Linear Universal Decoding for Compound Channels

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

Citation
Abbe, Emmanuel, and Lizhong Zheng. “Linear Universal Decoding for Compound Channels.” IEEE Transactions on Information Theory 56.12 (2010) : 5999-6013. Copyright © 2010, IEEE

As Published
http://dx.doi.org/10.1109/tit.2010.2080910

Publisher
Institute of Electrical and Electronics Engineers / IEEE Information Theory Society

Version
Final published version

Accessed
Fri Dec 07 21:55:42 EST 2018

Citable Link
http://hdl.handle.net/1721.1/65131

Terms of Use
Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.

Detailed Terms
Abstract—Over discrete memoryless channels (DMC), linear decoders (maximizing additive metrics) afford several nice properties. In particular, if suitable encoders are employed, the use of decoding algorithms with manageable complexities is permitted. For a compound DMC, decoders that perform well without the channel’s knowledge are required in order to achieve capacity. Several such decoders have been studied in the literature; however, there is no such known decoder which is linear. Hence, the problem of finding linear decoders achieving capacity for compound DMC is addressed, and it is shown that under minor concessions, such decoders exist and can be constructed. A geometric method based on the very noisy transformation is developed and used to solve this problem.

Index Terms—Additive decoders, compound channels, hypothesis testing, information geometry, mismatch, universal decoders.

I. INTRODUCTION

We consider a discrete memoryless channel with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$. The channel is described by the probability transition matrix $W$, each row of which is the conditional distribution of the output symbol conditioned on a particular input $x \in \mathcal{X}$. We are interested in the compound channel, where the exact value of $W$ is not known, either at the transmitter or the receiver. Such problems can be motivated by the wireless applications with unknown fading realizations, which are discussed in [5] and [18]. In the compound setting, the channel $W$ is not assumed to be known at the receiver and transmitter; instead, a set $S$ of possible channels is assumed to be known at the receiver and transmitter, and the goal is to design encoders and decoders that support reliable communication, no matter which channel in $S$ takes place for communication. Compound channels have been extensively studied in the literature [6], [24], [12], [13]. In particular, Blackwell et al. [6] showed that the highest achievable rate is given by the following expression:

$$C(S) \triangleq \max_P \inf_W I(P,W)$$  \hspace{1cm} (1)

where the maximization is over all probability distributions $P$ on $\mathcal{X}$ and where $I(P,W)$ denotes the mutual information for the input distribution $P$ and the channel $W$. Thus, $C(S)$ is referred to as the compound channel capacity. To achieve this capacity, i.i.d. random codes from the optimal input distribution, i.e., the distribution maximizing (1), are used in [6]. The random coding argument is commonly employed to prove achievability for a given channel, such as in Shannon’s original paper, where it is shown that, since the error probability averaged over the random ensemble can be made arbitrarily small, there exists “good” codes with low error probability. This argument is strengthened in [6] to show that with the random coding argument, the existence of codes that are good for all possible channels can also be deduced. Adopting this point of view in this paper, we will not be concerned with constructing the code book, or even finding the optimal input distribution, we will rather simply assume that an optimal random code realization is used, and focus on the design of efficient decoding algorithms.

In [6], a decoder that maximizes a uniform mixture of likelihoods over a dense set of representative channels in the compound set is used, and shown to achieve compound capacity. Another universal decoder is the maximum mutual information (MMI) decoder [11], which computes the empirical mutual information between each codeword and the received output and picks the codeword with highest score. The practical difficulty of implementing MMI decoders is obvious. As empirical distributions are used in computing the “score” of each codeword, it becomes challenging to efficiently store the exponentially many scores, and update the scores as symbols being received sequentially. Conceptually, when the empirical distribution of the received signals is computed, one can in principle estimate the channel $W$, making the assumption of lack in channel knowledge less meaningful. There has been a number of different universal decoders proposed in the literature, including the LZ based algorithm [20], or merged likelihood decoder [15]. In this paper, we try to find universal decoders in a class of particularly simple decoders: linear decoders.

Here, linear (or additive) decoders are defined to have the following structure. Upon receiving the $n$-symbol word $y$, the decoder compute a score/decoding metric $d^n(x_m,y)$ (note that the score of a codeword does not depend on other codewords) for each codeword $x_m$, $m = 1,2,\ldots, [2^{nR}]$, and decodes to the one codeword with the highest score (ties can be resolved arbitrarily). Moreover, the $n$-symbol decoding metric has the following additive structure:

$$d^n(x_m,y) = \sum_{i=1}^{n} d(x_m(i),y(i))$$

where $d : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ is a (single-letter) decoding metric. We call such decoders linear since the decoding metric it computes is indeed linear in the joint empirical distribution between the codeword and the received word:

$$d^n(x_m,y) = n \cdot \sum_{a \in \mathcal{X}, b \in \mathcal{Y}} \hat{P}(x_m,y)(a,b) \cdot d(a,b)$$
where $\hat{P}(x,y)$ denotes the joint empirical distribution of $(x_m,y)$. We also say that such a decoder is induced by the single-letter metric $d$. For example, the maximum likelihood (ML) decoder is a linear decoder, induced by the single-letter metric $d = \log W$, the log likelihood of the channel.

Linear decoders have been widely studied in [12], [21]. An additive decoding metric has several advantages. First, when used with appropriate structured codes, it allows the decoding complexity to be reduced; for example, when convolutional codes are used, Viterbi algorithm can then be employed to perform the decoding. Moreover, additive structures are also suitable for message passing algorithms, such as belief propagation. It is worth clarifying that the complexity reduction discussed here rely on certain structured codes being used, in the place of the random codes. In this paper, however, our analysis will be based on the random coding argument, with the implicit conjecture that there exist structured codes resembling the behavior of random codes for the considered problem. This kind of argument has indeed been developed in [14] and [7] for linear codes. Mathematically, as observed in [12] and [21], linear decoders are also more interesting in the sense that they create an interplay with different information theory and combinatorial problems, such as the graph-theoretical concept of Shannon capacity and Sperner capacity, or the zero-error (erasures-only) capacity.

It is not surprising that for some compound channels, a linear universal decoder does not exist. In [12] and [21], it is shown that $S$ being convex and compact is a sufficient condition for the existence of linear decoders which achieve compound capacity. In this paper, we give a more general sufficient condition for a set to admit a compound capacity achieving linear decoder, namely that $S$ is one-sided. For more general compound sets, in order to achieve the capacity, we have to resort to a relaxed restriction of the decoders, which we call generalized linear decoders. A generalized linear decoder, maximizes a finite number, $K$, of decoding metrics, $d_1, d_2, \ldots, d_K$. The decoding map can then be written as

$$\arg \max_{m} \sum_{k=1}^{K} d_k(x_m, y) = \arg \max_{m} \sum_{i=1}^{n} d_k(x_m(i), y(i))$$

where $\arg \max$ denotes the maximum operator. Note that the receiver calculates in parallel $K$ additive metrics for each codeword, and decodes to the codeword with the highest score among the total $2^{nR} \times K$ scores. We emphasize the restriction that $K$ has to be finite (not depending on the codeword length $n$) in order to preserve the complexity properties discussed previously for linear decoders. An example of generalized linear decoder is the well-known generalized loglikelihood ratio test (GLRT) when the number of likelihood ratio tests is finite. The decoder proposed in [6], a mixture of likelihoods over a dense set of channels, in general might require averaging over polynomial($n$) channels. In addition, optimizing the mixture of additive metrics, i.e., $\arg \max_{m} \sum_{k=1}^{K} d_k(x_m, y)$, cannot be solved by computing $K$ parallel additive metric optimizations: the codewords having the best scores for each of the $K$ metrics may not be the only candidates for the best score of the mixture of the metrics; on the other hand, if we consider a generalized linear decoder, the codewords having the best score for each of the $K$ metrics are the only one to be considered for the maximum of the $K$ metrics (hence parallel maximizations of the metrics can be performed).

The main result of this paper is the construction of generalized linear decoders that achieve the compound capacity on a large variety of compound sets. As shown in Section II, this construction requires solving some rather complicated optimization problems involving the Kullback-Leibler (KL) divergence (like almost every other information theoretical problem). To obtain insights to this problem, we introduce in the appendix a tool called the local geometric analysis: in a nutshell, we focus on the special cases where the two distributions of the KL divergence are “close” to each other, which can be thought in this context as approximating the given compound channels by very noisy channels. These channels can be used to model communication schemes at a low SNR regime, such as in Gaussian additive noise channels (with fading) for which power is small with respect to the noise variance. However our goal here is not to analyze the performance of linear decoders in the very noisy regime per se; we are interested in this local setting, because the information theoretical quantities can then be understood as quantities in an inner product space, where conditional distributions and decoding metrics correspond to vectors; divergence and mutual information correspond to squared norms and achievable rates with mismatched linear decoders can be understood with projections. The relationship between these quantities can then be understood geometrically and this provides insight to the local but also general problem. Indeed, while the results from such local approximations only apply to the special very noisy cases, we show in Appendix E that some of these results can be “lifted” to a naturally corresponding statement in the general setting.

The results in this paper can be summarized as follows.

- First we derive a new condition on compound sets called “one-sidedness”, cf. Definition 3, under which a linear decoder maximizing the log likelihood of the worst channel in the compound set achieves compound capacity. This condition is more general than the previously known one, which requires $S$ to be convex;
- Then, we show in our main result, that if the compound set $S$ can be written as a finite union of one sided sets, then a generalized linear decoder using the log a posteriori distribution of the worst channels of each one-sided component achieves the compound capacity; in contrast, GLRT using these worst channels is shown not to be a universal decoder.

One way to interpret this result is to compare the proposed decoder with the MMI. As discussed in more details in Section IV, the MMI decoder can be viewed as a “degenerated generalized linear decoder” with infinitely many metrics which are the MAP metrics of all possible DMCs. Our result says that, for a given compound set, instead of picking all possible DMCs metrics which is indeed wasteful, one can select carefully a much smaller subset of DMCs and still achieve the compound capacity. How to select the subset of DMCs is given by to the notion of one-sided components, and it is argued that the number of these special DMCs is in general finite. In the next

1Note that our goal here is limited to achieving the compound capacity and not the random coding error exponent.
section, we start with the precise problem formulation and the notation.

II. PROBLEM STATEMENT

We consider discrete memoryless channels with input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively. The channel is often written as a probability transition matrix, $W$, of dimension $|\mathcal{X}| \times |\mathcal{Y}|$, each row of which denotes the conditional distribution of the output, conditioned on a specific value of the input. We are interested in the compound channel, where $W$ can be any elements of a given set $\mathcal{S}$, referred to as the set of possible channels, or the compound set. For convenience, we assume $S$ to be compact. The value of the true channel is assumed to be fixed for the entire duration of communications, but not known at either the transmitter or the receiver; only the compound set $S$ is assumed to be known at both.

We assume that the transmitter and the receiver operates synchronously over blocks of $n$ symbols. In each block, a data message $m \in \{1,2,\ldots,|2^{nR}|\}$ is mapped by an encoder

$$F^n_m : \{1,2,\ldots,|2^{nR}|\} \mapsto \mathcal{X}^n$$

to $F^n_m(m) = x_m \in \mathcal{X}^n$, referred to as the $m$th codeword. The receiver observes the received word, drawn from the distribution

$$W^n(y|x_m) = \prod_{i=1}^n W(y(i)|x_m(i))$$

and applies a decoding map

$$G^n : \mathcal{Y}^n \mapsto \{1,2,\ldots,|2^{nR}|\}.$$

The average probability of error, averaged over a given code $(F^n_m, G^n)$, for a specific channel $W$, is written as

$$P_e(F^n_m, G^n, W) = \frac{1}{|2^{nR}|} \sum_{m=1}^{|2^{nR}|} \sum_{y \notin G^n_m(y) \neq m} W^n(y|x_m).$$

A rate $R$ is said to be achievable for the given compound set $S$ if it is $\varepsilon$-achievable for every $0 < \varepsilon < 1$, i.e., if for every $\delta > 0$ and every $n$ sufficiently large, there exists a block length $n$ and $(F^n_m, G^n)$ of rate at least $R - \delta$ such that $\sup_{W \in S} P_e(F^n_m, G^n, W) \leq \varepsilon$. The supremum of all achievable rates is called the \textit{compound channel capacity}, written as $C(S)$.

The result from Blackwell et al. gives the compound channel capacity with a single-letter expression.

\textbf{Lemma 1 (\cite{3})}:

$$C(S) = \max_{P_X} \inf_{W \in S} I(P_X, W).$$

As mentioned in previous section, a first difficulty encountered when going from the known channel to the compound channel setting is regarding the random coding argument. It is not enough to show that the ensemble average error probability is small for every $W$ in the compound set, since the “good”

\footnote{Since the set of all possible channels on fixed input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$ is compact subset of $[0,1]^{|\mathcal{X}| \times |\mathcal{Y}|}$, $\mathcal{S}$ is compact if it is closed, and this is just a minor technical assumption with no real impact on the results.}

codes for different channels can in principle be different, and hence this does not guarantee the existence of a single code that is universally good for all possible channels. However, it is shown in \cite{6} that this problem does not happen and the random coding argument can still be used. The second and bigger difficulty is regarding the decoding: since the channel is not known, there is no notion of typicality or likelihood. Hence, the decoding needs to be revisited in the compound setting. Before we proceed to define particular decoders, we introduce some notations:

- $W_0 \in S$ denotes the true channel under which communication takes place;
- for a joint distribution $\mu$ on $\mathcal{X} \times \mathcal{Y}$; $\mu_X$ and $\mu_Y$ denote respectively the marginal distributions on $\mathcal{X}$ and $\mathcal{Y}$; and $\mu^p = \mu_X \times \mu_Y$ the induced product distribution. Note that $\{\mu_X = (\mu_0)_X, \mu_Y = (\mu_0)_Y\} \Leftrightarrow \mu^p = \mu_0^p$;
- $\mu = P \circ W$ denotes the joint distribution with $P$ as the marginal distribution on $\mathcal{X}$ and $W$ as the conditional distribution. For example, the mutual information is expressed as

$$I(P; W) = D(P \circ W || (P \circ W)^p)$$

where $D(\cdot || \cdot)$ is the Kullback-Leibler divergence.

All decoders considered in this paper have the following form. Upon receiving $y$, it computes, for each codeword $x_m$, a score $d^n(x_m, y)$, and decodes to the message corresponding to the highest score. Note the restriction here is that the score for codeword $x_m$ does not depend on other codewords; such decoders are called $\alpha$-decoders in \cite{12}. As an example, the maximum mutual information (MMI) decoder has a score defined as

$$d^n_{\text{MMI}}(x_m, y) = I(\hat{P}(x_m,y))$$

where $\hat{P}$ denotes the empirical distribution

$$\hat{P}_{(x_m,y)}(a,b) = \frac{1}{n} \{i : (x_m(i), y(i)) = (a,b)\}$$

and $I(\mu)$ denotes the mutual information, as a function of the joint distribution $\mu$ on $\mathcal{X} \times \mathcal{Y}$.

It is well known that the MMI decoder is universal; when used with the optimal code, it achieves the compound channel capacity on any compound set. In fact, there are other advantages of the MMI decoder: it does not require the knowledge of $S$ and it achieves universally the random coding error exponent \cite{11}. Despite these advantages, the practical difficulties to implement an MMI decoder prevents it from becoming a real “universally used” decoder. As empirical distributions are used in computing the scores, it is difficult to store and update the scores, even when a structured codebook is used. The main goal of the current paper is to find decoders that are \textit{linear} and that also can, like the MMI decoder, be capacity achieving on compound channels.

\textbf{Definition 1 (Linear Decoder)}: We refer to a map

$$d : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$$
as a single-letter metric. A linear decoder induced by \( d \) is defined by the decoding map

\[
G_n(y) = \arg \max_m d^n(x_m, y)
\]

where

\[
d^n(x_m, y) = \frac{1}{n} \sum_{i=1}^n d(x_m(i), y(i)) = E_{P_{X^n,Y^n}[d_x^y]}
\]

Note that the reason why such decoders are called linear decoders (\( d \)-decoders in [12]) is to underline the fact that the decoding metric is additive, i.e., is a linear function of the empirical distribution \( \hat{P}_{X^n,Y^n} \).

The advantages of using linear decoders have been discussed thoroughly in [12], [21], and [17], and also briefly in the Introduction. Unfortunately, there are examples of compound sets for which no linear decoder can achieve the compound capacity. A well-known example is the compound set with two binary symmetric channels, with crossover probabilities \( 1/4 \) and \( 3/4 \). To address the decoding challenge of these cases, we will need a slightly more general version of linear decoders.

**Definition 2 (Generalized Linear Decoder):** Let \( d_1, d_2, \ldots, d_K \) be \( K \) single-letter metrics, where \( K \) is a finite number. A generalized linear decoder induced by these metrics is defined by the decoding map

\[
G_n(y) = \arg \max_m \bigvee_{k=1}^K d_k(x_m(i), y(i)) = \arg \max_m \bigvee_{k=1}^K E_{P_{X^n,Y^n}[d_k]}
\]

where \( \bigvee \) denotes the maximum. Note that we insist that \( K \) is a finite number, which does not depend on the code length \( n \).

The maximum likelihood decoder, of a given channel \( W \), is a linear decoder induced by

\[
d_{ML}(a,b) = \log W(b|a), \quad \forall a \in \mathcal{X}, b \in \mathcal{Y}
\]

For a given channel \( W \), the ML decoder, used with a random code from the optimal input distribution, is capacity achieving. If the channel knowledge is imperfect and if the decoder uses the ML rule for the channel \( W_{1} \) while the actual channel is \( W_{0} \), the mismatch in the decoding metric causes the achievable data rate to decrease. This effect has first been studied in [12] and [21], with a result quoted in the following Lemma. Mismatched decoding has also been investigated in [23] and [18].

**Lemma 2 ([12], [21]):** For a DMC \( W_{0} \), using a random codebook with input distribution \( P_{X} \), if the decoder is linear and induced by \( d \), the following rate can be achieved:

\[
R(P_{X}, W_{0}, d) = \inf_{\mu \in \mathcal{A}} D(\mu || \mu_{0}^*)
\]

where \( \mu_{0} = P_{X} \circ W_{0} \), and \( \mu_{0}^* \) is the product distribution with the same \( X \) and \( Y \) marginal distributions as \( \mu_{0} \) and the optimization is over the following set of joint distributions on \( \mathcal{X} \times \mathcal{Y} \):

\[
\mathcal{A} = \{ \mu : \mu_{X} = P_{X}, \mu_{Y} = (\mu_{0})_{Y}, \mathbb{E}_{\mu}[d] \geq \mathbb{E}_{\mu_{0}}[d] \}
\]

As discussed in [21], this expression, even for the optimal input distribution \( P_{X} \), does not give in general the highest achievable rate under the matched scenario. If the input alphabet is binary, [4] shows that this expression gives the highest achievable rate, otherwise it only gives the highest rate that can be achieved for codes that are drawn in a random ensemble [12], [21].

Similarly, when a generalized linear decoder induced by the single-letter metrics \( \{d_k\}_{k=1}^K \) is used, we can achieve the following rate:

\[
R(P_{X}, W_{0}, \{d_k\}_{k=1}^K) = \inf_{\mu \in \mathcal{A}} D(\mu || \mu_{0}^*)
\]

where

\[
\mathcal{A} = \{ \mu : \mu_{X} = P_{X}, \mu_{Y} = (\mu_{0})_{Y}, \mathbb{E}_{\mu}[d] \geq \mathbb{E}_{\mu_{0}}[d] \}
\]

\[
\bigvee_{k=1}^K E_{\mu}[d_k] \geq \mathbb{E}_{\mu_{0}}[d]\}
\]

The proof of this Lemma is similar to the proof of Lemma 2, details are provided in [1].

Note that \( R(P_{X}, W_{0}, \{d_k\}_{k=1}^K) \) can equivalently be expressed as

\[
R(P_{X}, W_{0}, \{d_k\}_{k=1}^K) = \inf_{\mu \in \mathcal{A}} D(\mu || \mu_{0}^*) \wedge \ldots \wedge \inf_{\mu \in \mathcal{A}_k} D(\mu || \mu_{0}^*_k)
\]

where \( \wedge \) denotes the minimum operator and

\[
\mathcal{A}_k = \{ \mu : \mu_{X} = P_{X}, \mu_{Y} = (\mu_{0})_{Y}, \mathbb{E}_{\mu}[d_k] \geq \mathbb{E}_{\mu_{0}}[d_k] \}
\]

Now we are ready for the main problem studied in this paper. For a given compound set \( S \), let the compound channel capacity be \( C(S) \) and the corresponding optimal input distribution be \( P_{X} \). We would like to find \( K \) and \( d_1, \ldots, d_K \), such that

\[
R(P_{X}, W_{0}, \{d_k\}_{k=1}^K) \geq C(S)
\]

for every \( W_{0} \in S \).

If this holds, the generalized linear decoder induced by the metrics \( \{d_k\}_{k=1}^K \) achieves compound capacity on the set \( S \) (i.e., using analogue arguments as for the achievability proof of the compound capacity in [6], there exists a code book that makes the overall coding scheme capacity achieving).

### III. RESULTS

Recall that the optimal input distribution of a set \( S \) is given by

\[
P_X = \arg \max_{P} \inf_{W \in S} I(P, W)
\]

where the maximization is over all probability distributions on \( \mathcal{X} \), and if the maximizer is not unique, we define \( P_{X} \) to be any arbitrary maximizer. We denote by \( c(S) \) the closure of \( S \).
Definition 3 (One-Sided Set): A set $S$ is one-sided with respect to an input distribution $P$ if the following minimizer is unique:

$$W_S = \arg \min_{W \in \mathfrak{g}(S)} J(P,W)$$

if $\text{Support}(\mu_0) \subseteq \text{Support}(\mu_S)$, for all $W_0 \in S$ and if

$$D(\mu_0 \| \mu_{S}^k) \geq D(\mu_0 \| \mu_S) + D(\mu_S \| \mu_{S}^k), \quad \forall W_0 \in S$$

where $\mu_0 = P \circ W_0$ and $\mu_S = P \circ W_S$.

A set $S$ is one-sided if it is one-sided with respect to $P_X$ (the optimal input distribution for $S$).

A set $S$ is a finite union of one-sided sets if each set in the union is one-sided with respect to $P_X$.

Lemma 4: Convex sets are one-sided and there exist one-sided sets that are not convex.

Remark 1: The intuition behind one-sided sets comes from a geometric argument, presented in details in Section D of the Appendix. Briefly, (8) is a Pythagorean inequality [10], and is an analogue of the inequality satisfied for the Euclidean squared norm of a triangle with vertices $\mu_0, \mu_S$ and $\mu_{S}^k$, with an obtuse angle at $\mu_S$. Hence, a set is one-sided, if it is contained in the region delimited by the “half-plane” passing by $\mu_S$ and on the other side of $\mu_{S}^k$. A basic example of set which is not one-sided, but is a union of two one-sided sets is the compound set containing two BSCs, with cross-over probabilities 1/4 and 3/4 (in this case each one-sided set is the singleton set containing each BSC). Note that the set consisting of the BSCs with cross-over probabilities $\{1/5, 1/4, 3/4, 4/5\}$ is also the union of two one-side sets (BSCs $\{1/5, 1/4\}$ and BSCs $\{3/4, 4/5\}$).

Lemma 5: For one-sided sets $S$, the linear decoder induced by the metric $d = \log W_S$ is compound capacity achieving. Note that in [12], the same linear decoder is proved to be compound capacity achieving for the case where $S$ is finite.

Lemma 6: For a set $S$ which is countable, the decoder maximizing the score function $\sup_{W \in S} \log W^y$, i.e., the GLRT maximizing the metrics $\{\log W : W \in S\}$, is compound capacity achieving, but generalized linear only if $S$ is finite.

The following theorem is the main result of the paper.

Theorem 1: For $S = \bigcup_{k=1}^K S_k$, which is a finite union of one-sided sets, the generalized linear decoder induced by the metrics $d_k = \log W_{S_k}$ (called MAP metrics), for $1 \leq k \leq K$ is compound capacity achieving.

Remark 2: For $S = \bigcup_{k=1}^K S_k$, which is a finite union of one-sided sets, the generalized linear decoder induced by the metrics $d_k = \log W_{S_k}$, for $1 \leq k \leq K$, (which is the GLRT decoder based on these metrics) is not compound capacity achieving in general. We provide a counterexample in Section F of the Appendix.

Interpretation of the results:

Lemma 6 tells us that for an arbitrary compound set $S$, by compactness of the set of possible channels on fixed finite alphabets, we could pick an increasing number of representative channels which are “dense” in $S$ and achieve rates that get arbitrarily close to the compound capacity with the GLRT decoder. However, this may require a very large number of channels and the attribute of using generalized linear decoders may be compromised. On the other hand, in view of Lemma 5, we know that when the compound set has a certain structure (one-sided), there is no need to go for this dense representation of $S$, indeed in this case, a simple maximum likelihood test is sufficient. Thus, a natural question is, can we combine these two results, and avoid going through a naive dense representation by picking the representational channels in a smart way. Using GLRT as described in Remark 2 has been our first attempt, but did not lead to the desired result. The right approach came by considering the decoder of Theorem 1, which is the GMAP instead of GLRT of the worst channels (the difference between these two decoders is discussed in Section G of the Appendix). While it is easy to construct an example of compound set with an infinite number of disconnected convex components, the notion of finite union of one-sided sets is general enough to include a large variety of compound sets (probably most compound sets that one may be exposed to). Yet, one can construct sets which are not even a finite union of one-sided sets. These points are further discussed in the next section.

IV. DISCUSSIONS

We raised the question whether it is possible for a decoder to be both linear and capacity achieving on compound channels. We showed that if the compound set is a union of one-sided sets, a generalized linear decoder which is capacity achieving exists. We constructed it as follows: if $W_1, \ldots, W_k$ are the worst channels of each component (cf. Fig. 1), use the generalized linear decoder induced by the MAP metrics $\log W_1, \ldots, \log W_k$, i.e., decode with

$$G_n(y) = \arg \max_{m \in \{1, \ldots, M\}} \max_{k=1}^K \mathbb{E}_{P_{f}(x_{m,y})} \log \frac{W_k}{P_X(y)}$$

where $\mu_k = P_X \circ W_k$, $P_X$ is the optimal input distribution on $S$, and $P_{f}(x_{m,y})$ is the joint empirical distribution of the $m$th codeword $x_m$ and the received word $y$. We denote this decoder by GMAP$(W_1, \ldots, W_K)$. We also found that using the ML metrics, instead of the MAP metrics for $W_1, \ldots, W_k$, i.e., GLRT$(W_1, \ldots, W_K)$, is not capacity achieving.

It is instrumental to compare our receiver with the MMI receiver. We observe that the empirical mutual information is given by

$$I(\hat{P}_{f}(x_{m,y})) = \sup_{W} \mathbb{E}_{P_{f}(x_{m,y})} \log \frac{W}{P_X \circ W} \quad (9)$$

where the maximization is taken over all possible DMC $W$, which means that the MMI is actually the GMAP decoders taking into account all DMCs. Our result says that we do not need to integrate all DMC metrics to achieve compound capacity; for a given compound set $S$, we can restrict ourselves to selecting carefully a subset of metrics and yet achieve the compound capacity. Those representative metrics are found by extracting the one-sided components of $S$, and taking the MAP metrics induced by the worst channel of these components.
When $S$ has a finite number of one-sided components, this decoder is generalized linear (the number of metrics is equal to the number of one-sided components).

In order do extract the one-sided components of an arbitrary compound set $S$, one can proceed as illustrated in Fig. 1. We first look for the worst channel $W_1$ of $S$. By decoding with the MAP metric of this channel, we can “support” all channels that are dominated by it (one-sided property). Then, one look at the remaining set of channels which are not dominated by $W_1$, and iterate the process, as illustrated in Fig. 1. This abstract figure, which is an accurate representation in the very noisy setting as explained in the Appendix, suggests that for most naturally picked compound set, the procedure of extracting one-sided components should stop with a finite number of iterations. Indeed, from a practical point of view, if the compound set is the result of an estimation of the channel, one may get directly the worst possible channels with the estimation procedure. Namely, assume that $\hat{W}_1, \ldots, \hat{W}_m$ are $m$ estimations of a channel $W$ over which one wishes to communicate reliably. One may decide to design a code for the average of the estimates $\hat{W}_1, \ldots, \hat{W}_m$, which may get close to $W$: this depends on how large $m$ is and on how ergodic the estimation process is. In general, this induces an outage probability, as in [22]. Another approach, more cautious, consist in designing the code by assuming that the true channel can be anyone of the $m$ estimates, i.e., a compound set containing all estimates. However, if one proceeds naively by using all the channel’s metrics, the attributes of generalized linear decoders may be lost (as $m$ is typically large). This is where our result can be used to circumvent this problem, since it says that it is sufficient to consider the “worst” estimated channels (the channels dominated each one-sided components as in Fig. 1), and that typically, these channels are much less numerous. This also shows that it is convenient to work with one-sided components instead convex components, since the number of convex components in a discrete set is given by the set cardinality. In contrast, to construct a set with infinitely many one-sided components, one should build the set $S$ such that the number of worst channel, in one of the iterations described in Fig. 1, is infinite. This can happen, if the set is perfectly overlapping with a level curve of constant mutual information; but this is mainly a pedagogical example. Moreover, if one is willing to give up a small fraction of the capacity, then a finite collection of linear decoding metrics would suffice for such cases too. We indeed believe that for this problem there is a graceful tradeoff between the number of metrics used, and the loss in the achievable rate (this represents a possible extension of this work).

Another interesting extension is to develop a notion of “blind” generalized linear decoder, which does not even require the knowledge of the compound set, yet guarantees to achieve a fraction of the compound capacity. We describe here such decoders in the very noisy setting, which is discussed in details in the Appendix. A reader who is not familiar with this setting or the appendix may simply consider the forthcoming figure as an abstract geometrical representation. As illustrated in Fig. 2, such decoders are induced by a set of metrics chosen in a “uniform” fashion. For a given compound set, we can then grow a polytope whose faces are the hyperplane orthogonal to these metrics and there will be a largest such polytope, that contains the entire compound set in its complement. This determines the rate that can be achieved with such a decoder on a given compound set, cf. $C_{\text{poly}}$ in Fig. 2. In general $C_{\text{poly}}$ is strictly less than the compound capacity, denoted by $C$ in Fig. 2; the only cases where $C = C_{\text{poly}}$ is if by luck, one of the uniform direction is along the worst channel (and if there are enough metrics to contain the whole compound set). Now, for a number $K$ of metrics, no matter what the compound set looks like, and not matter what its capacity is, the ratio between $C_{\text{poly}}$ and $C$ can be estimated: in the very noisy (VN) geometry, this is equivalent to picking a sphere with radius $C$ and to compute the ratio between $C$ and the “inner radius” of a $K$-polytope inscribed in the sphere. It is also clear that the higher the number of metrics is, the closer $C_{\text{poly}}$ to $C$ is, and this controls the tradeoff between the computational complexity and the achievable rate. Again, as suggested by the very noisy picture, there is a graceful tradeoff between the number of metrics used, and the loss in achievable rate. Note that this problem has been investigated in [2] for binary alphabet channels.
V. PROOFS

Proof of Lemma 4: Let $C$ be a convex set, then for any input distribution $P_X$ the set $D = \{ \mu | \mu(a,b) = P_X(a)W(b | a), W \in C \}$ is a convex set as well. For $\mu$ such that $\mu(a,b) = P_X(a)W(b | a)$, we have

$$D(\mu \parallel p_C^p) = I(P_X, W) + D(\mu_Y \parallel (\mu_C)_Y)$$

hence we obtain, by definition of $W_C$ being the worse channel of $cl(C)$

$$\mu_C = \min_{\mu \in cl(D)} D(\mu \parallel p_C^p).$$

Therefore, we can use theorem 3.1 in [10] and for any $\mu_0 \in D$, we have the Pythagorean inequality for convex sets

$$D(\mu_0 \parallel p_C^p) \geq D(\mu_0 \parallel \mu_C) + D(\mu_C \parallel p_C^p). \quad (10)$$

This concludes the proof of the first claim of the lemma. Now to construct a one-sided set that is not convex, one can simply take a convex set and remove one point in the interior, to create a “hole”. This does not affect the one-sidedness, but makes the set non-convex. It also shows that there are sets that are one-sided (and not convex) for all input distributions.

Proof of Lemma 5: The mismatched mutual information is given by

$$R(P_X, W_0, \log W_S) = \inf_{\mu \in A_S} D(\mu \parallel p^p) \quad (11)$$

where

$$A_S = \{ \mu_X = P_X, \mu_Y = (\mu_0)_Y, E_{\mu} \log W_S \geq E_{\mu_0} \log W_S \}.$$

Since we consider here a linear decoder, i.e., induced by only one single-letter metric, we can consider equivalently the ML or MAP metrics. We here work with the MAP metric and the constraint set is equivalently expressed as

$$A_S = \{ \mu_X = P_X, \mu_Y = (\mu_0)_Y, \quad E_{\mu} \log \frac{W_S}{(\mu_S)_Y} \geq E_{\mu_0} \log \frac{W_S}{(\mu_0)_Y} \}.$$

Expressing the quantities of interest in terms of divergences, we write

$$E_{\mu} \left[ \log \frac{W_S}{(\mu_S)_Y} \right] = E_{\mu} \left[ \log \frac{W_S}{(\mu_S)_Y} \mu^{p^p} \right] = D(\mu \parallel p^p) - D(\mu \parallel \mu_S) + D(p^p \parallel \mu_S).$$

Similarly, we have

$$E_{\mu_0} \log \frac{W_S}{(\mu_0)_Y} = D(\mu_0 \parallel p_0^p) - D(\mu_0 \parallel \mu_S) + D(p_0^p \parallel \mu_S).$$

Thus we can rewrite $A_S$ as

$$A_S = \{ \mu : \mu_X = P_X, \mu_Y = (\mu_0)_Y \times D(\mu \parallel p^p) - D(\mu \parallel \mu_S) + D(p^p \parallel \mu_S) \geq D(\mu_0 \parallel p_0^p) - D(\mu_0 \parallel \mu_S) + D(p_0^p \parallel \mu_S) \} \quad (12)$$

We then have that

$$D(\mu \parallel \mu_S) - D(p^p \parallel \mu_S^p) \geq 0$$

is a direct consequence of log-sum inequality, and with this, we can write for all $\mu \in A_S$

$$D(\mu \parallel p^p) \geq D(\mu_0 \parallel p_0^p) - D(\mu_0 \parallel \mu_S) + D(p_0^p \parallel \mu_S) \quad (13)$$

which is in turn lower bounded by $D(\mu_S \parallel \mu_S^p) = I(P_X, W_S)$, provided that the set $S$ is one-sided, cf. (8).

Proof of Lemma 6: We need to show that

$$\inf_{\mu \in A_S} \{ \mu_0, \mu \} \log W_1 \geq \inf_{\mu \in A_S} \{ \mu_0, \mu \} \log W_0 \geq I(P_X, W_1),$$

and we will see that the left-hand side of this inequality is equal to $I(P_X, W_0)$. Note that $\forall W_1 \in E_{\mu_0} \log W = E_{\mu_0} \log W_0 = I(P_X, W_0)$. Thus, the desired inequality is equivalent to ask that for any $W_1 \in S$

$$\inf_{\mu \in A_S} \{ \mu_0, \mu \} \log W_1 \geq \inf_{\mu_0, \mu} \{ \mu_0, \mu \} \log W_0 \geq I(P_X, W_1).$$

(15)

Using the marginal constraint $\mu^p = \mu_0^p$, we have

$$E_{\mu} \log W_1 \geq E_{\mu_0} \log W_0 \quad \Leftrightarrow \quad E_{\mu} \left[ \log \frac{W_1}{\mu^p} \right] \geq E_{\mu_0} \left[ \log \frac{W_0}{\mu_0^p} \right] \quad \Leftrightarrow \quad D(\mu \parallel p^p) - D(\mu \parallel \mu_1) \geq D(\mu_0 \parallel p_0^p) \quad (16)$$

hence

$$\inf_{\mu \in A_S} \{ \mu_0, \mu \} \log W_1 \geq \inf_{\mu \in A_S} \{ \mu_0, \mu \} \log W_0 \quad \Leftrightarrow \quad D(\mu \parallel p^p) \geq D(\mu \parallel \mu_1) \quad (17)$$

and $D(\mu \parallel p^p) \geq D(\mu_0 \parallel p_0^p) + D(\mu \parallel \mu_1)$ for all $\mu$ admissible, which implies, since $D(\mu \parallel \mu_1) \geq 0$, that (17) is lower bounded by

$$D(\mu_0 \parallel p_0^p) = I(P_X, W_0). \quad (18)$$

This concludes the proof of the lemma. In fact, one could get a tighter lower bound by expressing (16) as

$$E_{\mu} \log W_1 \geq E_{\mu_0} \log W_0 \quad \Leftrightarrow \quad D(\mu \parallel p^p) - (D(\mu \parallel \mu_1) - D(p^p \parallel \mu_1)) \geq D(\mu_0 \parallel p_0^p) + D(p_0^p \parallel \mu_1).$$
and using the log-sum inequality to show that $D(\mu \mid \mu_0) - D(\mu \mid \mu_1^0) \geq 0$, (17) is lower bounded by

$$D(\mu_0 \mid \mu_1^0) + D(\mu_1^0 \mid \mu_1^0).$$

Fig. 3 illustrates this gap.

**Proof of Theorem 1:** We need to show that

$$\inf_{\mu \in \mathcal{A}} D(\mu \mid \mu_0^0) \geq \sqrt[4]{k-1} I(P_X, W_k)$$

where $\mathcal{A}$ contains all joint distributions $\mu$ such that

$$\mu_X = P_X, \quad \mu_Y = (\mu_0)_Y,$$

$$\sqrt[k-1]{E_{\mu_0} \log \frac{W_k}{(\mu_0)_Y}} \geq \sqrt[k-1]{E_{\mu_0} \log \frac{W_k}{(\mu_0)_Y}}. \quad (20)$$

We can assume w.l.o.g. that $W_0 \in C_1$. We then have

$$D(\mu \mid \mu_0^0) \overset{(A)}{=} D(\mu \mid \mu_0^0) \geq \sqrt[k-1]{E_{\mu} \log \frac{W_k}{(\mu)_Y}} \geq E_{\mu_0} \log \frac{W_k}{(\mu_0)_Y} = I(P_X, W_1) \geq \sqrt[k-1]{E_{\mu} \log \frac{W_k}{(\mu)_Y}}.$$ \hspace{1cm} (21)

where (A) uses (20), (B) uses the log-sum inequality

$$E_{\mu} \log \frac{W_k}{(\mu)_Y} = \frac{D(\mu \mid \mu_0^0) + E_{\mu_0} \log \frac{W_k}{(\mu)_Y}}{2} = D(\mu \mid \mu_0^0) - \frac{D(\mu \mid \mu_0^0) - D(\mu \mid \mu_0^0)}{0}.$$ \hspace{1cm} (22)

(C) is simply (21) and (D) follows from the one-sided property

$$E_{\mu_0} \log \frac{W_k}{(\mu_0)_Y} - E_{\mu_1} \log \frac{W_k}{(\mu_1)_Y} = D(\mu_0 \mid \mu_1^0) - D(\mu_0 \mid \mu_1) - D(\mu_1 \mid \mu_1^0) \geq 0.$$ \hspace{1cm} (23)

**APPENDIX**

**A. Local Geometric Analysis**

In this section, we introduce a geometric tool which helped us in achieving the results of this paper. This tool is based on a well-known approximation of the K-L divergence for measures which are “close” to each other, and is further developed here with a geometric emphasis. As most of the information theory problems involve optimizations of K-L divergences, often between distributions with high dimensionality, we believe that the localization method used in this paper can be a generic tool to simplify these problems. Focusing on certain special cases, this method is obviously useful in providing counterexamples to disprove conjectures. However, we also hope to convince the reader that the insights provided by the geometric analysis can be also valuable in solving the general problem. For example, our definition of one-sided sets and the use of log a posteriori distributions as decoding metrics in Section III can be seen as “naturally” suggested by the local analysis.

The divergence is not a distance between two distributions, however, if its two arguments are close enough, the divergence is approximately a squared norm, namely for any probability distribution $p$ on $\mathcal{Z}$ (where $\mathcal{Z}$ is any alphabet) and for any $v$ s.t. $\sum_{z} v(z)p(z) = 0$, we have

$$D(p(1 + \varepsilon v) \mid p) = \frac{1}{2} \varepsilon^2 \sum_{z \in \mathcal{Z}} v^2(z)p(z) + o(\varepsilon^2). \quad (24)$$

This is the main approximation used in this section. For convenience, we define

$$||v||_p^2 = \sum_{z \in \mathcal{Z}} v^2(z)p(z)$$

which is the squared L2-norm of $v$, with weight measure $p$. Similarly, we can define the weighted inner product

$$\langle u, v \rangle_p = \sum_{z \in \mathcal{Z}} u(z)v(z)p(z).$$

With these notations, one can write the approximation (24) as

$$D(p(1 + \varepsilon v) \mid p) = \frac{1}{2} ||v||_p^2 + o(\varepsilon^2).$$

Ignoring the higher order term, the above approximation can greatly simplify many optimization problems involving K-L divergences. In information theoretic problems dealing with discrete channels, such approximation is tight for some special cases such as when the channel is very noisy. In general, a very
noisy channel implies that the channel output weakly depends on the input. There are many works which deal very noisy channels, since these channels can be considered as abstraction of fading channels in a low SNR regime. However, our focus here is not on a specific SNR regime but rather on the geometric insight which is gained by working with very noisy channels. Let us introduce more precise notations. If the conditional probability of observing any output does not depend on the input (i.e., the transition probability matrix has constant columns), we have a “pure noise” channel. So a very noisy channel should be somehow close to such a pure noise channel. Formally, we consider the following family of channels:

$$W_\epsilon(b|a) = P_N(b)(1 + \epsilon L(a, b))$$

where $L$ satisfies for any $a \in \mathcal{X}$

$$\sum_{b \in \mathcal{Y}} L(a, b)P_N(b) = 0. \quad (23)$$

We say that $W_\epsilon$ is a very noisy channel if $\epsilon \ll 1$. In this case, the conditional distribution of the output, conditioned on any input symbol, is close to a distribution $P_N$ (on $\mathcal{Y}$), which is a pure noise distribution (constant columns). Each of these channels, $W_\epsilon(\cdot|\cdot)$, can be viewed as a perturbation from a pure noise channel $P_N$, along the direction specified by $L(\cdot, \cdot)$.

This way of defining very noisy channel can be found in [19], [16]. In fact, there are many other possible ways to describe a perturbation of distribution. For example, readers familiar with [3] might feel it natural to perturb distributions along exponential families. Since we are interested only in small perturbations, it is not hard to verify that these different definitions are indeed equivalent.

When an input distribution $P_X$ is chosen, the corresponding output distribution, over the very noisy channel, can be written as, $\forall b \in \mathcal{Y}$

$$P_{X|\epsilon}(b) = \sum_{a \in \mathcal{X}} P_X(a)W_\epsilon(b|a)$$

$$= P_N(b)\left(1 + \epsilon \sum_{a} P_X(a)L(a, b)\right)$$

$$= P_N(b)(1 + \epsilon \tilde{L}(b))$$

where $\tilde{L}(b) = \sum_{a} P_X(a)L(a, b)$, $\forall a \in \mathcal{X}$. Hence, a codeword which is sent and the received output have components which are i.i.d. from the following distribution:

$$P_X \circ W_\epsilon = P_X \circ P_N(1 + \epsilon L)$$

(we will drop the $\circ$ symbol in the subsequent) and similarly, the codeword which is not sent and the received output have components which are i.i.d. from the following distribution:

$$(P_X \circ W_\epsilon)^{\perp} = P_X \times P_N(1 + \epsilon \tilde{L})$$

(we will drop the $\times$ symbol in the subsequent). Therefore, the mutual information for very noisy channels is given by

$$I(P_X, W_\epsilon) = D(P_XP_N(1 + \epsilon L)\| P_XP_N(1 + \epsilon \tilde{L}))$$

$$= \frac{\epsilon^2}{2} \| \tilde{L} \|^2 + o(\epsilon^2)$$

where

$$\| \cdot \| = \| \cdot \|_{P_X \times P_N}$$

and

$$\tilde{L}(a, b) \triangleq L(a, b) - L(b)$$

which we call the centered direction.

### B. Very Noisy Mismatched Rates

As stated in Lemma 2, for an input distribution $P_X$, a mismatched linear decoder induced by the metric $d$, when the true channel is $W_0$, can achieve the following rate

$$\inf_{\mu \in \mathcal{A}} D(\mu\|\mu_0^d) \quad (24)$$

where

$$\mathcal{A} = \{ \mu : \mu_X = P_X, \mu_Y = (\mu_0)_Y, E_{\mu}[d] \geq E_{\mu_0}[d]\}.$$  

Now, if the channels are very noisy, this achievable rate can be expressed in the following simple form.

**Lemma 7:** Let $W_{0,\epsilon} = P_X(1 + \epsilon L_0)$ and $d_\epsilon = \log W_{1,\epsilon}$, where $W_{1,\epsilon} = P_N(1 + \epsilon L_1)$. For a given input distribution $P_X$, we can achieve the following rate:

$$\lim_{\epsilon \to 0} \frac{2}{\epsilon^2} \mathbb{E}(P_X, W_{0,\epsilon}, d_\epsilon) = \begin{cases} \langle \tilde{L}_0, \tilde{L}_A \rangle^2, & \text{when } \langle \tilde{L}_0, \tilde{L}_A \rangle \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

Note that there is no loss of generality in considering single-letter metrics which are the log of some channels (cf. [1]), however, we do make the restriction that all channels are around a common $P_N$ distribution.

Previous result says that the mismatched mutual information obtained when decoding with the linear decoder induced by the mismatched metric $\log W_{1,\epsilon}$, whereas the true channel is $W_{0,\epsilon}$, is approximately the projections’ squared norm of the true channel centered direction $\tilde{L}_0$ onto the mismatched centered direction $\tilde{L}_A$. This result gives an intuitive picture of the mismatched mutual information, as expected, if the decoder is matched, i.e., $\tilde{L}_0 = \tilde{L}_A$, the projections’ squared norm is $\| \tilde{L}_0 \|^2$, which is the very noisy mutual information of $\tilde{L}_0$. The more angle there is between $\tilde{L}_A$ and $\tilde{L}_0$, the more mismatched the decoder is, with a lowest achievable rate of 0 if the two directions are orthogonal or negatively correlated—this is for example the case if the two directions correspond to two BSCs with cross-over probabilities $1/2 + \epsilon$ and $1/2 - \epsilon$. Also note that (24) is particular case of I-project, as defined in [8] and [9], and the result of Lemma 7 is in agreement with the geometric properties developed in this reference, which shows analogies between the I-projection and projections in euclidean spaces.

**Proof:** For each $\epsilon$, the minimizer $\mu_\epsilon$ can be expressed as

$$\mu_\epsilon = P_XP_N(1 + \epsilon L)$$

where $L$ is a function on $\mathcal{X} \times \mathcal{Y}$, satisfying

$$\sum_{a \in \mathcal{X}, b \in \mathcal{Y}} P_X(a)P_N(b)L(a, b) = 0$$
and the two marginal constraints, respectively

\[
(\mu_\varepsilon)_X = P_X \quad \iff \quad \sum_{b \in \mathcal{Y}} P_N(b)L(a, b) = 0, \quad \forall a \in \mathcal{X} \tag{25}
\]

\[
(\mu_\varepsilon)_Y = (\mu_0)_Y \quad \iff \quad \sum_{a \in \mathcal{X}} P_X(a)L(a, b)
\quad = \sum_{a \in \mathcal{X}} P_X(a)L_0(a, b), \quad \forall b \in \mathcal{Y}. \tag{26}
\]

Now the constraint \( E_\varepsilon[\log W_{1, \varepsilon}] \geq E_\mu[\log W_{1, \varepsilon}] \) can be written as

\[
\sum_{a \in \mathcal{X}, b \in \mathcal{Y}} P_X(a)P_N(b)(1 + \varepsilon L(a, b)) \
\quad \cdot \left[ \log P_N + \log(1 + \varepsilon L_1(x, y)) \right]
\quad \geq \sum_{a \in \mathcal{X}, b \in \mathcal{Y}} P_X(a)P_N(b)(1 + \varepsilon L_0(a, b)) \
\quad \cdot \left[ \log P_N + \log(1 + \varepsilon L_1(x, y)) \right].
\]

Using a first-order Taylor expansion for the two log terms, and the marginal constraint (26), we have that previous constraint is equivalent to

\[
\langle L, L_1 \rangle \geq \langle L_0, L_1 \rangle + o(1) \tag{27}
\]

where

\[
\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_{P_X \times P_N}. \tag{28}
\]

Finally, we can write the objective function as

\[
D(\mu_\varepsilon \| \mu^0_\varepsilon) = D(P_XP_N(1 + \varepsilon L) \| P_XP_N(1 + \varepsilon L_0)) \
\quad = \frac{\varepsilon^2}{2} \| L - L_0 \|_{P_X \times P_N}^2 + o(\varepsilon^2).
\]

So we have transformed the original optimization problem into the very noisy setting

\[
\lim_{\varepsilon \to 0} \frac{2}{\varepsilon^2} \inf_{\mu \in \mathcal{X}} D(\mu \| \mu^0_\varepsilon) = \inf_{L: \langle L, L_1 \rangle \geq \langle L_0, L_1 \rangle} \| L - L_0 \|_{P_X \times P_N}^2 \tag{29}
\]

where the optimization on the right-hand side is over \( L \) satisfying the marginal constraints (25) and (26).

Now this optimization can be further simplified. By noticing that (26) implies \( \bar{L} = L_0 \), we have that \( L - L_0 = L - \bar{L} \), which we defined to be \( \bar{L} \). So \( \bar{L} \) satisfies both marginal constraints and the constraint in (29) becomes

\[
\langle L, L_1 \rangle \geq \langle L_0, L_1 \rangle \iff \langle \bar{L}, L_1 \rangle \geq \langle L_0, L_1 \rangle - \langle L_0, L_\bar{L} \rangle \
\quad \iff \langle \bar{L}, L_1 \rangle \geq \langle \bar{L}_0, L_1 \rangle.
\]

That is, both the objective and the constraint functions are now written in terms of centered directions, \( \bar{L} \). Hence, (29) becomes

\[
\inf_{L: \langle L, L_1 \rangle \geq \langle L_0, L_1 \rangle} \| \bar{L} \|_{P_X \times P_N}^2
\]

and we can simply recognize that, if \( \langle \bar{L}_0, L_1 \rangle \geq 0 \), the minimizer of this expression is obtained by the projection of \( \bar{L}_0 \) onto \( \bar{L}_1 \), with a minimum given by the projections’ squared norm

\[
\frac{\langle \bar{L}_0, L_1 \rangle^2}{\| L_1 \|_{P_X \times P_N}^2}
\]

otherwise, if \( \langle \bar{L}_0, L_1 \rangle < 0 \), the minimizer is \( \bar{L} = 0 \), leading to a zero rate.

\textbf{Remark:} We have just seen two examples where in the very noisy limit, information theoretic quantities have a natural geometric meaning, in the previously described inner product space. The cases treated in this section are the ones relevant for the paper’s problem, however, following similar expansions, other information theoretic problems, in particular multi-user ones (e.g., broadcast or interference channels) can also be treated in this geometrical setting. To simplify the notation, since the very noisy expansions scale with \( \varepsilon^2 \) and have a factor \( \frac{1}{2} \) in the limit, we denote by \( \text{VN} \) the following operator:

\[
T(\varepsilon) \overset{\text{VN}}{\longrightarrow} \lim_{\varepsilon \to 0} \frac{2}{\varepsilon^2} T(\varepsilon).
\]

We use the abbreviation VN for very noisy. Note that the main reason why we use the VN limit in this paper is similar somehow to the reason why we consider infinite block length in information theory: it gives us a simpler model to analyze and helps us understanding the more complex (not necessarily very noisy) general model. This makes the VN limit more than just an approximation for a specific regime of interest, it makes it an analysis tool of our problems, by setting them in a geometric framework where notion of distance and angles are this time well defined. Moreover, as we will show in Section E, in some cases, results proven in the VN limit can in fact be “lifted” to results proven in the general cases.

\section{C. Linear Decoding for Compound Channel: The Very Noisy Case}

In this section, we study a special case of compound channel, by assuming that the compound set contains very noisy channels. The local geometric analysis introduced in the previous section can be immediately applied and we show how the important concepts used in Section III emerge naturally in this setting.

\begin{itemize}
  \item All the channels are very noisy, with the same pure noise distribution. That is, all considered channels are of the form
  \[ W_\varepsilon(b | a) = P_N(b)(1 + \varepsilon L(a, b)), \quad \forall a \in \mathcal{X}, b \in \mathcal{Y} \]
  for the same \( P_N \), where \( L \) is any direction satisfying \( \sum_x P_N(b)L(a, b) = 0, \quad \forall a \). The compound set is hence depending on \( \varepsilon \), and is expressed as \( S_\varepsilon = \{ P_N(1 + \varepsilon L) \mid L \in \mathcal{S} \} \), where \( \mathcal{S} \) is the set of all possible directions. Hence, \( \mathcal{S} \) together with the pure noise distribution \( P_N \), completely determine the compound set for any \( \varepsilon \). We refer to \( \mathcal{S} \) as the compound set in the VN setting. Note that \( \mathcal{S} \) being convex, resp. compact, is the sufficient and necessary condition for \( S_\varepsilon \) to be convex, resp. compact, for all \( \varepsilon \).
\end{itemize}
• $P_X$ is fixed (it is the optimal input distribution) and we write

$$\mu_{\varepsilon} = P_X P_N(1 + \varepsilon L), \quad L \in \mathcal{S}$$

as the joint distribution of the input and output over a particular channel in $\mathcal{S}_\varepsilon$. For a given channel $W_{\varepsilon}$, the output distribution is $P_N(1 + \varepsilon L)$, where

$$\hat{L}(b) = \sum_{a \in \mathcal{X}} I(a, b) P_X(a), \quad \forall b \in \mathcal{Y}$$

and as before, $\hat{L} = L - \bar{L}$. We then denote $\check{\mathcal{S}} = \{\hat{L} : L \in \mathcal{S}\}$. Again, the convexity and compactness of $\mathcal{S}$ is equivalent to those of $\check{\mathcal{S}}$. The only difference is that $\mathcal{S}$ depends on the channels only, whereas $\check{\mathcal{S}}$ depends on the input distribution as well. As we fix $P_X$ in this section, we use the conditions $L \in \mathcal{S}$ and $\hat{L} \in \check{\mathcal{S}}$ interchangeably.

• As a convention, we often give an index, $j$, to the possible channels, and we naturally associate the channel index (the joint distribution index) and the direction index, i.e., $W_{j\varepsilon} = P_N(1 + \varepsilon L_j)$ and $\mu_{j\varepsilon} = P_X P_N(1 + \varepsilon L_j)$. In particular, we reserve $W_{0\varepsilon} = P_N(1 + \varepsilon L_0)$ for the true channel and use other indices, $L_1, L_2$, etc., for other specific channels.

• We write all inner products and norms as weighted by $P_X \times P_N$, and omit the subscript:

$$\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_{P_X \times P_N}$$

• Finally,

$$\min_{W \in \mathcal{E}_\varepsilon} I(P_X, W) = \frac{\varepsilon^2}{2} \min_{L \in \check{\mathcal{S}}} ||\hat{L}||^2 + o(\varepsilon^2)$$

and we define

$$L_\varepsilon = \arg \min_{L \in \check{\mathcal{S}}} ||\hat{L}||^2$$

to be the worst direction and $||\hat{L}_\varepsilon||^2$ is referred to as the very noisy compound channel capacity (on $\mathcal{S}$).

We conclude this section with the following lemma, which will be frequently used in the subsequent.

**Lemma 8:** Let $L_1, L_j, L_k$ and $L_d$ be four directions and assume that $\sum_a P_X(a) L_1(a, b) = \sum_a P_X(a) L_k(a, b)$ for any $b$. We then have

$$E_{\mu_{\varepsilon}} \log W_{j\varepsilon} > E_{\mu_{\varepsilon}} \log W_{k\varepsilon}$$

where

$$\sum_{a, b} P_X(a) L_j(a, b) - \frac{1}{2} ||L_j||^2 > \sum_{a, b} P_X(a) L_k(a, b) - \frac{1}{2} ||L_k||^2.$$

**Proof:** Using a second order Taylor expansion for $\log(1 + \varepsilon L_j)$, we have

$$E_{\mu_{\varepsilon}} \log W_{j\varepsilon}$$

$$= \sum_a P_X P_N(1 + \varepsilon L_0) \log(P_N(1 + \varepsilon L_j))$$

$$= \sum_a P_X P_N \log P_N$$

$$+ \varepsilon \sum_a P_X P_N L_j \log P_N + \varepsilon \sum_a P_X P_N L_i L_j$$

$$+ \varepsilon^2 \sum_a P_X P_N L_i L_j - \varepsilon^2 \frac{1}{2} \sum_a P_X P_N L_j^2,$$

$$= \sum_a P_X P_N(1 + \varepsilon L_d) \log(P_N(1 + \varepsilon L_j))$$

$$= \sum_a P_X P_N \log P_N$$

$$+ \varepsilon \sum_a P_X P_N L_d \log P_N + \varepsilon \sum_a P_X P_N L_i L_d$$

$$+ \varepsilon^2 \sum_a P_X P_N L_i L_d - \varepsilon^2 \frac{1}{2} \sum_a P_X P_N L_d^2,$$

which proves the lemma.

**D. One-Sided Sets**

We consider for now the use of linear decoder (i.e., induced by only one metric). We recall that, as proved in previous section, for $W_{0\varepsilon} = P_N(1 + \varepsilon L_0)$ and $d_{\varepsilon} = \log W_{1\varepsilon}$, where $W_{1\varepsilon} = P_N(1 + \varepsilon L_1)$, we have

$$\lim_{\varepsilon \to 0} \frac{2}{\varepsilon} R(P_X, W_{0\varepsilon}, d_{\varepsilon}) = \begin{cases} \frac{\langle L_0, L_1 \rangle^2}{||L_0||^2}, & \text{when } \langle L_0, L_1 \rangle \geq 0, \\ 0, & \text{otherwise}. \end{cases}$$

This picture of the mismatched mutual information directly suggests a first result. Assume $\mathcal{S}$, hence $\check{\mathcal{S}}$, to be convex. By using the worse channel to be the only decoding metric, it is then clear that the VN compound capacity can be achieved. In fact, no matter what the true channel $L_0 \in \mathcal{S}$ is, the mismatched mutual information given by the projection squared norm of $L_0$ onto $\hat{L}_\varepsilon$ cannot be shorter than $||\hat{L}_\varepsilon||^2$, which is the very noisy compound capacity of $\mathcal{S}$ (cf. Fig. 4). This agrees with the result proved in [12] for convex sets.

However, with this picture we understand that the notion of convexity is not necessary. As long as the compound set is such that its projection in the direction of the minimal vector stays on one side, i.e., if the compound set is entirely contained in the half space delimited by the normal plan to the minimal vector, i.e., if for any $L_0 \in \mathcal{S}$, we have $\langle L_0, L_\varepsilon \rangle \geq 0$ and

$$\frac{\langle L_0, L_\varepsilon \rangle^2}{||L_\varepsilon||^2} \geq ||L_\varepsilon||^2$$

we will have $E_{\mu_{\varepsilon}} \log W_{j\varepsilon}$ as a direction (i.e., if for any $L_0 \in \mathcal{S}$, we have $\langle L_0, L_\varepsilon \rangle \geq 0$ and $||L_\varepsilon||^2$ of the set $\check{\mathcal{S}}$, which is a consequence of the convexity of $\mathcal{S}$).

![Fig. 4. Very noisy one-sided compound set: in this figure, $\check{\mathcal{S}}$ is the union of three sets. The linear decoder induced by the worst channel metric $\log W_{j\varepsilon}$ when the true channel is $L_0$ affords reliable communication for rates as large as the squared norm of the projection of $L_0$ onto $\hat{L}_\varepsilon$. From the one-sided shape of the compound set, this projection squared norm is always as large as the compound capacity given by the squared norm of $L_\varepsilon$.](image-url)
we will achieve compound capacity by using the linear decoder
induced by the worst channel metric (cf. Fig. 4 where $S$ is not
cvx but still verifies the above conditions). We call such sets
one-sided sets, as defined in the following.

Definition 4 (VN One-Sided Set): A VN compound set $S$ is
one-sided iff for any $L_0 \in S$, we have

$$\langle \tilde{L}_0, \tilde{L}_S \rangle \geq 0$$

(31)

$$\frac{\langle \tilde{L}_0, \tilde{L}_S \rangle^2}{||\tilde{L}_S||^2} \geq \frac{||\tilde{L}_S||^2}{2}.$$ 

(32)

Equivalently, a VN compound set $S$ is one-sided iff for any $L_0 \in S$,
we have

$$||\tilde{L}_0||^2 - ||\tilde{L}_S||^2 - ||\tilde{L}_0 - \tilde{L}_S||^2 \geq 0.$$ 

(33)

Lemma 9: In the VN setting, the linear decoder induced by
the worst channel metric $\log L_S$ is capacity achieving for one-
sided sets.

The very noisy picture also suggests that the one-sided prop-
erty is indeed necessary in order to be able to achieve the com-
ound capacity with a single linear decoder (cf. [1]). One can
then consider the problem of finding the “best metric” which
would lead to highest achievable rates, or which would mini-
mize an outage probability. This kind of approach is for example
considered in [22], in a setting where the channel is first esti-
imated and then the best linear decoder minimizing an outage
probability is looked for. However, our main goal here is not
motivated by results of this kind; we investigate instead whether
we can still achieve compound capacity on non one-sided com-
ound sets, by using generalized linear decoders.

E. Lifting Local to Global Results

In this section, we illustrate how the results and proofs ob-
tained in Section C in the very noisy setting can be lifted to
results and proofs in the general setting. We first consider the
case of one-sided sets. By revisiting the definitions made in
Section D, we will try to develop a “naturally” corresponding
notion of one-sidedness for the general problems.

By definition of a VN one-sided set, $S$ is such that

$$||\tilde{L}_0||^2 - ||\tilde{L}_S||^2 - ||\tilde{L}_0 - \tilde{L}_S||^2 \geq 0, \forall L_0 \in S.$$ 

(34)

Next, we find the divergences, for the general problems, whose
very noisy representations are these norms: recall that

$$D(\mu_0 || \mu_S) = \frac{\sum_{n=0}^{\infty} \langle L_0 - L_0 \rangle^2}{||L_0||^2}$$

(35)

$$D(\mu_S || \mu_S) = \frac{\sum_{n=0}^{\infty} \langle L_0 - L_S \rangle^2}{||L_S||^2}.$$ 

(36)

On the other hand, we also have

$$D(\mu_0 || \mu_S) = \sum_{n=0}^{\infty} ||L_0 - L_S||^2$$

and

$$D(\mu_S || \mu_S) = \sum_{n=0}^{\infty} ||L_0 - L_S||^2.$$ 

and hence

$$D(\mu_0 || \mu_S) = D(\mu_0 || \mu_S)$$

$$\sum_{n=0}^{\infty} ||L_0 - L_S||^2 - ||L_0 - L_S||^2 = ||L_0 - L_S||^2$$ 

(37)

where the last equality simply uses the projection principle, i.e.,
that the projection of $L_0$ onto the centered directions given by
$\tilde{L} = L - \tilde{L}$, is orthogonal to the projection’s height $\tilde{L}$, implying

$$||\tilde{L}||^2 = ||L||^2 - ||\tilde{L}||^2.$$ 

Now, by reversing the very noisy approximation in (35), (36),
and (37), we get that

$$D(\mu_0 || \mu_S) = D(\mu_S || \mu_S)$$

$$- D(\mu_0 || \mu_S) - D(\mu_0 || \mu_S) \geq 0$$

for all $W_0 \in S$, can be viewed as a “natural” counterpart of (33),
hence of the VN one-sided definition. Note that this inequality
is equivalent to

$$D(\mu_0 || \mu_S) \geq D(\mu_0 || \mu_S) + D(\mu_S || \mu_S), \forall W_0 \in S.$$ 

(38)

Clearly, as we mechanically generalized the notion of one-
sided sets from a special very noisy case to the general problem,
there is no reason to believe at this point that the resulting one-
sided sets will have the same property in the general setting, than
their counterparts in the very noisy case; namely, that the linear
decoder induced from the worst channel achieves the compound
capacity. However, this turns out to be true, and the proof again
follows closely the corresponding proof of the very noisy special
case.

Proof of Lemma 9: Recall that in the VN case, when the ac-
tual channel is $W_0 \epsilon$, and the decoder uses metric $d_k = \log W_L \epsilon$,
the achievable rate, in terms of the corresponding centered di-
rections $\tilde{L}_0, \tilde{L}_1$, is given by, cf. (29)

$$\lim_{\epsilon \rightarrow 0} \max_{\epsilon \geq 0} \min_{\epsilon \geq 0} \langle X, \epsilon \rangle = \inf \langle \tilde{L}, \tilde{L}_0 \rangle_{(\tilde{L}_0, \tilde{L}_1)} \frac{||\tilde{L}||^2}{2}.$$ 

(39)

The constraint of the optimization can be rewritten in norms as

$$\tilde{L} : \frac{||\tilde{L}||^2}{2} - ||\tilde{L} - \tilde{L}_0||^2 \geq \frac{||\tilde{L}_0||^2}{2} - ||\tilde{L}_0 - \tilde{L}_1||^2.$$ 

(40)

Now if $\tilde{L}_0$ lies in a one-sided set $S$, and we use decoding metric
as the worst channel $\tilde{L}_1 = \tilde{L}_S$, by using definition (34), and rec-
ognizing that $||\tilde{L} - \tilde{L}_0||^2$ is non-negative, this constraint implies

$$||\tilde{L}||^2 \geq ||\tilde{L}_0||^2 - ||\tilde{L}_0 - \tilde{L}_S||^2 \geq ||\tilde{L}_S||^2, \forall \tilde{L}_0 \in S$$ 

(41)

form which we conclude that the compound capacity is achiev-
able. The proof of Lemma 5 replicates these steps closely.

First, we write in the general setting, the mismatched mutual
information is given by

$$R(\mu, W_0 \log W) = \inf \mu \in \mathcal{A}_c D(\mu || \mu^c)$$ 

(42)
where
\[ A_S = \{ \mu_X = P_X, \mu_Y = (\mu_0)_Y, E_\mu \log W_S \geq E_{\mu_0} \log W_S \} \]

Since we consider here a linear decoder, i.e., induced by only one single-letter metric, we can consider equivalently the ML or MAP metrics. We then work with the MAP metric and the constraint set is equivalently expressed as
\[ A_S = \left\{ \mu_X = P_X, \mu_Y = (\mu_0)_Y, E_\mu \log \frac{W_S}{(\mu_s)_Y} \geq E_{\mu_0} \frac{W_S}{(\mu_s)_Y} \right\} \]

Expressing the quantities of interest in terms of divergences, we write
\[ E_\mu \left[ \frac{W_S}{(\mu_s)_Y} \right] = E_\mu \left[ \log \frac{W_S}{(\mu_s)_Y} \frac{\mu}{\mu_0} \right] = D(\mu || \mu_0) - D(\mu || \mu_0) + D(\mu || \mu_0) \cdot \]

Similarly, we have
\[ E_{\mu_0} \frac{W_S}{(\mu_s)_Y} = D(\mu_0 || \mu_0) - D(\mu_0 || \mu_0) + D(\mu_0 || \mu_0) \cdot \]

Thus, we can rewrite \( A_S \) as
\[ A_S = \{ \mu : \mu_X = P_X, \mu_Y = (\mu_0)_Y \times D(\mu || \mu_0) = D(\mu || \mu_0) + D(\mu || \mu_0) \}, \]
\[ (43) \]

It worth noticing that this is precisely the lifting of (40).

Now, in the VN limit, \( D(\mu || \mu_0) - D(\mu_0 || \mu_0) \) is given by
\[ ||L - L_0||^2 - ||L - L_0||^2 = ||L - L_0||^2, \]
which is clearly positive. Here, we have that
\[ D(\mu || \mu_0) - D(\mu || \mu_0) \geq 0 \]
is a direct consequence of log-sum inequality, and with this, we can write for all \( \mu \in A_S \)
\[ D(\mu || \mu_0) = D(\mu || \mu_0) + D(\mu || \mu_0) \]
which is in turn lower bounded by \( D(\mu_0 || \mu_0^2) = I(P_X, W_S) \), provided that the set \( S \) is one-sided, cf. (3) (note that last lines are again a lifting of (41)). Thus, the compound capacity is achieved.

This general proof can indeed be shortened. Here, we emphasize the correspondence with the proof for the very noisy case, in order to demonstrate the insights one obtains by using the local geometric analysis.

**F. Non Optimality of GLRT With One-Sided Components**

Assume
\[ S = S_1 \cup S_2 \]

where \( S_1 \) and \( S_2 \) are one-sided; in this section we consider only the VN setting, hence saying that \( S_1 \) is one sided really means that the VN compound set \( S_1 \) corresponding to \( S_{1,e} \) is one-sided according to Definition 4.

For a fixed input distribution \( P_X \), let \( W_1 = W_{S_1} \) and \( W_2 = W_{S_2} \) be the worst channel of \( S_1, S_2 \), respectively. A plausible candidate for a generalized linear universal decoder is the GLRT with metrics \( d_1 = \log W_1 \) and \( d_2 = \log W_2 \), hoping that a combination of earlier results for finite and one-sided sets would make this decoder capacity achieving. Say w.l.o.g. that \( W_0 \in S_1 \). Using (6), the following rate can be achieved with the proposed decoding rule:
\[ R(P_X, W_0, \{d_k\}_{k=1}^K) = R_1 \land R_2 \]

where
\[ R_k = \inf_{A_k} D(\mu || \mu_0^p), \quad k = 1, 2 \]

and for \( k = 1, 2 \),
\[ A_k = \{ \mu : \mu_1 = \mu_0, E_\mu \log W_k \geq \sqrt{2} \sqrt[\gamma]{E_{\mu_0} \log W_1} \}, \]
\[ (44) \]

Note that we are using similar notations for this section as for the previous one, although the sets \( A_k \) and rates \( R_k \) are now given by different expressions. We also use \( \mu^p = \mu_0^p \) to express in a more compact way that the marginals of \( \mu \) and \( \mu_0 \) are the same.

Since \( W_1 \) and \( W_2 \) are the worst channel for \( P_X \) in each component, the compound capacity over \( S = S_1 \cup S_2 \) is
\[ C(S) = I(P_X, W_1) \land I(P_X, W_2), \]

Note that the maximum in \( \sqrt{2} \sqrt[\gamma]{E_{\mu_0} \log W_1} \) cannot be identified in general and we need to consider two cases

**Case 1:** \[ E_{\mu_0} \log W_1 \geq E_{\mu_0} \log W_2 \]

**Case 2:** \[ E_{\mu_0} \log W_1 \leq E_{\mu_0} \log W_2 \]

In order to verify that the decoder is capacity achieving, we need to check if both \( R_1 \) and \( R_2 \) are greater than or equal to the compound capacity \( C(S) \), no matter which of case 1 or case 2 occurs. Note that the expression of \( R_2 \) in case 2 is

For case 2: \[ R_2 = R(P_X, W_0, \log W_2), \]

and the inequality we need to check in the very noisy case is

For case 2: \[ R_{2,e} \geq C(S_e) \sqrt[\gamma]{\frac{(\log p_0, L_2)^2}{||L_2||^2}} \]
\[ \geq ||L_1||^2 \land ||L_2||^2, \]
\[ (47) \]

However, the one-sided property does not apply anymore here, since we assumed that \( L_0 \) belongs to \( S_1 \) and not \( S_2 \). Indeed, if we have no restriction on the positions of \( L_0 \) and \( L_2 \), (47) can be zero. We next present an example where previous inequalities are not satisfied.
Lemma 10: In the VN setting and for compound sets having a finite number of one-sided components, GLRT with the worst channel of each component is not capacity achieving.

Counterexample: Let $X = Y = \{0, 1\}$, $P_X = P_N = \{1/2, 1/2\}$

$$L_0 = \begin{pmatrix} -2 & 2 \\ -7 & 7 \end{pmatrix}, \quad L_1 = \begin{pmatrix} 2 & -2 \\ 0 & 0 \end{pmatrix}, \quad L_2 = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}.$$  

Note that the fact that we use binary alphabets is important for the conclusion, since in this case the mismatched mutual information rates used below are the largest achievable rates in the mismatched scenario, as discussed in the introduction and proved in [4]. The achievable rate can be easily checked with this counterexample, and in fact there are many other examples that one can construct. We will, in the following, discuss the geometric insights that leads to these counterexamples (and check that it is indeed a counterexample).

We first use Lemma 8 to write

$$E_{\mu_0, \varepsilon} \log W_{1, \varepsilon} \leq \frac{\sum_{\varepsilon} \|L_0 - L_2\|}{\|L_0 - L_1\|} \leq \|L_0 - L_2\|$$

which can be use to rewrite (45) and (46) in the very noisy setting as

$$\text{Case 1:} \quad \|L_0 - L_2\| \geq \|L_0 - L_1\| \quad (48)$$
$$\text{Case 2:} \quad \|L_0 - L_2\| \leq \|L_0 - L_1\|. \quad (49)$$

Now to construct a counterexample, we consider the special case where $\|L_0 - L_2\| = \|L_0 - L_1\|$ and $\|L_1\| = \|L_2\|$. These assumptions are used to simplify our discussion, and are not necessary in constructing counterexamples. One can check that the above example satisfies both assumptions. Now (47) holds if and only if

$$\frac{\langle \tilde{L}_0, \tilde{L}_2 \rangle}{\|\tilde{L}_2\|} \geq \|\tilde{L}_2\|$$

which is equivalent to

$$\|\tilde{L}_0\|^2 - \|\tilde{L}_2\|^2 - \|\tilde{L}_0 - \tilde{L}_2\|^2 \geq 0.$$  

It is easy to check that the last inequality does not hold for the given counterexample, which completes the proof of Lemma 10.

In fact, one can write

$$\begin{align*}
\|\tilde{L}_0\|^2 - \|\tilde{L}_2\|^2 - \|\tilde{L}_0 - \tilde{L}_2\|^2 &= \|\tilde{L}_0\|^2 - \|\tilde{L}_1\|^2 - \|\tilde{L}_0 - \tilde{L}_1\|^2 \\
&\quad + \|\tilde{L}_0 - \tilde{L}_1\|^2 - \|\tilde{L}_0 - \tilde{L}_2\|^2. \quad (50)
\end{align*}$$

The term on the second line above is always positive (by the one-sided property), but we have a problem with the term on the last line: we assumed that $\|L_0 - L_2\| = \|L_0 - L_1\|$, and this does not imply that $\|\tilde{L}_0 - \tilde{L}_1\|^2 = \|\tilde{L}_0 - \tilde{L}_2\|^2$. The problem here is that when using log likelihood functions as decoding metrics, the constraints in (44), (45) and (46) are, in the very noisy case, given in terms of the perturbation directions $L_i$, $i = 0, 1, 2$, while the desired statement about achievable rates and the compound capacity are given in terms of the centered directions $\tilde{L}_i$. Thus, counterexamples can be constructed by carefully assign $\tilde{L}_i$’s to be different, hence the constraints on $L_i$’s cannot effectively regulate the behavior of $\tilde{L}_i$’s ([50] can be made negative). Fig. 5 gives a pictorial illustration of this phenomenon.

The above discussion also suggests a fix to the problem. If one could replace the constraints on $L_i$’s in (44), (45) and (46), by the corresponding constraints on $\tilde{L}_i$’s, that might at least allow better controls over the achievable rates. This is indeed possible by making a small change of the decoding metrics, and this precisely done by using the MAP metrics instead of the ML metrics, leading to our main result.

G. Difference Between GMAP and GLRT

For a linear decoder with a single metric $d : X \times Y \mapsto \mathbb{R}$, if one forms a different test by picking $d' \in \mathbb{R} \mapsto \mathbb{R}$, it is not hard to see that the resulting decision is exactly the same, for every possible received signal $y$. This is why the ML decoder and the MAP decoder, from the same mismatched channel $W_{1,1}$, are indeed equivalent, as they differ by a factor of $f = \log(P_X \circ W_{1,1})$. For a generalized linear decoder with multiple metrics, $d_1, d_2, \ldots, d_K$, if one changes the metrics to $d_1 + f, d_2 + f, \ldots, d_K + f$, for the same function $f$ on $Y$, again the resulting decoder is the same. Things are different, however, if one changes these metrics by different functions, to have $d_1 + f_1, d_2 + f_2, \ldots, d_K + f_K$. The problem is that this changes the balance between the metrics, which as we observed in the GMAP setting, is critical for the generalized linear decoder to work properly. For example, if one adds a big number on one of the metrics to make it always dominate the others, the purpose of using multiple metrics is defeated. GLRT differs from the GMAP decoder by factors of $\log(\mu_k)Y$ on the $k$th metric, which causes a bias depending the received signal $y$. The counterexample we presented in the previous section is in essence constructed to illustrate the effect of such bias. Through a similar approach, one can indeed show that the GMAP receiver is the unique generalized linear receiver, based on the worst channels of different one-sided components, in the sense...
that any non-trivial variation of these metrics, i.e., $f_1, \ldots, f_K$, which are not the same function, would result in a receiver that does not achieve the compound capacity in all cases. Counterexamples can always be constructed in a similar fashion. Further details on the very noisy geometry and on the GMAP decoding can be found in [1].

ACKNOWLEDGMENT

The authors wish to thank E. Telatar, for helpful comments and stimulating discussions.

REFERENCES

[1] E. Abbe, “Local to Global Geometric Methods in Information Theory,” Ph.D. dissertation, MIT, Cambridge, MA, 2008.
[2] E. Abbe and R. Pulikkoonattu, “Universal a posteriori metrics game,” in Information Theory Workshop (ITW), Dublin, Ireland, Aug. 2010.
[3] S. Amari and H. Nagaoka, Methods of Information Geometry. Providence, RI: American Mathematical Society, 2001.
[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, no. 6, pp. 1889–1902, Nov. 1995.
[5] E. Biglieri, J. Proakis, and S. Shamai, “Fading channels: Information-theoretic and communications aspects,” IEEE Trans. Inform. Theory, vol. 44, no. 10, pp. 2619–2692, Oct. 1998.
[6] D. Blackwell, L. Breiman, and A. Thomasian, “The capacity of a class of channels,” Ann. Mathemat. Stat., vol. 30, no. 4, pp. 1229–1241, Dec. 1959.
[7] G. Cozzo and F. Fagnani, “The capacity of abelian group codes over symmetric channels,” IEEE Trans. Inform. Theory, vol. 55, no. 5, pp. 2037–2054, May 2009.
[8] I. Csiszar, “Sanov property, generalized i-projection and a conditional limit theorem,” Ann. Probabil., vol. 12, pp. 768–793, 1984.
[9] I. Csiszar, “Information projections revisited,” IEEE Trans. Inform. Theory, vol. 49, no. 6, pp. 1474–1490, June 2003.
[10] I. Csiszar, “Information theory and statistics: A tutorial,” Found. and Trends in Commun. and Inform. Theory, vol. 1, no. 4, 2004.
[11] I. Csiszar and J. Korner, Information Theory: Coding Theorem for Discrete Memoryless Systems. Budapest, Hungary: Akademiai Kiado, 1986.
[12] I. Csiszar and P. Narayan, “Channel capacity for a given decoding metric,” IEEE Trans. Inform. Theory, vol. 41, no. 1, pp. 35–43, Jan. 1995.
[13] R. L. Dobrushin, “Optimum information transmission through a channel with unknown parameters,” Radio Eng. Electron., vol. 4, no. 12, 1959.
[14] P. Elias, “Coding for two noisy channels,” in Proc. 3rd London Symp. Information Theory, London, U.K., 1955, pp. 61–76.
[15] M. Feder and A. Lapidoth, “Universal decoding for channels with memory,” IEEE Trans. Inform. Theory, vol. 44, no. 9, pp. 1726–1745, Sep. 1998.
[16] R. G. Gallager, Information Theory and Reliable Communication. New York: Wiley, 1968.
[17] A. Lapidoth and P. Narayan, “Reliable communication under channel uncertainty,” IEEE Trans. Inform. Theory, vol. 44, no. 10, pp. 2148–2177, Oct. 1998.
[18] A. Lapidoth and S. Shamai, “Fading channels: How perfect need perfect side information be?” IEEE Trans. Inform. Theory, vol. 48, no. 5, pp. 1118–1134, May 2002.
[19] A. Lapidoth, E. Telatar, and R. Urbanke, “On wide-band broadcast channels,” IEEE Trans. Inform. Theory, vol. 49, no. 12, pp. 3250–3258, Dec. 2003.
[20] A. Lapidoth and J. Ziv, “On the universality of the LZ-based decoding algorithm,” IEEE Trans. Inform. Theory, vol. 44, no. 9, pp. 1746–1755, Sep. 1998.
[21] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai, “On information rates for mismatched decoders,” IEEE Trans. Inform. Theory, vol. 40, no. 6, pp. 1953–1967, Nov. 1994.
[22] F. Piantanida, S. Sadough, and P. Duhamel, “One the outage capacity of a practical decoder using channel estimation accuracy,” in IEEE Int. Symp. Information Theory (ISIT), Jun. 2007.
[23] O. Shayevitz and M. Feder, “Universal decoding for frequency-selective fading channels,” IEEE Trans. Inform. Theory, vol. 51, no. 8, pp. 2770–2790, Aug. 2005.
[24] J. Wolfowitz, “Simultaneous channels,” Arch. Rat. Mech. Analysis, vol. 4, no. 4, pp. 371–386, 1960.

Emmanuel Abbe received the M.S. degree in 2003 from the Mathematics Department, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, and the Ph.D. degree in 2008 from the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge. Since 2008, he has been a Postdoctoral Fellow and Lecturer in the Department of Communication and Computer Sciences, École Polytechnique Fédérale de Lausanne, and a Research Affiliate at Massachusetts Institute of Technology (MIT), Cambridge. His research interests include information theory and coding, wireless communications and graphical models.

Dr. Abbe received the CVCI Award for his Master’s thesis in Mathematics from the École Polytechnique Fédérale de Lausanne in 2003.

Lizhong Zheng received the B.S. and M.S. degrees from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 1994 and 1997, respectively, and the Ph.D. degree from the Department of Electrical Engineering and Computer Sciences, University of California (UC) Berkeley, in 2002. Since 2002, he has been with the Department of Electrical Engineering and Computer Sciences, Massachusetts Institute of Technology (MIT), Cambridge, where he is currently an Associate Professor. His research interests include information theory, wireless communications, and wireless networks.

Dr. Zheng received the Eli Jury Award from UC Berkeley in 2002, the IEEE Information Theory Society Paper Award in 2003, the National Science Foundation (NSF) CAREER award in 2004, and the AFOSR Young Investigator Award in 2007. He is currently an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION THEORY.