Refinements of the Third-Order Term in the Fixed Error Asymptotics of Constant-Composition Codes

Jonathan Scarlett
University of Cambridge
jms265@cam.ac.uk

Alfonso Martinez
Universitat Pompeu Fabra
alfonso.martinez@ieee.org

Albert Guillén i Fàbregas
ICREA & Universitat Pompeu Fabra
University of Cambridge
guillen@ieee.org

***Abstract*—This paper studies the fixed-error asymptotics of constant-composition codes for discrete memoryless channels. An achievable asymptotic expansion is derived with a third-order term that can be as high as $\frac{1}{2}\log n$, while being lower when (i) a certain feasibility-decoding condition fails, or (ii) the channel is a sum channel. Converse bounds are used to provide conditions under which each of these losses is unavoidable.**

I. Introduction

Fixed-error asymptotic studies of channel coding have recently regained significant interest following the works of Polyanskiy *et al.* [1] and Hayashi [2]. For discrete memoryless channels (DMCs), the highest number of messages $M^*(n,\epsilon)$ for a given block length n and error probability ϵ satisfies [3]

$$\log M^*(n,\epsilon) = nC - \sqrt{nV}\mathsf{Q}^{-1}(\epsilon) + o(\sqrt{n}), \tag{1}$$

where $\mathsf{Q}^{-1}(\cdot)$ is the inverse of the Q-function, C is the channel capacity, and V is known as the channel dispersion. While the third-order term is not yet completely characterized, it equals $\frac{1}{2}\log n$ for a wide class of channels satisfying a *non-singularity* or *feasibility decoding is suboptimal (FDIS)* condition [1], [4], [5], and $O(1)$ for a class of symmetric channels failing that condition [6].

In this paper, we present refined results on the third-order term for constant-composition codes, where each codeword has the same empirical distribution (i.e. type [7, Ch. 2]). Beyond their theoretical value, such codes are of particular interest for channels with cost constraints [8, Ch. 7], mismatched decoding settings [9], and multiuser settings [10], [11].

A. System Setup

Henceforth, the i-th entry of a vector (e.g. $\boldsymbol{x}$) is written using a subscript (e.g. x_i). The empirical distribution (i.e. type [7, Ch. 2]) of a vector $\boldsymbol{x}$ is denoted by $\hat{P}_{\boldsymbol{x}}$, and the set of all sequences of type Q is denoted by $T^n(Q)$.

The input and output alphabets are denoted by $\mathcal{X}$ and $\mathcal{Y}$ respectively, and are assumed to be finite. The channel transition law is denoted by $W(y|x)$, and we write $W^n(\boldsymbol{y}|\boldsymbol{x}) \triangleq \prod_{i=1}^n W(y_i|x_i)$. We fix an input distribution Q and let Q_n be a type such that $\|Q - Q_n\| \le \frac{1}{n}$, and $Q_n(x) = 0$ wherever $Q(x) = 0$. We assume that $Q(x) > 0$ for all x; this is without loss of generality for the results we provide, since in the general case the remaining inputs can be removed. We similarly assume that all outputs are reachable.

We say that $\mathcal{C} = \{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(M)}\}$ is a constant-composition codebook with input distribution Q if $\boldsymbol{x}^{(m)} \in T^n(Q_n)$ for all m. We let $M_Q^*(n,\epsilon)$ denote the highest number of messages of any constant-composition code for a given input distribution Q, block length n, and target error probability ϵ. We prove asymptotic lower bounds on $M_Q^*(n,\epsilon)$ using constant-composition random coding, where the codewords are randomly drawn from the distribution

$$P_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{|T^n(Q_n)|}\mathbb{1}\big\{\boldsymbol{x} \in T^n(Q_n)\big\}. \tag{2}$$

We define the information density

$$i(x,y) \triangleq \log \frac{W(y|x)}{\sum_{\overline{x}} Q(\overline{x})W(y|\overline{x})} \tag{3}$$

and its n-letter extension

$$i^n(\boldsymbol{x},\boldsymbol{y}) \triangleq \sum_{i=1}^n i(x_i,y_i). \tag{4}$$

The mean and conditional variance of $i(X,Y)$ are denoted by

$$I(Q) \triangleq \mathbb{E}[i(X,Y)] \tag{5}$$

$$V(Q) \triangleq \mathbb{E}\big[\mathrm{Var}[i(X,Y)\,|\,X]\big], \tag{6}$$

where $(X,Y) \sim Q \times W$. Observe that $I(Q)$ is simply the mutual information under $Q \times W$.

B. Previous Results and Contributions

Polyanskiy *et al.* [1] provided a converse bound on $M_Q^*(n,\epsilon)$ of the form (1) with $I(Q)$ and $V(Q)$ in place of C and V, and with a $\frac{1}{2}\log n$ third-order term. A matching second-order achievability result was given by Hayashi [2] with a third-order term of $o(\sqrt{n})$, and by Kostina-Verdú [12] with a third-order term of $-\frac{1}{2}(|\mathcal{X}|-1)\log n$. Moulin [13] derived a third-order term of $\frac{1}{2}\log n$ under various assumptions, and gave bounds on the fourth-order term. Converse results for general codes were presented in [4], [6], and refined results for the Gaussian channel were given in [1], [14].

The contributions of this paper are as follows: (i) We derive a third-order term of $\frac{1}{2}\log n$ for constant-composition codes using technical assumptions and analysis techniques differing from those of [13]; (ii) We study the reduction in the third-order term when these technical assumptions fail. In particular, we show that a loss in the third-order term for sum channels [15] is unavoidable in general.

II. Main Results

We define the following set of channels from $\mathcal{X}$ to $\mathcal{Y}$:

$$\mathcal{W}_0^{\text{cc}} \triangleq \Big\{W : \text{There exist functions } g(x), h(y) \text{ s.t. } W(y|x) = g(x)h(y) \text{ wherever } W(y|x) > 0\Big\}. \quad (7)$$

It was shown in [9] that if $W \in \mathcal{W}_0^{\text{cc}}$, then feasibility decoding (i.e. searching for a unique codeword whose likelihood is positive) is optimal for constant-composition codes. Thus, the condition $W \notin \mathcal{W}_0^{\text{cc}}$ is a natural generalization of the *feasibility decoding is suboptimal* (FDIS) condition from [5], which requires a similar condition with $g(x)$ equal to a constant (i.e. only a factor depending on y is allowed). Due to the additional structure of constant-composition codes, there is a wider class of channels for which feasibility decoding is optimal. For later reference, we define $\mathcal{W}_0$ in the same way as (7) with $g(x)$ replaced by 1; clearly $\mathcal{W}_0 \subseteq \mathcal{W}_0^{\text{cc}}$.

Our results also vary depending on whether W is a sum channel according to the following definition.

Definition 1. *A discrete memoryless channel $W(y|x)$ is said to be a sum of N subchannels $W_1, \dots, W_N$ if the alphabets $\mathcal{X}$ and $\mathcal{Y}$ can be partitioned into disjoint non-empty subsets $\mathcal{X}_1, \dots, \mathcal{X}_N$ and $\mathcal{Y}_1, \dots, \mathcal{Y}_N$ such that*

$$W(y|x) = \begin{cases} W_i(y|x) & x \in \mathcal{X}_i \text{ and } y \in \mathcal{Y}_i \, (i \in \{1, \dots, N\}) \\ 0 & \text{otherwise,} \end{cases} \quad (8)$$

and the same is not true for any higher value of N. We say that $W(y|x)$ is a sum channel if it is a sum of $N \geq 2$ subchannels.

The main result of this paper is the following.

Theorem 1. *Fix an input distribution Q such that $Q(x) > 0$ for all x. If W is the sum of N subchannels, then we have for any $\epsilon \in (0, 1)$ that*

$$\log M_Q^*(n, \epsilon) \geq nI(Q) - \sqrt{nV(Q)}\, \mathsf{Q}^{-1}(\epsilon) + \frac{1}{2}\big(\mathbb{1}\{W \notin \mathcal{W}_0^{\text{cc}}\} - (N-1)\big) \log n + O(1). \quad (9)$$

Proof: See Section III-A. ∎

In the special case that W is not a sum channel (i.e. $N = 1$), the coefficient to $\log n$ in (9) is $\frac{1}{2}$ when $W \notin \mathcal{W}_0^{\text{cc}}$, and 0 when $W \in \mathcal{W}_0^{\text{cc}}$. As mentioned above, a matching converse for the case $W \notin \mathcal{W}_0^{\text{cc}}$ was given in [1]. A matching converse for the case $W \in \mathcal{W}_0$ with an optimized input distribution was provided in [6, Prop. 2], but it is unclear whether our result is tight for $W \in \mathcal{W}_0^{\text{cc}} \backslash \mathcal{W}_0$. It is also unclear in general whether our result is tight for sum channels, but the following theorem provides cases in which the answer is affirmative. Recall that $M^*(n, \epsilon)$ is defined in the same way as $M_Q^*(n, \epsilon)$, but without the restriction to constant-composition codes.

Theorem 2. *Fix an input distribution Q such that $Q(x) > 0$ for all x. If W is the sum of N subchannels, each having an identical transition law, then we have for all $\epsilon \in (0, 1)$ that*

$$\log M_Q^*(n, \epsilon) \leq \log M^*(n, \epsilon) - \frac{1}{2}(N-1) \log n + O(1). \quad (10)$$

Proof: See Section III-B ∎

Theorem 2 can be combined with existing converse results to prove the tightness of (9) for a variety of sum channels with identical subchannels. For example, if $W \notin \mathcal{W}_0^{\text{cc}}$ and the capacity-achieving input distribution is unique, the tightness follows from [4], whereas if W is symmetric and $W \in \mathcal{W}_0$, the tightness follows from [6].

To our knowledge, the preceding observations provide the first known cases in which i.i.d. random coding provably outperforms constant-composition random coding, since the former yields the optimal third-order expansion of $M^*(n, \epsilon)$ [1], [6]. As mentioned in the introduction, the opposite is often true for mismatched and multi-user settings.

III. Proofs

We assume without loss of generality that $\mathcal{X} = \{1, \dots, |\mathcal{X}|\}$ and $\mathcal{Y} = \{1, \dots, |\mathcal{Y}|\}$. Furthermore, we present the proof assuming that Q is itself a type, and thus $Q_n = Q$; the results for the general case follow using the fact that $\|Q - Q_n\| = O\big(\frac{1}{n}\big)$. In particular, we have $\|I(Q) - I(Q_n)\| = O\big(\frac{1}{n}\big)$ since the mutual information is continuously differentiable in Q whenever $Q(x) > 0$ for all x.

Some details of the proof are omitted due to space limitations, and can be found in [16].

A. Proof of Theorem 1

Our analysis starts with the random-coding union (RCU) bound [1], which upper bounds the error probability by

$$\text{rcu}(n, M) \triangleq \mathbb{E}\Big[\min\Big\{1, (M-1)\mathbb{P}\big[i^n(\overline{\boldsymbol{X}}, \boldsymbol{Y}) \geq i^n(\boldsymbol{X}, \boldsymbol{Y}) \,\big|\, \boldsymbol{X}, \boldsymbol{Y}\big]\Big\}\Big], \quad (11)$$

where $(\boldsymbol{X}, \boldsymbol{Y}, \overline{\boldsymbol{X}}) \sim P_{\boldsymbol{X}}(\boldsymbol{x}) W^n(\boldsymbol{y}|\boldsymbol{x}) P_{\boldsymbol{X}}(\overline{\boldsymbol{x}})$. The main difference in the analysis compared to the i.i.d. case [17] is the way in which the inner probability in (11) is handled. This is done in the proof of the following lemma.

Lemma 1. *There exist constants K_1 and $\psi > 0$ such that*

$$\text{rcu}(n, M) \leq \mathbb{E}\Big[\min\Big\{1, MK_1 n^{-\frac{\eta}{2}} e^{-i^n(\boldsymbol{X}, \boldsymbol{Y})}\Big\}\Big] + e^{-\psi n} \quad (12)$$

for sufficiently large n, where

$$\eta \triangleq \mathbb{1}\{W \notin \mathcal{W}_0^{\text{cc}}\} - (N-1). \quad (13)$$

Once Lemma 1 is established, we can prove Theorem 1 in an identical fashion to [18, Thm. 5], [17, Sec. 3.4.5] using the Berry-Esseen theorem (respectively, Chebyshev's inequality) when $V(Q) > 0$ (respectively, $V(Q) = 0$). In particular, the first and second moments of i^n are given by

$$\mathbb{E}[i^n(\boldsymbol{x}, \boldsymbol{Y}) \,|\, \boldsymbol{X} = \boldsymbol{x}] = nI(Q) \quad (14)$$

$$\text{Var}[i^n(\boldsymbol{x}, \boldsymbol{Y}) \,|\, \boldsymbol{X} = \boldsymbol{x}] = nV(Q) \quad (15)$$

for any $\boldsymbol{x} \in T^n(Q)$. To avoid repetition with [17], [18], we focus our attention on proving Lemma 1.

1) Alternative Forms of the Codeword Distribution: By a symmetry argument, we can write (2) as

$$P_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{\mu_n} \prod_{i=1}^{n} Q(x_i) \mathbb{1}\{\boldsymbol{x} \in T^n(Q)\}, \tag{16}$$

where μ_n is a normalizing constant. Using the refined bounds on the size of $T^n(Q)$ in [7, Ex 2.2], it is readily verified that

$$\mu_n = \Theta\big(n^{-\frac{|\mathcal{X}|-1}{2}}\big). \tag{17}$$

As noted in [18], $T^n(Q)$ can be written as

$$T^n(Q) \triangleq \left\{ \boldsymbol{x} : \big|\hat{P}_{\boldsymbol{x}}(x) - Q(x)\big| \le \frac{\delta}{n},\, x = 1, \ldots, |\mathcal{X}| \right\}, \tag{18}$$

where $\delta \in [0, 1)$. We choose δ to be strictly positive.

It will prove useful to give yet another equivalent definition in terms of the product alphabet $\mathcal{X}^k$, where k is a fixed integer. We divide the sequence $\boldsymbol{x}$ into n/k blocks of length k:

$$\begin{aligned} x_1^{(k)} &\triangleq (x_1, \ldots, x_k) \\ x_2^{(k)} &\triangleq (x_{k+1}, \ldots, x_{2k}) \\ &\vdots \\ x_{n/k}^{(k)} &\triangleq (x_{n-k+1}, \ldots, x_n). \end{aligned} \tag{19}$$

For clarity of exposition, it is assumed here that n/k is an integer; the general case is handled similarly. For a given product symbol $x_j^{(k)}$, we define the following function that counts the number of occurrences of a given symbol x:

$$\xi_x^k(x_j^{(k)}) \triangleq \sum_{i=(j-1)k+1}^{jk} \mathbb{1}\{x_i = x\}. \tag{20}$$

We can now write (18) as

$$T^n(Q) = \left\{ \boldsymbol{x} : \Big| \sum_{i=1}^{n/k} \xi_x^k(x_i^{(k)}) - nQ(x) \Big| \le \delta,\, x = 1, \ldots, |\mathcal{X}| \right\}. \tag{21}$$

We provide some reasoning behind this block decomposition in the proof of Lemma 3 below.

We will also make use of a vector $\boldsymbol{\xi}^k(x^{(k)})$ containing $|\mathcal{X}| - N$ of the functions $\{\xi_x^{(k)}\}_{x \in \mathcal{X}}$. Specifically, we write

$$\boldsymbol{\xi}^k(x^{(k)}) \triangleq \begin{bmatrix} \xi_{\tilde{x}_1}^k(x^{(k)}) \\ \vdots \\ \xi_{\tilde{x}_{|\mathcal{X}|-N}}^k(x^{(k)}) \end{bmatrix}, \tag{22}$$

where the indices $\tilde{x}_1, \ldots, \tilde{x}_{|\mathcal{X}|-N}$ are chosen to include all of the symbols except one (e.g. the one with the highest index) from each of the N subchannels.

2) Further Auxiliary Lemmas: The reverse conditional distribution induced by Q and W is given by

$$\widetilde{P}(x|y) \triangleq \frac{Q(x)W(y|x)}{\sum_{\overline{x}} Q(\overline{x})W(y|\overline{x})}, \tag{23}$$

and we write $\widetilde{P}^n(\boldsymbol{x}|\boldsymbol{y}) \triangleq \prod_{i=1}^n \widetilde{P}(x_i|y_i)$. Furthermore, we define the random variables

$$(\boldsymbol{X}, \boldsymbol{Y}, \overline{\boldsymbol{X}}, \widetilde{\boldsymbol{X}}) \sim P_{\boldsymbol{X}}(\boldsymbol{x}) W^n(\boldsymbol{y}|\boldsymbol{x}) P_{\boldsymbol{X}}(\overline{\boldsymbol{x}}) \widetilde{P}^n(\widetilde{\boldsymbol{x}}|\boldsymbol{y}). \tag{24}$$

Lemma 2. *Let $k = |\mathcal{Y}|$, and fix the product symbol $y^{(k)} = (1, \ldots, |\mathcal{Y}|)$. The covariance matrix of $\boldsymbol{\xi}^k(\widetilde{X}^{(k)})$ has full rank under $\widetilde{X}^{(k)} \sim \widetilde{P}^k(\cdot|y^{(k)})$. Moreover, if $W \notin \mathcal{W}_0^{\mathrm{cc}}$, then the covariance matrix of*

$$\begin{bmatrix} i^k(\widetilde{X}^{(k)}, y^{(k)}) \\ \boldsymbol{\xi}^k(\widetilde{X}^{(k)}) \end{bmatrix} \tag{25}$$

has full rank, where i^k is defined analogously to (4).

Proof: See Appendix A. ∎

Next, we provide a large deviations result that plays the role of [17, Lemma 20] in the i.i.d. analysis of [17, Sec. 3.4.5].

Lemma 3. *Fix the integers $K > 0$ and $d \ge 2$, and for each n, let $(n_1, \ldots, n_K)$ be integers such that $\sum_j n_j = n$ and $\min_j n_j = \Theta(n)$. Fix the probability mass functions (PMFs) $Q_1, \ldots, Q_K$ on a finite subset of $\mathbb{R}^d$, let $\{\Sigma_j\}_{j=1}^K$ be the corresponding $d \times d$ covariance matrices, and let $\{\boldsymbol{Z}_i\}_{i=1}^n$ be independent d-dimensional random vectors, n_j of which are distributed according to Q_j for each j. Furthermore, let $\{\Sigma_j'\}_{j=1}^K$ be the submatrices of $\{\Sigma_j\}_{j=1}^K$ obtained by removing the first row and column of each, and let $\{\boldsymbol{Z}_i'\}_{i=1}^n$ be obtained from $\{\boldsymbol{Z}_i\}_{i=1}^n$ by removing the first entry:*

$$\boldsymbol{Z}_i' \triangleq [Z_{i,2},\, \ldots,\, Z_{i,d}]^T. \tag{26}$$

If $\det(\Sigma_j') > 0$ for some j, then for any constants t and $\delta > 0$, and any sequence of vectors $\boldsymbol{\gamma}_n \in \mathbb{R}^{d-1}$, we have

$$\mathbb{E}\left[e^{-\sum_{i=1}^n Z_{1,i}} \mathbb{1}\left\{ \sum_{i=1}^n Z_{1,i} \ge t \cap \left\| \sum_{i=1}^n \boldsymbol{Z}_i' - \boldsymbol{\gamma}_n \right\|_\infty \le \delta \right\} \right] \le \beta_n e^{-t}, \tag{27}$$

where $\beta_n = O\big(n^{-\frac{d-1}{2}}\big)$ uniformly in t and $\boldsymbol{\gamma}_n$. Furthermore, if $\det(\Sigma_j) > 0$ for some j, this can be strengthened to $\beta_n = O\big(n^{-\frac{d}{2}}\big)$ uniformly in t and $\boldsymbol{\gamma}_n$.

Proof: See Appendix B. ∎

3) Proof of Lemma 1: Recall that $Q(x) > 0$ for all x and all outputs are reachable. We define $P_Y(y) \triangleq \sum_x Q(x)W(y|x)$ and $p_{\min} \triangleq \min_y P_Y(y) > 0$. A standard argument (e.g. via the Chernoff bound) reveals that $\boldsymbol{Y}$ falls within the set

$$\mathcal{F}_n(\delta) \triangleq \left\{ \boldsymbol{y} : \min_y \hat{P}_{\boldsymbol{y}}(y) \ge \frac{p_{\min}}{2} \right\} \tag{28}$$

with probability approaching one exponentially fast. Using this observation, the lemma will follow once we show the following for all $\boldsymbol{y} \in \mathcal{F}_n(\delta)$:

$$\mathbb{P}\big[i^n(\overline{\boldsymbol{X}}, \boldsymbol{y}) \ge t\big] \le K_1 n^{-\frac{\eta}{2}} e^{-t} \tag{29}$$

for sufficiently large n and some constant K_1. To show this, we follow [1, Sec. 3.4.5] and note that the following holds

whenever $\widetilde{P}^n(\overline{\boldsymbol{x}}|\boldsymbol{y}) \neq 0$:

$$P_{\boldsymbol{X}}(\overline{\boldsymbol{x}}) = \frac{1}{\mu_n} Q^n(\overline{\boldsymbol{x}}) \frac{\widetilde{P}^n(\overline{\boldsymbol{x}}|\boldsymbol{y})}{\widetilde{P}^n(\overline{\boldsymbol{x}}|\boldsymbol{y})} \mathbb{1}\big\{\overline{\boldsymbol{x}} \in T^n(Q)\big\} \tag{30}$$

$$= \frac{1}{\mu_n} \widetilde{P}^n(\overline{\boldsymbol{x}}|\boldsymbol{y}) e^{-i^n(\overline{\boldsymbol{x}},\boldsymbol{y})} \mathbb{1}\big\{\overline{\boldsymbol{x}} \in T^n(Q)\big\}. \tag{31}$$

Summing both sides over all $\overline{\boldsymbol{x}}$ such that $i^n(\overline{\boldsymbol{x}},\boldsymbol{y}) \geq t$ yields

$$\mathbb{P}\Big[i^n(\overline{\boldsymbol{X}},\boldsymbol{y}) \geq t\Big] = \frac{1}{\mu_n} \mathbb{E}\Big[e^{-i^n(\widetilde{\boldsymbol{X}},\boldsymbol{y})} \times \mathbb{1}\big\{i^n(\widetilde{\boldsymbol{X}},\boldsymbol{y}) \geq t \cap \widetilde{\boldsymbol{X}} \in T^n(Q)\big\} \,\Big|\, \boldsymbol{Y} = \boldsymbol{y}\Big] \tag{32}$$

under the joint distribution in (24).

We set $k = |\mathcal{Y}|$ and let $y_1^{(k)}, \ldots, y_{n/k}^{(k)}$ denote the decomposition of $\boldsymbol{y}$ into blocks in the same way as (19), and similarly for $\widetilde{X}_1^{(k)}, \ldots, \widetilde{X}_{n/k}^{(k)}$. For sequences within the set $\mathcal{F}_n(\delta)$, each output symbol occurs at least $n\frac{p_{\min}}{2}$ times. By the symmetry of $P_{\boldsymbol{X}}$, $\mathbb{P}[i^n(\overline{\boldsymbol{X}},\boldsymbol{y}) \geq t]$ is invariant under permutations of $\boldsymbol{y}$, and we may therefore assume that $\boldsymbol{y}$ starts with $n\frac{p_{\min}}{2}$ repetitions of $(1,\ldots,|\mathcal{Y}|)$. Thus, using the decomposition in (19), the symbol $y^{(k)} = (1,\ldots,|\mathcal{Y}|)$ occurs at least $n\frac{p_{\min}}{2}$ times.

Using the form of $T^n(Q)$ in (21), we can write (32) as

$$\mathbb{P}\Big[i^n(\overline{\boldsymbol{X}},\boldsymbol{y}) \geq t\Big] = \frac{1}{\mu_n} \mathbb{E}\Bigg[e^{-\sum_{i=1}^{n/k} i^k(\widetilde{X}_i^{(k)}, y_i^{(k)})} \mathbb{1}\Bigg\{\sum_{i=1}^{n/k} i^k(\widetilde{X}_i^{(k)}, y_i^{(k)}) \geq t \cap \Bigg|\sum_{i=1}^{n/k} \xi_x^k(x_i^{(k)}) - nQ(x)\Bigg| \leq \delta,\, \forall x \in \mathcal{X}\Bigg\} \,\Bigg|\, \boldsymbol{Y} = \boldsymbol{y}\Bigg]. \tag{33}$$

We now prove (29) using Lemma 3, letting each Q_j therein correspond to the vector in (25) for some $y^{(k)}$, and replacing n therein by n/k.

When $W \notin \mathcal{W}_0^{\text{cc}}$, we upper bound (33) by requiring the second event in the indicator function to hold only for $x \in \{\tilde{x}_1, \ldots, \tilde{x}_{|\mathcal{X}|-N}\}$ (see (22)). Recalling the above observation that the symbol $y^{(k)} = (1,\ldots,|\mathcal{Y}|)$ occurs $\Theta(n)$ times, the second part of Lemma 2 reveals that Σ_j (defined in Lemma 3) has full rank for the corresponding value of j. Combining (27), the second part of Lemma 2 (with $d = |\mathcal{X}| - N + 1$) and (17), we can upper bound (33) by $O\big(n^{-\frac{1}{2}(2-N)}\big)e^{-t}$. This coincides with (29), since the condition $W \notin \mathcal{W}_0^{\text{cc}}$ implies that the indicator function in (13) equals one.

In the remaining case, i.e. $W \in \mathcal{W}_0^{\text{cc}}$, we follow similar steps, except that we also remove the first constraint (containing i^k) from the indicator function in (33), and we use the first part of Lemma 2 instead of the second part, leading to a bound of the form $O\big(n^{-\frac{1}{2}(1-N)}\big)e^{-t}$. This completes the proof of (29), which in turn completes the proof of Lemma 1.

B. Proof of Theorem 2

Let $\mathcal{U} = \{1,\ldots,N\}$ be the alphabet indexing the N subchannels, and let W_1 be one such subchannel. Recall that the subchannels are identical by assumption. Let $\mathcal{C} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(M)}\}$ be an arbitrary codebook (not necessarily constant-composition), and for a given codeword $\boldsymbol{x}$, let $\boldsymbol{u}(\boldsymbol{x})$ be a sequence of symbols in $\mathcal{U}$ indicating which subchannel is used for each channel use. The maximum-likelihood decoding rule can be interpreted as deterministically constructing the corresponding auxiliary sequence $\boldsymbol{u}$, and then choosing

$$\hat{m} = \underset{j\,:\,\boldsymbol{u}(\boldsymbol{x}^{(j)})=\boldsymbol{u}}{\arg\max}\; W^n(\boldsymbol{y}|\boldsymbol{x}^{(j)}). \tag{34}$$

We claim that if the total fraction of $\boldsymbol{u}$ sequences used (i.e. such that $\boldsymbol{u}(x^{(j)}) = \boldsymbol{u}$ for some j) is less than $\frac{1}{Z}$ for some integer $Z \geq 2$, then the number of messages can be increased by a factor of Z without changing the error probability. To see this, first note that since the $\boldsymbol{u}$ sequence is known with certainty at the decoder, the determination of the message j among $\{j : \boldsymbol{u}(\boldsymbol{x}^{(j)}) = \boldsymbol{u}\}$ simply corresponds to coding over W_1; the precise values in $\boldsymbol{u}$ play no further role. Thus, since less than a proportion $\frac{1}{Z}$ of the $\boldsymbol{u}$ sequences are utilized, we may take the original codebook structure and repeat it Z times across the remaining $\boldsymbol{u}$ sequences in an arbitrary fashion, without affecting the error probability.

We now specialize these observations to constant-composition codes. If the codewords $\{\boldsymbol{x}^{(j)}\}$ have the same composition, the same is true for the $\{\boldsymbol{u}(\boldsymbol{x}^{(j)})\}$. Thus, the total number of $\boldsymbol{u}$ sequences utilized is not N^n, but instead the size of some type class $T^n(Q_U)$ on $\mathcal{U}^n$. From [7, Ex. 2.2], the largest type class corresponds to Q_U being uniform on $\mathcal{U}$, and yields $|T^n(Q_U)| = \Theta\big(N^n n^{-\frac{N-1}{2}}\big)$. Thus, at most a fraction $\Theta\big(n^{-\frac{N-1}{2}}\big)$ of the $\boldsymbol{u}$ sequences are utilized, and we may increase the number of messages by a factor of $\Theta\big(n^{-\frac{N-1}{2}}\big)$ without affecting the error probability. Thus, for any sequence of constant-composition codes, there exists a sequence of general codes with an addition of $\frac{1}{2}(N-1)\log n + O(1)$ to the third-order term, as stated in (10).

Appendix

A. Proof of Lemma 2

From (22), the covariance matrix of $\boldsymbol{\xi}^k(\widetilde{X}^{(k)})$ has determinant zero if and only if

$$\text{Var}\Big[\gamma_{\tilde{x}_1}\xi_{\tilde{x}_1}^k(\widetilde{X}^{(k)}) + \cdots + \gamma_{\tilde{x}_{|\mathcal{X}|-N}}\xi_{\tilde{x}_{|\mathcal{X}|-N}}^k(\widetilde{X}^{(k)})\Big] = 0 \tag{35}$$

for some $(\gamma_{\tilde{x}_1}, \ldots, \gamma_{\tilde{x}_{|\mathcal{X}|-N}}) \neq (0,\ldots,0)$. Since $\widetilde{X}^{(k)} \sim \widetilde{P}^k(\cdot|y^{(k)})$ is a product distribution and $y^{(k)} = (1,\ldots,|\mathcal{Y}|)$ (see the lemma statement), we can write (35) as

$$\sum_y \text{Var}_{\widetilde{P}(\cdot|y)}\Big[\gamma_{\tilde{x}_1}\xi_{\tilde{x}_1}(\widetilde{X}) + \cdots + \gamma_{\tilde{x}_{|\mathcal{X}|-N}}\xi_{\tilde{x}_{|\mathcal{X}|-N}}(\widetilde{X})\Big] = 0, \tag{36}$$

where $\xi_x(\tilde{x}) \triangleq \mathbb{1}\{\tilde{x} = x\}$ by (20). We rewrite (36) as

$$\sum_y \text{Var}_{\widetilde{P}(\cdot|y)}\Big[\sum_x \gamma_x \xi_x(\widetilde{X})\Big] = 0, \tag{37}$$

where $\gamma_x = 0$ for all $x \notin \{\tilde{x}_1, \ldots, \tilde{x}_{|\mathcal{X}|-N}\}$. We proceed by assuming that (37) is true, and then arriving at a contradiction. Since $Q(x) > 0$ for all x, we have $\widetilde{P}(x|y) > 0$ if and only if $W(y|x) > 0$. It follows from (37) that $\gamma_x = \gamma_{\overline{x}}$ for any pair

$(x, \overline{x})$ sharing a common output (i.e. $W(y|x)W(y|\overline{x}) > 0$ for some y), since otherwise the argument to $\mathrm{Var}_{\widetilde{P}(\cdot|y)}$ would differ depending on whether $\widetilde{X} = x$ or $\widetilde{X} = \overline{x}$.

Since $\gamma_x = \gamma_{\overline{x}}$ for all $(x, \overline{x})$ sharing a common output, it also holds that $\gamma_x = \gamma_{\overline{x}}$ for all $\overline{x}$ that can be reached from x by following a path in the channel graph (i.e. by moving between inputs and outputs such that $W(\cdot|\cdot) \neq 0$). Since $\gamma_x = 0$ for at least one symbol of each of the N subchannels (see (22)), it follows that $\gamma_x = 0$ for all x, thus contradicting the assumption that $(\gamma_1, \ldots, \gamma_{|\mathcal{X}|-1}) \neq (0, \ldots, 0)$.

The claim regarding (25) follows similarly using a proof by contradiction. By the preceding argument, the submatrix of the covariance matrix obtained by removing the first row and column has full rank. The only remaining possibility for the matrix (25) to have determinant zero is that there exists a function $g(x)$ (equal to some linear combination of $\{\xi_x\}$) such that $\mathrm{Var}_{\widetilde{P}(\cdot|y)}[i(\widetilde{X}, y) + g(\widetilde{X})] = 0$ for all y. Recalling that $\widetilde{P}(x|y) > 0$ if and only if $W(y|x) > 0$, this condition is equivalent to $i(x, y) + g(x) + h(y)$ taking a constant value wherever $W(y|x) > 0$, for some function $h(y)$. We conclude from the definition of i in (3) that $W(y|x) = g'(x)h'(y)\mathbb{1}\{W(y|x) > 0\}$ for some functions $g'(x)$ and $h'(y)$. This contradicts the assumption that $W \notin \mathcal{W}_0^{\mathrm{cc}}$ (see (7)), and we conclude that the covariance matrix of (25) has full rank.

We can now provide some intuition as to why it is beneficial to work with blocks of size $k = |\mathcal{Y}|$ (cf. (19)). It is this decomposition that allowed us to write (36) with a summation over y, but with $\{\gamma_x\}$ not depending on y. In contrast, a similar argument considering one symbol at a time would lead to an analogous expression to (36) with γ_x replaced by $\gamma_{x,y}$.

B. Proof of Lemma 3

To prove the first part of the lemma, we write

$$\mathbb{E}\left[e^{-\sum_{i=1}^n Z_{1,i}}\mathbb{1}\left\{\sum_{i=1}^n Z_{1,i} \geq t \cap \left\|\sum_{i=1}^n \boldsymbol{Z}_i' - \boldsymbol{\gamma}_n\right\|_\infty \leq \delta\right\}\right] \leq e^{-t}\mathbb{P}\left[\left\|\sum_{i=1}^n \boldsymbol{Z}_i' - \boldsymbol{\gamma}_n\right\|_\infty \leq \delta\right], \quad (38)$$

which follows immediately from the constraint $\sum_{i=1}^n Z_{1,i} \geq t$. The probability on the right-hand side of (38) is that of a sum of independent vectors falling within a $(d-1)$-dimensional cube of side length 2δ, and may thus be upper bounded by the probability of the same sum falling within a sphere of radius $\delta\sqrt{d-1}$. Using the concentration function result of [19, Thm. 6.2] and the assumption that $\det(\Sigma_j') > 0$ for some $j \in \{1, \ldots, K\}$ corresponding to $\Theta(n)$ of the terms in the sum, this probability behaves as $\Theta\big(n^{-\frac{d-1}{2}}\big)$ uniformly in the location of the sphere (i.e. the vector $\boldsymbol{\gamma}_n$ in (38)). More precisely, this behavior is proved in the same way as [5, Appendix F] upon noting that the $\{\boldsymbol{Z}_i'\}$ are bounded since we are considering the case of finite alphabets.

The second part of the lemma is proved similarly by upper bounding the left-hand side of (38) by

$$\sum_{l=0}^{\infty} e^{-t-2l\delta}\mathbb{P}\left[t + 2l\delta \leq \sum_{i=1}^n Z_{1,i} < t + 2(l+1)\delta \cap \left\|\sum_{i=1}^n \boldsymbol{Z}_i' - \boldsymbol{\gamma}_n\right\|_\infty \leq \delta\right]. \quad (39)$$

Equation (39) gives the probability of a sum of independent vectors falling within a d-dimensional cube of side length 2δ. Using the same argument as the first part of the lemma and the assumption that $\det(\Sigma_j) > 0$ for some j, the probability behaves as $\Theta\big(n^{-\frac{d}{2}}\big)$ uniformly in t and $\boldsymbol{\gamma}_n$. The proof is concluded by using the geometric series to write $\sum_{l=1}^{\infty} e^{-2l\delta} = \frac{1}{1-e^{-2\delta}}$, which is a constant.

REFERENCES

[1] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[2] M. Hayashi, "Information spectrum approach to second-order coding rate in channel coding," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 4947–4966, Nov. 2009.

[3] V. Strassen, "Asymptotische Abschätzungen in Shannon's Informationstheorie," in *Trans. 3rd Prague Conf. on Inf. Theory*, 1962, pp. 689–723, [English Translation: http://www.math.wustl.edu/~luthy/strassen.pdf].

[4] M. Tomamichel and V. Tan, "A tight upper bound for the third-order asymptotics for most discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7041–7051, Nov 2013.

[5] Y. Altuğ and A. B. Wagner, "A refinement of the random coding bound," in *Allerton Conf. on Comm., Control and Comp.*, Monticello, IL, 2012.

[6] ——, "The third-order term in the normal approximation for singular channels," 2013, http://arxiv.org/abs/1309.5126.

[7] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge University Press, 2011.

[8] R. Gallager, *Information Theory and Reliable Communication*. John Wiley & Sons, 1968.

[9] I. Csiszár and P. Narayan, "Channel capacity for a given decoding metric," *IEEE Trans. Inf. Theory*, vol. 45, no. 1, pp. 35–43, Jan. 1995.

[10] Y. Liu and B. Hughes, "A new universal random coding bound for the multiple-access channel," *IEEE Trans. Inf. Theory*, vol. 42, no. 2, pp. 376–386, March 1996.

[11] J. Scarlett, A. Martinez, and A. Guillén i Fàabregas, "Second-order rate region of constant-composition codes for the multiple-access channel," *IEEE Trans. Inf. Theory*, vol. 61, no. 1, pp. 157–172, Jan. 2015.

[12] V. Kostina and S. Verdú, "Channels with cost constraints: Strong converse and dispersion," in *IEEE Int. Symp. Inf. Theory*, Istanbul, 2013.

[13] P. Moulin, "The log-volume of optimal codes for memoryless channels, within a few nats," 2013, http://arxiv.org/abs/1311.0181.

[14] V. Y. F. Tan and M. Tomamichel, "The third-order term in the normal approximation for the AWGN channel," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2430–2438, May 2015.

[15] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. Journal*, vol. 27, pp. 379–423, July and Oct. 1948.

[16] J. Scarlett, "Reliable communication under mismatched decoding," Ph.D. dissertation, University of Cambridge, 2014, [Online: http://itc.upf.edu/biblio/1061].

[17] Y. Polyanskiy, "Channel coding: Non-asymptotic fundamental limits," Ph.D. dissertation, Princeton University, 2010.

[18] J. Scarlett, A. Martinez, and A. Guillén i Fàbregas, "Mismatched decoding: Error exponents, second-order rates and saddlepoint approximations," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2647–2666, May 2014.

[19] C. Esseen, "On the concentration function of a sum of independent random variables," *Zeit. für Wahr. und Verwandte Gebiete*, vol. 9, no. 4, pp. 290–308, 1968.