The Likelihood Decoder: Error Exponents and Mismatch

Jonathan Scarlett
Laboratory for Information and Inference Systems
École Polytechnique Fédérale de Lausanne
jmscarlett@gmail.com

Alfonso Martinez
Universitat Pompeu Fabra
alfonso.martinez@ieee.org

Albert Guillén i Fàbregas
ICREA & Universitat Pompeu Fabra
University of Cambridge
guillen@ieee.org

***Abstract*—This paper studies likelihood decoding for channel coding over discrete memoryless channels. It is shown that the likelihood decoder recovers the same random-coding error exponents as the maximum-likelihood decoder for i.i.d. and constant-composition random codes. The role of mismatch in likelihood decoding is studied, and the notion of the mismatched likelihood decoder capacity is introduced. It is shown, both in the case of random coding and optimized codebooks, that the mismatched likelihood decoder can lead to strictly worse achievable rates and error exponents compared to the corresponding mismatched maximum-metric decoder.**

I. Introduction

Channel coding theorems can be proved using a variety of decoders, including joint typicality decoding [1], maximum-likelihood (ML) decoding [2], and threshold decoding [3]. Another alternative that has recently gained interest is the likelihood decoder [4], which is a stochastic decoder such that the probability of choosing a given codeword is proportional to its likelihood under the channel law. This decoder has been shown to simplify the derivations of a variety of asymptotic achievability bounds in network information theory [4], and analogous likelihood encoders have proved to be useful in the context of lossy compression [5]. The likelihood decoder is a special case of the *pretty good measurement* in quantum information theory [6], [7].

For the point-to-point channel coding problem, it was shown in [4] that this decoder not only yields the channel capacity of discrete memoryless channels (DMCs), but also the channel dispersion [8], [9]. On the other hand, existing bounds are not powerful enough to attain the best known random-coding error exponents [2], [10] in general. One of our contributions is a refined analysis of that in [4] that yields the random-coding error exponents of optimal ML decoding for both i.i.d. and constant-composition random coding.

The main focus of this paper is on the role of mismatch for the likelihood decoder. We let the decoder choose each codeword in proportion to an arbitrary function on the input and output alphabets, possibly differing from the channel likelihood. We introduce the notion of the mismatched likelihood decoder capacity, which is closely related to the notion of the mismatched capacity for maximum-metric decoding [11]–[14]. In contrast to the matched case, we show that the mismatched likelihood decoder can perform worse than the corresponding mismatched maximum-metric decoder, in terms of both achievable rates and error exponents.

Notation: The marginals of a joint distribution $P_{XY}(x,y)$ are denoted by $P_X(x)$ and $P_Y(y)$. We write $P_X = \widetilde{P}_X$ to denote element-wise equality between two distributions on the same alphabet. Given a distribution $Q(x)$ and a conditional distribution $W(y|x)$, we write $Q \times W$ to denote the joint distribution $Q(x)W(y|x)$. The set of all probability distributions on $\mathcal{X}$ is denoted by $\mathcal{P}(\mathcal{X})$, and the set of all empirical distributions for vectors in $\mathcal{X}^n$ (i.e. types [10], [15]) is denoted by $\mathcal{P}_n(\mathcal{X})$. For a given $Q \in \mathcal{P}_n(\mathcal{X})$, the type class $T^n(Q)$ is defined to be the set of all sequences in $\mathcal{X}^n$ with type Q. For two sequences $f(n)$ and $g(n)$, we write $f(n) \doteq g(n)$ if $\lim_{n\to\infty} \frac{1}{n} \log \frac{f(n)}{g(n)} = 0$. All rates are in units of nats except in the examples, where bits are used. We define $[\cdot]^+ = \max\{0,\cdot\}$, and denote the indicator function by $\mathbb{1}\{\cdot\}$.

A. Problem Setup

We consider a DMC on the alphabets $\mathcal{X}$ and $\mathcal{Y}$ described by $W(y|x)$, yielding an n-letter transition law given by $W^n(\boldsymbol{y}|\boldsymbol{x}) \triangleq \prod_{i=1}^n W(y_i|x_i)$. The encoder receives a message m equiprobable on $\{1,\dots,M\}$ and transmits the corresponding codeword from a codebook $\mathcal{C} = \{\boldsymbol{x}^{(1)},\dots,\boldsymbol{x}^{(M)}\}$. Given the random output sequence $\boldsymbol{Y}$ and the codebook, the (possibly mismatched) likelihood decoder randomly outputs a message according to the rule

$$\mathbb{P}[\hat{m} = j \,|\, \boldsymbol{Y} = \boldsymbol{y}] = \frac{q^n(\boldsymbol{x}^{(j)},\boldsymbol{y})}{\sum_{j'=1}^M q^n(\boldsymbol{x}^{(j')},\boldsymbol{y})}, \tag{1}$$

where $q^n(\boldsymbol{x},\boldsymbol{y}) \triangleq \prod_{i=1}^n q(x_i,y_i)$ for some non-negative function $q(\cdot,\cdot)$ called the decoding metric. The error probability is given by $p_{\mathrm{e}}(\mathcal{C}) = \mathbb{P}[\hat{m} \ne m]$, where the probability is over the message, the channel, and the decoder. Setting $q(x,y) = W(y|x)$ recovers the decoder studied in [4].

A rate R is said to be achievable if, for all $\delta > 0$, there exists a sequence of codebooks $\mathcal{C}_n$ with at least $\exp(n(R-\delta))$ codewords of length n such that $p_{\mathrm{e}}(\mathcal{C}_n) \to 0$ under the decoding rule in (1). The mismatched likelihood decoder capacity $\widetilde{C}_{\mathrm{M}}$ of (W,q) is defined to be the supremum of all

This work has been funded in part by the European Research Council under ERC grant agreement 259663, by the European Union's 7th Framework Programme under grant agreement 303633 and by the Spanish Ministry of Economy and Competitiveness under grants RYC-2011-08150 and TEC2012-38800-C03-03.

achievable rates. We say that $\widetilde{E}(R)$ is an achievable error exponent if there exist sequences of codebooks $\mathcal{C}_n$ at least $\exp(nR)$ codewords of length n such that

$$\liminf_{n\to\infty} -\frac{1}{n}\log p_{\mathrm{e}}(\mathcal{C}_n) \geq \widetilde{E}(R). \quad (2)$$

The decoding rule in (1) should be contrasted with the mismatched maximum-metric decoder, which chooses

$$\hat{m} = \arg\max_j q^n(\boldsymbol{x}^{(j)}, \boldsymbol{y}). \quad (3)$$

Throughout the paper, we distinguish between quantities associated with these decoders according to whether or not a tilde symbol is present; for example, the (classical) mismatched capacity is denoted by C_{M}, and is defied in the same way as $\widetilde{C}_{\mathrm{M}}$ according to the maximum-metric decoding rule. When $q(x,y) = W(y|x)$, (3) is the optimal ML decoding rule, and hence C_{M} is the (matched) capacity.

Our achievability results are based on random coding,where each codeword is independently generated according to some distribution $P_{\boldsymbol{X}}$. The random codewords are denoted by $\{\boldsymbol{X}^{(j)}\}_{j=1}^M$. The error probability averaged over the ensemble is denoted by $\overline{p}_e(n,M)$. We pay particular attention to constant-composition random coding, in which each codeword is drawn uniformly from the set of sequences having a given composition. That is,

$$P_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{|T^n(Q_n)|}\mathbb{1}\big\{\boldsymbol{x}\in T^n(Q_n)\big\}, \quad (4)$$

where $Q_n \in \mathcal{P}_n(\mathcal{X})$ is a type with the same support as Q such that $\max_x |Q(x) - Q_n(x)| \leq \frac{1}{n}$. Similarly to the case of mismatched maximum-metric decoding [16], all of our results have analogues for i.i.d. coding (i.e. $P_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^n Q(x_i)$) that can be proved in a similar (and often simpler) fashion.

II. Achievability via Dual-Domain Analysis

The following theorem presents a non-asymptotic bound obtained via a refinement of the analysis of Yassaee *et al.* [4], with an aim to improve the resulting error exponent. The main difference is the introduction of a parameter $s\in(0,1]$ that can be optimized. Our analysis is also related to that of Holevo [7], but there the parameter s is introduced into the decoding rule itself, rather than in the analysis alone.

Theorem 1. *For any channel W, metric q, and codebook distribution $P_{\boldsymbol{X}}$, the random-coding error probability satisfies*

$$\overline{p}_{\mathrm{e}}(n,M) \leq \min_{s\in(0,1]} \mathbb{E}\bigg[\min\bigg\{1, \frac{1}{s}(M-1)\frac{\mathbb{E}[q^n(\overline{\boldsymbol{X}},\boldsymbol{Y})^s\,|\,\boldsymbol{Y}]}{q^n(\boldsymbol{X},\boldsymbol{Y})^s}\bigg\}\bigg], \quad (5)$$

where $(\boldsymbol{X},\boldsymbol{Y},\overline{\boldsymbol{X}}) \sim P_{\boldsymbol{X}}(\boldsymbol{x})W^n(\boldsymbol{y}|\boldsymbol{x})P_{\boldsymbol{X}}(\overline{\boldsymbol{x}})$.

Proof: Fix $s\in(0,1]$. Assuming without loss of generality that $m=1$, the probability of a correct decision under random coding is given by

$$1-\overline{p}_{\mathrm{e}}$$
$$= \sum_{\boldsymbol{x},\boldsymbol{y}} P_{\boldsymbol{X}}(\boldsymbol{x})W^n(\boldsymbol{y}|\boldsymbol{x})\mathbb{E}\bigg[\frac{q^n(\boldsymbol{x},\boldsymbol{y})}{q^n(\boldsymbol{x},\boldsymbol{y}) + \sum_{\overline{m}\neq 1} q^n(\boldsymbol{y}|\boldsymbol{X}^{(\overline{m})})}\bigg] \quad (6)$$

$$= \sum_{\boldsymbol{x},\boldsymbol{y}} P_{\boldsymbol{X}}(\boldsymbol{x})W^n(\boldsymbol{y}|\boldsymbol{x})\mathbb{E}\Bigg[\frac{1}{1+\sum_{\overline{m}\neq 1}\frac{q^n(X^{(\overline{m})},\boldsymbol{y})}{q^n(\boldsymbol{x},\boldsymbol{y})}}\Bigg] \quad (7)$$

$$= \sum_{\boldsymbol{x},\boldsymbol{y}} P_{\boldsymbol{X}}(\boldsymbol{x})W^n(\boldsymbol{y}|\boldsymbol{x})\mathbb{E}\Bigg[\Bigg(\frac{1}{\Big(1+\sum_{\overline{m}\neq 1}\frac{q^n(X^{(\overline{m})},\boldsymbol{y})}{q^n(\boldsymbol{x},\boldsymbol{y})}\Big)^s}\Bigg)^{\frac{1}{s}}\Bigg] \quad (8)$$

$$\geq \sum_{\boldsymbol{x},\boldsymbol{y}} P_{\boldsymbol{X}}(\boldsymbol{x})W^n(\boldsymbol{y}|\boldsymbol{x})\Bigg(\frac{1}{\mathbb{E}\Big[\Big(1+\sum_{\overline{m}\neq 1}\frac{q^n(X^{(\overline{m})},\boldsymbol{y})}{q^n(\boldsymbol{x},\boldsymbol{y})}\Big)^s\Big]}\Bigg)^{\frac{1}{s}} \quad (9)$$

$$\geq \sum_{\boldsymbol{x},\boldsymbol{y}} P_{\boldsymbol{X}}(\boldsymbol{x})W^n(\boldsymbol{y}|\boldsymbol{x})\Bigg(\frac{1}{1+\sum_{\overline{m}\neq 1}\mathbb{E}\Big[\Big(\frac{q^n(X^{(\overline{m})},\boldsymbol{y})}{q^n(\boldsymbol{x},\boldsymbol{y})}\Big)^s\Big]}\Bigg)^{\frac{1}{s}} \quad (10)$$

$$= \sum_{\boldsymbol{x},\boldsymbol{y}} P_{\boldsymbol{X}}(\boldsymbol{x})W^n(\boldsymbol{y}|\boldsymbol{x})\Bigg(\frac{1}{1+(M-1)\mathbb{E}\Big[\Big(\frac{q^n(\overline{\boldsymbol{X}},\boldsymbol{y})}{q^n(\boldsymbol{x},\boldsymbol{y})}\Big)^s\Big]}\Bigg)^{\frac{1}{s}} \quad (11)$$

$$\geq 1 - \sum_{\boldsymbol{x},\boldsymbol{y}} P_{\boldsymbol{X}}(\boldsymbol{x})W^n(\boldsymbol{y}|\boldsymbol{x})\min\bigg\{1, \frac{1}{s}(M-1)\frac{\mathbb{E}[q^n(\overline{\boldsymbol{X}},\boldsymbol{y})^s]}{q^n(\boldsymbol{x},\boldsymbol{y})^s}\bigg\}, \quad (12)$$

where (9) follows from the convexity of $f(z) = \frac{1}{z^{1/s}}$ and Jensen's inequality, (10) follows since $\big(\sum_i a_i\big)^s \leq \sum_i a_i^s$ for $s\leq 1$, and (12) follows from the identity $1-\frac{1}{(1+z)^\alpha} \leq \min\{1,\alpha z\}$ for $\alpha > 0$ (the second term in the minimum is obtained by noting that the left-hand side is concave, equals zero at $z=0$, and has derivative α there). ■

The right-hand side of (5) coincides with a weakened version of the random-coding union (RCU) bound studied in [16], aside from the multiplicative factor of $\frac{1}{s}$. This factor does not affect the exponent; thus, choosing $P_{\boldsymbol{X}}$ as in (4) and following an identical analysis to [16], we have the following.

Theorem 2. *Under the constant-composition codeword distribution in* (4)*, the random-coding error probability satisfies* $\liminf_{n\to\infty} -\frac{1}{n}\log\overline{p}_{\mathrm{e}}(n,\lfloor e^{nR}\rfloor) \geq \widetilde{E}_{\mathrm{r}}(Q,R)$*, where*

$$\widetilde{E}_{\mathrm{r}}(Q,R) \triangleq \max_{\rho\in[0,1]} \widetilde{E}_0(Q,\rho) - \rho R, \quad (13)$$

and

$$\widetilde{E}_0(Q,\rho) \triangleq \sup_{s\in[0,1],a(\cdot)} -\sum_x Q(x) \times \log\sum_y W(y|x)\bigg(\frac{\sum_{\overline{x}} Q(\overline{x})q(\overline{x},y)^s e^{a(\overline{x})}}{q(x,y)^s e^{a(x)}}\bigg)^\rho. \quad (14)$$

Moreover, $\widetilde{E}_{\mathrm{r}}(Q,R) > 0$ *provided that* $R < \tilde{I}_{\mathrm{LM}}(Q)$*, where*

$$\widetilde{I}_{\mathrm{LM}}(Q) \triangleq \sup_{s\in[0,1],a(\cdot)} \sum_{x,y} Q(x)W(y|x)\log\frac{q(x,y)^s e^{a(x)}}{\sum_{\overline{x}} Q(\overline{x})q(\overline{x},y)^s e^{a(\overline{x})}}. \quad (15)$$

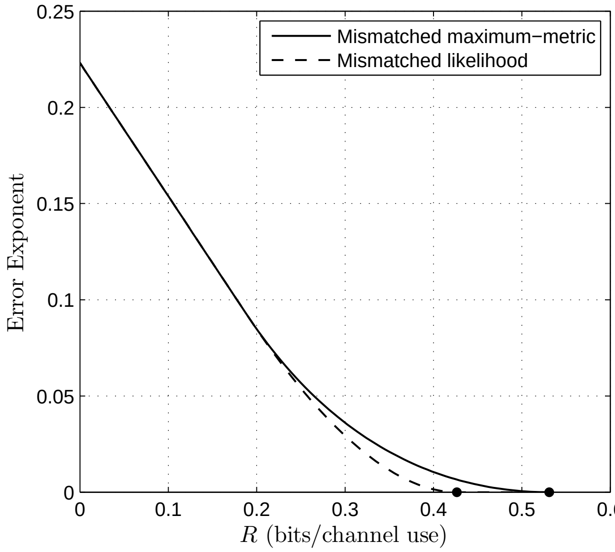

Figure 1. Mismatched error exponents $\widetilde{E}_{\rm r}$ and $E_{\rm r}$ for the binary symmetric channel with crossover probability 0.1, equiprobable inputs, and $q(x,y)=\sqrt{W(y|x)}$. The rates $\widetilde{I}_{\rm LM}$ and $I_{\rm LM}$ are marked on the horizontal axis.

These expressions resemble the dual-domain exponent and achievable rate given in [12], [16] for mismatched maximum-metric decoding; the latter are written as in (13)–(15) with the constraint $s\in[0,1]$ replaced by $s\geq 0$. For later reference, we denote these by $E_{\rm r}(Q,R)$ and $I_{\rm LM}(Q)$. We clearly have $\widetilde{E}_{\rm r}(Q,R)\leq E_{\rm r}(Q,R)$ and $\widetilde{I}_{\rm LM}(Q)\leq I_{\rm LM}(Q)$, i.e. the random-coding rates and exponents for the mismatched likelihood decoder never exceed those of the corresponding maximum-metric decoder.

In the case that $q=W$, the dual form of $\widetilde{E}_{\rm r}(Q,R)$ coincides with the constant-composition error exponent of Csiszár and Körner [10, Ch. 10] upon setting $s=\frac{1}{1+\rho}$ (which always lies in $[0,1]$) [16], and this exponent is the best possible under constant-composition random coding even for ML decoding [17]. Thus, the matched likelihood decoder achieves the same constant-composition random-coding exponent as the ML decoder. The same can be shown for the i.i.d. ensemble, with Gallager's exponent [2, Ch. 5] replacing that of [10, Ch. 10].

In contrast, in case of mismatched decoding, the restriction $s\in(0,1]$ does not always come without loss of optimality. As a simple example, if we set $q(x,y)=\sqrt{W(y|x)}$ then the optimal value is $s=\frac{2}{1+\rho}$. In Figure 1, we plot the achievable rates and error exponents associated with this choice, for the binary symmetric channel with crossover probability $\delta=0.1$ and an equiprobable input distribution Q. Under these choices, the mismatched maximum-metric decoder is the same as the ML decoder, and thus $I_{\rm LM}(Q)$ is the channel capacity and $E_{\rm r}(Q,R)$ equals the random-coding exponent of [10, Ch. 10]. We see that $\widetilde{E}_{\rm r}$ coincides with $E_{\rm r}$ at low rates (i.e. rates corresponding to $\frac{2}{1+\rho}\leq 1$), whereas the two behave differently at higher rates, in particular yielding $\widetilde{I}_{\rm LM}(Q)<I_{\rm LM}(Q)$.

III. Tightness via Primal-Domain Analysis

It is common in the literature on mismatched decoding and error exponents to consider both primal-domain expressions resembling those of Csiszár and Körner [10, Ch. 10], and dual-domain expressions resembling those of Gallager [2, Ch. 5]. In the previous section, we focused on the latter, whereas here we focus on the former. We begin with the following result formally stating the equivalent expressions.

Lemma 1. *For any DMC W, metric q, input distribution Q, and rate R, the exponent in* (13) *can be written as*

$$\widetilde{E}_{\rm r}(Q,R)=\min_{P_{XY}\,:\,P_X=Q}\ \min_{\widetilde{P}_{XY}\,:\,\widetilde{P}_X=Q,\widetilde{P}_Y=P_Y} D(P_{XY}\|Q\times W)+\Big[I_{\widetilde{P}}(X;Y)+\Big[\mathbb{E}\big[\log q(X,Y)\big]-\mathbb{E}_{\widetilde{P}}\big[\log q(\overline{X},Y)\big]\Big]^{+}-R\Big]^{+}, \quad (16)$$

Moreover, the achievable rate in (15) *can be written as*

$$\widetilde{I}_{\rm LM}(Q)=\min_{\widetilde{P}_{XY}\,:\,\widetilde{P}_X=Q,\widetilde{P}_Y=P_Y} I_{\widetilde{P}}(X;Y)+\Big[\mathbb{E}_P\big[\log q(X,Y)\big]-\mathbb{E}_{\widetilde{P}}\big[\log q(\overline{X},Y)\big]\Big]^{+}, \quad (17)$$

where $P_{XY}\triangleq Q\times W$.

Proof: The proof follows the approach of [12], [14], [16], and makes use of the identity $[\alpha]^+=\max_{s\in[0,1]}s\alpha$, Fan's minimax theorem [18], and Lagrange duality [19]. The details are omitted due to space constraints. ■

Equation (16) resembles the mismatched decoding error exponent of [20] for maximum-metric decoding, and (17) resembles the corresponding LM rate of Csiszár-Körner-Hui [20], [21]. More precisely, the latter are written as in (16)–(17) with the terms $\big[\mathbb{E}[\log q(X,Y)]-\mathbb{E}_{\widetilde{P}}[\log q(\overline{X},Y)]\big]^+$ removed, and with the constraint $\mathbb{E}_{\widetilde{P}}[\log q(X,Y)]\geq\mathbb{E}_P[\log q(X,Y)]$ added to the minimizations over $\widetilde{P}_{XY}$.

The main result of this section is the following theorem, providing a primal-domain analysis that proves that $\widetilde{E}_{\rm r}(Q,R)$ and $\widetilde{I}_{\rm LM}(Q)$ are tight with respect to the ensemble average for constant-composition random coding. That is, the random-coding error probability $\overline{p}_e$ cannot decay with an exponent exceeding $\widetilde{E}_{\rm r}$, and cannot vanish for rates exceeding $\widetilde{I}_{\rm LM}$. This concept was first studied in the context of mismatched decoding in [12]. We make use of type class enumerators, which have been shown to permit exponentially tight analyses in numerous source and channel coding problems (e.g. see [22]–[24]).

Theorem 3. *For any DMC W and metric q, under the constant-composition codeword distribution $P_{\boldsymbol{X}}$ in* (4)*, the random-coding error probability satisfies*

$$\lim_{n\to\infty}-\frac{1}{n}\log\overline{p}_{\rm e}(n,\lfloor e^{nR}\rfloor)=\widetilde{E}_{\rm r}(Q,R) \quad (18)$$

and

$$\sup\Big\{R\,:\,\lim_{n\to\infty}\overline{p}_e(n,\lfloor e^{nR}\rfloor)=0\Big\}=\widetilde{I}_{\rm LM}(Q). \quad (19)$$

Proof: From the results in the previous section, it suffices to prove only the converse parts. Due to space constraints, we focus our attention on (18). Equation (19) can be established by a straightforward combination of the arguments used in proving (18) and those used in [12, Thm. 1].

Applying the identity $\frac{1}{1+\alpha} \le 1 - \frac{1}{2}\min\{1,\alpha\}$ to (7) yields

$$\overline{p}_{\mathrm{e}} \ge \frac{1}{2}\sum_{\boldsymbol{x},\boldsymbol{y}} P_{\boldsymbol{X}}(\boldsymbol{x})W^n(\boldsymbol{y}|\boldsymbol{x})\mathbb{E}\bigg[\min\bigg\{1, \sum_{\overline{m}\neq 1}\frac{q^n(\boldsymbol{X}^{(\overline{m})},\boldsymbol{y})}{q^n(\boldsymbol{x},\boldsymbol{y})}\bigg\}\bigg]. \tag{20}$$

Consider a fixed pair such that $(\boldsymbol{x},\boldsymbol{y}) \in T^n(P_{XY})$, and let $N_{\boldsymbol{y}}(\widetilde{P}_{XY})$ be the random number of $\boldsymbol{X}^{(\overline{m})}$ ($\overline{m} \neq 1$) such that $(\boldsymbol{X}^{(\overline{m})},\boldsymbol{y}) \in T^n(\widetilde{P}_{XY})$. It follows that

$$\sum_{\overline{m}\neq 1}\frac{q^n(\boldsymbol{X}^{(\overline{m})},\boldsymbol{y})}{q^n(\boldsymbol{x},\boldsymbol{y})} = \frac{\sum_{\widetilde{P}_{XY}} N_{\boldsymbol{y}}(\widetilde{P}_{XY})q^n(\widetilde{P}_{XY})}{q^n(P_{XY})}, \tag{21}$$

where $q^n(\widetilde{P}_{XY})$ equals $q^n(\overline{\boldsymbol{x}},\boldsymbol{y})$ for an arbitrary pair $(\overline{\boldsymbol{x}},\boldsymbol{y}) \in T^n(\widetilde{P}_{XY})$. We proceed by applying standard exponentially tight steps based on the fact that there are only polynomially many terms in the summation [22]–[24]:

$$\begin{aligned}
&\mathbb{E}\bigg[\min\bigg\{1, \frac{\sum_{\widetilde{P}_{XY}} N_{\boldsymbol{y}}(\widetilde{P}_{XY})q^n(\widetilde{P}_{XY})}{q^n(P_{XY})}\bigg\}\bigg] \\
&\quad \doteq \mathbb{E}\bigg[\max_{\widetilde{P}_{XY}}\min\bigg\{1, \frac{N_{\boldsymbol{y}}(\widetilde{P}_{XY})q^n(\widetilde{P}_{XY})}{q^n(P_{XY})}\bigg\}\bigg] && (22)\\
&\quad \doteq \mathbb{E}\bigg[\sum_{\widetilde{P}_{XY}}\min\bigg\{1, \frac{N_{\boldsymbol{y}}(\widetilde{P}_{XY})q^n(\widetilde{P}_{XY})}{q^n(P_{XY})}\bigg\}\bigg] && (23)\\
&\quad \doteq \max_{\widetilde{P}_{XY}}\mathbb{E}\bigg[\min\bigg\{1, \frac{N_{\boldsymbol{y}}(\widetilde{P}_{XY})q^n(\widetilde{P}_{XY})}{q^n(P_{XY})}\bigg\}\bigg], && (24)
\end{aligned}$$

where the last step follows by first taking the summation outside the expectation.

Recalling that P_{XY} is the joint type of $(\boldsymbol{x},\boldsymbol{y})$ and $\widetilde{P}_{XY}$ is the joint type of an incorrect codeword with $\boldsymbol{y}$, we have

$$N_{\boldsymbol{y}}(\widetilde{P}_{XY}) = 0 \text{ if } \widetilde{P}_X \neq Q_n \text{ or } \widetilde{P}_Y \neq P_Y. \tag{25}$$

The first condition comes from the fact that all codewords have composition Q_n, and the second from the fact that the two joint types are associated with the same $\boldsymbol{y}$ sequence. We proceed by characterizing the behavior of the expectation in (24) for the joint types $\widetilde{P}_{XY}$ with $\widetilde{P}_X = Q_n$ and $\widetilde{P}_Y = P_Y$, using properties of type class enumerators. The arguments follow those of [24], [25], and are based on the fact that $N_{\boldsymbol{y}}(\widetilde{P}_{XY})$ has a Binomial distribution with exponentially many terms ($M-1 \doteq e^{nR}$) and an exponentially small success probability ($\mathbb{P}\big[(\overline{\boldsymbol{X}},\boldsymbol{y}) \in T^n(\widetilde{P}_{XY})\big] \doteq e^{-nI_{\widetilde{P}}(X;Y)}$). We omit the details and focus on the key ideas. Fixing $\delta > 0$ and writing $R = \frac{1}{n}\log M$, we have the following:

1) If $R > I_{\widetilde{P}}(X;Y) + \delta$, then $N_{\boldsymbol{y}}(\widetilde{P}_{XY}) \doteq Me^{-nI_{\widetilde{P}}(X;Y)}$ with probability approaching one super-exponentially fast, and hence the expectation in (24) has an error exponent of

$$\begin{aligned}
&-\frac{1}{n}\log\mathbb{E}\bigg[\min\bigg\{1, \frac{N_{\boldsymbol{y}}(\widetilde{P}_{XY})q^n(\widetilde{P}_{XY})}{q^n(P_{XY})}\bigg\}\bigg] \\
&\qquad \to \Big[I_{\widetilde{P}}(X;Y) - R + \Psi(P_{XY},\widetilde{P}_{XY})\Big]^+, && (26)
\end{aligned}$$

where

$$\Psi(P_{XY},\widetilde{P}_{XY}) \triangleq \mathbb{E}_P\big[\log q(X,Y)\big] - \mathbb{E}_{\widetilde{P}}\big[\log q(\overline{X},Y)\big]. \tag{27}$$

This case corresponds to the case that the mean of $N_{\boldsymbol{y}}(\widetilde{P}_{XY})$ is exponentially high, and hence sharp concentration is observed.

2) If $R < I_{\widetilde{P}}(X;Y) - \delta$, then $N_{\boldsymbol{y}}(\widetilde{P}_{XY}) \le e^{n2\delta}$ with probability approaching one super-exponentially fast. Combining the union bound with an associated tightness result for independent events [26, Lemma A.2], we have $\mathbb{P}\big[N_{\boldsymbol{y}}(\widetilde{P}_{XY}) > 0\big] \doteq Me^{-nI_{\widetilde{P}}(X;Y)}$. Combining these and taking $\delta \to 0$ gives

$$\begin{aligned}
&-\frac{1}{n}\log\mathbb{E}\bigg[\min\bigg\{1, \frac{N_{\boldsymbol{y}}(\widetilde{P}_{XY})q^n(\widetilde{P}_{XY})}{q^n(P_{XY})}\bigg\}\bigg] \\
&\qquad \to I_{\widetilde{P}}(X;Y) - R + \big[\Psi(P_{XY},\widetilde{P}_{XY})\big]^+. && (28)
\end{aligned}$$

Combining (25)–(28) and using a standard continuity argument to replace the minimizations over types by minimizations over all distributions [17], we obtain from (24) that

$$\begin{aligned}
&-\frac{1}{n}\log\mathbb{E}\bigg[\min\bigg\{1, \frac{\sum_{\widetilde{P}_{XY}} N_{\boldsymbol{y}}(\widetilde{P}_{XY})q^n(\widetilde{P}_{XY})}{q^n(P_{XY})}\bigg\}\bigg] \\
&\to \min\Bigg\{\min_{\substack{\widetilde{P}_{XY}\,:\,\widetilde{P}_X = Q,\\ \widetilde{P}_Y = P_Y,\\ I_{\widetilde{P}}(X;Y) \le R}} \Big[I_{\widetilde{P}}(X;Y) - R + \Psi(P_{XY},\widetilde{P}_{XY})\Big]^+, \\
&\qquad \min_{\substack{\widetilde{P}_{XY}\,:\,\widetilde{P}_X = Q,\\ \widetilde{P}_Y = P_Y,\\ I_{\widetilde{P}}(X;Y) \ge R}} I_{\widetilde{P}}(X;Y) - R + \big[\Psi(P_{XY},\widetilde{P}_{XY})\big]^+\Bigg\}. && (29)
\end{aligned}$$

We now note the following: (i) Surrounding the term $\Psi(P_{XY},\widetilde{P}_{XY})$ in the first $\min_{\widetilde{P}_{XY}}$ by $[\cdot]^+$ does not change the optimization, since the constraint $I_{\widetilde{P}}(X;Y) \le R$ means that the objective evaluates to zero whether this term is negative or zero; (ii) Surrounding the objective of the second $\min_{\widetilde{P}_{XY}}$ by $[\cdot]^+$ does not change the optimization, since the constraint $I_{\widetilde{P}}(X;Y) \ge R$ means that the objective is non-negative. Combining these observations with (20), recalling that P_{XY} denotes the joint type of $(\boldsymbol{x},\boldsymbol{y})$, and using the standard property of types [10, Ch. 2]

$$\mathbb{P}\big[(\boldsymbol{X},\boldsymbol{Y}) \in T^n(P_{XY})\big] \doteq e^{-nD(P_{XY}\|Q\times W)}, \tag{30}$$

we obtain (16). ∎

IV. Further Properties of $\widetilde{C}_{\mathrm{M}}$

In the following theorem, we present further properties comparing the mismatched likelihood decoder capacity $\widetilde{C}_{\mathrm{M}}$ and the classical mismatched capacity C_{M}.

Theorem 4. *The following statements hold:*

1) *For any (W,q) we have $\widetilde{C}_{\mathrm{M}} \le C_{\mathrm{M}}$.*
2) *There exist pairs (W,q) such that $\widetilde{C}_{\mathrm{M}} < C_{\mathrm{M}}$.*
3) *For any (W,q,Q), $\widetilde{I}_{\mathrm{LM}}(Q) > 0 \iff I_{\mathrm{LM}}(Q) > 0$.*
4) *For any (W,q), $\widetilde{C}_{\mathrm{M}} > 0 \iff C_{\mathrm{M}} > 0$.*

Proof: Let $P_{\boldsymbol{XY}}$ be the joint distribution induced by an arbitrary codebook and the channel. For the first part, we consider the set of all $(\boldsymbol{x},\boldsymbol{y})$ pairs such that $\boldsymbol{x}$ is not the unique maximum-metric codeword. Each such pair contributes at most $P_{\boldsymbol{XY}}(\boldsymbol{x},\boldsymbol{y})$ to the error probability of the maximum-metric decoder, and at least $\frac{1}{2}P_{\boldsymbol{XY}}(\boldsymbol{x},\boldsymbol{y})$ to the error probability of the likelihood decoder (since the best case scenario is that there is only one other codeword with an equal metric, and none with a higher metric). This means that the error probability for the likelihood decoder is at least half of that of the maximum metric decoder, and the claim $\widetilde{C}_{\mathrm{M}} \le C_{\mathrm{M}}$ follows.

For the second part, we consider W and q corresponding to binary symmetric channels with crossover probabilities δ and δ' respectively, with $0 < \delta < \delta' < 0.5$. The maximum-metric capacity is $\log 2 - H_2(\delta)$ (with H_2 denoting the binary entropy function), since (3) corresponds to minimum Hamming distance decoding for any such δ'. For the mismatched likelihood decoder, the probability of correct decoding is

$$1 - p_{\mathrm{e}} = \frac{1}{M}\sum_{\boldsymbol{x},\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x}) \frac{1}{1 + \frac{\sum_{\boldsymbol{x}' \ne \boldsymbol{x}} q^n(\boldsymbol{x}',\boldsymbol{y})}{q^n(\boldsymbol{x},\boldsymbol{y})}} \tag{31}$$

$$\le \frac{1}{M}\sum_{\boldsymbol{x},\boldsymbol{y}} W^n(\boldsymbol{y}|\boldsymbol{x}) \frac{1}{1 + \frac{\sum_{\boldsymbol{x}' \ne \boldsymbol{x}} (\delta')^n}{(1-\delta')^n}} \tag{32}$$

$$= \frac{1}{1 + (M-1)\frac{(\delta')^n}{(1-\delta')^n}}, \tag{33}$$

where (32) follows since $q(x,y)$ only takes values in $\{\delta', 1-\delta'\}$. We conclude that in order for $\overline{p}_e$ to vanish, it is necessary that $R < \log\frac{1-\delta'}{\delta'}$, regardless of the value of δ. This bound tends to zero as $\delta' \to \frac{1}{2}$, and we conclude that for any given δ, there exists a threshold $\delta_{\min} \in (\delta, 0.5)$ such that the mismatched likelihood capacity is less than the mismatched maximum-metric capacity whenever $\delta' \in (\delta_{\min}, 0.5)$.

For the third part, we note that the objective in (17) equals zero if and only if $I_{\widetilde{P}}(X;Y) = 0$ (and hence $\widetilde{P}_{XY} = Q \times P_Y$) and the expectation is non-positive. This yields the condition

$$\mathbb{E}_{Q\times P_Y}[\log q(X,Y)] \ge \mathbb{E}_{Q\times W}[\log q(X,Y)], \tag{34}$$

which matches the necessary and sufficient condition for $I_{\mathrm{LM}}(Q) = 0$ given in [11].

The fourth part follows immediately from the third due to the fact that $\max_Q I_{\mathrm{LM}}(Q) > 0$ if and only if $C_{\mathrm{M}} > 0$ [11]. ∎

The second part of Theorem 4 shows that the weakness of the mismatched likelihood decoder compared to the maximum-metric decoder is not limited to the random coding case. This can be understood at an intuitive level by considering the example given in the proof: As $\delta' \to 0.5$, the distribution in (1) becomes closer to uniform, and the uniform distribution yields an error probability of $1 - \frac{1}{M}$. However, this intuition does not establish the claim in the theorem, due to the ordering of the limits of $\delta' \to 0.5$ and $n \to \infty$.

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. Journal*, vol. 27, pp. 379–423, July and Oct. 1948.

[2] R. Gallager, *Information Theory and Reliable Communication*. John Wiley & Sons, 1968.

[3] A. Feinstein, "A new basic theorem of information theory," *IRE Prof. Group. on Inf. Theory*, vol. 4, no. 4, pp. 2–22, Sept. 1954.

[4] M. H. Yassaee, M. R. Aref, and A. Gohari, "A technique for deriving one-shot achievability results in network information theory," 2013, http://arxiv.org/abs/1303.0696.

[5] E. C. Song, P. Cuff, and H. V. Poor, "The likelihood encoder for lossy compression," http://arxiv.org/abs/1408.4522.

[6] P. Hausladen and W. K. Wootters, "A 'pretty good' measurement for distinguishing quantum states," *J. Mod. Optics*, vol. 41, no. 12, pp. 2385–2390, 1994.

[7] A. Holevo, "Reliability function of general classical-quantum channel," *IEEE Trans. Inf. Theory*, vol. 46, no. 6, pp. 2256–2261, Sep. 2000.

[8] V. Strassen, "Asymptotische Abschätzungen in Shannon's Informationstheorie," in *Trans. 3rd Prague Conf. on Inf. Theory*, 1962, pp. 689–723, [English Translation: http://www.math.wustl.edu/~luthy/strassen.pdf].

[9] 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.

[10] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge University Press, 2011.

[11] I. Csiszár and P. Narayan, "Channel capacity for a given decoding metric," *IEEE Trans. Inf. Theory*, vol. 45, no. 1, pp. 35–43, Jan. 1995.

[12] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai, "On information rates for mismatched decoders," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1953–1967, Nov. 1994.

[13] A. Ganti, A. Lapidoth, and E. Telatar, "Mismatched decoding revisited: General alphabets, channels with memory, and the wide-band limit," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2315–2328, Nov. 2000.

[14] J. Scarlett, "Reliable communication under mismatched decoding," Ph.D. dissertation, University of Cambridge, 2014, [Online: http://itc.upf.edu/biblio/1061].

[15] R. Gallager, "Fixed composition arguments and lower bounds to error probability," http://web.mit.edu/gallager/www/notes/notes5.pdf.

[16] J. Scarlett, A. Martinez, and A. Guillén i Fàbregas, "Mismatched decoding: Error exponents, second-order rates and saddlepoint approximations," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2647–2666, May 2014.

[17] A. G. D'yachkov, "Bounds on the average error probability for a code ensemble with fixed composition," *Prob. Inf. Transm.*, vol. 16, no. 4, pp. 3–8, 1980.

[18] K. Fan, "Minimax theorems," *Proc. Nat. Acad. Sci.*, vol. 39, pp. 42–47, 1953.

[19] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[20] I. Csiszár and J. Körner, "Graph decomposition: A new key to coding theorems," *IEEE Trans. Inf. Theory*, vol. 27, no. 1, pp. 5–12, Jan. 1981.

[21] J. Hui, "Fundamental issues of multiple accessing," Ph.D. dissertation, MIT, 1983.

[22] 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.

[23] ——, "Statistical physics and information theory," *Foundations and Trends in Comms. and Inf. Theory*, vol. 6, no. 1-2, pp. 1–212, 2009.

[24] 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.

[25] J. Scarlett, A. Martinez, and A. Guillén i Fàbregas, "Mismatched multi-letter successive decoding for the multiple-access channel," in *IEEE Int. Symp. Inf. Theory*, Honolulu, HI, 2014.

[26] N. Shulman, "Communication over an unknown channel via common broadcasting," Ph.D. dissertation, Tel Aviv University, 2003.