

# Streaming Data Transmission in the Moderate Deviations and Central Limit Regimes

Si-Hyeon Lee, *Member, IEEE*, Vincent Y. F. Tan, *Senior Member, IEEE*, and Ashish Khisti, *Member, IEEE*

**Abstract**—We consider streaming data transmission over a discrete memoryless channel. A new message is given to the encoder at the beginning of each block and the decoder decodes each message sequentially, after a delay of  $T$  blocks. In this streaming setup, we study the fundamental interplay between the rate and error probability in the central limit and moderate deviations regimes and show that i) in the moderate deviations regime, the moderate deviations constant improves over the block coding or non-streaming setup by a factor of  $T$  and ii) in the central limit regime, the second-order coding rate improves by a factor of approximately  $\sqrt{T}$  for a wide range of channel parameters. For both regimes, we propose coding techniques that incorporate a joint encoding of fresh and previous messages. In particular, for the central limit regime, we propose a coding technique with truncated memory to ensure that a summation of constants, which arises as a result of applications of the central limit theorem, does not diverge in the error analysis.

Furthermore, we explore interesting variants of the basic streaming setup in the moderate deviations regime. We first consider a scenario with an erasure option at the decoder, i.e., the decoder can output an erasure symbol instead of a message estimate, and show that both the exponents of the total error and the undetected error probabilities improve by factors of  $T$ . Next, by utilizing the erasure option, we show that the exponent of the total error probability can be improved to that of the undetected error probability (in the order sense) at the expense of a variable decoding delay.

**Index Terms**—Streaming communication, moderate deviations, central limit regime, second-order coding rates, channel dispersion

## I. INTRODUCTION

In many multimedia applications, a stream of data packets is required to be sequentially encoded and decoded under strict latency constraints. For such a streaming setup, both the fundamental limits and optimal schemes can differ from classical communication systems. In recent years, there has been a growing interest in the characterization of fundamental limits for streaming data transmission [2]–[10]. In [2]–[4], coding techniques based on tree codes were proposed for streaming setup with applications to control systems. The works [5]–[7] consider a streaming setup over packet erasure channels motivated by applications to interactive audio and video streaming. These papers focus on adversarial channel models involving burst erasures. In [8], Khisti and Draper

established the optimal diversity-multiplexing tradeoff (DMT) for streaming over a block-fading multiple-input multiple-output channel. In [9], the same authors proposed a coding technique using finite memory for streaming over discrete memoryless channels (DMCs) that attains the same reliability as previously known semi-infinite coding techniques with growing memory. In [10], the error exponent was studied in a streaming setup of distributed source coding. We note that many of the prior works assumed that the code operates in the large deviations regime in which the rate is bounded away from capacity (or the rate pair is strictly inside the optimal rate region for compression problems) and the error probability decays exponentially as the blocklength increases.

Other interesting asymptotic regimes include the central limit and moderate deviations regimes. Let  $n$  denote the blocklength of a single message henceforth. In the central limit regime, the rate approaches to the capacity at a speed proportional to  $\frac{1}{\sqrt{n}}$  and the error probability does not vanish as the blocklength increases. In the moderate deviations regime, the rate approaches to the capacity strictly slower than  $\frac{1}{\sqrt{n}}$  and the error probability decays sub-exponentially fast as the blocklength increases. For block coding problems, both regimes have received a fair amount of attention recently. These works aim to characterize the fundamental interplay between the coding rate and error probability. The most notable early work on channel coding in the central limit regime (also known as second-order asymptotics or the normal approximation regime) is that of Strassen [11], who considered DMCs and showed that the backoff from capacity scales as  $\sqrt{n}$  when the error probability is fixed. Strassen also deduced the constant of proportionality, which is related to the so-called *dispersion* [12]. Hayashi [13] considered DMCs with cost constraints. Polyanskiy *et al.* [12] refined the asymptotic expansions and also compared the normal approximation to the finite blocklength (non-asymptotic) fundamental limits. For a review and extensions to multi-terminal models, the reader is referred to [14]. For the moderate deviations regime, He *et al.* [15] considered fixed-to-variable length source coding with decoder side information. Altuğ and Wagner [16] initiated the study of moderate deviations for channel coding, specifically DMCs. Polyanskiy and Verdú [17] relaxed some assumptions in the conference version of Altuğ and Wagner’s work [18] and they also considered moderate deviations for additive white Gaussian noise (AWGN) channels. However, this line of research has not been extensively studied for the streaming setup. To the best of our knowledge, there has been no prior work on the streaming setup in the moderate deviations and

S.-H. Lee and A. Khisti are with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, Canada (e-mail: sihyeon.lee@utoronto.ca; akhisti@comm.utoronto.ca). V. Y. F. Tan is with the Department of Electrical and Computer Engineering and the Department of Mathematics, National University of Singapore, Singapore (e-mail: vtan@nus.edu.sg). The material in this paper was presented in part at IEEE ISIT 2016 [1]. The work of V. Y. F. Tan is supported in part by a Singapore Ministry of Education (MOE) Tier 2 grant (R-263-000-B61-113).

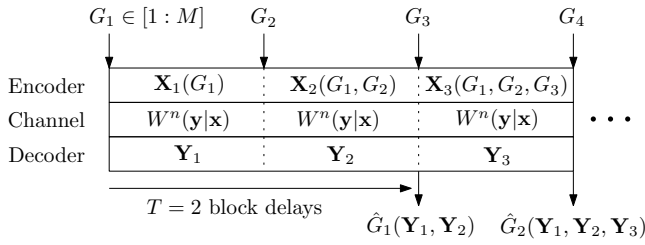


Figure 1. Our streaming setup is illustrated for the case with  $T = 2$ . In each block, a new message is given to the encoder in the beginning and the encoder generates a codeword as a function of all the past and current messages and transmits it over the channel. Since  $T = 2$ , the decoder decodes each message after two blocks, as a function of all the past received channel output sequences.

central limit regimes with the exception [19] where the focus is on source coding.

In this paper, we study streaming data transmission over a DMC in the moderate deviations and central limit regimes. Our streaming setup is illustrated in Fig. 1. In each block of length  $n$ , a new message is given to the encoder at the beginning, and the encoder generates a codeword as a function of all the past and current messages and transmits it over the channel. The decoder, given all the past received channel output sequences, decodes each message after a delay of  $T$  blocks. This streaming setup introduces a new dimension not present in the block coding problems studied previously. In the special case of  $T = 1$ , the setup reduces to the block channel coding problem. If  $T \geq 2$ , however, there exists an inherent tension in whether we utilize a block only for the fresh message or use it also for the previous messages with earlier deadlines. It is not difficult to see that due to the memoryless nature of the model, a time sharing scheme<sup>1</sup> will not provide any gain compared to the case of  $T = 1$ . A natural question is whether a joint encoding of fresh and previous messages would improve the performance when  $T \geq 2$ .

Our results indicate that the fundamental interplay between the rate and error probability can be greatly improved when delay is allowed in the streaming setup. In the moderate deviations regime, the moderate deviations constant is shown to improve over the block coding or non-streaming setup by a factor of  $T$ . In the central limit regime, the second-order coding rate is shown to improve by a factor of approximately  $\sqrt{T}$  for a wide range of channel parameters. For both asymptotic regimes, we propose coding techniques that incorporate a joint encoding of fresh and previous messages. For the moderate deviations regime, we propose a coding technique in which, for every block, the encoder jointly encodes all the previous and fresh messages and the decoder re-decodes all the previous messages in addition to the current target message. For the error analysis of this coding technique, we develop a refined and non-asymptotic version of the moderate deviations upper bound in [20, Theorem 3.7.1] that allows us to uniformly bound the error probabilities associated with the previous messages. On the other hand, for the central limit regime, we cannot apply such a coding technique whose

<sup>1</sup>In a time sharing scheme, some fraction of a block is used for a fresh message and some other fraction of the block is used for previous messages.

memory is linear in the block index. In the error analysis in the central limit regime, we encounter a summation of constants as a result of applications of the central limit theorem. If the memory is linear in the block index, this summation causes the upper bound on the error probability to diverge as the block index tends to infinity. Hence, for the central limit regime, we propose a coding technique with *truncated* memory where the memory at the encoder varies in a periodic fashion. Our proposed construction judiciously balances the rate penalty imposed due to the truncation and the growth in the error probability due to the contribution from previous messages. By analyzing the second-order coding rate of our proposed setup, we conclude that the channel dispersion parameter also decreases approximately by a factor of  $T$  for a wide range of channel parameters.

Furthermore, we explore interesting variants of the basic streaming setup in the moderate deviations regime. First, we consider a scenario where there is an erasure option at the decoder, i.e., the decoder can output an erasure symbol instead of a message estimate, and analyze the undetected error and the total error probabilities, extending a result by Hayashi and Tan [21]. Next, by utilizing the erasure option, we analyze the rate of decay of the error probability when a variable decoding delay is allowed. We show that such a flexibility in the decoding delay can dramatically improve the error probability in the streaming setup. This result is the analog of the classical results on variable-length decoding (see e.g., [22], [23]) to the streaming setup.

The rest of this paper is organized as follows. In Section II, we formally state our streaming setup. The main theorems are presented in Section III and proved in Section IV. In Section V, the moderate deviations result for the basic streaming setup is extended in various directions. We conclude this paper in Section VI.

#### A. Notation

The following notation is used throughout the paper. We reserve bold-font for vectors whose lengths are the same as blocklength  $n$ . For two integers  $i$  and  $j$ ,  $[i : j]$  denotes the set  $\{i, i+1, \dots, j\}$ . For constants  $x_1, \dots, x_k$  and  $S \subseteq [1 : k]$ ,  $x_S$  denotes the vector  $(x_j : j \in S)$  and  $x_i^j$  denotes  $x_{[i:j]}$  where the subscript is omitted when  $i = 1$ , i.e.,  $x^j = x_{[1:j]}$ . This notation is naturally extended for vectors  $\mathbf{x}_1, \dots, \mathbf{x}_k$ , random variables  $X_1, \dots, X_k$ , and random vectors  $\mathbf{X}_1, \dots, \mathbf{X}_k$ .  $\mathbb{1}\{\mathcal{E}\}$  for an event  $\mathcal{E}$  denotes the indicator function, i.e., it is 1 if  $\mathcal{E}$  is true and 0 otherwise.  $\lceil \cdot \rceil$  and  $\lfloor \cdot \rfloor$  denote the ceiling and floor functions, respectively.

For a DMC  $(\mathcal{X}, \mathcal{Y}, \{W(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\})$  and an input distribution  $P$ , we use the following standard notation and terminology in information theory:

- Information density:

$$i(x; y) := \log \frac{W(y|x)}{PW(y)}, \quad (1)$$

where  $PW(y) := \sum_{x \in \mathcal{X}} P(x)W(y|x)$  denotes the output distribution. We note that  $i(x; y)$  depends on  $P$  and  $W$  but this dependence is suppressed. The definition (1)

can be generalized for two vectors  $x^l$  and  $y^l$  of length  $l$  as follows:

$$i(x^l; y^l) := \sum_{j=1}^l i(x_j; y_j). \quad (2)$$

- Mutual information:

$$I(P, W) := \mathbb{E}[i(X; Y)] \quad (3)$$

$$= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x)W(y|x) \log \frac{W(y|x)}{PW(y)}. \quad (4)$$

- Unconditional information variance:

$$U(P, W) := \text{Var}[i(X; Y)]. \quad (5)$$

- Conditional information variance:

$$V(P, W) := \mathbb{E}[\text{Var}[i(X; Y)|X]]. \quad (6)$$

- Capacity:

$$C = C(W) := \max_{P \in \mathcal{P}} I(P, W), \quad (7)$$

where  $\mathcal{P}$  denotes the probability simplex on  $\mathbb{R}^{|\mathcal{X}|}$ .

- Set of capacity-achieving input distributions:

$$\Pi = \Pi(W) := \{P \in \mathcal{P} : I(P, W) = C(W)\}. \quad (8)$$

- Channel dispersion

$$V = V(W) := \min_{P \in \Pi} V(P, W) \quad (9)$$

$$\stackrel{(a)}{=} \min_{P \in \Pi} U(P, W), \quad (10)$$

where (a) is from [12, Lemma 62], where it is shown that  $V(P, W) = U(P, W)$  for all  $P \in \Pi$ .

## II. MODEL

Consider a DMC  $(\mathcal{X}, \mathcal{Y}, \{W(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\})$ . A streaming code is defined as follows:

**Definition 1** (Streaming code). An  $(n, M, \epsilon, T)$ -streaming code consists of

- a sequence of messages  $\{G_k\}_{k \geq 1}$  each distributed uniformly over  $\mathcal{G} := [1 : M]$ ,
- a sequence of encoding functions  $\phi_k : \mathcal{G}^k \rightarrow \mathcal{X}^n$  that maps the message sequence  $G^k \in \mathcal{G}^k$  to the channel input codeword  $\mathbf{X}_k \in \mathcal{X}^n$ , and
- a sequence of decoding functions  $\psi_k : \mathcal{Y}^{(k+T-1)n} \rightarrow \mathcal{G}$  that maps the channel output sequences  $\mathbf{Y}^{k+T-1} \in \mathcal{Y}^{(k+T-1)n}$  to a message estimate  $\hat{G}_k \in \mathcal{G}$ ,

that satisfies

$$\limsup_{N \rightarrow \infty} \sum_{k=1}^N \frac{\Pr(\hat{G}_k \neq G_k)}{N} \leq \epsilon, \quad (11)$$

i.e., the probability of error averaged over all block messages does not exceed  $\epsilon$ .

We note that a streaming code with a *fixed* blocklength  $n$  consists of a *sequence* of encoding and decoding functions since a stream of messages is sequentially encoded and decoded. Fig. 1 illustrates our streaming setup for the

case with  $T = 2$ . In the beginning of block  $k \in \mathbb{N}$ , new message  $G_k$  is given to the encoder. The encoder generates a codeword  $\mathbf{X}_k$  as a function of all the past and current messages  $G^k$  and transmits it over the channel in block  $k$ . Since  $T = 2$ , the decoder decodes message  $G_k$  at the end of block  $k+1$ , as a function of all the past received channel output sequences  $\mathbf{Y}^{k+1}$ . Note that the encoder and the decoder are implicitly assumed to have memories that possibly increase as the streaming communication proceeds.<sup>2</sup>

## III. MAIN RESULTS

In this section, we state our main results. The following two theorems present achievability bounds for the moderate deviations and the central limit regimes, respectively, which are proved in Section IV.

**Theorem 1** (Moderate deviations regime). Consider a DMC  $(\mathcal{X}, \mathcal{Y}, \{W(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\})$  with  $V > 0$  and any sequence of integers  $M_n$  such that  $\log M_n = nC - n\rho_n$ , where  $\rho_n > 0$ ,  $\rho_n \rightarrow 0$  and  $n\rho_n^2 \rightarrow \infty$ .<sup>3</sup> Then, there exists a sequence of  $(n, M_n, \epsilon_n, T)$ -streaming codes such that<sup>4</sup>

$$\limsup_{n \rightarrow \infty} \frac{1}{n\rho_n^2} \log \epsilon_n \leq -\frac{T}{2V}. \quad (12)$$

**Theorem 2** (Central limit regime). Consider a DMC  $(\mathcal{X}, \mathcal{Y}, \{W(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\})$  with  $V > 0$ . For any  $L > 0$  and  $0 < \delta < 1/2$ , there exists a sequence of  $(n, M_n, \epsilon_n, T)$ -streaming codes such that<sup>5</sup>

$$\log M_n = nC - L\sqrt{n} + O(n^\delta \log n) \quad (13)$$

and

$$\epsilon_n \leq \sum_{j=T}^{\infty} Q\left(\frac{\sqrt{j}}{\sqrt{V}}L\right) + O(n^{-\delta/2}). \quad (14)$$

The following corollary, whose proof is in Appendix A, elucidates a closed-form and interpretable expression for the upper bound on the error probability in (14).

**Corollary 3.** Consider a DMC  $(\mathcal{X}, \mathcal{Y}, \{W(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\})$  with  $V > 0$ . For any  $L > 0$ , there exists a sequence of  $(n, M_n, \epsilon_n, T)$ -streaming codes such that

$$\lim_{n \rightarrow \infty} \frac{nC - \log M_n}{\sqrt{n}} = L \quad (15)$$

and

$$\limsup_{n \rightarrow \infty} \epsilon_n < c_{L,V,T} Q\left(\sqrt{\frac{T}{V}}L\right), \quad (16)$$

<sup>2</sup>If the encoder has to encode only the fresh message in each block or the decoder is allowed to utilize the channel outputs in the recent block for decoding, the problem would reduce to the block coding or non-streaming setup.

<sup>3</sup>Throughput the paper, we ignore integer constraints on the number of codewords  $M_n$ .

<sup>4</sup>If  $\limsup_{n \rightarrow \infty} \frac{1}{n\rho_n^2} \log \epsilon_n \leq -\frac{1}{2\nu}$  for some  $\nu > 0$ ,  $\nu$  corresponds to an upper bound on the moderate deviations constant. In the special case of  $T = 1$ , the moderate deviations constant is shown to be the channel dispersion  $V$  in [16], [17].

<sup>5</sup> $L$  is termed second-order coding rate in this paper. This is slightly different from what is common in the literature where instead  $-L$  is known as the second-order coding rate [13].

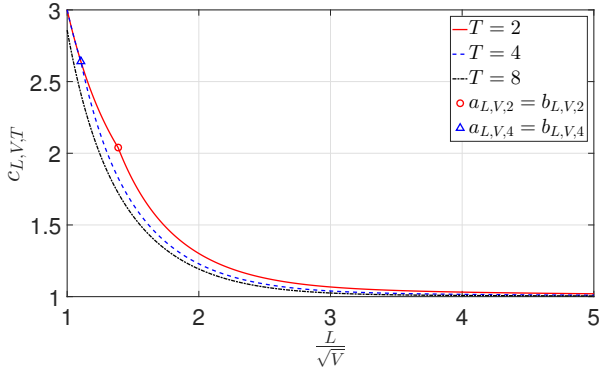


Figure 2. The constant  $c_{L,V,T}$  in Theorem 2 is illustrated in terms of  $\frac{L}{\sqrt{V}}$ . Note that  $c_{L,V,T}$  is the minimum of  $a_{L,V,T}$  and  $b_{L,V,T}$  where  $a_{L,V,T}$  dominates the minimum when  $\frac{L}{\sqrt{V}}$  is above some threshold value and otherwise  $b_{L,V,T}$  does. The threshold points where  $a_{L,V,2} = b_{L,V,2}$  and  $a_{L,V,4} = b_{L,V,4}$  are indicated by a circle and a triangle, respectively.

where  $c_{L,V,T}$  is defined in the following:

$$c_{L,V,T} := \min(a_{L,V,T}, b_{L,V,T}) \quad (17)$$

$$a_{L,V,T} := \frac{1 + (L/\sqrt{V})^2 T}{(L/\sqrt{V})^2 T} \cdot \frac{1}{1 - \exp\{-(L/\sqrt{V})^2/2\}} \quad (18)$$

$$b_{L,V,T} := 1 + \frac{2V}{L^2}. \quad (19)$$

**Remark 1.** Note that  $c_{L,V,T}$  defined in (17) has the property that for every  $T \in \mathbb{N}$ ,  $c_{L,V,T}$  tends to 1 as  $\frac{L}{\sqrt{V}}$  tends to infinity. The term  $\frac{L}{\sqrt{V}}$  becomes large when the second-order coding rate  $L$  is large and/or the channel dispersion  $V$  is small. The channel dispersion is small when, roughly speaking, there is less randomness in the channel, e.g., for binary erasure channel with erasure probability  $p$ , the channel dispersion  $V = p(1-p)$  is small when  $p$  is either close to 0 or 1. Note that the moderate deviations or large deviations regimes can roughly be interpreted as the limiting cases in which  $L$  tends to infinity.

Fig. 2 plots the constant  $c_{L,V,T}$  in Corollary 3 in terms of  $\frac{L}{\sqrt{V}}$ . Note that  $c_{L,V,T}$  is the minimum of  $a_{L,V,T}$  and  $b_{L,V,T}$  where  $a_{L,V,T}$  dominates the minimum when  $\frac{L}{\sqrt{V}}$  is above some threshold value and otherwise  $b_{L,V,T}$  does. In Fig. 2, we indicate by a circle and by a triangle the threshold points where  $a_{L,V,2} = b_{L,V,2}$  and  $a_{L,V,4} = b_{L,V,4}$ , respectively. We can see that the effect of  $c_{L,V,T}$  is not significant for a wide range of  $L, V$ , and  $T$ , e.g.,  $c_{L,V,2}$  is less than 1.1 when  $\frac{L}{\sqrt{V}} = 3$ .

Theorems 1 and 2 illustrate that the fundamental interplay between the rate and probability of error can be greatly improved when delay is allowed in the streaming setup. In the moderate deviations regime, the moderate deviations constant improves by a factor of  $T$ . Assuming that  $c_{L,V,T}$  can be approximated sufficiently well by 1, for the central limit regime, the second-order coding rate  $L$  is improved (reduced) by a factor of  $\sqrt{T}$ . Another way to view this via the lens of the channel dispersion  $V$ ; this parameter is approximately reduced by a factor of  $T$ .

## IV. PROOFS OF THE MAIN THEOREMS

### A. Proof of Theorem 1 for the moderate deviations regime

Consider a DMC  $(\mathcal{X}, \mathcal{Y}, \{W(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\})$  with  $V > 0$  and any sequence of integers  $M_n$  such that  $\log M_n = nC - n\rho_n$ , where  $\rho_n > 0, \rho_n \rightarrow 0$  and  $n\rho_n^2 \rightarrow \infty$ . We denote by  $P_X$  an input distribution that achieves the dispersion (9).

1) *Encoding:* For each  $k \in \mathbb{N}$  and  $g^k \in \mathcal{G}^k$ , generate  $\mathbf{x}_k(g^k)$  in an i.i.d. manner according to  $P_X$ . The generated codewords constitute the codebook  $\mathcal{C}_n$ . In block  $k$ , after observing the true message sequence  $G^k$ , the encoder sends  $\mathbf{x}_k(G^k)$ . This encoding procedure appeared in the literature in context of tree codes [2]–[4], [8], [10].

2) *Decoding:* Consider the decoding of  $G_k$  at the end of block  $T_k := k + T - 1$ . In our scheme, the decoder not only decodes  $G_k$ , but also re-decodes  $G_1, \dots, G_{k-1}$  at the end of block  $T_k$ .<sup>6</sup> Let  $\hat{G}_{T_k,j}$  denote the estimate of  $G_j$  at the end of block  $T_k$ . The decoder decodes  $G_j$  sequentially from  $j = 1$  to  $j = k$  as follows:

- Given  $\hat{G}_{T_k,[1:j-1]}$ , the decoder chooses  $\hat{G}_{T_k,j}$  according to the following rule.<sup>7</sup> If there is a unique index  $g_j \in \mathcal{G}$  that satisfies<sup>8</sup>

$$i(\mathbf{x}_{[j:T_k]}(\hat{G}_{T_k,[1:j-1]}, g_j, g_{[j+1:T_k]}), \mathbf{Y}_{[j:T_k]}) > (T_k - j + 1) \cdot \log M_n \quad (20)$$

for some  $g_{[j+1:T_k]}$ , let  $\hat{G}_{T_k,j} = g_j$ .<sup>9</sup> If there is none or more than one such  $g_j$ , let  $\hat{G}_{T_k,j} = 1$ .

- If  $j < k$ , repeat the above procedure by increasing  $j$  to  $j+1$ . If  $j = k$ , the decoding procedure terminates and the decoder declares that the  $k$ -th message is  $\hat{G}_k := \hat{G}_{T_k,k}$ .

3) *Error analysis:* We first consider the probability of error averaged over random codebook  $\mathcal{C}_n$ . The error event  $\{\hat{G}_k \neq G_k\}$  for  $k \in \mathbb{N}$  happens only if at least one of the following  $2k$  events occurs:

$$\mathcal{E}_{k,j} := \{i(\mathbf{X}_{[j:T_k]}(G^{T_k}), \mathbf{Y}_{[j:T_k]}) \leq (T_k - j + 1) \cdot \log M_n\}, \quad (21)$$

$$\tilde{\mathcal{E}}_{k,j} := \{i(\mathbf{X}_{[j:T_k]}(G^{j-1}, g_{[j:T_k]}), \mathbf{Y}_{[j:T_k]}) > (T_k - j + 1) \cdot \log M_n \text{ for some } g_{[j:T_k]} \text{ such that } g_j \neq G_j\} \quad (22)$$

for  $j \in [1 : k]$ .

Now, we have

$$\mathbb{E}_{\mathcal{C}_n}[\Pr(\hat{G}_k \neq G_k | \mathcal{C}_n)] \leq \sum_{j=1}^k (\Pr(\mathcal{E}_{k,j}) + \Pr(\tilde{\mathcal{E}}_{k,j})). \quad (23)$$

<sup>6</sup>We note that  $G_j$  for  $j \in [1 : k-1]$  has been already decoded at the end of block  $T_j$ . Nevertheless, the decoder re-decodes  $G^{k-1}$  at the end of  $T_k$ , because the decoder needs to decode  $G^{k-1}$  to decode  $G_k$  and the probability of error associated with  $G^{k-1}$  becomes lower (in general) by utilizing recent channel output sequences.

<sup>7</sup>When  $j = 1$ ,  $\hat{G}_{T_k}^{j-1}$  is null.

<sup>8</sup>We use the following notation for the set of codewords. Let  $\mathcal{K}_j$  for  $j \in \mathbb{N}$  denote the set of message indices mapped to the  $j$ -th codeword according to the encoding procedure. For  $\mathcal{J} \subseteq \mathbb{N}$  and  $\mathcal{K} \supseteq \bigcup_{j \in \mathcal{J}} \mathcal{K}_j$ , we denote by  $\mathbf{x}_{\mathcal{J}}(g_{\mathcal{K}})$  the set of codewords  $\{\mathbf{x}_j(g_{\mathcal{K}}) : j \in \mathcal{J}\}$ .

<sup>9</sup>We note that  $i(\cdot, \cdot)$  in (20) is defined in terms of  $P_X$  and  $W$ . This dependence is suppressed henceforth.

For each  $j \in [1 : k]$ , we have

$$\begin{aligned} & \Pr(\mathcal{E}_{k,j}) + \Pr(\tilde{\mathcal{E}}_{k,j}) \\ & \leq \Pr \left( \sum_{l=1}^{n(T_k-j+1)} i(X_l; Y_l) \leq (T_k - j + 1) \cdot \log M_n \right) \\ & \quad + M_n^{T_k-j+1} \Pr \left( \sum_{l=1}^{n(T_k-j+1)} i(X_l; \bar{Y}_l) \right. \\ & \quad \left. > (T_k - j + 1) \log M_n \right) \end{aligned} \quad (24)$$

$$\stackrel{(a)}{=} \mathbb{E} \left[ \exp \left\{ - \left[ \sum_{l=1}^{n(T_k-j+1)} i(X_l; Y_l) - (T_k - j + 1) \log M_n \right]^+ \right\} \right] \quad (25)$$

$$= \mathbb{E} \left[ \exp \left\{ - \left[ \sum_{l=1}^{n(T_k-j+1)} i(X_l; Y_l) - (T_k - j + 1)n(C - \rho_n) \right]^+ \right\} \right], \quad (26)$$

where  $(X_l, Y_l, \bar{Y}_l)$ 's are i.i.d. random variables each generated according to  $P_X(x_l)W(y_l|x_l)P_XW(\bar{y}_l)$  and (a) is from the identity [12, Eq. (69)] used to derive the DT bound.

Now, fix an arbitrary  $0 < \lambda < 1$ . By applying the chain of inequalities [17, Eq. (53)-(56)], we have

$$\begin{aligned} & \exp \left\{ - \left[ \sum_{l=1}^{n(T_k-j+1)} i(X_l; Y_l) - (T_k - j + 1)n(C - \rho_n) \right]^+ \right\} \\ & \leq \mathbb{1} \left\{ \sum_{l=1}^{n(T_k-j+1)} i(X_l; Y_l) \leq (T_k - j + 1)n(C - \lambda\rho_n) \right\} \\ & \quad + \exp \{ -(T_k - j + 1)n(1 - \lambda)\rho_n \}. \end{aligned} \quad (27)$$

Combining the bounds in (26) and (27), we obtain

$$\begin{aligned} & \Pr(\mathcal{E}_{k,j}) + \Pr(\tilde{\mathcal{E}}_{k,j}) \\ & \leq \Pr \left( \sum_{l=1}^{n(T_k-j+1)} i(X_l; Y_l) \leq (T_k - j + 1)n(C - \lambda\rho_n) \right) \\ & \quad + \exp \{ -(T_k - j + 1)n(1 - \lambda)\rho_n \} \end{aligned} \quad (28)$$

$$\stackrel{(a)}{\leq} \exp \left\{ -(T_k - j + 1)n \left( \frac{\lambda^2 \rho_n^2}{2V} - \lambda^3 \rho_n^3 \tau \right) \right\} + \exp \{ -(T_k - j + 1)n(1 - \lambda)\rho_n \} \quad (29)$$

for sufficiently large  $n$ , where  $\tau$  is some non-negative constant dependent only on the input distribution  $P_X(x)$  and channel statistics  $W(y|x)$  and (a) is from the moderate deviations upper bound in Lemma 4, which is relegated to the end of this subsection. Also see Remark 4.

Now, we have

$$\begin{aligned} & \mathbb{E}_{\mathcal{C}_n} [\Pr(\hat{G}_k \neq G_k | \mathcal{C}_n)] \\ & \leq \sum_{j=1}^k \left( \exp \left\{ -(T_k - j + 1)n\rho_n^2\lambda^2 \left( \frac{1}{2V} - \lambda\rho_n\tau \right) \right\} \right. \\ & \quad \left. + \exp \{ -(T_k - j + 1)n(1 - \lambda)\rho_n \} \right) \end{aligned} \quad (30)$$

$$\begin{aligned} & \leq \sum_{j=T}^{T_k} \left( \exp \left\{ -jn\rho_n^2\lambda^2 \left( \frac{1}{2V} - \lambda\rho_n\tau \right) \right\} \right. \\ & \quad \left. + \exp \{ -jn(1 - \lambda)\rho_n \} \right) \end{aligned} \quad (31)$$

$$\begin{aligned} & \leq \frac{\exp \{ -Tn\rho_n^2\lambda^2 (\frac{1}{2V} - \lambda\rho_n\tau) \}}{1 - \exp \{ -n\rho_n^2\lambda^2 (\frac{1}{2V} - \lambda\rho_n\tau) \}} \\ & \quad + \frac{\exp \{ -Tn(1 - \lambda)\rho_n \}}{1 - \exp \{ -n(1 - \lambda)\rho_n \}} \end{aligned} \quad (32)$$

for sufficiently large  $n$ , which leads to

$$\begin{aligned} & \limsup_{n \rightarrow \infty} \frac{1}{n\rho_n^2} \log \mathbb{E}_{\mathcal{C}_n} \left[ \limsup_{N \rightarrow \infty} \frac{\sum_{k=1}^N \Pr(\hat{G}_k \neq G_k | \mathcal{C}_n)}{N} \right] \\ & \leq -\frac{T\lambda^2}{2V}. \end{aligned} \quad (33)$$

Finally, by taking  $\lambda \rightarrow 1$ , we have

$$\begin{aligned} & \limsup_{n \rightarrow \infty} \frac{1}{n\rho_n^2} \log \mathbb{E}_{\mathcal{C}_n} \left[ \limsup_{N \rightarrow \infty} \frac{\sum_{k=1}^N \Pr(\hat{G}_k \neq G_k | \mathcal{C}_n)}{N} \right] \\ & \leq -\frac{T}{2V}. \end{aligned} \quad (34)$$

Hence, there must exist a sequence of codes  $\mathcal{C}_n$  that satisfies (12), which completes the proof.  $\blacksquare$

Our decoding procedure (20) is different from the best possible decoding technique, which is finding the maximum likelihood (ML) estimate of  $G_{[1:k]}$ , in two aspects; (i) evaluating and maximizing the likelihoods (over the messages) vs. threshold testing of the information density and (ii) joint decoding of  $G_{[1:k]}$  vs. sequential decoding from  $G_1$  to  $G_k$ . In the following, we make remarks on these two aspects.

**Remark 2.** We note that there are many possible decoding rules that attain some form of optimality. In the moderate deviations regime (also in the central limit regime), the threshold tests (of the information density) we propose are asymptotically optimal for the block coding or non-streaming setup [12], [17] and they are also easy to analyze. Hence, we choose to use these tests instead of ML which is optimal but harder to analyze for the two regimes of interest in this paper.

**Remark 3.** In our decoding procedure (20), we sequentially decode messages from  $G_1$  to  $G_k$  (at the end of block  $T_k$ ), instead of jointly decoding them. For both types of decoding, the dominant error is the error of the last message, i.e., message  $G_k$ , and hence the achievable moderate deviations constant would not be affected. We have employed the sequential decoding as we believe that the error events can be defined in a simpler manner and their probabilities can also be estimated easily.

The following lemma used in the proof of Theorem 1 corresponds to a non-asymptotic upper bound of the moderate deviations theorem [20, Theorem 3.7.1], whose proof is in Appendix B.

**Lemma 4.** Let  $\{Z_l\}_{l \geq 1}$  be a sequence of i.i.d. random variables such that  $\mathbb{E}[Z_1] = 0$ ,  $\text{Var}[Z_1] = \sigma^2 > 0$ , and its cumulant generating function  $h(s) := \log \mathbb{E}[\exp\{sZ_1\}]$  for  $s \geq 0$  is analytic around the origin and satisfies that  $K := \max_{s \in [0,1]} |h'''(s)|$  is finite. For a sequence  $\varepsilon_n > 0$  satisfying the moderate deviations constraints, i.e.,  $\varepsilon_n \rightarrow 0$  and  $n\varepsilon_n^2 \rightarrow \infty$ , the following bound holds:

$$\Pr\left(\frac{1}{n} \sum_{l=1}^n Z_l \geq \varepsilon_n\right) \leq \exp\left\{-n \left(\frac{\varepsilon_n^2}{2\sigma^2} - \frac{\varepsilon_n^3}{6\sigma^6} K\right)\right\} \quad (35)$$

for sufficiently large  $n$ .

**Remark 4.** Let us comment on the assumption in Lemma 4 that  $K$  is finite. In our application,

$$Z_l \equiv i(X_l; Y_l) - I(X_l; Y_l). \quad (36)$$

Then, we have

$$h(s) = \log \mathbb{E} \left[ \exp \left\{ s \left( \log \frac{W(Y_1|X_1)}{P_X W(Y_1)} - I(X_1; Y_1) \right) \right\} \right] \quad (37)$$

$$= -sI(X_1; Y_1) + \log \mathbb{E} \left[ \left( \frac{W(Y_1|X_1)}{P_X W(Y_1)} \right)^s \right]. \quad (38)$$

By differentiating thrice, we can show that  $h'''(s)$  is continuous in  $s$ .<sup>10</sup> Restricting  $s$  to  $[0, 1]$  means that  $h'''(s)$  is a continuous function over a compact set. Hence its maximum is attained and is necessarily finite.

### B. Proof of Theorem 2 for the central limit regime

Consider a DMC  $(\mathcal{X}, \mathcal{Y}, \{W(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\})$  with  $V > 0$ . We remark that in the moderate deviations regime, for every block, the encoder maps *all* the previous messages to a codeword. For the central limit regime, we propose a coding strategy where the encoder maps only *some* recent messages to the codeword in each block. Similar idea of incorporating truncated memory was used in [9] with the focus on reducing the complexity. Here, we use a different memory structure from [9]. Let  $A \in \mathbb{N}$  and  $B \in \mathbb{N}$  denote the maximum and the minimum numbers of messages that can possibly be mapped to a codeword in each block, respectively. We choose the size  $M_n$  of message alphabet as follows:

$$\log M_n = \frac{A - 2B + T + 2}{A} (nC - L\sqrt{n}) \quad (39)$$

for some  $L > 0$ . To make the above choice of  $M_n$  valid, we assume  $A \geq 2B - T - 2 \geq 0$ . Furthermore, we assume that the minimum encoding memory is at least  $T$ , i.e.,  $B \geq T$ . We denote by  $P_X$  an input distribution that achieves the dispersion (9).

<sup>10</sup>A detailed calculation follows similarly as in the proof of [16, Lemma 1].

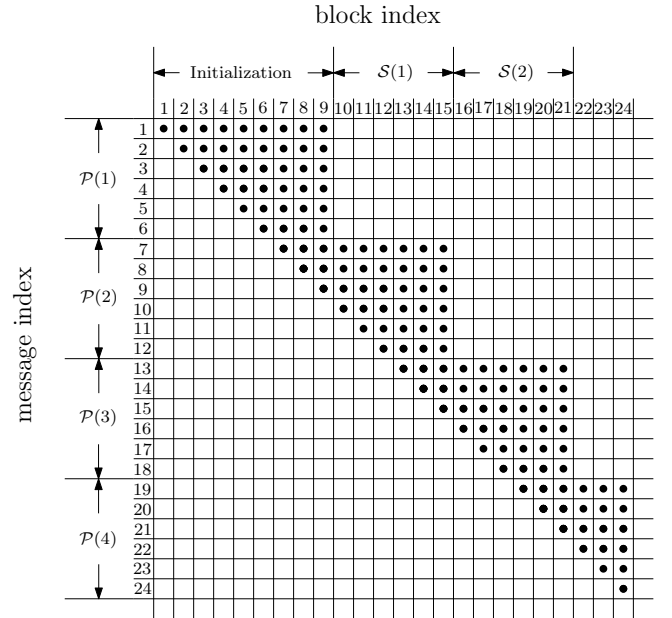


Figure 3. The proposed message-codeword mapping rule for the central limit regime is illustrated for the case of  $A = 9$  (maximum encoding memory) and  $B = 4$  (minimum encoding memory). For each block, the messages indicated by black dots are jointly encoded. After an initialization step of the first nine blocks, in which all the previous messages are mapped to a codeword, our encoder has a periodically time-varying memory from four to nine with a period of six blocks.

1) *Encoding*: Our encoder has a periodically time-varying memory  $m \in [B : A]$  with a period of  $A - B + 1$  blocks, after an initialization step of the first  $A$  blocks. For each  $k \in [1 : A]$  and  $g^k \in \mathcal{G}^k$ , generate  $\mathbf{x}_k(g^k)$  in an i.i.d. manner according to  $P_X$ . In block  $k \in [1 : A]$ , the encoder sends  $\mathbf{x}_k(G^k)$ . Since the maximum encoding memory is  $A$ , we *truncate* the messages that are mapped to a codeword on and after the  $A$ -th block, so that the encoding memory is periodically time-varying from  $B$  to  $A$  with a period of  $A - B + 1$  blocks. Let  $\mathcal{S}(q)$  for  $q \geq 1$  denote the set of  $(A - B + 1)$  block indices in the  $q$ -th period on and after the  $(A + 1)$ -st block, i.e.,  $\mathcal{S}(q) = \{(A - B + 1)q + B, \dots, (A - B + 1)(q + 1) + B - 1\}$ . For each  $k \in \mathcal{S}(q)$  and  $g^{k - q(A - B + 1)} \in \mathcal{G}^{k - q(A - B + 1)}$ ,<sup>11</sup> generate  $\mathbf{x}_k(g^{k - q(A - B + 1)})$  in an i.i.d. manner according to  $P_X$ . In block  $k \in \mathcal{S}(q)$ , the encoder sends  $\mathbf{x}_k(G_{[q(A - B + 1) + 1 : k]})$ .

On the other hand, we can group the messages according to the maximum block index to which a message is mapped. Let  $\mathcal{P}(q)$  for  $q \in \mathbb{N}$  denote the  $q$ -th group  $\{G_{(A - B + 1)(q - 1) + 1}, \dots, G_{(A - B + 1)q}\}$  of messages that are mapped to a codeword up to block  $(A - B + 1)q + B - 1$ . This grouping rule is useful for describing the decoding rule. Fig. 3 illustrates our message-codeword mapping rule for the case of  $A = 9$  and  $B = 4$ .

2) *Decoding*: The decoding rule of  $G_k \in \mathcal{P}(1)$  at the end of block  $T_k$  is exactly the same as that for the moderate deviations regime. Hence, from now on, let us focus on the decoding of  $G_k \in \mathcal{P}(q)$  for  $q \geq 2$  at the end of block  $T_k$ . At the end of block  $T_k$ , the decoder decodes not only  $G_k$ , but also all the

<sup>11</sup>In block  $k \in \mathcal{S}(q)$ , a total of  $k - q(A - B + 1)$  messages, i.e.,  $G_{q(A - B + 1) + 1}, \dots, G_k$ , are mapped to a codeword.

messages in the previous group and the previous messages in the current group,<sup>12</sup> i.e.,  $G_{(A-B+1)(q-2)+1}, \dots, G_{k-1}$ , according to the following three steps: (i) simultaneous non-unique decoding of the first  $B$  messages in the previous group, (ii) sequential decoding of the remaining  $A-2B+1$  messages in the previous group, and (iii) sequential decoding of the messages in the current group up to the current block. Let  $\hat{G}_{T_k, j}$  denote the estimate of  $G_j$  at the end of block  $T_k$ .

Let us describe the decoding rule when  $q = 2$  in the following:

- (i) If there is a unique index vector  $g^B$  that satisfies

$$i(\mathbf{x}_{[B:\min(A, T_k)]}(g^B, g_{[B+1:\min(A, T_k)]}), \mathbf{Y}_{[B:\min(A, T_k)]}) > \min(A, T_k) \cdot \log M_n \quad (40)$$

for some  $g_{[B+1:\min(A, T_k)]}$ , let  $\hat{G}_{T_k, [1:B]} = g^B$ . If there is none or more than one such  $g^B$ , let  $\hat{G}_{T_k, [1:B]} = (1, \dots, 1)$ .

- (ii) The decoder sequentially decodes  $g_j$  from  $j = B+1$  to  $j = A-B+1$  as follows:

- Given  $\hat{G}_{T_k, [1:j-1]}$ , the decoder chooses  $\hat{G}_{T_k, j}$  according to the following rule. If there is a unique index  $g_j \in \mathcal{G}$  that satisfies

$$i(\mathbf{x}_{[j:\min(A, T_k)]}(\hat{G}_{T_k, [1:j-1]}, g_j, g_{[j+1:\min(A, T_k)]}), \mathbf{Y}_{[j:\min(A, T_k)]}) > (\min(A, T_k) - j + 1) \cdot \log M_n \quad (41)$$

for some  $g_{[j+1:\min(A, T_k)]}$ , let  $\hat{G}_{T_k, j} = g_j$ . If there is none or more than one such  $g_j$ , let  $\hat{G}_{T_k, j} = 1$ .

- If  $j < A-B+1$ , repeat the above procedure by increasing  $j$  to  $j+1$ . If  $j = A-B+1$ , proceed to the next decoding procedure.

- (iii) The decoder sequentially decodes  $g_j$  from  $j = A-B+2$  to  $j = k$  as follows:

- Given  $\hat{G}_{T_k, [1:j-1]}$ , the decoder chooses  $\hat{G}_{T_k, j}$  according to the following rule. If there is a unique index  $g_j \in \mathcal{G}$  that satisfies

$$i(\mathbf{x}_{[j:T_k]}(\hat{G}_{T_k, [1:j-1]}, g_j, g_{[j+1:T_k]}), \mathbf{Y}_{[j:T_k]}) > (T_k - j + 1) \cdot \log M_n \quad (42)$$

for some  $g_{[j+1:T_k]}$ , let  $\hat{G}_{T_k, j} = g_j$ . If there is none or more than one such  $g_j$ , let  $\hat{G}_{T_k, j} = 1$ .

- If  $j < k$ , repeat the above procedure by increasing  $j$  to  $j+1$ . If  $j = k$ , the whole decoding procedure terminates and the decoder declares that the  $k$ -th message is  $\hat{G}_k := \hat{G}_{T_k, k}$ .

Note that the decoder does not utilize the channel output sequences in blocks  $1, \dots, B-1$  for decoding of messages  $G_1, \dots, G_B$ . To understand this, let us introduce a group  $\mathcal{P}(0)$  of (virtual) messages for symmetry, e.g., in Fig. 3, expand the pattern of black dots in the upper left direction. Then, the messages in group  $\mathcal{P}(0)$ , which we do not want to decode at this point, are also mapped to a codeword in

<sup>12</sup>Similarly as in the moderate deviations regime,  $G_j$  for  $j \in [1:k-1]$  has been already decoded at the end of block  $T_j$ . Nevertheless, the decoder re-decodes some of the previous messages at the end of  $T_k$ .

blocks  $1, \dots, B-1$ . Hence, those blocks are not considered for the decoding of messages  $G_1, \dots, G_B$ . We also note that for decoding of messages  $G^B$  and  $G_{B+1}$  to  $G_{A-B+1}$ , the decoder considers the channel output sequences up to block  $\min(A, T_k)$ , because it is the last available block to which those messages are mapped.

By exploiting the symmetry of the message-codeword mapping rule, the decoding rule for  $q \geq 3$  proceeds similarly.

3) *Error analysis:* We first consider the probability of error averaged over random codebook  $\mathcal{C}_n$ . Let us consider the decoding of  $G_k \in \mathcal{P}(2)$ . Let  $\alpha := \min(A, T_k)$ . The error event  $\{\hat{G}_k \neq G_k\}$  occurs only if at least one of the following  $2(k-B+1)$  events occurs:

$$\mathcal{E}_k^{(i)} := \{i(\mathbf{X}_{[B:\alpha]}(G^\alpha), \mathbf{Y}_{[B:\alpha]}) \leq \alpha \cdot \log M_n\} \quad (43)$$

$$\tilde{\mathcal{E}}_k^{(i)} := \{i(\mathbf{X}_{[B:\alpha]}(g^\alpha), \mathbf{Y}_{[B:\alpha]}) > \alpha \cdot \log M_n \text{ for some } g^\alpha \text{ such that } g^B \neq G^B\} \quad (44)$$

$$\mathcal{E}_{k,j}^{(ii)} := \{i(\mathbf{X}_{[j:\alpha]}(G^\alpha), \mathbf{Y}_{[j:\alpha]}) \leq (\alpha - j + 1) \cdot \log M_n \text{ for } j \in [B+1:A-B+1]\} \quad (45)$$

$$\tilde{\mathcal{E}}_{k,j}^{(ii)} := \{i(\mathbf{X}_{[j:\alpha]}(G^{j-1}, g_{[j:\alpha]}), \mathbf{Y}_{[j:\alpha]}) > (\alpha - j + 1) \cdot \log M_n \text{ for some } g_{[j:\alpha]} \text{ such that } g_j \neq G_j\} \text{ for } j \in [B+1:A-B+1] \quad (46)$$

$$\mathcal{E}_{k,j}^{(iii)} := \{i(\mathbf{X}_{[j:T_k]}(G^{T_k}), \mathbf{Y}_{[j:T_k]}) \leq (T_k - j + 1) \cdot \log M_n \text{ for } j \in [A-B+2:k]\} \quad (47)$$

$$\tilde{\mathcal{E}}_{k,j}^{(iii)} := \{i(\mathbf{X}_{[j:T_k]}(G^{j-1}, g_{[j:T_k]}), \mathbf{Y}_{[j:T_k]}) > (T_k - j + 1) \cdot \log M_n \text{ for some } g_{[j:T_k]} \text{ such that } g_j \neq G_j\} \text{ for } j \in [A-B+2:k]. \quad (48)$$

We note that the superscript in each error event represents the decoding step in which the error event is involved. Now, we have

$$\begin{aligned} & \mathbb{E}_{\mathcal{C}_n}[\Pr(\hat{G}_k \neq G_k | \mathcal{C}_n)] \\ & \leq \Pr(\mathcal{E}_k^{(i)}) + \Pr(\tilde{\mathcal{E}}_k^{(i)}) \\ & \quad + \sum_{j=B+1}^{A-B+1} \Pr(\mathcal{E}_{k,j}^{(ii)}) + \sum_{j=B+1}^{A-B+1} \Pr(\tilde{\mathcal{E}}_{k,j}^{(ii)}) \\ & \quad + \sum_{j=A-B+2}^k \Pr(\mathcal{E}_{k,j}^{(iii)}) + \sum_{j=A-B+2}^k \Pr(\tilde{\mathcal{E}}_{k,j}^{(iii)}). \end{aligned} \quad (49)$$

Let us bound each term in the RHS of (49). First,  $\Pr(\mathcal{E}_k^{(i)})$  is upper-bounded as follows:

$$\Pr(\mathcal{E}_k^{(i)}) = \Pr(i(\mathbf{X}_{[B:\alpha]}(G^\alpha), \mathbf{Y}_{[B:\alpha]}) \leq \alpha \cdot \log M_n) \quad (50)$$

$$\leq \Pr\left(\sum_{l=1}^{n(\alpha-B+1)} i(X_l; Y_l) \leq \alpha \cdot \log M_n\right) \quad (51)$$

$$\stackrel{(a)}{\leq} \Pr\left(\sum_{l=1}^{n(\alpha-B+1)} i(X_l; Y_l) \leq (\alpha - B + 1)(nC - L\sqrt{n})\right) \quad (52)$$

$$\stackrel{(b)}{\leq} Q\left(\frac{\sqrt{\alpha - B + 1}}{\sqrt{V}}L\right) + \frac{\tau_1}{\sqrt{(\alpha - B + 1)n}} \quad (53)$$

for some non-negative constant  $\tau_1$  that is dependent only on the input distribution  $P_X$  and the channel statistics  $W(y|x)$ , where  $(X_l, Y_l)$ 's are i.i.d. random variables each generated according to  $P_X(x_l)W(y_l|x_l)$ , (a) is from the choice of  $M_n$  in (39), and (b) is from the Berry-Esseen Theorem (e.g., [24]). Similarly, we can show

$$\begin{aligned} & \sum_{j=B+1}^{A-B+1} \Pr(\mathcal{E}_{k,j}^{(ii)}) \\ & \leq \sum_{j=B+1}^{A-B+1} Q\left(\frac{\sqrt{\alpha - j + 1}}{\sqrt{V}}L\right) + \frac{\tau_1}{\sqrt{(\alpha - j + 1)n}} \end{aligned} \quad (54)$$

and

$$\begin{aligned} & \sum_{j=A-B+2}^k \Pr(\mathcal{E}_{k,j}^{(iii)}) \\ & \leq \sum_{j=A-B+2}^k Q\left(\frac{\sqrt{T_k - j + 1}}{\sqrt{V}}L\right) + \frac{\tau_1}{\sqrt{(T_k - j + 1)n}}. \end{aligned} \quad (55)$$

Next,  $\Pr(\tilde{\mathcal{E}}_k^{(i)})$  is upper-bounded as follows:

$$\begin{aligned} & \Pr(\tilde{\mathcal{E}}_k^{(i)}) \\ & = \Pr(i(\mathbf{X}_{[B:\alpha]}(g^\alpha), \mathbf{Y}_{[B:\alpha]}) > \alpha \cdot \log M_n \\ & \quad \text{for some } g^\alpha \text{ such that } g^B \neq G^B) \end{aligned} \quad (56)$$

$$\leq M_n^\alpha \cdot \Pr\left(\sum_{l=1}^{n(\alpha-B+1)} i(X_l; \bar{Y}_l) > \alpha \cdot \log M_n\right) \quad (57)$$

$$\stackrel{(a)}{=} M_n^\alpha \cdot \mathbb{E}\left[\exp\left\{-\sum_{l=1}^{n(\alpha-B+1)} i(X_l; Y_l)\right\} \cdot \mathbb{1}\left\{\sum_{l=1}^{n(\alpha-B+1)} i(X_l; Y_l) > \alpha \log M_n\right\}\right] \quad (58)$$

$$\stackrel{(b)}{\leq} \frac{\tau_2}{\sqrt{(\alpha - B + 1)n}} \quad (59)$$

for some non-negative constant  $\tau_2$  that is dependent only on the input distribution  $P_X$  and channel statistics  $W(y|x)$ , where  $(X_l, Y_l, \bar{Y}_l)$ 's are i.i.d. random variables each generated according to  $P_X(x_l)W(y_l|x_l)P_XW(\bar{y}_l)$ , (a) is due to an elementary chain of equalities given in Appendix C, and (b) is from [12, Lemma 47].

Similarly, we can show

$$\sum_{j=B+1}^{A-B+1} \Pr(\tilde{\mathcal{E}}_{k,j}^{(ii)}) \leq \sum_{j=B+1}^{A-B+1} \frac{\tau_2}{\sqrt{(\alpha - j + 1)n}} \quad (60)$$

and

$$\sum_{j=A-B+2}^k \Pr(\tilde{\mathcal{E}}_{k,j}^{(iii)}) \leq \sum_{j=A-B+2}^k \frac{\tau_2}{\sqrt{(T_k - j + 1)n}}. \quad (61)$$

By substituting the above bounds into the RHS of (49), we obtain

$$\begin{aligned} & \mathbb{E}_{\mathcal{C}_n}[\Pr(\hat{G}_k \neq G_k | \mathcal{C}_n)] \\ & \leq \sum_{j=B}^{A-B+1} \left(Q\left(\frac{\sqrt{\alpha - j + 1}}{\sqrt{V}}L\right) + \frac{\tau_1 + \tau_2}{\sqrt{(\alpha - j + 1)n}}\right) \\ & \quad + \sum_{j=A-B+2}^k \left(Q\left(\frac{\sqrt{T_k - j + 1}}{\sqrt{V}}L\right) + \frac{\tau_1 + \tau_2}{\sqrt{(T_k - j + 1)n}}\right) \end{aligned} \quad (62)$$

$$\begin{aligned} & \leq \sum_{j=\alpha-A+B}^{\alpha-B+1} \left(Q\left(\frac{\sqrt{j}}{\sqrt{V}}L\right) + \frac{\tau_1 + \tau_2}{\sqrt{jn}}\right) \\ & \quad + \sum_{j=T}^{T_k-A+B-1} \left(Q\left(\frac{\sqrt{j}}{\sqrt{V}}L\right) + \frac{\tau_1 + \tau_2}{\sqrt{jn}}\right) \end{aligned} \quad (63)$$

$$\stackrel{(a)}{\leq} \sum_{j=B}^{A-B+1} \left(Q\left(\frac{\sqrt{j}}{\sqrt{V}}L\right) + \frac{\tau_1 + \tau_2}{\sqrt{jn}}\right) + \sum_{j=T}^{A-B+T} \left(Q\left(\frac{\sqrt{j}}{\sqrt{V}}L\right) + \frac{\tau_1 + \tau_2}{\sqrt{jn}}\right), \quad (64)$$

where (a) is because if  $\alpha = T_k$ , which implies  $T_k \leq A$ , the RHS of (63) is upper-bounded as follows:

$$\text{RHS of (63)} = \sum_{j=T}^{T_k-B+1} \left(Q\left(\frac{\sqrt{j}}{\sqrt{V}}L\right) + \frac{\tau_1 + \tau_2}{\sqrt{jn}}\right) \quad (65)$$

$$\leq \sum_{j=T}^{A-B+1} \left(Q\left(\frac{\sqrt{j}}{\sqrt{V}}L\right) + \frac{\tau_1 + \tau_2}{\sqrt{jn}}\right), \quad (66)$$

and if  $\alpha = A$ , which implies  $A \leq T_k$ , the RHS of (63) is upper-bounded as follows:

$$\begin{aligned} \text{RHS of (63)} & = \sum_{j=B}^{A-B+1} \left(Q\left(\frac{\sqrt{j}}{\sqrt{V}}L\right) + \frac{\tau_1 + \tau_2}{\sqrt{jn}}\right) \\ & \quad + \sum_{j=T}^{T_k-A+B-1} \left(Q\left(\frac{\sqrt{j}}{\sqrt{V}}L\right) + \frac{\tau_1 + \tau_2}{\sqrt{jn}}\right) \end{aligned} \quad (67)$$

$$\begin{aligned} & \leq \sum_{j=B}^{A-B+1} \left(Q\left(\frac{\sqrt{j}}{\sqrt{V}}L\right) + \frac{\tau_1 + \tau_2}{\sqrt{jn}}\right) \\ & \quad + \sum_{j=T}^{A-B+T} \left(Q\left(\frac{\sqrt{j}}{\sqrt{V}}L\right) + \frac{\tau_1 + \tau_2}{\sqrt{jn}}\right). \end{aligned} \quad (68)$$

Now, the RHS of (64) is bounded as follows:

$$\begin{aligned} & \text{RHS of (64)} \\ & = \sum_{j=B}^{A-B+1} \left(Q\left(\frac{\sqrt{j}}{\sqrt{V}}L\right) + \frac{\tau_1 + \tau_2}{\sqrt{jn}}\right) \\ & \quad + \sum_{j=T}^{A-B+T} \left(Q\left(\frac{\sqrt{j}}{\sqrt{V}}L\right) + \frac{\tau_1 + \tau_2}{\sqrt{jn}}\right) \end{aligned} \quad (69)$$

$$\begin{aligned} & \stackrel{(a)}{\leq} \sum_{j=B}^{A-B+1} Q\left(\frac{\sqrt{j}}{\sqrt{V}}L\right) + \sum_{j=T}^{A-B+T} Q\left(\frac{\sqrt{j}}{\sqrt{V}}L\right) \\ & \quad + 4(\tau_1 + \tau_2)\sqrt{\frac{A-B+T}{n}} \end{aligned} \quad (70)$$



$$\stackrel{(b)}{\leq} \frac{\sqrt{V}}{\sqrt{2\pi BL}} \cdot \frac{\exp\left\{-\frac{L^2 B}{2V}\right\}}{1 - \exp\left\{-\frac{L^2}{2V}\right\}} + \sum_{j=T}^{A-B+T} Q\left(\frac{\sqrt{j}}{\sqrt{V}}L\right) + 4(\tau_1 + \tau_2)\sqrt{\frac{A-B+T}{n}} \quad (71)$$

where (a) is from Lemma 5 (with the identification of  $f(j) \equiv 1/\sqrt{j}$ ), which is relegated to the end of this subsection, and (b) is obtained by applying similar steps as in the proof of Corollary 3.<sup>13</sup>

Now we can see a tension in choosing the maximum memory  $A$  and minimum memory  $B$ . First, the rate penalty due to truncation is proportional to  $\frac{B}{A}$  as we can see from (39) and it has to be small enough not to affect the second-order coding rate, i.e.,  $\frac{B}{A} = o(n^{-1/2})$ . On the other hand, in the bound on the error probability in the RHS of (71), the first term results from the decoding of the previous group, the second term comes from the decoding of the current group, and the third term corresponds to the summation of remainders as a result of applications of Berry-Esseen theorem. Thus, we want to make the first and the third terms negligible compared to the second term. To do so, it is required that  $B = \omega(1)$  and  $A = o(n)$ . In summary, the scalings of  $A$  and  $B$  in  $n$  have to be chosen appropriately to satisfy  $\frac{B}{A} = o(n^{-1/2})$ ,  $B = \omega(1)$ , and  $A = o(n)$  to obtain the desired result. Let us choose  $A = n^{1-\delta}$  and  $B = \frac{V}{L^2}\delta \log n$  for  $0 < \delta < \frac{1}{2}$ . By substituting this choice of  $A$  and  $B$  into the RHS of (39) and the RHS of (71), we obtain

$$\log M_n = nC - L\sqrt{n} + O(n^\delta \log n) \quad (72)$$

and

$$\begin{aligned} & \mathbb{E}_{\mathcal{C}_n}[\Pr(\hat{G}_k \neq G_k | \mathcal{C}_n)] \\ & \leq \sum_{j=T}^{\infty} Q\left(\frac{\sqrt{j}}{\sqrt{V}}L\right) + O(n^{-\delta/2}), \end{aligned} \quad (73)$$

respectively. Due to the symmetry of the decoding procedure, the bound (73) holds for  $G_k \in \mathcal{P}(q)$  for  $q \geq 3$ . For  $G_k \in \mathcal{P}(1)$ , by defining the error events in the same way as for the moderate deviations regime and then applying similar bounding techniques used in the above, it can be verified that

$$\begin{aligned} & \mathbb{E}_{\mathcal{C}_n}[\Pr(\hat{G}_k \neq G_k | \mathcal{C}_n)] \\ & \leq \sum_{j=T}^{T_k} Q\left(\frac{\sqrt{j}}{\sqrt{V}}L\right) + \frac{\tau_1 + \tau_2}{\sqrt{jn}} \end{aligned} \quad (74)$$

$$\leq \sum_{j=T}^{A-B+T} Q\left(\frac{\sqrt{j}}{\sqrt{V}}L\right) + \frac{\tau_1 + \tau_2}{\sqrt{jn}} \quad (75)$$

$$\leq \sum_{j=T}^{\infty} Q\left(\frac{\sqrt{j}}{\sqrt{V}}L\right) + O(n^{-\delta/2}). \quad (76)$$

Hence, there must exist a sequence of codes  $\mathcal{C}_n$  that satisfies (13) and (14), which completes the proof. ■

The following basic lemma, whose proof is omitted, is used in the proof of Theorem 2.

**Lemma 5.** Assume two integers  $a$  and  $b$  such that  $a \leq b$ . If  $f(x)$  is monotonically decreasing and integrable on  $[a, b]$ , we have

$$\sum_{j=a}^b f(j) \leq \int_a^{b+1} f(x-1)dx \quad (77)$$

$$= F(b) - F(a-1), \quad (78)$$

where  $F(x)$  denotes the antiderivative of  $f(x)$ .

## V. EXTENSIONS IN THE MODERATE DEVIATIONS REGIME

In this section, we explore interesting variations of the basic streaming setup in Section II. For the brevity of the results, we focus on the moderate deviations regime.

### A. Decoding with an erasure option

Consider the scenario where there is an erasure option at the decoder, i.e., the decoder can output an erasure symbol instead of a message estimate. In the presence of an erasure option, there are two types of error events: (i) the decoder declares an erasure and (ii) the decoder outputs an incorrect message, not an erasure. In many applications, the undetected error (the latter event) is more undesirable than an erasure (the former event). In the following, we define a streaming code with an erasure option by taking into account the undetected error and the total error probabilities separately.

**Definition 2** (Streaming code with an erasure option). An  $(n, M, \epsilon, \epsilon', T)$ -streaming code with an erasure option consists of

- a sequence of messages  $\{G_k\}_{k \geq 1}$  each distributed uniformly over  $\mathcal{G} := [1 : M]$ ,
- a sequence of encoding functions  $\phi_k : \mathcal{G}^k \rightarrow \mathcal{X}^n$  that maps the message sequence  $G^k \in \mathcal{G}^k$  to the channel input codeword  $\mathbf{X}_k \in \mathcal{X}^n$ , and
- a sequence of decoding functions  $\psi_k : \mathcal{Y}^{(k+T-1)n} \rightarrow \mathcal{G} \cup \{0\}$  that maps the channel output sequences  $\mathbf{Y}^{k+T-1} \in \mathcal{Y}^{(k+T-1)n}$  to a message estimate  $\hat{G}_k \in \mathcal{G}$  or an erasure symbol  $\hat{G}_k = 0$ ,

that satisfies

$$\limsup_{N \rightarrow \infty} \sum_{k=1}^N \frac{\Pr(\hat{G}_k \neq G_k)}{N} \leq \epsilon, \quad (79)$$

i.e., the total error probability does not exceed  $\epsilon$ , and

$$\limsup_{N \rightarrow \infty} \sum_{k=1}^N \frac{\Pr(\hat{G}_k \neq G_k, \hat{G}_k \neq 0)}{N} \leq \epsilon', \quad (80)$$

i.e., the undetected error probability does not exceed  $\epsilon'$ .

The following theorem presents upper bounds on the undetected error and the total error probabilities. The proof of this theorem is provided in Appendix D.

**Theorem 6.** Consider a DMC  $(\mathcal{X}, \mathcal{Y}, \{W(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\})$  with  $V > 0$  and any sequence of integers  $M_n$  such that  $\log M_n = nC - n\rho_n$ , where  $\rho_n > 0, \rho_n \rightarrow 0$  and  $n\rho_n^2 \rightarrow \infty$ . For any  $0 < \gamma < 1$ , there exists a sequence

<sup>13</sup>Step (b) can be obtained by replacing  $T$  by  $B$  in the RHS of (92).

of  $(n, M_n, \epsilon_n, \epsilon'_n, T)$ -streaming codes with an erasure option such that

$$\limsup_{n \rightarrow \infty} \frac{1}{n\rho_n^2} \log \epsilon_n \leq -\frac{T(1-\gamma)^2}{2V}. \quad (81)$$

$$\limsup_{n \rightarrow \infty} \frac{1}{n\rho_n} \log \epsilon'_n \leq -T\gamma. \quad (82)$$

Theorem 6 indicates that for our proposed scheme, the undetected error probability decays much faster than the total error probability, i.e., the exponent of the undetected error probability is the order of  $n\rho_n$ , whereas that of the total error probability is the order of  $n\rho_n^2$ . We note that when  $T = 1$  and  $\rho_n = an^{-t}$  for  $a > 0$  and  $0 < t < 1/2$ , Theorem 6 reduces to [21, Theorem 1]. In the streaming setup, both the exponents of the total error and the undetected error probabilities improve over the block coding or non-streaming setup in [21, Theorem 1] by factors of  $T$ .

### B. Decoding with average delay constraint

We note that the decoding delay is assumed to be fixed to  $T$  up to this point. In this subsection, we relax this constraint by requiring the *average* decoding delay not to exceed  $T$ . A streaming code with average delay constraint is defined as follows:

**Definition 3** (Streaming code with average delay constraint). An  $(n, M, \epsilon, T)$ -streaming code with average delay constraint consists of

- a sequence of messages  $\{G_k\}_{k \geq 1}$  each distributed uniformly over  $\mathcal{G} := [1 : M]$ ,
- a sequence of encoding functions  $\phi_k : \mathcal{G}^k \rightarrow \mathcal{X}^n$  that maps the message sequence  $G^k \in \mathcal{G}^k$  to the channel input codeword  $\mathbf{X}_k \in \mathcal{X}^n$ , and
- a sequence of decoding functions  $\psi_k : \mathcal{Y}^{kn} \rightarrow (\mathcal{G} \cup \{0\})^k$  that maps the channel output sequences  $\mathbf{Y}^k \in \mathcal{Y}^{kn}$  to a message estimate  $\hat{G}_{k,j} \in \mathcal{G}$  or an erasure symbol  $\hat{G}_{k,j} = 0$  for every  $j \in [1 : k]$

that satisfies

$$\limsup_{N \rightarrow \infty} \sum_{k=1}^N \frac{\Pr(\hat{G}_{k+D_k-1,k} \neq G_k)}{N} \leq \epsilon \quad (83)$$

and

$$\limsup_{N \rightarrow \infty} \sum_{k=1}^N \frac{\mathbb{E}[D_k]}{N} \leq T, \quad (84)$$

where  $D_k := \min\{d : \hat{G}_{k+d-1,k} \neq 0\}$  for  $k \in \mathbb{N}$  denotes the (random) decoding delay of the  $k$ -th message.<sup>14</sup>

For block channel coding with feedback, it is known that the error exponent can be significantly improved by allowing variable decoding delay, e.g., [22], [23]. For streaming setup, the following theorem, which is proved in Appendix E, shows

<sup>14</sup>Note that message  $G_k$  is required to be decoded at the end of every block on and after the  $k$ -th block in this definition. One may wonder why the decoder does not stop decoding  $G_k$  after it outputs an estimate of  $G_k$ , not an erasure. We note that our definition includes such a operation as a special case by letting the decoder simply fix the estimate of  $G_k$  once it outputs a message estimate.

that such an improvement can be obtained in the absence of feedback.

**Theorem 7.** Consider a DMC  $(\mathcal{X}, \mathcal{Y}, \{W(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\})$  with  $V > 0$  and any sequence of integers  $M_n$  such that  $\log M_n = nC - n\rho_n$ , where  $\rho_n > 0, \rho_n \rightarrow 0$  and  $n\rho_n^2 \rightarrow \infty$ . For any  $T \in \mathbb{N}$ , there exists a sequence of  $(n, M_n, \epsilon_n, T_n)$ -streaming codes with average delay constraint such that

$$\lim_{n \rightarrow \infty} T_n = T \quad (85)$$

$$\limsup_{n \rightarrow \infty} \frac{1}{n\rho_n} \log \epsilon_n \leq -T. \quad (86)$$

We note that the exponent of the error probability  $\epsilon_n$  is of the order  $n\rho_n$  (instead of  $n\rho_n^2$  as in (81)), and hence it is improved tremendously by allowing variable decoding delay. Our strategy here is to utilize upper bounds on the error probability in Theorem 6 to estimate the average decoding delay  $T$  and the overall error probability. The use of an errors-and-erasures code for the purpose of channel coding with variable-length feedback was done in Nakiboğlu and Gallager in [23] among others. Our strategy here is partly inspired by theirs.

## VI. CONCLUSION

In this paper, we studied the fundamental interplay between the rate and error probability for a streaming setup with a decoding delay of  $T$  blocks. In the moderate deviations regime, the moderate deviations constant was shown to improve by at least a factor of  $T$ . We proposed a coding technique with infinite memory such that all the previous and fresh messages are jointly encoded in each block. On the other hand, in the central limit regime, the second-order coding rate was shown to improve by approximately a factor of  $\sqrt{T}$  for a wide range of channel parameters. To ensure that the summation of Berry-Esseen constants (e.g., the last terms in the RHS of (53)-(55)) does not diverge in the error analysis, we proposed a coding technique with truncated memory such that the encoding and decoding memories do not grow with the block index. Furthermore, we generalized the moderate deviations result in various directions. We first considered a scenario with an erasure option at the decoder and showed that both the exponents of the total error and the undetected error probabilities improve by factors of  $T$ . Then, by utilizing the erasure option, we showed that the exponent of the total error probability can be improved to that of the undetected error probability (in the order sense) at the expense of a variable decoding delay. We note that all of our encoding strategies do not depend on  $T$ . Hence, our coding techniques are directly applicable for multicast scenario where a sender transmits a common stream of data packets to multiple receivers with possibly different decoding constraints. While the model in this paper assumes that the rate of each message is fixed, our encoding and decoding schemes as well as the error analysis can be easily extended to the case when the message rates are not fixed.

Let us conclude with a final remark on proving a converse in our streaming setup. Our problem appears to be closely related to the bit-wise unequal protection (UEP) problem in

the sense that we need to capture the tension that arises when a common channel is used for more than one messages with individual error criteria.<sup>15</sup> For the seemingly simpler bit-wise UEP problem [26] for the block channel coding with the *same* decoding deadline, however, tight characterizations of various asymptotic fundamental limits (e.g., error exponents) remain challenging open problems in general. This indicates that a highly-nontrivial converse technique would be needed for our streaming setup where the messages have *different* decoding deadlines. Very recently, in [27], a converse bound on the moderate deviations constant was obtained for a slightly different streaming setup, which turns out to be tight for output symmetric channels for a certain range of moderate deviations scalings.

#### ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewer for tightening the bound (16) in Corollary 3 by suggesting a bounding technique that yields  $b_{L,V,T}$ .

#### APPENDIX A PROOF OF COROLLARY 3

Let  $\mu := L/\sqrt{V}$ . Note that Corollary 3 is proved if we show

$$\sum_{j=T}^{\infty} Q(\mu\sqrt{j}) < \frac{1 + \mu^2 T}{\mu^2 T(1 - \exp\{-\mu^2/2\})} Q(\mu\sqrt{T}) \quad (87)$$

and

$$\sum_{j=T}^{\infty} Q(\mu\sqrt{j}) < \left(1 + \frac{2}{\mu^2}\right) Q(\mu\sqrt{T}). \quad (88)$$

To that end, we use the following bounds on the  $Q$ -function:

$$\frac{x\phi(x)}{1+x^2} < Q(x) < \frac{\phi(x)}{x} \quad \forall x > 0, \quad (89)$$

where  $\phi(x) := \frac{1}{\sqrt{2\pi}} \exp\{-\frac{x^2}{2}\}$ . First, (87) is derived as follows:

$$\sum_{j=T}^{\infty} Q(\mu\sqrt{j}) < \sum_{j=T}^{\infty} \frac{\phi(\mu\sqrt{j})}{\mu\sqrt{j}} \quad (90)$$

$$< \frac{1}{\mu\sqrt{T}} \sum_{j=T}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\{-\mu^2 j/2\} \quad (91)$$

$$= \frac{1}{\mu\sqrt{T}} \cdot \frac{1}{\sqrt{2\pi}} \frac{\exp\{-\mu^2 T/2\}}{1 - \exp\{-\mu^2/2\}} \quad (92)$$

$$< \frac{1 + \mu^2 T}{\mu^2 T} \cdot \frac{Q(\mu\sqrt{T})}{1 - \exp\{-\mu^2/2\}}. \quad (93)$$

Next, let us prove (88). Note that

$$\sum_{j=T}^{\infty} Q(\mu\sqrt{j}) = Q(\mu\sqrt{T}) + \sum_{j=T+1}^{\infty} Q(\mu\sqrt{j}) \quad (94)$$

<sup>15</sup>We note that there are two types of UEP problems, i.e., bit-wise and message-wise UEP, but our streaming setup is more related to the bit-wise UEP. For example, the message-wise UEP problem studied by Shkel et al. [25] simply considers partitioning a *single* message set into several sub-message sets with different error probability requirements.

$$\stackrel{(a)}{\leq} Q(\mu\sqrt{T}) + \int_T^{\infty} Q(\mu\sqrt{u}) du, \quad (95)$$

where (a) is from Lemma 5. The second term in the RHS of (95) can be rewritten as follows by the change of variables  $\mu\sqrt{u} = v$ :

$$\int_T^{\infty} Q(\mu\sqrt{u}) du = \frac{1}{\mu^2} \int_{\mu\sqrt{T}}^{\infty} 2vQ(v) dv. \quad (96)$$

Now by applying the technique of integration by parts twice, we have

$$\int_{\mu\sqrt{T}}^{\infty} 2vQ(v) dv = [v^2 Q(v)]_{\mu\sqrt{T}}^{\infty} + \int_{\mu\sqrt{T}}^{\infty} v^2 \frac{1}{\sqrt{2\pi}} e^{-v^2/2} dv \quad (97)$$

$$= -\mu^2 T Q(\mu\sqrt{T}) + \int_{\mu\sqrt{T}}^{\infty} v^2 \frac{1}{\sqrt{2\pi}} e^{-v^2/2} dv \quad (98)$$

$$= -\mu^2 T Q(\mu\sqrt{T}) + \left[-v \frac{1}{\sqrt{2\pi}} e^{-v^2/2}\right]_{\mu\sqrt{T}}^{\infty} + \int_{\mu\sqrt{T}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-v^2/2} dv \quad (99)$$

$$= -\mu^2 T Q(\mu\sqrt{T}) + \mu\sqrt{T} \frac{1}{\sqrt{2\pi}} e^{-\mu^2 T/2} + Q(\mu\sqrt{T}). \quad (100)$$

Note that

$$(1 + \mu^2 T) Q(\mu\sqrt{T}) \stackrel{(a)}{>} (1 + \mu^2 T) \frac{\mu\sqrt{T}}{1 + \mu^2 T} \frac{1}{\sqrt{2\pi}} e^{-\mu^2 T/2} \quad (101)$$

$$= \mu\sqrt{T} \frac{1}{\sqrt{2\pi}} e^{-\mu^2 T/2}, \quad (102)$$

where (a) is due to the lower bound on the  $Q$  function in (89). By combining (100) and (102), we obtain

$$\int_{\mu\sqrt{T}}^{\infty} 2vQ(v) dv < 2Q(\mu\sqrt{T}). \quad (103)$$

From (95), (96), and (103), we finally obtain (88).  $\blacksquare$

#### APPENDIX B PROOF OF LEMMA 4

Fix  $n \in \mathbb{N}$  and  $s \geq 0$ . Then, we have

$$\Pr\left(\frac{1}{n} \sum_{l=1}^n Z_l \geq \varepsilon_n\right) \leq \Pr\left(\exp\left\{s \sum_{l=1}^n Z_l\right\} \geq \exp\{ns\varepsilon_n\}\right) \quad (104)$$

$$\stackrel{(a)}{\leq} \exp\{-ns\varepsilon_n\} \mathbb{E}\left[\exp\left\{s \sum_{l=1}^n Z_l\right\}\right] \quad (105)$$

$$\stackrel{(b)}{=} \exp\{-n(s\varepsilon_n - \log \mathbb{E}[\exp\{sZ_1\}])\} \quad (106)$$

$$= \exp\{-n(s\varepsilon_n - h(s))\}. \quad (107)$$

where (a) follows from Markov's inequality and (b) follows from the independence of  $Z_l$ 's.

The third-order Taylor series expansion of the cumulant generating function  $h(s)$  can be written as

$$h(s) = h(0) + sh'(0) + \frac{s^2}{2}h''(0) + \frac{s^3}{6}h'''(\tilde{s}) \quad (108)$$

for some  $\tilde{s} \in [0, s]$ . It is easy to check that  $h(0) = 0$ ,  $h'(0) = E[Z_1] = 0$  and  $h''(0) = \text{Var}[Z_1] = \sigma^2$ . Now, we take

$$s := \frac{\varepsilon_n}{\sigma^2}. \quad (109)$$

Plugging this into (107) and (108) yields

$$\Pr \left( \frac{1}{n} \sum_{l=1}^n Z_l \geq \varepsilon_n \right) \leq \exp \left\{ -n \left( \frac{\varepsilon_n^2}{\sigma^2} - \frac{\varepsilon_n^2}{2\sigma^2} - \frac{\varepsilon_n^3}{6\sigma^6} h'''(\tilde{s}) \right) \right\} \quad (110)$$

$$\leq \exp \left\{ -n \left( \frac{\varepsilon_n^2}{2\sigma^2} - \frac{\varepsilon_n^3}{6\sigma^6} K \right) \right\}, \quad (111)$$

where the final inequality holds for all  $n$  sufficiently large since  $\varepsilon_n \rightarrow 0$  and  $\tilde{s} \rightarrow 0$  as  $n \rightarrow \infty$  and thus  $|h'''(\tilde{s})| \leq K$ . ■

#### APPENDIX C A CHAIN OF EQUALITIES

The following chain of equalities is used in the proof Theorem 2.

$$\Pr \left( \sum_{l=1}^{n(\alpha-B+1)} i(X_l; \bar{Y}_l) > \alpha \cdot \log M_n \right) = \sum_{s=1}^{n(\alpha-B+1)} \sum_{x_s, \bar{y}_s} \left( \prod_{t=1}^{n(\alpha-B+1)} P_X(x_t) P_X W(\bar{y}_t) \right) \cdot \mathbb{1} \left\{ \sum_{l=1}^{n(\alpha-B+1)} i(x_l; \bar{y}_l) > \alpha \cdot \log M_n \right\} \quad (112)$$

$$= \sum_{s=1}^{n(\alpha-B+1)} \sum_{x_s, \bar{y}_s} \left( \prod_{t=1}^{n(\alpha-B+1)} P_X(x_t) W(\bar{y}_t | x_t) \frac{P_X W(\bar{y}_t)}{W(\bar{y}_t | x_t)} \right) \cdot \mathbb{1} \left\{ \sum_{l=1}^{n(\alpha-B+1)} i(x_l; \bar{y}_l) > \alpha \cdot \log M_n \right\} \quad (113)$$

$$= \sum_{s=1}^{n(\alpha-B+1)} \sum_{x_s, \bar{y}_s} \left( \prod_{t=1}^{n(\alpha-B+1)} P_X(x_t) W(\bar{y}_t | x_t) \right) \cdot \exp \left\{ - \sum_{l=1}^{n(\alpha-B+1)} i(x_l; \bar{y}_l) \right\} \cdot \mathbb{1} \left\{ \sum_{l=1}^{n(\alpha-B+1)} i(x_l; \bar{y}_l) > \alpha \cdot \log M_n \right\} \quad (114)$$

$$= E \left[ \exp \left\{ - \sum_{l=1}^{n(\alpha-B+1)} i(X_l; Y_l) \right\} \cdot \mathbb{1} \left\{ \sum_{l=1}^{n(\alpha-B+1)} i(X_l; Y_l) > \alpha \log M_n \right\} \right]. \quad (115)$$

#### APPENDIX D PROOF OF THEOREM 6

Consider a DMC  $(\mathcal{X}, \mathcal{Y}, \{W(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\})$  with  $V > 0$  and any sequence of integers  $M_n$  such that  $\log M_n = nC - n\rho_n$ , where  $\rho_n > 0$ ,  $\rho_n \rightarrow 0$  and  $n\rho_n^2 \rightarrow \infty$ . We denote by  $P_X$  an input distribution that achieves the dispersion (9). Fix  $0 < \gamma < 1$ .

The encoding procedure is the same as that for the basic streaming setup in Section IV-A. Let us consider the decoding of  $G_k$  at the end of block  $T_k$ . The decoding procedure is modified from that for the basic streaming setup in Section IV-A as follows:

- The decoding test (20) is modified as follows:

$$i(\mathbf{x}_{[j:T_k]}(\hat{G}_{T_k, [1:j-1]}, g_{[j:T_k]}), \mathbf{y}_{[j:T_k]}) > (T_k - j + 1) \cdot (\log M_n + \gamma n \rho_n), \quad (116)$$

i.e., the threshold value is increased proportional to  $\gamma$ .

- If there is none or more than one  $g_j$  that satisfies the decoding test (116) for some  $g_{[j+1:T_k]}$ , the decoder declares an erasure, i.e.,  $\hat{G}_k = 0$ , and terminates the decoding procedure.

Similarly as in Section IV-A, we first consider the probability of error averaged over random codebook  $\mathcal{C}_n$ . The error event  $\{\hat{G}_k \neq G_k\}$  for  $k \in \mathbb{N}$  happens only if at least one of the following  $2k$  events occurs:

$$\mathcal{E}'_{k,j} := \{i(\mathbf{X}_{[j:T_k]}(G^{T_k}), \mathbf{Y}_{[j:T_k]}) \leq (T_k - j + 1) \cdot (\log M_n + \gamma n \rho_n)\}, \quad j \in [1 : k] \quad (117)$$

$$\mathcal{E}''_{k,j} := \{i(\mathbf{X}_{[j:T_k]}(G^{j-1}, g_{[j:T_k]}), \mathbf{Y}_{[j:T_k]}) > (T_k - j + 1) \cdot (\log M_n + \gamma n \rho_n) \text{ for some } g_{[j:T_k]} \text{ such that } g_j \neq G_j\}, \quad j \in [1 : k]. \quad (118)$$

We note that (117) and (118) are obtained by replacing  $\log M_n$  by  $\log M_n + \gamma n \rho_n$  in (21) and (22), respectively. Then, we have

$$E_{\mathcal{C}_n}[\Pr(\hat{G}_k \neq G_k | \mathcal{C}_n)] \leq \sum_{j=1}^k \left( \Pr(\mathcal{E}'_{k,j}) + \Pr(\mathcal{E}''_{k,j}) \right). \quad (119)$$

On the other hand, the undetected error event  $\{\hat{G}_k \neq G_k, \hat{G}_k \neq 0\}$  has the following relationship:

$$\{\hat{G}_k \neq G_k, \hat{G}_k \neq 0\} \subseteq \{\hat{G}_{T_k, [1:k]} \neq G_{[1:k]}, \hat{G}_k \neq 0\} \quad (120)$$

$$= \cup_{j \in [1:k]} \{\hat{G}_{T_k, [1:j-1]} = G_{[1:j-1]}, \hat{G}_{T_k, j} \neq G_j, \hat{G}_k \neq 0\}. \quad (121)$$

Hence, the undetected error probability is bounded as follows:

$$E_{\mathcal{C}_n}[\Pr(\hat{G}_k \neq G_k, \hat{G}_k \neq 0 | \mathcal{C}_n)] \leq \sum_{j=1}^k \Pr(\hat{G}_{T_k, [1:j-1]} = G_{[1:j-1]}, \hat{G}_{T_k, j} \neq G_j, \hat{G}_k \neq 0) \quad (122)$$

$$\leq \sum_{j=1}^k \Pr(\tilde{\mathcal{E}}'_{k,j}). \quad (123)$$

Now, for  $j \in [1 : k]$ , let us bound  $\Pr(\mathcal{E}'_{k,j})$  and  $\Pr(\tilde{\mathcal{E}}'_{k,j})$ . Similarly as in Section IV-A,  $(X_l, Y_l, \bar{Y}_l)$ 's denote i.i.d. random variables each generated according to  $P_X(x_l)W(y_l|x_l)P_XW(\bar{y}_l)$  in the following. First, we have

$$\begin{aligned} & \Pr(\mathcal{E}'_{k,j}) \\ & \leq \Pr \left( \sum_{l=1}^{n(T_k-j+1)} i(X_l; Y_l) \leq (T_k - j + 1) \right. \\ & \quad \left. \cdot (\log M_n + \gamma n \rho_n) \right) \end{aligned} \quad (124)$$

$$\begin{aligned} & \leq \Pr \left( \sum_{l=1}^{n(T_k-j+1)} i(X_l; Y_l) \leq (T_k - j + 1) \right. \\ & \quad \left. \cdot n(C - (1 - \gamma)\rho_n) \right) \end{aligned} \quad (125)$$

$$\begin{aligned} & \stackrel{(a)}{\leq} \exp \left\{ -(T_k - j + 1)n \left( \frac{(1 - \gamma)^2 \rho_n^2}{2V} \right. \right. \\ & \quad \left. \left. - (1 - \gamma)^3 \rho_n^3 \tau \right) \right\} \end{aligned} \quad (126)$$

for sufficiently large  $n$ , where  $\tau$  is some non-negative constant dependent only on the input distribution  $P_X(x)$  and channel statistics  $W(y|x)$  and (a) is from Lemma 4 in Section IV-A.

Next, we have

$$\begin{aligned} & \Pr(\tilde{\mathcal{E}}'_{k,j}) \\ & \leq M_n^{T_k-j+1} \cdot \Pr \left( \sum_{l=1}^{n(T_k-j+1)} i(X_l; \bar{Y}_l) \right. \\ & \quad \left. > (T_k - j + 1) \cdot (\log M_n + \gamma n \rho_n) \right) \end{aligned} \quad (127)$$

$$\begin{aligned} & \stackrel{(a)}{=} M_n^{T_k-j+1} \cdot \mathbb{E} \left[ \exp \left\{ - \sum_{l=1}^{n(T_k-j+1)} i(X_l; Y_l) \right\} \right. \\ & \quad \cdot \mathbb{1} \left\{ \sum_{l=1}^{n(T_k-j+1)} i(X_l; Y_l) > (T_k - j + 1) \right. \\ & \quad \left. \cdot (\log M_n + \gamma n \rho_n) \right\} \Big] \end{aligned} \quad (128)$$

$$\leq M_n^{T_k-j+1} \exp\{-(T_k - j + 1)(\log M_n + \gamma n \rho_n)\} \quad (129)$$

$$= \exp\{-(T_k - j + 1)\gamma n \rho_n\}, \quad (130)$$

where (a) is obtained by applying a chain of equalities similar to that in Appendix C.

Hence, we obtain

$$\begin{aligned} & \mathbb{E}_{\mathcal{C}_n}[\Pr(\hat{G}_k \neq G_k | \mathcal{C}_n)] \\ & \leq \sum_{j=1}^k \left( \exp \left\{ -(T_k - j + 1)n \rho_n^2 (1 - \gamma)^2 \right. \right. \\ & \quad \left. \left. \cdot \left( \frac{1}{2V} - (1 - \gamma)\rho_n \tau \right) \right\} \right. \\ & \quad \left. + \exp\{-(T_k - j + 1)n\gamma\rho_n\} \right) \end{aligned} \quad (131)$$

$$\begin{aligned} & \leq \sum_{j=T}^{T_k} \left( \exp \left\{ -jn \rho_n^2 (1 - \gamma)^2 \left( \frac{1}{2V} - (1 - \gamma)\rho_n \tau \right) \right\} \right. \\ & \quad \left. + \exp\{-jn\gamma\rho_n\} \right) \end{aligned} \quad (132)$$

$$\begin{aligned} & \leq \frac{\exp\{-Tn\rho_n^2(1-\gamma)^2(\frac{1}{2V} - (1-\gamma)\rho_n\tau)\}}{1 - \exp\{-n\rho_n^2(1-\gamma)^2(\frac{1}{2V} - (1-\gamma)\rho_n\tau)\}} \\ & \quad + \frac{\exp\{-Tn\gamma\rho_n\}}{1 - \exp\{-n\gamma\rho_n\}} \end{aligned} \quad (133)$$

and

$$\mathbb{E}_{\mathcal{C}_n}[\Pr(\hat{G}_k \neq G_k, \hat{G}_k \neq 0 | \mathcal{C}_n)] \leq \frac{\exp\{-Tn\gamma\rho_n\}}{1 - \exp\{-n\gamma\rho_n\}} \quad (134)$$

for sufficiently large  $n$ .

To show the existence of a deterministic code, we apply Markov's inequality as follows<sup>16</sup>:

$$\begin{aligned} & \Pr \left( \limsup_{N \rightarrow \infty} \sum_{k=1}^N \frac{\Pr(\hat{G}_k \neq G_k | \mathcal{C}_n)}{N} \right. \\ & \quad \left. > 2 \limsup_{N \rightarrow \infty} \sum_{k=1}^N \frac{\mathbb{E}_{\mathcal{C}_n}[\Pr(\hat{G}_k \neq G_k | \mathcal{C}_n)]}{N} \right) < \frac{1}{2} \end{aligned} \quad (135)$$

$$\begin{aligned} & \Pr \left( \limsup_{N \rightarrow \infty} \sum_{k=1}^N \frac{\Pr(\hat{G}_k \neq G_k, \hat{G}_k \neq 0 | \mathcal{C}_n)}{N} \right. \\ & \quad \left. > 2 \limsup_{N \rightarrow \infty} \sum_{k=1}^N \frac{\mathbb{E}_{\mathcal{C}_n}[\Pr(\hat{G}_k \neq G_k, \hat{G}_k \neq 0 | \mathcal{C}_n)]}{N} \right) \\ & < \frac{1}{2}. \end{aligned} \quad (136)$$

Then, from the union bound, we have

$$\begin{aligned} & \Pr \left( \limsup_{N \rightarrow \infty} \sum_{k=1}^N \frac{\Pr(\hat{G}_k \neq G_k | \mathcal{C}_n)}{N} \right. \\ & \quad \left. > 2 \limsup_{N \rightarrow \infty} \sum_{k=1}^N \frac{\mathbb{E}_{\mathcal{C}_n}[\Pr(\hat{G}_k \neq G_k | \mathcal{C}_n)]}{N} \right) \text{ or} \\ & \quad \limsup_{N \rightarrow \infty} \sum_{k=1}^N \frac{\Pr(\hat{G}_k \neq G_k, \hat{G}_k \neq 0 | \mathcal{C}_n)}{N} \\ & \quad \left. > 2 \limsup_{N \rightarrow \infty} \sum_{k=1}^N \frac{\mathbb{E}_{\mathcal{C}_n}[\Pr(\hat{G}_k \neq G_k, \hat{G}_k \neq 0 | \mathcal{C}_n)]}{N} \right) \\ & < 1. \end{aligned} \quad (137)$$

<sup>16</sup>Such a technique of applying Markov's inequality to derandomize the code was used in the proof of [21, Theorem 1].

Therefore, there must exist a sequence of codes  $\mathcal{C}_n$  that satisfies

$$\begin{aligned} & \limsup_{N \rightarrow \infty} \sum_{k=1}^N \frac{\Pr(\hat{G}_k \neq G_k | \mathcal{C}_n)}{N} \\ & \leq 2 \exp \left\{ -n\rho_n^2 \left( (1-\gamma)^2 \frac{T}{2V} + o(1) \right) \right\} \end{aligned} \quad (138)$$

and

$$\begin{aligned} & \limsup_{N \rightarrow \infty} \sum_{k=1}^N \frac{\Pr(\hat{G}_k \neq G_k, \hat{G}_k \neq 0 | \mathcal{C}_n)}{N} \\ & \leq 2 \exp \{-n\rho_n(T\gamma + o(1))\}, \end{aligned} \quad (139)$$

which completes the proof.  $\blacksquare$

#### APPENDIX E PROOF OF THEOREM 7

Consider a DMC  $(\mathcal{X}, \mathcal{Y}, \{W(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\})$  with  $V > 0$  and any sequence of integers  $M_n$  such that  $\log M_n = nC - n\rho_n$ , where  $\rho_n > 0, \rho_n \rightarrow 0$  and  $n\rho_n^2 \rightarrow \infty$ . We denote by  $P_X$  an input distribution that achieves the dispersion (9). Fix  $T \in \mathbb{N}$  and  $0 < \gamma < 1$ .

The encoding procedure is the same as that for the basic streaming setup in Section IV-A. Let us consider the decoding of message  $G_k$  at the end of block  $k + d - 1$  for  $d \in \mathbb{N}$ .<sup>17</sup> If  $d \in [1 : T - 1]$ , the decoder outputs  $\hat{G}_{k+d-1,k} = 0$ . For  $d \geq T$ , the decoder outputs a message estimate  $\hat{G}_{k+d-1,k} \in \mathcal{G}$  or an erasure symbol  $\hat{G}_{k+d-1,k} = 0$  according to the same decoding rule illustrated in Appendix D with delay  $d$ . Then, the error probability of  $G_k$  after the random decoding delay  $D_k = \min\{d : \hat{G}_{k+d-1,k} \neq 0\}$  (averaged over the random codebook generation) is bounded as follows:

$$\begin{aligned} & \mathbb{E}_{\mathcal{C}_n} [\Pr(\hat{G}_{k+D_k-1,k} \neq G_k | \mathcal{C}_n)] \\ & = \sum_{d=T}^{\infty} \mathbb{E}_{\mathcal{C}_n} [\Pr(D_k = d, \hat{G}_{k+d-1,k} \neq G_k, \\ & \quad \hat{G}_{k+d-1,k} \neq 0 | \mathcal{C}_n)] \end{aligned} \quad (140)$$

$$\leq \sum_{d=T}^{\infty} \mathbb{E}_{\mathcal{C}_n} [\Pr(\hat{G}_{k+d-1,k} \neq G_k, \hat{G}_{k+d-1,k} \neq 0 | \mathcal{C}_n)] \quad (141)$$

$$\stackrel{(a)}{\leq} \sum_{d=T}^{\infty} \frac{\exp\{-dn\gamma\rho_n\}}{1 - \exp\{-n\gamma\rho_n\}} \quad (142)$$

$$\leq \frac{\exp\{-Tn\gamma\rho_n\}}{(1 - \exp\{-n\gamma\rho_n\})^2}, \quad (143)$$

where (a) is from the upper bound (134) on the undetected error probability with delay  $d$ .

On the other hand, the excess of delay averaged over the random codebook generation is bounded as

$$\begin{aligned} & \mathbb{E}_{\mathcal{C}_n}[D_k - T | \mathcal{C}_n] \\ & = \Pr(D_k = T + 1 | \mathcal{C}_n) + 2\Pr(D_k = T + 2 | \mathcal{C}_n) + \dots \end{aligned} \quad (144)$$

<sup>17</sup>We note that in the definition of a streaming code with average delay constraint, the decoder decodes  $G_k$  at the end of every block  $k + d - 1$  for  $d \in \mathbb{N}$ .

$$\begin{aligned} & = \Pr(\hat{G}_{k+T-1,k} = 0, \hat{G}_{k+T,k} \neq 0 | \mathcal{C}_n) \\ & \quad + 2\Pr(\hat{G}_{k+T-1,k} = 0, \hat{G}_{k+T,k} = 0, \\ & \quad \quad \hat{G}_{k+T+1,k} \neq 0 | \mathcal{C}_n) + \dots \end{aligned} \quad (145)$$

$$\leq \sum_{d=T+1}^{\infty} (d - T) \cdot \Pr(\hat{G}_{k+d-2,k} = 0 | \mathcal{C}_n) \quad (146)$$

$$\leq \sum_{d=T+1}^{\infty} (d - T) \cdot \Pr(\hat{G}_{k+d-2,k} \neq G_k | \mathcal{C}_n) \quad (147)$$

$$\begin{aligned} & \stackrel{(a)}{\leq} \sum_{d=T+1}^{\infty} (d - T) \\ & \quad \cdot \left( \frac{\exp\{-(d-1)n\rho_n^2(1-\gamma)^2(\frac{1}{2V} - (1-\gamma)\rho_n\tau)\}}{1 - \exp\{-n\rho_n^2(1-\gamma)^2(\frac{1}{2V} - (1-\gamma)\rho_n\tau)\}} \right. \\ & \quad \left. + \frac{\exp\{-(d-1)n\gamma\rho_n\}}{1 - \exp\{-n\gamma\rho_n\}} \right), \end{aligned} \quad (148)$$

where (a) is from the upper bound (133) on the total error probability with delay  $d - 1$ .

By following similar statements using Markov's inequality in the proof of Theorem 6, we can obtain

$$\begin{aligned} & \Pr \left( \limsup_{N \rightarrow \infty} \sum_{k=1}^N \frac{\Pr(\hat{G}_{k+D_k-1,k} \neq G_k | \mathcal{C}_n)}{N} \right. \\ & \quad > 2 \limsup_{N \rightarrow \infty} \sum_{k=1}^N \frac{\mathbb{E}_{\mathcal{C}_n}[\Pr(\hat{G}_{k+D_k-1,k} \neq G_k | \mathcal{C}_n)]}{N} \\ & \quad \text{or } \limsup_{N \rightarrow \infty} \sum_{k=1}^N \frac{\mathbb{E}[D_k | \mathcal{C}_n]}{N} - T \\ & \quad \left. > 2 \limsup_{N \rightarrow \infty} \sum_{k=1}^N \frac{\mathbb{E}_{\mathcal{C}_n}[D_k - T | \mathcal{C}_n]}{N} \right) < 1. \end{aligned} \quad (149)$$

Therefore, there must exist a sequence of codes  $\mathcal{C}_n$  that satisfies

$$\begin{aligned} & \limsup_{N \rightarrow \infty} \sum_{k=1}^N \frac{\Pr(\hat{G}_{k+D_k-1,k} \neq G_k | \mathcal{C}_n)}{N} \\ & \leq 2 \exp\{-n\rho_n(T\gamma + o(1))\} \end{aligned} \quad (150)$$

and<sup>18</sup>

$$\limsup_{N \rightarrow \infty} \sum_{k=1}^N \frac{\mathbb{E}[D_k | \mathcal{C}_n]}{N} \leq T + o(1). \quad (151)$$

We note that (150) implies

$$\begin{aligned} & \limsup_{n \rightarrow \infty} \frac{1}{n\rho_n} \log \left( \limsup_{N \rightarrow \infty} \sum_{k=1}^N \frac{\Pr(\hat{G}_{k+D_k-1,k} \neq G_k | \mathcal{C}_n)}{N} \right) \\ & \leq -T\gamma. \end{aligned} \quad (152)$$

By taking  $\gamma \rightarrow 1$ , this completes the proof.  $\blacksquare$

<sup>18</sup>By calculating the infinite series in the RHS of (148), it can be verified that the RHS of (148) converges to 0 as  $n$  tends to infinity.

## REFERENCES

- [1] S.-H. Lee, V. Y. F. Tan, and A. Khisti, "Streaming data transmission in the moderate deviations and central limit regimes," in *2016 IEEE International Symposium on Information Theory (ISIT)*, July 2016, pp. 3072–3076.
- [2] L. Schulman, "Coding for interactive communication," *IEEE Trans. Inf. Theory*, vol. 42, pp. 1745–1756, Nov. 1996.
- [3] A. Sahai, "Anytime information theory," Ph.D. dissertation, Massachusetts Institute of Technology (MIT), 2001.
- [4] R. Sukhavasi and B. Hassibi, "Linear error correcting codes with anytime reliability," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, Jul.-Aug. 2011, pp. 1748–1752.
- [5] E. Martinian and C. E. W. Sundberg, "Burst erasure correction codes with low decoding delay," *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2494–2502, Oct 2004.
- [6] A. Badr, A. Khisti, W. t. Tan, and J. Apostolopoulos, "Streaming codes with partial recovery over channels with burst and isolated erasures," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 3, pp. 501–516, April 2015.
- [7] R. Mahmood, A. Badr, and A. Khisti, "Convolutional codes with maximum column sum rank for network streaming," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3039–3052, June 2016.
- [8] A. Khisti and S. C. Draper, "The streaming-DMT of fading channels," *IEEE Trans. Inf. Theory*, vol. 60, pp. 7058–7072, Nov. 2014.
- [9] S. C. Draper and A. Khisti, "Truncated tree codes for streaming data: Infinite-memory reliability using finite memory," in *Proc. International Symposium on Wireless Communication Systems (ISWCS)*, Nov. 2011, pp. 136–140.
- [10] S. C. Draper, C. Chang, and A. Sahai, "Lossless coding for distributed streaming sources," *IEEE Trans. Inf. Theory*, vol. 60, pp. 1447–1474, Mar. 2014.
- [11] V. Strassen, "Asymptotische Abschätzungen in Shannons Informationstheorie," in *Trans. Third Prague Conf. Inf. Theory*, Prague, 1962, pp. 689–723, <http://www.math.cornell.edu/~pmlut/strassen.pdf>.
- [12] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, pp. 2307–2359, May 2010.
- [13] M. Hayashi, "Information spectrum approach to second-order coding rate in channel coding," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 4947–4966, 2009.
- [14] V. Y. F. Tan, *Asymptotic estimates in information theory with non-vanishing error probabilities*. Foundations and Trends® in Communications and Information Theory, 2014, vol. 11, no. 1-2.
- [15] D.-K. He, L. Lastras-Montaña, E.-H. Yang, A. Jagmohan, and J. Chen, "On the redundancy of Slepian–Wolf coding," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5607–5627, 2009.
- [16] Y. Altug and A. B. Wagner, "Moderate deviations in channel coding," *IEEE Trans. Inf. Theory*, vol. 60, pp. 4417–4426, Aug. 2014.
- [17] Y. Polyanskiy and S. Verdú, "Channel dispersion and moderate deviations limits for memoryless channels," in *Proc. 48th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, 2010.
- [18] Y. Altug and A. Wagner, "Moderate deviation analysis of channel coding: Discrete memoryless case," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, Jun. 2010, pp. 265–269.
- [19] Z. Lin, V. Y. F. Tan, and M. Motani, "On error exponents and moderate deviations for lossless streaming compression of correlated sources," *IEEE Trans. Inf. Theory*, submitted for publication. [Online]. Available: <http://arxiv.org/abs/1507.03190>.
- [20] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*. New York: Springer Verlag., 2009.
- [21] M. Hayashi and V. Y. F. Tan, "Asymmetric evaluations of erasure and undetected error probabilities," *IEEE Trans. Inf. Theory*, vol. 61, pp. 6560–6577, Dec. 2015.
- [22] G. D. Forney, "Exponential error bounds for erasure, list and decision feedback schemes," *IEEE Trans. Inf. Theory*, vol. 14, pp. 206–220, Mar 1968.
- [23] B. Nakiboglu and R. G. Gallager, "Error exponents for variable-length block codes with feedback and cost constraints," *IEEE Transactions on Information Theory*, vol. 54, no. 3, pp. 945–963, March 2008.
- [24] W. Feller, *An Introduction to Probability Theory and its Applications*. New York: Wiley, 1971, vol. II.
- [25] Y. Y. Shkel, V. Y. F. Tan, and S. C. Draper, "Unequal message protection: Asymptotic and non-asymptotic tradeoffs," *IEEE Trans. Inf. Theory*, vol. 61, pp. 5396–5416, Oct. 2015.
- [26] S. Borade, B. Nakiboğlu, and L. Zheng, "Unequal error protection: An information-theoretic perspective," *IEEE Trans. Inf. Theory*, vol. 55, pp. 5511–5539, Dec. 2009.
- [27] S.-H. Lee, V. Y. F. Tan, and A. Khisti, "Exact moderate deviation asymptotics in streaming data transmission," *IEEE Trans. Inf. Theory*, submitted for publication. [Online]. Available: <http://arxiv.org/abs/1604.06848>.

**Si-Hyeon Lee** (S'08-M'13) received the B.S. (summa cum laude) and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2007 and 2013, respectively. She is now a Postdoctoral Fellow at the Department of Electrical and Computer Engineering, University of Toronto. Her research interests include network information theory, physical layer security, and wireless communication systems.

**Vincent Y. F. Tan** (S'07-M'11-SM'15) was born in Singapore in 1981. He is currently an Assistant Professor in the Department of Electrical and Computer Engineering (ECE) and the Department of Mathematics at the National University of Singapore (NUS). He received the B.A. and M.Eng. degrees in Electrical and Information Sciences from Cambridge University in 2005 and the Ph.D. degree in Electrical Engineering and Computer Science (EECS) from the Massachusetts Institute of Technology in 2011. He was a postdoctoral researcher in the Department of ECE at the University of Wisconsin-Madison and a research scientist at the Institute for Infocomm (I<sup>2</sup>R) Research, A\*STAR, Singapore. His research interests include information theory, machine learning and statistical signal processing.

Dr. Tan received the MIT EECS Jin-Au Kong outstanding doctoral thesis prize in 2011, the NUS Young Investigator Award in 2014, and was placed on the NUS Faculty of Engineering Teaching Award commendation list in 2016. He has authored a research monograph on "*Asymptotic Estimates in Information Theory with Non-Vanishing Error Probabilities*" in the Foundations and Trends in Communications and Information Theory Series (NOW Publishers). He is currently an Associate Editor of the IEEE Transactions on Communications.

**Ashish Khisti** (S'02-M'08) received his B.A.Sc Degree in Engineering Sciences (Electrical Option) from University of Toronto, and his S.M and Ph.D. Degrees in Electrical Engineering from the Massachusetts Institute of Technology. Between 2009-2015, he was an assistant professor in the Electrical and Computer Engineering department at the University of Toronto. He is presently an associate professor, and holds a Canada Research Chair in the same department. He is a recipient of an Ontario Early Researcher Award, the Hewlett-Packard Innovation Research Award and the Harold H. Hazen teaching assistant award from MIT. He presently serves as an associate editor for IEEE Transactions on Information Theory and is also a guest editor for the Proceedings of the IEEE (Special Issue on Secure Communications via Physical-Layer and Information-Theoretic Techniques).