Second-Order Capacities of Erasure and List Decoding

Vincent Y. F. Tan
Department of Electrical and Computer Engineering,
National University of Singapore
Email: vtan@nus.edu.sg

Pierre Moulin
Department of Electrical and Computer Engineering,
University of Illinois at Urbana-Champaign
Email: moulin@ifp.uiuc.edu

***Abstract*—We derive the second-order capacities (supremum of second-order coding rates) for erasure and list decoding. For erasure decoding, we show that second-order capacity is $\sqrt{V}\Phi^{-1}(\epsilon_{\mathrm{t}})$ where V is the channel dispersion and ϵ_{t} is the total error probability, i.e. the sum of the erasure and undetected errors. We show numerically that the expected rate at finite blocklength for erasures decoding can exceed the finite blocklength channel coding rate. For list decoding, we consider list codes of deterministic size $2^{\sqrt{n}l}$ and show that the second-order capacity is $l+\sqrt{V}\Phi^{-1}(\epsilon)$ where ϵ is the permissible error probability. Both coding schemes use the threshold decoder and converses are proved using variants of the meta-converse.**

I. Introduction

In many communication scenarios, it is advantageous to allow the decoder to have the option of either not deciding at all or putting out more than one estimate of the message. These are respectively known as *erasure* and *list* decoding respectively and have been studied extensively in [1]–[7]. The erasure and list options allow for smaller undetected error probabilities so these options are useful in practice.

In this paper, we revisit the problem or erasure and list decoding from the viewpoint of second-order asymptotics. The study of second- and higher-order asymptotics at fixed (non-vanishing) error probability was first done by Strassen [8] who showed for a well-behaved discrete memoryless channel (DMCs) W that the maximum number of codewords at error probability ϵ, namely $M^*(W^n,\epsilon)$, satisfies

$$\log M^*(W^n,\epsilon) = nC + \sqrt{nV}\Phi^{-1}(\epsilon) + O(\log n), \qquad (1)$$

where C and V are respectively the capacity and the dispersion of W. This line of work has been revisited by numerous authors recently [9]–[14].

A. Main Contributions

In this paper, for erasure decoding, we consider constant undetected and total (sum of undetected and erasure) error probabilities (numbers between 0 and 1) and we obtain the analogue of the second-order $\sqrt{n}$ term in (1). We show that the coefficient of the second-order term, termed the *second-order capacity*, is $\sqrt{V}\Phi^{-1}(\epsilon_{\mathrm{t}})$ where ϵ_{t} is the total error probability. We then compute the expected rate at finite blocklength allowing erasures and show that it can exceed the finite blocklength rate without the erasure option. For list decoding, we consider lists of deterministic size of order $2^{\sqrt{n}l}$ and show that the second-order capacity is $l+\sqrt{V}\Phi^{-1}(\epsilon)$. To the best of the authors' knowledge, this is the first time that lists of size other than constant or exponential have been considered in the literature.

B. Related Work

Previously, the study of erasure and list decoding has been primarily from the error exponents perspective. Forney [1] derived the optimal rules by generalizing the Neyman-Pearson lemma and also proved exponential upper bounds for the error probabilities using Gallager's techniques [15]. Shannon-Gallager-Berlekamp [2] considered lists of size 2^{nl} and showed that sphere packing error exponent (evaluated at the code rate minus l) is an upper bound on the reliability function. Bounds for the error probabilities were derived by Telatar [3] using a decoder parametrized by an asymmetric relation $\prec$ which is a function of the channel law. Blinovsky [4] studied the exponents of the list decoding problem at low (and even 0) rate. Csiszár-Körner [16, Thm. 10.11] present exponential upper bounds for universally attainable erasure decoding using the method of types. Moulin [5] generalized the treatment there and presented improved error exponents. Recently, Merhav also considered alternative methods of analysis [6] and expurgated exponents for these problems [7].

II. Problem Setting and Main Definitions

Let W be a channel from an input alphabet $\mathcal{X}$ to an output alphabet $\mathcal{Y}$. We denote length-n strings $\mathbf{x}=(x_1,\ldots,x_n)\in\mathcal{X}^n$ by boldface. If W^n satisfies $W^n(\mathbf{y}|\mathbf{x})=\prod_{i=1}^n W(y_i|x_i)$ for every $(\mathbf{x},\mathbf{y})\in\mathcal{X}^n\times\mathcal{Y}^n$ and the sets $\mathcal{X}$ and $\mathcal{Y}$ are finite, W^n is said to be a DMC. We focus on DMCs in this paper but extensions to other channels such as the AWGN channel are possible. For an finite alphabet $\mathcal{X}$, let $\mathcal{P}(\mathcal{X})$ and $\mathcal{P}_n(\mathcal{X})$ be the set of probability mass functions and n-types [16] respectively.

For information-theoretic quantities, we will mostly follow the notation in Csiszár and Körner [16]. For a DMC, we denote its *capacity* by $C := \max_{P\in\mathcal{P}(\mathcal{X})} I(P,W)$. We let Π be the set of *capacity-achieving input distributions*. If $(X,Y)\sim P\times W$, define $V(P,W) := \mathbb{E}_X\big[\mathsf{Var}\big(\log\frac{W(Y|X)}{PW(Y)}\,|\,X\big)\big]$ to be the *conditional information variance*. The *ϵ-dispersion* of the DMC W [8]–[10] is defined as

$$V_\epsilon := \begin{cases} V_{\min} := \min_{P\in\Pi} V(P,W) & \epsilon < 1/2 \\ V_{\max} := \max_{P\in\Pi} V(P,W) & \epsilon \ge 1/2 \end{cases}. \qquad (2)$$

For integers $l \leq m$, we denote $[l:m] := \{l, l+1, \ldots, m\}$ and $[m] := [1:m]$. Let $\Phi(x) := \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2}\, \mathrm{d}t$ be the cumulative distribution function of a standard Gaussian and $\Phi^{-1}(\cdot)$ be its inverse. We now define erasure codes.

Definition 1. *An M-*erasure code *for $W : \mathcal{X} \to \mathcal{Y}$ is a pair of mappings (f, φ) such that $f : [M] \to \mathcal{X}$ and $\varphi : \mathcal{Y} \to [0:M]$. The disjoint* decoding regions *are denoted as $\mathcal{D}_m := \varphi^{-1}(m)$ and the* conditional undetected, erasure *and* total error probabilities *are defined as*

$$\lambda_{\mathrm{u}}(m) := \sum_{\tilde{m} \in [M] \setminus \{m\}} W(\mathcal{D}_{\tilde{m}} | f(m)) \tag{3}$$

$$\lambda_{\mathrm{e}}(m) := W(\mathcal{D}_0 | f(m)) \tag{4}$$

$$\lambda_{\mathrm{t}}(m) := \sum_{\tilde{m} \in [0:M] \setminus \{m\}} W(\mathcal{D}_{\tilde{m}} | f(m)) \tag{5}$$

Note that $\lambda_{\mathrm{u}}(m) + \lambda_{\mathrm{e}}(m) = \lambda_{\mathrm{t}}(m)$. Typically, the code is designed so that $\lambda_{\mathrm{u}}(m) \ll \lambda_{\mathrm{e}}(m)$.

Definition 2. *An $(M, \epsilon_{\mathrm{u}}, \epsilon_{\mathrm{t}})_{\mathrm{a,b}}$-*erasure code *for W is an M-erasure code for the same channel where*

1) *If* $(\mathrm{a}, \mathrm{b}) = (\max, \max)$,
$$\max_{m \in [M]} \lambda_{\mathrm{u}}(m) \leq \epsilon_{\mathrm{u}}, \quad \max_{m \in [M]} \lambda_{\mathrm{t}}(m) \leq \epsilon_{\mathrm{t}}. \tag{6}$$
2) *If* $(\mathrm{a}, \mathrm{b}) = (\max, \mathrm{ave})$
$$\max_{m \in [M]} \lambda_{\mathrm{u}}(m) \leq \epsilon_{\mathrm{u}}, \quad \frac{1}{M} \sum_{m \in [M]} \lambda_{\mathrm{t}}(m) \leq \epsilon_{\mathrm{t}}. \tag{7}$$
3) *If* $(\mathrm{a}, \mathrm{b}) = (\mathrm{ave}, \max)$,
$$\frac{1}{M} \sum_{m \in [M]} \lambda_{\mathrm{u}}(m) \leq \epsilon_{\mathrm{u}}, \quad \max_{m \in [M]} \lambda_{\mathrm{t}}(m) \leq \epsilon_{\mathrm{t}}. \tag{8}$$
4) *If* $(\mathrm{a}, \mathrm{b}) = (\mathrm{ave}, \mathrm{ave})$,
$$\frac{1}{M} \sum_{m \in [M]} \lambda_{\mathrm{u}}(m) \leq \epsilon_{\mathrm{u}}, \quad \frac{1}{M} \sum_{m \in [M]} \lambda_{\mathrm{t}}(m) \leq \epsilon_{\mathrm{t}}. \tag{9}$$

Definition 3. *A number $r \in \mathbb{R}$ is an $(\epsilon_{\mathrm{u}}, \epsilon_{\mathrm{t}})_{\mathrm{a,b}}$-*achievable erasure second-order coding rate *for the DMC W^n with capacity C if there exists a sequence of $(M_n, \epsilon_{\mathrm{u},n}, \epsilon_{\mathrm{t},n})_{\mathrm{a,b}}$-*erasure codes *such that*

$$\liminf_{n \to \infty} \frac{1}{\sqrt{n}} (\log M_n - nC) \geq r, \tag{10}$$

$$\limsup_{n \to \infty} \epsilon_{\mathrm{u},n} \leq \epsilon_{\mathrm{u}}, \quad \textit{and} \quad \limsup_{n \to \infty} \epsilon_{\mathrm{t},n} \leq \epsilon_{\mathrm{t}}. \tag{11}$$

The $(\epsilon_{\mathrm{u}}, \epsilon_{\mathrm{t}})_{\mathrm{a,b}}$-erasure second-order capacity $r^*_{\mathrm{era,a,b}}(\epsilon_{\mathrm{u}}, \epsilon_{\mathrm{t}})$ *is the supremum of all $(\epsilon_{\mathrm{u}}, \epsilon_{\mathrm{t}})_{\mathrm{a,b}}$-achievable erasure second-order coding rates.*

We now turn our attention to codes which allow their decoders to output a *list* of messages. Let $\binom{[M]}{j}$ be the set of subsets of $[M]$ of size j. Furthermore, we use the notation $\binom{[M]}{\leq L} := \cup_{0 \leq j \leq L} \binom{[M]}{j}$ to denote the set of subsets of $[M]$ of size not exceeding L.

Definition 4. *An* (M, L)-list code *for $W : \mathcal{X} \to \mathcal{Y}$ is a pair of mappings (f, φ) such that $f : [M] \to \mathcal{X}$ and $\varphi : \mathcal{Y} \to \binom{[M]}{\leq L}$. The* decoding regions *are denoted as $\mathcal{D}_m := \{y \in \mathcal{Y} : m \in \varphi(y)\}$ and the* conditional error probability *is defined as*

$$\lambda(m) := W(\mathcal{Y} \setminus \mathcal{D}_m | f(m)). \tag{12}$$

Definition 5. *An* $(M, L, \epsilon)_{\mathrm{a}}$-list code *for W is an (M, L)-list code for the same channel where if* $\mathrm{a} = \max$,

$$\max_{m \in [M]} \lambda(m) \leq \epsilon, \tag{13}$$

or if $\mathrm{a} = \mathrm{ave}$,

$$\frac{1}{M} \sum_{m \in [M]} \lambda(m) \leq \epsilon. \tag{14}$$

Definition 6. *A number $r \in \mathbb{R}$ is an $(l, \epsilon)_{\mathrm{a}}$-*achievable list second-order coding rate *for the DMC W^n with capacity C if there exists a sequence of $(M_n, L_n, \epsilon_n)_{\mathrm{a}}$-*list codes *such that in addition to* (10)*, the following hold*

$$\limsup_{n \to \infty} \frac{1}{\sqrt{n}} \log L_n \leq l, \quad \textit{and} \quad \limsup_{n \to \infty} \epsilon_n \leq \epsilon. \tag{15}$$

The $(l, \epsilon)_{\mathrm{a}}$-list second-order capacity $r^*_{\mathrm{list,a}}(l, \epsilon)$ *is the supremum of all $(l, \epsilon)_{\mathrm{a}}$-achievable list second-order coding rates.*

According to (15), we stipulate that the list size grows as $2^{\Theta(\sqrt{n})}$. This differs from previous works in which the list size is either constant [7], [15] or exponential [2], [4]–[7], [16]. In fact if the list size grows polynomially (e.g. n^α for some $\alpha > 0$), the central object of study would the third-order coding rate [11]–[14]. We defer this to future work.

III. Main Results

In this section, we summarize the main results of this paper. For simplicity, we assume that the DMC W satisfies $V_{\min} > 0$.

A. Decoding with Erasure Option

Theorem 1. *For any $0 < \epsilon_{\mathrm{u}} \leq \epsilon_{\mathrm{t}} < 1$,*

$$r^*_{\mathrm{era,a,b}}(\epsilon_{\mathrm{u}}, \epsilon_{\mathrm{t}}) = \sqrt{V_{\epsilon_{\mathrm{t}}}}\, \Phi^{-1}(\epsilon_{\mathrm{t}}), \tag{16}$$

where (a, b) can be any element in $\{\max, \mathrm{ave}\}^2$.

The proof of Theorem 1 can be found in Section IV-A.

A few comments are in order: First, Theorem 1 can be interpreted as follows. Suppose $M^*(W^n; \epsilon_{\mathrm{u}}, \epsilon_{\mathrm{t}})$ is the maximum number of codewords that can be transmitted over W^n with undetected and total error ϵ_{u} and ϵ_{t} respectively, then

$$\log M^*(W^n; \epsilon_{\mathrm{u}}, \epsilon_{\mathrm{t}}) = nC + \sqrt{nV_{\epsilon_{\mathrm{t}}}}\, \Phi^{-1}(\epsilon_{\mathrm{t}}) + o(\sqrt{n}). \tag{17}$$

Second, the direct part of the proof of Theorem 1 uses threshold decoding, i.e. declare that m is sent if a certain score function, the empirical mutual information, is higher than a threshold. If no message's score exceeds the threshold, then an erasure is declared. This simple rule is not the optimal one. The optimal rule was derived using a generalized version of the Neyman-Pearson lemma by Forney [1, Thm. 1], i.e.

$$\mathcal{D}^*_m := \left\{ \mathbf{y} : W^n(\mathbf{y} | f(m)) \geq \gamma \sum_{\tilde{m} \in [M] \setminus \{m\}} W^n(\mathbf{y} | f(\tilde{m})) \right\}, \tag{18}$$

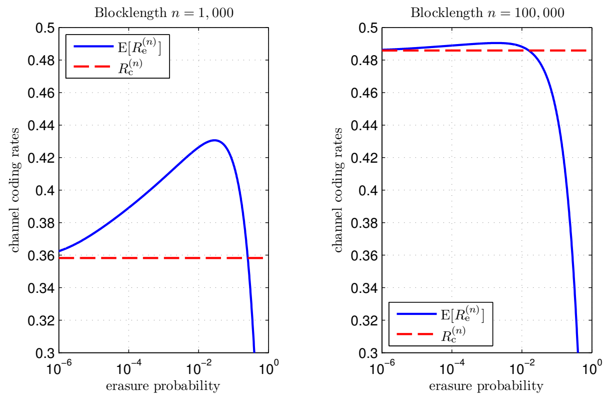

Fig. 1. Comparison of non-asymptotic rates with and without erasure option. Observe that $\mathbb{E}[R_{\mathrm{e}}^{(n)}]$ can be larger $R_{\mathrm{c}}^{(n)}$ for finite n.

for some $\gamma > 0$. However, this rule is difficult for second-order analysis. Forney also suggested the simpler rule [1, Eq. (11a)]

$$\mathcal{D}'_m := \left\{ \mathbf{y} : W^n(\mathbf{y}|f(m)) \geq \gamma \max_{\tilde{m} \in [M] \setminus \{m\}} W^n(\mathbf{y}|f(\tilde{m})) \right\}. \tag{19}$$

We analyzed this rule in the same way as one analyzes the random coding union (RCU) bound [9, Thm. 16] in the asymptotic setting but the analysis based on the rule in (19) could not be proved to be second-order optimal.

Third, the converse is based on Strassen's idea [8, Eq. (4.18)], connecting channel coding to hypothesis testing.

B. Expected Rate

From (17), we see that the backoff from the capacity at blocklength n is approximately $-\sqrt{V_{\epsilon_{\mathrm{t}}}/n}\,\Phi^{-1}(\epsilon_{\mathrm{t}})$ independent of ϵ_{u}. Thus, denoting $\epsilon_{\mathrm{e}} := \epsilon_{\mathrm{t}} - \epsilon_{\mathrm{u}}$ as the erasure probability, the *expected rate* in erasures decoding with blocklength n with undetected and total errors ϵ_{u} and ϵ_{t} respectively is

$$\mathbb{E}\big[R_{\mathrm{e}}^{(n)}\big] := (1 - \epsilon_{\mathrm{e}}) \left[C + \sqrt{\frac{V_{\epsilon_{\mathrm{t}}}}{n}} \Phi^{-1}(\epsilon_{\mathrm{t}}) \right], \tag{20}$$

assuming the Gaussian approximation is sufficiently accurate. It was numerically shown in [9] that the Gaussian approximation is accurate for moderate blocklengths. The reduction in rate by $1 - \epsilon_{\mathrm{e}}$ was observed by Forney [1, Eq. (49)]. One may use an *automatic repeat request (ARQ)* scheme[1] to resend information if it is erased. For channel coding with error ϵ_{u}, the non-asymptotic channel coding rate is approximated [9] by

$$R_{\mathrm{c}}^{(n)} := C + \sqrt{\frac{V_{\epsilon_{\mathrm{u}}}}{n}} \Phi^{-1}(\epsilon_{\mathrm{u}}). \tag{21}$$

Clearly if $0 < \epsilon_{\mathrm{u}} < \epsilon_{\mathrm{e}} < \epsilon_{\mathrm{t}} < 1$ are constants,

$$C = \lim_{n \to \infty} R_{\mathrm{c}}^{(n)} > \lim_{n \to \infty} \mathbb{E}\big[R_{\mathrm{e}}^{(n)}\big] = (1 - \epsilon_{\mathrm{e}})C, \tag{22}$$

so there is no advantage in allowing for erasures. However, in finite blocklength regime, for small enough ϵ_{e}, we may have

$$R_{\mathrm{c}}^{(n)} < \mathbb{E}\big[R_{\mathrm{e}}^{(n)}\big] \tag{23}$$

so erasure decoding may be advantageous *in expectation*. We illustrate this with an example.

In Fig. 1, we consider a binary symmetric channel (BSC) with crossover probability $p = 0.11$ so $C = 0.5$ bits/channel use and $V = 0.891$ bits^2/channel use. We keep the undetected error probability at $\epsilon_{\mathrm{u}} = 10^{-6}$ and vary the erasure error probability $\epsilon_{\mathrm{e}} \in [10^{-6}, 10^{-0.5}]$ and blocklength $n \in \{10^3, 10^5\}$. We observe that for some moderate erasure probabilities, the gain of erasure coding over channel coding is rather pronounced. This gain is reduced if (i) ϵ_{e} is increased because we retransmit the whole block more often on average via the use of decision feedback or (ii) n becomes large because the second-order term becomes less significant (cf. (22)).

C. List Decoding

Theorem 2. *For any $0 < \epsilon < 1$,*

$$r^*_{\mathrm{list,a}}(l, \epsilon) = l + \sqrt{V_\epsilon}\Phi^{-1}(\epsilon). \tag{24}$$

where a *can be any element in* $\{\max, \mathrm{ave}\}$.

The proof of Theorem 2 can be found in Section IV-B.

Theorem 2 shows that if we allow the list to grow as $2^{\sqrt{n}l}$, then the second-order capacity is increased by l. This is unsurprising and concurs with the intuition we obtain from the error exponent analysis in [2].

IV. Proofs

In this section, we prove both theorems. For the direct parts, we show that the second-order capacities are also *universally attainable*–i.e. the codes do not require channel knowledge.[2]

A. Decoding with Erasure Option: Proof of Theorem 1

1) Converse Part: Before we start, let us define the function

$$\beta_\alpha(P, Q) := \min_{\mathcal{D} \subset \mathcal{Y} : P(\mathcal{D}) \geq \alpha} Q(\mathcal{D}), \tag{25}$$

where P, Q are two probability measures on $\mathcal{Y}$. Let us fix any $(M_n, \epsilon_{\mathrm{u},n}, \epsilon_{\mathrm{t},n})_{\max,\max}$-erasure code for W^n. This means that

$$\min_{m \in [M_n]} W^n(\mathcal{D}_m | f(m)) \geq 1 - \epsilon_{\mathrm{t},n}. \tag{26}$$

In addition, we assume that the code is constant composition, i.e. all codewords are of the same type P. Then, for any permutation invariant output distribution $Q \in \mathcal{P}(\mathcal{Y}^n)$, we have

$$1 \geq \sum_{m=1}^{M_n} Q(\mathcal{D}_m) \geq \sum_{m=1}^{M_n} \min_{\substack{\mathcal{D} \subset \mathcal{Y}^n : \\ W^n(\mathcal{D}|f(m)) \geq 1 - \epsilon_{\mathrm{t},n}}} Q(\mathcal{D}) \tag{27}$$

$$= \sum_{m=1}^{M_n} \beta_{1-\epsilon_{\mathrm{t},n}}(W^n(\cdot|f(m)), Q) = M_n \beta_{1-\epsilon_{\mathrm{t},n}}(W^n(\cdot|\mathbf{x}), Q), \tag{28}$$

where the final equality follows from the fact that $\beta_{1-\epsilon_{\mathrm{t},n}}(W^n(\cdot|f(m)), Q)$ does not depend on m for permutation invariant Q (which is what we choose Q to be) and constant composition codes [8]. We also used $\mathbf{x}$ to denote any

[1] Forney [1] calls this *decision feedback*, but nowadays, the term ARQ is more common.

[2] A simpler proof based on thresholding the likelihood can also be be used; however, channel knowledge is required.

element in the type class $\mathcal{T}_P$. Choose $Q = (PW)^n$. Since $\epsilon_{\mathrm{t},n}$ satisfies (11), for every $\eta \in (0, 1-\epsilon_{\mathrm{t}})$, there exists sufficiently large n such that $\epsilon_{\mathrm{t},n} \le \epsilon_{\mathrm{t}} + \eta$. Hence, from (28),

$$M_n \le \beta^{-1}_{1-\epsilon_{\mathrm{t}}-\eta}(W^n(\cdot|\mathbf{x}), Q). \tag{29}$$

Now one may use the asymptotic expansion of $-\log \beta_{1-\epsilon}(W^n(\cdot|\mathbf{x}), Q)$ [8, Sec. 2] and continuity arguments [11, Lem. 7] to complete the converse for the $(\max, \max)$ setting. Removing the constant composition assumption simply adds an $O(\frac{\log n}{n})$ term to the rate which is inconsequential for second-order asymptotics. To get to the $(\mathrm{ave}, \mathrm{ave})$ setting, we use an expurgation argument as in [9, Eq. (284)] or the converse technique in [11] which is directly applicable to the average error criterion.

2) Direct Part: It suffices to prove that $\sqrt{V_{\epsilon_{\mathrm{e}}}}\Phi^{-1}(\epsilon_{\mathrm{e}})$ is an achievable $(0, \epsilon_{\mathrm{e}})_{\mathrm{ave,ave}}$-erasure second-order coding rate for W^n. Indeed, here the total error $\epsilon_{\mathrm{t}} = \epsilon_{\mathrm{e}}$. However, any achievable $(0, \epsilon_{\mathrm{t}})_{\mathrm{ave,ave}}$-second-order coding rate is also an achievable $(\epsilon_{\mathrm{u}}, \epsilon_{\mathrm{t}})_{\mathrm{ave,ave}}$-second-order coding rate for any $\epsilon_{\mathrm{u}} \in [0, \epsilon_{\mathrm{t}})$. Furthermore, by an expurgation argument, the same statement can be proved under the $(\max, \max)$ setting.

Fix a type $P \in \mathcal{P}_n(\mathcal{X})$. Generate M_n codewords uniformly at random from the type class $\mathcal{T}_P$. Denote the random codebook as $\{\mathbf{X}(m) : m \in [M_n]\} \subset \mathcal{T}_P$. The number M_n is to be chosen later. At the receiver, we decode to $\hat{m}$ if and only if $\hat{m}$ is the unique message to satisfy

$$\hat{I}(\mathbf{X}(\hat{m}) \wedge \mathbf{Y}) \ge \gamma \tag{30}$$

where $\hat{I}(\mathbf{x} \wedge \mathbf{y})$ is the empirical mutual information of $(\mathbf{x}, \mathbf{y})$, i.e. the mutual information of the random variables $(\tilde{X}, \tilde{Y})$ whose distribution is the joint type $P_{\mathbf{x},\mathbf{y}}$ and γ is to be chosen later. Assume as usual that the true message $m = 1$. We use the following elementary result which can be proved using the method of types. See the proof of [16, Lem. 10.1] for details.

Lemma 3. *Let $P \in \mathcal{P}_n(\mathcal{X})$ be any n-type. Let $\mathbf{X}(1)$ and $\mathbf{X}(2)$ be selected independently and uniformly at random from the type class $\mathcal{T}_P$. Let $\mathbf{Y}$ be the channel output when $\mathbf{X}(1)$ is the input, i.e. $\mathbf{Y} \sim \prod_{i=1}^n W(\,\cdot\,|X_i(1))$. Then, for every $\gamma > 0$,*

$$\Pr\left[\hat{I}(\mathbf{X}(2) \wedge \mathbf{Y}) > \gamma\right] \le (n+1)^{|\mathcal{X}|+|\mathcal{X}||\mathcal{Y}|} 2^{-n\gamma}. \tag{31}$$

The undetected error probability is bounded as

$$\begin{aligned}\Pr[\mathcal{E}_{\mathrm{u}}] &\le \Pr\left[\max_{m\ge 2} \hat{I}(\mathbf{X}(m) \wedge \mathbf{Y}) \ge \gamma\right] && (32)\\ &\le (M_n - 1)\Pr\left[\hat{I}(\mathbf{X}(2) \wedge \mathbf{Y}) \ge \gamma\right] && (33)\\ &\le (M_n - 1)(n+1)^{|\mathcal{X}||\mathcal{Y}|+|\mathcal{X}|} 2^{-n\gamma} && (34)\end{aligned}$$

where (34) follows from Lemma 3.

We let the random conditional type of $\mathbf{Y}$ given $\mathbf{X}(1)$ be V. The erasure probability can be bounded as

$$\begin{aligned}\Pr[\mathcal{E}_{\mathrm{e}}] &= \Pr\left[\hat{I}(\mathbf{X}(1 \wedge \mathbf{Y}) \le \gamma\right] && (35)\\ &= \Pr\Bigg[I(P, W) + \sum_{x,y}(V(y|x) - W(y|x)) I'_W(y|x) \\ &\qquad\qquad + O(\|V - W\|^2) \le \gamma\Bigg] && (36)\end{aligned}$$

where the final step follows by Taylor expanding $V \mapsto I(P, V)$ around $V = W$ and $I'_W(y|x) := \frac{\partial I(P,V)}{\partial V(y|x)}\big|_{V=W}$. We also can bound the remainder term uniformly [17] yielding

$$\begin{aligned}\Pr[\mathcal{E}_{\mathrm{e}}] \le \Pr\Bigg[&I(P, W) + \sum_{x,y}(V(y|x) - W(y|x)) I'_W(y|x) \\ &\le \gamma + O\left(\frac{\log n}{n}\right)\Bigg] + O(n^{-2}). && (37)\end{aligned}$$

Wang-Ingber-Kochman [17] computed the relevant first- and second-order statistics of the random variable $\sum_{x,y}(V(y|x) - W(y|x)) I'_W(y|x)$ allowing us to applying the Berry-Esseen theorem [18, Ch. XVI.5], leading to

$$\Pr[\mathcal{E}_{\mathrm{e}}] \le \Phi\left(\frac{\gamma + O(\frac{\log n}{n}) - I(P, W)}{\sqrt{V(P, W)/n}}\right) + O(n^{-1/2}). \tag{38}$$

Hence, we set

$$\gamma = I(P, W) + \sqrt{\frac{V(P, W)}{n}}\Phi^{-1}(\epsilon'_{\mathrm{e}}) \tag{39}$$

to assert that

$$\Pr[\mathcal{E}_{\mathrm{e}}] \le \epsilon'_{\mathrm{e}} + O(n^{-1/2}). \tag{40}$$

We then set M_n to be the smallest integer satisfying

$$\log M_n \ge n\gamma - \left(|\mathcal{X}||\mathcal{Y}| + |\mathcal{X}| + \frac{1}{2}\right)\log n. \tag{41}$$

Then we may assert from (34) that

$$\Pr[\mathcal{E}_{\mathrm{u}}] \le n^{-1/2}. \tag{42}$$

Hence, we have that

$$\log M_n \ge nI(P, W) + \sqrt{nV(P, W)}\Phi^{-1}(\epsilon'_{\mathrm{e}}) + O(\log n). \tag{43}$$

By choosing P to be a type that is the closest to an input distribution achieving $V_{\epsilon_{\mathrm{e}}}$, we obtain, by the usual approximation arguments [11, Lem. 7],

$$\log M_n \ge nC + \sqrt{nV_{\epsilon'_{\mathrm{e}}}}\Phi^{-1}(\epsilon'_{\mathrm{e}}) + O(\log n). \tag{44}$$

We have proved the the random ensemble satisfies (40) and (42) but it is not clear yet there exists a single deterministic code that satisfies the same two bounds. To show this, let $\theta \in (0, 1)$. Set $\epsilon_{\mathrm{e}} := \frac{1}{1-\theta}(\epsilon'_{\mathrm{e}} + O(n^{-1/2}))$ and $\epsilon_{\mathrm{u}} := \frac{1}{\theta}n^{-1/2}$. Then, making the expectation over the random code $\mathcal{C}$ explicit, (40) and (42) can be written as

$$\mathbb{E}[\Pr[\mathcal{E}_{\mathrm{e}}|\mathcal{C}]] \le (1-\theta)\epsilon_{\mathrm{e}}, \quad \mathbb{E}[\Pr[\mathcal{E}_{\mathrm{u}}|\mathcal{C}]] \le \theta\epsilon_{\mathrm{u}}. \tag{45}$$

Put $\eta := \theta/2$. By Markov's inequality,

$$\Pr(\mathcal{A}) \le 1 - \theta, \quad \Pr(\mathcal{B}) \le \theta - \eta, \tag{46}$$

where event $\mathcal{A} := \{\Pr[\mathcal{E}_{\mathrm{e}}|\mathcal{C}] > \frac{1}{1-\theta}\mathbb{E}[\Pr[\mathcal{E}_{\mathrm{e}}|\mathcal{C}]]\}$ and event $\mathcal{B} := \{\Pr[\mathcal{E}_{\mathrm{u}}|\mathcal{C}] > \frac{1}{\theta-\eta}\mathbb{E}[\Pr[\mathcal{E}_{\mathrm{u}}|\mathcal{C}]]\}$. This implies that there exists a *deterministic* code C^*_n for which

$$\Pr[\mathcal{E}_{\mathrm{e}}|\mathsf{C}^*_n] \le \frac{1}{1-\theta}\mathbb{E}[\Pr[\mathcal{E}_{\mathrm{e}}|\mathcal{C}]] \le \epsilon_{\mathrm{e}}, \quad \text{and} \tag{47}$$

$$\Pr[\mathcal{E}_{\mathrm{u}}|\mathsf{C}^*_n] \le \frac{1}{\theta-\eta}\mathbb{E}[\Pr[\mathcal{E}_{\mathrm{u}}|\mathcal{C}]] \le \frac{\theta}{\theta-\eta}\epsilon_{\mathrm{u}} = \frac{2}{\theta\sqrt{n}}. \tag{48}$$

Letting $n \to \infty$, we conclude there exists a sequence of deterministic codes $\{\mathsf{C}_n^*\}_{n\geq 1}$ such that

$$\limsup_{n\to\infty} \Pr[\mathcal{E}_{\mathrm{e}}|\mathsf{C}_n^*] \leq \epsilon_{\mathrm{e}}, \quad \lim_{n\to\infty} \Pr[\mathcal{E}_{\mathrm{u}}|\mathsf{C}_n^*] = 0 \tag{49}$$

and per (44) and the relation between ϵ_{e} and ϵ_{e}',

$$\liminf_{n\to\infty} \frac{1}{\sqrt{n}}(\log M_n - nC) \geq \sqrt{V_{(1-\theta)\epsilon_{\mathrm{e}}}}\Phi^{-1}((1-\theta)\epsilon_{\mathrm{e}}). \tag{50}$$

Now take $\theta \downarrow 0$ to complete the proof.

B. List Decoding: Proof of Theorem 2

1) Converse Part: First by the definition of an (M_n, L_n, ϵ_n)-list code, for every $\mathbf{y} \in \mathcal{Y}^n$, we have

$$L_n \geq \sum_{m=1}^{M_n} \mathbf{1}\{\mathbf{y} \in \mathcal{D}_m\} \tag{51}$$

where $\mathcal{D}_m \subset \mathcal{Y}^n$ is the set of all $\mathbf{y}$ such that m is contained in the list $\varphi(\mathbf{y})$. Let $Q \in \mathcal{P}(\mathcal{Y}^n)$ be any output distribution. Multiplying by $Q(\mathbf{y})$ in (51) and summing over all $\mathbf{y}$ yields

$$L_n \geq \sum_{\mathbf{y}} Q(\mathbf{y}) \sum_{m=1}^{M_n} \mathbf{1}\{\mathbf{y} \in \mathcal{D}_m\} = \sum_{m=1}^{M_n} Q(\mathcal{D}_m). \tag{52}$$

Now emulate the erasures setting in Sec. IV-A1. A similar non-asymptotic bound for list decoding was also obtained by Kostina-Verdú in [19, Thm. 4].

2) Direct Part: The codebook generation is the same as in Sec. IV-A2. Now the decoder outputs *all* messages m whose empirical mutual information exceeds $\gamma \geq 0$, i.e. the list is

$$\mathcal{L} := \{m \in [M_n] : \hat{I}(\mathbf{X}(m) \wedge \mathbf{Y}) \geq \gamma\}. \tag{53}$$

We now analyze the error probability assuming that message $m = 1$ was sent. The two error events are

$$\mathcal{E}_1 := \{\hat{I}(\mathbf{X}(1) \wedge \mathbf{Y}) < \gamma\}, \text{ and } \mathcal{E}_2 := \{|\mathcal{L}| > 2^{\sqrt{n}l}\}. \tag{54}$$

The probability of $\mathcal{E}_1$ can be analyzed using the Berry-Esseen theorem [18, Ch. XVI.5] just as for the erasures case. Indeed from the steps leading to (38), we have

$$\Pr[\mathcal{E}_1] \leq \Phi\left(\frac{\gamma + O(\frac{\log n}{n}) - I(P,W)}{\sqrt{V(P,W)/n}}\right) + O(n^{-1/2}). \tag{55}$$

Consider the expectation of the size of the list. Indeed

$$\begin{aligned} \mathbb{E}[|\mathcal{L}|] &= \mathbb{E}\left[\sum_{m=1}^{M_n} \mathbf{1}\{\hat{I}(\mathbf{X}(m) \wedge \mathbf{Y}) \geq \gamma\}\right] & (56) \\ &\leq 1 + \sum_{m=2}^{M_n} \Pr\left[\hat{I}(\mathbf{X}(m) \wedge \mathbf{Y}) \geq \gamma\right] & (57) \\ &\leq 1 + M_n(n+1)^{|\mathcal{X}||\mathcal{Y}|+|\mathcal{X}|}2^{-n\gamma} & (58) \end{aligned}$$

where (58) follows from Lemma 3. By Markov's inequality, the probability of the second error event can be bounded as

$$\Pr[\mathcal{E}_2] \leq \frac{\mathbb{E}[|\mathcal{L}|]}{2^{\sqrt{n}l}} \leq \frac{1 + M_n(n+1)^{|\mathcal{X}||\mathcal{Y}|+|\mathcal{X}|}2^{-n\gamma}}{2^{\sqrt{n}l}}. \tag{59}$$

Now choose for some large enough $K > 0$,

$$\gamma = I(P,W) + \sqrt{\frac{V(P,W)}{n}}\Phi^{-1}(\epsilon) - \frac{K \log n}{n}. \tag{60}$$

This ensures that $\Pr[\mathcal{E}_1] \leq \epsilon - \frac{2}{\sqrt{n}}$ for large enough n. Furthermore, choose M_n to be smallest integer satisfying

$$\log M_n \geq n\gamma + \sqrt{n}l - \left(|\mathcal{X}||\mathcal{Y}| + |\mathcal{X}| + \frac{1}{2}\right)\log n, \tag{61}$$

ensuring that $\Pr[\mathcal{E}_2] \leq \frac{2}{\sqrt{n}}$ and thus $\Pr[\mathcal{E}_1 \cup \mathcal{E}_2] \leq \epsilon$. Approximating P^* achieving V_ϵ with a type P completes the proof as we have demonstrated that there exists a code with list size no larger than $2^{\sqrt{n}l}$, error probability not exceeding ϵ, and second-order coding rate at least $l + \sqrt{V_\epsilon}\Phi^{-1}(\epsilon)$.

Note that unlike the erasures setting, in this case we do not need to augment the proof with the argument involving Markov's inequality (cf. argument leading to (48)) because here, there is only a single error criterion.

Acknowledgements: VT would like to acknowledge helpful discussions with M. Hayashi, M. Tomamichel and J. Scarlett.

REFERENCES

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

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

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

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

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

[6] N. Merhav. Error exponents of erasure list decoding revisited via moments of distance enumerators. *IEEE Trans. on Inf. Th.*, 54:4439–4447, Oct 2008.

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

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

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

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

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

[12] P. Moulin. The log-volume of optimal codes for memoryless channels, within a few nats. Nov 2013. arXiv:1311.0181.

[13] V. Y. F. Tan and M. Tomamichel. The third-order term in the normal approximation for the AWGN channel. 2013. arXiv:1311.2237v2.

[14] Y. Altuğ and A. Wagner. The third-order term in the normal approximation for singular channels. 2013. arXiv:309.5126.

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

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

[17] D. Wang, A. Ingber, and Y. Kochman. The dispersion of joint source-channel coding. In *Allerton Conference*, 2011. arXiv:1109.6310.

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

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