Mismatched Decoding Revisited: General Alphabets, Channels with Memory, and the Wide-Band Limit

Anand Ganti, Amos Lapidoth, *Member, IEEE*, and İ. Emre Telatar, *Member, IEEE*

***Abstract*—The mismatch capacity of a channel is the highest rate at which reliable communication is possible over the channel with a given (possibly suboptimal) decoding rule. This quantity has been studied extensively for single-letter decoding rules over discrete memoryless channels (DMCs). Here we extend the study to memoryless channels with general alphabets and to channels with memory with possibly non-single-letter decoding rules. We also study the wide-band limit, and, in particular, the mismatch capacity per unit cost, and the achievable rates on an additive-noise spread-spectrum system with single-letter decoding and binary signaling.**

***Index Terms*—Capacity per unit cost, channels with memory, general alphabets, mismatched decoding, nearest neighbor decoding, spread spectrum.**

I. Introduction

THIS paper deals with the rates at which reliable communication is possible over a given channel with a given—possibly suboptimal—decoding rule. This scenario arises naturally when, due to imprecise channel measurement, the receiver performs maximum-likelihood decoding with respect to the wrong channel law, or when the receiver is intentionally designed to perform a suboptimal decoding rule so as to simplify its implementation. This problem has been studied extensively, and we refer the reader to [1], [2] for relevant references.

In the problem's simplest form, the channel under consideration is a memoryless channel over finite input and output alphabets, and the decoding rule is a single-letter rule. Even for this simple case, the mismatch capacity, which is defined as the supremum of all achievable rates, is unknown. In fact, it has been demonstrated in [10] that a general solution to this problem would yield, as a special case, a solution to the long-standing problem of computing the zero-error capacity of a channel.

Other than the trivial bound that bounds the mismatch capacity by the matched capacity, to the best of our knowledge, no general upper bounds on the mismatch capacity were reported. See, however, [4] for binary input channels. Lower bounds on the mismatch capacity were derived using random coding arguments. Such arguments are based on the analysis of the probability of error of the mismatched decoder averaged over some ensemble of codebooks. For each block length one typically picks some distribution on the set of all codebooks of a given rate, and one then studies the highest rate for which the average probability of error—averaged over this ensemble—decays to zero as the block length tends to infinity. This rate is then achievable by Shannon's classical random coding argument, as there must by some family of codes in the ensemble for which the probability of error decays to zero.

Different choices of the code distribution lead to different bounds on the mismatch capacity. A distribution over the codebooks under which the codewords are independent and each codeword is chosen according to a product distribution leads to a bound that is referred to in [5] as the Generalized Mutual Information (GMI); see also [6]. A tighter lower bound to the mismatch capacity can be derived by considering code distributions under which the different codewords are still independent, but rather than drawing each codeword according to a product distribution, each codeword is chosen from a type class [7]–[9]. Further improvements can be made by choosing other distributions on the codewords [10] or by considering code distributions where different codewords are not drawn independently [11].

Although the GMI is the loosest of the above bounds, it has the benefit of being applicable to channels over nonfinite alphabets. Indeed, its derivation does not rely on the method of types [12] but rather on Gallager's bounds [13], thus making it applicable to channels over continuous alphabets as well. (See [14] for an alternative derivation of the GMI via information spectrum techniques.) On the other hand, the bound based on equi-type ensembles, while superior to the GMI, relies heavily on the method of types and is thus essentially limited to channels over finite alphabets. More critically, the method of types is of limited applicability to channels with memory, rendering the bound inapplicable to such channels. See, however, [2] for some extensions to memoryless channels of an exponential type and to some channels with memory.

In this paper, we extend the bound obtained by equi-type ensembles to memoryless channels with general alphabets and even to channels with memory. This is accomplished by using an alternative derivation that does not require the method of types. Using our bound we extend some of Verdú's [15] and Gallager's [16] results on the capacity per unit cost to the mismatched decoding scenario. Certain applications to spread-spectrum communication with unknown jamming statistics are also discussed.

Manuscript received July 18, 1999; revised May 23, 2000. The material in this paper was presented in part at the International Technion Communications Day in Honor of Professor Israel Bar-David, March 25, 1999, Technion, Haifa, Israel.

A. Ganti is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139-4307 USA.

A. Lapidoth was with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139-4307 USA. He is now with the Swiss Federal Institute of Technology, ETH-Zentrum, CH-8092 Zurich, Switzerland (e-mail: lapidoth@isi.ee.ethz.ch).

İ. E. Telatar was with Lucent Technologies, Murray Hill, NJ 07974 USA. He is now with the Swiss Federal Institute of Technology, EPFL-DSC-LTHI, CH-1015 Lausanne, Switzerland (e-mail: Emre.Telatar@epfl.ch).

Communicated by S. Shamai, Associate Editor for Shannon Theory.

Publisher Item Identifier S 0018-9448(00)09664-4.

It should be noted that the extension of mismatch results from finite alphabets to continuous channels cannot, in general, be accomplished using a limiting argument applied to ever finer channel quantizations. This approach, while applicable to optimal decoding scenarios [13], becomes quite tricky in the presence of decoding mismatch. Indeed, in the matched case it is clear that the optimal decoder for the general channel performs at least as well as a decoder that first quantizes the output and then performs optimal processing on the quantized samples. Under mismatched decoding, however, it is unclear how to relate the performance of the mismatched decoder on the original channel to its performance on the output-quantized channel.

The study of the various random-coding bounds to the mismatch capacity is sometimes of interest not only as a means of studying the mismatch capacity, but also in its own right. In some applications where the mismatch conditions are not taken into consideration in designing the codebook, some engineering insight into the performance of a "typical" codebook may be gained from the study of the average performance of a random codebook chosen from an appropriately defined ensemble. In such situations, the exact mismatch capacity may not give the right engineering intuition, because it is better suited for applications where the nature of the mismatch is taken into consideration in designing the optimal codebook.

The rest of this paper is organized as follows. In Section II, we formulate the mismatch problem for memoryless channels and describe some of the known results that are special to memoryless channels over finite alphabets. In Section III, we derive the lower bound for memoryless channels over infinite alphabets, and in Section IV, we extend these results to channels with memory. Section V studies the mismatch capacity per unit cost, and Section VI studies a spread-spectrum example [17]. We conclude the paper with a discussion of the various bounds, and with a discussion of some of the peculiarities of the mismatch capacity per unit cost.

II. The Mismatch Problem

Consider a memoryless channel of law $W(\cdot\,|\,x)$ over the general input and output alphabets $\mathcal{X}, \mathcal{Y}$. Such a channel is thus a mapping from the input alphabet $\mathcal{X}$ to probability measures on the output alphabet $\mathcal{Y}$. We shall assume throughout that both $\mathcal{X}$ and $\mathcal{Y}$ are Polish (i.e., complete separable metric) spaces endowed with the Borel σ-algebra. We endow $\mathcal{X} \times \mathcal{Y}$ with the product σ-algebra. Following [18], we shall assume throughout that the mapping $x \mapsto W(\cdot\,|\,x)$ from $\mathcal{X}$ to the set of probability measures on $\mathcal{Y}$ is Borel measurable, i.e., that for any Borel subset B of $\mathcal{Y}$ the mapping $x \mapsto W(B\,|\,x)$ is measurable.

Thus for any probability measure P_X on $\mathcal{X}$ we can define the probability measure $P_{X,Y} = P_X \circ W$ on $\mathcal{X} \times \mathcal{Y}$ by

$$P_{X,Y}(A \times B) = \int_A W(B\,|\,x)\, dP_X(x) \tag{1}$$

where A, B are Borel sets in $\mathcal{X}$ and $\mathcal{Y}$, respectively.

We similarly define the output distribution $P_Y = P_X W$ by

$$P_Y(B) = \int W(B\,|\,x)\, dP_X(x) \tag{2}$$

for any Borel subset $B \subset \mathcal{Y}$. Finally, the product law $P_X P_Y$ is defined by

$$P_X P_Y(A \times B) = P_X(A) P_Y(B). \tag{3}$$

We shall associate with every input symbol $x \in \mathcal{X}$ a nonnegative cost $g(x)$, where

$$g : \mathcal{X} \to [0, \infty) \tag{4}$$

is Borel measurable. We extend the domain of the definition of the cost function to n-tuples in an additive way so that

$$g(\boldsymbol{x}) = \frac{1}{n} \sum_{k=1}^{n} g(x_k), \quad \boldsymbol{x} \in \mathcal{X}^n.$$

A rate-R block length-n codebook $\mathcal{C}$ of cost Γ maps each message $m \in \mathcal{M}$ to some n-tuple

$$\boldsymbol{x}(m) = (x_1(m), \ldots, x_n(m)) \in \mathcal{X}^n$$

satisfying $g(\boldsymbol{x}(m)) \le \Gamma$. Here

$$\mathcal{M} = \{1, \ldots, \lfloor e^{nR} \rfloor\} \tag{5}$$

denotes the set of messages.

We now turn to the decoder. Let

$$d : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \tag{6}$$

be some measurable function to which we shall refer as the "decoding metric" even though it need not be a metric in the topological sense. Given a codebook $\mathcal{C}$ and a decoding metric $d(x, y)$, the decoder ϕ_d is defined as the mapping

$$\phi_d : \mathcal{Y}^n \to \mathcal{M} \cup \{0\} \tag{7}$$

that maps the received sequence $\boldsymbol{y}$ to $m \in \mathcal{M}$ if

$$\sum_{k=1}^{n} d(x_k(m), y_k) < \sum_{k=1}^{n} d(x_k(m'), y_k), \qquad m' \in \mathcal{M} \setminus \{m\} \tag{8}$$

and if no such $m \in \mathcal{M}$ exists (as can only be due to ties), we set $\phi_d(\boldsymbol{y}) = 0$.

If message $m \in \mathcal{M}$ is transmitted then we shall say that an error has occurred if $\phi_d(\boldsymbol{y}) \neq m$.

Definition 1: A rate R is achievable over the channel $W(\cdot\,|\,\cdot)$ with cost Γ and decoding rule ϕ_d if for every $\epsilon > 0$ and all sufficiently large n there exists a block length-n rate-R codebook of cost Γ that when decoded over the channel $W(\cdot\,|\,\cdot)$ using the decoder ϕ_d results in a maximal (over messages) probability of error smaller than ϵ.

The mismatch capacity is the supremum of achievable rates and is denoted[1] $C_M(\Gamma)$.

Setting

$$\Gamma_{\min} = \inf_{x \in \mathcal{X}} g(x) \tag{9}$$

we have the following lemma.

Lemma 1: For a memoryless channel $W(\cdot\,|\,x)$ with general alphabets, the function $C_M(\Gamma)$ is a nonnegative nondecreasing function of Γ. It is concave and continuous in the interval $(\Gamma_{\min}, +\infty)$.

[1] The mismatch capacity depends on the channel law, the decoding metric, and the cost Γ. The dependence on the former two quantities is not, however, made explicit in our notation.

Proof: The nonnegativity of $C_M(\Gamma)$ follows from its definition.

Consider a codebook of parameters $(n, \mathcal{M}, \Gamma, \epsilon)$, where n denotes the block length, Γ is the cost constraint, and ϵ is the maximal probability of error incurred over the channel $W(\cdot \,|\, x)$ using the mismatched decoder ϕ_d. Consider also a second codebook of parameters $(n', \mathcal{M}', \Gamma', \epsilon')$. From these two codebooks we can form the product codebook that consists of all possible way by which a codeword from the first codebook can be concatenated with a codeword from the second codebook. The product codebook is thus of block length $n + n'$, rate

$$\frac{\log|\mathcal{M}|}{n}\frac{n}{n+n'} + \frac{\log|\mathcal{M}'|}{n'}\frac{n'}{n+n'}$$

and has the cost parameter

$$\Gamma\frac{n}{n+n'} + \Gamma'\frac{n'}{n+n'}.$$

The mismatched decoder will err in decoding the product codebook only if either the first n symbols of the received sequence would cause it to err on the first codebook, or if the last n' received symbols would cause it to err on the second codebook. The union of events bound thus demonstrates that on the product code, the mismatched decoder errs with probability at most $\epsilon + \epsilon'$. This establishes the concavity of $C_M(\Gamma)$.

By [19, Theorem 10.1] it follows from the concavity of $C_M(\Gamma)$ that $C_M(\Gamma)$ is continuous for $\Gamma > \Gamma_{\min}$. □

For channels over finite alphabets and in the absence of cost constraints the following holds [7]–[9].

Theorem 1: For memoryless channels over finite alphabets and in the absence of cost constraints, the mismatch capacity can be lower-bounded by

$$\max_{P_X} I_{\rm LM}(P_X)$$

where the maximization is over all probability distributions P_X on $\mathcal{X}$, and

$$I_{\rm LM}(P_X) = \min_{\nu \in \mathcal{F}} D(\nu \,\|\, P_X P_Y) \tag{10}$$

where $D(\cdot \,\|\, \cdot)$ denotes the relative entropy functional [20] and the set $\mathcal{F}$ denotes the set of all probability mass functions ν on $\mathcal{X} \times \mathcal{Y}$ that satisfy

$$\sum_{y\in\mathcal{Y}} \nu(x,y) = P_X(x)$$
$$\sum_{x\in\mathcal{X}} \nu(x,y) = P_Y(y)$$

and

$$\sum_{(x,y)\in\mathcal{X}\times Y} \nu(x,y)d(x,y) \leq \sum_{(x,y)\in\mathcal{X}\times Y} P_{X,Y}(x,y)d(x,y).$$

Note that this lower bound to the mismatch capacity is, in general, not tight [10]. It is, however, tight if the input alphabet is binary, i.e., if $|\mathcal{X}| = 2$ [4].

Using Lagrange multipliers and duality theory one can give an alternative expression for $I_{\rm LM}(P_X)$ [2]

$$I_{\rm LM}(P_X) = \sup \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_X(x)W(y\,|\,x) \times \log \frac{e^{-sd(x,y)-a(x)}}{\sum_{x'\in\mathcal{X}} P_X(x')e^{-sd(x',y)-a(x')}} \tag{11}$$

where the supremum is over all $s \geq 0$ and over all functions $a : \mathcal{X} \to \mathbb{R}$. Here $\mathbb{R}$ denotes the set of real numbers.

We shall refer to (10) as the primal problem, and to (11) as the dual problem. The dual problem has two advantages. First, the dual problem need not be solved in order to obtain a lower bound on the mismatch capacity. Any choice of the parameter s and the function $a(\cdot)$ yields a lower bound to the mismatch capacity. This should be contrasted with the primal expression where an arbitrary feasible ν only gives an upper bound to $I_{\rm LM}$, i.e, an upper bound to a lower bound on the mismatch capacity.

The second advantage of the dual expression is that it generalizes more easily to general alphabets. Indeed, in this paper, rather than relying on the method of types to obtain the primal expression and then using duality theory to derive the dual expression, we shall derive the dual expression directly without using types.

Before doing so, we conclude this section with two alternative description of the GMI bound on the mismatch capacity. For any input distribution P_X the primal expression for the GMI is given by

$$I_{\rm GMI}(P_X) = \min_{\nu\in\mathcal{G}} D(\nu \,\|\, P_X P_Y)$$

where $\mathcal{G}$ denotes the set of all probability mass functions ν on $\mathcal{X} \times \mathcal{Y}$ that satisfy

$$\sum_{x\in\mathcal{X}} \nu(x,y) = P_Y(y)$$

and

$$\sum_{(x,y)\in\mathcal{X}\times Y} d(x,y)\nu(x,y) \leq \sum_{(x,y)\in\mathcal{X}\times Y} d(x,y)P_{X,Y}(x,y).$$

Since $\mathcal{F} \subseteq \mathcal{G}$ it is apparent that for any P_X

$$I_{\rm GMI}(P_X) \leq I_{\rm LM}(P_X).$$

The better known expression for the GMI is actually the dual expression and is given by

$$I_{\rm GMI}(P_X) = \sup_{s\geq 0} \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} P_X(x)W(y\,|\,x) \times \log \frac{e^{-sd(x,y)}}{\sum_{x'\in\mathcal{X}} P_X(x')e^{-sd(x',y)}}. \tag{12}$$

Notice that the dual expression to the GMI is obtained from the dual expression to $I_{\rm LM}$ simply by choosing $a(\cdot) \equiv 0$, thus demonstrating again that

$$I_{\rm GMI}(P_X) \leq I_{\rm LM}(P_X).$$

III. General Alphabets

In this section we extend (11) to memoryless channels with general alphabets. The general idea of the derivation is as follows. To derive the GMI bound (12) for general alphabets is fairly simple, because it is typically derived using Gallager's bounding technique, which does not rely on the method of types. If for any $s > 0$ and function $a(x)$ the decoding rule induced by the metric $d'(x,y) = sd(x,y) + a(x)$ were equivalent to the decoding rule induced by the metric $d(x,y)$, then (11) would follow from (12) simply by applying (12) to the decoding rule $d'(x,y)$ (with the parameter s in (12) set to one). The problem, however, is that the decoding rule induced by $d'(x,y) = sd(x,y) + a(x)$ is, in general, not equivalent to the one induced by $d(x,y)$, unless

$$n^{-1} \sum_{k=1}^{n} a(x_k(m))$$

does not depend on the message m. Imposing this condition on the ensemble, brings us back to notions of types and away from the independent and identically distributed (i.i.d.) codebooks that are so amenable to analysis. Instead, we impose a different condition on the codewords, namely, that

$$\left| n^{-1} \sum_{k=1}^{n} a(x_k(m)) - \mathbb{E}_{P_X}[a(X)] \right|$$

shall be very small. In this case, the decoding rule induced by $d(x,y)$ is not worse than the decoding rule induced by $d'(x,y)$ with a threshold decoder. Since the latter is simple to analyze and since the highest rate it can achieve approaches (11) as the threshold approaches zero, we can prove (11).

The specifics are described next. Fix some input distribution P_X satisfying

$$\mathbb{E}_{P_X}[g(X)] < \Gamma$$

where $\mathbb{E}$ denotes the expectation functional with its subscript denoting the law with respect to which the expectation is taken. Let $P_{X,Y}$, P_Y, and $P_X P_Y$ be defined as in (1), (2), and (3), respectively.

Fix an $s \geq 0$ and an $a(\cdot) \in L_1(P_X)$ for which

$$\log \int e^{-sd(x,y)-a(x)} \, dP_X(x) \in L_1(P_Y). \tag{13}$$

Here $L_1(\mu)$ denotes the set of all functions[2] that are integrable with respect to μ. For such $s \geq 0$ and $a(\cdot)$ let

$$\tilde{q}(x,y) = e^{-sd(x,y)-a(x)}, \qquad (x,y) \in \mathcal{X} \times \mathcal{Y}$$
$$b(y) = \log \int e^{-sd(x,y)-a(x)} \, dP_X(x) \tag{14}$$

and

$$q(x,y) = e^{-sd(x,y)-a(x)-b(y)}.$$

Thus $a(x) \in L_1(P_X)$, $b(y) \in L_1(P_Y)$, and for every bounded $h(y)$

$$\int h(y) q(x,y) \, dP_X P_Y(x,y) = \int h(y) \, dP_Y(y). \tag{15}$$

That is, the measure whose Radon–Nikodym derivative with respect to $P_X P_Y$ is $q(x,y)$ has y-marginal identical to P_Y. Consequently,

$$\mathbb{E}_{P_X}[q(X,Y)] = 1 \quad P_Y \text{ a.s.} \tag{16}$$

i.e.,

$$P_Y \left(\left\{ y : \int q(x,y) \, dP_X(x) \neq 1 \right\} \right) = 0.$$

Extend the definition of $\tilde{q}(x,y)$ and $q(x,y)$ to sequences by defining

$$\tilde{q}^{(n)}(\boldsymbol{x},\boldsymbol{y}) = \prod_{k=1}^{n} \tilde{q}(x_k, y_k), \qquad (\boldsymbol{x},\boldsymbol{y}) \in \mathcal{X}^n \times \mathcal{Y}^n$$
$$q^{(n)}(\boldsymbol{x},\boldsymbol{y}) = \prod_{k=1}^{n} q(x_k, y_k), \qquad (\boldsymbol{x},\boldsymbol{y}) \in \mathcal{X}^n \times \mathcal{Y}^n.$$

Consider a threshold decoder

$$\phi_{\text{Th}} : \mathcal{Y}^n \to \mathcal{M} \cup \{0\} \tag{17}$$

that for a given codebook and for some given $\tau > 0$ maps the received sequence $\boldsymbol{y} \in \mathcal{Y}^n$ to $m \in \mathcal{M}$ if

$$\tilde{q}^{(n)}(\boldsymbol{x}(m),\boldsymbol{y}) > e^{n\tau} \tilde{q}^{(n)}(\boldsymbol{x}(m'),\boldsymbol{y}), \qquad m' \in \mathcal{M} \setminus \{m\} \tag{18}$$

and if no such $m \in \mathcal{M}$ exists, maps $\boldsymbol{y}$ to 0.

By taking the logarithm of both sides of (18) one establishes that if

$$\left| \frac{1}{n} \sum_{k=1}^{n} a(x_k(m)) - \int a(x) \, dP_X(x) \right| < \frac{\tau}{2}, \qquad m \in \mathcal{M} \tag{19}$$

then the mismatched decoder ϕ_d (7), (8) errs only if the threshold decoder ϕ_{Th} (17), (18) errs. It is thus instructive to investigate the performance of the threshold decoder. To this end we prove the following lemma, which is analogous to [21, Lemma 6.9].

Lemma 2: Consider an ensemble of block length-n rate-R codebooks whose $\lfloor e^{nR} \rfloor$ codewords are drawn independently, each according to an n-fold product distribution P_{X^n} on $\mathcal{X}^n$ of marginal P_X.

Let $\bar{e}(\phi_{\text{Th}})$ denote the average (over messages and codebooks) probability of error incurred by the threshold decoder (17), (18) over the channel $W(\cdot \mid \cdot)$. Let $\epsilon > 0$ be fixed. Then

$$\bar{e}(\phi_{\text{Th}}) \leq P_{X^n,Y^n} \left(\left\{ (\boldsymbol{x},\boldsymbol{y}) \in \mathcal{X}^n \times \mathcal{Y}^n : q^{(n)}(\boldsymbol{x},\boldsymbol{y}) < \frac{e^{n(R+\tau)}}{\epsilon} \right\} \right) + \epsilon$$

where P_{X^n,Y^n} is the n-fold product distribution on $\mathcal{X}^n \times \mathcal{Y}^n$ of marginal $P_{X,Y}$.

[2] The domain of definition of these functions is determined by the argument μ. For example, $L_1(P_X)$ denotes the class of integrable functions from $\mathcal{X}$ to $\mathbb{R}$.

Proof: The proof is very similar to the proof of [21, Lemma 6.9]. We have

$$\bar{e}(\phi_{\mathrm{Th}}) = \int \Pr\left(\max_{m'\neq m} q^{(n)}(\boldsymbol{X}_{m'},\boldsymbol{y}) \geq e^{-n\tau} q^{(n)}(\boldsymbol{x},\boldsymbol{y})\right) \cdot dP_{X^n,Y^n}(\boldsymbol{x},\boldsymbol{y})$$

where $\boldsymbol{X}_m$ and $\boldsymbol{Y}$ are distributed on $\mathcal{X}^n \times \mathcal{Y}^n$ according to P_{X^n,Y^n} independently of $\{\boldsymbol{X}_{m'}\}_{m'\neq m}$ that are independently distributed on $\mathcal{X}^n$ according to P_{X^n}.

For pairs $(\boldsymbol{x},\boldsymbol{y})$ for which $q^{(n)}(\boldsymbol{x},\boldsymbol{y}) < e^{n(R+\tau)}/\epsilon$ we upper-bound the integrand by 1. For other pairs, i.e, pairs such that $q^{(n)}(\boldsymbol{x},\boldsymbol{y}) \geq e^{n(R+\tau)}/\epsilon$ we note that

$$\begin{aligned}\Pr&\left(\max_{m'\neq m} q^{(n)}(\boldsymbol{X}_{m'},\boldsymbol{y}) \geq e^{-n\tau} q^{(n)}(\boldsymbol{x},\boldsymbol{y})\right)\\ &\leq \Pr\left(\max_{m'\neq m} q^{(n)}(\boldsymbol{X}_{m'},\boldsymbol{y}) \geq \frac{e^{nR}}{\epsilon}\right)\\ &\leq (e^{nR}-1)\Pr\left(q^{(n)}(\boldsymbol{X}_{m'},\boldsymbol{y}) \geq \frac{e^{nR}}{\epsilon}\right), \qquad m'\neq m\\ &\leq \epsilon\end{aligned}$$

where the inequality before last follows from the union of events bound, and the last inequality follows from Markov's inequality and the fact that by (16)

$$P_{Y^n}\left\{\boldsymbol{y} : \int q^{(n)}(\boldsymbol{x},\boldsymbol{y})\, dP_{X^n}(\boldsymbol{x}) \neq 1\right\} = 0. \qquad \square$$

To simplify the analysis we now define a modified threshold decoder ϕ'_{Th} that given a codebook maps the received sequence $\boldsymbol{y}$ to 0 if the transmitted codeword $\boldsymbol{x}$ violates

$$\frac{1}{n}\sum_{k=1}^{n} g(x_k) < \Gamma \tag{20}$$

or if it violates

$$\left|\frac{1}{n}\sum_{k=1}^{n} a(x_k) - \int a(x)\, dP_X(x)\right| < \frac{\tau}{2} \tag{21}$$

and, otherwise, if both conditions are satisfied, maps $\boldsymbol{y}$ to $\phi_{\mathrm{Th}}(\boldsymbol{y})$. The modified decoder thus agrees with the threshold decoder if the transmitted codeword satisfies both (20) and (21), and declares an error otherwise.

Lemma 3: Consider an ensemble of blocklength-n rate-R codebooks whose codewords are drawn independently, each according to an n-fold product distribution P_{X^n} of marginal P_X. Let $\bar{e}(\phi'_{\mathrm{Th}})$ denote the average (over messages and codebooks) probability of error incurred by the modified threshold decoder ϕ'_{Th} over the channel $W(\cdot\,|\,\cdot)$. Then

$$\begin{aligned}\bar{e}(\phi'_{\mathrm{Th}}) \leq \bar{e}(\phi_{\mathrm{Th}}) &+ P_{X^n}\left(\frac{1}{n}\sum_{k=1}^{n} g(X_k) > \Gamma\right)\\ &+ P_{X^n}\left(\left|\frac{1}{n}\sum_{k=1}^{n} a(X_k) - \int a(x)\, dP_X(x)\right| \geq \frac{\tau}{2}\right).\end{aligned} \tag{22}$$

Proof: Follows directly from the union of events bound. □

We can now state the main result of this section regarding the mismatch capacity of a memoryless channel over general alphabets.

Theorem 2: The mismatch capacity $C_M(\Gamma)$, $\Gamma > \Gamma_{\min}$ of a channel $W(\cdot\,|\,\cdot)$ with cost function $g(x)$ and decoding rule $d(x,y)$ can be bounded by

$$C_M(\Gamma) \geq C_{\mathrm{LM}}(\Gamma)$$

where

$$C_{\mathrm{LM}}(\Gamma) = \sup I_{\mathrm{LM}}(P_X) \tag{23}$$

where the supremum is over all input distributions P_X satisfying

$$\begin{aligned}\mathbb{E}_{P_X}[g(X)] &\leq \Gamma\\ \mathbb{E}_{P_{X,Y}}[d(X,Y)] &< \infty.\end{aligned} \tag{24}$$

Here $P_{X,Y}$ is the joint distribution defined by (1), and

$$I_{\mathrm{LM}}(P_X) = \sup \int \log \frac{e^{-sd(x,y)-a(x)}}{\int e^{-sd(x',y)-a(x')}\, dP_X(x')}\, dP_{X,Y}(x,y) \tag{25}$$

where the supremum is over $s \geq 0$, and $a(x) \in L_1(P_X)$ satisfying (13).

Proof: We first claim that it suffices to prove that $C_M(\Gamma)$ is no smaller than $I_{\mathrm{LM}}(P_X)$ for distributions P_X for which (24) holds with strict inequality. To see this, consider the bounds on $C_M(\cdot)$ derived from $I_{\mathrm{LM}}(P_X)$ applied to distributions P_X that satisfy (24) with strict inequality. Since the mismatched capacity $C_M(\Gamma)$ is concave in the cost Γ for all $\Gamma > \Gamma_{\min}$ (see Lemma 1), the concave envelope of these bounds is also a lower bound to $C_M(\cdot)$. Being concave in Γ, this envelope is continuous in Γ for $\Gamma > \Gamma_{\min}$, and the claim follows.

Fix then some distribution P_X satisfying the strict inequality

$$\mathbb{E}_{P_X}[g(X)] < \Gamma. \tag{26}$$

Consider a block length-n rate-R codebook whose codewords are chosen independently according to the n-fold product distribution of marginal P_X. Fix some $\tau > 0$. It follows from Lemma 2 and the law of large numbers that as long as

$$R + \tau < \mathbb{E}_{P_{X,Y}}[\log q(X,Y)]$$

the ensemble averaged probability of error of the threshold decoder will decrease to zero as the block length n tends to infinity. By Lemma 3 and (26) the same is also true for the ensemble averaged probability of error for the modified threshold decoder ϕ'_{Th}. Given any $\epsilon > 0$ we can use the random coding argument to find, for all sufficiently large block length n, a codebook $\mathcal{C}$ of rate R for which the average probability of error incurred by the decoder ϕ'_{Th} is smaller than ϵ. By throwing away half its codewords, we can find a code $\mathcal{C}'$ of rate $R - n^{-1}\log 2$ for which the maximal probability of error with the decoder ϕ'_{Th} is smaller than 2ϵ.

Since any codeword that violates the cost constraint is incorrectly decoded by ϕ'_{Th}, it follows that all the codewords in $\mathcal{C}'$ satisfy the cost constraint (20), as well as the constraint (21).

Since the codewords in $\mathcal{C}'$ satisfy (21), it follows that a received sequence $\boldsymbol{y}$ will cause the mismatched decoder ϕ_d to err, only if it causes the threshold decoder to err also. Thus on the code $\mathcal{C}'$ the probability of error of the mismatched decoder cannot exceed the probability of error of the threshold decoder, i.e., 2ϵ. The result now follows by letting τ tend to zero. □

Remark 1: The lower bound $I_{\mathrm{LM}}(P_X)$ to the mismatch capacity is unchanged when the decoding metric $d(x,y)$ is replaced with the decoding metric

$$d'(x,y) = d(x,y) + f(y) + h(x) \tag{27}$$

for arbitrary $f(y) \in L_1(P_Y)$ and $h(x) \in L_1(P_X)$.

With regard to the above remark one should note that for discrete memoryless channels (DMCs), replacing $d(x,y)$ with $d'(x,y)$ as in (27) not only does not change the value of $I_{\mathrm{LM}}(P_X)$, but it also does not change the value of the mismatch capacity [9], [10]. This is because if $\mathcal{X}$ is finite, then the mismatch capacity can be achieved with constant composition codes, and for such codes $d(x,y)$ and $d'(x,y)$ yield identical decoding rules. It is not clear whether this also holds for general input alphabets.

We conclude this section with a condition under which $I_{\mathrm{LM}}(P_X)$ is strictly positive. It is clear from the primal expression (10) that for DMCs over finite alphabets, $I_{\mathrm{LM}}(P_X)$ is zero if, and only if

$$\mathbb{E}_{P_X P_Y}[d(X,Y)] \le \mathbb{E}_{P_{X,Y}}[d(X,Y)].$$

This is also true for memoryless channels over general alphabets.

Proposition 1: Let P_X be some input distribution to a memoryless channel over the general alphabets $\mathcal{X}$, $\mathcal{Y}$ with a nonnegative[3] decoding metric $d(x,y)$. Let $P_{X,Y}$ and $P_X P_Y$ be defined as in (1) and (3). Let $I_{\mathrm{LM}}(P_X)$ be, as in (25), the random coding lower bound to the mismatch capacity corresponding to the input distribution P_X. Then

$$I_{\mathrm{LM}}(P_X) = 0, \quad \text{iff} \quad \mathbb{E}_{P_X P_Y}[d(X,Y)] \le \mathbb{E}_{P_{X,Y}}[d(X,Y)].$$

Proof: The choice of $s = 0$ and $a(x) \equiv 0$ demonstrates that $I_{\mathrm{LM}}(P_X) \ge 0$. Next, by Jensen's inequality [22, Proposition 2.12]

$$\log \int e^{-sd(x',y) - a(x')}\, dP_X(x') \ge \int \left(-sd(x',y) - a(x')\right) dP_X(x')$$

and thus

$$\int \log \frac{e^{-sd(x,y)-a(x)}}{\int e^{-sd(x',y)-a(x')}\, dP_X(x')}\, dP_{X,Y}(x,y) \le s\left(\mathbb{E}_{P_X P_Y}[d(X,Y)] - \mathbb{E}_{P_{X,Y}}[d(X,Y)]\right). \tag{28}$$

Consequently, if

$$\mathbb{E}_{P_X P_Y}[d(X,Y)] \le \mathbb{E}_{P_{X,Y}}[d(X,Y)]$$

then $I_{\mathrm{LM}}(P_X) = 0$.

We now prove the reverse implication. Choose $a(x) \equiv 0$. Let

$$\begin{aligned} I(s) &= \int \log \frac{e^{-sd(x,y)}}{\int e^{-sd(x',y)}\, dP_X(x')}\, dP_{X,Y}(x,y) \\ &= -s\mathbb{E}_{P_{X,Y}}[d(X,Y)] \\ &\quad - \int \log \int e^{-sd(x,y)}\, dP_X(x)\, dP_Y(y). \end{aligned}$$

Noting that $I(0) = 0$

$$\begin{aligned} I'(0) = \lim_{s\downarrow 0} I(s)/s &= -\mathbb{E}_{P_{X,Y}}[d(X,Y)] \\ &\quad - \lim_{s\downarrow 0} \frac{1}{s} \int \log \int e^{-sd(x,y)}\, dP_X(x)\, dP_Y(y). \end{aligned}$$

Observe that

$$-\frac{1}{s} \log \int e^{-sd(x,y)}\, dP_X(x)$$

is nonnegative. Using Fatou's lemma

$$\begin{aligned} &-\lim_{s\downarrow 0} \frac{1}{s} \int \log \int e^{-sd(x,y)}\, dP_X(x)\, dP_Y(y) \\ &\quad \ge -\int \lim_{s\downarrow 0} \frac{1}{s} \log \int e^{-sd(x,y)}\, dP_X(x)\, dP_Y(y) \\ &\quad \ge \int \lim_{s\downarrow 0} \int \frac{1 - e^{-sd(x,y)}}{s}\, dP_X(x)\, dP_Y(y). \end{aligned}$$

Using Fatou's lemma once more

$$\begin{aligned} &\int \lim_{s\downarrow 0} \int \frac{1 - e^{-sd(x,y)}}{s}\, dP_X(x)\, dP_Y(y) \\ &\qquad \ge \iint d(x,y)\, dP_X(x)\, dP_Y(y) \end{aligned}$$

and thus

$$I'(0) \ge \mathbb{E}_{P_X P_Y}[d(X,Y)] - \mathbb{E}_{P_{X,Y}}[d(X,Y)].$$

This shows that if

$$\mathbb{E}_{P_X P_Y}[d(X,Y)] > \mathbb{E}_{P_{X,Y}}[d(X,Y)]$$

then, $I(s)$ is positive for sufficiently small s. □

It should be noted that for DMCs the mismatch capacity is positive only if $I_{\mathrm{LM}}(P_X) > 0$ for some input distribution P_X such that $P_X(x) = P_X(x') = 1/2$ for some $x, x' \in \mathcal{X}$, see [10]. It is unclear whether a similar statement can be made for memoryless channels over general alphabets.

For nondeterministic input distributions P_X that are concentrated at two points, the condition for the positivity of $I_{\mathrm{LM}}(P_X)$ takes on a particularly simple form.

Corollary 1: If P_X is nondeterministic and concentrated on $\{x_0, x_1\} \subseteq \mathcal{X}$ so that

$$P_X(x_0) + P_X(x_1) = 1, \qquad 0 < P_X(x_0) < 1,$$

then $I_{\mathrm{LM}}(P_X) > 0$ if, and only if,

$$\int \Delta(y)\, dP_{Y\,|\,X=x_0}(y) > \int \Delta(y)\, dP_{Y\,|\,X=x_1}(y) \tag{29}$$

where

$$\Delta(y) = d(x_1,y) - d(x_0,y).$$

IV. More General Channels

In this section, we study the mismatch capacity for channels with memory and non-single-letter decoders. Our results can be viewed as the mismatched decoding counterparts of the results of Verdú and Han [23] on channels with memory with optimal decoding.

As before, we denote the channel input and output alphabets by $\mathcal{X}$ and $\mathcal{Y}$. We assume that for any block length n the product sets $\mathcal{X}^n$ and $\mathcal{Y}^n$ are complete separable metric spaces endowed

[3] By Remark 1, if a decoding metric d is such that there exists integrable functions $f: \mathcal{X} \to \mathbb{R}$ and $g: \mathcal{Y} \to \mathbb{R}$ such that $d(x,y) + f(x) + g(y)$ is nonnegative, the proposition will hold for this d as well.

with the Borel σ-algebras, and that $\mathcal{X}^n \times \mathcal{Y}^n$ is endowed with the product σ-algebra. We further assume that for each block length n there corresponds a Borel measurable channel mapping $W^{(n)}(\cdot \,|\, \boldsymbol{x})$ that maps n-length input sequences to probability distributions on $\mathcal{Y}^n$. For example, for a DMC

$$W^{(n)}(\boldsymbol{y} \,|\, \boldsymbol{x}) = \prod_{k=1}^{n} W(y_k \,|\, x_k).$$

We let $\{P_{X^n}\}_{n=1}^{\infty}$ be a sequence of probability measures, where P_{X^n} is a probability measure on $\mathcal{X}^n$. For example, for a DMC we might consider

$$P_{X^n}(\boldsymbol{x}) = \prod_{k=1}^{n} P_X(x_k).$$

Note, however, that even for a memoryless channel, an i.i.d. input distribution may not be optimal; see [10] for the improved bounds on the mismatch capacity of DMC obtained by considering product spaces. As in (1), we denote the joint law on $\mathcal{X}^n \times \mathcal{Y}^n$ induced by the input distribution P_{X^n} and the channel $W^{(n)}(\cdot \,|\, \cdot)$ by P_{X^n,Y^n}. Similarly, as in (2), we let P_{Y^n} denote the law induced on $\mathcal{Y}^n$.

We assume as given a sequence of decoding metrics $\{d^{(n)}(\boldsymbol{x}, \boldsymbol{y})\}$ where

$$d^{(n)} : \mathcal{X}^n \times \mathcal{Y}^n \to \mathbb{R}.$$

For example, for single-letter decoding

$$d^{(n)}(\boldsymbol{x}, \boldsymbol{y}) = n^{-1} \sum_{k=1}^{n} d(x_k, y_k)$$

for some $d : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$. Similarly, we assume a sequence of cost functions $g^{(n)}$ where

$$g^{(n)} : \mathcal{X}^n \mapsto [0, \infty).$$

For example, in the DMC case we might have

$$g^{(n)}(\boldsymbol{x}) = n^{-1} \sum_{k=1}^{n} g(x_k).$$

We let $\{s^{(n)}\}$ denote a sequence of nonnegative real numbers. For the DMC we would typically set $s^{(n)} = s$, i.e., a constant sequence. Finally, we consider a sequence of functions

$$a^{(n)} : \mathcal{X}^n \to \mathbb{R}$$

which for the DMC case could be given by

$$a^{(n)}(\boldsymbol{x}) = n^{-1} \sum_{k=1}^{n} a(x_k)$$

for some single-letter function $a : \mathcal{X} \to \mathbb{R}$.

Theorem 3: Let the sequences $\{a^{(n)}(\cdot)\}$ and $\{s^{(n)}\}$ be such that:

- $a^{(n)}(\cdot) \in L_1(P_{X^n})$;
- $\lim_{n\to\infty} P_{X^n}(|a^{(n)}(\boldsymbol{X}) - \mathbb{E}_{P_{X^n}}[a^{(n)}(\boldsymbol{X})]| > \delta) = 0, \delta > 0$;
- the function $b^{(n)} : \mathcal{Y}^n \to \mathbb{R}$ formally defined by $$b^{(n)}(\boldsymbol{y}) = \frac{1}{n} \log \int e^{n\left[-s^{(n)} d^{(n)}(\boldsymbol{x}, \boldsymbol{y}) - a^{(n)}(\boldsymbol{x})\right]} \, dP_{X^n}$$ is defined and is in $L_1(P_{Y^n})$.

Let the sequence of input distributions $\{P_{X^n}\}$ satisfy the cost constraints with strict inequality

$$\mathbb{E}_{P_{X^n}}\left[g^{(n)}(\boldsymbol{X})\right] < \Gamma, \qquad n \geq 1.$$

Then, the mismatch capacity $C_M(\Gamma)$ with cost Γ is lower-bounded by

$$C_M(\Gamma) \geq \liminf \text{ in } P_{X^n,Y^n} \text{ of}$$
$$-s^{(n)} d^{(n)}(\boldsymbol{X}, \boldsymbol{Y}) - a^{(n)}(\boldsymbol{X}) - b^{(n)}(\boldsymbol{Y}).$$

Note: The liminf in probability should be interpreted as the supremum of real numbers α that satisfy

$$\lim_{n\to\infty} P_{X^n,Y^n}\left(-s^{(n)} d^{(n)}(\boldsymbol{X}, \boldsymbol{Y}) - a^{(n)}(\boldsymbol{X}) - b^{(n)}(\boldsymbol{Y}) < \alpha\right) = 0.$$

Proof: The proof is almost identical to the proof of Theorem 2. We define

$$q^{(n)}(\boldsymbol{x}, \boldsymbol{y}) = e^{n\left[-s^{(n)} d^{(n)}(\boldsymbol{x}, \boldsymbol{y}) - a^{(n)}(\boldsymbol{x}) - b^{(n)}(\boldsymbol{y})\right]}$$

and note that the measure whose Radon–Nikodym derivative with respect to $P_{X^n} P_{Y^n}$ is given by $q^{(n)}(\boldsymbol{x}, \boldsymbol{y})$ has marginal P_{Y^n} and, consequently,

$$P_{Y^n}\left(\left\{\boldsymbol{y} : \int q^{(n)}(\boldsymbol{x}, \boldsymbol{y}) \, dP_{X^n}(\boldsymbol{x}) \neq 1\right\}\right) = 0.$$

The theorem now follows from Lemma 2 in much the same way that Theorem 2 follows from that lemma. □

V. The Wide-Band Limit

Consider a memoryless channel $W(\cdot \,|\, x)$ on the input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$. Let $g(x)$ be a cost function on $\mathcal{X}$ and let $d(x, y)$ be some fixed decoding metric. As before, we denote by $C_M(\Gamma)$ the mismatch capacity with cost Γ. We define the mismatch capacity per unit cost as

$$C_M = \sup_{\Gamma > 0} \frac{C_M(\Gamma)}{\Gamma}. \tag{30}$$

In this section, we study C_M in an attempt to extend some of the results of Verdú [15] on the matched capacity per unit cost. Note, however, that the definition of the capacity per unit cost in [15] is somewhat different from the definition we adopt. Verdú's definition allows for the number of codewords to grow subexponentially in the block length. Nevertheless, he shows that in the matched case, the two definitions yield identical capacities.

In general, very little can be said about the supremum in (30). However, in the case where there exists an input symbol of zero cost, one can show that the supremum is achieved in the limit as $\Gamma \downarrow 0$. Before we can state and prove this result, we need the following lemma.

Lemma 4: Let the nonnegative function $f : (0, \infty) \to [0, +\infty]$ be monotonically nondecreasing and concave in the interval $[0, \infty)$. Then

$$\sup_{\xi > 0} \frac{f(\xi)}{\xi} = \lim_{\xi \downarrow 0} \frac{f(\xi)}{\xi} \tag{31}$$

where $\alpha/0 = +\infty$ for $\alpha > 0$.

Proof: If the limit of $f(\xi)$ as $\xi \downarrow 0$ is positive, then both sides of (31) are infinite, and equality thus holds. Otherwise, if this limit is zero, then $f(\xi)$ is concave and continuous in $[0, +\infty)$ with $f(0) = 0$. Consequently,

$$\begin{aligned} f(\alpha\xi) &\geq \alpha f(\xi) + (1 - \alpha) f(0), \qquad 0 \leq \alpha \leq 1 \\ &= \alpha f(\xi) \end{aligned}$$

so that the function $f(\xi)/\xi$ is monotonically nonincreasing in $(0, +\infty)$. □

Proposition 2: If there exists some input symbol x_0 of zero cost, i.e, $g(x_0) = 0$, then

$$\sup_{\Gamma>0} \frac{C_M(\Gamma)}{\Gamma} = \lim_{\Gamma\downarrow 0} \frac{C_M(\Gamma)}{\Gamma} \tag{32}$$

where $\alpha/0 = +\infty$, $\alpha > 0$.

Proof: In the presence of a zero-cost symbol $\Gamma_{\min} = 0$, so that by Lemma 1 $C_M(\Gamma)$ is monotonically nondecreasing and concave for $\Gamma > 0$. The result thus follows from Lemma 4. □

We focus now on two lower bounds to C_M

$$C_{\text{LM}} = \sup_{\Gamma>0} \frac{C_{\text{LM}}(\Gamma)}{\Gamma} \tag{33}$$

and

$$C^0_{\text{LM}} = \lim\sup_{\Gamma\downarrow 0} \frac{C_{\text{LM}}(\Gamma)}{\Gamma}. \tag{34}$$

By Theorem 2, it follows that

$$C^0_{\text{LM}} \le C_{\text{LM}} \le C_M.$$

It is not surprising that the inequality $C_{\text{LM}} \le C_M$ can be strict. Indeed, if all input symbols are of unit cost, then an example is provided by [10, Example 4].

Perhaps more surprising is the fact that even in the presence of a zero-cost symbol, C^0_{LM} can be strictly smaller than C_{LM}. (An example demonstrating this phenomenon is presented later in the section.) Thus while in the presence of a zero-cost symbol the mismatch capacity per unit cost C_M is always achieved in the limit as the cost goes to zero, this is not the case for its random coding lower bound C_{LM}.

In the presence of a zero-cost input symbol x_0 one can further lower-bound C_M by limiting oneself to binary-input distributions concentrated on x_0 and on some other arbitrary input symbol. This approach leads to the following bound

Lemma 5: For a memoryless channel with general alphabets and in the presence of a zero-cost symbol $x_0 \in \mathcal{X}$

$$C^0_{\text{LM}} \ge \sup_{x_1\ne x_0} \sup_{s\ge 0} \frac{1}{g(x_1)} \left[-s \int \Delta(y)\, dW(y\,|\,x_1) - \log \int e^{-s\Delta(y)}\, dW(y\,|\,x_0) \right] \tag{35}$$

where

$$\Delta(y) = d(x_1, y) - d(x_0, y), \tag{36}$$

and the maximization is over all symbols $x_1 \in \mathcal{X} \setminus \{x_0\}$ such that

$$\min\{\Delta(y), 0\} \in L_1(W(\cdot\,|\,x_0)) \cap L_1(W(\cdot\,|\,x_1)). \tag{37}$$

If two symbols x_0 and x_1 have zero cost then the capacity per unit cost is infinite if (29) holds.

Proof: Let $x_1 \in \mathcal{X}$ be any symbol of positive cost and satisfying (37). Consider the input distribution

$$P_{X,\Gamma}(X = x_1) = 1 - P_{X,\Gamma}(X = x_0) = \beta$$

where

$$\beta = \frac{\Gamma}{g(x_1)}.$$

In studying $I_{\text{LM}}(P_{X,\Gamma})$ there is no loss of generality in assuming that the decoding metric is given by

$$d(x, y) = \begin{cases} 0, & \text{if } x = x_0 \\ \Delta(y), & \text{if } x = x_1 \end{cases}$$

see Remark 1. Assume that Γ is sufficiently small so that $0 < \beta < 1/2$. Fix some $s \ge 0$ and $\alpha > 0$. Let

$$a(x) = \begin{cases} 0, & \text{if } x = x_0 \\ -\log\alpha, & \text{if } x = x_1. \end{cases}$$

With this choice of $a(x)$ we have

$$e^{-sd(x,y)-a(x)} = \begin{cases} 1, & \text{if } x = x_0 \\ \alpha e^{-s\Delta(y)}, & \text{if } x = x_1. \end{cases}$$

With these definitions we have that if

$$b(y) = \log \int e^{-sd(x,y)-a(x)}\, dP_{X,\Gamma}$$

then

$$b(y) = \log\Big(1 + e^{-s\Delta(y)+\log\alpha+\log\beta} - \beta\Big)$$

and

$$P_Y(\cdot) = \beta W(\cdot\,|\,x_1) + (1-\beta) W(\cdot\,|\,x_0).$$

It can be readily verified that (37) guarantees that $b(y) \in L_1(P_Y)$. Indeed, one can write

$$\int_{y\in\mathcal{Y}} |b(y)|\, dP_Y(y) = \int_{y:-s\Delta(y)+\log\alpha+\log\beta<0} |b(y)|\, dP_Y(y) + \int_{y:-s\Delta(y)+\log\alpha+\log\beta\ge 0} |b(y)|\, dP_Y(y) \tag{38}$$

and note that the integrand in the first integral is bounded, and use the bound

$$\log(1 + e^x - \beta) \le x + \log 2, \qquad \text{for } x \ge 0 \text{ and } \beta \ge 0$$

in the other.

For any $s \ge 0$ and α, by (25)

$$\frac{1}{\Gamma} I_{\text{LM}}(P_{X,\Gamma}) \ge \frac{1}{g(x_1)} \left\{ \int \log \frac{\alpha e^{-s\Delta(y)}}{1-\beta+\beta\alpha e^{-s\Delta(y)}}\, dW(y\,|\,x_1) + \frac{1}{\beta}(1-\beta) \int \log \frac{1}{1-\beta+\beta\alpha e^{-s\Delta(y)}}\, dW(y\,|\,x_0) \right\}. \tag{39}$$

Using the inequality $\log(1+x) \le x$ we obtain

$$\frac{1}{\Gamma} I_{\text{LM}}(P_{X,\Gamma}) \ge \frac{1}{g(x_1)} \left\{ \log\alpha - s\int \Delta(y)\, dW(y\,|\,x_1) - \beta \int \left(\alpha e^{-s\Delta(y)} - 1\right) dW(y\,|\,x_1) - (1-\beta) \int \left(\alpha e^{-s\Delta(y)} - 1\right) dW(y\,|\,x_0) \right\}. \tag{40}$$

By letting β tend to zero we obtain

$$\lim_{\beta\downarrow 0} \frac{1}{\Gamma} I_{\text{LM}}(P_{X,\Gamma}) \ge \frac{1}{g(x_1)} \left\{ \log\alpha - s\int \Delta(y)\, dW(y\,|\,x_1) - \int \left(\alpha e^{-s\Delta(y)} - 1\right) dW(y\,|\,x_0) \right\}. \tag{41}$$

The result now follows by choosing

$$\alpha = \frac{1}{\int e^{-s\Delta(y)}\, dW(y\,|\,x_0)}.$$

□

Remark 2: For a DMC, an alternative expression (primal) for the right-hand side of (35) is

$$\max_{x_1\neq x_0}\min_{f} D(f(\cdot)\,\|\,W(\cdot\,|\,x_0)) \tag{42}$$

where the minimization is over all probability mass functions (PMFs) $f(\cdot)$ on $\mathcal{Y}$ satisfying

$$\sum_{y\in\mathcal{Y}} f(y)\Delta(y)=\sum_{y\in\mathcal{Y}} W(y\,|\,x_1)\Delta(y).$$

The following theorem explores conditions under which the bound on C^0_{LM} provided by Lemma 5 is tight. See [15] for the analogous statement about the matched capacity per unit cost.

Theorem 4: Consider a DMC $W(y\,|\,x)$ over the finite input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$. Let $g(x)$ be a cost function on $\mathcal{X}$, and assume the existence of a unique input symbol x_0 of zero cost. Further assume that the matched capacity per unit cost

$$\max_{x\neq x_0}\{D(W(\cdot\,\|\,x)\,\|\,W(\cdot\,\|\,x_0))/g(x)\}$$

is finite. Then (35) holds with equality and, in particular, there exists some input symbol $x_1\in\mathcal{X}$ such that

$$C^0_{\mathrm{LM}}=\lim_{\Gamma\downarrow 0}\frac{I_{\mathrm{LM}}(P_{X,\Gamma})}{\Gamma}$$

where the input distribution $P_{X,\Gamma}$ satisfies $\mathbb{E}_{P_{X,\Gamma}}[g(X)]=\Gamma$ and $P_{X,\Gamma}(\{x_0,x_1\})=1$.

Proof: See the Appendix. □

The following example will demonstrate some of the differences between the behavior of the matched and mismatched capacities per unit cost.

Consider a noiseless channel $W(\cdot\,|\,\cdot)$ over the input alphabet $\mathcal{X}=\{0,1,2,3\}$ and the output alphabet $\mathcal{Y}=\{0,1,2,3\}$, with law

$$W(j\,|\,i)=I\{i=j\},\qquad 0\leq i,j\leq 3.$$

Here we use the notation

$$I\{\text{statement}\}=\begin{cases}1, & \text{if "statement" is true}\\ 0, & \text{if "statement" is false.}\end{cases}$$

Associate with every input symbol $x\in\mathcal{X}$ the cost $g(x)$ defined by

$$g(i)=I\{i\neq 0\},\qquad i=0,1,2,3.$$

Thus all symbols have unit cost except for the symbol 0, which has zero cost.

We now choose the decoding metric to discourage the use of the symbol 0 in spite of its zero cost. Since an input 0 results in the output 0, and since the mismatched decoder minimizes the accumulated metric, this is achieved by setting

$$\begin{aligned}&d(0,0)=3\\ &d(j,0)=0,\qquad j=1,2,3.\end{aligned}$$

Next, we guarantee that if a codebook does not contain the symbol 0, then our decoding rule will allow for error-free communication. We thus set

$$\begin{aligned}&d(i,i)=1, && i=1,2,3\\ &d(i,j)=2, && \text{whenever } 1\leq i,j\leq 3 \text{ and } i\neq j.\end{aligned}$$

By using a codebook containing all the 3^n distinct n-length sequences over $\{1,2,3\}$ we guarantee a rate of $\log 3$ bits per symbol. Thus

$$\sup_{\Gamma>0}\frac{C_M(\Gamma)}{\Gamma}\geq\left.\frac{C_M(\Gamma)}{\Gamma}\right|_{\Gamma=1}\geq\log 3. \tag{43}$$

In fact, it can be shown that $I_{\mathrm{LM}}(Q)=\log 3$ if Q is uniform over $\{1,2,3\}$. Thus

$$\sup_{\Gamma>0}\frac{C_{\mathrm{LM}}(\Gamma)}{\Gamma}\geq\left.\frac{C_{\mathrm{LM}}(\Gamma)}{\Gamma}\right|_{\Gamma=1}\geq\log 3. \tag{44}$$

We next show that the mismatch capacity per unit cost is not achievable using binary signaling. To this end, we first consider binary signaling where neither of the signals in use is the zero-cost symbol 0. In this case, the average cost of each codewords is 1, and since we are using only two symbols, the communication rate is no bigger than 1 bit/symbol. Thus we cannot achieve a rate-per-cost larger than $\log 2$, whereas the mismatch capacity per unit cost is at least $\log 3$; see (43).

The other form of binary signaling is when one of the symbols is the zero-cost symbol 0. Without loss of generality we shall assume that the other symbol is 1. We will show that with the above decoding metric any such binary code yields an average probability of error of one. Indeed, consider two codewords $\boldsymbol{x}=(x_1,\ldots,x_n)$ and $\boldsymbol{x}'=(x'_1,\ldots,x'_n)$ over $\{0,1\}^n$. Assume that $\boldsymbol{x}$ is transmitted. If $\boldsymbol{x}=\boldsymbol{x}'$ then both codewords accumulate the same metric, thus leading to an error. Assume now that $\boldsymbol{x}\neq\boldsymbol{x}'$. In computing the difference between the metric accumulated by the two codewords, we may ignore components k for which $x_k=x'_k$. Consider then some k for which $x_k\neq x'_k$. If $x_k=0$ then the corresponding received symbol is 0 and the metric added to the correct codeword is $d(0,0)=3$ while the metric added to the incorrect codeword is $d(1,0)=0$. In the other case, if $x_k=1$ and, consequently, $y_k=1$ then the metric added to the correct codeword is $d(1,1)=1$ while the metric added to the incorrect codeword is $d(0,1)=0$. In either case the metric accumulated by the incorrect codeword is lower than the metric accumulated by the correct codeword, and an error results.

The above argument also demonstrates that the average probability of error of the mismatched decoder over an ensemble of binary codes consisting of the symbols 0 and 1 is also one.

VI. A Spread-Spectrum Example

Consider an additive (not necessarily Gaussian) noise channel where the output Y_k at time k is given by

$$Y_k=x_k+Z_k \tag{45}$$

where x_k denotes the channel input at time k and Z_k is the corresponding noise sample. Note that we do not assume that the noise samples are of zero mean.

In [25] it was demonstrated that if the noise process $\{Z_k\}$ has an ergodic law (that does not depend on the input sequence) and if the decoder performs nearest neighbor decoding (i.e., the decoding rule that would have been optimal if the noise were i.i.d. zero-mean Gaussian) then for a Gaussian ensemble of power P

$$I_{\mathrm{GMI}}(P_G)=\frac{1}{2}\log\left(1+\frac{P}{N}\right) \tag{46}$$

where

$$N=\mathbb{E}\left[Z_k^2\right] \tag{47}$$

and where P_G denotes here a zero-mean, variance-P Gaussian distribution. In the wide-band limit we obtain

$$\lim_{P\downarrow 0}\frac{1}{P}I_{\rm GMI}(P_G)=\frac{1}{2N}\log e. \tag{48}$$

For wide-band systems, one rarely uses Gaussian codebooks, and a more practical approach is to use binary spread-spectrum signaling so that

$$x_k\in\{\delta,-\delta\}. \tag{49}$$

If one considers an ensemble of codebooks whose codewords are chosen i.i.d. according to a product Bernoulli$(1/2)$ distribution, one obtains a limit quite similar to (48), i.e.,

$$\lim_{\delta\downarrow 0}\frac{1}{\delta^2}I_{\rm GMI}(P_B)=\frac{1}{2N}\log e \tag{50}$$

where P_B denotes the Bernoulli distribution that takes on the values $\pm\delta$ equiprobably.

We shall, however, demonstrate that if in the ensemble the codewords are chosen uniformly over the Bernoulli$(1/2)$ type then

$$\lim_{\delta\downarrow 0}\frac{1}{\delta^2}I_{\rm LM}(P_B)=\frac{1}{2\tilde{N}}\log e \tag{51}$$

where

$$\tilde{N}=\mathbb{E}\left[Z_k^2\right]-(\mathbb{E}[Z_k])^2 \tag{52}$$

is the variance of the noise.

Rather than deriving this limit directly, we shall first derive a more general result applicable to general single-letter decoding rules, and only then specialize to the Euclidean distance decoding metric.

Recall that by (25)

$$\begin{aligned}&I_{\rm LM}(P_B)\\&\quad=\sup_{s\geq 0,a}\mathbb{E}\log\frac{e^{-sd(X,Y)-a(X)}}{\frac{1}{2}\left[e^{-sd(\delta,Y)-a(\delta)}+e^{-sd(-\delta,Y)-a(-\delta)}\right]}.\end{aligned}$$

We can, without loss of generality, choose $a(\delta)=-a(-\delta)=a$, and rewrite the above as

$$I_{\rm LM}(P_B)=\sup_{s\geq 0,a}\mathbb{E}\log\frac{e^{-sd(X,Y)}}{\frac{1}{2}\left[e^{-sd(\delta,Y)-a}+e^{-sd(-\delta,Y)+a}\right]}.$$

Let

$$\begin{aligned}A(y)&=\frac{1}{2}[d(\delta,y)+d(-\delta,y)]\\B(y)&=\frac{1}{2}[d(-\delta,y)-d(\delta,y)].\end{aligned}$$

Then, for $x\in\{\delta,-\delta\}$

$$\begin{aligned}d(\delta,y)&=A(y)-B(y)\\d(-\delta,y)&=A(y)+B(y)\\d(x,y)&=A(y)-\frac{x}{\delta}B(y).\end{aligned}$$

We thus see that

$$I_{\rm LM}(P_B) \tag{53}$$

$$=\sup_{s\geq 0,a}\mathbb{E}\log\frac{e^{s(X/\delta)B(Y)}}{\cosh(sB(Y)-a)} \tag{54}$$

$$=(\log e)\sup_{s\geq 0,a}\left[\frac{s}{\delta}\mathbb{E}[XB(Y)]-\mathbb{E}\ln\cosh(sB(Y)-a)\right] \tag{55}$$

$$\geq(\log e)\sup_{s\geq 0,a}\left[\frac{s}{\delta}\mathbb{E}[XB(Y)]-\frac{1}{2}\mathbb{E}[(sB(Y)-a)^2]\right] \tag{56}$$

$$=(\log e)\sup_{s\geq 0}\left[\frac{s}{\delta}\mathbb{E}[XB(Y)]-\frac{1}{2}s^2\,{\rm Var}[B(Y)]\right] \tag{57}$$

where the inequality follows from $\cosh x\leq e^{x^2/2}$. For the supremum over s we need to consider the following two possibilities.

1) $\mathbb{E}[XB(Y)]\leq 0$: The supremum over s is achieved at $s=0$ with value zero.

2) $\mathbb{E}[XB(Y)]>0$: This implies ${\rm Var}[B(Y)]>0$ for otherwise $B(Y)$ would equal a constant with probability 1 and $\mathbb{E}[XB(Y)]=0$. The supremum is achieved at

$$s=\mathbb{E}[(X/\delta)B(Y)]/{\rm Var}[B(Y)]$$

with value

$$(\log e)\frac{\mathbb{E}[(X/\delta)B(Y)]^2}{2\,{\rm Var}[B(Y)]}.$$

Let us now specialize to the Euclidean distance decoding metric, for which $d(x,y)=(x-y)^2$. Then $B(y)=2\delta y$, and thus $\mathbb{E}[(X/\delta)B(Y)]=2\delta^2$, ${\rm Var}[B(Y)]=4\delta^2(\delta^2+{\rm Var}[Z])$, and

$$I_{\rm LM}(P_B)\geq(\log e)\frac{1}{2}\delta^2\frac{1}{{\rm Var}[Z]+\delta^2}.$$

Furthermore, we claim that if $0<{\rm Var}[Z]<\infty$, the lower bound to $I_{\rm LM}(P_B)$ is tight as δ approaches zero, so that

$$\lim_{\delta\to 0}\frac{I_{\rm LM}(P_B)}{\delta^2}=(\log e)\frac{1}{2\,{\rm Var}[Z]}.$$

To prove the claim, we need to show that the inequality

$$E\ln\cosh(2s\delta Y-a))\leq(1/2)E(2s\delta Y-a)^2$$

is asymptotically tight as δ gets small. To that end, let $a(\delta)$ and $s(\delta)$ be the values of a and s which achieve the supremum in (55). We will first show that as δ approaches zero, $a(\delta)$ and $\delta s(\delta)$ both approach zero. By differentiating (55) with respect to a and s we obtain the following equations for optimal $a(\delta)$ and $s(\delta)$:

$$\mathbb{E}[\tanh(2s(\delta)\delta Y-a(\delta))]=0 \tag{58}$$

$$\mathbb{E}[2\delta Y\tanh(2s(\delta)\delta Y-a(\delta))]=\mathbb{E}[2XY]=2\delta^2. \tag{59}$$

Now, if we could prove that $\delta s(\delta)$ approaches zero as δ approaches zero, it would follow from the first equality that $a(\delta)$ approaches zero: $\delta s(\delta)Y$ would approach 0 in probability, and since $\tanh$ is continuous and bounded,

$$\tanh(2s(\delta)\delta Y-a(\delta))=0$$

implies that $\tanh(a(\delta))\to 0$, which yields that $a(\delta)\to 0$. It is also easy to see that $a(\delta)\approx 2s(\delta)\delta\mathbb{E}[Y]$ to the first order in δ. It is thus sufficient to prove that $\delta s(\delta)\to 0$. To that end, consider the second equation with $a(\delta)=2s(\delta)\delta\zeta(\delta)$. Using the inequality $\tanh(x)\geq x/(1+x)$ for $x\geq 0$ (which follows from $e^x\geq 1+x$), we obtain

$$2\delta s(\delta)E\left[\frac{(Y-\zeta(\delta))^2}{1+2\delta s(\delta)|Y-\zeta(\delta)|}\right]\leq\delta.$$

If $\delta s(\delta)$ is not approaching zero, we can extract a subsequence δ_n for which $\delta_n s(\delta_n)$ remains bounded away from zero. Since $\mathrm{Var}[Z] > 0$, this implies that the left-hand side remains bounded away from zero. However, the right-hand side approaches zero, leading to a contradiction. We have thus shown that $\delta s(\delta)$ approaches zero. Moreover, given that $\delta s(\delta) \to 0$, the expectation above remains bounded away from zero, which implies that $s(\delta)$ remains bounded.

Given that $\delta s(\delta) \to 0$ and $a(\delta) \to 0$, we now use the facts that i) $\ln\cosh(x) > (1-\eta(\epsilon))(1/2)x^2$ for $|x| < \epsilon, \eta(\epsilon) \to 0$ as $\epsilon \to 0$, and ii) since $\mathrm{Var}[Z] < \infty, \mathbb{E}[Y^2 \mathbf{1}_{|Y|<A}(Y)] \to \mathbb{E}[Y^2]$ as $A \to \infty$. These imply that for any $\eta > 0$, for sufficiently small δ

$$\mathbb{E}[\ln\cosh(2s(\delta)\delta Y - a(\delta)] \geq (1-\eta)\frac{1}{2}E(2s\delta Y - a(\delta))^2.$$

This inequality implies that our lower bound to $I_{\mathrm{LM}}(P_B)$ is tight.

VII. Discussion

The tightness of the random-coding lower bounds on the mismatch capacity depend on two factors: the distribution according to which the codebooks are drawn, and the inequalities that are used to upper-bound the respective ensemble-averaged probabilities of error.

The GMI bound (12) is based on a codebook distribution according to which codewords are chosen independently of each other, each according to an i.i.d. distribution. The analysis of the average probability of error is based on Gallager's bounding techniques.

On the other hand, the tighter bound (Theorem 1) is based on a different code distribution. Here the codewords are still chosen independently of each other, but each codeword is now drawn uniformly over a type class. The analysis of the average probability of error is performed using the method of types.

A natural question to ask is whether the GMI bound is inferior because of the code distribution (i.i.d. versus uniform over a type) or because of the performance analysis method (Gallager's bounds versus the method of type). It turns out that for DMCs the fault lies with the code distribution and not with the bounding technique: Gallager's bounding technique is tight for i.i.d. ensembles in the sense that for rates above $I_{\mathrm{GMI}}(P_X)$ the average probability of error for an i.i.d. ensemble tends to one, as the block length tends to infinity.[4]

Similarly, subject to some minor technical conditions, the method of types technique is tight for ensembles where the codewords are drawn uniformly over a type class. Thus for this code distribution and for all rates about $I_{\mathrm{LM}}(P_X)$, the ensemble-averaged probability of error tends to one, as the block length tends to infinity.[5] When C_{LM} is strictly smaller than the mismatch capacity it is not because the method of types is inadequate, but rather because the codebook distribution is inappropriate. The "average codebook" in this ensemble is simply not good enough to achieve the mismatch capacity.

We turn now to some remarks about the mismatch capacity per unit cost. In contrast to the behavior of the matched capacity [15] and to the behavior of the random coding lower bound to the mismatch capacity per unit cost (Theorem 4) we have the following.

Remark 3: Even in the presence of a zero-cost symbol, the mismatch capacity per unit cost C_M need not be achievable using codebooks over a subset $\mathcal{X}' \subset \mathcal{X}$ of cardinality two.

Proof: This is demonstrated by the example in Section V, where binary signaling cannot achieve a rate per cost greater than $\log 2$, whereas ternary signaling can achieve a rate per cost of $\log 3$. □

Since the input alphabet in the above example contains a zero-cost symbol, it follows from Theorem 4 that C^0_{LM} is achieved by binary signaling with one of the symbols being the zero-cost symbol. For the example at hand this implies that $C^0_{\mathrm{LM}} = 0$. On the other hand, it is demonstrated that C_{LM} is no smaller than $\log 3$, see (44). We thus conclude.

Remark 4: Even in the presence of a zero-cost symbol, the random coding lower bound to the mismatch capacity per unit cost need not be attained in the limit of zero cost. That is, $C^0_{\mathrm{LM}} \leq C_{\mathrm{LM}}$ can hold with a strict inequality.

If the function $C_{\mathrm{LM}}(\Gamma)$ were concave for $\Gamma \in (0,\infty)$ then by Lemma 4 (applied to function $C_{\mathrm{LM}}(\cdot)$) it would have followed that C^0_{LM} and C_{LM} are identical, in contradiction to Remark 4. We can therefore only conclude the following.

Remark 5: The random coding lower bound to the mismatch capacity $C_{\mathrm{LM}}(\Gamma)$ need not be a concave function of the cost Γ.

This has the following consequence, which was observed earlier in [24].

Remark 6: The random coding lower bound $I_{\mathrm{LM}}(P_X)$ defined in (25) need not be a concave function of the input distribution P_X.

Proof: To arrive at a contradiction to Remark 5 we shall assume that $I_{\mathrm{LM}}(P_X)$ is concave. Let $\Gamma_1, \Gamma_2 \geq \Gamma_{\min}$ be otherwise arbitrary. Let $\{P^{(n)}_{X,\Gamma_1}\}$ satisfy

$$\mathbb{E}_{P^{(n)}_{X,\Gamma_1}}[g(X)] \leq \Gamma_1$$

and

$$\sup_n I_{\mathrm{LM}}\left(P^{(n)}_{X,\Gamma_1}\right) = C_{\mathrm{LM}}(\Gamma_1).$$

Similarly, define a sequence $\{P^{(n)}_{X,\Gamma_2}\}$. Let $\lambda \in (0,1)$ be arbitrary, and let $\bar{\lambda} = 1 - \lambda$.

If $I_{\mathrm{LM}}(P_X)$ were concave then we would have

$$\begin{aligned} C_{\mathrm{LM}}(\lambda\Gamma_1 + \bar{\lambda}\Gamma_2) &\geq I_{\mathrm{LM}}\left(\lambda P^{(n)}_{X,\Gamma_1} + \bar{\lambda} P^{(n)}_{X,\Gamma_2}\right) \\ &\geq \lambda I_{\mathrm{LM}}\left(P^{(n)}_{X,\Gamma_1}\right) + \bar{\lambda} I_{\mathrm{LM}}\left(P^{(n)}_{X,\Gamma_2}\right) \end{aligned}$$

from which a contradiction to Remark 5 results upon letting n approach infinity. □

[4] This claim can be proved as in [25, Theorem 1] using the primal expression for the GMI, or using techniques similar to those used in [25, Appendix].

[5] See [11, Theorem 3] for the multiple-access channel version of this claim, or [2, Theorem 1] for a slightly weaker single-user version of this claim.

APPENDIX
PROOF OF THEOREM 4

Without loss of generality (see Remark 1) suppose that $d(x_0, y) = 0$ for all $y \in \mathcal{Y}$. Let $\Gamma'_{\min} = \inf_{x \neq x_0} g(x)$. Since $\mathcal{X}$ is finite and x_0 is the only symbol of zero cost, $\Gamma'_{\min} > 0$. Since

$$C^0_{\mathrm{LM}} = \limsup_{\Gamma \downarrow 0} \frac{C_{\mathrm{LM}}(\Gamma)}{\Gamma}$$

we can find sequences $\Gamma_n \downarrow 0$ and P_{X,Γ_n} such that

$$\sum_x P_{X,\Gamma_n}(x) g(x) = \Gamma_n \text{ and } \lim_{n\to\infty} \frac{I_{\mathrm{LM}}\left(P_{X,\Gamma_n}\right)}{\Gamma_n} = C^0_{\mathrm{LM}}. \tag{60}$$

Let P_{Y,Γ_n} denote the output distribution on Y that corresponds to the input distribution P_{X,Γ_n}.

For $x \neq x_0$ let $\Lambda_n(x) = P_{X,\Gamma_n}(x)/\Gamma_n$. From (60) it follows that $0 \le \Lambda_n(x) \le 1/\Gamma'_{\min}$. Note that $\Lambda(x) \ge 0$ for $x \neq x_0$ and $\sum_{x: x \neq x_0} g(x)\Lambda(x) = 1$.

Recall that I_{LM} is given by either the primal or the dual expressions

$$I_{\mathrm{LM}}\left(P_{X,\Gamma_n}\right) = \inf_{\nu} D\left(\nu \,\|\, P_{X,\Gamma_n} P_{Y,\Gamma_n}\right) \tag{61}$$

$$I_{\mathrm{LM}}\left(P_{X,\Gamma_n}\right) = \sup_{s \ge 0, a} \sum_{x,y} P_{X,\Gamma_n}(x) W(y \,|\, x) \tag{62}$$

$$\log \frac{e^{sd(x,y)-a(x)}}{\sum_{x'} e^{sd(x',y)-a(x')} P_{X,\Gamma_n}(x')} \tag{63}$$

where in (61), the infimum is over all ν's for which

$$\sum_x \nu(x,y) = P_{Y,\Gamma_n}(y) \qquad \sum_y \nu(x,y) = P_{X,\Gamma_n}(x)$$

and

$$\sum_{x,y} \nu(x,y) d(x,y) \le P_{X,\Gamma_n}(x) W(y \,|\, x) d(x,y).$$

Let ν_n be the distribution that achieves the infimum in the primal expression (61). By duality, ν_n is of the form

$$\nu_n(x,y) = e^{-s_n d(x,y) - a_n(x) - b_n(y)} P_{X,\Gamma_n}(x) P_{Y,\Gamma_n}(y)$$

with $s_n \ge 0$. Without loss of generality we can fix $a_n(x_0) = 0$. It follows from the condition on the Y marginal that

$$\nu_n(x,y) = \frac{e^{-s_n d(x,y) - a_n(x)} P_{X,\Gamma_n}(x) P_{Y,\Gamma_n}(y)}{\sum_{x'} e^{-s_n d(x',y) - a_n(x')} P_{X,\Gamma_n}(x')}$$

and thus

$$\nu_n(y \,|\, x) = \frac{e^{-s_n d(x,y) - a_n(x)} P_{Y,\Gamma_n}(y)}{\sum_{x'} e^{-s_n d(x',y) - a_n(x')} P_{X,\Gamma_n}(x')}. \tag{64}$$

Since $\nu_n(x,y)$ and $\nu_n(y \,|\, x)$ are in $[0,1]$ for all n, and since $\mathcal{X}$ and $\mathcal{Y}$ are finite, we can extract a subsequence $\{n_k\}$ for which the limits $\lim_{k\to\infty} \nu_{n_k}(x,y)$ and $\lim_{k\to\infty} \nu_{n_k}(y \,|\, x)$ exist for each $(x,y) \in \mathcal{X} \times \mathcal{Y}$. To simplify notation, we shall assume that the original sequences Γ_n and P_{X,Γ_n} were chosen such that this convergence already held. Let these limits above be $\nu(x,y)$ and $\nu(y \,|\, x)$, respectively. Since $P_{X,\Gamma_n}(x_0) \to 1$, the Y marginal condition implies

$$\nu(x_0, y) = \nu(y \,|\, x_0) = W(y \,|\, x_0).$$

From (64) we observe that

$$e^{-s_n d(x,y) - a_n(x)} = \frac{\nu_n(y \,|\, x)}{\nu_n(y \,|\, x_0)}.$$

Taking limits on both sides we see that

$$\lim_{n\to\infty} e^{-s_n d(x,y) - a_n(x)}$$

exists, and observing that $\nu(y|x) \le 1$ and $\nu(y|x_0) = W(y|x_0)$

$$\lim_{n\to\infty} e^{-s_n d(x,y) - a_n(x)} \le \frac{1}{W(y \,|\, x_0)}. \tag{65}$$

Now

$$\begin{aligned}
I_{\mathrm{LM}}&\left(P_{X,\Gamma_n}\right) \\
&= \sum_{x,y} P_{X,\Gamma_n}(x) W(y \,|\, x) \\
&\quad \times \log \frac{e^{-s_n d(x,y) - a_n(x)}}{\sum_{x'} e^{-s_n d(x',y) - a_n(x')} P_{X,\Gamma_n}(x')} \\
&= -\sum_{x: x \neq x_0} P_{X,\Gamma_n}(x) \sum_y W(y \,|\, x)[s_n d(x,y) + a_n(x)] \\
&\quad - \sum_y P_{Y,\Gamma_n}(y) \log \left[\sum_x e^{-s_n d(x,y) - a_n(x)} P_{X,\Gamma_n}(x) \right] \\
&= -\Gamma_n \sum_{x \neq x_0} \Lambda_n(x) \sum_y W(y \,|\, x)[s_n d(x,y) + a_n(x)] \\
&\quad - \sum_y P_{Y,\Gamma_n}(y) \log \\
&\quad \times \left(1 + \Gamma_n \left[\sum_{x \neq x_0} \left(e^{-s_n d(x,y) - a_n(x)} - 1 \right) \Lambda_n(x) \right] \right).
\end{aligned}$$

Since $\Lambda_n(x)$ is bounded, and $\Gamma_n \to 0$, for large enough n the argument of the logarithm above exceeds $1/2$. Since $\log(1+x) \ge x - x^2$ for $x > -1/2$

$$\begin{aligned}
I_{\mathrm{LM}}&\left(P_{X,\Gamma_n}\right) \\
&\le -\Gamma_n \sum_{x \neq x_0} \Lambda_n(x) \sum_y W(y \,|\, x)[s_n d(x,y) + a_n(x)] \\
&\quad - \Gamma_n \sum_y P_{Y,\Gamma_n}(y) \\
&\quad \times \left[\sum_{x \neq x_0} \left(e^{-s_n d(x,y) - a_n(x)} - 1 \right) \Lambda_n(x) \right] \\
&\quad + \Gamma_n^2 \sum_y P_{Y,\Gamma_n}(y) \\
&\quad \times \left[\sum_{x \neq x_0} \left(e^{-s_n d(x,y) - a_n(x)} - 1 \right) \Lambda_n(x) \right]^2.
\end{aligned} \tag{66}$$

The finiteness of the matched capacity implies that the Y-measures $W(y\,|\,x)$ are absolutely continuous with respect to $W(y\,|\,x_0)$, i.e.,

$$W(y\,|\,x) > 0 \text{ implies } W(y\,|\,x_0) > 0.$$

In particular, $P_{Y,\Gamma_n}(y) > 0$ implies $W(y\,|\,x_0) > 0$, and thus for those y and for large enough n

$$e^{-s_n d(x,y) - a_n(x)}$$

is bounded (by (65)). Hence the third term in (66) is $O(\Gamma_n^2)$. Thus

$$\begin{aligned}\lim_{n\to\infty} & \frac{I_{\rm LM}\left(P_{X,\Gamma_n}\right)}{\Gamma_n} \\ &\le \lim_{n\to\infty}\left[-\sum_{x\neq x_0}\Lambda_n(x)\sum_y W(y\,|\,x)[s_n d(x,y)+a_n(x)]\right. \\ &\qquad \left. -\sum_{x\neq x_0}\Lambda_n(x)\sum_y P_{Y,\Gamma_n}(y)\left(e^{-s_n d(x,y)-a_n(x)}-1\right)\right].\end{aligned} \tag{67}$$

The boundedness of $e^{-s_n d(x,y)-a_n(x)}$ and $\Lambda_n(x)$ allows us to replace $P_{Y,\Gamma_n}(y)$ above by its limiting value, $W(y\,|\,x_0)$, to obtain

$$\begin{aligned}C_{\rm LM}^0 &= \lim_{n\to\infty}\frac{I_{\rm LM}\left(P_{X,\Gamma_n}\right)}{\Gamma_n} \\ &\le \lim_{n\to\infty}\left[-\sum_{x\neq x_0}\Lambda_n(x)\sum_y W(y\,|\,x)[s_n d(x,y)+a_n(x)]\right. \\ &\qquad -\sum_{x\neq x_0}\Lambda_n(x)\sum_y W(y\,|\,x_0) \\ &\qquad \left. \times\left(e^{-s_n d(x,y)-a_n(x)}-1\right)\right].\end{aligned} \tag{68}$$

Observe now that the term inside the braces is a linear function of Λ_n, and using the fact that $\Lambda_n(x) \le 1/g(x)$ is less than

$$\begin{aligned}\max_{x:x\neq x_0} -\frac{1}{g(x)}\sum_y &\Big[W(y\,|\,x)[s_n d(x,y)+a_n(x)] \\ &+ W(y\,|\,x_0)\left(e^{-s_n d(x,y)-a_n(x)}-1\right)\Big].\end{aligned}$$

Let x_n^* be the x that achieves this maximum. Since $\mathcal{X}$ is finite, we can extract a subsequence n_k for which $x_{n_k}^*$ is constant. Letting that value to be x^*, we thus obtain

$$\begin{aligned}C_{\rm LM}^0 \le -\frac{1}{g(x^*)}\lim_{n\to\infty}\sum_y &\Big[W(y\,|\,x^*)[s_n d(x^*,y)+a_n(x^*)] \\ &+ W(y\,|\,x_0)\left(e^{-s_n d(x^*,y)-a_n(x^*)}-1\right)\Big].\end{aligned} \tag{69}$$

We will now show that the above upper bound to $C_{\rm LM}^0$ can be achieved by a binary distribution. To that end, consider a sequence of distributions Q_{X,Γ_n} for which $Q_{X,\Gamma_n}(x)$ is nonzero only for $x = x_0$ and $x = x^*$, and $Q_{X,\Gamma_n}(x^*) = \Gamma_n/g(x^*)$. Let Q_{Y,Γ_n} denote the corresponding output distribution.

By using the dual expression for $I_{\rm LM}$

$$\begin{aligned}I_{\rm LM}(Q_{X,\Gamma_n}) &= \sup_{s\ge 0,a}\sum_{x,y}Q_{X,\Gamma_n}(x)W(y\,|\,x) \\ &\qquad\times\log\frac{e^{-sd(x,y)-a(x)}}{\sum_{x'}e^{-sd(x',y)-a(x')}Q_{X,\Gamma_n}(x')} \\ &\ge \sum_{x,y}Q_{X,\Gamma_n}(x)W(y\,|\,x) \\ &\qquad\times\log\frac{e^{-s_n d(x,y)-a_n(x)}}{\sum_{x'}e^{-s_n d(x',y)-a_n(x')}Q_{X,\Gamma_n}(x')}\end{aligned}$$

where s_n and a_n are those obtained from ν_n. Thus

$$\begin{aligned}I_{\rm LM}&(Q_{X,\Gamma_n}) \\ &\ge -Q_{X,\Gamma_n}(x^*)\sum_y W(y\,|\,x^*)[s_n d(x^*,y)+a_n(x^*)] \\ &\quad -\sum_y Q_{Y,\Gamma_n}(y)\log\Big(1+Q_{X,\Gamma_n}(x^*) \\ &\quad \times\left[\left(e^{-s_n d(x^*,y)-a_n(x^*)}-1\right)\right]\Big) \\ &\ge -Qxs_{X,\Gamma_n}(x^*)\sum_y W(y\,|\,x^*)[s_n d(x^*,y)+a_n(x^*)] \\ &\quad -Q_{X,\Gamma_n}(x^*)\sum_y Q_{Y,\Gamma_n}(y) \\ &\quad \times\left(e^{-s_n d(x^*,y)-a_n(x^*)}-1\right),\end{aligned}$$

where the second inequality follows from $\log(1+x)\le x$. We thus see that

$$\begin{aligned}\frac{I_{\rm LM}(Q_{X,\Gamma_n})}{\Gamma_n} \ge -\frac{1}{g(x^*)}\sum_y &\Big[W(y\,|\,x^*)[s_n d(x^*,y)+a_n(x^*)] \\ &+ Q_{Y,\Gamma_n}(y)\left(e^{-s_n d(x^*,y)-a_n(x^*)}-1\right)\Big].\end{aligned}$$

The boundedness of $e^{-s_n d(x^*,y)-a_n(x^*)}$ allows us to conclude that

$$\begin{aligned}\lim_{n\to\infty}&\frac{I_{\rm LM}(Q_{X,\Gamma_n})}{\Gamma_n} \\ &\ge -\frac{1}{g(x^*)}\lim_{n\to\infty}\sum_y\Big[W(y\,|\,x^*)[s_n d(x^*,y)+a_n(x^*)] \\ &\quad + W(y\,|\,x_0)\left(e^{-s_n d(x^*,y)-a_n(x^*)}-1\right)\Big].\end{aligned} \tag{70}$$

Comparing this expression to the upper bound on $C_{\rm LM}^0$ we see that $C_{\rm LM}^0$ can be achieved by a binary input distribution.

Acknowledgment

The authors wish to thank the anonymous referees for their comments, which improved the paper's clarity.

References

[1] A. Lapidoth and P. Narayan, "Reliable communication under channel uncertainty," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2148–2177, Oct. 1998.

[2] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai (Shitz), "On information rates for mismatched decoders," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1953–1967, Nov. 1994.

[3] I. Csiszár and P. Narayan, “Capacity and decoding rules for classes of arbitrarily varying channels,” *IEEE Trans. Inform. Theory*, vol. 35, pp. 752–769, July 1989.

[4] V. B. Balakirsky, “A converse coding theorem for mismatched decoding at the output of binary-input memoryless channels,” *IEEE Trans. Inform. Theory*, vol. 41, pp. 1889–1902, Nov. 1995.

[5] G. Kaplan and S. Shamai (Shitz), “Information rates of compound channels with application to antipodal signaling in a fading environment,” *AËU*, vol. 47, no. 4, pp. 228–239, 1993.

[6] I. G. Stiglitz, “Coding for a class of unknown channels,” *IEEE Trans. Inform. Theory*, vol. IT-12, pp. 189–195, Apr. 1966.

[7] I. Csiszár and J. Körner, “Graph decomposition: A new key to coding theorems,” *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 5–12, Jan. 1981.

[8] J. Y. N. Hui, “Fundamental issues of multiple accessing,” Ph.D. dissertation, Mass. Inst. Technol., Cambridge, MA, 1983.

[9] V. B. Balakirsky, “Coding theorem for discrete memoryless channels with given decision rules,” in *Proc. 1st French–Sov. Workshop on Algebraic Coding (Lecture Notes in Computer Science)*, G. Cohen, S. Litsyn, A. Lobstein, and G. Zémor, Eds. Berlin, Germany: Springer-Verlag, July 1991, vol. 573, pp. 142–150.

[10] I. Csiszár and P. Narayan, “Channel capacity for a given decoding metric,” *IEEE Trans. Inform. Theory*, vol. 41, pp. 35–43, Jan. 1995.

[11] A. Lapidoth, “Mismatched decoding and the multiple-access channel,” *IEEE Trans. Inform. Theory*, vol. 42, pp. 1439–1452, Sept. 1996.

[12] I. Csiszár, “The method of types,” *IEEE Trans. Inform. Theory*, vol. 44, pp. 2505–2523, Oct. 1998.

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

[14] R. Sundaresan and S. Verdú, “Robust decoding for timing channels,” *IEEE Trans. Inform. Theory*, vol. 46, pp. 405–419, Mar. 2000.

[15] S. Verdú, “On channel capacity per unit cost,” *IEEE Trans. Inform. Theory*, vol. 36, pp. 1019–1030, Sept. 1990.

[16] R. G. Gallager, “Energy Limited Channels: Coding, Multiaccess, and Spread Spectrum,” Laboratory for Information and Decision Systems, Mass. Inst. Technol., Tech. Rep. LIDS-P-1714, Nov. 1988.

[17] M. K. Simon, J. K. Omura, R. A. Scholz, and B. K. Levitt, *Spread Spectrum Communications Handbook*, revised ed. New York: McGraw-Hill, 1994.

[18] I. Csiszár, “Arbitrarily varying channel with general alphabets and states,” *IEEE Trans. Inform. Theory*, vol. 38, pp. 1725–1742, Nov. 1992.

[19] R. T. Rockafellar, *Convex Analysis*. Princeton, NJ: Princeton Univ. Press, 1970.

[20] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[21] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.

[22] I. Vajda, *Theory of Statistical Inference and Information*. Boston, MA: Kluwer, 1989.

[23] S. Verdú and T. S. Han, “A general formula for channel capacity,” *IEEE Trans. Inform. Theory*, vol. 40, pp. 1147–1157, July 1994.

[24] İ. E. Telatar, “Multi-access communications with decision feedback decoding,” Ph.D. dissertation, Mass. Inst. Technol., Cambridge, MA, May 1992.

[25] A. Lapidoth, “Nearest-neighbor decoding for additive non-Gaussian noise channels,” *IEEE Trans. Inform. Theory*, vol. 42, pp. 1520–1529, Sept. 1996.