Third-order Analysis of Channel Coding in the Moderate Deviations Regime

Recep Can Yavas, Victoria Kostina, and Michelle Effros

2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834841

***Abstract*—The channel coding problem in the moderate deviations regime is studied; here, the error probability sub-exponentially decays to zero, and the rate approaches the capacity slower than $O(1/\sqrt{n})$. The main result refines Altuğ and Wagner's moderate deviations result by deriving lower and upper bounds on the third-order term in the asymptotic expansion of the maximum achievable message set size. The third-order term of the expansion employs a new quantity called the channel skewness. For the binary symmetric channel and most practically important (n, ϵ) pairs, including $n \in [100, 500]$ and $\epsilon \in [10^{-10}, 10^{-1}]$, an approximation up to the channel skewness is the most accurate among several expansions in the literature.**

I. Introduction

The fundamental limit of channel coding is the maximum achievable message size $M^*(n, \epsilon)$ given a channel $P_{Y|X}$, a blocklength n, and an average error probability ϵ. Since determining $M^*(n, \epsilon)$ exactly is difficult for arbitrary triples $(P_{Y|X}, n, \epsilon)$, the literature investigating the behavior of $M^*(n, \epsilon)$ studies two asymptotic regimes: the central limit theorem (CLT) and the large deviations (LD) regimes. Given a sum of n independent and identically distributed (i.i.d.) random variables, the probability that this sum deviates from the mean by order-$\sqrt{n}$ is characterized by a Gaussian distribution whose parameters are constant with respect to n. This classical result is known as the CLT. The probability that the sum of n i.i.d. random variables deviates from the mean by order-n is characterized by Cramér's theorem [1], which shows that the probability decays exponentially with n if Cramér's condition is satisfied. This result is commonly known as the LD theorem. Any deviation from the mean with order strictly greater than $\sqrt{n}$ and strictly smaller than n is said to fall in the moderate deviations (MD) regime. A bound on the probability that an i.i.d. sum deviates from the mean by some amount in the MD regime appears in [2, Ch. 8]. This work focuses on channel coding in the MD regime.

Given a channel $P_{Y|X}$, error probability ϵ, and blocklength n, we define the *non-Gaussianity* of the channel as

$$\zeta(n, \epsilon) \triangleq \log M^*(n, \epsilon) - (nC - \sqrt{nV_\epsilon} Q^{-1}(\epsilon)), \qquad (1)$$

where C is the capacity, and $V_\epsilon > 0$ is the ϵ-dispersion of the channel [3, Sec. IV]; $\zeta(n, \epsilon)$ is the third-order term in a second-order optimal characterization of $\log M^*(n, \epsilon)$.

R. C. Yavas, V. Kostina, and M. Effros are with the Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125, USA (e-mail: ryavas, vkostina, effros@caltech.edu). This work was supported in part by the National Science Foundation (NSF) under grant CCF-1817241 and CCF-1956386.

Channel coding analyses in the CLT regime fix a target error probability $\epsilon \in (0, 1)$ and characterize $M^*(n, \epsilon)$ as the blocklength n approaches infinity. Strassen's results [4] for discrete memoryless channels (DMCs) under the maximal error probability constraint fall in this domain, giving $\zeta(n, \epsilon) = O(\log n)$. More recently, Polyanskiy *et al.* [3] and Hayashi [5] revive Strassen's result [4], showing that the same asymptotic expansion holds for the average error probability constraint, bounding the coefficient of the $\log n$ term from below and above, and extending the result to the Gaussian channel with maximal or average power constraint. CLT-style analyses also exist for channels with feedback [6], lossy [7] and lossless data compression [4], [8], network information theory [9], random access channels [10], [11], and many other scenarios.

For channel coding scenarios in the LD regime, which is commonly known as the error exponent regime, we fix a rate $R = \frac{\log M}{n}$ strictly below the channel capacity and seek to characterize the rate of exponential decay of the minimum achievable error probability $\epsilon^*(n, R)$ as the blocklength n approaches infinity. In this case, $\epsilon^*(n, R)$ decays exponentially with n, and [12, Ch. 5] derives the optimal error exponent $\lim_{n\to\infty} -\frac{1}{n} \log \epsilon^*(n, R)$ for R above the critical rate.

In [13], Altuğ and Wagner seek to bridge the gap between the CLT and LD regimes. In their results, the error probability ϵ_n decays sub-exponentially to zero, i.e., $\epsilon_n \to 0$ and $-\frac{1}{n} \log \epsilon_n \to 0$, and the rate approaches the capacity with a gap of order strictly greater than $\frac{1}{\sqrt{n}}$. Altuğ and Wagner argue that this formulation is more practically relevant since it simultaneously considers low error probabilities and high achievable rates. For DMCs with positive ϵ_n-dispersion V_{ϵ_n} and a sequence of sub-exponentially decaying ϵ_n, they show

$$\zeta(n, \epsilon_n) = o(\sqrt{n} Q^{-1}(\epsilon_n)). \qquad (2)$$

In [14], Polyanskiy and Verdú give an alternative proof of (2) and extend the result to the Gaussian channel with a maximal power constraint. In [15], Chubb *et al.* extend the second-order result in (2) to quantum channels.

Emerging applications with tight delay constraints motivate increasing interest in refining the asymptotic expansions of maximum achievable channel coding rate. For small blocklength n, the non-Gaussianity $\zeta(n, \epsilon)$ in (1) can be quite large when compared to the second-order term $O(\sqrt{n})$. For example, in the CLT regime, given a DMC with finite input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$, [3] bounds the non-

Gaussianity as

$$O(1) \le \zeta(n,\epsilon) \le \left(|\mathcal{X}| - \frac{1}{2}\right)\log n + O(1). \tag{3}$$

A variety of refinements follow. For nonsingular channels, the random coding union bound improves the lower bound to $\frac{1}{2}\log n + O(1)$ [16, Cor. 54]. For DMCs with positive ϵ-dispersion, Tomamichel and Tan [17] improve the upper bound to $\frac{1}{2}\log n + O(1)$. For nonsingular channels with positive ϵ-dispersion, where the information density random variable has a non-lattice distribution, Moulin [18] shows

$$\zeta(n,\epsilon) \ge \frac{1}{2}\log n + \underline{S}\, Q^{-1}(\epsilon)^2 + \underline{B} + o(1) \tag{4}$$

$$\zeta(n,\epsilon) \le \frac{1}{2}\log n + \overline{S}\, Q^{-1}(\epsilon)^2 + \overline{B} + o(1), \tag{5}$$

where $\underline{S}$, $\overline{S}$. $\underline{B}$, and $\overline{B}$ are constants depending on the channel parameters. For Gallager-symmetric [12, p. 94], singular channels, [19] shows $\zeta(n,\epsilon) = O(1)$. Bounds on the sub-exponential factors in the LD regime appear in [19]–[22]. Similar to the CLT regime results reviewed above, these results depend on whether the channel is singular or nonsingular.

A sequence of error probabilities $\{\epsilon_n\}_{n=1}^{\infty}$ is said to be an *MD sequence* if for every $c > 0$, there exists an $n_0(c)$ such that for all $n \ge n_0(c)$,

$$\exp\{-cn\} \le \epsilon_n \le 1 - \exp\{-cn\}, \tag{6}$$

or, equivalently $Q^{-1}(\epsilon_n) = o(\sqrt{n})$. This definition extends the MD error probability region in [13] to include the sequences that sub-exponentially approach 1 and the constant values. We study channel coding for nonsingular channels and average error probability satisfying (6), refining the lower and upper bounds in (2). Our result generalizes (4)–(5) to non-constant error probability ϵ_n at the expense of not bounding the constant term. We show that for nonsingular channels with positive minimum dispersion and ϵ_n satisfying (6), $\zeta(n,\epsilon_n)$ in (2) is bounded below and above as

$$\frac{1}{2}\log n + \underline{S}\, Q^{-1}(\epsilon_n)^2 + O\left(\frac{Q^{-1}(\epsilon_n)^3}{\sqrt{n}}\right) + O(1) \le \zeta(n,\epsilon_n) \tag{7}$$

$$\le \frac{1}{2}\log n + \overline{S}\, Q^{-1}(\epsilon_n)^2 + O\left(\frac{Q^{-1}(\epsilon_n)^3}{\sqrt{n}}\right) + O(1), \tag{8}$$

where the constants $\underline{S}$ and $\overline{S}$ are the same ones as in (4) and (5). We show that the non-Gaussianity gets arbitrarily close to $O(\sqrt{n})$ as ϵ_n approaches an exponential decay, rivaling the dispersion term in (1). Thus, refining the third-order term as we do in (7)–(8) is especially significant in the MD regime. Our achievability bound applies the standard random coding bound used both in the CLT [3], [18] and LD [20] regimes, and our converse bound combines the result in [17, Prop. 6], which is a relaxation of the meta-converse bound [3, Th. 27], and a saddlepoint result of a maximin problem in [18, Lemma 14]. For the ϵ_n behavior studied in the MD regime, neither the Berry-Esseen theorem used in [3] nor the refined Edgeworth expansion used in [18] to treat the constant ϵ case is sharp enough for the $O(1)$ precision in (7)–(8). We replace these tools with the moderate deviations bounds found in [2, Ch. 8].

We define the *channel skewness* operationally as

$$S \triangleq \lim_{\epsilon\to 0} \limsup_{n\to\infty} \frac{\zeta(n,\epsilon) - \frac{1}{2}\log n}{Q^{-1}(\epsilon)^2}. \tag{9}$$

For the maximum achievable rate, the channel skewness serves as the third-order fundamental channel characterization after channel capacity and dispersion [3, Sec. IV]. The skewness of the information density random variable (see (10), below) plays a critical role in characterizing the channel skewness.

For Cover-Thomas symmetric channels [23], our result gives the channel skewness S exactly. For the binary symmetric channel (BSC), in Fig. 1 in Section III below, we compare our asymptotic approximation for the maximum achievable rate using terms up to the channel skewness, i.e., $\zeta(n,\epsilon) \approx \frac{1}{2}\log n + S\, Q^{-1}(\epsilon)^2$ with Moulin's bounds in (4)–(5), the normal approximation, which takes $\zeta(n,\epsilon) \approx \frac{1}{2}\log n$, and the saddlepoint approximations in [21], [22].

The paper is organized as follows. Section II includes notation definitions and other preliminaries. Section III presents and discusses the main result. The complete proof is relegated to the extended version [24].

II. Notation and Preliminaries

A. Notation

For any $k \in \mathbb{N}$, we denote $[k] \triangleq \{1,\dots,k\}$. We denote random variables by capital letters (e.g., X) and individual realizations of random variables by lowercase letters (e.g., x). We use boldface letters (e.g., $\mathbf{x}$) to denote length-n vectors, calligraphic letters (e.g., $\mathcal{X}$) to denote sets, and sans serif font (e.g., A) to denote matrices. The i-th entry of a vector $\mathbf{x}$ is denoted by x_i, and (i,j)-th entry of a matrix A is denoted by $\mathsf{A}_{i,j}$.

The sets of all distributions on the channel input alphabet $\mathcal{X}$ and the channel output alphabet $\mathcal{Y}$ are denoted by $\mathcal{P}$ and $\mathcal{Q}$, respectively. We write $X \sim P_X$ to indicate that X is distributed according to $P_X \in \mathcal{P}$. Given a distribution $P_X \in \mathcal{P}$ and a conditional distribution $P_{Y|X}$ on $\mathcal{Y}$ given $\mathcal{X}$, we write $P_X \times P_{Y|X}$ to indicate the joint distribution of (X,Y), and P_Y to indicate the marginal distribution of Y. The skewness of a random variable X, denoted by $\mathrm{Sk}(X)$, is defined as

$$\mathrm{Sk}(X) \triangleq \frac{\mathbb{E}\left[X^3\right]}{\mathrm{Var}\left[X\right]^{3/2}}. \tag{10}$$

We measure information in nats, and logarithms and exponents have base e.

We use notation $O(\cdot)$ and $o(\cdot)$ as $f(n) = O(g(n))$ if $\limsup_{n\to\infty} |f(n)/g(n)| < \infty$, and $f(n) = o(g(n))$ if $\lim_{n\to\infty} |f(n)/g(n)| = 0$. We use $Q(\cdot)$ to represent the complementary Gaussian cumulative distribution function $Q(x) \triangleq \frac{1}{\sqrt{2\pi}}\int_x^\infty \exp\left\{-\frac{t^2}{2}\right\} dt$ and $Q^{-1}(\cdot)$ to represent its functional inverse.

B. Definitions related to information density

The relative entropy and divergence variance between two distributions $P, Q \in \mathcal{P}$ are denoted by $D(P\|Q) \triangleq \mathbb{E}\left[\log \frac{P(X)}{Q(X)}\right]$ and $V(P\|Q) \triangleq \mathrm{Var}\left[\log \frac{P(X)}{Q(X)}\right]$ where $X \sim P$. The conditional relative entropy and conditional divergence variance are denoted by

$$D(P_{Y|X}\|Q_Y|P_X) \triangleq \sum_{x\in\mathcal{X}} P_X(x) D(P_{Y|X=x}\|Q_Y) \tag{11}$$

$$V(P_{Y|X}\|Q_Y|P_X) \triangleq \sum_{x\in\mathcal{X}} P_X(x) V(P_{Y|X=x}\|Q_Y). \tag{12}$$

Let $(X,Y) \sim P_X \times P_{Y|X}$. The information density is defined as

$$\imath(x;y) \triangleq \log \frac{P_{Y|X}(y|x)}{P_Y(y)}, \quad \forall\, x \in \mathcal{X}, y \in \mathcal{Y}. \tag{13}$$

We define the following moments of the random variable $\imath(X;Y)$:

- the mutual information

$$I(P_X, P_{Y|X}) \triangleq \mathbb{E}\left[\imath(X;Y)\right] = D(P_{Y|X}\|P_Y|P_X), \tag{14}$$

- the unconditional information variance

$$V_{\mathrm{u}}(P_X, P_{Y|X}) \triangleq V(P_X \times P_{Y|X}\|P_X \times P_Y), \tag{15}$$

- the conditional information variance

$$V(P_X, P_{Y|X}) \triangleq V(P_{Y|X}\|P_Y|P_X), \tag{16}$$

- the reverse dispersion [16, Sec. 3.4.5]

$$V^{\mathrm{r}}(P_X, P_{Y|X}) \triangleq \mathbb{E}\left[\mathrm{Var}\left[\imath(X;Y)|Y\right]\right], \tag{17}$$

- the unconditional information skewness

$$\mathrm{Sk}_{\mathrm{u}}(P_X, P_{Y|X}) \triangleq \mathrm{Sk}(\imath(X;Y)). \tag{18}$$

C. Discrete memoryless channel

A discrete memoryless channel (DMC) is characterized by a finite input alphabet $\mathcal{X}$, a finite output alphabet $\mathcal{Y}$, and a probability transition matrix $P_{Y|X}$. A DMC satisfies

$$P_{Y^n|X^n}(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{n} P_{Y|X}(y_i|x_i). \tag{19}$$

Below, we define the channel code.

Definition 1: An (n, M, ϵ)-code for a DMC $P_{Y|X}$ comprises an encoding function

$$\mathsf{f}\colon [M] \to \mathcal{X}^n, \tag{20}$$

and a decoding function

$$\mathsf{g}\colon \mathcal{Y}^n \to [M], \tag{21}$$

that satisfy an average error probability constraint

$$1 - \frac{1}{M}\sum_{m=1}^{M} P_{Y|X}^n(\mathsf{g}^{-1}(m)|\mathsf{f}(m)) \le \epsilon. \tag{22}$$

The maximum achievable message size $M^*(n,\epsilon)$ under the average error probability criterion is defined as

$$M^*(n,\epsilon) \triangleq \max\{M\colon \exists \text{ an } (n, M, \epsilon)\text{-code}\}. \tag{23}$$

D. Definitions related to the optimal input distribution

The capacity of a DMC $P_{Y|X}$ is

$$C(P_{Y|X}) \triangleq \max_{P_X\in\mathcal{P}} I(P_X, P_{Y|X}). \tag{24}$$

We denote the set of capacity-achieving input distributions by

$$\mathcal{P}^\dagger \triangleq \{P_X \in \mathcal{P}\colon I(P_X, P_{Y|X}) = C(P_{Y|X})\}. \tag{25}$$

The capacity-achieving output distribution is unique [12, Cor. 1 to Th. 4.5.2] and here denoted by $P_Y^* \in \mathcal{Q}$. For any $P_X^\dagger \in \mathcal{P}^\dagger$, it holds that $V(P_X^\dagger, P_{Y|X}) = V_{\mathrm{u}}(P_X^\dagger, P_{Y|X})$ [3, Lemma 62].

Define $V_{\min} \triangleq \min_{P_X^\dagger\in\mathcal{P}^\dagger} V(P_X^\dagger, P_{Y|X})$ and $V_{\max} \triangleq \max_{P_X^\dagger\in\mathcal{P}^\dagger} V(P_X^\dagger, P_{Y|X})$. The ϵ-dispersion [3] of a DMC is defined as

$$V_\epsilon \triangleq \begin{cases} V_{\min} & \text{if } \epsilon < \frac{1}{2} \\ V_{\max} & \text{if } \epsilon \ge \frac{1}{2}. \end{cases} \tag{26}$$

The set of dispersion-achieving input distributions is defined as

$$\mathcal{P}^* \triangleq \begin{cases} \{P_X^\dagger \in \mathcal{P}^\dagger\colon V(P_X^\dagger, P_{Y|X}) = V_\epsilon\} & \text{if } \epsilon \ne \frac{1}{2} \\ \mathcal{P}^\dagger & \text{if } \epsilon = \frac{1}{2}. \end{cases} \tag{27}$$

Any $P_X^\dagger \in \mathcal{P}^\dagger$ satisfies $D(P_{Y|X=x}\|P_Y^*) = C$ for any $x \in \mathcal{X}$ with $P_X^\dagger(x) > 0$, and $D(P_{Y|X=x}\|P_Y^*) \le C$ for all $x \in \mathcal{X}$ [12, Th. 4.5.1]. Hence, the support of any capacity-achieving input distribution is a subset of

$$\mathcal{X}^\dagger = \{x \in \mathcal{X}\colon D(P_{Y|X=x}\|P_Y^*) = C\}. \tag{28}$$

The support of any dispersion-achieving input distribution is a subset of

$$\mathcal{X}^* \triangleq \bigcup_{P_X^*\in\mathcal{P}^*} \mathrm{supp}(P_X) \subseteq \mathcal{X}^\dagger. \tag{29}$$

The quantities that follow appear in the third-order term evaluated with the skewness-achieving input distribution perturbed away from $P_X^* \in \mathcal{P}^*$ (see (39), below). Define

$$\mathsf{J}_{x,x'} \triangleq \begin{cases} -\nabla^2 I(P_X^\dagger, P_{Y|X})_{x,x'} & \text{if } x, x' \in \mathcal{X}^\dagger, \\ 0 & \text{otherwise.} \end{cases} \tag{30}$$

$$\tilde{\mathsf{J}} \triangleq \mathsf{J}^+ - \frac{1}{\mathbf{1}^\top \mathsf{J}^+ \mathbf{1}} (\mathsf{J}^+\mathbf{1})(\mathsf{J}^+\mathbf{1})^\top \tag{31}$$

$$\mathbf{v}(P_X)_x \triangleq \nabla V(P_X, P_{Y|X})_x \tag{32}$$

$$\bar{\mathbf{v}}(P_X)_x \triangleq \mathbb{E}\left[\frac{\partial V(P_{Y|X=\tilde{X}}\|P_Y)}{\partial P_X(x)}\right], \ x \in \mathcal{X}, \text{ and } \tilde{X} \sim P_X \tag{33}$$

$$A_0(P_X) \triangleq \frac{1}{8V_\epsilon} \mathbf{v}(P_X)^\top \tilde{\mathsf{J}} \mathbf{v}(P_X) \tag{34}$$

$$A_1(P_X) \triangleq \frac{1}{8V_\epsilon} \bar{\mathbf{v}}(P_X)^\top \tilde{\mathsf{J}} \bar{\mathbf{v}}(P_X), \tag{35}$$

where ∇ and ∇^2 denote the gradient and Hessian operators with respect to P_X, J^+ denotes the Moore-Penrose pseudo-inverse of J, and $\mathbf{1}$ denotes the all-one vector. See the extended version of this paper [24] and [18, Lemma 2] for properties of these quantities.

E. Singularity of a DMC

The following definition divides DMCs into two groups, for which $M^*(n,\epsilon_n)$ behaves differently in the non-Gaussianity term (even in the CLT regime). An input distribution-channel pair $(P_X, P_{Y|X})$ is *singular* [20, Def. 1] if for all $(x,\overline{x},y)$ such that $P_X \times P_{Y|X}(x,y) > 0$ and $P_X \times P_{Y|X}(\overline{x},y) > 0$, it holds that

$$P_{Y|X}(y|x) = P_{Y|X}(y|\overline{x}). \tag{36}$$

We define the singularity parameter [18, eq. (2.25)]

$$\eta(P_X, P_{Y|X}) \triangleq 1 - \frac{V^{\mathrm{r}}(P_X, P_{Y|X})}{V_{\mathrm{u}}(P_X, P_{Y|X})}, \tag{37}$$

which is a constant in $[0,1]$. The pair $(P_X, P_{Y|X})$ is singular if and only if $\eta(P_X, P_{Y|X}) = 1$ [25, Remark 1]. A channel $P_{Y|X}$ is called singular if $\eta(P_X, P_{Y|X}) = 1$ for all $P_X^* \in \mathcal{P}^*$, and nonsingular otherwise. The binary erasure channel is an example singular channel. Our focus in this paper is on nonsingular channels. For brevity, we drop $P_{Y|X}$ in the notations for capacity, dispersion, skewness, and singularity parameter of the channel.

III. Main Result

Theorems 1 and 2 are our achievability and converse results, respectively.

Theorem 1: Suppose that ϵ_n satisfies (6) and that $P_{Y|X}$ is a nonsingular channel with $V_{\epsilon_n} > 0$ for all n and $\mathcal{X}^\dagger = \mathcal{X}^*$. It holds that

$$\zeta(n,\epsilon_n) \geq \frac{1}{2}\log n + Q^{-1}(\epsilon_n)^2 \max_{P_X^* \in \mathcal{P}^*} \left(\frac{\mathrm{Sk}_{\mathrm{u}}(P_X^*)\sqrt{V_{\epsilon_n}}}{6} + \frac{1-\eta(P_X^*)}{2(1+\eta(P_X^*))} + A_0(P_X^*) \right) + O\left(\frac{Q^{-1}(\epsilon_n)^3}{\sqrt{n}}\right) + O(1). \tag{38}$$

Proof: The proof consists of two parts and extends the argument in [18][1] to include ϵ_n that decreases to 0 or increases to 1 as permitted by (6). The first part is a standard random coding bound. It is used in the CLT regime for a third-order analysis in [16] and a fourth-order analysis in [18]; it also comes up in the LD regime [20]. We set an arbitrary distribution $P_X \in \mathcal{P}$ for the i.i.d. random codebook and employ a maximum likelihood decoder. To bound the probability $\mathbb{P}\left[\imath(\mathbf{X};\mathbf{Y}) \leq \tau_n\right]$, we replace the Edgeworth expansion in [18, eq. (5.30)], which gives the refined asymptotics of the Berry-Esseen theorem, with its moderate deviations version from [2, Ch. 8, Th. 2]. Note that the Edgeworth expansion yields an additive remainder term $O(1/n)$ to the normality; this term becomes too large for the entire range of probabilities in (6). Therefore, a moderate deviation result that yields a multiplicative remainder term $(1+o(1))$ is desired. In the proof, we also derive the inverse of the moderate deviations theorem [2, Ch. 8, Th. 2] that gives the quantile value τ_n given that $\mathbb{P}\left[\imath(\mathbf{X};\mathbf{Y}) \leq \tau_n\right]$ equals a target MD sequence. We apply the large deviations result in [26, Th. 3.4] to bound the probability $\mathbb{P}\left[\imath(\bar{\mathbf{X}};\mathbf{Y}) \geq \imath(\mathbf{X};\mathbf{Y}) \geq \tau_n\right]$, where $\mathbf{X}$ and $\bar{\mathbf{X}}$ denote a transmitted codeword and an independent codeword drawn from the same distribution, respectively. This bound replaces the bounds in [18, eq. (7.25)-(7.27)] and refines the large deviations bound [3, Lemma 47] used in [16, Th. 53]. We show an achievability result as a function of $I(P_X, P_{Y|X})$, $V_{\mathrm{u}}(P_X, P_{Y|X})$, and $\mathrm{Sk}_{\mathrm{u}}(P_X, P_{Y|X})$. If $P_X = P_X^* \in \mathcal{P}^*$, the resulting bound is (38) with $A_0(P_X^*)$ replaced by zero. We then optimize the bound over P_X using the second-, first- and zeroth-order Taylor series expansions around $P_X^* \in \Pi^*$ of $I(P_X, P_{Y|X})$, $V_{\mathrm{u}}(P_X, P_{Y|X})$, and $\mathrm{Sk}_{\mathrm{u}}(P_X, P_{Y|X})$, respectively. Interestingly, the right-hand side of (38) is achieved using i.i.d. random codewords drawn from

$$P_X = P_X^* - \frac{Q^{-1}(\epsilon_n)}{2\sqrt{nV_{\epsilon_n}}} \tilde{\jmath}\mathbf{v}(P_X^*) \in \mathcal{P} \tag{39}$$

instead of a dispersion-achieving input distribution $P_X^* \in \mathcal{P}^*$. In the second-order MD result in [13], Altuğ and Wagner apply the non-asymptotic bound in [12, Cor. 2 on p. 140], which turns out to be insufficiently sharp for the derivation of the third-order term. ∎

Theorem 2: Under the conditions of Theorem 1,

$$\zeta(n,\epsilon_n) \leq \frac{1}{2}\log n + Q^{-1}(\epsilon_n)^2 \max_{P_X^* \in \mathcal{P}^*} \left(\frac{\mathrm{Sk}_{\mathrm{u}}(P_X^*)\sqrt{V_{\epsilon_n}}}{6} + \frac{1}{2} + A_0(P_X^*) - A_1(P_X^*) \right) + O\left(\frac{Q^{-1}(\epsilon_n)^3}{\sqrt{n}}\right) + O(1). \tag{40}$$

Proof: The proof of Theorem 2 combines the converse bound from [17, Prop. 6], which is derived from the meta-converse bound [3, Th. 27], and a saddlepoint result in [18, Lemma 14], which involves a maximization over an input distribution $P_X \in \mathcal{P}$ and a minimization over an auxiliary output distribution $Q_Y \in \mathcal{Q}$. Combining these results and not deriving the $O(1)$ term in (40) yield a much simpler proof than that in [18]. While [18, proof of Th. 4] relies on the asymptotic expansion of the Neyman-Pearson Lemma, i.e., the $\beta_{1-\epsilon}(P,Q)$ function defined in [3, eq. (100)], the use of [17, Prop. 6] allows us to bypass this part. After carefully choosing the parameter δ in [17, Prop. 6], the problem reduces to a max-imin problem involving the quantities $D(P_{Y|X}\|Q_Y|P_X)$ and $V(P_{Y|X}\|Q_Y|P_X)$, where the maximization is over $P_X \in \mathcal{P}$ and the minimization is over $Q_Y \in \mathcal{Q}$. Then, similar to the steps in [18, eq. (8.22)], for the maximization over P_X, we separate the cases where $\|P_X - P_X^*\|_2 = c_0 \frac{Q^{-1}(\epsilon_n)}{\sqrt{n}}$ or not, where $P_X^* \in \mathcal{P}^*$ and $c_0 > 0$. Applying [18, Lemmas 14 and 9-iii] completes the proof. ∎

The constant terms $\underline{B}$ and $\overline{B}$ in [18] differ depending on whether the information density random variable $\imath(X;Y)$ is a lattice or non-lattice random variable because both the Edgeworth expansion and the large deviation result used in [18] take distinct forms for lattice and non-lattice random variables. The BSC is analyzed separately in [18, Th. 7] since

[1] There is a sign error in [18, eq. (3.1)-(3.2)], which is carried out throughout the paper. The sign of the terms with $S(P_X)$ should be positive instead of negative in both equations. The error in the achievability result originates in [18, eq. (7.15) and (7.19)], where it is missed that $\mathrm{Sk}(-X) = -\mathrm{Sk}(X)$ for any random variable X. The error in the converse result stems from the same sign error in [18, eq. (6.8)].

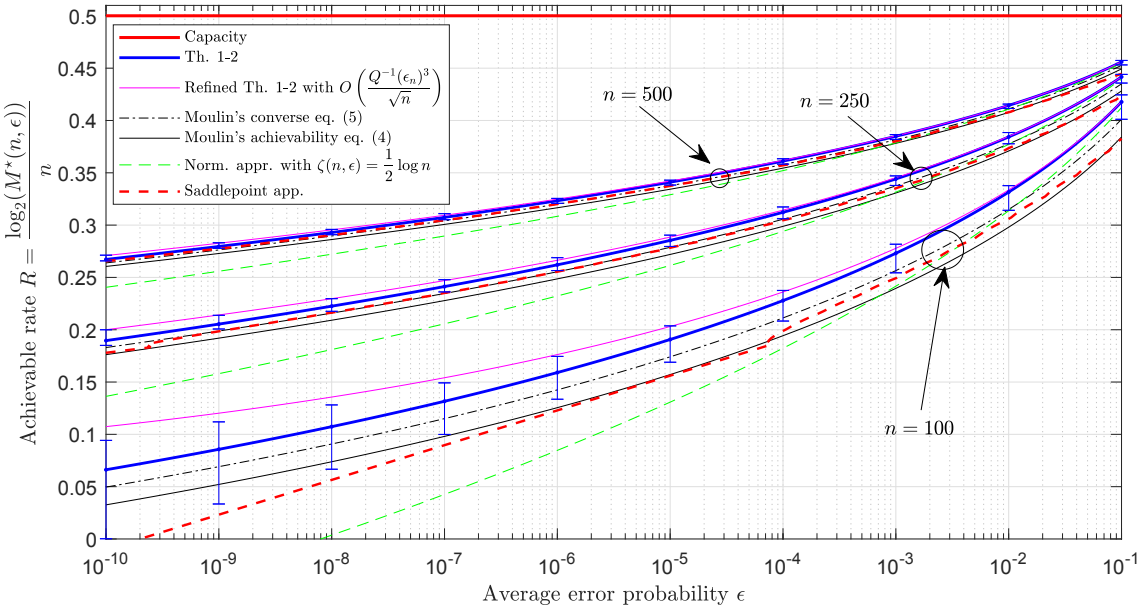

Fig. 1. The expansion from Theorems 1 and 2, excluding the $O(\cdot)$ terms, is shown for the BSC(0.11) with $\epsilon \in [10^{-10}, 10^{-1}]$ and $n = \{100, 250, 500\}$. The error bars correspond to the non-asymptotic achievability and converse bounds from [3, Th. 33 and 35]; the normal approximation, which achieves $\zeta(n,\epsilon) = \frac{1}{2}\log n$, is from [16, Th. 53]; Moulin's results are from [18, Th. 7]; the saddlepoint approximation for the achievability bound is from [21], [22].

the information density for the BSC is lattice. A single proof holds for lattice and non-lattice cases if we do not attempt to bound the $O(1)$ term as in this paper.

A. The tightness of Theorem 1 and Theorem 2

If the channel satisfies $|\mathcal{P}^*| = 1$, $A_0(P_X^*) = A_1(P_X^*) = 0$, and $\eta(P_X^*) = 0$, then achievability (38) and converse (40) bounds yield the channel skewness term

$$S = \frac{\mathrm{Sk_u}(P_X^*)\sqrt{V_{\min}}}{6} + \frac{1}{2}. \tag{41}$$

Cover-Thomas symmetric channels [23, p. 190] satisfy these conditions;[2] the BSC is an example. Further, if ϵ_n satisfies $Q^{-1}(\epsilon_n) = O(n^{1/6})$, then the $O\left(\frac{Q^{-1}(\epsilon_n)^3}{\sqrt{n}}\right)$ in (38) and (40) is dominated by the $O(1)$ term, giving that for Cover-Thomas symmetric channels, $\zeta(n,\epsilon_n) = \frac{1}{2}\log n + S\,Q^{-1}(\epsilon_n)^2 + O(1)$. For the BSC with crossover probability 0.11, Fig. 1 compares asymptotic expansions for the maximum achievable rate, $\frac{\log_2 M^*(n,\epsilon_n)}{n}$, dropping $o(\cdot)$ and $O(\cdot)$ terms except where noted otherwise. The curves plotted in Fig. 1 include Theorems 1 and 2 both with and without the leading term of $O\left(\frac{Q^{-1}(\epsilon_n)^3}{\sqrt{n}}\right)$ computed,[3] various other asymptotic expansions in the CLT and LD regimes, and the non-asymptotic bounds from [3, Th. 33 and 35]. Unlike the normal approximation from [16, Th. 53] and Theorems 1 and 2 with the leading term of $O\left(\frac{Q^{-1}(\epsilon_n)^3}{\sqrt{n}}\right)$, Theorems 1 and 2 without the $O(\cdot)$ terms lie in between the non-asymptotic bounds from [3] for all (n,ϵ) pairs shown, which highlights the accuracy of the channel skewness in explaining the fundamental limit of channel coding.

In [19], Altuğ and Wagner show that in the LD regime, for Gallager-symmetric channels, the prefactors in the lower and upper bounds on the error probability have the same order; that order depends on whether the channel is singular or non-singular. Extending the analysis in [18, Sec. III-C-2)] to any Gallager-symmetric channel shows that Gallager-symmetric channels satisfy $A_0(P_X^*) = A_1(P_X^*) = 0$; $\eta(P_X^*)$ is not necessarily zero (see [18, Sec. III-C-2)] for a counterexample), which means that (38) and (40) are not tight up to the $O(1)$ term for some Gallager-symmetric channels. The findings in [19] suggest that Theorem 1 or Theorem 2 or both could be improved for some channels. The main difference between the achievability bounds in [19], [20] and ours is that [20] bounds the error probability by

$$\mathbb{P}\left[\mathcal{D}\right] + (M-1)\mathbb{P}\left[\mathcal{D}^{\mathrm{c}} \cap \left\{\imath(\bar{\mathbf{X}};\mathbf{Y}) \geq \imath(\mathbf{X};\mathbf{Y})\right\}\right], \tag{42}$$

where

$$\mathcal{D} \triangleq \left\{\log \frac{P_{Y|X}^n(\mathbf{Y}|\mathbf{X})}{Q_Y^n(\mathbf{Y})} < \tau_n\right\} \tag{43}$$

$$Q_Y(y) \triangleq c\left(\sum_{x\in\mathcal{X}} P_X(x) P_{Y|X}(y|x)^{1/1+\rho}\right)^{1+\rho}, \quad y \in \mathcal{Y}. \tag{44}$$

Here Q_Y is the tilted output distribution, $\rho \in [0,1]$ and c are some constants, and τ_n is a sequence. Our achievability bound uses a special case of (44) with $\rho = 0$, giving $Q_Y = P_Y$. Whether the more general bound in (44) yields an improved bound in the MD regime is a question for future work.

[2]Channels that (i) are Cover-Thomas weakly symmetric, (ii) have $|\mathcal{X}| = |\mathcal{Y}|$, and (iii) have a positive definite J satisfy the same conditions [18, Prop. 6].

[3]In general, $O\left(\frac{Q^{-1}(\epsilon_n)^3}{\sqrt{n}}\right)$ is in the form of $\sum_{i=1}^{\infty} c_i \frac{Q^{-1}(\epsilon_n)^{i+2}}{n^{i/2}}$. The leading term refers to the first term in that infinite series. See [24, Sec. VI] for its derivation.

REFERENCES

[1] H. Cramér, "Sur un nouveau théorème-limite de la théorie des probabilités," *Actualités Sci. Ind.*, vol. 736, pp. 2–23, 1938.

[2] V. V. Petrov, *Sums of independent random variables*. New York, USA: Springer, Berlin, Heidelberg, 1975.

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

[4] V. Strassen, "Asymptotische abschätzugen in Shannon's informationstheorie," in *Trans. Third Prague Conf. Inf. Theory*, Prague, 1962, pp. 689–723.

[5] 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.

[6] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Feedback in the non-asymptotic regime," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 4903–4925, Aug. 2011.

[7] V. Kostina and S. Verdú, "Fixed-length lossy compression in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 58, no. 6, pp. 3309–3338, June 2012.

[8] I. Kontoyiannis and S. Verdú, "Optimal lossless data compression: Non-asymptotics and asymptotics," *IEEE Trans. Inf. Theory*, vol. 60, no. 2, pp. 777–795, Feb. 2014.

[9] V. Y. Tan and O. Kosut, "On the dispersions of three network information theory problems," *IEEE Trans. Inf. Theory*, vol. 60, no. 2, pp. 881–903, Feb. 2014.

[10] Y. Polyanskiy, "A perspective on massive random-access," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, June 2017, pp. 2523–2527.

[11] R. C. Yavas, V. Kostina, and M. Effros, "Random access channel coding in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 67, no. 4, pp. 2115–2140, Apr. 2021.

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

[13] Y. Altuğ and A. B. Wagner, "Moderate deviations in channel coding," *IEEE Trans. Inf. Theory*, vol. 60, no. 8, pp. 4417–4426, Aug. 2014.

[14] Y. Polyanskiy and S. Verdú, "Channel dispersion and moderate deviations limits for memoryless channels," in *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sep. 2010, pp. 1334–1339.

[15] C. Chubb, V. Y. F. Tan, and M. Tomamichel, "Moderate deviation analysis for classical communication over quantum channels," *Commun. Math. Phys.*, vol. 355, p. 1283–1315, Aug. 2017.

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

[17] M. Tomamichel and V. Y. F. Tan, "A tight upper bound for the third-order asymptotics of discrete memoryless channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Istanbul, Turkey, July 2013, pp. 1536–1540.

[18] P. Moulin, "The log-volume of optimal codes for memoryless channels, asymptotically within a few nats," *IEEE Trans. Inf. Theory*, vol. 63, no. 4, pp. 2278–2313, Apr. 2017.

[19] Y. Altuğ and A. B. Wagner, "On exact asymptotics of the error probability in channel coding: Symmetric channels," *IEEE Trans. Inf. Theory*, vol. 67, no. 2, pp. 844–868, Feb. 2021.

[20] ——, "Refinement of the random coding bound," *IEEE Trans. Inf. Theory*, vol. 60, no. 10, pp. 6005–6023, Oct. 2014.

[21] J. Honda, "Comprehensive analysis on exact asymptotics of random coding error probability," *arXiv:1707.04401*, July 2017.

[22] J. Font-Segura, A. Martinez, and A. G. i Fàbregas, "Asymptotics of the random coding union bound," in *Int. Symp. Inf. Theory and Its Applications (ISITA)*, Singapore, Oct. 2018, pp. 125–129.

[23] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. NJ, USA: Wiley, 2006.

[24] R. C. Yavas, V. Kostina, and M. Effros, "Third-order analysis of channel coding in the moderate deviations regime," *arXiv:2203.01418*, Mar. 2022.

[25] Y. Altuğ and A. B. Wagner, "The third-order term in the normal approximation for singular channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Honolulu, HI, USA, July 2014, pp. 1897–1901.

[26] N. R. Chaganty and J. Sethuraman, "Multidimensional strong large deviation theorems," *Journal of Statistical Planning and Inference*, vol. 55, no. 3, pp. 265–280, Nov. 1996.