Asymmetric Evaluations of Erasure and Undetected Error Probabilities

Masahito Hayashi† and Vincent Y. F. Tan‡

Abstract

The problem of channel coding with the erasure option is revisited for discrete memoryless channels. The interplay between the code rate, the undetected and total error probabilities is characterized. Using the information spectrum method, a sequence of codes of increasing blocklengths \( n \) is designed to illustrate this tradeoff. Furthermore, for additive discrete memoryless channels, the ensemble performance of a sequence of random codes is also analyzed to demonstrate the optimality of the above-mentioned codes. The tradeoff between the code rate, undetected and total errors as well as the threshold in a generalized likelihood ratio test is characterized. Two asymptotic regimes are studied. First, the code rate tends to the capacity of the channel at a rate slower than \( n^{-1/2} \) corresponding to the moderate deviations regime. In this case, both error probabilities decay subexponentially and asymmetrically. The precise decay rates are characterized. Second, the code rate tends to capacity at a rate of \( n^{-1/2} \). In this case, the total error probability is asymptotically a positive constant while the undetected error probability decays as \( \exp(-bn^{1/2}) \) for some \( b > 0 \). The proof techniques involve applications of a modified (or “shifted”) version of the Gärtner-Ellis theorem and the type class enumerator method to characterize the asymptotic behavior of a sequence of cumulant generating functions.

Index Terms

Channel coding, Erasure decoding, Moderate deviations, Second-order coding rates, Large deviations, Gärtner-Ellis theorem

I. INTRODUCTION

A. Background

In channel coding, we are interested in designing a code that can reliably decode a message sent through a noisy channel. However, when the effect of the noise is so large such that the decoding system is not sufficiently confident of which message was sent, it is preferable to declare that an erasure event has occurred. In this way, the system avoids declaring that an incorrect message was sent, a costly mistake, and may use an automatic repeat request (ARQ) protocol or decision feedback system to resend the intended message. This paper revisits the information-theoretic limits of channel coding with the erasure option.

It has long been known since Forney’s seminal paper on decoding with the erasure option and list decoding [1] that the optimum decoder for a given codebook has the following structure: It outputs the message for which the likelihood of that message given the channel output exceeds a multiple \( \exp(nT) \) (where \( n \) is the blocklength of the code) of the sum of all the other likelihoods. This is a generalization of the likelihood ratio test which underlies the Neyman-Pearson lemma for binary hypothesis testing. For erasure decoding, the threshold \( T \) is set to a positive number so that the decoding regions are disjoint and furthermore, the erasure region is non-empty. Among our other contributions in this paper, we examine other possibly suboptimal decoding regions.

If the threshold \( T \) in Forney’s decoding regions is a fixed positive number not tending to zero, then it is known from his analysis [1] and many follow-up works [2]–[9] that both the undetected error probability and the erasure probability decay exponentially fast in \( n \) for an appropriately chosen codebook. Typically, and following in the spirit of Shannon’s seminal work [10], the codebook is randomly chosen. The constant \( T \) serves to tradeoff between the two error probabilities. This exponential decay in both error probabilities corresponds to large deviations analysis. However, there is substantial motivation to study other asymptotic regimes to gain greater insights about

† M. Hayashi is with the Graduate School of Mathematics, Nagoya University, and the Center for Quantum Technologies (CQT), National University of Singapore (Email: masahito@math.nagoya-u.ac.jp).
‡ V. Y. F. Tan is with the Department of Electrical and Computer Engineering and the Department of Mathematics, National University of Singapore (Email: vtan@nus.edu.sg).
the fundamental limits of channel codes with the erasure option. This corresponds to setting the threshold $T$ to be a positive sequence that tends to zero as the blocklength $n$ grows.

Strassen [11] pioneered the fixed error probability or second-order asymptotic analysis for discrete memoryless channels (DMCs) without the erasure option. There have been prominent works recently in this area by Hayashi [12] and Polyanskiy, Poor and Verdú [13]. Altu˘g and Wagner [14] pioneered the moderate deviations analysis for DMCs and Tan [15] considered the rate-distortion counterpart for discrete and Gaussian sources. Second-order and moderate deviations analyses respectively correspond to operating at coding rates that have a deviation of $\Theta(n^{-1/2})$ and $\omega(n^{-1/2})$ from the first-order fundamental limit, i.e., the capacity or the rate-distortion function. Tan and Moulin [16] recently studied the information-theoretic limits of channel coding with erasures where both the undetected and total error probabilities are fixed at positive constants.

B. Main Contributions

In this work, we study different regimes for the errors and erasure problems. In particular, we analyze the the moderate deviations [14], [17] and mixed regimes. For moderate deviations, the code rate tends towards capacity but deviates from it by a sequence that grows slower than $n^{-1/2}$. For the mixed regime, the undetected error is designed to decay as $\exp(-bn^{1/2})$ for some $b > 0$, but the total error is asymptotically a positive constant governed by the Gaussian distribution. Our main contributions are detailed as follows.

First, for the achievability results, we draw on ideas from information spectrum analysis [18] to present a sequence of block codes with the erasure option that demonstrate the above-mentioned asymmetric tradeoff between the undetected and total error probabilities.

Second, and most importantly, we show that ensemble-wise, our so-constructed codes above are optimal for additive DMCs. To do so, we consider Forney’s decoding regions [1] where the threshold parameter $T$ depends on $n$ and, in particular, is set to be a decaying sequence $\Theta(n^{-t})$ where $t \in (0, 1/2)$. We show that both the undetected and total error probabilities decay subexponentially (i.e., the moderate deviations regime [14], [15], [17], [19]) and asymmetrically in the sense that their decay rates are different. These decay rates depend on $t$ and also the implied constant the $\Theta(n^{-t})$ notation. In fact, we characterize the precise tradeoff between these error probabilities, the code rate as well as the threshold. Our technique, which is based the type class enumerator method [6]–[9], carry over to the mixed regime in which the total error probability is asymptotically a constant [11]–[13] while the undetected error decays as $\exp(-bn^{1/2})$. Just as for the pure moderate deviations setting, we characterize the precise tradeoffs between the different parameters in the system. The decay rates turn out to the same as for the achievability results showing that the decoder designed based on information spectrum analysis is, in fact, optimal.

Finally, an auxiliary contribution of the present work is a new mathematical tool. We develop of a modified (“shifted”) version of the Gärtner-Ellis theorem [20, Sec. 2.3] to prove our results concerning the asymptotics of the undetected and total error probabilities under both the moderate and mixed regimes. This generalization, presented in Theorem 8, appears to be distinct from other generalizations of the Gärtner-Ellis theorem in the literature (e.g., [21], [22]). It turns out to be very useful for our application and may be of independent interest in other information-theoretic settings.

C. Paper Organization

This paper is organized as follows: In Section II, we state our notation and the problem setup precisely. The main results are detailed in Section III where the direct results are in Section III-A and the ensemble converse results in Section III-B. The proofs the main results are deferred to Section IV. We conclude our discussion and suggest avenues for future work in Section V. The appendices contain some auxiliary mathematical tools including the modification of the Gärtner-Ellis theorem for general orders, which we use to estimate the both errors. This is presented as Theorem 8 in the Appendices.

II. Notation and Problem Setting

In this paper, we adopt standard notation in information theory, particularly in the book by Csiszár and Körner [23]. Random variables are denoted by upper case (e.g., $X$) and their realizations by lower case (e.g., $x$). All alphabets of the random variables are finite sets and are denoted by calligraphic font (e.g., $\mathcal{X}$). A sequence of letters from the $n$-fold Cartesian product $\mathcal{X}^n$ is denoted by boldface $\mathbf{x} = (x_1, \ldots, x_n)$. A sequence of random variables is denoted using
Let the set of capacity-achieving input distributions be function where the memoryless (and stationary), this means that given a sequence of input letters \( x = (x_1, \ldots, x_n) \) \( X \) is denoted by \( X \). All logs and exps are with respect to the natural base e.

We consider a DMC \( W \) with input alphabet \( \mathcal{X} \) and output alphabet \( \mathcal{Y} \). This is denoted as \( W : \mathcal{X} \to \mathcal{Y} \). By memoryless (and stationary), this means that given a sequence of input letters \( x = (x_1, \ldots, x_n) \in \mathcal{X}^n \) the probability of the output letters \( y = (y_1, \ldots, y_n) \in \mathcal{Y}^n \) is the product \( \prod_{i=1}^n W(y_i|x_i) \). The capacity of the DMC is denoted as

\[
C = C(W) := \max\{I(P_X, W) : P_X \in \mathcal{P}(\mathcal{X})\}.
\]

Let the set of capacity-achieving input distributions be

\[
\Pi = \Pi(W) := \{P_X \in \mathcal{P}(\mathcal{X}) : I(P_X, W) = C(W)\}
\]

A DMC is called additive if \( \mathcal{X} = \mathcal{Y} = \{0,1,\ldots,d-1\} \) for some \( d \in \mathbb{N} \) and there exists a probability mass function \( P \in \mathcal{P}(\mathcal{X}) \), the noise distribution, such that

\[
W(y|x) = P(y - x)
\]

where the \( - \) in (3) is understood to be modulo \( d \), i.e., the subtraction operation in the additive group \( \{0,1,\ldots,d-1\} \). In other words, \( Y = X + Z \) (mod \( d \)) where the noise \( Z \) has distribution \( P \). The capacity of the additive channel \( W \) is \( C = \log d - H(P) \) and the (unique) capacity-achieving input distribution is the uniform distribution on \( \{0,1,\ldots,d-1\} \).

We consider a channel coding problem in which a message taking values in \( \{1,\ldots,M_n\} \) uniformly at random is to be transmitted across a noisy channel \( W^n \). An encoder \( f : \{1,\ldots,M_n\} \to \mathcal{X}^n \) transforms the message to a codeword. The codebook \( \mathcal{C}_n = \{x_1,\ldots,x_{M_n}\} \) where \( x_m = f(m) \) is the set of all codewords. The channel \( W^n \) then applies a random transformation to the chosen codeword \( x_m \in \mathcal{X}^n \) resulting in \( y \in \mathcal{Y}^n \). A decoder \( d : \mathcal{Y}^n \to \{0,1,\ldots,M_n\} \) either declares a estimate of the message or outputs an erasure symbol, denoted as 0. The decoding operation can thus be regarded as partition of the output space \( \mathcal{Y}^n \) into \( M_n + 1 \) disjoint decoding regions \( \mathcal{D}_0, \mathcal{D}_1, \ldots, \mathcal{D}_{M_n} \subset \mathcal{Y}^n \) where \( \mathcal{D}_m := d^{-1}(m) \). The set of all \( y \in \mathcal{D}_0 \) leads to an erasure event.

Given a codebook \( \mathcal{C}_n \), one can define two undesired error events for \( n \) uses of the DMC. The first is the event in which the decoder does not make the correct decision, i.e., if message \( m \) is sent, it declares either an erasure 0 or outputs an incorrect message \( m' \neq m \) (more precisely, \( m \in \{1,\ldots,M_n\} \setminus \{m\} \)). The probability of this event \( \mathcal{E}_1 \) can be written as

\[
\Pr(\mathcal{E}_1|\mathcal{C}_n) = \frac{1}{M_n} \sum_{m=1}^{M_n} \sum_{y \in \mathcal{D}_m} W^n(y|x_m).
\]

This is the total error probability. The more serious error event is \( \mathcal{E}_2 \), which is defined as the event of declaring an incorrect message, i.e., if \( m \) is sent, the decoder declares that \( m' \neq m \) is sent instead. This undetected error probability can be written as

\[
\Pr(\mathcal{E}_2|\mathcal{C}_n) = \frac{1}{M_n} \sum_{m=1}^{M_n} \sum_{y \in \mathcal{D}_m} \sum_{m' \neq m} W^n(y|x_{m'}).\]

One usually designs the codebook \( \mathcal{C}_n \) and the decoder \( d \) such that \( \Pr(\mathcal{E}_2|\mathcal{C}_n) \) is much smaller than \( \Pr(\mathcal{E}_1|\mathcal{C}_n) \).

III. MAIN RESULTS

A. Direct Results

We now state our main result in this paper concerning the asymmetric evaluation of \( \Pr(\mathcal{E}_1|\mathcal{C}_n) \) and \( \Pr(\mathcal{E}_2|\mathcal{C}_n) \) which correspond to the total error probability and the undetected error probability respectively. Define the conditional information variance of an input distribution \( P_X \) and the channel \( W \) as

\[
V(P_X, W) := \sum_{x \in \mathcal{X}} P_X(x) \sum_{y \in \mathcal{Y}} W(y|x) \left[ \log \frac{W(y|x)}{P_X W(y)} - D(W(\cdot|x)||P_X W) \right]^2,
\]
where \( P_X W(y) = \sum_x P_X(x)W(y|x) \) is the output distribution induced by \( P_X \) and \( W \). We further define the minimum and maximum conditional information variances as

\[
V_{\text{max}}(W) := \max_{P_X \in \Pi} V(P_X, W) \quad \text{and} \quad V_{\text{min}}(W) := \min_{P_X \in \Pi} V(P_X, W).
\]

(7)

(8)

Note that for all \( P_X \in \Pi \), we have \( V(P_X, W) = U(P_X, W) \) [13, Lem. 62], where the unconditional information variance \( U(P_X, W) \) is defined as

\[
U(P_X, W) := \sum_{x \in \mathcal{X}} P_X(x) \sum_{y \in \mathcal{Y}} W(y|x) \left[ \log \left( \frac{W(y|x)}{P_X W(y)} - C \right) \right]^2.
\]

(9)

We assume that the channel \( W \) satisfies \( V_{\text{min}}(W) > 0 \) throughout.

**Theorem 1 (Moderate Deviations Regime Direct).** Let \( 0 < t < 1/2 \) and \( a > b > 0 \). Set the number of codewords \(^1 M_n \) to satisfy

\[
\log M_n = nC - an^{1-t}.
\]

(10)

There exists a sequence of codebooks \( \mathcal{C}_n \) with \( M_n \) codewords such that the two error probabilities satisfy

\[
\lim_{n \to \infty} \frac{1}{n^{1-2t}} \log \Pr(\mathcal{E}_1|\mathcal{C}_n) = \frac{(a - b)^2}{2V_{\text{min}}(W)}, \quad \text{and} \quad \liminf_{n \to \infty} \frac{1}{n^{1-t}} \log \Pr(\mathcal{E}_2|\mathcal{C}_n) \geq b
\]

(11)

(12)

The proof of this result can be found in Section IV-A. Interestingly, we do not analyze the optimal decoding regions prescribed by Forney [1] and described in (30) in the sequel. We consider the following regions \( \{\hat{\mathcal{D}}_m\}_{m=1}^{M_n} \) motivated by information spectrum analysis [18]:

\[
\hat{\mathcal{D}}_m := \left\{ y : \log \frac{W^n(y|x_m)}{(P_X W)^n(y)} \geq \log M_n + bn^{1-t} \right\},
\]

(13)

where \( P_X \) is a capacity-achieving input distribution. We choose \( P_X \) to achieve either \( V_{\text{min}}(W) \) or \( V_{\text{max}}(W) \) in the proofs. Now we define the set of all \( y \in \mathcal{Y}^n \) that leads to an erasure event in terms of \( \{\hat{\mathcal{D}}_m\}_{m=1}^{M_n} \) as

\[
\hat{\mathcal{D}}_0 := \left( \bigcap_{m=1}^{M_n} \hat{\mathcal{D}}_m^c \right) \cup \left( \bigcup_{m \neq m'} (\hat{\mathcal{D}}_m \cap \hat{\mathcal{D}}_{m'}) \right).
\]

(14)

Then, the decoding region for message \( m = 1, \ldots, M_n \) is defined to be

\[
\hat{\mathcal{D}}_m := \hat{\mathcal{D}}_m \setminus \hat{\mathcal{D}}_0.
\]

(15)

The erasure region is \( \hat{\mathcal{D}}_0 \) described in (14). A moment’s of thought reveals that \( \hat{\mathcal{D}}_0, \hat{\mathcal{D}}_1, \ldots, \hat{\mathcal{D}}_{M_n} \) are mutually disjoint and furthermore \( \cup_{m=0}^{M_n} \hat{\mathcal{D}}_m = \mathcal{Y}^n \).

Theorem 1 corresponds to the so-called moderate deviations regime in channel coding considered by Altu˘g and Wagner [14] and Polyanskiy and Verdú [17]. Thus, the appearance of the term \( V_{\text{min}}(W) \) in the results is natural. However, notice that the error probabilities \( \Pr(\mathcal{E}_1|\mathcal{C}_n) \) and \( \Pr(\mathcal{E}_2|\mathcal{C}_n) \) decay asymmetrically. By that, we mean that the rates of decay are different—\( \Pr(\mathcal{E}_1|\mathcal{C}_n) \) decays as \( \exp(-\Theta(n^{1-2t})) \) while \( \Pr(\mathcal{E}_2|\mathcal{C}_n) \) decays as \( \exp(-O(n^{1-t})) \).

When \( t = 1/2 \), we observe different asymptotic scaling from that in Theorem 3. Define

\[
\varphi(w) := \frac{1}{\sqrt{2\pi}} \exp\left( -\frac{w^2}{2} \right)
\]

(16)

\[
\Phi(\alpha) := \int_{-\infty}^{\alpha} \varphi(w) \, dw
\]

(17)

\(^1\)As is usual in information theory, we ignore integer constraints on the number of codewords \( M_n \). We simply set \( M_n \) to the nearest integer to the number satisfying (10).
to be the cumulative distribution function of a standard Gaussian.

**Theorem 2** (Mixed Regime Direct). Let $b > 0$, $a \in \mathbb{R}$, and $M_n$ chosen as in (10) with $t = 1/2$. There exists a sequence of codebooks $C_n$ with $M_n$ codewords such that $\Pr(\mathcal{E}_2|C_n)$ satisfies

$$
\lim_{n \to \infty} \Pr(\mathcal{E}_1|C_n) = \begin{cases} 
\Phi \left( \frac{b-a}{\sqrt{V_{\min}(W)}} \right) & \text{if } a \leq 0, \\
\Phi \left( \frac{b-a}{\sqrt{V_{\max}(W)}} \right) & \text{if } a > 0,
\end{cases}
$$

and

$$
\lim_{n \to \infty} -\frac{1}{\sqrt{n}} \log \Pr(\mathcal{E}_2|C_n) \geq b
$$

The proof of this result can be found in Section IV-B. Observe that the first error probability is in the central limit regime [11]-[13] while the second scales as $\exp(-\sqrt{n} b)$, which is in the moderate deviations regime [14], [17]. Thus, we call this the mixed regime.

**B. Ensemble Converse Results**

It is, at this point, not clear that the codes we proposed in Section III-A are optimal. In this section, we demonstrate the tightness of our code for additive DMCs. We consider an ensemble evaluation of the two error probabilities. Similarly to (10), the sizes of the codes we consider $\{M_n\}_{n \in \mathbb{N}}$ take the form

$$
\log M_n = nC - an^{1-t}
$$

where $C = \log d - H(P)$ is the capacity of the additive channel and $0 < t \leq 1/2$. When $t < 1/2$ (resp. $t = 1/2$), the code size is in the moderate deviations (resp. central limit or mixed) regime.

We now state our main results in this paper concerning the asymmetric evaluation of $\Pr(\mathcal{E}_1|C_n)$ and $\Pr(\mathcal{E}_2|C_n)$ corresponding to the total error probability and the undetected error probability respectively. We define the varentropy [24] or source dispersion of the additive noise $P$ as

$$
V(P) := \sum_{z=0}^{d-1} P(z) \left[ \log \frac{1}{P(z)} - H(P) \right]^2.
$$

This is simply the variance of the self-information random variable $-\log P(Z)$ where $Z$ is distributed as $P$. We assume that $V(P) > 0$ throughout. It is easy to see that because of the additivity of the channel, the $\epsilon$-dispersion [13] of $W$ is $V(P)$ for every $\epsilon \in (0, 1)$, i.e., $V_{\min}(W) = V_{\max}(W) = V(P)$.

**Theorem 3** (Moderate Deviations Regime Converse). Let $0 < t < 1/2$ and $a > b > 0$. Consider a sequence of random codebooks $C_n$ with $M_n$ codewords where each codeword is drawn uniformly at random from $\{0, 1, \ldots, d-1\}^n$ and $M_n$ satisfies (20). When the expectation of the total error satisfies

$$
\lim_{n \to \infty} -\frac{1}{n^{1-2t}} \log \mathbb{E}_{C_n} \left[ \Pr(\mathcal{E}_1|C_n) \right] \geq \frac{(a-b)^2}{2V(P)},
$$

then the expectation of the undetected error satisfies

$$
\lim_{n \to \infty} -\frac{1}{n^{1-t}} \log \mathbb{E}_{C_n} \left[ \Pr(\mathcal{E}_2|C_n) \right] \leq b.
$$

Conversely, when the expectation of the undetected error satisfies

$$
\lim_{n \to \infty} -\frac{1}{n^{1-t}} \log \mathbb{E}_{C_n} \left[ \Pr(\mathcal{E}_2|C_n) \right] \geq b,
$$

then the expectation of the total error satisfies

$$
\lim_{n \to \infty} -\frac{1}{n^{1-2t}} \log \mathbb{E}_{C_n} \left[ \Pr(\mathcal{E}_1|C_n) \right] \leq \frac{(a-b)^2}{2V(P)}.
$$

**Theorem 4** (Mixed Regime Converse). Let $b > 0$, $a \in \mathbb{R}$ and $M_n$ chosen according to (20) with $t = 1/2$. Consider a sequence of random codebooks $C_n$ with $M_n$ codewords where each codeword is drawn uniformly at random from
\{0, 1, \ldots, d - 1\}_n$, the decoding regions are chosen according to (30) with thresholds (32). When the expectation of the total error satisfies
\[
\limsup_{n \to \infty} \mathbb{E}_{\mathcal{C}_n} [\Pr(\mathcal{E}_1|\mathcal{C}_n)] \leq \Phi \left( \frac{b - a}{\sqrt{2V(P)}} \right)
\]  
then the expectation of the undetected error satisfies
\[
\limsup_{n \to \infty} -\frac{1}{\sqrt{n}} \log \mathbb{E}_{\mathcal{C}_n} [\Pr(\mathcal{E}_2|\mathcal{C}_n)] \leq b.
\]  
Conversely, when the expectation of the undetected error satisfies
\[
\liminf_{n \to \infty} -\frac{1}{\sqrt{n}} \log \mathbb{E}_{\mathcal{C}_n} [\Pr(\mathcal{E}_2|\mathcal{C}_n)] \geq b,
\]  
then the expectation of the total error satisfies
\[
\liminf_{n \to \infty} \mathbb{E}_{\mathcal{C}_n} [\Pr(\mathcal{E}_1|\mathcal{C}_n)] \geq \Phi \left( \frac{b - a}{\sqrt{2V(P)}} \right).
\]  
These theorems imply that if we generate our encoder according to the uniform distribution even if we improve our decoder, we cannot improve both errors. That is, these theorems show the optimality of our codes for the additive channel. The proofs of these theorems follow immediately from Lemmas 5 and 6 to follow.

To prove these theorems we need to develop Lemmas 5 and 6 in the following. We recall Forney’s result in [1] that for a given codebook \( \mathcal{C}_n := \{x_1, \ldots, x_{M_n}\} \), the optimal decoding region for each message \( m \in \{1, \ldots, M_n\} \) is given by
\[
\mathcal{D}_m := \left\{ y : \frac{W_n(y|x_m)}{\sum_{m' \neq m} W_n(y|x_{m'})} \geq \exp(nT_n) \right\},
\]  
where \( T_n > 0 \) is a threshold parameter that serves to trade off between the two error probabilities \( \Pr(\mathcal{E}_1|\mathcal{C}_n) \) and \( \Pr(\mathcal{E}_2|\mathcal{C}_n) \). This is simply generalization of the Neyman-Pearson lemma. Because \( T_n > 0 \), the regions are disjoint. We let \( \mathcal{D}_0 \) denote the set of all \( y \) that leads to an erasure, i.e.,
\[
\mathcal{D}_0 := y^n \setminus \bigsqcup_{m=1}^{M_n} \mathcal{D}_m.
\]  
In the literature on decoding with an erasure option (e.g., [1]–[8]), \( T_n \) is usually kept at a constant (not depending on \( n \)), leading to results concerning tradeoffs between the exponential decay rates of \( \Pr(\mathcal{E}_1|\mathcal{C}_n) \) and \( \Pr(\mathcal{E}_2|\mathcal{C}_n) \), i.e., the error exponents of the total and undetected error probabilities. Our treatment is different. We let \( T_n \) in the definitions of the decision regions \( \mathcal{D}_m \) in (30) depend on \( n \) and show that the error probabilities \( \Pr(\mathcal{E}_1|\mathcal{C}_n) \) and \( \Pr(\mathcal{E}_2|\mathcal{C}_n) \) decay subexponentially and in an asymmetric manner, i.e., at different speeds.

**Lemma 5** (Moderate Deviations Regime Ensemble). Let \( 0 < t < 1/2 \) and \( a > b > 0 \). Consider a sequence of random codebooks \( \mathcal{C}_n \) with \( M_n \) codewords where each codeword is drawn uniformly at random from \( \{0, 1, \ldots, d - 1\}_n \) and \( M_n \) satisfies (20). Let the decoding regions be chosen as in (30) with thresholds
\[
T_n := \frac{b}{n^t},
\]  
Then the expectation of the two error probabilities satisfy
\[
\lim_{n \to \infty} -\frac{1}{n^{1-2t}} \log \mathbb{E}_{\mathcal{C}_n} [\Pr(\mathcal{E}_1|\mathcal{C}_n)] = \frac{(a - b)^2}{2V(P)}; \quad \text{and}
\]  
\[
bn^{1-t} + \frac{(a - b)^2}{2V(P)} n^{1-2t} + o(n^{1-2t}) \leq -\log \mathbb{E}_{\mathcal{C}_n} [\Pr(\mathcal{E}_2|\mathcal{C}_n)] \leq bn^{1-t} + o(n^{1-t})
\]  
The proof of this lemma is provided in Section IV-C. At this point, a few comments concerning this lemma are in order:

Since the decoder given in (30) is optimal, Theorem 3 is proven using Lemma 5 as follows. If \( T_n = b'n^{-t} + o(n^{-t}) \) with \( b' < b \), (24) does not hold. This fact can be shown by applying Lemma 5 (and, in particular, (34)) to the case
b := b'. So, to satisfy (24), we need to choose $T_n \geq bn^{-t} + o(n^{-t})$. Applying Lemma 5 to this case, we obtain (25). That is, (24) implies (25). Conversely, to satisfy (22), we need to choose $T_n \leq bn^{-t} + o(n^{-t})$. Hence, due to Lemma 5, we have (23). That is, (22) implies (23).

This result again corresponds to the so-called moderate deviations regime in channel coding considered by Altuğ and Wagner [14] and Polyanskiy and Verdú [17]. Thus, the appearance of the varentropy term $V(P)$ in the results is very natural. The total and undetected error probabilities in (33) and (34) can be written as

\[
\mathbb{E}_{c_n} \left[ \Pr(\mathcal{E}_1|c_n) \right] \approx \exp \left( -\frac{(a-b)^2}{2V(P)} n^{1-2t} \right), \quad \text{and}
\]

\[
\mathbb{E}_{c_n} \left[ \Pr(\mathcal{E}_2|c_n) \right] \approx \exp \left( -bn^{1-t} \right)
\]

respectively. This scaling is also different from those found in the literature which primarily focus on exponentially decaying probabilities [1]–[8] or constant non-zero errors [16]. Both our total and undetected error probabilities are designed to decay subexponentially fast in the blocklength $n$. Our proof technique involves estimating appropriately-defined cumulant generating functions and invoking a modified version of the Gärtner-Ellis theorem [20, Sec. 2.3] (cf. Theorem 8 in Appendix A). Similarly to the work by Somekh-Baruch and Merhav [8], the two probabilities in (33)–(34) are asymptotic equalities (if we consider the normalization $n^{1-2t}$ and $n^{1-t}$ rather than inequalities [cf. [1], [6]]). In fact for the lower bound in (34), we can even calculate a higher-order asymptotic term scaling as $n^{1-2t}$ (but unfortunately, we do not yet have a matching upper bound).

Next, observe that the undetected error decays much faster than the total error as expected because the former is much more undesirable than an erasure. If $a$ is increased for fixed $b$, the effective number of codewords is decreased so commensurately, the total error probability $\Pr(\mathcal{E}_1|c_n)$ is also reduced. Also, if $b$ is increased (tending towards $a$ from below), the probability of an erasure increases and so the probability of an undetected error decreases. This is evident in (35) where the coefficient $(a-b)^2/(2V(P))$ decreases and in (36) where the leading coefficient $b$ increases. Thus, we observe a delicate interplay between $a$ governing the code size and $b$, the parameter in the threshold.

Finally, if $b$ is negative (a case not allowed by Lemma 5), so is $T_n$. This corresponds to list decoding [1] where the decoder is allowed to output more than one message (i.e., a list of messages) and an error event occurs if and only if the transmitted message is not in the list. In this case, $\Pr(\mathcal{E}_2|c_n)$ no longer corresponds to the probability of undetected error. Rather, the expression for $\Pr(\mathcal{E}_2|c_n)$ in (5) corresponds to the average number of incorrect codewords in the list corresponding to the overlapping (non-disjoint) decision regions $\{D^{m}_{m=1}\}$.

**Lemma 6 (Mixed Regime Ensemble).** Let $b > 0$, $a \in \mathbb{R}$ and $M_n$ chosen according to (20) with $t = 1/2$. Consider a sequence of random codebooks $C_n$ with $M_n$ codewords where each codeword is drawn uniformly at random from $\{0, 1, \ldots, d-1\}^n$. The decoding regions are chosen according to (30) with thresholds (32). Then, the expectation of the two error probabilities satisfy

\[
\lim_{n \rightarrow \infty} \mathbb{E}_{c_n} \left[ \Pr(\mathcal{E}_1|c_n) \right] = \Phi \left( \frac{b-a}{\sqrt{V(P)}} \right)
\]

and

\[
b\sqrt{n} + \frac{(a-b)^2}{2V(P)} + o(1) \leq - \log \mathbb{E}_{c_n} \left[ \Pr(\mathcal{E}_2|c_n) \right] \leq b\sqrt{n} + o(\sqrt{n}).
\]

The proof of this lemma is provided in Section IV-D. It is largely similar to that for Lemma 5 but for the total error probability in (37), instead of invoking the Gärtner-Ellis theorem [20, Sec. 2.3], we use the fact that if the cumulant generating function of a sequence of random variables $\{K_n\}_{n \in \mathbb{N}}$ converges in distribution to a Gaussian random variable. However, this is not completely straightforward as we can only prove that the cumulant generating function converges pointwise for positive parameters (cf. Lemma 7). We thus need to invoke a result by Mukherjee et al. [25, Thm. 2] (building on initial work by Curtiss [26]) to assert weak convergence. (See Lemma 9 in Appendix B.) The asymptotic bounds in (38) are proved using a modified version of the Gärtner-Ellis theorem.

Since the decoder given in (30) is optimal, Theorem 4 is proven using Lemma 6 as follows. If $T_n = b'n^{-1/2} + o(n^{-1/2})$ with $b' < b$, (28) does not hold. This fact can be shown by applying Lemma 6 (and, in particular, (38)) to the case $b := b'$. So, to satisfy (28), we need to choose $T_n \geq bn^{-1/2} + o(n^{-1/2})$. Applying Lemma 6 to this case,
we obtain (29). That is, (28) implies (29). Conversely, to satisfy (26), we need to choose $T_n \leq bn^{-1/2} + o(n^{-1/2})$. Hence, due to Lemma 6, we have (27). That is, (26) implies (27).

Here, ignoring the constant term in the lower bound, the undetected error probability in (38) decays as

$$\mathbb{E}_{C_n} \left[ \Pr(\mathcal{E}_2|C_n) \right] \approx \exp \left( -b\sqrt{n} \right).$$

(39)

The total (and hence, erasure) error probability in (37) is asymptotically a constant depending on the varentropy of the noise distribution $P$, the threshold parametrized by $b$ and the code size parametrized by $a$. Similarly to Lemma 5, if $b$ increases for fixed $a$, the likelihood of an erasure event occurring also increases but this decreases the undetected error probability as evidenced by (38). The situation in which $b \downarrow 0$ for fixed $a$ recovers a special case of a recent result by Tan and Moulin [16, Thm. 1] where the total error probability is kept constant at a positive constant and the undetected error probability vanishes. Note that for this result, we do not require that $a > b$ unlike what we assumed for the pure moderate deviations setting of Lemma 5.

IV. PROOFS OF THE MAIN RESULTS

A. Proof of Theorem 1

Choose any input distribution $P_X \in \Pi(W)$ achieving $V_{\min}(W)$ in (8). We consider choosing each codeword $x_m, m \in \{1, \ldots, M_n\}$ with the product distribution $P_X^n \in \mathcal{P}(X^n)$. The expectation over this random choice of codebook is denoted as $\mathbb{E}_{C_n}[.]$. Now, we first consider $\Pr(\mathcal{E}_1|C_n)$. Define the (capacity-achieving) output distribution $P_Y := P_X W$ and its $n$-fold memoryless extension $P_Y^n$. Next, we consider regions $\mathcal{D}_m$ defined in (13). The expectation over the code of the $W^n(\cdot|X^n_m)$-probability of $\mathcal{D}_m$ can be evaluated as

$$\mathbb{E}_{C_n} \left[ \sum_y W^n(y|X^n_m) \mathbf{1} \left\{ W^n(y|X^n_m) \geq M_n \exp(bn^{1-t})P_Y^n(y) \right\} \right]$$

$$= \mathbb{E}_{X_m^n} \left[ \sum_y P^n_Y(y) \mathbf{1} \left\{ W^n(y|X^n_m) \geq M_n \exp(bn^{1-t})P_Y^n(y) \right\} \right]$$

$$\leq \mathbb{E}_{X_m^n} \left[ \sum_y M_n^{-1} \exp(-bn^{1-t})W^n(y|X^n_m) \mathbf{1} \left\{ W^n(y|X^n_m) \geq M_n \exp(bn^{1-t})P_Y^n(y) \right\} \right]$$

$$\leq M_n^{-1} \exp(-bn^{1-t})$$

(40)

for $m' \neq m$, where (40) is because of independence of codeword generation and $\mathbb{E}_{X_m^n}[W^n(y|X^n_m)] = P_Y^n(y)$. Since $\mathcal{D}_m \subset \hat{\mathcal{D}}_m$, by the definition of $\hat{\mathcal{D}}_m$ in (13), the expectation of the undetected error probability over the random codebook can be written as

$$\mathbb{E}_{C_n} \left[ \Pr(\mathcal{E}_2|C_n) \right] \leq \mathbb{E}_{C_n} \left[ \frac{1}{M_n} \sum_{m=1}^{M_n} \sum_y \sum_{m' \neq m} W^n(y|X^n_m) \mathbf{1} \left\{ W^n(y|X^n_m) \geq M_n \exp(bn^{1-t})P_Y^n(y) \right\} \right]$$

$$\leq \frac{1}{M_n} \sum_{m=1}^{M_n} \sum_y \sum_{m' \neq m} M_n^{-1} \exp(-bn^{1-t})$$

$$= \frac{M_n - 1}{M_n} \exp(-bn^{1-t})$$

$$\leq \exp(-bn^{1-t}).$$

(43)

(44)

(45)

(46)

Hence, this bound verifies (12).

By the definition of $\hat{\mathcal{D}}_m$ for $m = 0, 1, \ldots, M_n$ in (13) and (14), we know that

$$\hat{\mathcal{D}}_m^c = \hat{\mathcal{D}}_m \cup \hat{\mathcal{D}}_0 = \hat{\mathcal{D}}_m \cup \bigcup_{m' \neq m} \left( \hat{\mathcal{D}}_m \cap \hat{\mathcal{D}}_{m'} \right) \subset \hat{\mathcal{D}}_m \cup \bigcup_{m' \neq m} \hat{\mathcal{D}}_{m'}.$$
The expectation of the total error probability over the random codebook can be written as

\[
\mathbb{E}_C \left[ \Pr(\mathcal{E}_1 | C_n) \right] \\
= \mathbb{E}_C \left[ \frac{1}{M_n} \sum_{m=1}^{M_n} \sum_{y \in \tilde{D}_m^c} W^n(y | X_m^n) \right] \\
\leq \mathbb{E}_C \left[ \frac{1}{M_n} \sum_{m=1}^{M_n} \sum_{y} W^n(y | X_m^n) 1 \{ y \in \tilde{D}_m^c \} \right] + \mathbb{E}_C \left[ \frac{1}{M_n} \sum_{m=1}^{M_n} \sum_{y \neq m} W^n(y | X_m^n) 1 \{ y \in \tilde{D}_{m'} \} \right] \\
\leq \mathbb{E}_C \left[ \frac{1}{M_n} \sum_{m=1}^{M_n} \sum_{y} W^n(y | X_m^n) 1 \{ y \in \tilde{D}_m^c \} \right] + \exp(-bn^{1-t}), \\
= \sum_{x,y} P^n_X(x) W^n(y|x) 1 \left\{ \log \frac{W^n(y|x)}{P^n_Y(y)} - nC < -(a-b)n^{1-t} \right\} + \exp(-bn^{1-t}),
\]

where (49) follows from (47), (50) follows from similar calculations that led to (46), and (51) follows from the definition of \( \tilde{D}_m \) and the choice of \( M_n \) in (10). In fact, by using the bound \( \tilde{D}_m^c \supset \tilde{D}_m^c \) from the first equality in (47), we see that the upper bound on \( \mathbb{E}_C \left[ \Pr(\mathcal{E}_1 | C_n) \right] \) in (51) is tight in the sense that it can also be lower bounded as

\[
\mathbb{E}_C \left[ \Pr(\mathcal{E}_1 | C_n) \right] \geq \sum_{x,y} P^n_X(x) W^n(y|x) 1 \left\{ \log \frac{W^n(y|x)}{P^n_Y(y)} - nC < -(a-b)n^{1-t} \right\}.
\]

Recall that \( a > b \). By the moderate deviations theorem [20, Thm. 3.7.1], the sums on the right-hand-sides of (51) and (52) behave as

\[
\exp \left( -n^{1-2t} \frac{(a-b)^2}{2U(P_X, W)} + o(n^{1-2t}) \right),
\]

which is much larger than (i.e., dominates) the second term in (51), namely \( \exp(-bn^{1-t}) \). Since \( U(P_X, W) = V_{\min}(W) \) [13, Lem. 62], we have the asymptotic equality in (11).

Finally, by employing a standard Markov inequality argument to (46) and (53) to derandomize the code (e.g., see the proof of [16, Thm. 1]), we see that there exist a sequence of deterministic codes \( C_n \) satisfying the conditions of the theorem.

**B. Proof of Theorem 2**

In this case, \( t = 1/2 \). We first consider the case where \( a \leq 0 \). Choose \( P_X \) that achieves \( V_{\max}(W) \). In this case by the Berry-Esseen theorem [27, Sec. XVI.7], the right-hand-sides of (51) and (52) behave as

\[
\Phi \left( \frac{b-a}{\sqrt{V_{\max}(W)}} \right) + O \left( \frac{1}{n} \right).
\]

Thus, by a standard Markov inequality argument to derandomize the code, for any sequence \( \{ \theta_n \}_{n \in \mathbb{N}} \subset (0,1) \), there exists a sequence of deterministic codes \( C_n \) satisfying \( \Pr(\mathcal{E}_1 | C_n) \approx (1-\theta_n)^{-1} \Phi \left( \frac{b-a}{\sqrt{V_{\max}(W)}} \right) \) and \( \Pr(\mathcal{E}_2 | C_n) \approx \theta_n^{-1} \exp(-\sqrt{nb}) \). Choose \( \theta_n := 1/n \) to complete the proof of the theorem for \( a \leq 0 \). For \( a > 0 \), choose the input distribution \( P_X \) to achieve \( V_{\min}(W) \) and proceed in exactly the same way.

**C. Proof of Lemma 5**

Proof: We consider choosing each codeword \( x_m, m \in \{1, \ldots, M_n \} \) uniformly at random from \( \{0,1, \ldots, d-1\}^n \). Indeed, the capacity-achieving input distribution of the additive channel is uniform on \( \{0,1, \ldots, d-1\} \). As above, the expectation over this random choice of codebook is denoted as \( \mathbb{E}_C[\cdot] \). Now, we first consider \( \Pr(\mathcal{E}_1 | C_n) \).
From the definition in (4), the expectation of the error probability over the random codebook can be written as
\[
\mathbb{E}_{C_n} \left[ \Pr(\mathcal{E}_1|C_n) \right] = \mathbb{E}_C \left[ \frac{1}{M_n} \sum_{m=1}^{M_n} \sum_{y} \mathbb{W}^n(y|X^n_m) \mathbf{1} \left\{ \frac{\sum_{m' \neq m} \mathbb{W}^n(y|X^n_{m'})}{\mathbb{W}^n(y|X^n_m)} \geq \exp(-nT_n) \right\} \right] \quad (55)
\]
\[
= \mathbb{E}_C \left[ \frac{1}{M_n} \sum_{m=1}^{M_n} \Pr \left( \log \left( \sum_{m' \neq m} \mathbb{W}^n(Y^n|X^n_{m'}) \right) - \log \mathbb{W}^n(Y^n|X^n_m) \geq -nT_n \left| C_n \right. \right) \right] \quad (56)
\]
\[
= \frac{1}{M_n} \sum_{m=1}^{M_n} \Pr \left( \log \left( \sum_{m' \neq m} \mathbb{W}^n(Y^n|X^n_{m'}) \right) - \log \mathbb{W}^n(Y^n|X^n_m) \geq -bn^{1-t} \right) \quad (57)
\]

In (56), the inner probability is over $Y^n \sim \mathbb{W}^n(\cdot|x_m)$ for a fixed code $C_n$ and in (57), the probability is over both the random codebook $C_n$ and the channel output $Y^n$ given message $m$ was sent. By symmetry of the codebook generation, it is sufficient to study the behavior of the random variable

\[
F_n := \log \left( \sum_{m' \neq m} \mathbb{W}^n(Y^n|X^n_{m'}) \right) - \log \mathbb{W}^n(Y^n|X^n_m)
\]

for any $m \in \{1, \ldots, M_n\}$, say $m = 1$. In particular, to estimate the probability $\Pr(F_n \geq 0)$ in (57), it suffices to estimate the cumulant generating function of $F_n$. We denote the cumulant generating function as

\[
\phi_n(s) := \log \mathbb{E}[\exp(sF_n)]
\]

\[
= \log \mathbb{E}_C \left[ \sum_{y} \mathbb{W}^n(y|X^n_1)^{1-s} \left( \sum_{m' \neq m} \mathbb{W}^n(y|X^n_{m'}) \right)^s \right] \quad (59)
\]

\[
= \log \sum_{y} \mathbb{E}_C \left[ \mathbb{W}^n(y|X^n_1)^{1-s} \cdot \mathbb{E}_C \left[ \left( \sum_{m' \neq m} \mathbb{W}^n(y|X^n_{m'}) \right)^s \right] \right]. \quad (60)
\]

The final equality follows from the independence in the codeword generation procedure. We have the following important lemma which is proved in Section IV-E.

**Lemma 7** (Asymptotics of Cumulant Generating Functions). Fix $t \in (0, 1/2]$. Given the condition on the code size in (20), the cumulant generating function satisfies

\[
\phi_n \left( \frac{u}{n^t} \right) = \left( -au + u^2 \frac{V(P)}{2} \right) n^{1-2t} + O(n^{1-3t}) + o(1)
\]

for any constant $u > 0$.

Now, we apply the Gärtner-Ellis theorem with the general order, i.e., Case (ii) of Theorem 8 in Appendix A with the identifications $\alpha_n \equiv 0$, $\beta_n \equiv n^{1-t}$, and $\gamma_n \equiv n^{-t}$. Now, we can also make the additional identifications $X_n \equiv -F_n$, $p_n(\cdot) \equiv \frac{1}{M_n} \sum_{m=1}^{M_n} \Pr(\cdot)$, $\mu_n(\cdot) \equiv \phi_n(\cdot)$, $\theta_0 \equiv 0$, $\nu_1 \equiv 0$, $y \equiv s/\gamma_n$, $x \equiv b$. Thus, $\nu_2(y) = \lim_{n \to \infty} \nu_2,n(y) = -ya + y^2V(P)/2$ according to Lemma 7. The conditions of Case (ii) Theorem 8 are satisfied so we can readily apply it here, thus,

\[
- \log \mathbb{E}_C \left[ \Pr(\mathcal{E}_1|C_n) \right] = - \log \Pr \left( F_n > -bn^{1-t} \right) = \frac{(a-b)^2}{2V(P)} n^{1-2t} + o(n^{1-2t}),
\]

which implies (33).

Now we estimate $\mathbb{E}_{C_n}[\Pr(\mathcal{E}_2|C_n)]$. Using the same calculations that led to (57), one finds that

\[
\mathbb{E}_{C_n} \left[ \Pr(\mathcal{E}_2|C_n) \right] = \mathbb{E}_C \left[ \frac{1}{M_n} \sum_{m=1}^{M_n} \sum_{y} \sum_{m' \neq m} \mathbb{W}^n(y|X^n_m) \mathbf{1} \left\{ \frac{\sum_{m' \neq m} \mathbb{W}^n(y|X^n_{m'})}{\mathbb{W}^n(y|X^n_m)} < \exp(-nT_n) \right\} \right] \quad (64)
\]

\[
= \mathbb{E}_C \left[ \frac{1}{M_n} \sum_{m=1}^{M_n} \mathbb{Q} \left( \left\{ y : \log \sum_{m' \neq m} \mathbb{W}^n(y|X^n_{m'}) - \log \mathbb{W}^n(y|X^n_m) < -bn^{1-t} \right\} \left| C_n \right. \right) \right] \quad (65)
\]
where in (65), we defined the (unnormalized) conditional measure $Q(\mathcal{A}|\mathcal{C}_n = \{x_m\}_{m=1}^{M_n}) := \sum_{m' \neq m} W^n(\mathcal{A}|x_{m'})$ where $\mathcal{A} \subset \mathcal{Y}^n$. Given $Q$, we can define a normalized probability measure $Q' := Q/(M_n - 1)$. Since the form of (65) is similar to the starting point for the calculation of $\mathbb{E}_{\mathcal{C}_n} [\text{Pr}(\mathcal{E}_2|\mathcal{C}_n)]$ in (57), we may estimate $\mathbb{E}_{\mathcal{C}_n} [\text{Pr}(\mathcal{E}_2|\mathcal{C}_n)]$ using similar steps to the above. Define another probability measure $P(\mathcal{A}|\mathcal{C}_n = \{x_m\}_{m=1}^{M_n}) := W^n(\mathcal{A}|x_m)$. Note by the definition of $F_n$ in (58), and the measures above that for all $\mathcal{A} \subset \mathcal{Y}^n$,  

\[
\exp(F_n) = \frac{Q'(\mathcal{A}|\mathcal{C}_n)}{P(\mathcal{A}|\mathcal{C}_n)} \cdot (M_n - 1).
\]

Observe that the random variable involved in (65), namely $\log \sum_{m' \neq m} W^n(Y^n|X^n_m) - \log W^n(Y^n|X^n_m)$, is exactly $F_n$ defined in (58) where $Y^n$ now has conditional law $Q(\cdot|\mathcal{C}_n)$ instead of $P(\cdot|\mathcal{C}_n)$. The cumulant generating function of $F_n$ under the probability measure $Q'$ is 

\[
\lambda_n(s) := \log \mathbb{E}_{\mathcal{C}_n,Q'}[\exp(sF_n)] = \log \left( \frac{\mathbb{E}_{\mathcal{C}_n,P}[\exp((1+s)F_n)]}{M_n-1} \right) = \phi_n(1+s) - \log(M_n-1)
\]

where (68) follows from (66) and (69) from the definition of $\phi_n(s)$ in (59). Now, we apply Case (i) of Theorem 8 in Appendix A with the identifications $\alpha_n \equiv 0$, $\beta_n \equiv n^{-t}$, and $\gamma_n \equiv n^{-t}$. Furthermore, from (65) and (69), one can also make the additional identifications $X_n = F_n$, $p_n(\cdot) = \mathbb{E}_{\mathcal{C}_n}[Q(\cdot|\mathcal{C}_n)]$, $\mu_n(\cdot) = \lambda_n(\cdot)$, $\theta_0 \equiv -1$, $\nu_1 \equiv 0$, $y \equiv s/\gamma_n$, $x \equiv -b$, and $\nu_2(y) \equiv n^{2t-1} \phi_n(y/n^t)$. Thus, $\nu_2(y) = \lim_{n \to \infty} \nu_{2,n}(y) = -ya + y^2 V(P)/2 =: \nu_2(y)$ according to Lemma 7. Thus, by relating $Q$ to $Q'$ and using (113) in Case (i) of Theorem 8, we obtain 

\[
-\log \mathbb{E}_{\mathcal{C}_n}[\text{Pr}(\mathcal{E}_2|\mathcal{C}_n)] = -\log \mathbb{E}_{\mathcal{C}_n}[Q(F_n < -bn^{1-t}|\mathcal{C}_n)] = -\log \mathbb{E}_{\mathcal{C}_n}[Q'(F_n < -bn^{1-t}|\mathcal{C}_n)] - \log(M_n-1) \leq bn^{1-t} + o(n^{1-t}),
\]

which implies the upper bound in (34). The lower bound in (34) follows by invoking (112), from which we obtain 

\[
-\log \mathbb{E}_{\mathcal{C}_n}[\text{Pr}(\mathcal{E}_2|\mathcal{C}_n)] \geq bn^{1-t} + \frac{(a-b)^2}{2V(P)} n^{1-2t} + o(n^{1-2t}).
\]

This completes the proof of Lemma 5.

**Remark 1.** Observe that to evaluate the probabilities in (63) and (70), we employed Theorem 8 in Appendix A, which is a modified ("shifted") version of the usual Gärtner-Ellis theorem [20, Sec. 2.3]. Theorem 8 assumes a sequence of random variables $X_n$ has cumulant generating functions $\mu_n(\theta)$ that additionally satisfy the expansion $\mu_n(\theta_0 + \gamma_n y) = \alpha_n + \beta_n \nu_1 + \beta_n \gamma_n \nu_2(y)$ for some vanishing sequence $\gamma_n$. This generalization and the application to the erasure problem appears to the authors to be novel. In particular, since $Q$ in (65) above is not a (normalized) probability measure, the usual Gärtner-Ellis theorem does not apply readily and we have to define the new probability measure $Q'$. This, however, is not the crux of the contributions of which there are three.

1) First, our Theorem 8 also has to take into account the offsets $\theta_0 = -1$ and $\alpha_n = -\log(M_n-1)$ in our application of the Gärtner-Ellis theorem.

2) Second, an interesting feature of our result is that the “exponent” $b$ is not governed by the first-order term $\alpha_n$ (which is the offset) but instead the second-order term $\beta_n \nu_1 = -bn^{1-t}$ leading to (72)–(73).

3) Finally, Theorem 8 also allows us to obtain an additional term scaling as $n^{1-2t}$ in (73).

**D. Proof of Lemma 6**

**Proof:** The exact same steps in the proof of Lemma 5 follow even if $t = 1/2$. In particular, in this setting, Lemma 7 with $t = 1/2$ yields 

\[
\lim_{n \to \infty} \phi_n \left( \frac{u}{\sqrt{n}} \right) = -ua + u^2 V(P) / 2
\]

(74)
for any constant $u > 0$. By appropriate translation, scaling, and Lemma 9 in Appendix B, the sequence of random variables $\{F_n n^{-1/2}\}_{n \in \mathbb{N}}$ converges in distribution to a Gaussian random variable with mean $-a$ and variance $V(P)$. This implies that the following asymptotic statement holds true

$$\lim_{n \to \infty} E_{C_n} [\Pr(E_1 | C_n)] = \lim_{n \to \infty} \Pr \left( \frac{F_n}{\sqrt{n}} > -b \right)$$

$$= \int_{-b}^{\infty} \frac{1}{\sqrt{2\pi V(P)}} \exp \left( -\frac{(w + a)^2}{2 V(P)} \right) \, dw$$

$$= \Phi \left( \frac{b - a}{\sqrt{V(P)}} \right).$$

To calculate $E_{C_n} [\Pr(E_2 | C_n)]$, we can adopt the same change of measure and Gártner-Ellis arguments (Case (i) of Theorem 8) as in the steps leading from (64) to (73) to assert that (38) is true. Note that in this situation, we take $\gamma_n \equiv n^{-1/2}$ and $\beta_n \equiv 1/2$.

E. Proof of Lemma 7: Asymptotics of Cumulant Generating Functions

Proof: To estimate $\phi_n(s)$ in (61), we define

$$A := E_{C_n} \left[ W^n(y | X_n^1)^{1-s} \right] \quad \text{and}$$

$$B := E_{C_n} \left[ \left( \sum_{m' \neq m} W^n(y | X_m^{n}) \right)^s \right].$$

The first term $A$ is easy to handle. Indeed, by the additivity of the channel, we have

$$A = E_{X_m^n} \left[ P^n(y - X_m^1)^{1-s} \right]$$

$$= E_{X_m^n} \left[ P^n(X_m^1)^{1-s} \right]$$

where the shifted codewords are defined as $X_m^n := y - X_m^1$. By using the product structure of $P^n$, we see that regardless of $y$, the term $A$ can be written as

$$A = \frac{1}{d^n} \exp \left( -n \psi(s) \right)$$

where

$$\psi(s) := -\log \sum_z P(z)^{1-s}.$$  (83)

This function is related to the Rényi entropy as follows: $s \psi(s) = -H_{1-s}(P)$ where $H_{1-s}(P)$ is the usual Rényi entropy of order $\alpha$ (e.g., [23, Prob. 1.15]). Now, for a fixed $u > 0$, we make the choice

$$s = \frac{u}{n^t},$$

where recall that $t$ is a fixed parameter in $(0, 1/2]$. It is straightforward to check that $\psi(0) = 0$, $\psi'(0) = -H(P)$ and $\psi''(0) = -V(P)$. By a second-order Taylor expansion of $\psi(s)$ around $s = 0$, we have

$$A = \frac{1}{d^n} \exp \left( n \left( s H(P) + s^2 \frac{V(P)}{2} + O(s^3) \right) \right)$$

$$= \frac{1}{d^n} \exp \left( un^{1-t} H(P) + u^2 n^{1-2t} \frac{V(P)}{2} + O(n^{1-3t}) \right),$$

where (86) follows from the definition of $s$ in (84).

Now we estimate $B$ in (79). Define the random variable $N_{C_n}(Q)$ which represents the number of shifted codewords excluding that indexed by $m$ with type $Q \in \mathcal{P}_n(X)$, i.e., $N_{C_n}(Q) := |\{m' \neq m : \text{type}(X_m^{n}) = Q\}|$. This plays the
role of the type class enumerator or distance enumerator in Merhav [6], [9]. Then, \( B \) can be written as

\[
B = \mathbb{E}_{C_n}\left[ \left( \sum_{m' \neq m} P^n(y - X^n_{m'}) \right)^s \right] \quad (87)
\]

\[
= \mathbb{E}_{C_n}\left[ \left( \sum_{m' \neq m} P^n(\tilde{X}^n_{m'}) \right)^s \right] \quad (88)
\]

\[
= \mathbb{E}_{C_n}\left[ \left( \sum_{Q \in \mathcal{P}_n(X)} N_{C_n}(Q) \exp \left( - n[D(Q\|P) + H(Q)] \right) \right)^s \right]. \quad (89)
\]

In (87), we again used the additivity of the channel and introduced the noise distribution \( P \). In (88), we used the definition of the shifted codewords \( \tilde{X}^n_{m'} \). In (89), we introduced the type class enumerators \( N_{C_n}(Q) \). We also recall from [23, Lem. 2.6] that \( \exp(-n[D(Q\|P) + H(Q)]) \) is the exact \( P^n \)-probability of a sequence of type \( Q \). Note that the bound in (89) is independent of \( y \), just as for the calculation of \( A \) in (86). In the following, we find bounds on \( B \) that turn out to tight in the sense that the analysis yield the final result in Theorem 3. We start with lower bounding \( B \) by as follows:

\[
B \geq \mathbb{E}_{C_n}\left[ \left( \max_{Q' \in \mathcal{P}_n(X)} N_{C_n}(Q') \exp \left( - n[D(Q'\|P) + H(Q')] \right) \right)^s \right] \quad (90)
\]

\[
= \mathbb{E}_{C_n}\left[ \left( \max_{Q' \in \mathcal{P}_n(X)} N_{C_n}(Q') \right)^s \exp \left( - ns[D(Q'\|P) + H(Q')] \right) \right] \quad (91)
\]

\[
\geq \max_{Q' \in \mathcal{P}_n(X)} \mathbb{E}_{C_n}[N_{C_n}(Q')] \exp \left( - ns[D(P_n\|P) + H(P_n)] \right) \quad (92)
\]

\[
\geq \mathbb{E}_{C_n}[N_{C_n}(P_n)^s] \exp \left( - ns[D(P_n\|P) + H(P_n)] \right), \quad (93)
\]

where \( P_n \in \mathcal{P}_n(X) \) is defined as

\[
P_n \in \arg \min_{Q \in \mathcal{P}_n(X)} \{\|Q - P\|_1 : H(Q) \geq H(P) + 2an^{-1} \} \quad (94)
\]

Then, \( H(P_n) = H(P) + 2an^{-1} + O(n^{-1} \log n) \) due to Fannes inequality [23, Lem. 2.7], i.e., continuity of Shannon entropy. One immediately finds that

\[
D(P_n\|P) = O(\|P_n - P\|_1^2) = O(n^{-2t}) \quad (95)
\]

which is negligible. More precisely,

\[
-ns[D(P_n\|P) + H(P_n)] = -un^{1-t}H(P) - 2aun^{1-2t} + O(n^{1-3t}) \quad (96)
\]

as \( n \) grows.

Next, apply Lemma 10 in Appendix C to the case with \( L = M_n - 1, M_1 = d^n, M_2 = |\mathcal{T}_{P_n}^{(n)}|, \{X_1, \ldots, X_L\} = \{X^n_{m'}\}_{m' \neq m}, A = \mathcal{T}_{P_n}^{(n)}, s = un^{-t}, \) and a fixed positive constant \( \epsilon > 0 \). Since \( \log |\mathcal{T}_{P_n}^{(n)}| \geq nH(P_n) - (d-1) \log(n+1) = nH(P) + 2aun^{1-2t} + (d-1) \log(n+1) + O(1) \), we have \( \log L + \log M_2 - \log M_1 \geq an^{1-t} - (d-1) \log(n+1) + O(1) \). Thus, \( \log[1 - \exp(-L\epsilon^2/(2M_1))] = o(1) \). Also, we have \( s \log(1 - \epsilon) = un^{-t} \log(1 - \epsilon) = o(1) \) and \( s(\log L + \log M_2 - \log M_1) \geq aun^{1-2t} + o(1) \). Therefore, Lemma 10 says that

\[
\log \mathbb{E}_{C_n}[N_{C_n}(P_n)^s] = \log \mathbb{E}[N^s] \geq aun^{1-2t} + o(1). \quad (97)
\]

Combining (93), (96) and (97), we find that

\[
\log B \geq -n^{1-t}uH(P) - an^{1-2t}u + O(n^{1-3t}) + o(1). \quad (98)
\]
Now, we proceed to upper bound $B$ in (89). Since the map $x \mapsto x^s$ for $0 < s < 1$ is concave and $|T^{(n)}_{Q'}| \leq \exp(nH(Q'))$, we have

$$B = \mathbb{E}_{C_n}\left[\left(\sum_{Q' \in \mathcal{P}_n(x)} N_{C_n}(Q') \exp\left(-n[D(Q'\|P) + H(Q')]\right)\right)^s\right]$$

(99)

$$\leq \left(\mathbb{E}_{C_n}\left[\sum_{Q' \in \mathcal{P}_n(x)} N_{C_n}(Q') \exp\left(-n[D(Q'\|P) + H(Q')]\right)\right]\right)^s$$

(100)

$$= \left(\sum_{Q' \in \mathcal{P}_n(x)} \mathbb{E}_{C_n}\left[N_{C_n}(Q')\right] \exp\left(-n[D(Q'\|P) + H(Q')]\right)\right)^s$$

(101)

$$= \left(\sum_{Q' \in \mathcal{P}_n(x)} \frac{M_n|\mathcal{T}^{(n)}_{Q'}|}{d^n} \exp\left(-n[D(Q'\|P) + H(Q')\right]\right)^s$$

(102)

$$\leq (n + 1)^{d-1} \max_{Q' \in \mathcal{P}_n(x)} \left(\frac{M_n|\mathcal{T}^{(n)}_{Q'}|}{d^n}\right) \exp\left(-ns[D(Q'\|P) + H(Q')\right]\right)^s$$

(103)

$$= (n + 1)^{(d-1)s} \max_{Q' \in \mathcal{P}_n(x)} \left(\frac{M_n|\mathcal{T}^{(n)}_{Q'}|}{d^n}\right) \exp\left(-s[nH(P) - an^{1-t} + nH(Q')] - ns[D(Q'\|P) + H(Q')\right]\right)^s$$

(104)

$$\leq (n + 1)^{(d-1)s} \max_{Q' \in \mathcal{P}_n(x)} \left(\frac{M_n|\mathcal{T}^{(n)}_{Q'}|}{d^n}\right) \exp\left(-s[nH(P) - an^{1-t} - nsD(Q'\|P)]\right)^s$$

(105)

$$\leq (n + 1)^{(d-1)s} \exp\left(-s[nH(P) - an^{1-t} + \max_{Q' \in \mathcal{P}(x)} -nsD(Q'\|P)]\right)^s$$

(106)

$$\leq (n + 1)^{(d-1)s} \exp\left(-s[nH(P) - an^{1-t}]\right)^s$$

(107)

$$\leq (n + 1)^{(d-1)s} \exp\left(-s[nH(P) - an^{1-t}]\right)^s$$

(108)

$$= (n + 1)^{(d-1)s} \exp\left(-an^{1-t}H(P) - aun^{1-2t}\right)^s$$

(109)

Thus, we find that

$$\log B \leq -n^{1-t}uH(P) - an^{1-2t}u + o(1).$$

(110)

Combining the evaluations of $A$ and $B$ together in (61), we see that the sum over $y$ cancels the $1/d^n$ term in (86) and the first-order entropy terms also cancel. The final expression for the cumulant generating function of $F_n$ satisfies (62) as desired.

V. DISCUSSION AND FUTURE WORK

In this paper, we analyzed channel coding with the erasure option where we designed both the undetected and total errors to decay subexponentially and asymmetrically. We analyzed two regimes, namely, the pure moderate deviations and mixed regimes. We proposed an information spectrum-type decoding rule [18] and showed using an ensemble converse argument that this simple decoding rule is, in fact, optimal for additive DMCs. To do so, we estimated appropriate cumulant generating functions of the total and undetected errors. We also developed a modified version of the Gärtner-Ellis theorem that is particularly useful for our problem. In contrast to previous works on erasure (and list) decoding [1]–[8], we do not evaluate the rate of exponential decay of the two error probabilities. In our work, the two error probabilities decay subexponentially (and asymmetrically) for the pure moderate deviations setting. For the mixed regime, the total (and hence erasure) error is non-vanishing while the undetected error decays as $\exp(-bn^{1/2})$ for some $b > 0$.

In the future, it would be useful to remove the assumption that the DMC is additive for the ensemble converse result in Section III-B. However, it appears that this is not straightforward and it is likely that we have to make an assumption like that for Theorem 1 of Merhav’s work [6]. In addition, it would be useful from a mathematical standpoint to tighten the higher-order asymptotics for the expansions of the log-probabilities in (34) and (38), but this appears to require some independence assumptions which are not available in the Gärtner-Ellis theorem (so
a new concentration bound may be required). In addition, this seems to require tedious calculus to evaluate the higher-order asymptotic terms of the cumulant generating function in Lemma 7. A refinement of the type class enumerator method [6]–[9] seems to be necessary for this purpose.

APPENDIX A
MODIFIED GÄRTNER-Ellis THEOREM

Here we present a modified form of the Gärtner-Ellis theorem with general order.

**Theorem 8** (Modified Gärtner-Ellis theorem). We consider three sequences $\alpha_n, \beta_n, \gamma_n$ satisfying $\beta_n \to \infty$, $\gamma_n \to 0$. Let $p_n$ be a sequence of distributions, and $X_n$ be a sequence of random variables. Define the cumulant generating function $\mu_n(\theta) := \log \mathbb{E}_{p_n}[\exp(\theta X_n)]$. Assume that

$$\mu_n(\theta_0 + \gamma_n y) = \alpha_n + \beta_n \nu_1 + \beta_n \gamma_n \nu_2(y)$$

(111)

where $\theta_0 \leq 0$ and $\nu_1$ are constants, $\lim_{n \to \infty} \nu_2(y) = \nu_2(y)$ and $\nu_2(y)$ is continuous. We also fix $x \in \mathbb{R}$ and assume that $y_0$ is such that $\nu_2(y_0) = x$.

• Case (i): If $\theta_0 < 0$, we have

$$(y_0 x - \nu_2(y_0)) \beta_n \gamma_n + o(\beta_n \gamma_n) \leq -\log p_n \left\{ \frac{X_n}{\beta_n} \leq x \right\} + \alpha_n - (\theta_0 x - \nu_1) \beta_n \leq o(\beta_n)$$

(112)

(113)

• Case (ii): If $\theta_0 = \alpha_n = \nu_1 = 0$, $y_0 < 0$, and $\beta_n \gamma_n \to \infty$, we have

$$\lim_{n \to \infty} -\frac{1}{\beta_n \gamma_n} \log p_n \left\{ \frac{X_n}{\beta_n} \leq x \right\} = y_0 x - \nu_2(y_0).$$

(114)

We only prove Case (i) as Case (ii) is essentially the standard Gärtner-Ellis theorem with normalization $\beta_n \gamma_n \to \infty$ (instead of $n$). Observe that in the lower bound in (112), we can characterize a third-order term scaling as $\beta_n \gamma_n$ but unfortunately, we were not able to do the same for the upper bound in (113). We leave the tightening the bounds for future work.

Proof: For the lower bound, first note that $\theta_0 < 0$ and $\gamma_n \to 0$, so for sufficiently large $n$, we have $\theta_0 + \gamma_n y_0 < 0$. Thus, using Markov’s inequality,

$$p_n \left\{ \frac{X_n}{\beta_n} \leq x \right\} = p_n \left\{ \exp \left[ \left( \frac{X_n}{\beta_n} - x \right) \beta_n (\theta_0 + \gamma_n y_0) \right] \geq 1 \right\} \leq \mathbb{E}_{p_n} \left\{ \exp \left[ \left( \frac{X_n}{\beta_n} - x \right) \beta_n (\theta_0 + \gamma_n y_0) \right] \right\}.$$

(115)

(116)

In other words,

$$-\log p_n \left\{ \frac{X_n}{\beta_n} \leq x \right\} \geq x \beta_n \theta_0 + x \gamma_n \beta_n y_0 - \mu_n(\theta_0 + \gamma_n y_0)$$

(117)

$$= -\alpha_n + (\theta_0 x - \nu_1) \beta_n + (y_0 x - \nu_2(y_0)) \beta_n \gamma_n$$

(118)

where (118) follows from the expansion of $\mu_n(\theta_0 + \gamma_n y)$ in (111). So from the assumption that $\nu_2(y_0) \to \nu_2(y_0)$, we obtain

$$-\log p_n \left\{ \frac{X_n}{\beta_n} \leq x \right\} \geq -\alpha_n + (\theta_0 x - \nu_1) \beta_n + (y_0 x - \nu_2(y_0)) \beta_n \gamma_n + o(\beta_n \gamma_n)$$

(119)

as desired.

For the upper bound in (113), first notice that

$$p_n \left\{ \frac{X_n}{\beta_n} \leq x \right\} \geq p_n \left\{ \frac{X_n}{\beta_n} < x \right\}.$$

(120)
So by the same reasoning as the lower bound in the proof of Cramér’s theorem [20, Thm. 2.2.3], it suffices to further lower bound (120) by considering the \( p_n \)-probability of open balls \( \mathcal{B}_{n,\delta}(x_0) := \{ \omega : X_n(\omega) \in \beta_n(x_0 - \delta, x_0 + \delta) \} \) for fixed \( x_0 < x \) and \( \delta \in (0, x - x_0) \). Define the tilted probability measure

\[
\tilde{p}_{n,\delta}(\omega) := p_n(\omega) \exp(\theta X_n(\omega) - \mu_n(\theta)).
\] (121)

It follows by straightforward calculations (cf. proof of [20, Thm. 2.3.6(b)]) involving the tilted probability measure that

\[
\frac{1}{\beta_n} \log p_n(\mathcal{B}_{n,\delta}(x_0)) \geq \frac{1}{\beta_n} \mu_n(\theta_0) - \theta_0 x_0 - |\theta_0| \delta + \frac{1}{\beta_n} \log \tilde{p}_{n,\delta_0}(\mathcal{B}_{n,\delta}(x_0))
\] (122)

Using the definition of \( \mu_n \) in (111) with \( y = 0 \), this implies that

\[
\frac{1}{\beta_n} \left[ - \log p_n(\mathcal{B}_{n,\delta}(x_0)) + \alpha_n \right] \leq \gamma_n \nu_{2,n}(0) + \theta_0 x_0 - \nu_1 + |\theta_0| \delta - \frac{1}{\beta_n} \log \tilde{p}_{n,\delta_0}(\mathcal{B}_{n,\delta}(x_0))
\] (123)

Since \( \gamma_n \to 0 \) and \( \nu_2(0) < \infty \), we have

\[
\lim_{\delta \downarrow 0} \limsup_{n \to \infty} \frac{1}{\beta_n} \left[ - \log p_n(\mathcal{B}_{n,\delta}(x_0)) + \alpha_n \right] \leq \theta_0 x_0 - \nu_1 - \lim_{\delta \downarrow 0} \liminf_{n \to \infty} \frac{1}{\beta_n} \log \tilde{p}_{n,\delta_0}(\mathcal{B}_{n,\delta}(x_0)).
\] (124)

Using the same steps as in [20, Thm. 2.3.6(b)], we will show that \( \tilde{p}_{n,\delta_0}(\mathcal{B}_{n,\delta}(x_0)) \to 1 \) for every \( \delta > 0 \). Consequently, the \( \lim \inf \) in the rightmost term in (124) is zero for every \( \delta > 0 \) and so the rightmost term is zero. Define

\[
\tilde{\mu}_n(\lambda) := \log E_{\tilde{p}_{n,\delta_0}}[\exp(\lambda X_n)],
\] (125)

where \( \tilde{p}_{n,\delta_0} \) is defined in (121). From direct calculations, we obtain

\[
\tilde{\mu}_n(\gamma_n y) = \mu_n(\theta_0 + \gamma_n y) - \mu_n(\theta_0).
\] (126)

Consequently, from the definition of \( \mu_n \) in (111), the first two terms (involving \( \alpha_n \) and \( \beta_n \nu_1 \)) cancel and we have

\[
\tilde{\mu}_n(\gamma_n y) = \beta_n \gamma_n \left[ \nu_{2,n}(y) - \nu_{2,n}(0) \right].
\] (127)

By assumption, \( \nu_{2,n}(y) \to \nu_2(y) \) so the following limit exists

\[
\lim_{n \to \infty} \frac{1}{\beta_n} \gamma_n \tilde{\mu}_n(\gamma_n y) = \nu_2(y).
\] (128)

Hence, the sequence of cumulant generating functions \( \{ y \mapsto \xi_n(y) := \tilde{\mu}_n(\gamma_n y) / \gamma_n \}_{n \in \mathbb{N}} \) satisfies Assumption 2.3.2 in [20] with normalization \( \beta_n \to \infty \). So by applying Lemma 2.3.9 in [20], we see that the Fenchel-Legendre transform of \( y \mapsto \lim_{n \to \infty} \xi_n(y) / \beta_n \) is a good rate function. Following the rest of the proof of the usual Gärtner-Ellis theorem [20, pp. 50], we can apply the large deviations upper bound to the set \( \mathcal{B}_{n,\delta}(x_0)^c \) to conclude that

\[
\limsup_{n \to \infty} \frac{1}{\beta_n} \log \tilde{p}_{n,\delta_0}(\mathcal{B}_{n,\delta}(x_0)^c) < 0.
\] (129)

In other words, \( \tilde{p}_{n,\delta_0}(\mathcal{B}_{n,\delta}(x_0)) \geq 1 - \exp(-c\beta_n) \) for some constant \( c > 0 \). So we have shown that the final term in (124) is zero, and thus

\[
\limsup_{n \to \infty} \frac{1}{\beta_n} \left[ - \log p_n \left( \frac{X_n}{\beta_n} \leq x \right) + \alpha_n \right] \leq \theta_0 x_0 - \nu_1.
\] (130)

Now let \( x_0 \uparrow x \) (note \( \theta_0 < 0 \)) to obtain

\[
\limsup_{n \to \infty} \frac{1}{\beta_n} \left[ - \log p_n \left( \frac{X_n}{\beta_n} \leq x \right) + \alpha_n \right] \leq \theta_0 x - \nu_1,
\] (131)

yielding the tightest possible upper bound. Thus, we have shown the upper bound in (113).

\[\blacksquare\]

**Remark 2.** The crux in the proof of the (more challenging) upper bound is to shift the offset term \( \alpha_n \) to the left-hand-side of (123) and the rest of the proof proceeds similarly to the usual Gärtner-Ellis proof with normalization (order) \( \beta_n \) instead of \( n \).
APPENDIX B

CONVERGENCE IN DISTRIBUTION BASED ON CONVERGENCE OF CUMULANT GENERATING FUNCTIONS

Lemma 9. Let \( p_n \) be a sequence of distributions (probability measures) on \( \mathbb{R} \). Suppose that
\[
\log \int_\mathbb{R} \exp(sx) p_n(dx) \to f(s) := \frac{s^2}{2}, \quad \forall s > 0. \tag{132}
\]
Then, \( p_n \) converges (weakly) to the standard normal distribution.

Notice that in (132), the assumption pertains only to \( s > 0 \). In particular, it is not assumed that the convergence holds for all \( s \in \mathbb{R} \), in which case convergence of \( p_n \) to the standard normal is an elementary fact (cf. Lévi’s continuity theorem [28, Thm. 18.21]).

See Mukherjea et al. [25, Thm. 2] for a statement (and accompanying proof) more general than Lemma 9. We provide a proof sketch of Lemma 9 for completeness.

Proof Sketch of Lemma 9: Due to the assumption in (132), the sequence of probability measures
\[
q_n(dx) := \exp \left( x - f(1) \right) p_n(dx) \tag{133}
\]
has a finite cumulant generating function for all \( s > -1 \). Indeed, for all \( s > -1 \), the cumulant generating function of \( \{q_n\}_{n \in \mathbb{N}} \) converges to that for a normal distribution with mean 1 and variance 1, i.e.,
\[
\lim_{n \to \infty} \log \int_\mathbb{R} \exp(sx) q_n(dx) = s + \frac{s^2}{2}, \quad \text{for} \quad s > -1. \tag{134}
\]
The sequence of probability measures \( \{q_n\}_{n \in \mathbb{N}} \) is tight and has a weak limit. Clearly, from (134), the sequence of cumulant generating functions of \( \{q_n\}_{n \in \mathbb{N}} \) converges pointwise on an interval containing the origin. Thus, by Curtiss’ Theorem [26] and (134), \( \{q_n\}_{n \in \mathbb{N}} \) converges weakly to a normal distribution with mean 1 and variance 1. Since \( q_n \) is simply an exponential tilting of \( p_n \) per (133), we can invert this exponential tilting (cf. [25, Proof of Thm. 2]) to conclude that \( \{p_n\}_{n \in \mathbb{N}} \) converges weakly to the standard normal distribution. ■

APPENDIX C

A BASIC CONCENTRATION BOUND

Lemma 10. Let \( X_1, \ldots, X_L \) be independent random variables, each subject to the uniform distribution on \( \{1, \ldots, M_1\} \). We fix a subset \( A \subset \{1, \ldots, M_1\} \) whose cardinality is \( M_2 \). We denote the random number \( |\{i \in \{1, \ldots, L\} : X_i \in A\}| \) by \( N \). For every \( 0 < s < 1 \),
\[
\mathbb{E}[N^s] \geq \left[ \frac{LM_2}{M_1} (1 - \epsilon) \right]^s \left[ 1 - \exp \left( -L \frac{M_2}{2M_1} \epsilon^2 \right) \right] \tag{135}
\]
where \( 0 < \epsilon < 1 \) is also an arbitrary number.

Proof: By straightforward calculations, we have
\[
\mathbb{E}[N^s] = \sum_{l=0}^L l^s \Pr(N = l) \tag{136}
\]
\[
\geq \sum_{l \geq LM_2(1-\epsilon)/M_1} l^s \Pr(N = l) \tag{137}
\]
\[
\geq \left[ \frac{LM_2}{M_1} (1 - \epsilon) \right]^s \Pr \left( N \geq \frac{LM_2(1-\epsilon)}{M_1} \right) \tag{138}
\]
\[
= \left[ \frac{LM_2}{M_1} (1 - \epsilon) \right]^s \left[ 1 - \Pr \left( \frac{N}{L} < (1 - \epsilon) \frac{M_2}{M_1} \right) \right]. \tag{139}
\]

Now since the event in probability in (139) implies that the relative frequency of the number of events \( \{X_i \in A\}, i = 1, \ldots, L \) is less than \( 1 - \epsilon \) multiplied by the mean \( \mathbb{E}[1\{X_i \in A\}] = M_2/M_1 \) of each indicator \( 1\{X_i \in A\} \), we can invoke the Chernoff bound for independent Poisson trials (e.g., [29, Thm. 4.5]) to conclude that
\[
\Pr \left( \frac{N}{L} < (1 - \epsilon) \frac{M_2}{M_1} \right) = \Pr \left( \frac{1}{L} \sum_{i=1}^L 1\{X_i \in A\} < (1 - \epsilon) \frac{M_2}{M_1} \right) \leq \exp \left( -L \frac{M_2}{2M_1} \epsilon^2 \right). \tag{140}
\]
Combining this with (139) concludes the proof. ■
Acknowledgements: MH is grateful for Prof. Nobuo Yoshida for clarifying Lemma 9. VT is grateful to Prof. Rongfeng Sun for clarifying the same lemma.

MH is partially supported by a MEXT Grant-in-Aid for Scientific Research (A) No. 23246071. MH is also partially supported by the National Institute of Information and Communication Technology (NICT), Japan. The Centre for Quantum Technologies is funded by the Singapore Ministry of Education and the National Research Foundation as part of the Research Centres of Excellence programme. VT’s research is supported by NUS startup grants WBS R-263-000-A98-750 (FoE) and WBS R-263-000-A98-133 (ODPRT).

REFERENCES

[1] G. D. Forney. Exponential error bounds for erasure, list, and decision feedback schemes. *IEEE Trans. on Inform. Th.*, IT-14(2):206–220, 1968.

[2] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp. Lower bounds to error probability for coding in discrete memoryless channels I-II. *Information and Control*, 10:65–103,522–552, 1967.

[3] I. E. Telatar and R. G. Gallager. New exponential upper bounds to error and erasure probabilities. In *Proc. of Intl. Symp. on Inform. Th.*, page 379, Trondheim, Norway, 1994.

[4] V. M. Blinovsky. Error probability exponent of list decoding at low rates. *Problems of Information Transmission*, 27(4):277–287, 2001.

[5] P. Moulin. A Neyman-Pearson approach to universal erasure and list decoding. *IEEE Trans. on Inform. Th.*, 55(10):4462–4478, Oct 2009.

[6] N. Merhav. Error exponents of erasure/list decoding revisited via moments of distance enumerators. *IEEE Trans. on Inform. Th.*, 54(10):4439–4447, 2008. arXiv:0711.2501.

[7] N. Merhav. List decoding–random coding exponents and expurgated exponents. Submitted to the *IEEE Trans. on Inform. Th.*, 2013. arXiv:1311.7298.

[8] A. Somekh-Baruch and N. Merhav. Exact random coding exponent for erasure decoding. *IEEE Trans. on Inform. Th.*, 57(10):6444–6454, 2011.

[9] N. Merhav. *Statistical Physics and Information Theory*, volume 6 of *Foundations and Trends in Communications and Information Theory*. Now Publishers Inc, 2010.

[10] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 1948.

[11] V. Strassen. Asymptotische Abschätzungen in Shannon’s Informationstheorie. In *Trans. Third Prague Conf. Inf. Theory*, pages 689–723, Prague, 1962. http://www.math.cornell.edu/~pmlut/strassen.pdf.

[12] M. Hayashi. Information spectrum approach to second-order coding rate in channel coding. *IEEE Trans. on Inform. Th.*, 55(11):4947–4966, 2009.

[13] Y. Polyanskiy, H. V. Poor, and S. Verdú. Channel coding rate in the finite blocklength regime. *IEEE Trans. on Inform. Th.*, 56(5):2307–2359, 2010.

[14] Y. Altuğ and A. B. Wagner. Moderate deviations in channel coding. *IEEE Trans. on Inform. Th.*, 60(8):4417–4426, 2014.

[15] V. Y. F. Tan. Moderate-deviations of lossy source coding for discrete and Gaussian sources. In *Proc. of Intl. Symp. on Inform. Th.*, Cambridge, MA, 2012. arXiv:1111.2217.

[16] V. Y. F. Tan and P. Moulin. Fixed error probability asymptotics for erasure and list decoding. *IEEE Trans. on Inform. Th.*, 2014. arXiv:1402.4881.

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

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

[19] Y. Altuğ, A. B. Wagner, and I. Kontoyiannis. Lossless compression with moderate error probability. In *Proc. of Intl. Symp. on Inform. Th.*, 2013.

[20] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Springer, 2nd edition, 1998.

[21] P.-N. Chen. Generalization of Gartner-Ellis theorem. *IEEE Trans. on Inform. Th.*, 46(7):2752–2760, 2000.

[22] C. Joutard. A strong large deviation theorem. *Mathematical Methods of Statistics*, 22(2):155–164, April 2013.

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

[24] S. Verdú and I. Kontoyiannis. Optimal lossless data compression: Non-asymptotics and asymptotics. *IEEE Trans. on Inform. Th.*, 60(2):777–795, 2014.

[25] A. Mukherjea, M. Rao, and S. Suen. A note on moment generating functions. *Statistics & Probability Letters*, 76(1):1185–1189, Jun 2006.

[26] J. H. Curtiss. A note on the theory of moment generating functions. *Annals of Mathematical Statistics*, 13(4):430–433, 1942.

[27] W. Feller. *An Introduction to Probability Theory and Its Applications*. John Wiley and Sons, 2nd edition, 1971.

[28] B. E. Fristedt and L. F. Gray. *A modern approach to probability theory*. Birkhäuser Boston, 1996.

[29] M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.