Moderate Deviations for Joint Source-Channel Coding of Systems With Markovian Memory

Vincent Y. F. Tan[⋆], Shun Watanabe[†], Masahito Hayashi[‡]

[⋆]Department of Electrical and Computer Engineering, National University of Singapore. Email: vtan@nus.edu.sg
[†]Department of Information Science and Intelligent Systems, University of Tokushima. Email: shun-wata@is.tokushima-u.ac.jp
[‡]Graduate School of Mathematics, Nagoya University, and CQT, NUS, Singapore. Email: masahito@math.nagoya-u.ac.jp

***Abstract*—We study the (almost lossless) joint source-channel coding problem from the moderate deviations perspective where the bandwidth expansion ratio tends towards the ratio of the channel capacity and source entropy at a rate larger than $n^{-1/2}$ (n being the channel blocklength) and the error probability decays subexponentially. We consider the stationary ergodic Markov (SEM) source as well as discrete memoryless and additive SEM channels. We also discuss the loss due to separation in the moderate deviations setting.**

I. Introduction

We study the (almost lossless) joint source-channel coding (JSCC) problem in which a source of block length k is transmitted over a channel using a code of blocklength n. The study of JSCC has a long history, and here we only refer to literatures that are directly related to our work. For the discrete memoryless source (DMS) and the discrete memoryless channel (DMC), Gallager derived an achievable bound for the error exponent [1, Prob. 5.16]. In [2], Csiszár derived an alternative achievable bound for the error exponent, and he also derived a converse bound for the error exponent. Csiszár's bounds were shown to be tight when a certain rate is above the critical rate of the DMC. Zhong-Alajaji-Campbell derived an achievable bound and a converse bound for the error exponent in the DMS and DMC case [3] and for the stationary ergodic Markov (SEM) source and the additive SEM noise channel [4]. They also compared their exponents to the tandem setting, i.e. when source and channel coding are done separately. They showed that the error exponent for JSCC E_{J}, while generally larger than the error exponent in the tandem setting E_{T}, does not exceed $2E_{\mathrm{T}}$.

The error exponent is the central object of study in the so-called large deviation (LD) regime. Other than the LD regime, the second- and third-order regime have also been attracting attention in information theory lately [5]–[8]. For DMS and DMC, Wang-Ingber-Kochman [9] and Kostina-Verdú [10] derived the second-order result of JSCC.[1] Their result states that a JSCC code with error probability smaller than ε exists if and only if the bandwidth expansion ratio satisfies

$$\frac{k}{n} = \frac{C(\mathbf{W})}{H(P_{\mathbf{S}})} - \sqrt{\frac{V(\mathbf{W},P_{\mathbf{S}})}{n}}Q^{-1}(\varepsilon) + O\left(\frac{\log n}{n}\right), \quad (1)$$

where $V(\mathbf{W},P_{\mathbf{S}})$ is the JSC dispersion defined later.

[1] In fact, they also derived the result for lossy JSCC.

In this paper, we study the moderate deviation (MD) regime of JSCC, i.e., the bandwidth expansion ratio behaves like

$$r_n := \frac{k}{n} = \frac{C(\mathbf{W})}{H(P_{\mathbf{S}})} - \epsilon_n \quad (2)$$

for some ϵ_n that vanishes slower than $n^{-1/2}$. We consider SEM source and both DMC or the additive SEM channel. The reason to study the MD regime is clear. In communication systems, we want to maximize r_n so that it reaches its fundamental limit $C(\mathbf{W})/H(P_{\mathbf{S}})$ *and* minimize the error probability *at the same time*. In the second-order studies (1) the error probability ε is bounded away from zero while in the LD regime, $\limsup_{n\to\infty} r_n$ is bounded away from $C(\mathbf{W})/H(P_{\mathbf{S}})$.

The MD regime interpolates between the LD and second-order regimes, and has been studied for several information-theoretic problems. Altuğ-Wagner studied MD for the channel coding of DMCs with positive[2] entries [11]. This was subsequently extended by Polyanskiy-Verdú to general DMCs [12]. He *et al.* [13] and Kuzuoka [14] studied MD for Slepian-Wolf coding. Sason studied MD for hypothesis testing [15]. One of the authors studied MD of lossy source coding for DMS and Gaussian sources [16]. The other authors studied MD of source coding for SEM and channel coding for additive SEM [17].

One of difficulties in MD analysis is that the error probability may converge to zero very slowly and thus the standard arguments of the method of type does not work since polynomial factors of the blocklength cannot be ignored. Our converse proof is based on Csiszár's idea to split the JSCC error probability into the source and channel error probabilities [2, Lem. 2]. Although we use the method of (Markov) types [18], [19], we need a delicate argument to avoid the aforementioned difficulty, which is one of technical contributions of this paper (cf. Sec. IV-C for details).

The rest of the paper is organized as follows: we introduce the problem setup in Sec. II. The main results and their proofs are given in Sec. III and Sec. IV respectively. Sec. V discusses the loss due to a separation (or tandem [3], [4]) coding scheme. We conclude in Sec. VI and discuss future work.

II. Problem Setup and Definitions

Let $\mathcal{X},\mathcal{Y}$ and $\mathcal{S}$ be finite alphabets. We consider throughout this paper a communication system with *transmission rate* (or

[2] In the full paper version of [11], the positivity condition was removed.

bandwidth expansion ratio) r_n in source symbols per channel use. This consists of a source represented by k-dimensional distributions $P_{\mathbf{S}} = \{P_{S^k} \in \mathcal{P}(\mathcal{S}^k)\}_{k=1}^{\infty}$ and a discrete channel represented by a sequence of n-dimensional transition matrices $\mathbf{W} = \{W^n : \mathcal{X}^n \to \mathcal{Y}^n\}_{n=1}^{\infty}$ where $r_n := k/n$. A (k,n)-*joint source-channel (JSC) code* consists of an encoder $f_k : \mathcal{S}^k \to \mathcal{X}^n$ and a decoder $\varphi_n : \mathcal{Y}^n \to \mathcal{S}^k$. The *error probability* of a JSC code (f_k, φ_n) is given by

$$e(f_k, \varphi_n) := \sum_{\mathbf{s} \in \mathcal{S}^k} P_{S^k}(\mathbf{s}) W^n(\mathcal{Y}^n \setminus \varphi_n^{-1}(\mathbf{s}) \mid f_k(\mathbf{s})). \tag{3}$$

The minimum error probability over all (k,n)-JSC codes is denoted as $e(f_k^*, \varphi_n^*)$.

We consider two different models for the channel. First, we assume that $\mathbf{W}$ is a DMC with positive capacity and dispersion. In this case, $W^n(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^n W(y_i|x_i)$. Second, we assume that $\mathbf{W}$ represents a discrete channel with additive noise in which case $\mathcal{X} = \mathcal{Y} = \{0, 1, \dots, B-1\}$ and the noise, with alphabet $\mathcal{Z} = \mathcal{X}$, is represented by n-dimensional distributions $P_{\mathbf{Z}} = \{P_{Z^n} \in \mathcal{P}(\mathcal{Z}^n)\}_{n=1}^{\infty}$ such that

$$Y_i = X_i \oplus Z_i \pmod{B}. \tag{4}$$

We assume that $P_{\mathbf{Z}}$ represents SEM noise as described and analyzed in [4]. In particular, for any tuple $\mathbf{z} \in \mathcal{Z}^n$, we have

$$P_{Z^n}(z^n) = P_{Z_1}(z_1) \prod_{i=2}^n \Gamma_Z(z_i|z_{i-1}) \tag{5}$$

for some initial distribution P_{Z_1} and transition distribution Γ_Z which is assumed to be irreducible. We call this the *additive SEM channel*. Let $\tilde{P}_Z$ be the stationary distribution of $P_{\mathbf{Z}}$. For $P_{\mathbf{S}}$, we assume it is a discrete SEM source with stationary distribution $\tilde{P}_S$ and the transition matrix Γ_S is also positive everywhere. This subsumes the case where $P_{\mathbf{S}}$ is a discrete memoryless source (DMS).

We now define several information quantities. For a SEM source $P_{\mathbf{S}}$, its *entropy rate* is $H(P_{\mathbf{S}}) := \sum_{s'} \tilde{P}(s') \sum_s \Gamma_S(s|s') \log \frac{1}{\Gamma_S(s|s')}$ [20, Thm. 4.2.4], where $\tilde{P}$ is the stationary distribution of the transition matrix Γ_S. For transition matrix $\Gamma_S(s|s')$ and parameter $\theta \in (-1, \infty)$, we introduce the *tilted matrix* $\Gamma_{S,\theta}(s|s') = \Gamma_S(s|s')^{1+\theta}$. Then, the normalized (in the sense of probability) eigenvector corresponding to the Perron-Frobenius eigenvalue [21] of $\Gamma_{S,\theta}$ is denoted as $\{\tilde{P}_{S,\theta}(s)\}_{s \in \mathcal{S}}$. The *Rényi entropy* rate is given by

$$H_{1+\theta}(P_{\mathbf{S}}) = -\frac{1}{\theta} \log \sum_{s,s'} \tilde{P}_{S,\theta}(s') \Gamma_{S,\theta}(s|s'). \tag{6}$$

The *dispersion* of the SEM source is defined as $V(P_{\mathbf{S}}) := \lim_{\theta \to 0} \frac{2[H(P_{\mathbf{S}}) - H_{1+\theta}(P_{\mathbf{S}})]}{\theta}$ [17, Eq. (20) and Thm. 14], which is assumed to be positive. Note that in the special case that the source is a DMS [5], $V(P_{\mathbf{S}}) = \mathsf{Var}(-\log P_S(S))$.

For a DMC $W : \mathcal{X} \to \mathcal{Y}$, the *channel capacity* is $C(\mathbf{W}) := \max_P I(P, W)$ and the *channel dispersion* [6], [7] is $V(\mathbf{W}) := \min \sum_x P^*(x) \mathsf{Var}\left(\log \frac{W(\cdot|x)}{P^* W(\cdot)}\right)$, where the minimum is taken over all capacity-achieving input distribution P^*. For an additive SEM channel with $P_{\mathbf{Z}}$, the capacity [4] is $C(\mathbf{W}) = \log B - H(P_{\mathbf{Z}})$ and the dispersion is $V(\mathbf{W}) = V(P_{\mathbf{Z}})$ [17, Thm. 54].

For the MD regime, we assume that r_n converges to the asymptotic limit $C(\mathbf{W})/H(P_{\mathbf{S}})$ as in (2). Furthermore, the *backoff sequence* ϵ_n satisfies the two conditions

$$\lim_{n \to \infty} \epsilon_n = 0, \qquad \lim_{n \to \infty} n\epsilon_n^2 = \infty. \tag{7}$$

This setting "interpolates" between the LD (error exponent) regime [2]–[4] where ϵ_n is a constant and the central limit (or second-order) regime [9], [10] where $\epsilon_n = an^{-1/2}$ for $a \in \mathbb{R}$.

III. Main Results

Let us define a fundamental quantity before stating our main theorem. If the source has entropy rate $H(P_{\mathbf{S}})$ and dispersion $V(P_{\mathbf{S}})$ and the channel has capacity $C(\mathbf{W})$ and dispersion $V(\mathbf{W})$, define the *JSC dispersion*

$$V(\mathbf{W}, P_{\mathbf{S}}) := \frac{1}{H(P_{\mathbf{S}})^2} \left[V(\mathbf{W}) + \frac{C(\mathbf{W})}{H(P_{\mathbf{S}})} V(P_{\mathbf{S}}) \right]. \tag{8}$$

Theorem 1 (Moderate Deviations for JSC Coding)**.** *Let $P_{\mathbf{S}}$ be a SEM source and $\mathbf{W}$ be either a DMC (under the assumptions in Sec. II) or an additive SEM. Then,*

$$\lim_{n \to \infty} -\frac{1}{n\epsilon_n^2} \log e(f_k^*, \varphi_n^*) = \frac{1}{2V(\mathbf{W}, P_{\mathbf{S}})} \tag{9}$$

IV. Proof of Theorem 1

We first consider the case where $\mathbf{W}$ is a DMC. Subsequently, we make the necessary changes to handle the case when $\mathbf{W}$ is an additive SEM channel. Finally, we provide some remarks concerning the proof.

A. Channel is a DMC

Direct Part: Define *Gallager's channel function* [1, Sec. 5.4] for the DMC $W : \mathcal{X} \to \mathcal{Y}$ as

$$E_{\mathrm{o}}(\tau, P, W) := -\log \sum_x P(x) \left(\sum_y W(y|x)^{\frac{1}{1+\tau}} \right)^{1+\tau}, \tag{10}$$

where $0 \le \tau \le 1$ and $P \in \mathcal{P}(\mathcal{X})$. Then [1, Prob. 5.16] tells us that there exists a (k,n)-JSC code (f_k, φ_n) whose error probability $e(f_k, \varphi_n)$ is upper bounded by

$$\left(\sum_{\mathbf{s} \in \mathcal{S}^k} P_{S^k}(\mathbf{s})^{\frac{1}{1+\tau}} \right)^{1+\tau} \sum_{\mathbf{y} \in \mathcal{Y}^n} \left(\sum_{\mathbf{x} \in \mathcal{X}^n} P^n(\mathbf{x}) W^n(\mathbf{y}|\mathbf{x})^{\frac{1}{1+\tau}} \right)^{1+\tau} \tag{11}$$

for any $0 \le \tau \le 1$ and $P \in \mathcal{P}(\mathcal{X})$. Using the reparametrization $\tau = \frac{-\theta}{1+\theta}$ for the source term and the approximation of the Rényi entropy [17, Lem. 5], we see that (11) can be overbounded by

$$\exp\left(\frac{-(k-1)\theta H_{1+\theta}(P_{\mathbf{S}}) - \underline{\delta}(\theta)}{1+\theta} \right) \exp\big(-nE_{\mathrm{o}}(\tau, P, W) \big) \tag{12}$$

where $H_{1+\theta}(P_{\mathbf{S}})$ is defined in (6) and $\underline{\delta}(\theta)$ is a constant that does not depend on n. We know that by Taylor expanding

$\theta \mapsto \theta H_{1+\theta}(P_{\mathbf{S}})$ and $\tau \mapsto E_{\mathrm{o}}(\tau, P, W)$ in the neighborhood of $\theta, \tau = 0$ that

$$\theta H_{1+\theta}(P_{\mathbf{S}}) = \theta H(P_{\mathbf{S}}) - \frac{\theta^2}{2} V(P_{\mathbf{S}}) + O(\theta^3), \tag{13}$$

$$E_{\mathrm{o}}(\tau, P, W) = \tau I(P, W) - \frac{\tau^2}{2} V(P, W) + O(\tau^3). \tag{14}$$

The first Taylor expansion is by [17, Eqs. (19) and (20)] and the second by [11, Lem. 2.1]. We now set P to be capacity-achieving, meaning that $I(P, W) = C(\mathbf{W})$ and $V(P, W) = V(\mathbf{W})$. Plugging these expansions into the exponent in (12), using the rate relation in (2) and the fact that $\tau = \frac{-\theta}{1+\theta}$, we obtain

$$e(f_k, \varphi_n) \le \exp\Bigg(- n\Big[- \frac{\tau^2}{2}\Big(V(\mathbf{W}) + \frac{C(\mathbf{W})}{H(P_{\mathbf{S}})} V(P_{\mathbf{S}})\Big) + \epsilon_n \tau H(P_{\mathbf{S}}) + \frac{\epsilon_n \tau^2}{2} V(P_{\mathbf{S}}) + O(\tau^3)\Big]\Bigg). \tag{15}$$

Now we set τ to be

$$\tau := \frac{\epsilon_n H(P_{\mathbf{S}})}{V(\mathbf{W}) + \frac{C(\mathbf{W})}{H(P_{\mathbf{S}})} V(P_{\mathbf{S}})} \tag{16}$$

which is in $[0, 1]$ for large enough n. Uniting (15)–(16) yields

$$e(f_k, \varphi_n) \le \exp\Bigg(- n\Big[\frac{\epsilon_n^2}{2V(\mathbf{W}, P_{\mathbf{S}})} + O(\epsilon_n^3)\Big]\Bigg). \tag{17}$$

By taking the logarithm and normalizing by $-n\epsilon_n^2$, we obtain

$$\liminf_{n\to\infty} -\frac{1}{n\epsilon_n^2} \log e(f_k, \varphi_n) \ge \frac{1}{2V(\mathbf{W}, P_{\mathbf{S}})}, \tag{18}$$

which completes the proof of the direct part.

Converse Part: For the converse, we make use of the method of types for Markov sources [18], [19]. For a string $\mathbf{s} \in \mathcal{S}^k$, let $k_{ij}(\mathbf{s})$ be the number of transitions from $i \in \mathcal{S}$ to $j \in \mathcal{S}$ in $\mathbf{s}$ with the cyclic convention that s_1 follows s_k. The matrix $[k_{ij}(\mathbf{s})/k]_{i,j\in\mathcal{S}}$ is called the *Markov type* of $\mathbf{s}$. The set of all Markov types formed from length-k sequences is denoted as $\mathcal{P}_k^{(2)}(\mathcal{S})$. Let $\mathcal{T}_P^k$ be the set of all length-k sequences with Markov type P, i.e. the *Markov type class*. We may further partition $\mathcal{T}_P^k$ into the following subsets:

$$\mathcal{T}_P^k(i, j) := \big\{\mathbf{s} = (s_1, \ldots, s_k) \in \mathcal{T}_P^k : s_1 = i, s_k = j\big\}. \tag{19}$$

All sequences in a given subset are equiprobable under the SEM source P_{S^k}.

We define for some $R > 0$ to be specified later,

$$\mathcal{P}_k^{(2)}(\mathcal{S}; R) := \big\{(i, j, P) \in \mathcal{S}^2 \times \mathcal{P}_k^{(2)}(\mathcal{S}) : |\mathcal{T}_P^k(i, j)| \le 2^{kR}\big\}. \tag{20}$$

Let $\mathrm{P}_e^*(M; P_S)$ to be the error probability of the optimal source code of size smaller than or equal to M for source P_S. We have the following lemma which is proven in Appendix A.

Lemma 2. *We have*

$$\mathrm{P}_e^*\left(|\mathcal{S}|^2(k+1)^{|\mathcal{S}|^2} 2^{kR}; P_{S^k}\right) \le \sum_{(i,j,P)\notin\mathcal{P}_k^{(2)}(\mathcal{S};R):\mathcal{T}_P^k(i,j)\neq\emptyset} \; \sum_{\tilde{\mathbf{s}}\in\mathcal{T}_P^k(i,j)} P_{S^k}(\tilde{\mathbf{s}}). \tag{21}$$

Let $\mathrm{P}_e^*(M; W)$ to be the average error probability of the optimal channel code of size larger than or equal to M for channel W. By using Lemma 2, we have the following lemma which says that the JSC error can be decomposed into a source and channel error. See Appendix B for the proof.

Lemma 3. *For any $R \ge 0$, we have*

$$e(f_k^*, \varphi_n^*) \ge \mathrm{P}_e^*\left(|\mathcal{S}|^2(k+1)^{|\mathcal{S}|^2} 2^{kR}; P_{S^k}\right) \cdot \mathrm{P}_e^*\left(2^{kR}; W^n\right). \tag{22}$$

Now we can lower bound the error probabilities of source coding and channel coding separately. From the MD result for single-terminal almost lossless source coding [17, Thm. 13], we have the following:

Lemma 4. *Let $H(P_{\mathbf{S}})$ be the entropy rate of the Markov chain. Let $R := H(P_{\mathbf{S}}) + \delta\epsilon_n$ for some $\delta > 0$. For arbitrarily small $\xi > 0$ and sufficiently large n, we have*[3]

$$-\log \mathrm{P}_e^*\left(|\mathcal{S}|^2(k+1)^{|\mathcal{S}|^2} 2^{kR}; P_{S^k}\right) \le \frac{(\delta+\xi)^2}{2V(P_{\mathbf{S}})} k\epsilon_n^2. \tag{23}$$

On the other hand, from Wolfowitz's strong converse [1, Thm. 5.8.5] and Haroutunian's sphere-packing bound [22] (see also [11, Eq. (24)]) we have the following:

Lemma 5. *For every $\gamma > 0$ and $\psi < \infty$, there exists a sufficiently large n such that*

$$\mathrm{P}_e^*(M; W^n) \ge \exp\left(-\frac{n}{1-\gamma}\Big[E\Big(\frac{\log M}{n} - \frac{\psi}{\sqrt{n}}\Big) + \frac{1}{n}\Big]\right), \tag{24}$$

where $E(R) := \max_P \min_{V: I(P,V)\le R} D(V\|W|P)$ is the sphere-packing exponent of the DMC W.

By combining Lemmas 3, 4 and 5 and by setting $R = H(P_{\mathbf{S}}) + \delta\epsilon_n$, we have

$$\begin{aligned} &e(f_k^*, \varphi_n^*) \\ &\ge \exp\left(- \frac{(\delta+\xi)^2}{2V(P_{\mathbf{S}})} k\epsilon_n^2\right) \\ &\quad \cdot \exp\left(-\frac{n}{1-\gamma}\Big[E\Big(\frac{k(H(P_{\mathbf{S}}) + \delta\epsilon_n)}{n} - \frac{\psi}{\sqrt{n}}\Big) + \frac{1}{n}\Big]\right) \end{aligned} \tag{25}$$

Now, we approximate the sphere-packing exponent as in [6, Sec. 5] [11, Lem. 4.2] and we also apply (2). Note that δ is a constant and is chosen in the following (cf. (30)). By noting $V(\mathbf{W}) > 0$ and using the Taylor approximation, we have

$$\begin{aligned} &E\Big(\frac{k(H(P_{\mathbf{S}}) + \delta\epsilon_n)}{n} - \frac{\psi}{\sqrt{n}}\Big) \\ &= E\Big(C(\mathbf{W}) + \epsilon_n\Big(\frac{C(\mathbf{W})}{H(P_{\mathbf{S}})}\delta - H(P_{\mathbf{S}})\Big) + \delta\epsilon_n^2 + \frac{\psi}{\sqrt{n}}\Big) && (26) \\ &\le \frac{\epsilon_n^2\big(\frac{C(\mathbf{W})}{H(P_{\mathbf{S}})}\delta - H(P_{\mathbf{S}})\big)^2}{2V(\mathbf{W})} + o(\epsilon_n^2), && (27) \end{aligned}$$

[3] Although ϵ_n in the statement of [17, Thm. 13] is $\epsilon_n = n^{-t}$ for some $t \in (0, 1/2)$, the statement can be extended for arbitrary ϵ_n satisfying (7) without any modification in the proof.

where (27) follows from the fact that $n\epsilon_n^2 \to \infty$ as $n \to \infty$. Uniting (25)–(27), normalizing and taking the limit, we obtain

$$\limsup_{n\to\infty} -\frac{1}{n\epsilon_n^2}\log e(f_k^*,\varphi_n^*) \le \frac{(\delta+\xi)^2 C(\mathbf{W})}{2V(P_{\mathbf{S}})H(P_{\mathbf{S}})} + \frac{\big(\frac{C(\mathbf{W})}{H(P_{\mathbf{S}})}\delta - H(P_{\mathbf{S}})\big)^2}{2(1-\gamma)V(\mathbf{W})}. \tag{28}$$

Since $\xi,\gamma > 0$ are arbitrary, we can let these constants tend to zero. As such the right-hand-side can be expressed as

$$\frac{\delta^2 C(\mathbf{W})}{2V(P_{\mathbf{S}})H(P_{\mathbf{S}})} + \frac{\big(\frac{C(\mathbf{W})}{H(P_{\mathbf{S}})}\delta - H(P_{\mathbf{S}})\big)^2}{2V(\mathbf{W})} \tag{29}$$

By choosing the constant

$$\delta := \frac{C(\mathbf{W})}{\big(\frac{1}{V(P_{\mathbf{S}})}\frac{C(\mathbf{W})}{H(P_{\mathbf{S}})} + \frac{1}{V(\mathbf{W})}\frac{C(\mathbf{W})^2}{H(P_{\mathbf{S}})^2}\big)V(\mathbf{W})} \tag{30}$$

we see, after some algebra, that (29) reduces to $(2V(\mathbf{W},P_{\mathbf{S}}))^{-1}$ as desired. It is worth noting that

$$\frac{C(\mathbf{W})}{H(P_{\mathbf{S}})}\delta - H(P_{\mathbf{S}}) < 0, \tag{31}$$

and thus the approximation in (27) is justified.

B. Channel is an additive SEM

Conceptually, the proof for additive SEM channels is similar to that for the DMC case. For the direct part, the only change involves the approximation of the random coding exponent in (14). This is justified by [17, Thm. 13] since the channel Gallager function reduces to the Gallager function of the single-terminal source coding in the case of additive channels. For the converse part, we use the result in [17, Thm. 53] to bound $\mathrm{P}_e^*\left(2^{kR}; W^n\right)$ in (22). Then, after a similar algebra as in (29) and (30), we can get the desired MD bound.

C. Remarks on the proof

We observe that in the converse part, we basically considered those source sequences belonging to the complement of $\mathcal{P}_k^{(2)}(\mathcal{S};R)$ in the lower bounding of the error probability in Lemma 3. In Zhong-Alajaji-Campbell [4, Thm. 5] and Csiszár [2, Lem. 2], only a single dominant type was identified. This was because the authors focused on finding the error exponent of JSC coding and the analysing performance of the dominant type was sufficient. In our work, if we had just identified the dominant Markov type, we would require the condition $\epsilon_n^2 n/\log n \to \infty$ (cf. [16]) instead of the weaker condition in (7). Thus, the coarser partitioning of sequences into two classes, namely $\mathcal{P}_k^{(2)}(\mathcal{S};R)$ and its complement, allows us to prove a stronger MD result.

V. The Loss Due to Separation

One of Shannon's main contributions is to show that rate-wise, *separating* the tasks of source and channel coding is optimal. However, there is a loss in the LD regime [2]–[4] the the second-order (dispersion) regime [9], [10]. Here we examine the analogue of Theorem 1 if we use a separation or tandem coding scheme. We focus on the DMS-DMC case (with (W,P_S)) noting that the generalization to the SEM source and the additive SEM channel is straightforward.

The error probability for the separation scheme [2] satisfies

$$e_{\mathrm{sep}}(f_k,\varphi_n) \le \exp\left(-n\sup_{R\ge 0}\min\Big\{r_n E_{\mathrm{s}}\Big(\frac{R}{r_n}\Big), E(R)\Big\}\right), \tag{32}$$

where r_n is given in (2), $E_{\mathrm{s}}(R) := \max_{Q:H(Q)\ge R} D(Q\|P)$ is the almost lossless source coding error exponent [23, Prob. 2.7(c)] and $E(R)$ is the channel error exponent. The polynomial factor for the source term may be omitted and the channel term involving $E(R)$ can be obtained using Gallager's method [1, Sec. 5.4]. We parametrize the optimization variable R above as follows:

$$R = C(W) - \eta_n \tag{33}$$

where W is the DMC and η_n is a vanishing sequence that we optimize. We will see that $\eta_n \to 0$ as $n\to\infty$. By using the facts that

$$E_{\mathrm{s}}(H(P_S)+\xi) = \frac{\xi^2}{2V(P_S)} + O(\xi^3), \tag{34}$$

$$E(C(W)-\xi) = \frac{\xi^2}{2V(W)} + O(\xi^3), \tag{35}$$

we can simplify the exponent in (32). Indeed, to find the optimal η_n, we substitute (33)–(35) into the exponent in (32) and equate two contributions (to maximize the minimum). The optimal η_n is found to be a linear function of ϵ_n as follows:

$$\eta_n = \frac{H(P_S)\sqrt{\frac{H(P_S)}{C(W)}}}{\sqrt{\frac{V(P_S)}{V(W)}} + \sqrt{\frac{H(P_S)}{C(W)}}}\,\epsilon_n + o(\epsilon_n) \tag{36}$$

As such, by evaluating $\eta_n^2/(2V(W))$, we can deduce that

$$\liminf_{n\to\infty} -\frac{1}{n\epsilon_n^2}\log e_{\mathrm{sep}}(f_k,\varphi_n) \ge \frac{1}{2V_{\mathrm{sep}}(W,P_S)}, \tag{37}$$

where the *separation JSC dispersion* is

$$V_{\mathrm{sep}}(W,P_S) := V(W,P_S) + \frac{2\sqrt{\frac{C(W)}{H(P_S)}V(W)V(P_S)}}{H(P_S)^2} \tag{38}$$

and $V(W,P_S)$ is the DMS-DMC version of the JSC dispersion in (8). Since $V_{\mathrm{sep}}(W,P_S) > V(W,P_S)$ for channels with positive capacity and dispersion and non-deterministic sources, separation incurs a penalty of the (positive) second term in (38). This also corroborates the derivation by Wang-Ingber-Kochman in [9, Sec. 5] in which the authors used a different method that is based on second-order asymptotics (with excess distortion probability ε tending to zero) instead of starting from the LD technique which we did in (32).

VI. Conclusion and Further Work

Theorem 1 in this paper showed that if the difference between $C(\mathbf{W})/H(P_{\mathbf{S}})$ and the bandwidth expansion ratio r_n is ϵ_n satisfying (7), the optimal error probability is roughly

$$e(f_k^*,\varphi_n^*) \approx \exp\left(-\frac{n\epsilon_n^2}{2V(\mathbf{W},P_{\mathbf{S}})}\right) \tag{39}$$

where $V(\mathbf{W}, P_{\mathbf{S}})$, defined in (8), is a fundamental quantity of the channel and source known as the JSC dispersion.

There are several avenues for further research. Firstly, we can extend the above analysis to the lossy JSC coding setting. This may be achieved using the methods of Wang-Ingber-Kochman [9] or Kostina-Verdú [10]. Second, the proof above hinged on the (Markov) method of types. It may also be possible to use information spectrum methods [24, Sec. 3.8] to obtain the same conclusion as in Theorem 1. This has the added advantage of providing easily computable finite blocklength bounds that hold for all n. Lastly, the use of information spectrum methods will also yield second-order results for the SEM source and general additive SEM channel, which includes the channel coding with SEM side-information. We note that Tomamichel and Tan [25, Thm. 8] have already derived the second-order coding rate of channel coding with SEM side-information using information spectrum methods.

Appendix A
Proof of Lemma 2

Consider a lossless source code that encodes all length-k sequences $\mathbf{s}$ belonging to

$$\bigcup_{(i,j,P)\in\mathcal{P}_k^{(2)}(\mathcal{S};R):\mathcal{T}_P^k(i,j)\neq\emptyset} \mathcal{T}_P^k(i,j). \tag{40}$$

This set has size no larger than $|\mathcal{S}|^2(k+1)^{|\mathcal{S}|^2}2^{kR}$ by the definition of $\mathcal{P}_k^{(2)}(\mathcal{S};R)$ and the fact that the number of Markov types does not exceed $(k+1)^{|\mathcal{S}|^2}$ [18], [19]. Furthermore the probability of error is the P_{S^k}-probability of the complement of the set in (40) which can be upper bounded as (21) by the union bound. This completes the proof. $\square$

Appendix B
Proof of Lemma 3

Define $\mathcal{D}_{\mathbf{s}} := \{\mathbf{y} \in \mathcal{Y}^n : \varphi_n(\mathbf{y}) = \mathbf{s}\}$ be the decoding region for source sequence $\mathbf{s} \in \mathcal{S}^k$. The error probability of any code (f_k, φ_n) can be written as follows:

$$\begin{aligned}
&e(f_k,\varphi_n)\\
&\geq \sum_{\substack{(i,j,P)\notin\mathcal{P}_k^{(2)}(\mathcal{S};R):\\ \mathcal{T}_P^k(i,j)\neq\emptyset}} \sum_{\mathbf{s}\in\mathcal{T}_P^k(i,j)} P_{S^k}(\mathbf{s}) W^n(\mathcal{D}_{\mathbf{s}}^c|f_k(\mathbf{s})) &(41)\\
&= \sum_{\substack{(i,j,P)\notin\mathcal{P}_k^{(2)}(\mathcal{S};R):\\ \mathcal{T}_P^k(i,j)\neq\emptyset}} \left(\sum_{\tilde{\mathbf{s}}\in\mathcal{T}_P^k(i,j)} P_{S^k}(\tilde{\mathbf{s}})\right)\\
&\qquad\cdot \sum_{\mathbf{s}\in\mathcal{T}_P^k(i,j)} \frac{P_{S^k}(\mathbf{s})}{\sum_{\tilde{\mathbf{s}}\in\mathcal{T}_P^k(i,j)} P_{S^k}(\tilde{\mathbf{s}})} W^n(\mathcal{D}_{\mathbf{s}}^c|f_k(\mathbf{s})) &(42)\\
&= \sum_{\substack{(i,j,P)\notin\mathcal{P}_k^{(2)}(\mathcal{S};R):\\ \mathcal{T}_P^k(i,j)\neq\emptyset}} \sum_{\tilde{\mathbf{s}}\in\mathcal{T}_P^k(i,j)} P_{S^k}(\tilde{\mathbf{s}}) e_{\mathcal{T}_P^k(i,j)}(f_k,\varphi_n). &(43)
\end{aligned}$$

In (43), we used the notation $e_{\mathcal{M}}(f_k, \varphi_n)$ to denote the average error probability of a *channel code* with f_k as encoder and φ_n as decoder restricted to the message set $\mathcal{M}$. Now, since $\mathcal{T}_P^k(i,j)$ satisfies $|\mathcal{T}_P^k(i,j)| > 2^{kR}$ for $(i,j,P) \notin \mathcal{P}_k^{(2)}(\mathcal{S};R)$ (cf. the definition of $\mathcal{P}_k^{(2)}(\mathcal{S};R)$ in (20)), we have

$$e_{\mathcal{T}_P^k(i,j)}(f_k,\varphi_n) \geq \mathrm{P}_e^*\left(2^{kR}, W^n\right). \tag{44}$$

Thus, by substituting this inequality into (43) and by using Lemma 2, we complete the proof. $\square$

References

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

[2] I. Csiszár. Joint source-channel error exponent. *Prob. Contr. and Info. Theory*, 9(5):315–328, 1980.

[3] Y. Zhong, F. Alajaji, and L. L. Campbell. On the joint source-channel coding error exponent for discrete memoryless systems. *IEEE Trans. on Inf. Th.*, 52(4):1450–1468, Apr 2006.

[4] Y. Zhong, F. Alajaji, and L. L. Campbell. Joint source-channel coding error exponent for discrete communication systems with Markovian memory. *IEEE Trans. on Inf. Th.*, 53(12):4457–4472, Dec 2007.

[5] V. Strassen. Asymptotische Abschätzungen in Shannons Informationstheorie. In *Trans. Third Prague Conf. Inf. Theory*, pages 689–723, Prague, 1962.

[6] M. Hayashi. Information spectrum approach to second-order coding rate in channel coding. *IEEE Trans. on Inf. Th.*, 55:4947–4966, Nov 2009.

[7] Y. Polyanskiy, H. V. Poor, and S. Verdú. Channel coding rate in the finite blocklength regime. *IEEE Trans. on Inf. Th.*, 56:2307–2359, May 2010.

[8] M. Tomamichel and V. Y. F. Tan. A tight upper bound for the third-order asymptotics of most discrete memoryless channels. *IEEE Trans. on Inf. Th.*, 59(11):7041–7051, Nov 2013.

[9] D. Wang, A. Ingber, and Y. Kochman. The dispersion of joint source-channel coding. In *Allerton Conference*, 2011.

[10] V. Kostina and S. Verdú. Lossy joint source-channel coding in the finite blocklength regime. *IEEE Trans. on Inf. Th.*, 59(5):2545–2575, May 2013.

[11] Y. Altuğ and A. B. Wagner. Moderate deviation analysis of channel coding: Discrete memoryless case. In *Int. Symp. Inf. Th.*, Austin, TX, 2010. `arXiv:1208.1924 [cs.IT]`.

[12] Y. Polyanskiy and S. Verdú. Channel dispersion and moderate deviations limits for memoryless channels. In *Allerton Conference*, 2010.

[13] D.-K. He, L. A. Lastras-Montaño, E.-H. Yang, A. Jagmohan, and J. Chen. On the redundancy of Slepian-Wolf coding. *IEEE Trans. on Inf. Th.*, 55(12):5607–5627, Dec 2009.

[14] S. Kuzuoka. A simple technique for bounding the redundancy of source coding with side information. In *Int. Symp. Inf. Th.*, Boston, MA, 2012.

[15] I. Sason. On refined versions of the Azuma-Hoeffding inequality with applications in information theory. `arXiv:1111.1977`, Nov 2011.

[16] V. Y. F. Tan. Moderate-deviations of lossy source coding for discrete and Gaussian sources. In *Int. Symp. Inf. Th.*, Cambridge, MA, 2012.

[17] M. Hayashi and S. Watanabe. A unified approach to non-asymptotic and asymptotic analyses of information processing on markov chain. `arXiv:1309.7528 [cs.IT]`, Sep 2013.

[18] L. D. Davisson, G. Longo, and A. Sgarro. The error exponent for the noiseless encoding of finite Markov sources. *IEEE Trans. on Inf. Th.*, 27(4):431–438, Jul 1981.

[19] S. Natarajan. Large deviations, hypothesis testing and source coding for finite Markov chains. *IEEE Trans. on Inf. Th.*, 31(3):360–365, 1985.

[20] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

[21] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[22] E. A. Haroutunian. Estimates of the error exponent for the semi-continuous memoryless channel. *Problems of Information Transmission*, 4(4):37–48, 1968.

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

[24] T. S. Han. *Information-Spectrum Methods in Information Theory*. Springer Berlin Heidelberg, Feb 2003.

[25] M. Tomamichel and V. Y. F. Tan. ε-capacities and second-order coding rates for channels with general state. `arXiv:1305.6789 [cs.IT]`, May 2013.