The third-order term in the normal approximation for singular channels

Yücel Altuğ Member and Aaron B. Wagner Senior Member

Abstract

For a singular and symmetric discrete memoryless channel with positive dispersion, the third-order term in the normal approximation is shown to be upper bounded by a constant. This finding completes the characterization of the third-order term for symmetric discrete memoryless channels. The proof method is extended to asymmetric and singular channels with constant composition codes, and its connection to existing results, as well as its limitation in the error exponents regime, are discussed.

I. INTRODUCTION

Decades after its introduction in information theory (e.g., [1], [2]), the normal approximation has recently enjoyed a surge in interest. See, for instance, [3]–[32] for a partial list of recent work that is most closely related to the present paper. When particularized to coding over a discrete memoryless channel (DMC), say W, the normal approximation states that for any positive integer n and \( \epsilon \in (0, 1) \), the logarithm of the maximum number of messages that can be communicated with an error probability not larger than \( \epsilon \) behaves asymptotically as

\[
\frac{1}{n} C(W) + \sqrt{n} V_e(W) \Phi^{-1}(\epsilon) + O(\ln n),
\]

where \( C(W) \) and \( V_e(W) \) are the capacity and the channel error exponent, respectively, and \( \Phi(\cdot) \) denotes the distribution of the standard normal random variable.

Although the first two terms in (1) are well-understood, the third-order term has proven to be more elusive (e.g., [3], [7] Sec. 3.4.5], [32]). Some bounds are available, however. The third-order term is known to be no greater than \( \ln \sqrt{n} \) and no smaller than a constant (i.e., it cannot diverge to negative infinity) [6, Theorem 45]. Each bound is tight for some channel. The upper bound is tight for a class of channels that includes the binary symmetric channel (BSC) [7, Sec. 3.4.5]. [32] while the lower bound is tight for the binary erasure channel (BEC) [6, Theorem 53]. It is not known, however, whether these two extremes are the only possibilities.

In this paper, we prove that for symmetric channels, these are indeed the only two possibilities. Specifically, we show that for a singular and symmetric DMC with positive dispersion, the third-order term is upper bounded by a constant (see Proposition 1). By combining this finding with existing results in the literature, we can conclude that the third-order term for a symmetric DMC with positive dispersion is either \( \ln \sqrt{n} \) or a constant depending whether the channel is nonsingular or singular (see Theorem 1 to follow).

It is worth noting that the analogous result for error exponents is already known [33]. For symmetric channels, the optimal order of the sub-exponential factor in the error exponents regime is \( \Theta(n^{-0.5}) \) in the singular case and \( \Theta(n^{-0.5(1+E'(R))}) \) in the nonsingular case, where \( E'(\cdot) \) is the derivative of the reliability function [33]. In fact, the proof for the result presented here is based on the proof of this error exponent result. In Section IV-A we show how our main result can also be proven via the “minimax converse” (e.g., [22, Theorem 1]) in which a non-product output distribution is utilized. Polyanskiy [22, Section VI.D] had earlier showed how the constant upper bound on the third-order term in the normal approximation for the BEC could be obtained via the minimax converse with a particular non-product output distribution. Hence, an ancillary contribution of this paper is to show how the proof technique of Polyanskiy [22, Section VI.D] can be extended to all singular and symmetric channels.

Our proof technique can also be extended to asymmetric and singular channels, provided that attention is restricted to constant composition codes (see Proposition 2). In Section IV-B we discuss the difficulty in dropping this assumption for asymmetric and singular channels. Even with the constant composition assumption, our proof technique does not carry over easily to the error exponents regime if the channel is asymmetric, as we discuss in Section IV-C.

The authors are with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853. E-mail: yalu6@cornell.edu, wagner@ece.cornell.edu.

1Throughout the paper, we use nats as the unit of information.
2For a definition of symmetric channels, see Definition 2.
3For a definition of singular channels, see Definition 3.
4A code is constant composition if all the codewords have the same empirical distribution.
II. NOTATION, DEFINITIONS AND STATEMENT OF THE RESULTS

A. Notation

\(\mathbb{Z}^+, \mathbb{R}, \mathbb{R}^+\) and \(\mathbb{R}_+\) denote the set of positive integers, real, positive real and non-negative real numbers, respectively. Boldface letters denote vectors, regular letters with subscripts denote individual components of vectors. Furthermore, capital letters represent random variables and lowercase letters denote individual realizations of the corresponding random variable.

For a finite set \(\mathcal{X}\), \(\mathcal{P}(\mathcal{X})\) (resp. \(U_\mathcal{X}\)) denotes the set of all probability measures (resp. the uniform probability measure) on \(\mathcal{X}\). Similarly, for two finite sets \(\mathcal{X}\) and \(\mathcal{Y}\), \(\mathcal{P}(\mathcal{Y}|\mathcal{X})\) denotes the set of all stochastic matrices from \(\mathcal{X}\) to \(\mathcal{Y}\). Given any \(P \in \mathcal{P}(\mathcal{X})\), \(S(P) := \{x \in \mathcal{X} : P(x) > 0\}\). For any finite set \(\mathcal{X}\) and \(n \in \mathbb{Z}^+\), \(P^n_x\) denotes the type of the sequence \(x^n\) and \(P^n\) (\(\mathcal{X}\)) denotes the set of all types on \(\mathcal{X}^n\). \(\mathbb{1}\{\cdot\}\) denotes the standard indicator function. \(\phi(\cdot)\) denotes the density of the standard Gaussian random variable. For a set \(\mathcal{S}\), \(\text{cl}(\mathcal{S})\) and \(\mathcal{S}^c\) denotes the closure of \(\mathcal{S}\) and the complementary set, respectively. We follow the notation of the book of Csiszár-Körner [44] for standard information theoretic quantities.

B. Definitions

Definition 1. For any \(n \in \mathbb{Z}^+\) and \(\epsilon \in (0, 1)\),

\[
M^*(n, \epsilon) := \max\{[e^{nR}] \in \mathbb{R}_+ : \bar{P}_\epsilon(n, R) \leq \epsilon\},
\]

where \(\bar{P}_\epsilon(n, R)\) denotes the minimum average error probability attainable by any \((n, R)\) code. Further, for any \(n \in \mathbb{Z}^+\) and \(\epsilon \in (0, 1)\),

\[
M^*_\epsilon(n, \epsilon) := \max\{[e^{nR}] \in \mathbb{R}_+ : \bar{P}_{\epsilon, \epsilon}(n, R) \leq \epsilon\},
\]

where \(\bar{P}_{\epsilon, \epsilon}(n, R)\) denotes the minimum average error probability attainable by any \((n, R)\) constant composition code. \(\diamondsuit\)

Definition 2. (Gallager [45] pg. 94) A discrete memoryless channel is symmetric if the channel outputs can be partitioned into subsets such that within each subset, the matrix of transition probabilities satisfies the following: every row (resp. column) is a permutation of every other row (resp. column). A channel that is not symmetric is called asymmetric. \(\diamondsuit\)

Definition 3. (Gallager Definition 2) A channel \(W \in \mathcal{P}(\mathcal{Y}|\mathcal{X})\) is singular if

\[
\forall (x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{X} \text{ with } W(y|x)W(y|z) > 0, W(y|x) = W(y|z).
\]

A channel that is not singular is called nonsingular. \(\diamondsuit\)

Given \(W \in \mathcal{P}(\mathcal{Y}|\mathcal{X}), C(W)\) denotes the capacity of the channel. For any \(P \in \mathcal{P}(\mathcal{X})\), define \(q_P(y) := \sum_{x \in \mathcal{X}} P(x)W(y|x)\). For convenience, let \(q\) denote \(q_{U_\mathcal{X}}\). Given any \(W \in \mathcal{P}(\mathcal{Y}|\mathcal{X}), P \in \mathcal{P}(\mathcal{X})\) and \(\epsilon \in (0, 1)\), define (e.g., [7] Sec. 3.4)

\[
V(P, W) := \sum_{x,y} P(x)W(y|x) \left[ \ln \frac{W(y|x)}{q_P(y)} - \sum_b W(b|x) \ln \frac{W(b|x)}{q_P(b)} \right]^2.
\]

\[
V_\epsilon(W) := \begin{cases} \min_{Q:Q(\mathcal{Y}|\mathcal{W})=C(W)} V(Q, W), & \text{if } \epsilon \in (0, 1/2), \\ \max_{Q:Q(\mathcal{Y}|\mathcal{W})=C(W)} V(Q, W), & \text{if } \epsilon \in [1/2, 1). \end{cases}
\]

\[
V^*(P, W) := \sum_{x,y} P(x)W(y|x) \left[ \ln \frac{W(y|x)}{q_P(y)} - \sum_z P(z)W(y|z) \ln \frac{W(y|z)}{q_P(y)} \right]^2.
\]

The following result gives an equivalent definition of singularity in terms of the quantity defined in (7).

Lemma 1. Consider a channel \(W \in \mathcal{P}(\mathcal{Y}|\mathcal{X})\) with \(P(x) > 0\) for all \(x \in \mathcal{X}\). \(V^*(P, W) = 0\) if and only if \(W\) is singular. \(\diamondsuit\)

Proof: We note that

\[
[V^*(P, W) = 0] \iff \left[ \forall y \in \mathcal{Y}, \ln W(y|x) = \sum_z \frac{P(z)W(y|z)}{q_P(y)} \ln W(y|z), \forall x \in \mathcal{X} \text{ with } W(y|x) > 0 \right],
\]

which is a direct consequence of the definition of \(V^*(U_\mathcal{X}, W)\), i.e., (7). In light of Definition 3 the right side of (8) is equivalent to saying \(W\) is singular. \(\square\)

Remark 1. In [7] Lemma 52, it is claimed that

\[
[V^*(P, W) = 0] \iff \left[ \forall (x, y, y') : W(y|x) = W(y'|x) \text{ or } P(x)W(y|x) = 0 \right].
\]

By choosing \(P = U_\mathcal{X}\) and \(W\) as BEC with parameter \(\delta \in (0, 1)\), one can verify that \(V^*(P, W) = 0\) by elementary calculation. Evidently, this \((P, W)\) pair does not satisfy the right side of (8) and hence (9) is incorrect.
Proposition 1. \( \forall (x, y, z) \in X \times Y \times X \) with \( P(x)W(y|x)P(z)W(y|z) > 0, W(y|x) = W(y|z) \),

then, by noticing

\[
[V^+(P, W) = 0] \iff \left[ \forall y \in Y, \ln W(y|x) = \frac{P(z)W(y|z)}{q_P(y)} \ln W(y|z), \forall x \in X \text{ with } P(x)W(y|x) > 0 \right]
\]

it is easy to see that \((9)\) holds. \(\diamond\)

C. Results

**Proposition 1.** Given \( \epsilon \in (0, 1) \) and a singular, symmetric \( W \) with \( V_\epsilon(W) > 0 \), we have

\[
\ln M^*(n, \epsilon) \leq nC(W) + \sqrt{nV_\epsilon(W)}\Phi^{-1}(\epsilon) + \tilde{K}(\epsilon, W),
\]

where \( K(\epsilon, W) \in \mathbb{R}^+ \) is a constant that depends on \( \epsilon \) and \( W \).

**Proof:** Given in Section III-A \( \blacksquare \)

**Theorem 1.** Given a symmetric \( W \) and \( \epsilon \in (0, 1) \), we have the following:

(i) If \( W \) is nonsingular and \( V_\epsilon(W) > 0 \), then

\[
\ln M^*(n, \epsilon) = nC(W) + \sqrt{nV_\epsilon(W)}\Phi^{-1}(\epsilon) + \ln \sqrt{n} + \Theta(1).
\]

(ii) If \( W \) is singular and \( V_\epsilon(W) > 0 \), then

\[
\ln M^*(n, \epsilon) = nC(W) + \sqrt{nV_\epsilon(W)}\Phi^{-1}(\epsilon) + \Theta(1).
\]

(iii) If \( V_\epsilon(W) = 0 \), then

\[
\ln M^*(n, \epsilon) = nC(W) + \Theta(1).
\]

\(\blacklozenge\)

**Proof:** We point out the existing results that justify the cases except the converse statement of item (ii), which follows from Proposition [1] Achievability of item (i) follows from \([7\text{ Corollary 54]}\) that is applicable due to Lemma [1] Converse of item (i) follows from \([7\text{ Theorem 55]}\). Achievability of item (ii) follows from \([7\text{ Theorem 47]}\), coupled with item (ii) of Lemma [4] Item (iii) is proved in \([7\text{ Corollary 57]}\). \(\blacksquare\)

**Proposition 2.** Given a singular and asymmetric \( W \), we have

(i) If \( \epsilon \in (0, 1/2) \), then

\[
\ln M^*_\epsilon(n, \epsilon) \leq nC(W) + \sqrt{nV_\epsilon(W)}\Phi^{-1}(\epsilon) + \tilde{K}(\epsilon, W),
\]

where \( \tilde{K}(\epsilon, W) \in \mathbb{R}^+ \) is a constant that depends on \( \epsilon \) and \( W \).

(ii) If \( \epsilon \in (1/2, 1) \) and \( V_\epsilon(W) > 0 \), then

\[
\ln M^*_\epsilon(n, \epsilon) \leq nC(W) + \sqrt{nV_\epsilon(W)}\Phi^{-1}(\epsilon) + \tilde{K}'(\epsilon, W),
\]

where \( \tilde{K}'(\epsilon, W) \in \mathbb{R}^+ \) is a constant that depends on \( \epsilon \) and \( W \). \(\blacklozenge\)

**Proof:** Given in Section III-B \(\blacksquare\)

**Remark 2.** (i) The set of asymmetric and singular channels is not empty. For an example, let \( X := \{0, 1, 2\} \), \( Y := \{0, 1, 2, 3\} \) and consider

\[
W(y|x) := \begin{cases} 2/3, & \text{if } (x, y) = (0, 0) \\ 1/6, & \text{if } (x, y) \in \{(0, 1), (0, 3), (1, 3), (2, 1)\} \\ 5/6, & \text{if } (x, y) \in \{(1, 2), (2, 2)\} \\ 0, & \text{else}. \end{cases}
\]

(ii) We do not analyze the zero-dispersion case for \( \epsilon \in (1/2, 1) \), because the third-order term also depends on whether the channel is exotic (e.g., \([6\text{ pg. 2331]}\)) and the main purpose of this paper is to investigate the effect of singularity on the third-order term. Similarly, we do not consider \( \epsilon = 1/2 \) case, since the third-order term also depends on whether the channel is exotic. See \([22 \text{ Section III]}\) for a detailed discussion on the effect of the exotic property of the channel on the third-order term when \( \epsilon \in [1/2, 1) \). \(\blacklozenge\)

\(^5\)Note that \([10] \) is the definition of singularity given in \([59 \text{ Definition 1]}\).
III. PROOFS

First, we prove two lemmas that will be used in the proofs of both Proposition[1] and [2]. To this end, consider any \( Q \in \mathcal{P}(\mathcal{X}) \) and define
\[
\alpha_y(Q) := \sum_{x : W(y|x) > 0} Q(x).
\]  
(19)
Consider any singular \( W \in \mathcal{P}(\mathcal{Y} | \mathcal{X}) \). As a direct consequence of the singularity of the channel, for any \( y \in \mathcal{Y} \), \( W(y|x) \) is either 0 or a column specific positive constant, say \( \delta_y \). For any \( y \in \mathcal{Y} \), \( q_y(y) = \delta_y \alpha_y(Q) \). The following set will be instrumental in our analysis:
\[
S_R(Q) := \left\{ y^n : \frac{1}{n} \sum_{i=1}^{n} \ln \frac{1}{\alpha_y(Q)} \leq R \right\},
\]  
(20)
for any \( R \in \mathbb{R}_+ \).

**Lemma 2.** Consider a singular \( W \in \mathcal{P}(\mathcal{Y} | \mathcal{X}) \). Consider any \((n, R)\) code, say \((f, \varphi)\), with codewords \( \{x^n(m)\}_{m=1}^{\left|\mathcal{M}\right|} \) where \( \mathcal{M} := \{1, \ldots, [e^{nR}]\} \) denotes the set of messages. Let \( \bar{P}_e(f, \varphi) \) denote the average error probability of this code. Fix some \( Q \in \mathcal{P}(\mathcal{X}) \) and \( x^n \in \mathcal{X}^n \) and assume that for all \( m \in \mathcal{M} \), \( W(S_R(Q)|x^n(m)) = W(S_R(Q)|z^n) \) and \( q_Q \) dominates \( W(-|x) \) for all \( x \in S(P_{x^n(m)}) \). Then we have
\[
\bar{P}_e(f, \varphi) \geq W(S_R(Q)|z^n) - \sum_{y^n \in S_R(Q)} q_Q(y^n)e^{-n\left[R - \frac{1}{n} \sum_{i=1}^{n} \ln \frac{1}{\alpha_y(Q)}\right]}.
\]  
(21)

**Proof:** Assume \( W(S_R(Q)|z^n) > 0 \), otherwise (21) is trivial. For any \( x^n \in \mathcal{X}^n \) with \( W(S_R(Q)|x^n) > 0 \), define
\[
P_{Y|X,S_R(Q)}(y^n|x^n, S_R(Q)) := \begin{cases} \frac{W(y^n|x^n)}{W(S_R(Q)|x^n)}, & \text{if } y^n \in S_R(Q), \\ 0, & \text{else}. \end{cases}
\]  
(22)
Evidently, \( P_{Y|X,S_R(Q)}(\cdot|x^n, S_R(Q)) \) is a well-defined probability measure. Let \( \{A_m\}_{m=1}^{\left|\mathcal{M}\right|} \) denote the decoding regions of the code. We have
\[
\bar{P}_e(f, \varphi) = \frac{1}{\left|\mathcal{M}\right|} \sum_{m \in \mathcal{M}} \sum_{y^n \in A_m} W(y^n|x^n(m))
\]  
(23)
\[
\geq \frac{1}{\left|\mathcal{M}\right|} \sum_{m \in \mathcal{M}} \sum_{y^n \in A_m} W(S_R(Q)|x^n(m)) P_{Y|X,S_R(Q)}(y^n|x^n(m), S_R(Q))
\]  
(24)
\[
\geq W(S_R(Q)|z^n) \left[ 1 - \frac{1}{\left|\mathcal{M}\right|} \sum_{m \in \mathcal{M}} \sum_{y^n \in A_m} P_{Y|X,S_R(Q)}(y^n|x^n(m), S_R(Q)) \right],
\]  
(25)
where (24) follows from (22) and (25) follows from the assumption that \( W(S_R(Q)|x^n(m)) = W(S_R(Q)|z^n) \), for all \( m \in \mathcal{M} \). Define \( P_{D|Y}(y^n) := \{y^n \in A_m\} \), for all \( m \in \mathcal{M} \). Since the decoding regions are mutually exclusive and collectively exhaustive on \( \mathcal{M} \), \( P_{D|Y}(\cdot|y^n) \) is a well-defined probability measure. Hence, (25) implies that
\[
\bar{P}_e(f, \varphi) \geq W(S_R(Q)|z^n) \left[ 1 - \frac{e^{-nR}}{W(S_R(Q)|z^n)} \sum_{m \in \mathcal{M}} \sum_{y^n} P_{D|Y}(m|y^n)W(y^n|x^n(m))\mathbb{I}\{y^n \in S_R(Q)\} \right]
\]  
(26)
\[
\geq W(S_R(Q)|z^n) \left[ 1 - \frac{e^{-nR}}{W(S_R(Q)|z^n)} \sum_{m \in \mathcal{M}} \sum_{y^n} P_{D|Y}(m|y^n)\mathbb{I}\{y^n \in S_R(Q)\} q_Q(y^n)e^{-n\left[R - \frac{1}{n} \sum_{i=1}^{n} \ln \frac{1}{\alpha_y(Q)}\right]} \right]
\]  
(27)
\[
\geq W(S_R(Q)|z^n) \left[ 1 - \frac{e^{-nR}}{W(S_R(Q)|z^n)} \sum_{y^n} \mathbb{I}\{y^n \in S_R(Q)\} q_Q(y^n)e^{-n\left[R - \frac{1}{n} \sum_{i=1}^{n} \ln \frac{1}{\alpha_y(Q)}\right]} \right],
\]  
(28)
where (27) follows from the fact that \( q_Q(y) = \delta_y \alpha_y(Q) \) and the assumption that for all \( m \in \mathcal{M} \), \( q_Q \) dominates \( W(-|x) \) for all \( x \in S(P_{x^n(m)}) \).

**Remark 3.** Lemma [2] has the following intuitive interpretation. For simplicity, consider an \((n, R)\) constant composition code \((f, \varphi)\) with the common composition \( Q \). We write
\[
\bar{P}_e(f, \varphi) = \Pr(S_R(Q))\bar{P}_e(f, \varphi|S_R(Q)),
\]  
(29)
where \( \bar{P}_e(f, \varphi|S_R(Q)) \) denotes the average error probability of \((f, \varphi)\) conditioned on \( S_R(Q) \).
Given any $\mathbf{x}^n \in \mathcal{X}^n$ with $P_{\mathbf{x}^n} = Q$, $S_R(Q)$ captures the event that the empirical mutual information, i.e., $\frac{1}{n} \sum_{i=1}^n \ln \frac{W(Y_i|x^n)}{q(y)}$, is smaller than $R$ as a direct consequence of the singularity of $W$. Intuitively, the code will make an error if the channel realization is such that the resulting empirical mutual information is not large enough to support the coding rate, and this is our rationale in writing (29). Since $(f, \varphi)$ is a constant composition code, one can write $W(S_R(Q)|\mathbf{x}^n)$ in place of $\Pr(S_R(Q))$ and

$$1 - \frac{1}{W(S_R(Q)|\mathbf{x}^n)} \sum_{\mathbf{y}^n \in \mathcal{S}_R(Q)} q(Q) e^{-n \left[ R - \frac{1}{n} \sum_{i=1}^n \ln \frac{W(Y_i|x^n)}{q(y)} \right]}$$

(30)

can be viewed as a lower bound on $\bar{P}_e(f, \varphi|S_R(Q))$. Therefore, (21) can be considered as a lower bound on the right side of (29). ♦

We continue with a simple result for sums of independent random variables whose proof is inspired by the proof of [6, Lemma 47]. The reason of its inclusion is the fact that the bound in (31) is tighter than the one that follows by a direct application of [6, Lemma 47], at least by a factor of 2.

Lemma 3. Let $\{Z_i\}_{i=1}^n$ be independent with $m_{2,n} := \sum_{i=1}^n \text{Var}[Z_i] > 0$ and $m_{3,n} := \sum_{i=1}^n \text{E} \left[ (Z_i - \text{E}[Z_i])^3 \right] < \infty$. Then for any $r \in \mathbb{R}$ and $n \in \mathbb{Z}^+$

$$\text{E} \left[ I \left( \sum_{i=1}^n Z_i \leq r \right) e^{-[r - \sum_{i=1}^n Z_i]} \right] \leq \frac{1}{2\sqrt{\text{m}_{2,n}}} + \frac{2m_{3,n}}{m_{2,n}^{3/2}}.$$ (31)

Further, if the random variables are also identically distributed, then

$$\text{E} \left[ I \left( \sum_{i=1}^n Z_i \leq r \right) e^{-[r - \sum_{i=1}^n Z_i]} \right] \leq \frac{1}{2\sqrt{\text{m}_{2,n}}} + \frac{m_{3,n}}{m_{2,n}^{3/2}}.$$ (32)

Proof: Define $S_n := \sum_{i=1}^n Z_i$ and let $F_n$ denote the distribution function of $S_n$. For convenience, let $B_n(r)$ denote the left side of (31) and $m_{1,n} := \sum_{i=1}^n \text{E}[Z_i]$. We have

$$B_n(r) = e^{-r} \int_{-\infty}^r e^z dF_n(z)$$

(33)

$$= F_n(r) - \int_{-\infty}^r e^{(z-r)} F_n(z) dz$$

(34)

$$= \int_0^\infty e^{-x} \left[ F_n(r) - F_n(r-x) \right] dx$$

(35)

$$\leq \int_0^\infty e^{-x} \left\{ \int_{-\infty}^{r-m_{1,n}} e^{-\frac{z^2}{2m_{2,n}}} + \frac{m_{3,n}}{m_{2,n}^{3/2}} \right\} dx$$

(36)

$$\leq \frac{1}{2\sqrt{\text{m}_{2,n}}} + c \frac{m_{3,n}}{m_{2,n}^{3/2}}.$$ (37)

where (34) follows from integration by parts, (36) follows from the Berry-Esseen theorem\(^6\) and $c = 2$ (resp. $c = 1$) if the random variables are independent (resp. i.i.d.). ■

A. Proof of Proposition 1

Let $W \in \mathcal{P}(Y|X)$ be a symmetric and singular channel. Without loss of generality, assume $W$ has no all-zero column. Consider any $\epsilon \in (0, 1)$. Define

$$\forall x \in \mathcal{X}, M_2(\lambda) := \text{E}_{W(\cdot|x)} \left[ e^{\lambda \ln \frac{W(Y|x)}{q(Y)}} \right], m_3(x) := \text{E}_{W(\cdot|x)} \left[ \left| \ln \frac{W(Y|x)}{q(Y)} - C(W) \right|^3 \right],$$

(38)

for any $\lambda \in \mathbb{R}$. For convenience, let $\mathcal{S}_R$ denotes $\mathcal{S}_R(U_X)$, which is defined in (20).

Lemma 4. Let $W \in \mathcal{P}(Y|X)$ be a symmetric and singular channel. Fix an arbitrary $x_o \in \mathcal{X}$.

(i) For any $x \in \mathcal{X}$, $M_2(\lambda) = M_{2o}(\lambda)$ for all $\lambda \in \mathbb{R}$.

\(^6\)We take the constant in the Berry-Esseen theorem as 1 (resp. 1/2) if the random variables are independent (resp. i.i.d.), although neither choice is the best possible (e.g., [37]).
(ii) For all \( x \in \mathcal{X} \),
\[
E_{W(\cdot|x)} \left[ \ln \frac{W(Y|x)}{q(Y)} \right] = E_{W(\cdot|x_0)} \left[ \ln \frac{W(Y|x_0)}{q(Y)} \right] = C(W),
\]
(39)
\[
\text{Var}_{W(\cdot|x)} \left[ \ln \frac{W(Y|x)}{q(Y)} \right] = \text{Var}_{W(\cdot|x_0)} \left[ \ln \frac{W(Y|x_0)}{q(Y)} \right] =: V(W) = V_0(W),
\]
(40)
\[
m_3(x) = m_3(x_0).
\]
(41)

(iii) For any \( x^n \in \mathcal{X}^n \), \( W(S_R|x^n) = W(S_R|x^n_o) \), where \( x^n_o \) denotes the element of \( \mathcal{X}^n \) consisting of all \( x_o \).

(iv) \( E_q[-\ln \alpha_Y] = C(W) \), \( \text{Var}[{-\ln \alpha_Y}] = V(W) \) and \( E_q[-\ln \alpha_Y - C(W)] = m_3(x_o). \)
\( \diamondsuit \)

Proof: Since \( U_X \) is a capacity achieving input distribution of \( W \) (e.g., [33] Theorem 4.5.2) and the unique capacity achieving output distribution has full support (e.g., [33] Corollary 1 and 2 to Theorem 4.5.1), we conclude that \( \alpha_y > 0 \), for all \( y \in \mathcal{Y} \).

(i) Let \( \{Y_l\}_{l=1}^L \) be a partition\( \diamondsuit \) mentioned in Definition 2. Let \( \{W_l\}_{l=1}^L \) denote the sub-channel associated with each \( Y_l \), which is the same matrix formed by using the columns of \( Y_l \). Evidently, for any \( l \in \{1, \ldots, L\} =: \mathcal{L} \) and \( y_1, y_2 \in Y_l \), \( \delta_{y_1} = \delta_{y_2} \) and hence, for any \( l \in \mathcal{L} \), any entry of \( W_l \) can take only two values, either 0 or \( \delta_l \) with \( \delta_l := \delta_y \) for some \( y \in Y_l \). Following similar reasoning, \( \alpha(Y_l) \) is also constant along \( Y_l \) and we define \( \alpha_l := \alpha_y \) for all \( y \in Y_l \). For any \( l \in \mathcal{L} \), let \( \nu_l \) denote the number of positive elements in a row of \( W_l \).

For any \( x \in \mathcal{X} \) and \( \lambda \in \mathbb{R} \), we have
\[
M_x(\lambda) = \sum_{y:W(y|x)>0} \delta_y \nu_l \alpha_l^{-1},
\]
(42)
where the second equality follows due to the symmetry of the channel. Evidently, (42) ensures that \( M_x(\cdot) \) is finite on \( \mathbb{R} \) and also implies item (i).

(ii) Item (i), along with the uniqueness theorem for moment generating functions (e.g., [38] Ex. 26.7]), directly implies (39), (41) and
\[
\text{Var}_{W(\cdot|x)} \left[ \ln \frac{W(Y|x)}{q(Y)} \right] = V(W), \forall x \in \mathcal{X}.
\]
(43)
The last equality of (40) is evident in light of (43) and the fact that \( q \) is the unique capacity achieving output distribution of \( W \).

(iii) The claim is a direct consequence of item (i) of this lemma.

(iv) The claim directly follows from item (ii) of this lemma on account of the definition of \( q \) and the fact that \( q(y) = \delta_y \alpha_y \).
\( \blacksquare \)

Remark 4. Equations (39) and (40) are proved in [7] Theorem 55 for the set of weakly input symmetric channels that subsume symmetric channels. \( \diamondsuit \)

We conclude the proof as follows. First, we define
\[
k(W) := \frac{m_3(x_o)}{V(W)^{3/2}},
\]
(44)
\[
K(\epsilon, W) := \frac{k(W) \sqrt{V(W)}}{\phi(\Phi^{-1}(\epsilon))} + \frac{2}{\phi(\Phi^{-1}(\epsilon))} \left( \frac{1}{\sqrt{2\pi}} + \frac{m_3(x_o)}{V(W)} \right).
\]
(45)
Evidently, \( K(\epsilon, W) \in \mathbb{R}^+ \). Choose some \( n_0(\epsilon, W) \in \mathbb{Z}^+ \) such that for all \( n \geq n_0(\epsilon, W) \),
\[
\left\{ 1 - \frac{K(\epsilon, W)}{2\phi(\Phi^{-1}(\epsilon)) \sqrt{nV(W)}} \right\} > 1/2.
\]
(46)
Consider any \( n \geq n_0(\epsilon, W) \) and define
\[
R := C(W) + \sqrt{\frac{V(W)}{n}} \Phi^{-1}(\epsilon) + \frac{K(\epsilon, W)}{n}.
\]
(47)
Consider any \((n, R)\) code, say \((f, \varphi)\). Due to the fact that \( q(y) > 0 \) for all \( y \in \mathcal{Y} \) and item (iii) of Lemma 4 we can apply Lemma 2 to deduce
\[
\tilde{P}_e(f, \varphi) \geq W(S_R|x^n_o) - \sum_{y^n \in S_R} q(y^n) e^{-n \left[ \frac{R - \frac{1}{n} \sum_{i=1}^n \ln \frac{1}{\alpha_i}}{\Phi^{-1}(\epsilon)} \right]}.
\]
(48)
\( ^{7} \)The choice of the partition will be immaterial in the sequel.
Since $V(W) > 0$, item (iv) of Lemma 4 ensures that we can apply Lemma 3 to have
\[
\sum_{y^n \in S_R} q(y^n) e^{-n \left[ R - \frac{1}{2} \sum_{i=1}^n \ln \frac{1}{q(y_i)} \right]} \leq \frac{1}{\sqrt{2\pi n V(W)}} + \frac{k(W)}{\sqrt{n}}. \tag{49}
\]

Next, we claim that
\[
W(S_R | x_0^n) \geq \epsilon + \frac{K(\epsilon, W) \phi(\Phi^{-1}(\epsilon))}{\sqrt{n V(W)}} \left( 1 - \frac{K(\epsilon, W)}{\phi(\Phi^{-1}(\epsilon)) 2 \sqrt{n V(W)}} \right) - \frac{k(W)}{2 \sqrt{n}}. \tag{50}
\]
To see (50), we note that
\[
W(S_R | x_0^n) = W \left\{ \frac{1}{n} \sum_{i=1}^n \ln \frac{W(Y_i|x_0)}{q(Y_i)} \leq R \right\} \tag{51}
\]
\[
\geq \Phi \left( \Phi^{-1}(\epsilon) + \frac{K(\epsilon, W)}{\sqrt{n V(W)}} \right) - \frac{k(W)}{2 \sqrt{n}}, \tag{53}
\]
where (51) follows since $q(y) = \delta_y \alpha_y$, along with the singularity of the channel, (52) follows from the definition of $R$, i.e., (47), and (53) follows from the Berry-Esseen theorem, whose applicability is ensured by item (ii) of Lemma 4 and the fact that $V(W) > 0$. Via a straightforward power series approximation, one can check that (53) implies (50).

By plugging (49) and (50) into (48), along with (46) and noticing the fact that the code is arbitrary, we deduce that
\[
\forall n \geq n_0(\epsilon, W), \ln M^*(n, \epsilon) \leq nC(W) + \sqrt{n V(W)} \Phi^{-1}(\epsilon) + K(\epsilon, W), \tag{54}
\]
which, in turn, implies the desired result.

\[\square\]

B. Proof of Proposition 2

We separately analyze three different possibilities for the composition of the code $P$: large $I(P; W)$ with large $V(P, W)$, large $I(P; W)$ with small $V(P, W)$, and small $I(P; W)$. This idea originated in Strassen’s classical paper [2] and is frequently used in the normal approximation regime.

Specifically, given any $\delta, \nu \in \mathbb{R}^+$, we define
\[
S_1(\delta, \nu) := \left\{ \begin{array}{l} P \in \mathcal{P}(\mathcal{X}) : \min_{P^* \in \mathcal{P}_W} ||P - P^*||_2 \leq \delta \text{ and } V(P, W) \geq \nu \end{array} \right\}, \tag{55}
\]
\[
S_2(\delta, \nu) := \left\{ \begin{array}{l} P \in \mathcal{P}(\mathcal{X}) : \min_{P^* \in \mathcal{P}_W} ||P - P^*||_2 \leq \delta \text{ and } V(P, W) < \nu \end{array} \right\}, \tag{56}
\]
\[
S_3(\delta) := \left\{ \begin{array}{l} P \in \mathcal{P}(\mathcal{X}) : \min_{P^* \in \mathcal{P}_W} ||P - P^*||_2 > \delta \end{array} \right\}, \tag{57}
\]
where $\mathcal{P}_W := \{ P \in \mathcal{P}(\mathcal{X}) : I(P; W) = C(W) \}$. Throughout the section, $\bar{P}_e(f, \varphi)$ denotes the average error probability of the code $(f, \varphi)$.

**Lemma 5.** Fix some $W \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$ with $C(W) > 0$, $\delta \in \mathbb{R}^+$ and $\epsilon \in (0, 1)$. Consider a sequence of constant composition $(n, R_n)$ codes $\{(f_n, \varphi_n)\}_{n \geq 1}$ with the common composition $Q_n \in S_2(\delta)$ and $R_n := C(W) + \sqrt{\frac{V(W)}{n}} \Phi^{-1}(\epsilon)$. For some $n_0(W, \epsilon, \delta) \in \mathbb{Z}^+$, we have
\[
\bar{P}_e(f_n, \varphi_n) > \epsilon, \text{ for all } n \geq n_0(W, \epsilon, \delta). \tag{58}
\]

\[\diamondsuit\]

**Proof:** Define $\gamma(\delta) \in \mathbb{R}^+$ as
\[
\gamma(\delta) := C(W) - \max_{Q \in \mathcal{P}(\mathcal{X})} I(Q; W). \tag{59}
\]
For any message $m$, let
\[
G_n(m) := \left\{ y^n : \frac{1}{n} \sum_{i=1}^n \ln \frac{W(Y_i|x_0(m))}{q(Q_n(Y_i))} > I(Q_n; W) + \gamma(\delta) \right\}. \tag{60}
\]

\[\diamondsuit\] Since $I(\cdot, W)$ is continuous over $\mathcal{P}(\mathcal{X})$, $\gamma(\delta)$ is well-defined and positive.
Define $V_{\text{max}} := \max_{P \in \mathcal{P}(\mathcal{X})} V(P, W) \in \mathbb{R}^+$.

The following arguments are essentially the ones used in [39, Appendix B], which we outline here for completeness. We have

$$
\bar{P}_e(f_n, \varphi_n) = 1 - \frac{1}{|M_n|} \sum_{m \in M_n} \sum_{y^n \in A_m \cap G_n(m)} W(y^n|x^n(m)) - \frac{1}{|M_n|} \sum_{m \in M_n} \sum_{y^n \in A_m \cap G_n^c(m)} W(y^n|x^n(m)).
$$

(61)

Since $q_{Q_n}$ is a probability measure on $\mathcal{Y}^n$ and the decoding regions are disjoint, one can verify that

$$
\frac{1}{|M_n|} \sum_{m \in M_n} \sum_{y^n \in A_m \cap G_n(m)} W(y^n|x^n(m)) \leq e^{-n \left[ \frac{a}{2} + \sqrt{\frac{a n}{2}} \Phi^{-1}(\epsilon) \right]}.
$$

(62)

Moreover, via an application of Chebyshev’s inequality, it is easy to verify that

$$
\frac{1}{|M_n|} \sum_{m \in M_n} \sum_{y^n \in A_m \cap G_n^c(m)} W(y^n|x^n(m)) \leq \frac{V(Q; W)}{n^2} \leq \frac{4V_{\text{max}}}{n^2}.
$$

(63)

By plugging (62) and (63) into (61) and choosing $n_0(W, \epsilon, \delta) \in \mathbb{Z}^+$ such that for all $n \geq n_0(W, \epsilon, \delta)$, we have

$$
1 - e^{-n \left[ \frac{a}{2} + \sqrt{\frac{a n}{2}} \Phi^{-1}(\epsilon) \right]} - \frac{4V_{\text{max}}}{n^2} > \epsilon,
$$

we conclude that $\text{(58)}$ holds.

Lemma 6. Fix some $\epsilon \in (0, 0.5, 1)$, $W \in \mathcal{P}(\mathcal{Y}, \mathcal{X})$ with $V_c(W) > 0$, and $a \in \mathbb{R}^+$ with $a > \frac{2}{\epsilon}$. Consider an $(a, R_n)$ constant composition code $(f, \varphi)$ with $R_n = C(W) + \sqrt{\frac{V_c(W)}{n} \Phi^{-1}(\epsilon)} - \frac{\ln(1 - \epsilon)}{n}$ and the common composition $Q$ satisfying $V(Q, W) < V_c(W)\Phi^{-1}(\epsilon))^2$. We have

$$
\bar{P}_e(f, \varphi) > \epsilon.
$$

(64)

Proof: For any message $m$, define

$$
G_n(m) := \left\{ y^n : \frac{1}{n} \sum_{i=1}^n \ln \frac{W(y_i|x_i(m))}{q_n(Y_i)} > C(W) + \sqrt{\frac{V_c(W)}{n} \Phi^{-1}(\epsilon)} \right\}.
$$

(65)

Via arguments similar to the ones given in the proof of Lemma [5] one can verify that

$$
\bar{P}_e(f, \varphi) \geq 1 - \left( 1 - \epsilon - \frac{2}{a} \right) \frac{V(Q, W)}{n(C(W) - I(Q; W)) + \sqrt{nV_c(W)\Phi^{-1}(\epsilon)}}^2.
$$

(66)

$$
\geq \epsilon + \frac{a}{2} > \epsilon.
$$

(67)

For any $Q \in \mathcal{P}(\mathcal{X})$, define

$$
U(Q, W) := \sum_{x, y} Q(x)W(y|x) \left[ \ln \frac{W(y|x)}{q_Q(y)} - I(Q; W) \right]^2.
$$

(68)

$$
m_3(Q, W) := \sum_{x \in \mathcal{X}} Q(x)E_{W|x}[W(x)] \left[ \ln \frac{W(Y|x)}{q_Q(Y)} - E_{W|x}[Y] \right]^3.
$$

(69)

Choose $\delta > 0$ such that

$$
S(q_Q) = \mathcal{Y}, \text{ for all } Q \in \mathcal{P}(\mathcal{X}) \setminus S_3(\delta).
$$

(70)

Such a choice is possible due to the evident continuity of $x_y(\cdot)$ for any $y \in \mathcal{Y}$ and the fact that the unique capacity achieving output distribution has full support, as noted before. The following has been shown by Polyanskiy et al. [6, Lemma 46]

$$
m_3(Q, W) := \sum_{x \in \mathcal{X}} Q(x)W(y|x) \left[ \ln \frac{W(Y|x)}{q_Q(Y)} - I(Q; W) \right]^3 \leq \left( \frac{3(|\mathcal{X}|^{1/3} + |\mathcal{Y}|^{1/3})}{e} + \ln \min\{|\mathcal{X}|, |\mathcal{Y}|\} \right)^3 =: \kappa(W) \in \mathbb{R}^+.
$$

(71)
Fix some $\nu \in \mathbb{R}^+$ and $\epsilon \in (0, 1)$. Assume $S_1(\delta, \nu) \neq \emptyset$ and define\footnote{Since $m_3(\cdot, W)$ and $V(\cdot, W)$ are continuous over $\mathcal{P}(\mathcal{X})$ (e.g., Lemma 62), $K(W, \epsilon, \delta, \nu)$ is well-defined, positive and finite.}

$$K(W, \epsilon, \delta, \nu) := \frac{2}{\phi(\Phi^{-1}(\epsilon))} \left[ \max_{P \in S_1(\delta, \nu)} \frac{m_3(P, W)}{V(P, W)} + \left( \frac{1}{\sqrt{2\pi}} + \frac{\kappa(W)}{\nu} \right) \right] \in \mathbb{R}^+. \quad (72)$$

**Lemma 7.** Fix an asymmetric and singular $W \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$, $\epsilon \in (0, 1)$ and $\nu \in \mathbb{R}^+$. Choose $\delta \in \mathbb{R}^+$ such that \footnote{This holds. For some $\bar{n}_o(W, \epsilon, \delta, \nu) \in \mathbb{Z}^+$ and any $n \geq \bar{n}_o(W, \epsilon, \delta, \nu)$, consider an $(n, R_n)$ constant composition code $(f, \varphi)$ with common composition $Q \in S_1(\delta, \nu)$ and $R_n = I(Q, W) + \sqrt{\frac{V(Q, W)}{n}} \Phi^{-1}(\epsilon) + \frac{K(W, \epsilon, \delta, \nu)}{n}$. We have\footnote{In light of (72), the existence of such a choice is evident.}

$$\bar{P}_\epsilon(f, \varphi) > \epsilon. \quad (73)$$

\begin{itemize}
  \item **Proof:** Assume $S_1(\delta, \nu) \neq \emptyset$, because otherwise the claim is void. The proof is similar to the proof of Proposition\footnote{Consider any $(n, R_n)$ constant composition code, say $(f, \varphi)$, with the common composition $Q$. Assume $Q$ and $R_n$ are as in the statement of the lemma. Consider any $x^n \in \mathcal{X}^n$ and define

$$M_{\mathcal{X}_n}(\lambda) := \mathbb{E}_{\mathcal{W}(x^n)} \left[ e^{\lambda \ln \frac{W(y^n|x^n)}{q^n_{\alpha_\lambda}}} \right], \forall \lambda \in \mathbb{R}. \quad (75)$$

We claim that for any $x^n, z^n \in \mathcal{X}^n$ with $P_{x^n} = P_{z^n}$, we have

$$M_{\mathcal{X}_n}(\lambda) = M_{\mathcal{Z}_n}(\lambda), \forall \lambda \in \mathbb{R}. \quad (76)$$

To see this, simply note that

$$M_{\mathcal{X}_n}(\lambda) = \sum_{y^n : W(y^n|x^n) > 0} e^{\lambda \ln \frac{W(y^n|x^n)}{q^n_{\alpha_\lambda}}} \phi(\Phi^{-1}(\epsilon)) \left[ \frac{m_3(P, W)}{V(P, W)} + \frac{1}{\sqrt{2\pi}} + \frac{\kappa(W)}{\nu} \right] \in \mathbb{R}^+. \quad (72)$$

\end{itemize}
Further, define $P_{X,Y}(x, y) := Q(x)W(y|x)$ and $P_{X,Y}(x^n, y^n) := \prod_{i=1}^n P_{X,Y}(x_i, y_i)$. Evidently,

$$
\sum_{y^n \in S, (x^n, y^n)} q_Q(y^n) e^{-n\left[R_n - \frac{1}{n} \sum_{i=1}^n \ln \frac{1}{q_Q(y_i^n)}\right]} = \sum_{(x^n, y^n)} P_{X,Y}(x^n, y^n) \mathbb{I}\left\{\frac{1}{n} \sum_{i=1}^n \ln \frac{W(y_i|x_i)}{q_Q(y_i)} \leq R_n\right\} e^{-n\left[R_n - \frac{1}{n} \sum_{i=1}^n \ln \frac{W(y_i|x_i)}{q_Q(y_i)}\right]}
$$

(84)

$$
\leq \frac{1}{\sqrt{2\pi nU(Q, W)}} + \frac{\tilde{m}_3(Q, W)}{\sqrt{n}U(Q, W)^{3/2}} 
\leq \frac{1}{\sqrt{2\pi nV(Q, W)}} \left(\frac{1}{\sqrt{2\pi nV(Q, W)}} + \kappa(W)\right),
$$

(85)

(86)

where (85) follows from Lemma 3 whose application is ensured by the fact that $U(Q, W) \geq V(Q, W)$ (e.g., [6 Lemma 62]), which, along with (71), also implies (86).

By plugging (83) and (86) into (81), along with the definitions of $K(W, \epsilon, \delta, \nu)$ and $n_o(W, \epsilon, \delta, \nu)$, one can verify that

$$
P_\epsilon(f, \varphi) > \epsilon + \frac{1}{\sqrt{nV(Q, W)}} \left(\max_{P \in S_1(\delta, \nu)} \frac{m_3(P, W)}{V(P, W)} - \frac{m_3(Q, W)}{V(Q, W)}\right) \geq \epsilon,
$$

(87)

which, in turn, implies (73).

To prove item (i) of Proposition 2, fix some $\epsilon \in (0, 0.5)$ and assume $V_\epsilon(W) > 0$, because otherwise [32 Proposition 9] implies (16). Fix some $\delta > 0$ such that (70) holds and $S_2(\delta, V_\epsilon(W)/2) = \emptyset$. Such a choice is possible since $V_\epsilon(W)$ is continuous over $P(\mathcal{X})$, as noted before. For any $P \in P(\mathcal{X})$, let $P^*(P) := \arg \min_{Q \in P_\epsilon} ||P - Q||_2$. Fix some $\beta_1, \beta_2 \in \mathbb{R}^+$ such that

$$
I(P; W) \leq C(W) - \beta_1 ||P - P^*(P)||^2_2, \quad \left|\sqrt{V(P, W)} - \sqrt{V(P^*(P), W)}\right| \leq \beta_2 ||P - P^*(P)||_2,
$$

(88)

for any $P \in S_1(\delta, V_\epsilon(W)/2)$, whose existence is ensured by [32 Lemma 7]. In light of (88), for all $P \in S_1(\delta, V_\epsilon(W)/2)$ and for any $n \in \mathbb{Z}^+$,

$$
nI(P; W) + \sqrt{nV(P, W)}\Phi^{-1}(\epsilon) \leq nC(W) + \sqrt{nV_\epsilon(W)}\Phi^{-1}(\epsilon) - \beta_1 n||P - P^*(P)||^2_2 + \beta_2 |\Phi^{-1}(\epsilon)|\sqrt{n}||P - P^*(P)||_2
$$

(89)

$$
\leq nC(W) + \sqrt{nV_\epsilon(W)}\Phi^{-1}(\epsilon) + \frac{(||\beta_2|\Phi^{-1}(\epsilon)||)^2}{4\beta_1},
$$

(90)

where (90) follows from elementary calculus. Consider any $n \in \mathbb{Z}^+$ such that

$$
n \geq \max\{n_o(W, \epsilon, \delta, \tilde{n}_o(W, \epsilon, \delta, V_\epsilon(W)/2)\},
$$

(91)

where $n_o$ and $\tilde{n}_o$ are as given in Lemmas 5 and 7 respectively. Define

$$
R_n := C(W) + \sqrt{\frac{V_\epsilon(W)}{n}}\Phi^{-1}(\epsilon) + \frac{(||\beta_2|\Phi^{-1}(\epsilon)||)^2}{4\beta_1} + K(W, \epsilon, \delta, V_\epsilon(W)/2),
$$

(92)

and consider any $(n, R_n)$ constant composition code $(f, \varphi)$ with the common composition $Q$. Now, if $Q \in S_3(\delta)$, then Lemma 5 implies that $P_\epsilon(f, \varphi) > \epsilon$. Similarly, if $Q \in S_1(\delta, V_\epsilon(W)/2)$, then Lemma 7 and (90) implies that $P_\epsilon(f, \varphi) > \epsilon$. Since the code is arbitrary, we conclude the proof of item (i) of the proposition.

To prove item (ii) of Proposition 2, fix some $\epsilon \in (0.5, 1)$ and $\delta > 0$ such that (70) holds. Choose some $\alpha \in \mathbb{R}^+$ that satisfies $\alpha > \frac{2}{\epsilon}$ and $\nu \in \mathbb{R}^+$ such that $\nu \leq \frac{V_\epsilon(W)|\Phi^{-1}(\epsilon)|^2}{\alpha}$. Similar to (88), choose $\beta_1, \beta_2 \in \mathbb{R}^+$ such that

$$
I(P; W) \leq C(W) - \beta_1 ||P - P^*(P)||^2_2, \quad |\sqrt{V(P, W)} - \sqrt{V(P^*(P), W)}| \leq \beta_2 ||P - P^*(P)||_2,
$$

(93)

for any $P \in S_1(\delta, \nu)$. From (93), similar to (90), we deduce that for all $P \in S_1(\delta, \nu)$ and $n \in \mathbb{Z}^+$,

$$
nI(P; W) + \sqrt{nV(P, W)}\Phi^{-1}(\epsilon) \leq nC(W) + \sqrt{nV_\epsilon(W)}\Phi^{-1}(\epsilon) + \frac{(||\beta_2|\Phi^{-1}(\epsilon)||)^2}{4\beta_1}.
$$

(94)

Consider any $n \in \mathbb{Z}^+$ such that $n \geq \max\{n_o(W, \epsilon, \delta, \tilde{n}_o(W, \epsilon, \delta, \nu))\}$, where $n_o$ and $\tilde{n}_o$ are as given in Lemmas 5 and 7 respectively. Consider any $(n, R_n)$ constant composition code $(f, \varphi)$ with the common composition $Q$ and define

$$
R_n := C(W) + \sqrt{\frac{V_\epsilon(W)}{n}}\Phi^{-1}(\epsilon) + \frac{(||\beta_2|\Phi^{-1}(\epsilon)||)^2}{4\beta_1} + K(W, \epsilon, \delta, \nu) - \ln \left(1 - \frac{2}{\alpha}\right).
$$

(95)

If $Q \in S_3(\delta)$, then $P_\epsilon(f, \varphi) > \epsilon$ due to Lemma 5. If $Q \in S_2(\delta, \nu)$, then $P_\epsilon(f, \varphi) > \epsilon$ because of Lemma 6. Finally, if $Q \in S_1(\delta, \nu)$, then Lemma 7 along with (94), implies that $P_\epsilon(f, \varphi) > \epsilon$. Since the code is arbitrary, we conclude that (17) holds.
IV. Discussion

A. Relation to the minimax converse

One can interpret the arguments leading to the proof of Proposition 1 in terms of the minimax converse (e.g., [22, Theorem 1]), which we illustrate next. To this end, we fix a symmetric and singular $W \in \mathcal{P}(Y|X)$ and note that [22, Eq. (9) and (11)] imply that for any $n \in \mathbb{Z}^+$ and $\epsilon \in (0, 1),$

$$M^*(n, \epsilon) \leq \min_{P_{X^n}} \max_{Q_{Y^n}} \frac{1}{\beta_1 - \epsilon P_{X^n, Y^n}, P_{X^n} \times Q_{Y^n}},$$

(96)

where $P_{X^n, Y^n}(x^n, y^n) := P_{X^n}(x^n)W(y^n|x^n)$, $(P_{X^n} \times Q_{Y^n})(x^n, y^n) := P_{X^n}(x^n)Q_{Y^n}(y^n)$ and $\beta_1 - \epsilon (P_{X^n, Y^n}, P_{X^n} \times Q_{Y^n})$ denotes the minimum probability of error under $P_{X^n} \times Q_{Y^n}$, subject to the constraint that the error probability under hypothesis $P_{X^n, Y^n}$ does not exceed $\epsilon$. Due to [22, Theorem 21], the minimum on the right side of (96) is attained by $U_{Y^n}$. Consider some $n \in \mathbb{Z}^+$ such that (46) holds and let $R$ be as in (47). With these choices, we define

$$Q_{Y^n}(y^n) := e^{n \sum_y P_{Y^n}(y) \ln \delta_y} \mathbb{1} \{y^n \in S_R\}$$

(97)

where $\delta_y$ and $S_R$ are as defined before. Evidently, $Q_{Y^n}^* \in \mathcal{P}(Y^n)$. With a slight abuse of notation, let $\beta_1 - \epsilon (U_{X^n}, Q_{Y^n})$ denote the cost function of the optimization problem in the denominator of (96) when $P_{X^n} = U_{X^n}$ and $Q_{Y^n} = Q_{Y^n}^*$. Evidently,

$$M^*(n, \epsilon) \leq \frac{1}{\beta_1 - \epsilon (U_{X^n}, Q_{Y^n}^*)}.$$

(98)

From the Neyman-Pearson Lemma (e.g., [40]), the right side of (98) is attained by a randomized threshold test with the randomization parameter $\tau \in (0, 1)$ that satisfies

$$\tau W(S_R|X^n_0) = \epsilon \quad \text{and} \quad \beta_1 - \epsilon (U_{X^n}, Q_{Y^n}^*) = \frac{(1 - \tau)W(S_R|X^n_0)}{e^{nR} \sum_{y^n \in S_R} q(y^n) e^{-n \left[ R - \frac{1}{2} \sum_{i=1}^{n} \ln \frac{1}{\pi_{yi}} \right]}}.$$

(99)

Equation (99) can be verified via elementary algebra by noticing the fact that $W$ is singular and symmetric, and we omit the details for brevity. Finally, (49) and (50), along with (46) and (47), imply that

$$W(S_R|X^n_0) - \sum_{y^n \in S_R} q(y^n) e^{-n \left[ R - \frac{1}{2} \sum_{i=1}^{n} \ln \frac{1}{\pi_{yi}} \right]} > \epsilon.$$

(100)

Equations (98), (99) and (100) imply that $M^*(n, \epsilon) < e^{nR}$, which, in turn, implies Proposition 1.

In light of the above discussion, the proof of Proposition 1 would be shorter had we used the minimax converse with the output distribution given in (77). However, we opt to use Lemma 2 because it makes the role of $S_R$ more transparent, as explained in Remark 5.

B. On dropping the constant composition assumption

As noted before, Proposition 2 gives an $O(1)$ upper bound on the third-order term of the normal approximation for asymmetric and singular DMCs only if we consider constant composition codes. Although this restriction is undesirable, it is quite common in converse results. Indeed, the usual proof of the converse statement of (1) involves first showing it for constant composition codes, and then arguing that this restriction at most results in an extra $O(\ln n)$ term.

It should be noted that if the channel has sufficient symmetry, then the constant composition step is not necessary and one can derive an $\ln \sqrt{n}$ upper bound on the third-order term [7, Sec. 3.4.5]. Recently, Tomamichel-Tan [32] have shown an $\ln \sqrt{n}$ upper bound on the third-order term in general by dispensing with the constant composition code restriction in the first step. This result, coupled with the existing results in the literature, gives the third-order term for a broad class of channels, which includes positive channels with positive capacity [32], but does not include asymmetric and singular channels. The method of [32] is essentially based on relating the channel coding problem to a binary hypothesis test by using an auxiliary output distribution, which is in the same vein as the so-called meta-converse of Polyanskiy et al. (e.g., [6, Section III.E and III.F]). As opposed to the classical applications of this idea, which use a product auxiliary output distribution and result in the aforementioned two-step procedure, the authors of [32] uses an appropriately chosen non-product output distribution to dispense with the constant composition step. However, their non-product distribution is different from the one used in the previous subsection and it is an interesting future research topic to investigate how to combine the analysis of [32] and the viewpoint in Section IV-A to drop the constant composition assumption in Proposition 1.

12The following non-product distribution is inspired by [22, Eq. (168)]. In particular, if $W$ is BEC then (97) reduces to [22, Eq. (168)].
C. Limitation in the error exponents regime

One might conjecture that by following the same program used to prove Proposition 2, one could prove the following lower bound for asymmetric and singular channels

\[ \liminf_{n \to \infty} \frac{P_{e,n}(n, R)}{\sqrt{n}} \geq K(R, W), \]  

where \( K(R, W) \) is a positive constant that depends on \( R \) and \( W \), and \( \text{ESP}(R, W) \) is the sphere-packing exponent (e.g., Eq. (5.8.2))

\[ \text{ESP}(R, W) := \max_{Q \in \mathcal{P}(\mathcal{X})} \text{ESP}(R, Q, W), \quad \text{ESP}(R, Q, W) := \sup_{\rho \geq 0} \left\{ -\rho R - \ln \sum_{y \in \mathcal{Y}} \left( \sum_{x \in \mathcal{X}} Q(x) W(y|x)^{1/(1+\rho)} \right)^{1+\rho} \right\}. \]  

However, a proof of (101) seems to be more involved than its counterpart in the normal approximation regime, i.e., Proposition 2. The main technical difficulty is proving the continuity properties of \( \text{ESP}(R, \cdot, W) \) that are required to distinguish between the “good types”, for which \( \text{ESP}(R, Q, W) \approx \text{ESP}(R, W) \) and hence one can use a result like Lemma 7 to deduce an \( \Omega(1/\sqrt{n}) \) sub-exponential term directly, and the “bad types”, for which \( \text{ESP}(R, Q, W) \) is bounded away from \( \text{ESP}(R, W) \) and hence one can utilize this inferiority of the exponent to deduce an \( \Omega(1/\sqrt{n}) \) sub-exponential term. Indeed, the proof of these continuity properties appears to be quite intricate.

ACKNOWLEDGMENT

The authors thank Sergio Verdú for raising the question of whether the technique used in [33] could be applied to the normal approximation regime. The first author thanks Paul Cuff for his hospitality while portions of this work were being completed during his visit to Princeton University. This research is supported by the National Science Foundation under grant CCF-1218578.

REFERENCES

[1] A. A. Yushkevich, “On limit theorems connected with the concept of entropy of Markov chains,” (in Russian) Uspekhi Mat. Nauk, vol. 8, no. 5(57), pp. 177–180, 1953.
[2] V. Strassen, “Asymptotische abschätzungen in Shannon’s informationstheorie,” Trans. Third Prague Conf. Information Theory, 1962, Czechoslovak Academy of Sciences, Prague, pp. 689–723.
[3] M. Hayashi, “Information spectrum approach to second-order coding rate in channel coding,” IEEE Trans. on Information Theory, vol. IT 55, no. 11, pp. 4947–4966, Nov. 2009.
[4] L. Wang, R. Renner, and R. Colbeck, “Simple channel coding bounds,” in Proc. 2009 IEEE Int. Symp. Inf. Theory, Seoul, S. Korea, Jul. 2009, pp. 1804–1808.
[5] Y. Polyanskiy, H. V. Poor and S. Verdú, “Dispersion of Gaussian channels,” in Proc. 2009 IEEE Int. Symp. Inf. Theory, Seoul, S. Korea, Jul. 2009, pp. 2204–2208.
[6] Y. Polyanskiy, H. V. Poor and S. Verdú, “Channel coding rate in the finite blocklength regime,” IEEE Trans. on Information Theory, vol. IT–56, no. 5, pp. 2307–2359, May 2010.
[7] Y. Polyanskiy, H. V. Poor and S. Verdú, “Channel coding: non-asymptotic fundamental limits,” Ph.D. dissertation, Princeton Univ., Princeton, NJ, Nov. 2010.
[8] Y. Polyanskiy and S. Verdú, “Ariimoto channel coding converse and Rényi divergence,” in Proc. 48th Annual Allerton Conf. Communications, Control, and Computing, Monticello, IL, Oct. 2010, pp. 1327–1333.
[9] Y. Polyanskiy, H. V. Poor and S. Verdú, “Dispersion of the Gilbert-Elliott channel,” IEEE Trans. on Information Theory, vol. IT–57, no. 4, pp. 1829–1848, Apr. 2011.
[10] A. Martinez and A. Guillén i Fàbregas, “Random-coding bounds for threshold decoders: error exponent and saddlepoint approximation,” in Proc. 2011 IEEE Int. Symp. Inf. Theory, St. Petersburg, Russia, Aug. 2011, pp. 2899–2903.
[11] Y. Polyanskiy and S. Verdú, “Scalar coherent fading channel: dispersion analysis,” in Proc. 2011 IEEE Int. Symp. Inf. Theory, St. Petersburg, Russia, Aug. 2011, pp. 2959–2963.
[12] P. N. Chen, H. Y. Lin and S. M. Moser, “Ultra-small block-codes for binary discrete memoryless channels,” in Proc. IEEE Inf. Theory Workshop, Paraty, Brazil, Oct. 2011, pp. 175–179.
[13] T. J. Kiedl, T. J. Coleman and A. C. Singer, “Finite block-length achievable rates for queuing timing channels,” in Proc. IEEE Inf. Theory Workshop, Paraty, Brazil, Oct. 2011, pp. 200–204.
[14] J. Hoydis, J. Coullet, R. Piantanida and M. Debbah, “A random matrix approach to the finite blocklength regime of MIMO fading channels,” in Proc. 2012 IEEE Int. Symp. Inf. Theory, Boston, MA, Jul. 2012, pp. 2181–2185.
[15] P. Moulin, “The log-volume of optimal constant-composition codes for memoryless channels, within O(1) bits,” in Proc. 2012 IEEE Int. Symp. Inf. Theory, Boston, MA, Jul. 2012, pp. 826–830.
[16] L. R. Varshney, S. J. Mitter and V. K. Goyal, “An information-theoretic characterization of channels that die,” IEEE Trans. on Information Theory, vol. IT–58, no. 9, pp. 5711–5724, Sep. 2012.
[17] W. Yang, W. Durisi, G. Koch and Y. Polyanskiy, “Diversity versus channel knowledge at finite block-length,” in Proc. IEEE Inf. Theory Workshop, Lausanne, Switzerland, Sep. 2012, pp. 572–576.
[18] Y. Polyanskiy and S. Verdú, “Empirical distribution of good channel codes with non-vanishing error probability,” submitted to IEEE Trans. on Information Theory.

Available from: http://people.lids.mit.edu/yp/homepage/data/optcodes_journal.pdf
[19] Y. Polyanskiy and Y. Wu, “Peak-to-average power ratio of good codes for Gaussian channel,” submitted to IEEE Trans. on Information Theory. Available from: http://arxiv.org/pdf/1302.0084v1.pdf
[20] A. Ingber, R. Zamir and M. Feder, “Finite-dimensional infinite constellations,” IEEE Trans. on Information Theory, vol. IT–59, no. 3, pp. 1630–1656, Mar. 2013.
[21] J. Scarlett, A. Martínez, A. Guillén i Fàbregas, “Mismatched decoding: finite-length bounds, error exponents and approximations,” submitted to IEEE Trans. on Information Theory. Available from: http://arxiv.org/pdf/1303.6166v1.pdf
[22] Y. Polyanskiy, “Saddle point in the minimax converse for channel coding,” *IEEE Trans. on Information Theory*, vol. IT–59, no. 5, pp. 2576–2595, May 2013.

[23] E. Haim, Y. Kochman and U. Erez, “The importance of tie-breaking in finite-blocklength bounds,” in *Proc. 2013 IEEE Int. Symp. Inf. Theory*, Istanbul, Turkey, Jul. 2013.

[24] J. Hoydis, R. Couillet and P. Piantanida, “Bounds on the second-order coding rate of the MIMO Rayleigh block-fading channel,” in *Proc. 2013 IEEE Int. Symp. Inf. Theory*, Istanbul, Turkey, Jul. 2013.

[25] V. Kostina and S. Verdú, “Channels with cost constraints: strong converse and dispersion,” in *Proc. 2013 IEEE Int. Symp. Inf. Theory*, Istanbul, Turkey, Jul. 2013.

[26] P. Moulin, “Asymptotic Neyman-Pearson games for converse to the channel coding theorem,” in *Proc. 2013 IEEE Int. Symp. Inf. Theory*, Istanbul, Turkey, Jul. 2013.

[27] M. Raginsky and I. Sason, “Refined bounds on the empirical distribution of good channel codes via concentration inequalities,” in *Proc. 2013 IEEE Int. Symp. Inf. Theory*, Istanbul, Turkey, Jul. 2013.

[28] Y. Y. Shkel, V. Y. F. Tan and S. C. Draper, “Converse bounds for assorted codes in the finite blocklength regime,” in *Proc. 2013 IEEE Int. Symp. Inf. Theory*, Istanbul, Turkey, Jul. 2013.

[29] G. Vazquez-Vilar, A. Tauste Campo, A. Guillén i Fàbregas and A. Martinez, “The meta-converse bound is tight,” in *Proc. 2013 IEEE Int. Symp. Inf. Theory*, Istanbul, Turkey, Jul. 2013.

[30] M. Tomamichel and V. Y. F. Tan, “ǫ-capacities and second-order coding rates for channels with general state,” submitted to *IEEE Trans. on Information Theory*. Available from: http://arxiv.org/pdf/1305.6789v1.pdf

[31] W. Yang, G. Durisi, T. Koch and Y. Polyanskiy, “Quasi-static SIMA fading channels at finite blocklength,” in *Proc. 2013 IEEE Int. Symp. Inf. Theory*, Istanbul, Turkey, Jul. 2013.

[32] M. Tomamichel and V. Y. F. Tan, “A tight upper bound for the third-order asymptotics for most discrete memoryless channels,” to appear *IEEE Trans. on Information Theory*.

[33] Y. Altuğ and A. B. Wagner, “Exact asymptotics of the error probability in channel coding: symmetric channels,” to be submitted to *IEEE Trans. on Information Theory*.

[34] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic Press, 1981.

[35] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[36] Y. Altuğ and A. B. Wagner, “Refinement of the random coding bound,” to be submitted to *IEEE Trans. on Information Theory*.

[37] V. Yu. Korolev and I. G. Shevtsova, “A new moment-type estimate of convergence rate in the Lyapunov theorem,” *Theory Probab. Appl.* vol. 55, no. 3, pp. 505–509, 2011.

[38] P. Billingsley, *Probability and measure, 3rd edition*. Hoboken, NJ: Wiley, 1995.

[39] Y. Altuğ and A. B. Wagner, “Moderate deviation analysis of channel coding,” submitted to *IEEE Trans. on Information Theory*. Available from: http://arxiv.org/pdf/1208.1924v1.pdf

[40] J. Neyman and E. S. Pearson, “On the problem of the most efficient tests of statistical hypothesis,” *Philos. Trans. Roy. Soc. London. Ser. A*, vol. 231, pp. 289–337, 1933.

[41] C. E. Shannon, R. G. Gallager and E. R. Berlekamp, “Lower bounds to error probability for coding on discretememoryless channels,” *Inform. Contr.*, vol. 10, pp. 65–103, Jan. 1967

[42] Y. Altuğ and A. B. Wagner, “Refinement of the sphere-packing bound: asymmetric channels,” submitted to *IEEE Trans. on Information Theory*. Available from: http://arxiv.org/pdf/1211.6697v1.pdf