Asymmetric Evaluations of Erasure and Undetected Error Probabilities

Masahito Hayashi, *Senior Member, IEEE*, and Vincent Y. F. Tan, *Senior Member, IEEE*

***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 with uniform input distribution, we establish that our analysis is tight with respect to the ensemble average. This is done by analyzing the ensemble performance in terms of a tradeoff between the code rate, the undetected and total errors. This tradeoff is parameterized by the threshold in a generalized likelihood ratio test. 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 the 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]–[10] 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 [11], 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 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 [12] 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 [13] and Polyanskiy *et al.* [14]. See [15] for a review. Altuğ and Wagner [16] pioneered the *moderate deviations* analysis for DMCs and Tan [17] 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 [18] recently studied the information-theoretic limits of channel coding with erasures where both the undetected and total error probabilities are fixed at positive constants.

Manuscript received September 6, 2014; revised September 15, 2015; accepted October 20, 2015. Date of publication October 27, 2015; date of current version November 18, 2015. M. Hayashi was supported in part by the National Institute of Information and Communication Technology, Japan, in part by a Grant-in-Aid for Scientific Research (A) under Grant 23246071, in part by a Grant-in-Aid for Young Scientists (A) through MEXT under Grant 20686026, in part by the Singapore Ministry of Education, and in part by the National Research Foundation through the Research Centres of Excellence Programme. V. Y. F. Tan was supported by the National University of Singapore under Grant R-263-000-A98-750/133. This paper was presented in part at the 2015 International Symposium on Information Theory.

M. Hayashi is with the Graduate School of Mathematics, Nagoya University, Nagoya 464-860, Japan, and also with the Centre for Quantum Technologies, National University of Singapore, Singapore 117542 (e-mail: 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, Singapore 119077 (e-mail: vtan@nus.edu.sg).

Communicated by A. B. Wagner, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2015.2495258

B. Main Contributions

In this work, we study different regimes for the errors and erasure problems. In particular, we analyze the *moderate deviations* [16], [19] 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 [20] 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 equally importantly, we show that our so-constructed codes above are *tight with respect to the ensemble average*, or more succinctly, *ensemble-tight* for additive DMCs with the uniform random coding distribution. This means that our ensemble evaluation of the two error probabilities (averaged over the random codebook) is tight in some asymptotic sense to be made precise in the statements. To prove these statements, 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 [16], [17], [19], [21], [22]) 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 on the type class enumerator method [6]–[10], carries over to the mixed regime in which the total error probability is asymptotically a constant [12]–[14] 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 be the same as for the achievability results showing that the decoder designed based on information spectrum analysis is, in fact, asymptotically optimal, i.e., Forney's decoding regions (together with our analyses) trade off the Pareto-optimal curve between the two error probabilities.

Finally, an auxiliary contribution of the present work is a new mathematical tool. We develop a modified ("shifted") version of the Gärtner-Ellis theorem [23, Th. 2.3.6] 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., [24], [25]). It turns out to be very useful for our application and may be of independent interest in other information-theoretic settings. A self-contained proof containing some novel proof techniques is contained in Appendix A.

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 of 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 Appendix A.

II. Notation and Problem Setting

A. Notation

In this paper, we adopt standard notation in information theory, particularly in the book by Csiszár and Körner [26]. 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 a superscript, i.e., $X^n = (X_1, \ldots, X_n)$. Information-theoretic quantities are denoted in the usual way, e.g., $H(P)$ is the entropy of the random variable X with distribution P. The set of all probability mass functions on a finite set $\mathcal{X}$ is denoted by $\mathcal{P}(\mathcal{X})$ while the subset of types (empirical distributions) with denominator n is denoted as $\mathcal{P}_n(\mathcal{X})$. The set of all sequences with type $P \in \mathcal{P}_n(\mathcal{X})$, the *type class*, is denoted as $\mathcal{T}_P = \{\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n : \sum_{i=1}^n \mathbf{1}\{x_i = a\} = nP(a), \forall a \in \mathcal{X}\}$. The ℓ_1 (twice the variational) distance between $P, Q \in \mathcal{P}(\mathcal{X})$ is denoted as $\|P - Q\|_1 = \sum_{x \in \mathcal{X}} |P(x) - Q(x)|$. All logs and exps are with respect to the natural base e.

B. Discrete Memoryless Channels (DMCs)

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 $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$ the probability of the output letters $\mathbf{y} = (y_1, \ldots, y_n) \in \mathcal{Y}^n$ is the product $\prod_{i=1}^n W(y_i|x_i)$. The capacity of the DMC is denoted as

$$C = C(W) := \max\{I(P_X, W) : P_X \in \mathcal{P}(\mathcal{X})\}. \tag{1}$$

Let the set of capacity-achieving input distributions be

$$\Pi = \Pi(W) := \{P_X \in \mathcal{P}(\mathcal{X}) : I(P_X, W) = C(W)\}. \tag{2}$$

This set is compact.

C. Additive DMCs

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 $P \in \mathcal{P}(\mathcal{X})$ with positive entries $P(x) > 0, x \in \mathcal{X}$ such that

$$W(y|x) = P(y - x) \tag{3}$$

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, $Y = X + Z \pmod d$ where the noise Z has distribution P. Consequently, P is also called the *noise distribution*. The capacity of the additive channel W is $C = \log d - H(P)$ and which is achieved (possibly non-uniquely) by the uniform distribution on $\{0, 1, \ldots, d-1\}$ [27, Th. 7.2.1]. This class of channels, while somewhat restrictive, includes important DMCs such as the binary symmetric channel (BSC) where $d = 2$ and $P(0) = q$ and $P(1) = 1 - q$ and $q \in (0, 1)$ is the crossover probability. Also, additive DMCs simplify analyses in other problems in Shannon theory such as in the error exponent analysis of the performance of linear codes [28].

D. Channel Coding With the Erasure Option

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* $\mathsf{f} : \{1, \ldots, M_n\} \to \mathcal{X}^n$ transforms the message to a codeword. The *codebook* $\mathcal{C}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_{M_n}\}$ where $\mathbf{x}_m = \mathsf{f}(m)$ is the set of all codewords. The channel W^n then applies a random transformation to the chosen codeword $\mathbf{x}_m \in \mathcal{X}^n$ resulting in $\mathbf{y} \in \mathcal{Y}^n$. A *decoder* $\mathsf{d} : \mathcal{Y}^n \to \{0, 1, \ldots, M_n\}$ either declares an 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 $\mathcal{D}_m := \mathsf{d}^{-1}(m)$. The set of all $\mathbf{y} \in \mathcal{D}_0$ leads to an *erasure* event.

E. Total and Undetected Error Probabilities

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

$$\Pr(\mathcal{E}_1 | \mathcal{C}_n) = \frac{1}{M_n} \sum_{m=1}^{M_n} \sum_{\mathbf{y} \in \mathcal{D}_m^c} W^n(\mathbf{y} | \mathbf{x}_m). \tag{4}$$

This is the *total error probability*.

The other 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

$$\Pr(\mathcal{E}_2 | \mathcal{C}_n) = \frac{1}{M_n} \sum_{m=1}^{M_n} \sum_{\mathbf{y} \in \mathcal{D}_m} \sum_{m' \neq m} W^n(\mathbf{y} | \mathbf{x}_{m'}). \tag{5}$$

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 $\Pr(\mathcal{E}_1 | \mathcal{C}_n)$, because undetected errors are usually more undesirable than erasures.

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

$$V(P_X, W) := \sum_{x \in \mathcal{X}} P_X(x) \sum_{y \in \mathcal{Y}} W(y|x) \times \left[\log \frac{W(y|x)}{P_X W(y)} - D(W(\cdot|x) \| P_X W) \right]^2, \tag{6}$$

where $P_X W(y) = \sum_x P_X(x) W(y|x)$ is the output distribution induced by P_X and W. This quantity is finite whenever $W(\cdot|x) \ll P_X W$ for all x. We further define the minimum and maximum conditional information variances as

$$V_{\max}(W) := \max_{P_X \in \Pi} V(P_X, W) \quad \text{and} \tag{7}$$

$$V_{\min}(W) := \min_{P_X \in \Pi} V(P_X, W). \tag{8}$$

Since Π is compact and $P_X \mapsto V(P_X, W)$ is continuous, there exists capacity-achieving input distributions $P_X \in \Pi$ that achieve both $V_{\min}(W)$ and $V_{\max}(W)$ and so they are finite. The P_X that achieves $V_{\min}(W)$ may not be the same as that achieving $V_{\max}(W)$. Note that for all $P_X \in \Pi$, we have $V(P_X, W) = U(P_X, W)$ [14, Lemma 62], where the unconditional information variance $U(P_X, W)$ is defined as

$$U(P_X, W) := \sum_{x \in \mathcal{X}} P_X(x) \sum_{y \in \mathcal{Y}} W(y|x) \left[\log \frac{W(y|x)}{P_X W(y)} - C \right]^2. \tag{9}$$

We assume that the channel W satisfies $V_{\min}(W) > 0$ throughout. This holds for all interesting DMCs (except some degenerate cases) and we make this assumption which is standard in moderate deviations analysis [16], [19]. If $V_{\min}(W) = 0$, the conclusion from the moderate deviations theorem [23, Th. 3.7.1] fails to hold.

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*

$$\log M_n = nC - a n^{1-t}. \tag{10}$$

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

$$\lim_{n \to \infty} -\frac{1}{n^{1-2t}} \log \Pr(\mathcal{E}_1 | \mathcal{C}_n) = \frac{(a-b)^2}{2 V_{\min}(W)}, \quad \textit{and} \tag{11}$$

$$\liminf_{n \to \infty} -\frac{1}{n^{1-t}} \log \Pr(\mathcal{E}_2 | \mathcal{C}_n) \geq b. \tag{12}$$

The proof of this result can be found in Section IV-A. We assume that $a > b$ because if we demand that the undetected error probability decays as in (12), we must have that the rate $\frac{1}{n} \log M_n$ backs off further from capacity per (10).

[1] 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).

Furthermore, $b < 0$ corresponds to the list region which we do not discuss in detail in this paper.

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 $\{\tilde{\mathcal{D}}_m\}_{m=1}^{M_n}$ motivated by information spectrum analysis [20]:

$$\tilde{\mathcal{D}}_m := \left\{ \mathbf{y} : \log \frac{W^n(\mathbf{y}|\mathbf{x}_m)}{(P_X W)^n(\mathbf{y})} \ge \log M_n + bn^{1-t} \right\}, \quad (13)$$

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

$$\hat{\mathcal{D}}_0 := \left(\bigcap_{m=1}^{M_n} \tilde{\mathcal{D}}_m^c \right) \cup \left(\bigcup_{m \neq m'} (\tilde{\mathcal{D}}_m \cap \tilde{\mathcal{D}}_{m'}) \right). \quad (14)$$

Then, the *decoding region for message* $m = 1, \ldots, M$ is defined to be

$$\hat{\mathcal{D}}_m := \tilde{\mathcal{D}}_m \setminus \hat{\mathcal{D}}_0. \quad (15)$$

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 $\cup_{m=0}^{M_n} \hat{\mathcal{D}}_m = \mathcal{Y}^n$. The intuition behind the decoding regions in (14)–(15) is as follows: The erasure region in (14) is the union of all the complements of nominal regions $\tilde{\mathcal{D}}_m^c$ and the sets of pairwise intersections which potentially cause confusion in decoding, namely, $\tilde{\mathcal{D}}_m \cap \tilde{\mathcal{D}}_{m'}$. After defining the erasure region $\hat{\mathcal{D}}_0$, we remove this from the nominal regions $\tilde{\mathcal{D}}_m$ to form the actual decoding region for each message $\hat{\mathcal{D}}_m$. Note that in the ensemble tightness results to be presented in Section III-B we do not analyze the information spectrum decoding regions in (13)–(15). Rather we analyze the *optimal decoder* suggested by Forney [1]. Hence, the decoding regions in (13)–(15), in general, may not be asymptotically optimal, unlike Forney's decoding regions. However, we do show that these decoders are asymptotically optimal for additive DMCs.

Theorem 1 corresponds to the so-called moderate deviations regime in channel coding considered by Altuğ and Wagner [16] and Polyanskiy and Verdú [19]. Thus, the appearance of the term $V_{\min}(W)$ 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 $\exp(-\Omega(n^{1-t}))$.

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

$$\varphi(w) := \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{w^2}{2}\right) \quad (16)$$

to be the probability density function of a standard Gaussian annd

$$\Phi(\alpha) := \int_{-\infty}^{\alpha} \varphi(w)\,\mathrm{d}w \quad (17)$$

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* $\mathcal{C}_n$ *with* M_n *codewords such that* $\Pr(\mathcal{E}_2|\mathcal{C}_n)$ *satisfies*

$$\lim_{n\to\infty} \Pr(\mathcal{E}_1|\mathcal{C}_n) = \begin{cases} \Phi\left(\frac{b-a}{\sqrt{V_{\max}(W)}}\right) & \text{if } a \le 0 \\ \Phi\left(\frac{b-a}{\sqrt{V_{\min}(W)}}\right) & \text{if } a > 0. \end{cases}, \quad \text{and} \quad (18)$$

$$\liminf_{n\to\infty} -\frac{1}{\sqrt{n}} \log \Pr(\mathcal{E}_2|\mathcal{C}_n) \ge b. \quad (19)$$

The proof of this result can be found in Section IV-B. Observe that the first error probability is in the central limit regime [12]–[14] while the second scales as $\exp(-\sqrt{n}\,b)$, which is in the moderate deviations regime [16], [19]. Thus, we call this the *mixed regime*.

B. Tightness With Respect to the Ensemble Average

It is, at this point, not clear that the codes we proposed in Section III-A are asymptotically optimal. In this section, we demonstrate the tightness of the decoder for *additive* DMCs with uniform input distribution (which is a capacity-achieving input distribution for additive DMCs). We consider an ensemble evaluation of the two error probabilities. That is, we evaluate the probabilities of total and undetected errors averaged over the random code and show that this evaluation is tight in some asymptotic sense to be made precise in the statements. For brevity, we also call this evaluation *ensemble tightness* or *ensemble converse*. Similarly to (10), the sizes of the codes we consider $\{M_n\}_{n\in\mathbb{N}}$ take the form

$$\log M_n = nC - an^{1-t} \quad (20)$$

where $C = \log d - H(P)$ is the capacity of the additive channel and $0 < t \le 1/2$. 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 $\Pr(\mathcal{E}_1|\mathcal{C}_n)$ and $\Pr(\mathcal{E}_2|\mathcal{C}_n)$ corresponding to the total error probability and the undetected error probability respectively. We define the *varentropy* [29] or *source dispersion* [30] of the additive noise P as

$$V(P) := \sum_{z=0}^{d-1} P(z) \left[\log \frac{1}{P(z)} - H(P) \right]^2. \quad (21)$$

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 ϵ-dispersion [14] of W is $V(P)$ for every $\epsilon \in (0, 1)$, i.e., $V_{\min}(W) = V_{\max}(W) = V(P)$.

In the following, we emphasize that the uniform distribution will be chosen as the input distribution of the code. This is equivalent to choosing the M_n codewords where each codeword is drawn uniformly at random from $\{0, 1, \ldots, d-1\}^n$.

Theorem 3 (Moderate Deviations Regime Converse): 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 W be an additive DMC. When the expectation of the total error satisfies*

$$\liminf_{n\to\infty} -\frac{1}{n^{1-2t}} \log \mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_1|\mathcal{C}_n)\big] \ge \tfrac{(a-b)^2}{2V(P)}, \tag{22}$$

then the expectation of the undetected error satisfies

$$\limsup_{n\to\infty} -\frac{1}{n^{1-t}} \log \mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_2|\mathcal{C}_n)\big] \le b. \tag{23}$$

Conversely, when the expectation of the undetected error satisfies

$$\liminf_{n\to\infty} -\frac{1}{n^{1-t}} \log \mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_2|\mathcal{C}_n)\big] \ge b, \tag{24}$$

then the expectation of the total error satisfies

$$\limsup_{n\to\infty} -\frac{1}{n^{1-2t}} \log \mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_1|\mathcal{C}_n)\big] \le \tfrac{(a-b)^2}{2V(P)}. \tag{25}$$

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 $\mathcal{C}_n$ 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). *Let W be an additive DMC. When the expectation of the total error satisfies*

$$\limsup_{n\to\infty} \mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_1|\mathcal{C}_n)\big] \le \Phi\left(\tfrac{b-a}{\sqrt{V(P)}}\right) \tag{26}$$

then the expectation of the undetected error satisfies

$$\limsup_{n\to\infty} -\frac{1}{\sqrt{n}} \log \mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_2|\mathcal{C}_n)\big] \le b. \tag{27}$$

Conversely, when the expectation of the undetected error satisfies

$$\liminf_{n\to\infty} -\frac{1}{\sqrt{n}} \log \mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_2|\mathcal{C}_n)\big] \ge b, \tag{28}$$

then the expectation of the total error satisfies

$$\liminf_{n\to\infty} \mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_1|\mathcal{C}_n)\big] \ge \Phi\left(\tfrac{b-a}{\sqrt{V(P)}}\right). \tag{29}$$

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 asymptotic 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 $\mathcal{C}_n := \{\mathbf{x}_1, \ldots, \mathbf{x}_{M_n}\}$, the Pareto-optimal decoding region for each message $m \in \{1, \ldots, M_n\}$ is given by

$$\mathcal{D}_m := \left\{\mathbf{y} : \frac{W^n(\mathbf{y}|\mathbf{x}_m)}{\sum_{m'\neq m} W^n(\mathbf{y}|\mathbf{x}_{m'})} \ge \exp(nT_n)\right\}, \tag{30}$$

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 a generalization of the Neyman-Pearson lemma. Because $T_n > 0$, the regions are disjoint. We let $\mathcal{D}_0$ denote the set of all $\mathbf{y}$ that leads to an erasure, i.e.,

$$\mathcal{D}_0 := \mathcal{Y}^n \setminus \bigsqcup_{m=1}^{M_n} \mathcal{D}_m. \tag{31}$$

In the literature on decoding with an erasure option (e.g., [1]–[9]), 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*

$$T_n := \frac{b}{n^t}, \tag{32}$$

Let W be an additive DMC. Then, the expectation of the two error probabilities satisfy

$$\lim_{n\to\infty} -\frac{1}{n^{1-2t}} \log \mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_1|\mathcal{C}_n)\big] = \frac{(a-b)^2}{2V(P)}, \quad \textit{and} \tag{33}$$

$$bn^{1-t} + \frac{(a-b)^2}{2V(P)} n^{1-2t} + o(n^{1-2t}) \le -\log \mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_2|\mathcal{C}_n)\big] \le bn^{1-t} + o(n^{1-t}) \tag{34}$$

The proof of this lemma is provided in Section IV-C. From this lemma, we can show Theorem 3 by a simple argument which we defer to Section IV-D. At this point, a few other comments concerning are in order.

This result again corresponds to the so-called moderate deviations regime in channel coding considered by Altuğ and Wagner [16] and Polyanskiy and Verdú [19]. Thus, the appearance of the varentropy term $V(P)$ in the results is natural. The total and undetected error probabilities in (33) and (34) can be written as

$$\mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_1|\mathcal{C}_n)\big] = \exp\left(-\frac{(a-b)^2}{2V(P)} n^{1-2t} + o(n^{1-2t})\right), \quad \text{and} \tag{35}$$

$$\mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_2|\mathcal{C}_n)\big] = \exp\big(-bn^{1-t} + o(n^{1-t})\big). \tag{36}$$

respectively. This scaling is also different from those found in the literature which primarily focus on exponentially decaying probabilities [1]–[9] or non-vanishing error probabilities [18]. 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 [23, Th. 2.3.6]. The statement of this modified form of the Gärtner-Ellis theorem is presented as Theorem 8 in Appendix A and we provide a self-contained proof therein. Similarly to the work by Somekh-Baruch and Merhav [9], the two probabilities in (33)–(34) are asymptotic *equalities* (if we consider the normalizations 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 for the higher-order term).

Next, observe that the undetected error decays faster than the total error because the former is 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 $\Pr(\mathcal{E}_1|\mathcal{C}_n)$ 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 $\frac{(a-b)^2}{2V(P)}$ 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 T_n is negative (a case not allowed by Lemma 5). 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, $\Pr(\mathcal{E}_2|\mathcal{C}_n)$ no longer corresponds to the probability of undetected error. Rather, the expression for $\Pr(\mathcal{E}_2|\mathcal{C}_n)$ in (5) corresponds to the average number of incorrect codewords in the list corresponding to the *overlapping* (non-disjoint) decision regions $\{\mathcal{D}_m\}_{m=1}^{M_n}$.

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 $\mathcal{C}_n$ 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). *Let W be an additive DMC. Then, the expectation of the two error probabilities satisfy*

$$\lim_{n\to\infty} \mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_1|\mathcal{C}_n)\big] = \Phi\left(\frac{b-a}{\sqrt{V(P)}}\right) \quad \textit{and} \tag{37}$$

$$\begin{aligned} b\sqrt{n} + \frac{(a-b)^2}{2V(P)} + o(1) &\le -\log \mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_2|\mathcal{C}_n)\big] \\ &\le b\sqrt{n} + o(\sqrt{n}). \end{aligned} \tag{38}$$

The proof of this lemma is provided in Section IV-E. It is largely similar to that for Lemma 5 but for the total error probability in (37), instead of invoking the Gärtner-Ellis theorem [23, Th. 2.3.6], we use the fact that if the cumulant generating function of a sequence of random variables $\{K_n\}_{n\in\mathbb{N}}$ converges to a quadratic function, $\{K_n\}_{n\in\mathbb{N}}$ 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 Mukherjea *et al.* [31, Th. 2] (building on initial work by Curtiss [32]) 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.

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

$$\mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_2|\mathcal{C}_n)\big] = \exp\big(-b\sqrt{n} + o(\sqrt{n})\big). \tag{39}$$

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 [18, Th. 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.

In the same way as we can show Theorem 3 from Lemma 5, we can also use the exact same argument to show Theorem 4 from Lemma 6. Thus, we omit the details here.

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 $\mathbf{x}_m$, $m \in \{1, \ldots, M_n\}$ with the product distribution $P_X^n \in \mathcal{P}(\mathcal{X}^n)$. The expectation over this random choice of codebook is denoted as $\mathbb{E}_{\mathcal{C}_n}[\cdot]$. Now, we first consider $\Pr(\mathcal{E}_1|\mathcal{C}_n)$. Define the (capacity-achieving) output distribution $P_Y := P_X W$ and its n-fold memoryless extension P_Y^n. Next, we consider regions $\tilde{\mathcal{D}}_m$ defined in (13). The expectation over the code of the $W^n(\cdot|X_{m'}^n)$-probability of $\tilde{\mathcal{D}}_m$ can be evaluated as

$$\begin{aligned} &\mathbb{E}_{\mathcal{C}_n}\left[\sum_{\mathbf{y}} W^n(\mathbf{y}|X_{m'}^n)\mathbf{1}\left\{W^n(\mathbf{y}|X_m^n) \ge M_n \exp(bn^{1-t}) P_Y^n(\mathbf{y})\right\}\right] \\ &= \mathbb{E}_{X_m^n}\left[\sum_{\mathbf{y}} P_Y^n(\mathbf{y})\mathbf{1}\left\{W^n(\mathbf{y}|X_m^n) \ge M_n \exp(bn^{1-t}) P_Y^n(\mathbf{y})\right\}\right] \end{aligned} \tag{40}$$

$$\begin{aligned} &\le \mathbb{E}_{X_m^n}\Bigg[\sum_{\mathbf{y}} M_n^{-1}\exp(-bn^{1-t}) W^n(\mathbf{y}|X_m^n) \\ &\qquad \times \mathbf{1}\left\{W^n(\mathbf{y}|X_m^n) \ge M_n \exp(bn^{1-t}) P_Y^n(\mathbf{y})\right\}\Bigg] \end{aligned} \tag{41}$$

$$\le M_n^{-1}\exp(-bn^{1-t}) \tag{42}$$

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

$$\mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_2|\mathcal{C}_n)\big]$$
$$\leq \mathbb{E}_{\mathcal{C}_n}\left[\frac{1}{M_n}\sum_{m=1}^{M_n}\sum_{\mathbf{y}}\sum_{m'\neq m} W^n(\mathbf{y}|X_{m'}^n) \times \mathbf{1}\left\{W^n(\mathbf{y}|X_m^n) \geq M_n \exp(bn^{1-t}) P_Y^n(\mathbf{y})\right\}\right] \tag{43}$$
$$\leq \frac{1}{M_n}\sum_{m=1}^{M_n}\sum_{m'\neq m} M_n^{-1}\exp(-bn^{1-t}) \tag{44}$$
$$= \frac{M_n-1}{M_n}\exp(-bn^{1-t}) \tag{45}$$
$$\leq \exp(-bn^{1-t}). \tag{46}$$

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

$$\hat{\mathcal{D}}_m^c = \tilde{\mathcal{D}}_m^c \cup \hat{\mathcal{D}}_0 = \tilde{\mathcal{D}}_m^c \cup \bigcup_{m'\neq m}\left(\tilde{\mathcal{D}}_m \cap \tilde{\mathcal{D}}_{m'}\right) \subset \tilde{\mathcal{D}}_m^c \cup \bigcup_{m'\neq m}\tilde{\mathcal{D}}_{m'}. \tag{47}$$

The expectation of the total error probability over the random codebook can be written as

$$\mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_1|\mathcal{C}_n)\big]$$
$$= \mathbb{E}_{\mathcal{C}_n}\left[\frac{1}{M_n}\sum_{m=1}^{M_n}\sum_{\mathbf{y}\in\hat{\mathcal{D}}_m^c} W^n(\mathbf{y}|X_m^n)\right] \tag{48}$$
$$\leq \mathbb{E}_{X_m^n}\left[\frac{1}{M_n}\sum_{m=1}^{M_n}\sum_{\mathbf{y}} W^n(\mathbf{y}|X_m^n)\mathbf{1}\{\mathbf{y}\in\tilde{\mathcal{D}}_m^c\}\right] + \mathbb{E}_{\mathcal{C}_n}\left[\frac{1}{M_n}\sum_{m=1}^{M_n}\sum_{\mathbf{y}}\sum_{m'\neq m} W^n(\mathbf{y}|X_m^n)\mathbf{1}\{\mathbf{y}\in\tilde{\mathcal{D}}_{m'}\}\right] \tag{49}$$
$$\leq \mathbb{E}_{X_m^n}\left[\frac{1}{M_n}\sum_{m=1}^{M_n}\sum_{\mathbf{y}} W^n(\mathbf{y}|X_m^n)\mathbf{1}\{\mathbf{y}\in\tilde{\mathcal{D}}_m^c\}\right] + \exp(-bn^{1-t}), \tag{50}$$
$$= \sum_{\mathbf{x},\mathbf{y}} P_X^n(\mathbf{x})W^n(\mathbf{y}|\mathbf{x})\mathbf{1}\left\{\log\frac{W^n(\mathbf{y}|\mathbf{x})}{P_Y^n(\mathbf{y})} - nC < -(a-b)n^{1-t}\right\} + \exp(-bn^{1-t}), \tag{51}$$

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

$$\mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_1|\mathcal{C}_n)\big] \geq \sum_{\mathbf{x},\mathbf{y}} P_X^n(\mathbf{x})W^n(\mathbf{y}|\mathbf{x}) \times \mathbf{1}\left\{\log\frac{W^n(\mathbf{y}|\mathbf{x})}{P_Y^n(\mathbf{y})} - nC < -(a-b)n^{1-t}\right\}. \tag{52}$$

Recall that $a > b$. By the moderate deviations theorem [23, Th. 3.7.1], the sums on the right-hand-sides of (51) and (52) behave as

$$\exp\left(-n^{1-2t}\frac{(a-b)^2}{2U(P_X, W)} + o(n^{1-2t})\right), \tag{53}$$

which is much larger than (i.e., dominates) the second term in (51), namely $\exp(-bn^{1-t})$. Since $U(P_X, W) = V_{\min}(W)$ [14, Lemma 62], we have the asymptotic equality in (11).

To derandomize the code, fix $\theta \in (0, 1)$. By employing Markov's inequality to (46) and (53) (cf. [18, Proof of Theorem 1]), we obtain

$$\Pr\left(\Pr(\mathcal{E}_1|\mathcal{C}_n) > \frac{1}{\theta}\mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_1|\mathcal{C}_n)\big]\right) < \theta, \quad \text{and} \tag{54}$$
$$\Pr\left(\Pr(\mathcal{E}_2|\mathcal{C}_n) > \frac{1}{1-\theta}\mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_2|\mathcal{C}_n)\big]\right) < 1-\theta. \tag{55}$$

Thus,

$$\Pr\left(\Pr(\mathcal{E}_1|\mathcal{C}_n) > \frac{1}{\theta}\mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_1|\mathcal{C}_n)\big] \text{ or } \Pr(\mathcal{E}_2|\mathcal{C}_n) > \frac{1}{1-\theta}\mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_2|\mathcal{C}_n)\big]\right) < 1. \tag{56}$$

Thus, by taking $\theta = \frac{1}{2}$, there exists a deterministic code satisfying

$$\Pr(\mathcal{E}_1|\mathcal{C}_n) \leq 2\exp\left(-n^{1-2t}\frac{(a-b)^2}{2U(P_X, W)} + o(n^{1-2t})\right), \quad \text{and} \tag{57}$$
$$\Pr(\mathcal{E}_2|\mathcal{C}_n) \leq 2\exp(-bn^{1-t}). \tag{58}$$

This completes the proof.

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 [33, Sec. XVI.7], the right-hand-sides of (51) and (52) behave as

$$\Phi\left(\frac{b-a}{\sqrt{V_{\max}(W)}}\right) + O\left(\frac{1}{\sqrt{n}}\right). \tag{59}$$

Thus, by the same Markov inequality argument to derandomize the code as above, for any sequence $\{\theta_n\}_{n\in\mathbb{N}} \subset (0, 1)$, there exists a sequence of deterministic codes $\mathcal{C}_n$ satisfying

$$\Pr(\mathcal{E}_1|\mathcal{C}_n) \leq \frac{1}{\theta_n}\left[\Phi\left(\frac{b-a}{\sqrt{V_{\max}(W)}}\right) + O\left(\frac{1}{\sqrt{n}}\right)\right] \quad \text{and} \tag{60}$$
$$\Pr(\mathcal{E}_2|\mathcal{C}_n) \leq \frac{1}{1-\theta_n}\exp(-\sqrt{n}b). \tag{61}$$

Choose $\theta_n := 1 - 1/n$ to complete the proof of the theorem for $a \le 0$. 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 $\mathbf{x}_m$, $m \in \{1, \ldots, M_n\}$ uniformly at random from $\{0, 1, \ldots, d-1\}^n$. Indeed, a capacity-achieving input distribution of the additive channel is the uniform distribution on $\{0, 1, \ldots, d-1\}$. As above, the expectation over this random choice of codebook is denoted as $\mathbb{E}_{\mathcal{C}_n}[\cdot]$. Now, we first consider $\Pr(\mathcal{E}_1|\mathcal{C}_n)$. From the definition in (4), the expectation of the error probability over the random codebook can be written as

$$\begin{aligned} &\mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_1|\mathcal{C}_n)\big] \\ &= \mathbb{E}_{\mathcal{C}}\left[\frac{1}{M_n}\sum_{m=1}^{M_n}\sum_{\mathbf{y}} W^n(\mathbf{y}|X_m^n) \right.\\ &\qquad \left.\times \mathbf{1}\left\{\frac{\sum_{m'\neq m} W^n(\mathbf{y}|X_{m'}^n)}{W^n(\mathbf{y}|X_m^n)} \ge \exp(-nT_n)\right\}\right] \quad (62)\\ &= \mathbb{E}_{\mathcal{C}_n}\left[\frac{1}{M_n}\sum_{m=1}^{M_n}\Pr\left(\log\left(\sum_{m'\neq m} W^n(Y^n|X_{m'}^n)\right)\right.\right.\\ &\qquad \left.\left. - \log W^n(Y^n|X_m^n) \ge -nT_n \middle| \mathcal{C}_n\right)\right] \quad (63)\\ &= \frac{1}{M_n}\sum_{m=1}^{M_n}\Pr\left(\log\left(\sum_{m'\neq m} W^n(Y^n|X_{m'}^n)\right)\right.\\ &\qquad \left. - \log W^n(Y^n|X_m^n) \ge -bn^{1-t}\right) \quad (64)\end{aligned}$$

In (63), the inner probability is over $Y^n \sim W^n(\cdot|\mathbf{x}_m)$ for a fixed code $\mathcal{C}_n$ and in (64), the probability is over both the random codebook $\mathcal{C}_n$ 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

$$F_n := \log\left(\sum_{m'\neq m} W^n(Y^n|X_{m'}^n)\right) - \log W^n(Y^n|X_m^n) \quad (65)$$

for any $m \in \{1, \ldots, M_n\}$, say $m = 1$. In particular, to estimate the probability $\Pr(F_n \ge -bn^{1-t})$ in (64), it suffices to estimate the cumulant generating function of F_n. We denote the cumulant generating function as

$$\begin{aligned}\phi_n(s) &:= \log \mathbb{E}\big[\exp(sF_n)\big] \quad (66)\\ &= \log \mathbb{E}_{\mathcal{C}_n}\left[\sum_{\mathbf{y}} W^n(\mathbf{y}|X_m^n)^{1-s}\left(\sum_{m'\neq m} W^n(\mathbf{y}|X_{m'}^n)\right)^s\right] \quad (67)\\ &= \log \sum_{\mathbf{y}} \mathbb{E}_{\mathcal{C}_n}\left[W^n(\mathbf{y}|X_m^n)^{1-s}\right]\\ &\qquad \times \mathbb{E}_{\mathcal{C}_n}\left[\left(\sum_{m'\neq m} W^n(\mathbf{y}|X_{m'}^n)\right)^s\right]. \quad (68)\end{aligned}$$

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

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*

$$\phi_n\left(\frac{u}{n^t}\right) = \left(-au + u^2\frac{V(P)}{2}\right)n^{1-2t} + O(n^{1-3t}) + o(1) \quad (69)$$

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, to (64) with the identifications

$$\alpha_n \equiv 0 \quad (70)$$
$$\beta_n \equiv n^{1-t}, \quad \text{and} \quad (71)$$
$$\gamma_n \equiv n^{-t} \quad (72)$$

Now, we can also make the additional identifications

$$\nu_1 \equiv 0 \quad (73)$$
$$X_n \equiv -F_n \quad (74)$$
$$p_n(\cdot) \equiv \frac{1}{M_n}\sum_{m=1}^{M_n}\Pr(\cdot), \quad (75)$$
$$\mu_n(\theta) \equiv \phi_n(-\theta), \quad (76)$$
$$\theta_0 \equiv 0, \quad \text{and} \quad (77)$$
$$x \equiv b. \quad (78)$$

Now, applying (69) of Lemma 7 to the case with $u \equiv -y$ (with $y < 0$), we have

$$\nu_2(y) = \lim_{n\to\infty}\nu_{2,n}(y) = ya + y^2\frac{V(P)}{2}. \quad (79)$$

Then, defining y_0 to the (unique) real number satisfying $\nu_2'(y_0) = x$, we have

$$a + y_0V(P) = b \iff y_0 = \frac{b-a}{V(P)}. \quad (80)$$

which implies by simple algebra that

$$\begin{aligned}y_0x - \nu_2(y_0) &= \left(\frac{b-a}{V(P)}\right)b - \left[\left(\frac{b-a}{V(P)}\right)a + \left(\frac{b-a}{V(P)}\right)^2\frac{V(P)}{2}\right]\\ &= \frac{(b-a)^2}{2V(P)}. \quad (81)\end{aligned}$$

Because $a > b$, y_0 is negative. We can also verify that all the conditions of Case (ii) in Theorem 8 ($\theta_0 = 0, y_0 < 0, \alpha_n = \nu_1 = 0, \beta_n\gamma_n \to \infty$) are satisfied so we can readily apply it here. Thus,

$$\begin{aligned}-\log \mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_1|\mathcal{C}_n)\big] &= -\log\Pr\left(F_n > -bn^{1-t}\right)\\ &= \frac{(a-b)^2}{2V(P)}n^{1-2t} + o(n^{1-2t}), \quad (82)\end{aligned}$$

which implies (33).

Now we estimate $\mathbb{E}_{\mathcal{C}_n}[\Pr(\mathcal{E}_2|\mathcal{C}_n)]$. Using the same calculations that led to (64), one finds that

$$\mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_2|\mathcal{C}_n)\big] = \mathbb{E}_{\mathcal{C}_n}\left[\frac{1}{M_n}\sum_{m=1}^{M_n}\sum_{\mathbf{y}}\sum_{m'\neq m} W^n(\mathbf{y}|X_{m'}^n) \times \mathbf{1}\left\{\frac{\sum_{m'\neq m} W^n(\mathbf{y}|X_{m'}^n)}{W^n(\mathbf{y}|X_m^n)} < \exp(-nT_n)\right\}\right] \tag{83}$$

$$= \mathbb{E}_{\mathcal{C}_n}\left[\frac{1}{M_n}\sum_{m=1}^{M_n}\mathbb{Q}\left(\left\{\mathbf{y} : \log\sum_{m'\neq m} W^n(\mathbf{y}|X_{m'}^n) - \log W^n(\mathbf{y}|X_m^n) < -bn^{1-t}\right\}\middle|\mathcal{C}_n\right)\right] \tag{84}$$

where in (84), we defined the (unnormalized) conditional measure $\mathbb{Q}(\mathcal{A}|\mathcal{C}_n = \{\mathbf{x}_m\}_{m=1}^{M_n}) := \sum_{m'\neq m} W^n(\mathcal{A}|\mathbf{x}_{m'})$ where $\mathcal{A} \subset \mathcal{Y}^n$. Given $\mathbb{Q}$, we can define a *normalized probability* measure

$$\mathbb{Q}'(\mathcal{A}|\mathcal{C}_n) := \frac{\mathbb{Q}(\mathcal{A}|\mathcal{C}_n)}{M_n - 1}. \tag{85}$$

Since the form of (84) is similar to the starting point for the calculation of $\mathbb{E}_{\mathcal{C}_n}[\Pr(\mathcal{E}_1|\mathcal{C}_n)]$ in (64), we may estimate $\mathbb{E}_{\mathcal{C}_n}[\Pr(\mathcal{E}_2|\mathcal{C}_n)]$ using similar steps to the above. Define another probability measure $\mathbb{P}(\mathcal{A}|\mathcal{C}_n = \{\mathbf{x}_m\}_{m=1}^{M_n}) := W^n(\mathcal{A}|\mathbf{x}_m)$. Note by the definition of F_n in (65), and the measures above that for all $\mathcal{A} \subset \mathcal{Y}^n$,

$$\exp(F_n) = \frac{\mathbb{Q}'(\{Y^n\}|\mathcal{C}_n)}{\mathbb{P}(\{Y^n\}|\mathcal{C}_n)} \cdot (M_n - 1). \tag{86}$$

Observe that the random variable involved in (84), namely $\log\sum_{m'\neq m} W^n(Y^n|X_{m'}^n) - \log W^n(Y^n|X_m^n)$, is exactly F_n defined in (65) where Y^n now has conditional law $\mathbb{Q}(\cdot|\mathcal{C}_n)$ instead of $\mathbb{P}(\cdot|\mathcal{C}_n)$. The cumulant generating function of F_n under the probability measure $\mathbb{Q}'$ is

$$\lambda_n(s) := \log\mathbb{E}_{\mathcal{C}_n,\mathbb{Q}'}\big[\exp(sF_n)\big] \tag{87}$$

$$= \log\left(\frac{\mathbb{E}_{\mathcal{C}_n,\mathbb{P}}\big[\exp((1+s)F_n)\big]}{M_n - 1}\right) \tag{88}$$

$$= \phi_n(1+s) - \log(M_n - 1) \tag{89}$$

where (88) follows from (86) and (89) from the definition of $\phi_n(s)$ in (66). Now, we apply Case (ia) of Theorem 8 in Appendix A with the identifications

$$\alpha_n \equiv -\log(M_n - 1), \tag{90}$$

$$\beta_n \equiv n^{1-t}, \quad \text{and} \tag{91}$$

$$\gamma_n \equiv n^{-t}. \tag{92}$$

Furthermore, from (84) and (89), one can also make the additional identifications

$$X_n \equiv F_n, \tag{93}$$

$$p_n(\cdot) \equiv \mathbb{E}_{\mathcal{C}_n}[\mathbb{Q}'(\cdot|\mathcal{C}_n)], \tag{94}$$

$$\mu_n(\theta) \equiv \lambda_n(\theta), \tag{95}$$

$$\theta_0 \equiv -1, \tag{96}$$

$$\nu_1 \equiv 0, \quad \text{and} \tag{97}$$

$$x \equiv -b. \tag{98}$$

Then we have

$$\nu_{2,n}(y) \equiv \lambda_n(-1 + yn^{-t}) + \log(M_n - 1) = n^{2t-1}\phi_n\left(\frac{y}{n^t}\right) \tag{99}$$

Thus, using (69),

$$\nu_2(y) = \lim_{n\to\infty}\nu_{2,n}(y) = -ya + y^2\frac{V(P)}{2}. \tag{100}$$

So, we have $\theta_0 x - \nu_1 = b$ and

$$y_0 = \frac{a-b}{V(P)}, \tag{101}$$

which is positive. This implies that

$$y_0 x - \nu_2(y_0) = \frac{(a-b)^2}{2V(P)}. \tag{102}$$

Thus, by the relation between $\mathbb{Q}$ to $\mathbb{Q}'$ in (85) and the bound in (170) (in Case (ia) of Theorem 8), we obtain

$$-\log\mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_2|\mathcal{C}_n)\big]$$

$$= -\log\mathbb{E}_{\mathcal{C}_n}\left[\mathbb{Q}\left(F_n < -bn^{1-t}\middle|\mathcal{C}_n\right)\right] \tag{103}$$

$$= -\log\mathbb{E}_{\mathcal{C}_n}\left[\mathbb{Q}'\left(F_n < -bn^{1-t}\middle|\mathcal{C}_n\right)\right] - \log(M_n - 1) \tag{104}$$

$$\leq bn^{1-t} + o(n^{1-t}), \tag{105}$$

which implies the upper bound in (34). The lower bound in (34) follows by invoking (169), from which we obtain

$$-\log\mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_2|\mathcal{C}_n)\big] \geq bn^{1-t} + \frac{(a-b)^2}{2V(P)}n^{1-2t} + o(n^{1-2t}). \tag{106}$$

This completes the proof of Lemma 5. ■

Remark 1: Observe that to evaluate the probabilities in (82) *and* (103)*, we employed Theorem 8 in Appendix A, which is a modified ("shifted") version of the usual Gärtner-Ellis theorem [23, Th. 2.3.6]. Theorem 8 assumes a sequence of random variables X_n has cumulant generating functions $\mu_n(\theta)$ that additionally satisfy the expansion $\mu_n(\theta_0 + \gamma_n y) = \alpha_n + \beta_n\nu_1 + \beta_n\gamma_n\nu_{2,n}(y)$ for some vanishing sequence γ_n. This generalization and the application to the erasure problem appears to the authors to be novel. In particular, since $\mathbb{Q}$ in* (84) *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 $\mathbb{Q}'$ as in* (85)*. 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 α_n (which is the offset) but instead the second-order term $-(\theta_0 x - \nu_1)\beta_n = -bn^{1-t}$ leading to* (105)–(106)*.*
3) *Finally, Theorem 8 also allows us to obtain an additional term scaling as n^{1-2t} in* (106)*, but we cannot obtain the coefficient of the higher-order term scaling as n^{1-2t} in* (105)*.*

D. Proof of Theorem 3

Proof: First, observe that for (24) to be satisfied, i.e., that the undetected error probability decays as

$$\mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_2|\mathcal{C}_n)\big] \le \exp(-bn^{1-t} + o(n^{1-t})), \tag{107}$$

we need the threshold to be of the form

$$T_n \ge bn^{-t} + o(n^{-t}). \tag{108}$$

To show this formally, suppose, to the contrary,

$$T_n = b'n^{-t} + o(n^{-t}). \tag{109}$$

for some $0 < b' < b$. So the constant in front of n^{-t} is strictly smaller than b. Then by (34) in Lemma 5, and since the decoder is asymptotically optimal, the undetected error probability decays as

$$\mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_2|\mathcal{C}_n)\big] = \exp(-b'n^{1-t} + o(n^{1-t})), \tag{110}$$

an asymptotic equality. This is a contradiction. Hence, (108) must hold. Intuitively, the thresholds in Forney's test in (30) must be large enough so that the decoding regions corresponding to the messages are sufficiently small so that the undetected error probability decays at least as fast as in (107). Consequently, by the asymptotically tight result in (33), we know that (24) implies (25).

Conversely, to satisfy the condition (22), i.e., that the total error probability decays as

$$\mathbb{E}_{\mathcal{C}_n}\big[\Pr(\mathcal{E}_1|\mathcal{C}_n)\big] \le \exp\left(-\frac{(a-b)^2}{2V(P)}n^{1-2t} + o(n^{1-2t})\right), \tag{111}$$

we need to choose

$$T_n \le bn^{-t} + o(n^{-t}) \tag{112}$$

by the same argument as the above. This means that the thresholds must be small enough so that the decoding regions corresponding to the messages are sufficiently large. Hence, due to the asymptotically tight result in Lemma 5, we have (23). ■

E. 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

$$\lim_{n\to\infty} \phi_n\left(\frac{u}{\sqrt{n}}\right) = -ua + u^2\frac{V(P)}{2} \tag{113}$$

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

$$\lim_{n\to\infty} \mathbb{E}_{\mathcal{C}_n}[\Pr(\mathcal{E}_1|\mathcal{C}_n)] = \lim_{n\to\infty} \Pr\left(\frac{F_n}{\sqrt{n}} > -b\right) \tag{114}$$

$$= \int_{-b}^{\infty} \frac{1}{\sqrt{2\pi V(P)}} \exp\left(-\frac{(w+a)^2}{2V(P)}\right) \mathrm{d}w \tag{115}$$

$$= \Phi\left(\frac{b-a}{\sqrt{V(P)}}\right). \tag{116}$$

To calculate $\mathbb{E}_{\mathcal{C}_n}[\Pr(\mathcal{E}_2|\mathcal{C}_n)]$, we can adopt the same change of measure and Gärtner-Ellis arguments and employ Case (ib) of Theorem 8. We follow exactly the steps leading from (83) to (105) to assert that (38) is true. Note that in this situation, we take $\gamma_n \equiv n^{-1/2}$ and $\beta_n \equiv n^{1/2}$. To apply Case (ib) of Theorem 8, we verify that $\beta_n = \gamma_n^{-1}$ and ν_2, derived in (100), is indeed a quadratic function. ■

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

Proof: To estimate $\phi_n(s)$ in (68), we define

$$A := \mathbb{E}_{\mathcal{C}_n}\left[W^n(\mathbf{y}|X_m^n)^{1-s}\right] \quad \text{and} \tag{117}$$

$$B := \mathbb{E}_{\mathcal{C}_n}\left[\left(\sum_{m'\neq m} W^n(\mathbf{y}|X_{m'}^n)\right)^s\right]. \tag{118}$$

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

$$A = \mathbb{E}_{X_m^n}\left[P^n(\mathbf{y} - X_m^n)^{1-s}\right] \tag{119}$$

$$= \mathbb{E}_{X_m^n}\left[P^n(\tilde{X}_m^n)^{1-s}\right] \tag{120}$$

where the *shifted codewords*[2] are defined as $\tilde{X}_m^n := \mathbf{y} - X_m^n$. By using the product structure of P^n, we see that regardless of $\mathbf{y}$, the term A can be written as

$$A = \frac{1}{d^n}\exp\big(-n\psi(s)\big) \tag{121}$$

where

$$\psi(s) := -\log\sum_z P(z)^{1-s}. \tag{122}$$

This function is related to the Rényi entropy as follows: $s\psi(s) = -H_{1-s}(P)$ where $H_\alpha(P)$ is the usual Rényi entropy of order α (e.g., [26, Problem 1.15]). Now, for a fixed $u > 0$, we make the choice

$$s = \frac{u}{n^t}, \tag{123}$$

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

$$A = \frac{1}{d^n}\exp\left(n\left(s\,H(P) + s^2\,\frac{V(P)}{2} + O(s^3)\right)\right) \tag{124}$$

$$= \frac{1}{d^n}\exp\left(un^{1-t}\,H(P) + u^2n^{1-2t}\,\frac{V(P)}{2} + O(n^{1-3t})\right), \tag{125}$$

where (125) follows from the definition of s in (123).

Now we estimate B in (118). Define the random variable $N_{\mathcal{C}_n}(Q)$ which represents the number of shifted codewords excluding that indexed by m with type $Q \in \mathcal{P}_n(\mathcal{X})$, i.e., $N_{\mathcal{C}_n}(Q) := |\{m' \neq m : \mathrm{type}(\tilde{X}_{m'}^n) = Q\}|$. This plays

[2] The shifted codewords need not be codewords *per se*, so this is a slight abuse of terminology.

the role of the *type class enumerator* or *distance enumerator* in Merhav [6], [10]. Then, B can be written as

$$B = \mathbb{E}_{\mathcal{C}_n}\left[\left(\sum_{m'\neq m} P^n(\mathbf{y} - X^n_{m'})\right)^s\right] \tag{126}$$
$$= \mathbb{E}_{\mathcal{C}_n}\left[\left(\sum_{m'\neq m} P^n(\tilde{X}^n_{m'})\right)^s\right] \tag{127}$$
$$= \mathbb{E}_{\mathcal{C}_n}\left[\left(\sum_{Q\in\mathcal{P}_n(\mathcal{X})} N_{\mathcal{C}_n}(Q)\exp\left(-n[D(Q\|P)+H(Q)]\right)\right)^s\right]. \tag{128}$$

In (126), we again used the additivity of the channel and introduced the noise distribution P. In (127), we used the definition of the shifted codewords $\tilde{X}^n_m$. In (128), we introduced the type class enumerators $N_{\mathcal{C}_n}(Q)$. We also recall from [26, Lemma 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 expression in (128) is independent of $\mathbf{y}$, just as for the calculation of A in (125). 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:

$$B \geq \mathbb{E}_{\mathcal{C}_n}\left[\left(\max_{Q'\in\mathcal{P}_n(\mathcal{X})} N_{\mathcal{C}_n}(Q') \times \exp\left(-n[D(Q'\|P)+H(Q')]\right)\right)^s\right] \tag{129}$$
$$= \mathbb{E}_{\mathcal{C}_n}\left[\max_{Q'\in\mathcal{P}_n(\mathcal{X})} N_{\mathcal{C}_n}(Q')^s \exp\left(-ns[D(Q'\|P)+H(Q')]\right)\right] \tag{130}$$
$$\geq \max_{Q'\in\mathcal{P}_n(\mathcal{X})} \mathbb{E}_{\mathcal{C}_n}\left[N_{\mathcal{C}_n}(Q')^s\right]\exp\left(-ns[D(Q'\|P)+H(Q')]\right) \tag{131}$$
$$\geq \mathbb{E}_{\mathcal{C}_n}\left[N_{\mathcal{C}_n}(P_n)^s\right]\exp\left(-ns[D(P_n\|P)+H(P_n)]\right), \tag{132}$$

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

$$P_n \in \operatorname*{argmin}_{Q\in\mathcal{P}_n(\mathcal{X})} \{\|Q-P\|_1 : H(Q) \geq H(P) + 2an^{-t}\}. \tag{133}$$

Fannes inequality [26, Lemma 2.7] (uniform continuity of Shannon entropy) says that

$$|H(P) - H(Q)| \leq \|P-Q\|_1 \log\frac{\|P-Q\|_1}{|\mathcal{X}|} \tag{134}$$

if $\|P-Q\|_1 \leq \frac{1}{2}$. Since P_n must be an n-type,

$$H(P_n) = H(P) + 2an^{-t} + O(n^{-1}\log n). \tag{135}$$

Because $P(z) > 0$ for all $z \in \mathcal{X}$, one immediately finds that

$$D(P_n\|P) = O(\|P_n - P\|_1^2) = O(n^{-2t}), \tag{136}$$

which is negligible. Combining the above estimates, we obtain

$$-ns[D(P_n\|P) + H(P_n)] = -un^{1-t}H(P) - 2aun^{1-2t} + O(n^{1-3t}) \tag{137}$$

as n grows.

Next, apply Lemma 10 in Appendix C to the expectation (132) to the case with

$$L = M_n - 1, \tag{138}$$
$$M_1 = d^n, \tag{139}$$
$$M_2 = |\mathcal{T}^{(n)}_{P_n}|, \tag{140}$$
$$\{X_1,\ldots,X_L\} = \{X^n_{m'}\}_{m'\neq m}, \tag{141}$$
$$\mathcal{A} = \mathcal{T}^{(n)}_{P_n}, \quad \text{and,} \tag{142}$$
$$s = un^{-t} \tag{143}$$

and a fixed positive constant $\epsilon > 0$. We now perform a series of steps to bound the terms in (243). By a standard property of types [26] and the estimate in (135),

$$\log|\mathcal{T}^{(n)}_{P_n}| \geq nH(P_n) - (d-1)\log(n+1) \tag{144}$$
$$= nH(P) + 2an^{1-t} + O(\log n). \tag{145}$$

Thus, using the definition of L in (138), the definition M_1 in (139), and the number of codewords M_n in (20), we also have

$$\log L + \log M_2 - \log M_1 \geq an^{1-t} + O(\log n). \tag{146}$$

Consequently,

$$\log\left[1 - \exp\left(-\frac{LM_2}{2M_1}\epsilon^2\right)\right] = o(1). \tag{147}$$

Also, we have

$$s\log(1-\epsilon) = un^{-t}\log(1-\epsilon) = o(1). \tag{148}$$

and by (143),

$$s(\log L + \log M_2 - \log M_1) \geq aun^{1-2t} + o(1). \tag{149}$$

Therefore, Lemma 10 says that

$$\log\mathbb{E}_{\mathcal{C}_n}\left[N_{\mathcal{C}_n}(P_n)^s\right] = \log\mathbb{E}[N^s] \geq aun^{1-2t} + o(1). \tag{150}$$

Combining (132), (137) and (150), we find that

$$\log B \geq -n^{1-t}uH(P) - an^{1-2t}u + O(n^{1-3t}) + o(1). \tag{151}$$

Now, we proceed to upper bound B in (128). Note that we consider the case when $0 < s < 1$ because we substitute un^{-t} into s per (143). Consider,

$$B = \mathbb{E}_{\mathcal{C}_n}\left[\left(\sum_{Q'\in\mathcal{P}_n(\mathcal{X})} N_{\mathcal{C}_n}(Q')\exp\left(-n[D(Q'\|P)+H(Q')]\right)\right)^s\right] \tag{152}$$
$$\leq \left(\mathbb{E}_{\mathcal{C}_n}\left[\sum_{Q'\in\mathcal{P}_n(\mathcal{X})} N_{\mathcal{C}_n}(Q')\exp\left(-n[D(Q'\|P)+H(Q')]\right)\right]\right)^s \tag{153}$$
$$= \left(\sum_{Q'\in\mathcal{P}_n(\mathcal{X})} \mathbb{E}_{\mathcal{C}_n}\left[N_{\mathcal{C}_n}(Q')\right]\exp\left(-n[D(Q'\|P)+H(Q')]\right)\right)^s \tag{154}$$
$$= \left(\sum_{Q'\in\mathcal{P}_n(\mathcal{X})} \frac{M_n|\mathcal{T}^{(n)}_{Q'}|}{d^n}\exp\left(-n[D(Q'\|P)+H(Q')]\right)\right)^s \tag{155}$$

$$\leq \Bigg((n+1)^{d-1} \max_{Q' \in \mathcal{P}_n(\mathcal{X})} \frac{M_n |\mathcal{T}_{Q'}^{(n)}|}{d^n} \times \exp\big(-n[D(Q'\|P) + H(Q')]\big)\Bigg)^s \tag{156}$$

$$= (n+1)^{s(d-1)} \max_{Q' \in \mathcal{P}_n(\mathcal{X})} \left(\frac{M_n |\mathcal{T}_{Q'}^{(n)}|}{d^n}\right)^s \times \exp\big(-ns[D(Q'\|P) + H(Q')]\big) \tag{157}$$

$$\leq (n+1)^{s(d-1)} \max_{Q' \in \mathcal{P}_n(\mathcal{X})} \exp\big(-s[nH(P) - an^{1-t} - nH(Q')] - ns[D(Q'\|P) + H(Q')]\big) \tag{158}$$

$$= (n+1)^{s(d-1)} \max_{Q' \in \mathcal{P}_n(\mathcal{X})} \exp\big(-snH(P) - asn^{1-t} - nsD(Q'\|P)\big) \tag{159}$$

$$= (n+1)^{s(d-1)} \exp\Bigg(-snH(P) - asn^{1-t} + \max_{Q' \in \mathcal{P}(\mathcal{X})} -nsD(Q'\|P)\Bigg) \tag{160}$$

$$\leq (n+1)^{s(d-1)} \exp\big(-snH(P) - asn^{1-t}\big) \tag{161}$$

$$= (n+1)^{\frac{u(d-1)}{n^t}} \exp\big(-un^{1-t}H(P) - aun^{1-2t}\big) \tag{162}$$

where (153) follows from Jensen's inequality applied to the concave function $x \mapsto x^s$, (155) follows from the fact that

$$\mathbb{E}_{\mathcal{C}_n}\big[N_{\mathcal{C}_n}(Q')\big] = M_n \Pr\left(\mathbf{y} - X_1^n \in \mathcal{T}_{Q'}^{(n)}\right) = M_n \Pr\left(\tilde{X}_1^n \in \mathcal{T}_{Q'}^{(n)}\right) = \frac{M_n |\mathcal{T}_{Q'}^{(n)}|}{d^n} \tag{163}$$

and (158) follows from the choice of $\log M_n = n(\log d - H(P)) - an^{1-t}$ and the fact that $|\mathcal{T}_{Q'}^{(n)}| \leq \exp(nH(Q'))$. Thus, we find that

$$\log B \leq -n^{1-t}uH(P) - an^{1-2t}u + o(1). \tag{164}$$

Combining the evaluations of A and B together in (68), we see that the sum over $\mathbf{y}$ cancels the $1/d^n$ term in (125) and the first-order entropy terms also cancel. The final expression for the cumulant generating function of F_n satisfies (69) as desired. ∎

V. Conclusion 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 [20] and showed using an ensemble tightness argument that this simple decoding rule is, in fact, asymptotically optimal for additive DMCs with uniform input distribution. 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]–[9], 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 $b > 0$.

Possible extensions of this work include:

1) Removing the assumption that the DMC is additive for the ensemble tightness results 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]. This assumption seems necessary using our techniques to establish the asymptotics of the cumulant generating function in Lemma 7. It heavily relies on the fact that input distribution is uniform so that, by symmetry, the statistics of the codewords $\{X_m^n : m = 1, \ldots, M_n\}$ are the same as that of the shifted codewords $\{\mathbf{y} - X_m^n : m = 1, \ldots, M_n\}$ for any $\mathbf{y}$.
2) Extending the analysis to the list decoding case where $T_n = bn^{-t}$ where $b < 0$ and $t \in (0, 1/2]$. Our information spectrum-style decoding regions and subsequent analysis of their probabilities only works for the case $b > 0$. See the argument after (53). Hence it would be useful to develop alternative threshold decoders and more refined tools to analyze the list decoding setting.
3) Tightening the higher-order asymptotics for the expansions of the log-probabilities in (34) and (38). This would be interesting from a mathematical standpoint. However, 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]–[10] seems to be necessary for this purpose.

Appendix A
Modified Gärtner-Ellis Theorem

Here we present and prove a modified form of the Gärtner-Ellis theorem with a shift and a general order (normalization).

Some of the ideas (for example the proof of (170) for the case $\beta_n\gamma_n \to \infty$ and (171)) are contained in [23, Th. 2.3.6] but other elements of the proof are novel. To keep the exposition self-contained, we provide all the details of the proof for the event of interest $\{X_n \leq x\beta_n\}$. The standard Gärtner-Ellis theorem [23, Th. 2.3.6] applies in full generality to open and closed sets but here we are only interested in events of the form $\{X_n \leq x\beta_n\}$.

Theorem 8 (Modified Gärtner-Ellis Theorem): We consider three sequences $\alpha_n, \beta_n, \gamma_n$. The sequence α_n is arbitrary and β_n and γ_n are positive sequences that additionally satisfy $\beta_n \to \infty$ and $\gamma_n \to 0$. Let p_n be a sequence of distributions, and X_n be the sequence of random variables with distribution p_n. Define the cumulant generating function

$$\mu_n(\theta) := \log \mathbb{E}_{p_n}[\exp(\theta X_n)]. \tag{165}$$

Let $\theta_0 \leq 0$ and ν_1 be constants. Assume that

$$\mu_n(\theta_0 + \gamma_n y) = \alpha_n + \beta_n\nu_1 + \beta_n\gamma_n\nu_{2,n}(y) \tag{166}$$

for some sequence of functions $\nu_{2,n}$. Assume that

$$\nu_{2,n}(y) \to \nu_2(y) \quad \text{pointwise,} \tag{167}$$

and the limiting function $\nu_2(y)$ satisfies

1) $\nu_2(0) = 0$;
2) $\nu_2(y)$ *is strictly convex;*
3) $\nu_2(y)$ *is C^2-continuous on an open subset $G \subset \mathbb{R}$*

We also fix x and the real number $y_0 \in G$ satisfying[3]

$$\nu_2'(y_0) = x. \tag{168}$$

(i) *When $\theta_0 < 0$, we consider two subcases:*
 (a) $\beta_n \gamma_n \to \infty$;
 (b) $\beta_n = \gamma_n^{-1}$ *and ν_2 is a quadratic function.*[4]

 In both cases, we have the lower bound

$$-\log p_n\left\{\frac{X_n}{\beta_n} \le x\right\} \ge -\alpha_n + (\theta_0 x - \nu_1)\beta_n + (y_0 x - \nu_2(y_0))\beta_n\gamma_n + o(\beta_n\gamma_n) \tag{169}$$

 and the upper bound

$$-\log p_n\left\{\frac{X_n}{\beta_n} \le x\right\} \le -\alpha_n + (\theta_0 x - \nu_1)\beta_n + o(\beta_n). \tag{170}$$

(ii) *When $\theta_0 = 0$, $y_0 < 0$, $\alpha_n = \nu_1 = 0$, and $\beta_n\gamma_n \to \infty$,*

$$-\log p_n\left\{\frac{X_n}{\beta_n} \le x\right\} = (y_0 x - \nu_2(y_0))\beta_n\gamma_n + o(\beta_n\gamma_n). \tag{171}$$

Note that in Case (i), y_0 may take any real value. However, in Case (ii), y_0 always takes a negative value.

The proofs of these statements are split into four parts. In Appendix A-A, we prove the lower bounds to all statements. In Appendix A-B, we prove the upper bound for Case (ia) with $\beta_n\gamma_n \to \infty$. In Appendix A-C, we prove the upper bound for Case (ib) with $\beta_n = \gamma_n^{-1}$ and ν_2 is a quadratic. In Appendix A-D, we prove the upper bound for Case (ii).

A. Proof of Lower Bounds of Both Cases in (169) and (171)

Proof: The proofs of the lower bounds all cases are common.

First note that for Case (i), $\theta_0 < 0$ and $\gamma_n \to 0$, so for sufficiently large n, we have $\theta_0 + \gamma_n y_0 < 0$. For Case (ii), even though $\theta_0 = 0$, $y_0 < 0$ so similarly, we have $\theta_0 + \gamma_n y_0 < 0$. Thus, using Markov's inequality,

$$p_n\left\{\frac{X_n}{\beta_n} \le x\right\} = p_n\left\{\exp\left[\left(\frac{X_n}{\beta_n} - x\right)\beta_n(\theta_0 + \gamma_n y_0)\right] \ge 1\right\} \tag{172}$$

$$\le \mathbb{E}_{p_n}\left\{\exp\left[\left(\frac{X_n}{\beta_n} - x\right)\beta_n(\theta_0 + \gamma_n y_0)\right]\right\}. \tag{173}$$

[3]Such a real number y_0 is guaranteed to exist because ν_2 is strictly convex so ν_2' is strictly increasing.

[4]Since ν_2 is strictly convex and quadratic in this case, it must be of the form $\nu_2(y) = \varrho_0 + \varrho_1 y + \varrho_2 y^2$ for some constants $\varrho_2 > 0$ and $\varrho_0, \varrho_1 \in \mathbb{R}$.

In other words,

$$-\log p_n\left\{\frac{X_n}{\beta_n} \le x\right\} \ge x\beta_n\theta_0 + x\gamma_n\beta_n y_0 - \mu_n(\theta_0 + \gamma_n y_0) \tag{174}$$

$$= -\alpha_n + (\theta_0 x - \nu_1)\beta_n + (y_0 x - \nu_{2,n}(y_0))\beta_n\gamma_n \tag{175}$$

where (175) follows from the expansion of $\mu_n(\theta_0 + \gamma_n y)$ in (166). So from the assumption that $\nu_{2,n}(y_0) \to \nu_2(y_0)$ (cf. (167)), we obtain

$$-\log p_n\left\{\frac{X_n}{\beta_n} \le x\right\} \ge -\alpha_n + (\theta_0 x - \nu_1)\beta_n + (y_0 x - \nu_2(y_0))\beta_n\gamma_n + o(\beta_n\gamma_n) \tag{176}$$

as desired. This completes the proofs of the lower bounds in (169) and (171). ■

B. Proof of Upper Bound of Case (ia) in (170), i.e., $\beta_n\gamma_n \to \infty$

Proof: The proof of this upper bound proceeds in three distinct steps.

Step 1 (Measure Tilting): First fix a constant $\delta > 0$. For the sake of brevity, define the set

$$\mathcal{D}_{x,\delta} := \left\{\omega : x - 2\delta \le \frac{X_n(\omega)}{\beta_n} \le x\right\}. \tag{177}$$

It suffices to lower bound the p_n-probability of the set $\mathcal{D}_{x,\delta}$ because $\{\omega : X_n(\omega)/\beta_n \le x\} \supset \mathcal{D}_{x,\delta}$. Next given the constant $\delta > 0$, we can define the point

$$x' := x - \delta. \tag{178}$$

Correspondingly, also define the point y' such that

$$\nu_2'(y') = x'. \tag{179}$$

Let θ be defined as

$$\theta := \theta_0 + \gamma_n y', \tag{180}$$

where θ_0 and γ_n are fixed in the theorem statement. Define the tilted probability measure

$$\tilde{p}_{n,\theta}(\omega) := p_n(\omega)\exp(\theta X_n(\omega) - \mu_n(\theta)). \tag{181}$$

Now, for all n large enough $\theta < 0$ since $\theta_0 < 0$ and $\gamma_n \to 0$. Thus, from the definition of the tilted measure $\tilde{p}_{n,\theta}(\omega)$, for those $\omega \in \mathcal{D}_{x,\delta}$, we have

$$\tilde{p}_{n,\theta}(\omega) \le p_n(\omega)\exp\left(\theta\beta_n(x - 2\delta) - \mu_n(\theta)\right). \tag{182}$$

Integrating over all $\omega \in \mathcal{D}_{x,\delta}$, taking the logarithm and normalizing by β_n we obtain

$$\frac{1}{\beta_n}\log\tilde{p}_{n,\theta}\{\mathcal{D}_{x,\delta}\} \le \frac{1}{\beta_n}\log p_n\{\mathcal{D}_{x,\delta}\} + \theta(x - 2\delta) - \frac{\mu_n(\theta)}{\beta_n}. \tag{183}$$

Substituting the definition of θ in (180) into the above, we obtain

$$\frac{1}{\beta_n}\log\tilde{p}_{n,\theta_0+\gamma_n y'}\big\{\mathcal{D}_{x,\delta}\big\} \leq \frac{1}{\beta_n}\log p_n\big\{\mathcal{D}_{x,\delta}\big\}+(\theta_0+\gamma_n y')(x-2\delta)-\frac{\mu_n(\theta_0+\gamma_n y')}{\beta_n}. \quad (184)$$

Using the expansion of $\mu_n(\cdot)$ in (166) in the above, we obtain

$$\frac{1}{\beta_n}\log\tilde{p}_{n,\theta_0+\gamma_n y'}\big\{\mathcal{D}_{x,\delta}\big\} \leq \frac{1}{\beta_n}\log p_n\big\{\mathcal{D}_{x,\delta}\big\}+(\theta_0+\gamma_n y')(x-2\delta) -\frac{\alpha_n+\beta_n\nu_1+\beta_n\gamma_n\nu_{2,n}(y')}{\beta_n}. \quad (185)$$

Rearranging, we obtain

$$\frac{1}{\beta_n}\Big[-\log p_n\big\{\mathcal{D}_{x,\delta}\big\}+\alpha_n-\beta_n(\theta_0 x-\nu_1)\Big] \leq -2\delta(\theta_0+\gamma_n y')+\gamma_n\big(xy'-\nu_{2,n}(y')\big) -\frac{1}{\beta_n}\log\tilde{p}_{n,\theta_0+\gamma_n y'}\big\{\mathcal{D}_{x,\delta}\big\}. \quad (186)$$

Step 2 (Bounding the Probability in (186)): Now, our aim is to lower bound the probability in the final term in (186) namely $\tilde{p}_{n,\theta_0+\gamma_n y'}\big\{\mathcal{D}_{x,\delta}\big\}$. To this end, define the cumulant generating function with respect to the tilted measure $\tilde{p}_{n,\theta_0+\gamma_n y'}$ as follows:

$$\tilde{\mu}_n(\lambda):=\log\mathbb{E}_{\tilde{p}_{n,\theta_0+\gamma_n y'}}[\exp(\lambda X_n)]. \quad (187)$$

Now observe that for any $s\in\mathbb{R}$,

$$\begin{aligned}
\tilde{\mu}_n(\gamma_n s) &= \log\int_{\mathbb{R}}\exp\big(\gamma_n sX(\omega)\big)\tilde{p}_{n,\theta_0+\gamma_n y'}(\mathrm{d}\omega) && (188)\\
&= \log\int_{\mathbb{R}}\exp\big(\gamma_n sX(\omega)\big)\exp\big((\theta_0+\gamma_n y')X(\omega)\big) \times\exp\big(-\mu_n(\theta_0+\gamma_n y')\big)p_n(\mathrm{d}\omega) && (189)\\
&= \bigg[\log\int_{\mathbb{R}}\exp\big(\gamma_n sX(\omega)\big)\exp\big((\theta_0+\gamma_n y')X(\omega)\big) p_n(\mathrm{d}\omega)\bigg]-\mu_n(\theta_0+\gamma_n y') && (190)\\
&= \mu_n(\theta_0+\gamma_n(s+y'))-\mu_n(\theta_0+\gamma_n y') && (191)\\
&= \beta_n\gamma_n\big[\nu_{2,n}(s+y')-\nu_{2,n}(y')\big], && (192)
\end{aligned}$$

where (189) is due to (181), and (192) is due to (166). Thus, by normalizing by $\beta_n\gamma_n\to\infty$, and noting that $\nu_{2,n}$ converges to ν_2 pointwise (cf. (167)), we have

$$\lim_{n\to\infty}\frac{1}{\beta_n\gamma_n}\tilde{\mu}_n(\gamma_n s)=\nu_2(s+y')-\nu_2(y'). \quad (193)$$

Thus, from this calculation, we can conclude that

$$\xi(s;y'):=\nu_2(s+y')-\nu_2(y') \quad (194)$$

as a function of s, is the limiting cumulant generating function of the sequence of measures $\tilde{p}_{n,\theta_0+\gamma_n y'}$. By the union bound,

$$\tilde{p}_{n,\theta_0+\gamma_n y'}\big\{\mathcal{D}_{x,\delta}^c\big\}\leq\tilde{p}_{n,\theta_0+\gamma_n y'}\big\{\mathcal{E}_1\big\}+\tilde{p}_{n,\theta_0+\gamma_n y'}\big\{\mathcal{E}_2\big\} \quad (195)$$

where

$$\mathcal{E}_1:=\bigg\{\omega:\frac{X_n(\omega)}{\beta_n}<x-2\delta\bigg\},\quad\text{and} \quad (196)$$

$$\mathcal{E}_2:=\bigg\{\omega:\frac{X_n(\omega)}{\beta_n}>x\bigg\}. \quad (197)$$

We analyze the $\tilde{p}_{n,\theta_0+\gamma_n y'}$-probability of $\mathcal{E}_2$ first. By Markov's inequality, for any fixed $s\geq 0$,

$$\begin{aligned}
\tilde{p}_{n,\theta_0+\gamma_n y'}\big\{\mathcal{E}_2\big\} &= \tilde{p}_{n,\theta_0+\gamma_n y'}\big\{X_n>\beta_n x\big\} && (198)\\
&= \tilde{p}_{n,\theta_0+\gamma_n y'}\big\{\exp(s\gamma_n X_n)>\exp(s\beta_n\gamma_n x)\big\} && (199)\\
&\leq \frac{\mathbb{E}_{\tilde{p}_{n,\theta_0+\gamma_n y'}}[\exp(s\gamma_n X_n)]}{\exp(s\beta_n\gamma_n x)}. && (200)
\end{aligned}$$

Taking logarithms, normalizing by $\beta_n\gamma_n\to\infty$ and taking the lim sup, we obtain

$$\limsup_{n\to\infty}\frac{1}{\beta_n\gamma_n}\log\tilde{p}_{n,\theta_0+\gamma_n y'}\big\{\mathcal{E}_2\big\} \leq -sx+\limsup_{n\to\infty}\frac{1}{\beta_n\gamma_n}\log\mathbb{E}_{\tilde{p}_{n,\theta_0+\gamma_n y'}}[\exp(s\gamma_n X_n)] \quad (201)$$

Now using the definition of $s\mapsto\xi(s;y')$ in (193)–(194), we conclude that

$$\limsup_{n\to\infty}\frac{1}{\beta_n\gamma_n}\log\tilde{p}_{n,\theta_0+\gamma_n y'}\big\{\mathcal{E}_2\big\}\leq -sx+\xi(s;y'). \quad (202)$$

Since $s\geq 0$ is arbitrary, we have

$$\limsup_{n\to\infty}\frac{1}{\beta_n\gamma_n}\log\tilde{p}_{n,\theta_0+\gamma_n y'}\big\{\mathcal{E}_2\big\}\leq-\sup_{s\geq 0}\big\{sx-\xi(s;y')\big\}. \quad (203)$$

Define the Fenchel-Legendre transform

$$\xi^*(x;y'):=\sup_{s\geq 0}\big\{sx-\xi(s;y')\big\}. \quad (204)$$

Now, we claim that $\xi^*(x;y')>0$. Let s_x^* achieve the supremum in (204). Consider the following optimality condition for the convex optimization problem in (204):

$$x-\nu_2'(s_x^*+y')=0. \quad (205)$$

Since ν_2 is assumed to be strictly convex, so ν_2' is strictly increasing. Furthermore, it has the property (cf. (179)) that $\nu_2'(y')=x'$. Since $x'<x$, by continuity of ν_2', the optimal s_x^* in (205) is positive. Since the strict convexity of ν_2 means the same for $s\mapsto\xi(s;y')$, this implies that $\xi^*(x;y')$ is positive. So we conclude that

$$\tilde{p}_{n,\theta_0+\gamma_n y'}\big\{\mathcal{E}_2\big\}\leq\exp(-\beta_n\gamma_n\tau_2) \quad (206)$$

for some $\tau_2>0$ and for all n large enough. In a completely analogous way, we can show that

$$\limsup_{n\to\infty}\frac{1}{\beta_n\gamma_n}\log\tilde{p}_{n,\theta_0+\gamma_n y'}\big\{\mathcal{E}_1\big\}\leq-\sup_{s\leq 0}\big\{s(x-2\delta)-\xi(s;y')\big\}. \quad (207)$$

Defining

$$\tilde{\xi}^*(x-2\delta;y'):=\sup_{s\leq 0}\big\{s(x-2\delta)-\xi(s;y')\big\}, \quad (208)$$

and examining the optimality condition for s in (208), we see that $\tilde{\xi}^*(x-2\delta; y') > 0$ and so

$$\tilde{p}_{n,\theta_0+\gamma_n y'}\{\mathcal{E}_1\} \le \exp(-\beta_n\gamma_n\tau_1) \tag{209}$$

for some $\tau_1 > 0$ and for all n large enough. Consequently, we have

$$\tilde{p}_{n,\theta_0+\gamma_n y'}\{\mathcal{D}_{x,\delta}\} \ge 1 - 2\exp(-\beta_n\gamma_n\tau) \tag{210}$$

where $\tau := \min\{\tau_1, \tau_2\} > 0$.

Step 3 (Considering Asymptotics): Now substituting (210) back into (186), we obtain

$$\begin{aligned}\frac{1}{\beta_n}&\left[-\log p_n\{\mathcal{D}_{x,\delta}\} + \alpha_n - \beta_n(\theta_0 x - \nu_1)\right]\\ &\le -2\delta(\theta_0+\gamma_n y') + \gamma_n\big(xy' - \nu_{2,n}(y')\big)\\ &\quad - \frac{1}{\beta_n}\log\big(1 - 2\exp(-\beta_n\gamma_n\tau)\big).\end{aligned} \tag{211}$$

Since $\beta_n\gamma_n \to \infty$, the final term vanishes. Consequently, taking limits, we have

$$\limsup_{n\to\infty}\frac{1}{\beta_n}\left[-\log p_n\{\mathcal{D}_{x,\delta}\} + \alpha_n - \beta_n(\theta_0 x - \nu_1)\right] \le -2\delta\theta_0. \tag{212}$$

Since $\delta > 0$ is arbitrary, we may take $\delta \downarrow 0$ (also recall that $\theta_0 < 0$ so the term on the right-hand-side of (212) is non-negative) to conclude that

$$\limsup_{n\to\infty}\frac{1}{\beta_n}\left[-\log p_n\{\mathcal{D}_{x,\delta}\} + \alpha_n - \beta_n(\theta_0 x - \nu_1)\right] \le 0. \tag{213}$$

This concludes the proof of the upper bound of (170) for the Case (ib), i.e., $\beta_n\gamma_n \to \infty$. ∎

C. Proof of Upper Bound of Case (ib) in (170), i.e., $\beta_n = \gamma_n^{-1}$

Proof: Recall that $\theta_0 < 0$ and $\beta_n = \gamma_n^{-1}$ consequently $\beta_n\gamma_n = 1$. Define the tilted probability measure

$$\tilde{p}_n(\omega) := p_n(\omega)\exp\big((\theta_0+\gamma_n y_0)X_n(\omega) - \mu_n(\theta_0+\gamma_n y_0)\big). \tag{214}$$

Fix $s \in \mathbb{R}$. Note from the strict convexity of ν_2 that $\nu_2''(y_0) > 0$ for all y_0. Then, using $\mathbb{E}_{\tilde{p}_n}[\cdot]$ to denote the expectation with respect to the distribution $\tilde{p}_n$ in (214), we have

$$\begin{aligned}\log&\,\mathbb{E}_{\tilde{p}_n}\left[\exp\left(\sqrt{\frac{\gamma_n}{\beta_n\nu_2''(y_0)}}s(X_n - x\beta_n)\right)\right]\\ &= \log\int_{\mathbb{R}}\exp\left(\sqrt{\frac{\gamma_n}{\beta_n\nu_2''(y_0)}}s(X_n(\omega) - x\beta_n)\right)\mathrm{d}\tilde{p}_n(\omega)\end{aligned} \tag{215}$$

$$\begin{aligned}&= \log\int_{\mathbb{R}}\exp\left(\sqrt{\frac{\gamma_n}{\beta_n\nu_2''(y_0)}}s(X_n(\omega) - x\beta_n)\right)\\ &\qquad\times\exp\big((\theta_0+\gamma_n y_0)X_n(\omega) - \mu_n(\theta_0+\gamma_n y_0)\big)\mathrm{d}p_n(\omega)\end{aligned} \tag{216}$$

$$\begin{aligned}&= \mu_n\left(\theta_0 + \gamma_n\left(y_0 + \sqrt{\frac{1}{\gamma_n\beta_n\nu_2''(y_0)}}s\right)\right)\\ &\quad - \mu_n(\theta_0+\gamma_n y_0) - \sqrt{\frac{\gamma_n}{\beta_n\nu_2''(y_0)}}sx\beta_n\end{aligned} \tag{217}$$

$$= \nu_{2,n}\left(y_0 + \sqrt{\frac{1}{\nu_2''(y_0)}}s\right) - \nu_{2,n}(y_0) - \sqrt{\frac{1}{\nu_2''(y_0)}}sx \tag{218}$$

$$= \nu_2\left(y_0 + \sqrt{\frac{1}{\nu_2''(y_0)}}s\right) - \nu_2(y_0) + o(1) - \sqrt{\frac{1}{\nu_2''(y_0)}}sx \tag{219}$$

where

1) (216) follows from the change of measure per (214);
2) (217) follows from the definition of $\mu_n(\cdot)$ in (165);
3) (218) follows from the expansion of $\mu_n(\theta_0+\gamma_n y)$ in (166) and the fact that $\beta_n\gamma_n = 1$;
4) and (219) follows from the pointwise convergence of $\nu_{2,n}$ to ν_2 in (167).

Hence, taking limits and noting that $x = \nu_2'(y_0)$ (cf. (168)), we obtain

$$\begin{aligned}\lim_{n\to\infty}&\log\mathbb{E}_{\tilde{p}_n}\left[\exp\left(\sqrt{\frac{\gamma_n}{\beta_n\nu_2''(y_0)}}s(X_n - x\beta_n)\right)\right]\\ &= \nu_2\left(y_0 + \sqrt{\frac{1}{\nu_2''(y_0)}}s\right) - \nu_2(y_0) - s\nu_2'(y_0)\sqrt{\frac{1}{\nu_2''(y_0)}}.\end{aligned} \tag{220}$$

Since we assumed that ν_2 is a quadratic function, the second-order Taylor approximation of ν_2 at y_0 is exactly a quadratic so

$$\lim_{n\to\infty}\log\mathbb{E}_{\tilde{p}_n}\left[\exp\left(\sqrt{\frac{\gamma_n}{\beta_n\nu_2''(y_0)}}s(X_n - x\beta_n)\right)\right] = \frac{1}{2}s^2. \tag{221}$$

The relation in (221) says that the sequence of random variables $\sqrt{\frac{\gamma_n}{\beta_n\nu_2''(y_0)}}(X_n - x\beta_n)$ with corresponding distribution $\tilde{p}_n$ has a sequence of cumulant generating functions that converges pointwise to the quadratic function $\frac{1}{2}s^2$. Thus, we can conclude that it converges in distribution to the standard Gaussian by Lévi's continuity theorem [34, Th. 18.21].

Furthermore, from (214), we have

$$\begin{aligned}p_n&\{X_n \le x\beta_n\}\exp\big(-\mu_n(\theta_0+\gamma_n y_0)\big)\exp\big((\theta_0+\gamma_n y_0)x\beta_n\big)\\ &= \int_\Omega\exp\big(-(\theta_0+\gamma_n y_0)(X_n(\omega) - x\beta_n)\big)\\ &\qquad\times\mathbf{1}\{X_n(\omega) \le x\beta_n\}\,\mathrm{d}\tilde{p}_n(\omega).\end{aligned} \tag{222}$$

Now, for notational brevity, define

$$a_n := -(\theta_0+\gamma_n y_0)\sqrt{\frac{\beta_n\nu_2''(y_0)}{\gamma_n}}. \tag{223}$$

Because $\theta_0 < 0$, $\gamma_n \to 0$ and $\beta_n \to \infty$, we conclude that $a_n \ge 0$ for n large enough and also $a_n \to \infty$.

Let the probability measure corresponding to the random variable $\sqrt{\frac{\gamma_n}{\beta_n\nu_2''(y_0)}}(X_n - x\beta_n)$ with respect to the distribution $\tilde{p}_n$ in (214) be denoted as $\mathbb{P}_n$. More precisely, for every

Borel measurable set $\mathcal{E}$, we have the relation

$$\mathbb{P}_n(\mathcal{E}) := \int_\Omega \mathbf{1}\left\{\sqrt{\frac{\gamma_n}{\beta_n \nu_2''(y_0)}}(X_n(\omega) - x\beta_n) \in \mathcal{E}\right\} \mathrm{d}\tilde{p}_n(\omega). \tag{224}$$

This is the relation between the measures $\tilde{p}_n$ and $\mathbb{P}_n$, and it is this change-of-measure step (from $\tilde{p}_n$ to $\mathbb{P}_n$) that is crucial in this proof. Note from the calculation leading to (221) that $\mathbb{P}_n$ converges weakly to the standard Gaussian measure $\mathbb{P}(\mathcal{A}) := \int_{\mathcal{A}} \varphi(w)\,\mathrm{d}w$. Thus, through a change of variables

$$\omega \mapsto z = \sqrt{\frac{\gamma_n}{\beta_n \nu_2''(y_0)}}(X_n(\omega) - x\beta_n), \tag{225}$$

the quantity in (222) can be expressed as the integral

$$\Psi_n := \int_{-\infty}^{0} \exp(a_n z)\,\mathrm{d}\mathbb{P}_n(z). \tag{226}$$

We now provide upper and lower bounds for this integral to understand its asymptotic behavior. We have

$$\Psi_n \le \int_{-\infty}^{0} \mathrm{d}\mathbb{P}_n(z) \le \frac{3}{4} \tag{227}$$

where the last bound follows for all n sufficiently large due to Lévi's continuity theorem [34, Th. 18.21], i.e., $\mathbb{P}_n \to \mathbb{P}$ in distribution where $\mathbb{P}$ is a standard Gaussian distribution. Obviously, $\mathbb{P}((\infty, 0]) = \frac{1}{2}$.

To lower bound Ψ_n, we fix $\varepsilon > 0$ consider

$$\Psi_n \ge \int_{-\infty}^{-\varepsilon} \exp(a_n z)\,\mathrm{d}\mathbb{P}_n(z) \tag{228}$$

$$\ge \exp(-\varepsilon a_n) \int_{-\infty}^{-\varepsilon} \mathrm{d}\mathbb{P}_n(z) \tag{229}$$

$$\ge \frac{1}{4}\exp(-\varepsilon a_n) \tag{230}$$

where in (229) we substituted the upper limit into the integrand and in (230) we assumed ε is chosen small enough so that $\mathbb{P}_n((-\infty, -\varepsilon]) \ge \frac{1}{4}$ for all n large enough. In sum, (227) and (230) yield

$$\Psi_n = O(1) \cap \Omega\big(\exp(-\varepsilon a_n)\big). \tag{231}$$

Recall from (222) that

$$p_n\{X_n \le x\beta_n\} \exp\big(-\mu_n(\theta_0 + \gamma_n y_0)\big) \times \exp\big((\theta_0 + \gamma_n y_0)x\beta_n\big) = \Psi_n. \tag{232}$$

Therefore,

$$\begin{aligned} &-\log p_n\{X_n \le x\beta_n\} \\ &= -\mu_n(\theta_0 + \gamma_n y_0) + (\theta_0 + \gamma_n y_0)x\beta_n - \log \Psi_n \qquad (233)\\ &= -\alpha_n - \beta_n \nu_1 - \beta_n\gamma_n \nu_{2,n}(y_0) + (\theta_0 + \gamma_n y_0)x\beta_n - \log \Psi_n \qquad (234)\\ &= -\alpha_n + \beta_n(\theta_0 x - \nu_1) + \beta_n\gamma_n(y_0 x - \nu_2(y_0) + o(1)) - \log \Psi_n, \qquad (235)\\ &= -\alpha_n + \beta_n(\theta_0 x - \nu_1) + O(1) - \log \Psi_n, \qquad (236) \end{aligned}$$

where (235) holds from the pointwise convergence of $\nu_{2,n}$ to ν_2 (cf. (167)), and (236) holds because $\beta_n\gamma_n = 1$ and $y_0 x - \nu_2(y_0) = O(1)$. Now using the the asymptotic behavior of Ψ_n in (231), we obtain

$$-\kappa \le \big[-\log p_n\{X_n \le x\beta_n\} + \alpha_n - \beta_n(\theta_0 x - \nu_1)\big] + O(1) \le \kappa\varepsilon a_n \tag{237}$$

for some finite constant $\kappa > 0$. Since $\theta_0 < 0$, from the definition of a_n in (223), we know that a_n is of the order $O(\sqrt{\beta_n \gamma_n^{-1}})$. Additionally, since $\varepsilon > 0$ is arbitrarily small,

$$-\kappa \le \big[-\log p_n\{X_n \le x\beta_n\} + \alpha_n - \beta_n(\theta_0 x - \nu_1)\big] + O(1) \le o\Big(\sqrt{\beta_n \gamma_n^{-1}}\Big). \tag{238}$$

Now, recall that $\beta_n = \gamma_n^{-1}$. So $o\big(\sqrt{\beta_n\gamma_n^{-1}}\big) = o(\beta_n)$ and we have finished the proof of the upper bound in (170) for the case $\beta_n = \gamma_n^{-1}$ and ν_2 being a quadratic. ■

D. Proof of Upper Bound of Case (ii) in (171)

Proof: The proof of this case proceeds similarly to that in Appendix V-B. We only highlight the main differences here. The measure tilting step proceeds similarly with the exception that $\theta = \gamma_n y'$ (cf. (180)). Because δ can be chosen arbitrarily small, $y' < 0$ by continuity since $y_0 < 0$. Thus similarly to the proof in Appendix V-B, θ is a negative sequence.

Since $\theta_0 = 0$, (211) reduces to

$$\begin{aligned} &\frac{1}{\beta_n\gamma_n}\big[-\log p_n\{\mathcal{D}_{x,\delta}\} + \alpha_n - \beta_n(\theta_0 x - \nu_1)\big] \\ &\quad \le -2\delta y' + \big(xy' - \nu_{2,n}(y')\big) \\ &\qquad - \frac{1}{\beta_n\gamma_n}\log\big(1 - 2\exp(-\beta_n\gamma_n b)\big). \end{aligned} \tag{239}$$

Since $\beta_n\gamma_n \to \infty$, the final term vanishes when we take limits, yielding

$$\limsup_{n\to\infty} \frac{1}{\beta_n\gamma_n}\big[-\log p_n\{\mathcal{D}_{x,\delta}\} + \alpha_n - \beta_n(\theta_0 x - \nu_1)\big] \le -2\delta y' + \big(xy' - \nu_2(y')\big). \tag{240}$$

Now take $\delta \downarrow 0$, we obtain by the continuity of ν_2' that $y' \to y_0$. Thus, we have

$$\limsup_{n\to\infty} \frac{1}{\beta_n\gamma_n}\big[-\log p_n\{\mathcal{D}_{x,\delta}\} + \alpha_n - \beta_n(\theta_0 x - \nu_1)\big] \le xy_0 - \nu_2(y_0) \tag{241}$$

as desired. ■

Remark 2: Observe from the above proof that we used two different techniques for the cases $\beta_n\gamma_n \to \infty$ *(Appendices A-B and A-D) and* $\beta_n = \gamma_n^{-1}$ *(Appendix A-C). The former case follows essentially from the same steps as in the standard proof of the Gärtner-Ellis theorem in [23, Th. 2.3.6] (however, see Remark 3). The latter case cannot be handled using the technique for* $\beta_n\gamma_n \to \infty$ *because the final term in (211) fails to vanish with* $\beta_n\gamma_n = $ const. *Hence, we develop a novel technique based on the weak convergence of* $\sqrt{\frac{\gamma_n}{\beta_n\nu_2''(y_0)}}(X_n - x\beta_n)$ *(under the tilted measure* $\tilde{p}_n$*) to handle the case where* $\beta_n\gamma_n = $ const *(with the added assumption that* ν_2 *is a quadratic). This technique may be of independent* *interest to other problems in probability theory. We note, though, that this technique based on weak convergence cannot be used to handle the case in which* $\beta_n \gamma_n \to \infty$. *This is because in this case, a careful examination of the steps from* (215) *to* (221) *would show that this technique leads to an approximation of the cumulant generating function* $\log \mathbb{E}_{\tilde{p}_n}\left[\exp\left(\sqrt{\frac{\gamma_n}{\beta_n v_2''(y_0)}} s(X_n - x\beta_n)\right)\right]$ *that is too coarse for our needs.*

Remark 3: Readers familiar with the standard proof of the Gärtner-Ellis theorem in [23, Th. 2.3.6] would notice the subtle difference of the proof in Appendix A-B vis-à-vis the standard one. The titled measure is $\tilde{p}_{n,\theta_0+\gamma_n y'}$ *and* y' *is defined in terms of* x' *in* (178). *In contrast, in [23, Th. 2.3.6], the tilting parameter is chosen to be a fixed exposed hyperplane [23, Definition 2.3.3] of the analogue of* x'. *Our tilting parameter, and hence also the tilting distribution, is allowed to vary with* n.

Appendix B
Convergence in Distribution Based on Convergence of Cumulant Generating Functions

Lemma 9: Let μ_n *be a sequence of probability measures on* $\mathbb{R}$. *Suppose that for some* $0 < a < b$,

$$\log \int_{\mathbb{R}} \exp(sx)\mu_n(\mathrm{d}x) \to f(s) := \frac{s^2}{2}, \quad \forall s \in (a, b). \tag{242}$$

Then, μ_n *converges (weakly) to the standard Gaussian* $\mu(\mathcal{A}) = \int_{\mathcal{A}} \varphi(w)\,\mathrm{d}w$.

Notice that in (242), the assumption pertains only to s in the open interval (a, b). 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 [34, Th. 18.21]).

See Mukherjea *et al.* [31, Th. 2] for the proof of Lemma 9.

Appendix C
A Basic Concentration Bound

Lemma 10: Let $X_1, \ldots, X_L$ *be independent random variables, each distributed according to the uniform distribution on* $\{1, \ldots, M_1\}$. *We fix a subset* $\mathcal{A} \subset \{1, \ldots, M_1\}$ *whose cardinality is* M_2. *We denote the random number* $|\{i \in \{1, \ldots, L\} : X_i \in \mathcal{A}\}|$ *by* N. *For every* $s > 0$,

$$\mathbb{E}[N^s] \geq \left\lfloor \frac{LM_2}{M_1}(1-\epsilon)\right\rfloor^s \left[1 - \exp\left(-L\frac{M_2}{2M_1}\epsilon^2\right)\right] \tag{243}$$

where $0 < \epsilon < 1$ *is also an arbitrary number.*

Proof: By straightforward calculations, we have

$$\begin{aligned}
\mathbb{E}[N^s] &= \sum_{l=0}^{L} l^s \Pr(N = l) &(244)\\
&\geq \sum_{l \geq LM_2(1-\epsilon)/M_1} l^s \Pr(N = l) &(245)\\
&\geq \left\lfloor \frac{LM_2}{M_1}(1-\epsilon)\right\rfloor^s \Pr\left(N \geq \frac{LM_2(1-\epsilon)}{M_1}\right) &(246)\\
&= \left\lfloor \frac{LM_2}{M_1}(1-\epsilon)\right\rfloor^s \left[1 - \Pr\left(\frac{N}{L} < (1-\epsilon)\frac{M_2}{M_1}\right)\right]. &(247)
\end{aligned}$$

Now since the event in probability in (247) implies that the relative frequency of the number of events $\{X_i \in \mathcal{A}\}, i = 1, \ldots, L$ is less than $1-\epsilon$ multiplied by the mean $\mathbb{E}[\mathbf{1}\{X_i \in \mathcal{A}\}] = M_2/M_1$ of each indicator $\mathbf{1}\{X_i \in \mathcal{A}\}$, we can invoke the Chernoff bound for independent Bernoulli trials (e.g., [35, Th. 4.5]) to conclude that

$$\begin{aligned}
\Pr\left(\frac{N}{L} < (1-\epsilon)\frac{M_2}{M_1}\right) &= \Pr\left(\frac{1}{L}\sum_{i=1}^{L} \mathbf{1}\{X_i \in \mathcal{A}\} < (1-\epsilon)\frac{M_2}{M_1}\right)\\
&\leq \exp\left(-L\frac{M_2}{2M_1}\epsilon^2\right). &(248)
\end{aligned}$$

Combining this with (247) concludes the proof. ■

Acknowledgements

The authors would like to acknowledge the Associate Editor (Prof. Aaron Wagner) and the anonymous reviewers for their extensive and useful comments during the revision process.

MH is grateful for Prof. Nobuo Yoshida for clarifying Lemma 9. VT is grateful to Prof. Rongfeng Sun for clarifying the same lemma.

References

[1] G. D. Forney, Jr., "Exponential error bounds for erasure, list, and decision feedback schemes," *IEEE Trans. Inf. Theory*, vol. 14, no. 2, pp. 206–220, Mar. 1968.

[2] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, "Lower bounds to error probability for coding on discrete memoryless channels. II," *Inf. Control*, vol. 10, no. 5, pp. 522–552, May 1967.

[3] E. I. Telatar and R. G. Gallager, "New exponential upper bounds to error and erasure probabilities," in *Proc. Int. Symp. Inf. Theory*, Trondheim, Norway, Jun./Jul. 1994, p. 379.

[4] V. M. Blinovsky, "Error probability exponent of list decoding at low rates," *Problems Inf. Transmiss.*, vol. 37, no. 4, pp. 277–287, Oct./Dec. 2001.

[5] P. Moulin, "A Neyman–Pearson approach to universal erasure and list decoding," *IEEE Trans. Inf. Theory*, vol. 55, no. 10, pp. 4462–4478, Oct. 2009.

[6] N. Merhav, "Error exponents of erasure/list decoding revisited via moments of distance enumerators," *IEEE Trans. Inf. Theory*, vol. 54, no. 10, pp. 4439–4447, Oct. 2008.

[7] N. Merhav, "List decoding—Random coding exponents and expurgated exponents," *IEEE Trans. Inf. Theory*, vol. 60, no. 11, pp. 6749–6759, Nov. 2014.

[8] N. Merhav, "Erasure/list exponents for Slepian–Wolf decoding," *IEEE Trans. Inf. Theory*, vol. 60, no. 8, pp. 4463–4471, Aug. 2014.

[9] A. Somekh-Baruch and N. Merhav, "Exact random coding exponents for erasure decoding," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6444–6454, Oct. 2011.

[10] N. Merhav, "Statistical physics and information theory," *Found. Trends Commun. Inf. Theory*, vol. 6, nos. 1–2, pp. 1–212, 2010.

[11] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, Jul./Oct. 1948.

[12] V. Strassen, "Asymptotische Abschätzungen in Shannons Informationstheorie," in *Proc. Trans. 3rd Prague Conf. Inf. Theory*, Prague, Czech Republic, 1962, pp. 689–723. [Online]. Available: http://www.maTheorycornell.edu/~pmlut/strassen.pdf

[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, Nov. 2009.

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

[15] V. Y. F. Tan, "Asymptotic estimates in information theory with non-vanishing error probabilities," *Found. Trends Commun. Inf. Theory*, vol. 11, nos. 1–2, pp. 1–184, Sep. 2014.

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

[17] V. Y. F. Tan, "Moderate-deviations of lossy source coding for discrete and Gaussian sources," in *Proc. Int. Symp. Inf. Theory*, Cambridge, MA, USA, May 2012, pp. 920–924. [Online]. Available: http://arxiv.org/abs/1111.2217
[18] V. Y. F. Tan and P. Moulin, "Second-order capacities of erasure and list decoding," in *Proc. Int. Symp. Inf. Theory*, Honolulu, HI, USA, Jun./Jul. 2014, pp. 1887–1891.
[19] Y. Polyanskiy and S. Verdú, "Channel dispersion and moderate deviations limits for memoryless channels," in *Proc. Allerton Conf.*, Sep./Oct. 2010, pp. 1334–1339.
[20] T. S. Han, *Information-Spectrum Methods in Information Theory*. Berlin, Germany: Springer-Verlag, Feb. 2003.
[21] Y. Altuğ, A. B. Wagner, and I. Kontoyiannis, "Lossless compression with moderate error probability," in *Proc. Int. Symp. Inf. Theory*, Jul. 2013, pp. 1744–1748.
[22] V. Y. F. Tan, S. Watanabe, and M. Hayashi, "Moderate deviations for joint source-channel coding of systems with Markovian memory," in *Proc. Int. Symp. Inf. Theory*, Honolulu, HI, USA, Jun./Jul. 2014, pp. 1687–1691.
[23] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. Berlin, Germany: Springer-Verlag, 1998.
[24] P.-N. Chen, "Generalization of Gartner-Ellis theorem," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2752–2760, Nov. 2000.
[25] C. Joutard, "A strong large deviation theorem," *Math. Methods Statist.*, vol. 22, no. 2, pp. 155–164, Apr. 2013.
[26] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2011.
[27] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York, NY, USA: Wiley, 2006.
[28] I. Csiszár, "Linear codes for sources and source networks: Error exponents, universal coding," *IEEE Trans. Inf. Theory*, vol. 28, no. 4, pp. 585–592, Jul. 1982.
[29] S. Verdú and I. Kontoyiannis, "Optimal lossless data compression: Non-asymptotics and asymptotics," *IEEE Trans. Inf. Theory*, vol. 60, no. 2, pp. 777–795, Feb. 2014.
[30] V. Kostina and S. Verdú, "Fixed-length lossy compression in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 58, no. 6, pp. 3309–3338, Jun. 2012.
[31] A. Mukherjea, M. Rao, and S. Suen, "A note on moment generating functions," *Statist. Probab. Lett.*, vol. 76, no. 11, pp. 1185–1189, Jun. 2006.
[32] J. H. Curtiss, "A note on the theory of moment generating functions," *Ann. Math. Statist.*, vol. 13, no. 4, pp. 430–433, 1942.
[33] W. Feller, *An Introduction to Probability Theory and Its Applications*, 2nd ed. New York, NY, USA: Wiley, 1971.
[34] B. E. Fristedt and L. F. Gray, *A Modern Approach to Probability Theory*. Boston, MA, USA: Birkhäuser, 1996.
[35] M. Mitzenmacher and E. Upfal, *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

Masahito Hayashi (M'06–SM'13) was born in Japan in 1971. He received the B.S. degree from the Faculty of Sciences in Kyoto University, Japan, in 1994 and the M.S. and Ph.D. degrees in Mathematics from Kyoto University, Japan, in 1996 and 1999, respectively.

He worked in Kyoto University as a Research Fellow of the Japan Society of the Promotion of Science (JSPS) from 1998 to 2000, and worked in the Laboratory for Mathematical Neuroscience, Brain Science Institute, RIKEN from 2000 to 2003, and worked in ERATO Quantum Computation and Information Project, Japan Science and Technology Agency (JST) as the Research Head from 2000 to 2006. He also worked in the Superrobust Computation Project Information Science and Technology Strategic Core (21st Century COE by MEXT) Graduate School of Information Science and Technology, The University of Tokyo as Adjunct Associate Professor from 2004 to 2007. In 2006, he published the book "*Quantum Information: An Introduction*" from Springer. He worked in the Graduate School of Information Sciences, Tohoku University as Associate Professor from 2007 to 2012. In 2012, he joined the Graduate School of Mathematics, Nagoya University as Professor. He also worked in Centre for Quantum Technologies, National University of Singapore as Visiting Research Associate Professor from 2009 to 2012 and as Visiting Research Professor from 2012 to now. In 2011, he received Information Theory Society Paper Award (2011) for "Information-Spectrum Approach to Second-Order Coding Rate in Channel Coding".

He is on the Editorial Board of *International Journal of Quantum Information* and *International Journal on Advances in Security*. His research interests include classical and quantum information theory and classical and quantum statistical inference.

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 (I2R) Research, A*STAR, Singapore. His research interests include information theory, machine learning and signal processing.

Dr. Tan received the MIT EECS Jin-Au Kong outstanding doctoral thesis prize in 2011 and the NUS Young Investigator Award in 2014. 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.