sqrt{nV_n^3}}.
\]
Hence, we obtain
\[
D_\varepsilon^x(P_1 \times \ldots P_n \| Q^n) \leq nD_n + \sqrt{nV_n} \Phi^{-1}(\varepsilon + \frac{6T_n}{\sqrt{nV_n^3}}),
\]
which concludes the proof. \hfill \Box

B. Converse Bounds on General Channels

Here, we give a new converse bound on the code size for general channels.

Proposition 6. Let $\varepsilon \in (0, 1)$ and let $W$ be any channel. Then, for any $\delta \in (0, 1 - \varepsilon)$, we have
\[
\log M^*(W, \varepsilon) \leq \inf_{Q \in \mathcal{P}(Y)} \max_{x \in X} D_\varepsilon^x(W(\cdot|x)\|Q) + \log \frac{1}{\delta}.
\]

Note that the first part of the proof of the converse is similar to the meta-converse of PPV [4]; however, we give a conceptually simple alternative proof along the lines of Wang-Colbeck-Renner [9] and Wang-Renner [10].

Proof. For any code $C = \{M, e, d\}$ with $p_{err}(C) \leq \varepsilon$ and any $Q \in \mathcal{P}(Y)$, the following holds.

Starting from a uniform distribution over $M$, the Markov chain $M \xrightarrow{e} X \xrightarrow{W} Y \xrightarrow{d} M'$ induces a joint probability distribution $P_{MXYM'}$. Due to the data-processing inequality for $D_\varepsilon^x$, we immediately find $D_\varepsilon^x(P\times W\| P\times Q) = D_\varepsilon^x(P_{XY}\| P_X\times Q_Y) \geq D_\varepsilon^x(P_{MM'}\| P_M\times Q_{M'})$, where $Q_{M'}$
is the distribution induced by $d$ applied to $Q_Y = Q$. Moreover, using the test $\xi(m, m') = \delta_{m, m'}$, we can readily see that

$$E_{P_{MM'}} [\xi(M, M')] = \Pr_{P_{MM'}} [M = M'] \geq 1 - \varepsilon$$

Hence, $D^\varepsilon_h (P_{MM'} \| P_M \times Q_{M'}) \geq \log |C| + \log (1 - \varepsilon)$ by definition of the $\varepsilon$-hypothesis testing divergence. Finally, applying Lemmas 2 and 4, we find

$$\max_{x \in X} D^\varepsilon_{\delta} (W^\cdot | x)^{Q_n}) \geq D^\varepsilon_{\delta} (P \times W \| P \times Q) \geq D^\varepsilon_h (P \times W \| P \times Q) - \log \frac{1 - \varepsilon}{\delta} \geq \log |C| - \log \frac{1}{\delta}.$$ 

This yields the converse bound upon minimizing over $Q \in P(Y)$. \hfill \Box

C. A Suitable Choice of Output Distribution $Q$

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

$$\log M^*(W^n, \varepsilon) \leq \min_{Q(n) \in P(Y \times n)} \max_{x \in X^n} D^\varepsilon_{\delta} (W^n \cdot | x)^{Q(n)}) + \log \frac{1}{\delta},$$

and it is important to find a suitable choice of $Q(n) \in P(Y \times n)$ 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 $\zeta := |Y|(|Y| - 1)$ and let $\gamma > 0$ be a constant which is to be chosen later. Consider the following convex combination of product distributions:

$$Q_n(y) := \frac{1}{2} \sum_{k \in K} \exp \left( - \gamma \frac{\|k\|_2^2}{2} \right) \prod_{i=1}^{n} Q_k(y_i) + \frac{1}{2} \sum_{P_k \in P_n(X)} \frac{1}{|P_n(X)|} \prod_{i=1}^{n} P_k W(y_i),$$

(6)

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

$$Q_k(y) := Q^*(y) + \frac{k_y}{\sqrt{n\zeta}}, \quad K := \left\{ k \in \mathbb{Z}^{|Y|} \mid \sum_y k_y = 0 \land k_y \geq -Q^*(y) \sqrt{n\zeta} \right\}.$$
In (6), $F$ is a normalization constant that ensures $\sum_y Q^{(n)}(y) = 1$. Note that what we have done in our choice of $Q_k$ is to uniformly quantize the simplex $\mathcal{P}(\mathcal{Y})$ along axis-parallel directions. The constraint that each $k$ belongs to $\mathcal{K}$ ensures that each $Q_k$ is a valid probability mass function. See Fig. 1. We find that

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

is a finite constant. Furthermore, by construction, the representation points $\{Q_k\}_k$ form an $\epsilon$-net with $\epsilon = n^{-\frac{1}{2}}$ for $\mathcal{P}(\mathcal{Y})$. Namely, for every $Q \in \mathcal{P}(\mathcal{Y})$, there exists a $k$ such that $\|Q - Q_k\|_2 \leq n^{-\frac{1}{2}}$. This can be verified easily since by choosing a $k$ that minimizes the distance in all but one direction (say the last), yielding

$$\|Q - Q_k\|_2^2 = \sum_{y=1}^{|\mathcal{Y}|-1} (Q(y) - Q_k(y))^2 + (Q(|\mathcal{Y}|) - Q_k(|\mathcal{Y}|))^2$$

$$= \sum_{y=1}^{|\mathcal{Y}|-1} (Q(y) - Q_k(y))^2 + \left( \sum_{y=1}^{|\mathcal{Y}|-1} Q_k(y) - Q(y) \right)^2$$

$$\leq \sum_{y=1}^{|\mathcal{Y}|-1} \left( \frac{1}{\sqrt{n\zeta}} \right)^2 + \left( \sum_{y=1}^{|\mathcal{Y}|-1} \frac{1}{\sqrt{n\zeta}} \right)^2 = \frac{1}{n}.$$

D. Continuity around the CAIDs and the unique CAOD

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

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

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

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

Note that $\cap_{\eta>0} \Gamma_\mu^\eta = \Pi_\mu W$ if the stochastic matrix $W$ has full rank. See Fig. 2 for an illustration. The following Lemma summarizes known results about these sets.

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

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


Proof. Properties 1 and 2 hold for small enough $\mu$ and $\eta$ by continuity since $Q^*$ has full support \cite[Cor. 1 to Thm. 4.5.1]{14} and $V(W\|P^*W\|P^*) \geq V_{\text{min}} > 0$. Property 3 was established by Strassen \cite{3} as well as PPV \cite[Eq. (501)]{4}. Since $D(P \times W\|P \times Q) = I(P,W) + D(PW\|Q)$, Property 4 follows immediately from the fact that $D(PW\|Q) \leq \frac{1}{\min_{y \in Y} Q(y)} \|PW - Q\|_2^2$ (see, e.g., \cite[Lem. 6.3]{16}).

To verify Properties 5 and 6, note that the quotient $W(y|x)/Q(y) < \infty$ by Property 1. If $W(y|x)/Q(y) = 0$, the corresponding terms in the sums defining $V(P,W)$ and $V(W\|P)$ are excluded because $\vartheta \log^k \vartheta \to 0$ as $\vartheta \to 0$ for all $k > 0$. Hence, $P \mapsto V(P,W)$ and $Q \mapsto V(W\|Q)$ are continuously differentiable on $\Pi_{\mu}$ and $\Gamma_{\mu}^\eta$ respectively. Because $t \mapsto \sqrt{t}$ is continuously differentiable away from 0, by Property 2, $P \mapsto \sqrt{V(P,W)}$ and $Q \mapsto \sqrt{V(W\|Q)}$ are Lipschitz on $\Pi_{\mu}$ and $\Gamma_{\mu}^\eta$ respectively.

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 $\delta = n^{-\frac{1}{4}}$, which satisfies $0 < \delta < 1 - \varepsilon$ for sufficiently large $n$. Substitute the output distribution $Q^{(n)}$ in (6) to get

$$\log M^*(W^n,\varepsilon) \leq \max_{x \in \mathcal{X}} D^{\varepsilon + \delta}_x(W^n(\cdot|x)\|Q^{(n)}) + \frac{1}{2} \log n.$$ 

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

We apply Lemma 7 that supplies us with constants $\mu$, $\eta$, $y_{\text{min}}$, $\alpha$ and $\beta$ and distinguish between two cases for the following; either a) $x$ satisfies $P_x \notin \Pi_{\mu}$ or b) $x$ satisfies $P_x \in \Pi_{\mu}$. This strategy in which we partition input types into two classes was proposed by Strassen \cite[Sec. 4]{3}. See also PPV \cite[Appendix I]{4}.
Case a): $P_x \notin \Pi_\mu$

The mutual information outside $\Pi_\mu$ is bounded away from the capacity, i.e., $I(P_x, W) \leq C'_x < C$ for all $P_x \notin \Pi_\mu$.

We first apply Lemma 3 and then Lemma 5 to bound

$$cv(x) \leq D^+ + \delta(W^n(\cdot|x)| (P_xW)^{\times n}) + \log(2|\mathcal{P}_n(\mathcal{X}))$$

$$\leq nI(P_x, W) + \sqrt{nV(P_x, W) \over 1 - \varepsilon - \delta} + \log(2|\mathcal{P}_n(\mathcal{X}))$$

For the second inequality, we note that $D_n$ in Lemma 3 evaluates to

$$D_n = {1 \over n} \sum_{i=1}^n E_{W(x)} \left[ \log W(x_i) \right] = E_{P_xW} \left[ \log W \right] = D(P_x \times W||P_x \times P_W) = I(P_x, W),$$

and similar calculation can be done to show that $V_n = V(P_x, W)$. Invoking [4, Lem. 62] and [13, Rmk. 3.1.1] yields the uniform bound $V(P_x, W) \leq 8 \log^2 \varepsilon |\mathcal{Y}| \leq 2.3 |\mathcal{Y}|$. Hence,

$$cv(x) \leq nC' + \sqrt{n} \log \left( 2.3 |\mathcal{Y}| \right) + (|\mathcal{X}| - 1) \log (n + 1) + \log 2.$$

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 $cv(x) \leq nC + \sqrt{n}V_x \Phi^{-1}(\varepsilon)$ for sufficiently large $n$.

Case b): $P_x \in \Pi_\mu$

Before we commence, let us define the third absolute moment of the log-likelihood ratio between $P$ and $Q$ to be $T(P||Q) := E_P \left[ |\log F - D(P||Q)|^3 \right]$. Also define

$$V_+ := \max_{P \in \mathcal{P}(\mathcal{X}, Q \in \mathcal{G}_P)} V(W||Q|P), \quad T_+ := \max_{P \in \mathcal{P}(\mathcal{X}) Q \in \mathcal{G}_P} T(W||Q|P),$$

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 $x$, we denote by $Q_{k(x)}$ the element of the $\epsilon$-net (constructed in Section III C) closest to $P_xW$. We note that since $\|Q_{k(x)} - P_xW\|_2 \leq \epsilon = n^{-1/2}$, we have $Q_{k(x)} \in \Gamma^\alpha_n$ for sufficiently large $n$, which enables us to apply the properties described in Lemma 7 extensively below.

We first use Lemma 3 to bound

$$cv(x) \leq D^++\delta(W^n(\cdot|x)|| (Q_{k(x)})^{\times n}) + \gamma\|k(x)\|^2_2 + \log(2F).$$

We now employ Lemma 5, where we choose $P_1 = W(\cdot|x_i)$ and may set $V_- = V_{\text{min}}$ due to Lemma 7. The bound on the third moment, $T_n$ in Lemma 5, still depends on $Q_{k(x)}$. However, we can upper-bound $T(Q_{k(x)})$ by $T_+$ which is finite. We then introduce the finite constant $B := 1 + 6T_+ / V_+^{3/2}$, while substituting for $\delta = n^{-1/2}$, to get

$$cv(x) \leq nD(P_xW||P_x \times Q_{k(x)}) + \sqrt{nV(W||Q_{k(x)}|P_x)} \Phi^{-1} \left( \varepsilon + {B \over \sqrt{n}} \right) + \gamma\|k(x)\|^2_2 + \log(2F).$$
We now require that \( n \geq N \), where \( N \) is chosen large enough such that \( \varepsilon + \frac{B}{\sqrt{n}} < 1 \). This ensures that the term growing as \( \sqrt{n} \) in the above expression is finite. Next, we use the fact that \( \Phi^{-1} \) is infinitely differentiable and \( V(W||Q_{k(x)}|P_x) \leq V_+ \) is finite to bound

\[
\sqrt{nV(W||Q_{k(x)}|P_x)} \Phi^{-1} \left( \varepsilon + \frac{B}{\sqrt{n}} \right) \leq \sqrt{nV(W||Q_{k(x)}|P_x)} \Phi^{-1}(\varepsilon) + G_1,
\]

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

\[
cv(x) \leq nD(P_X \times W|P_X \times Q_{k(x)}) + \sqrt{nV(W||Q_{k(x)}|P_x)} \Phi^{-1}(\varepsilon) + \gamma \|k(x)\|_2^2 + G_2,
\]

Next, we would like to replace \( Q_{k(x)} \) with \( P_x W \) in the above bound. This can be done without too much loss due to Lemma 7, which states that

\[
\text{infinitely differentiable and } \| \Phi^{-1}(\varepsilon) \|_{Lip} \leq G_1.
\]

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

\[
cv(x) \leq nI(P_X, W) + \sqrt{nV(P_X, W)} \Phi^{-1}(\varepsilon) + \gamma \|k(x)\|_2^2 + G_3.
\]

In the following, we use the fact that all distributions (and types) \( P_x \) in \( \mathcal{P}_\mu \) satisfy \( I(P_X, W) \leq C - \alpha \xi^2 \) and \( |\sqrt{V(P_X, W)} - \sqrt{V(P^*, W)}| \leq \beta \xi \), where \( P^* := \arg \min_{P \in \Pi} \|P - P^*\|_2 \) (which is unique) and \( \xi := \|P_X - P^*\|_2 \). Hence,

\[
cv(x) \leq nC + \sqrt{nV(P^*, W)} \Phi^{-1}(\varepsilon) + \left( -\alpha \xi^2 n + \beta |\Phi^{-1}(\varepsilon)| \xi \sqrt{n} + \gamma \|k(x)\|_2^2 \right) + G_3.
\]

It thus remains to show that the term in the bracket is upper bounded by a constant, for an appropriate choice of \( \gamma \). Let \( \|W\|_2 := \max\{\|uW\|_2 \mid \|u\|_2 \leq 1\} \) be the spectral norm of the matrix \( W \). From the construction of the \( \varepsilon \)-net in Section III C,

\[
\|k(x)\|_2 = \sqrt{n\xi} \|Q_{k(x)} - Q^*\|_2 \\
\leq \sqrt{n\xi} \left( \|Q_{k(x)} - P_x W\|_2 + \|P_x W - Q^*\|_2 \right) \\
\leq \sqrt{n\xi} \left( \frac{1}{\sqrt{n}} + \|W\|_2 \right).
\]

Substituting this bound into (7), we find that the term in the bracket evaluates to

\[
\left( \gamma \xi \|W\|_2^2 - \alpha \right) \xi^2 n + \left( \beta |\Phi^{-1}(\varepsilon)| + 2\gamma \xi \|W\|_2 \right) \xi \sqrt{n} + \gamma \xi
\]

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

\[
cv(x) \leq nC + \sqrt{nV(P^*, W)} \Phi^{-1}(\varepsilon) + G_4
\]

for an appropriate constant \( G_4 \) and \( n \geq N \).

Summarizing the bounds for Cases a) and b), we thus have the following asymptotic expansion for all \( n \) sufficiently large:

\[
\log M^*(W^n, \varepsilon) \leq \max_{P^* \in \Pi} nC + \sqrt{nV(P^*, W)} \Phi^{-1}(\varepsilon) + \frac{1}{2} \log n + G_4
\]

\[= nC + \frac{1}{2} \log n + G_4,
\]

where the last equality follows by definition of \( V_\varepsilon \). \( \square \)
Acknowledgements

MT thanks Ligong Wang for helpful explanations. 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.