Upper Bounds on the Mismatched Reliability Function and Capacity Using a Genie Receiver

Anelia Somekh-Baruch®, Member, IEEE

Abstract—We develop a novel framework for proving converse theorems for channel coding, which is based on the analysis technique of multicast transmission with an additional auxiliary receiver, which serves as a genie to the original receiver. The genie provides the original receiver a certain narrowed list of codewords to choose from that includes the transmitted one. This technique is used to derive upper bounds on the mismatch capacity of discrete memoryless channels as well as the reliability function with a mismatched decoding metric. Unlike previous works, our bounding technique exploits also the inherent symmetric requirement from the codewords, leading to these new upper bounds. Since the computations of most of the known bounds on the mismatch capacity are rather complicated, we further present a method to obtain relaxed bounds that are easier to compute. As an example, we analyze the obtained bounds in the binary-input channels case. We conclude by presenting simpler bounds on the reliability function, and provide sufficient conditions for their tightness in certain ranges of rates.

Index Terms—Mismatched decoding, mismatch capacity, channel coding, reliability function, error exponents.

I. INTRODUCTION

This paper addresses the problem of determining the fundamental limits of reliable communication over a discrete memoryless channel $W$ with a given decoding metric $q$. This setup is usually referred to as mismatched decoding, since the decoding metric differs from the optimal maximum likelihood (ML) metric that is matched to the communication channel. While the ML decoder minimizes the error probability, in certain cases it is not applicable for various reasons such as channel estimation errors or for practical decoder implementation considerations. The highest achievable rate with decoding metric $q$ is referred to as the mismatch capacity. The problem of determining the fundamental limits of channels with mismatched decoding is also related to other information-theoretic setups such as zero-error transmission over communication channels.

There have been quite a few works on achievable rates for channels with mismatched decoding from the information-theoretic view point. A partial list of works is [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], and [15], and for a survey on the subject see [16], and references therein.

Early results with a converse flavor were derived in [17], where a necessary and sufficient condition for the positivity of the mismatch capacity was determined, as well as a single-letter expression for the mismatch capacity in the case of the binary input binary output channel.

Later works [18], [19] presented among other results, a tight multi-letter soft converse result, and initiated multi-letter bounds, which involve a max-min form, where the maximum is taken over input distributions, and a minimum is taken over a set of auxiliary channels that satisfy certain constraints (see [19, Theorem 4] and [18, Theorems 4-5]).

The work [20] proved a max-min upper bound having a single-letter form. The bound of [20] was obtained by constructing a graph in the output space of the channel and using graph-theoretic and large deviations tools.

Tighter max-min single-letter upper bounds on the mismatch capacity were derived in [21] (see also [22]), using a proof technique based on multicast transmission of only one common message over a broadcast channel. This technique relies on extending the single-user channel $W$ with output $Y$ to a channel that has an additional output $Z$, with the property that the intersection event of correct $q$-decoding of the $Y$-receiver and erroneous decoding of the auxiliary $Z$-receiver has zero probability for any codebook of a certain composition $P$. This approach led to a strictly tighter bound with a significantly simpler proof, which holds also for continuous alphabet channels. A tighter bound of a more involved form was presented in [21], that relied on considering a $Z$-receiver which is genie-aided and is informed of the actual joint empirical distribution of the transmitted codeword and the $Z$-output sequence. Equivalence classes of isomorphic channel-metric pairs $(W,q)$ were further introduced in [21], that enabled to derive a sufficient condition for the tightness of the bound.

In a later work, [23], the class of broadcast channels was enlarged to include channels satisfying that if the $Z$-receiver makes an error, then with high probability (approaching 1 as the block length tends to infinity) so does the $Y$-receiver. An improved bound was derived in [24] and [25], which

Manuscript received 16 March 2022; revised 31 August 2022; accepted 27 October 2022. Date of publication 15 March 2023; date of current version 16 June 2023. This work was supported by the Israel Science Foundation (ISF) under Grant 631/17. An earlier version of this paper was presented in part at the Information Theory Workshop (ITW) 2022 and in part at the Oberwolfach 2022 Workshop titled “New Mathematical Techniques in Information Theory [DOI: 10.1109/ITW54588.2022.9965803].”

The author is with the Faculty of Engineering, Bar-Ilan University, Ramat Gan 5290002, Israel (e-mail: somekha@biu.ac.il).

Communicated by L. Wang, Associate Editor for Shannon Theory and Information Measures.

Digital Object Identifier 10.1109/TIT.2023.3257189

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
further enlarged the class of channels. The above mentioned bounds are described in detail in this paper, but most of them are quite complicated to compute in the sense that it is required to solve an optimization problem in order to determine whether a certain channel belongs to the set or not. The tightest bound known to date which is easily computable in this sense is the basic bound of [21].

The study of the reliability function (error exponents) of channels with ML decoding has been quite extensive (see, e.g., [26], [27], [28], [29]). Clearly, the known upper bounds are applicable also to mismatched decoding. Lower bounds on the exponents with mismatched decoding were derived in several works such as [3], [5], and [14] (see also [16], and references therein). Recently, an upper bound on the reliability function with mismatched decoding for zero-rate codes was derived in [30]. For a wide class of channel-metric pairs, this bound was shown to be tight at $R = 0^+$. In [24], an upper bound was derived for all rates up to the aforementioned bound of [24] on the mismatch capacity.

In this paper, we refine our multicast approach to allow the genie-aided auxiliary $Z$-receiver of the channel to serve as a genie for the original $Y$-receiver. We call this approach “transmission with a genie-aided-genie”. The idea is very simple: the genie-aided auxiliary $Z$-receiver informs the original $Y$-receiver of the list of all the codewords that share the same empirical statistics (joint type-class) with the channel $Z$-output as that of the actual transmitted codeword. Doing so, it narrows down the list of competing hypothesized codewords that the original mismatched decoder needs to choose from, and yields a lower bound on the probability of error. This leads to a basic upper bound on the mismatch capacity and the reliability function. We further present possibly looser bounds which are easily computable.

This paper is organized as follows: In Section II we present notation conventions. A formal statement of the mismatched decoding setup appears in Section III. In Section IV-A, we present the transmission with a Genie-Aided-Genie proof technique. Section IV-B summarizes our main results. In Section V, we study the case of binary-input DMCs. In Section VI we compare our results to former results. Sections VII and VIII are dedicated to the proofs of the main theorems regarding reliability function and mismatch capacity theorems, respectively. In Section IX, we establish some simpler bounds on the reliability function and sufficient conditions for tightness. Section X discusses some concluding remarks. Proofs of additional results and lemmas appear in the appendix.

II. NOTATION

Throughout this paper, scalar random variables (RVs) are denoted by capital letters, their sample values are denoted by their respective lower case letters, and their alphabets are denoted by their respective calligraphic letters; e.g. $X$, $x$, and $\mathcal{X}$, respectively. A similar convention applies to random vectors of dimension $n$ and their sample values, which are denoted in boldface; e.g., $\mathbf{x}$. The set of all $n$-vectors with components taking values in a certain finite alphabet are denoted by the same alphabet superscripted by $n$, e.g., $\mathcal{X}^n$.

Logarithms are taken to the natural base $e$, unless stated otherwise.

For a given sequence $x \in \mathcal{X}^n$, where $\mathcal{X}$ is a finite alphabet, $\hat{P}_x$ denotes the empirical distribution on $\mathcal{X}$ extracted from $x$; in other words, $\hat{P}_x$ is the vector $\{\hat{P}_x(x), x \in \mathcal{X}\}$, where $\hat{P}_x(x)$ is the relative frequency of the symbol $x$ in the vector $x$. The type-class of $x$ is the set of $x' \in \mathcal{X}^n$ such that $\hat{P}_x = \hat{P}_{x'}$, which is denoted $T_x(\hat{P}_x)$. Similarly, the joint empirical distribution of two sequences $x, y$, denoted $\hat{P}_{xy}$, is the vector $\{\hat{P}_{xy}(x, y), (x, y) \in \mathcal{X} \times \mathcal{Y}\}$, where $\hat{P}_{xy}(x, y)$ is the relative frequency of the pair of symbols $(x, y)$ in the vector $(x, y)$; i.e. the number of indices $i$ such that $(x_i, y_i) = (x, y)$ normalized by $n$. The conditional type-class of $y$ given $x$ is the set of $y$’s such that $\hat{P}_{xy} = \hat{P}_{x}y$, which is denoted $T_{x}(\hat{P}_{xy}|x)$. The set of all probability distributions on $\mathcal{X}$ is denoted by $\mathcal{P}(\mathcal{X})$, the set of conditional distributions from $\mathcal{X}$ to $\mathcal{Y}$ is denoted $\mathcal{P}(\mathcal{Y}|\mathcal{X})$, and the set of empirical distributions of order $n$ on alphabet $\mathcal{X}$ is denoted $\mathcal{P}_n(\mathcal{X})$.

Information theoretic quantities, such as entropy, conditional entropy, and mutual information are denoted as follows. Throughout this paper, scalar random variables (RVs) are denoted by calligraphic letters; e.g. $X$, $Y$, $Z$. A similar convention applies to random vectors of dimension $n$, respectively. A similar convention applies to random vectors of dimension $n$ and their sample values, which are denoted in boldface; e.g., $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$. An encoder maps a message $\mathbf{W}$ to an output sequence $\mathbf{Y}$. The dependence of a quantity on a certain underlying distribution is denoted by a bar; e.g., $P(\mathbf{X}|\mathbf{Y})$. The conditional type-class of $\mathbf{Y}$ given $\mathbf{X}$ is denoted $\mathcal{T}_\mathbf{x}(\mathcal{P}_\mathbf{y}|\mathbf{x})$. The set of all probability distributions on $\mathcal{X}$ is denoted by $\mathcal{P}(\mathcal{X})$, the set of empirical distributions of order $n$ on alphabet $\mathcal{X}$ is denoted $\mathcal{P}_n(\mathcal{X})$.

Information theoretic quantities, such as entropy, conditional entropy, and mutual information are denoted as follows. Throughout this paper, scalar random variables (RVs) are denoted by calligraphic letters; e.g. $X$, $Y$, $Z$. A similar convention applies to random vectors of dimension $n$, respectively. A similar convention applies to random vectors of dimension $n$ and their sample values, which are denoted in boldface; e.g., $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$. An encoder maps a message $\mathbf{W}$ to an output sequence $\mathbf{Y}$. The dependence of a quantity on a certain underlying distribution is denoted by a bar; e.g., $P(\mathbf{X}|\mathbf{Y})$. The conditional type-class of $\mathbf{Y}$ given $\mathbf{X}$ is denoted $\mathcal{T}_\mathbf{x}(\mathcal{P}_\mathbf{y}|\mathbf{x})$. The set of all probability distributions on $\mathcal{X}$ is denoted by $\mathcal{P}(\mathcal{X})$, the set of empirical distributions of order $n$ on alphabet $\mathcal{X}$ is denoted $\mathcal{P}_n(\mathcal{X})$. The cardinality of a finite set $A$ is denoted by $|A|$. The indicator function of an event $E$ is denoted by $\mathbb{1}_E$.

For two measures $P, Q$ defined on the same measurable space $(\Omega, \mathcal{F})$ the measure $P$ is said to be absolutely continuous w.r.t. $Q$ if for every $E \in \mathcal{F}$ such that $Q(E) = 0$ it also holds that $P(E) = 0$; this is denoted $P \ll Q$.

III. A FORMAL STATEMENT OF THE PROBLEM

Consider transmission over a memoryless channel described by a conditional probability mass function $W(y|x)$, with input $x \in \mathcal{X}$ and output $y \in \mathcal{Y}$ finite alphabets $\mathcal{X}$ and $\mathcal{Y}$. We define $W^n(y|x) = \prod_{k=1}^n W(y_k|x_k)$ for input/output sequences $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and $y = (y_1, \ldots, y_n) \in \mathcal{Y}^n$. The corresponding RVs are denoted by $\mathbf{X}$ and $\mathbf{Y}$.

An encoder maps a message $m \in \{1, \ldots, M_n\}$ to a channel input sequence $\mathbf{x}_m \in \mathcal{X}$, where the number of messages is denoted by $M_n$. The message, represented by the random variable $M$, is assumed to take values in $\{1, \ldots, M_n\}$ equiprobably. This mapping induces an $(n, M_n)$-codebook $C_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_{M_n}\}$ with rate $R_n = \frac{1}{n} \log M_n$.

Upon observing the channel output $\mathbf{y}$, the decoder produces an estimate of the transmitted message $\hat{m} \in \{1, \ldots, M_n\}$. We consider the decoding rule

$$\hat{m} = \arg \max_{i \in \{1, \ldots, M_n\}} q(x_i, y),$$

(1)

where $q(x_i, y)$ is a certain additive decoding metric [27, Ch. 2] defined by a single-letter mapping $q : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ such
that
\[
q(x, y) = \frac{1}{n} \sum_{i=1}^{n} q(x_i, y_i) = \mathbb{E}_{\hat{P}_{XY}}(q(X, Y)) \triangleq q(\hat{P}_{XY}),
\]
where for convenience we slightly abuse notation using \( q \) for both the per-letter metric \( q(x, y) \) and the \( n \)-letter metric \( q(\hat{P}_{XY}) \). Throughout the paper it is assumed that ties are broken uniformly between the maximizers.

Denoting the RV corresponding to the decoded message by \( \hat{M}_q(Y) \), we denote the average error probability as \( P_e(W, C_n, q) = \mathbb{P}[\hat{M}_q(Y) \neq M] \).

A rate \( R \) is said to be achievable with decoding metric \( q \) if there exists a sequence of codebooks \( C_n, n = 1, 2, \ldots \) such that \( \frac{1}{n} \log |C_n| \geq R \) and \( \lim_{n \to \infty} P_e(W, C_n, q) = 0 \). The channel capacity w.r.t. metric \( q \), denoted \( C_q(W) \), is defined as the supremum of achievable rates, and is referred to as the mismatch capacity.

Since the optimal decoding rule w.r.t. average error probability (of equiprobable messages) is ML, which is additive for DMCs, Shannon’s channel capacity \( C(W) \) can be viewed in fact as the channel capacity w.r.t. the metric \( q \), which observes \( z \), serves as a genie to the \( Y \)-receiver by providing it with the list
\[
L(z, \hat{P}_{xz}) \triangleq C_n \cap T_n(\hat{P}_{xz} \mid z) \triangleq \{ \mathcal{x}_1(z, \hat{P}_{xz}), \ldots, \mathcal{x}_r(z, \hat{P}_{xz}) \} \quad (7)
\]
of all the codewords, which lie in the empirical distribution of the input signal \( z \). The \( Y \)-receiver compares the metrics of all the codewords in the list and outputs the \( \hat{m} \) as
\[
\hat{m} = \arg \max_{i \in \mathcal{x}_i \in C(z, \hat{P}_{xz})} q(\mathcal{x}_i, y),
\]
where ties are broken uniformly between the maximizers. This setup is depicted in Fig. 1.

Since by definition the true codeword belongs to this narrowed down list, the error probability in mismatched decoding of the \( Y \) receiver cannot exceed that of the original single-user setup.

The narrowing of the codebook to the list enables to lower bound the average pairwise error probability within the list since within the list, the joint pairwise empirical distribution of codewords is more constrained compared to the entire codebook. The actual size of the list does not play a major role in the proof, except for exceeding 1; i.e., containing at least one additional message except the transmitted one.

We show that for rates which exceed our upper bound on the mismatch capacity, \( C_q(W) \), the average error probability is bounded away from zero. As for rates below \( C_q(W) \), we upper bound the exponent of the average error probability.

Note that we use the terminology of multicast transmission over a broadcast channel similar to [21], and we also refer
to it as a two-output channel in certain cases where this terminology is more suitable. We emphasize that this should not be confused with the ordinary broadcast channel over which two separate messages are transmitted, and each is intended to a different receiver.

A summary of the main differences between the proof technique of this work and previous ones can be found in Section VI-1.

B. An Overview of the Main Results

In this section we present the new upper bounds on the achievable error exponent using decoding metric \( q \), and the mismatch capacity \( C_q(W) \). Throughout this paper, we adopt the shorthand notation that \( W \) without subscript signifies the original single-user channel \( W_{Y|X} \). Whenever we refer to other marginal distributions; i.e., \( W_{Y|X,Z}, W_{Z|X,Y} \) or \( W_{Y,Z|X} \), the subscript is mentioned explicitly.

1) The Main Upper Bound on \( C_q(W) \): Consider the set of two-output (broadcast) channels:

\[
\mathcal{W}_q(P_X) = \left\{ P_{Y|X}: \min_{U: \mathbb{X} = X(U)} \mathbb{E}\left( X, Y \right) \geq \mathbb{E}(X, Y) \right\}
\]

where \( U \) is an auxiliary RV \(^1\) with alphabet size \( |U| \leq |X|^2 \). and the condition \( \mathbb{V}_{U,XZ} = \mathbb{V}_{V,U,Z} \) signifies that for all \( (u, z, x, z) \), \( \mathbb{V}_{U,XZ}(u, z, x, z) = \mathbb{V}_{U,XZ}(u, x, z) \).

Note that the RVs \( U, X, Y, Z \) have the following intuitive explanation originating from the proofs: \( X \) - a typical channel input symbol, \( Y, Z \) - typical channel output symbols, \( X \) - a worst-case competing codeword symbol, and \( U \) - a time-sharing RV.

Next, let

\[
\mathcal{C}_q(W) = \max_{P_X} \min_{P_{Y|X} \in \mathcal{W}_q(P_X)} \min_{P_{Y|X} \in \mathcal{W}_q(P_X)} I(X; Z).
\]

where throughout this paper, we adopt the convention that the minimum and maximum over an empty set equal \( \infty \) and \( -\infty \), respectively.

Our basic bound on \( C_q(W) \) of the following theorem is proved in Section VII.

Theorem 1: For any \( W \), additive metric \( q \in \mathbb{R} \cup \{-\infty, \infty\} \), and finite alphabet \( Z \),

\[
C_q(W) \leq \mathcal{C}_q(W).
\]

The bound \( \mathcal{C}_q(W) \) is tighter compared to previously known bounds, see Section VI for a comparison to previous works. Note that \( \mathcal{C}_q(W) \) is quite difficult to compute, since in order to determine whether a two-output channel \( P_{Y|X} \) belongs to the set \( \mathcal{W}_q(P_X) \) or not, one needs to solve the minimization problem in (12). A similar problem arises with the computation of many of the previous bounds (e.g., those of \([23],[24]\)).

It is worth mentioning though, that denoting

\[
\Delta^q(P_{XZU}, P_{Y|XZ}) \triangleq \sum_{x,z,u,x,y} P_{U|Z}(u, z)P_{X|U|Z}(x|u, z)
\]

\[
\times P_{X|U}(\bar{x}|u)P_{Y|XZ}(y|x, z)[q(\bar{x}, y) - q(x, y)],
\]

the bound \( \mathcal{C}_q(W) \) \((10)\) can also be expressed as

\[
\mathcal{C}_q(W) = \max_{P_X} \min_{P_{Y|X}: \min_{P_{U|X}} \Delta^q(P_{XZU}, P_{Y|XZ}) \geq 0, P_{V|X} = W} I(X; Z).
\]

In this form of the bound, for every given \( P_X, P_{U|X} \), it is easy to determine whether the channel \( P_{Y,Z|X} \) satisfies \( \Delta^q(P_{XZU}, P_{Y|XZ}) \geq 0 \) or not, nevertheless, one still needs to optimize over the pair \( (P_X, P_{U|X}) \).

2) Possibly Looser Easier to Compute Bounds on \( C_q(W) \): We next present several sets of two-output channels, \( \mathcal{W}^{sym}_q(P_X) \), and \( \mathcal{W}^{psd}_q(P_X) \), \( \mathcal{W}_q(P_X) \), \( \mathcal{W}^{sym}_q(P_X) \), which are subsets of \( \mathcal{W}_q(P_X) \).

Consider the following set of symmetric distributions:

\[
\mathcal{P}_{sym}(X^2 \times Z) \triangleq \left\{ P_{XZ} \in \mathcal{P}(X^2 \times Z) : \forall(x, z, z), \ P_{XZ}(x, z, z) = P_{XZ}(x, z, \bar{z}) \right\},
\]

and further consider sets of joint distributions over \( \mathcal{P}(X^2 \times Z \times Z) \):

\[
\Lambda(P_{XYZ}) \triangleq \left\{ P_{XZY} \in \Lambda(P_{XYZ}) : \forall(x, z), \ P_{X|Y,Z}(x|z) \geq P_{X|Z}(x|z) \right\}.
\]

Next, define two sets of two-output channels.

\[
\mathcal{W}^{sym}_q(P_X) \triangleq \left\{ P_{Y|X}: \min_{P_{V|X} \in \mathcal{P}(X^2 \times Z)} \min_{P_{XZU} \in \Lambda(P_{XY})} \mathbb{E}(X, Y) \geq \mathbb{E}(X, Y) \right\}
\]

\[
\mathcal{W}^{psd}_q(P_X) \triangleq \left\{ P_{Y|X}: \min_{P_{V|X} \in \mathcal{P}(X^2 \times Z)} \min_{P_{XZU} \in \Lambda(P_{XY})} \mathbb{E}(X, Y) \geq \mathbb{E}(X, Y) \right\}.
\]

\(^1\)By inspecting (9) it is evident that without loss of generality one can take \( U \) such that \( Z \) is a deterministic function of \( U \) (perhaps with larger alphabet for \( U \)), because one can replace any \( U \) by \( U' = (U, Z) \). Therefore, one can add the constraint \( H(Z|U) = 0 \) to the set of the minimization in (9) without changing the resulting bound (10).

\(^2\)In fact, the sets, as well as \( \mathcal{W}_q(P) \) are also functions of the alphabet cardinality \( |Z| \), but for the sake of simplicity we omit this dependence from our notation.
and a third set of two-output channels
\[ \tilde{W}_q(P_X) \triangleq \{ P_{Y|Z|X|Y|Z:Z \in Z}: \exists P_{Z|X}: P_{Z|X} \preceq P_{Z|X}, \]
\[ \sum_{x,y,z} V(z|x) V(y|z) P_{Y|X,Y}(y|x,z) q(x,y) - q(x,y) \geq 0 \} \]
\[ = \{ P_{Y|Z|X|Y|Z:Z \in Z}, P_{Z|X}: P_{Z|X} \preceq P_{Z|X} \}
\quad \tilde{E}_q(V|X,Y) \geq E_q(V|X,Y) \}
\] (20)
(21)

where as mentioned in Section II, \( \preceq \) denotes absolute continuity.

Now, denote w.l.o.g. \( \mathcal{X} = \{1, \ldots, |\mathcal{X}|\} \), and consider the collection of symmetric \( |\mathcal{X}| \times |\mathcal{X}| \) matrices \( \{ D^q(P_{Y|X,Z=z}) \}_{i,j} \), indexed by \( z \in \mathcal{Z} \), whose \((i,j)\)-th entries are given by:
\[ D^q(P_{Y|X,Z=z}) \]
\[ = \sum_y \left( P_{Y|X|Z}(y|i, z)[q(j, y) - q(i, y)] + P_{Y|X|Z}(y|j, z)[q(i, y) - q(j, y)] \right) \]
\[ = \sum_y \left[ P_{Y|X|Z}(y|i, z) - P_{Y|X|Z}(y|j, z) \right] \cdot [q(j, y) - q(i, y)] \] (22)
(23)

Denoting that a matrix \( D \) is positive semi-definite (p.s.d.) by \( D \succeq 0 \), we define the last set of two-output channels:
\[ \tilde{W}_{q,psd}(P_X) \triangleq \{ P_{Y|Z}: \forall z \in Z, D^q(P_{Y|X,Z=z}) \succeq 0 \} \] (24)

Let
\[ C^{sym}_{q}(W) \triangleq \max_{P_X} \min_{P_{Y|Z}: I(X;Z) \leq R} I(X;Z) \]
\[ C^{sym}_{q}(W) \triangleq \max_{P_X} \min_{P_{Y|Z}: I(X;Z) \leq R} I(X;Z) \]
\[ C_{q}(W) \triangleq \max_{P_X} \min_{P_{Y|Z}: I(X;Z) \leq R} I(X;Z) \]
\[ C^{psd}_{q}(W) \triangleq \max_{P_X} \min_{P_{Y|Z}: I(X;Z) \leq R} I(X;Z) \] (25)
(26)
(27)
(28)

The possibly looser bounds compared to \( \overline{C}_q(W) \) are presented in the following proposition, which is proved in Appendix A.

**Proposition 1:** For any \( P \in \mathcal{P}(\mathcal{X}) \),
\[ W_{q}(P) \subseteq \tilde{W}_{q}(P) \subseteq \tilde{W}_{q,psd}(P), \]
\[ \overline{W}_{q,psd}(P_X) \subseteq \overline{W}_{q,psd}(P) \subseteq \tilde{W}_{q,psd}(P) \] (29)
(30)

and therefore
\[ \overline{C}_q(W) \leq \overline{C}_q(W) \leq C^{psd}_{q}(W), \]
\[ \overline{C}_q(W) \leq C^{sym}_{q}(W) \leq \overline{C}_q(W). \] (31)
(32)

As mentioned before, the bound \( C^{psd}_{q}(W) \) has a significant advantage over \( \overline{C}_q(W) \), since it is easier to compute in the sense that determining whether a two-output channel \( P_{Z|X} \) belongs to the set \( W_{q,psd}(P_X) \) requires a simple calculation. In particular, one needs to verify that \( \forall z \in Z, D^q(P_{Y|X,Z=z}) \succeq 0 \) by checking that the determinants of the \(|\mathcal{X}| - 1\) minors of each of these \(|\mathcal{Z}| \) matrices are all non-negative [31]. This is in contrast to the calculation of \( \overline{C}_q(W) \) and similarly several other previous bounds (e.g., those of [23] and [24]), which require to solve a certain minimization problem, such as the minimization in (9), in order to determine whether a two-output channel belongs to the set \( W_{q}(P_X) \) (see Section VI).

Furthermore, the bound \( C^{sym}_{q}(W) \) is also easier to compute numerically compared to \( \overline{C}_q(W) \) and previous bounds, since the constraints in the definition of the set \( W_{sym}(P) \) limit the range and the number of degrees of freedom of the solution of the optimization problem. In Section V, we analyze the various bounds for the binary-input channel case \(|\mathcal{X}| = 2, |\mathcal{Z}| \leq \infty \).

3) **Upper Bounds on the Reliability Function With Decoding Metric** \( q \): In this section we present our main bound \( E_{sp}(R, W) \) on the reliability function with mismatched decoding, \( E_{q}(R, W) \), and similar to the mismatched capacity, we present looser bounds that are easier to compute.

For \( P \in \mathcal{P}(\mathcal{X}) \) define
\[ E_{sp}(R, P, W) \triangleq \min_{P_{Y|Z}: I(X;Z) \leq R} D(P_{Y|X}||W|P) \] (33)

Due to (6), our main result concerning the reliability function is presented in terms of upper bounds on \( e_{n}(R, P, W) \).

**Theorem 2:** Let \( |\mathcal{Z}| \leq \infty \), then for all \( n \), and any \( P \in \mathcal{P}(\mathcal{X}) \)
\[ e_{n}(R, P, W) \leq E_{sp}(R - \epsilon_{n,a}, P, W) + \epsilon_{n,b}, \] (34)
where \( \epsilon_{n,a} = O\left(\frac{\log n}{n}\right) \), and \( \epsilon_{n,b} = O\left(\frac{\log n}{n}\right) \).

The main idea of the proof of Theorem 2 was presented in Section IV-A, the full proof can be found in Section VII, as well as the exact quantities \( \epsilon_{n,a} \) and \( \epsilon_{n,b} \). Note that the proof of Theorem 2 can be shortened to yield the same result as (34) with the exception that \( E_{q} \) is replaced\(^3\) by \( \overline{E}_{sp,psd} \).

Now, define further
\[ \overline{E}_{sp}(R, P, W) \triangleq \min_{P_{Y|Z}: I(X;Z) \leq R} D(P_{Y|X}||W|P) \] (35)
\[ E_{sp}(R, P, W) \triangleq \min_{P_{Y|Z}: I(X;Z) \leq R} D(P_{Y|X}||W|P) \] (36)
\[ E_{sp}(R, P, W) \triangleq \min_{P_{Y|Z}: I(X;Z) \leq R} D(P_{Y|X}||W|P) \] (37)
\[ \overline{E}_{sp}(R, P, W) \triangleq \min_{P_{Y|Z}: I(X;Z) \leq R} D(P_{Y|X}||W|P). \] (38)

Further, let \( E_{sp}(R, W), \overline{E}_{sp}(R, W), E_{sp}(R, W) \) denote the maximum over \( P \in \mathcal{P}(\mathcal{X}) \) of the above quantities (33)-(38), respectively. The following theorem states the additional bounds.

\(^3\)This is mentioned in Eq. (102) that can replace the stronger result (103), which requires a slightly more involved proof.
Theorem 3: The following inequalities hold for all $(R, P, W)$

\[ E_q^q(R, P, W) \leq \bar{E}_q^q(R, P, W) \leq E_{q,p}^q(R, P, W) \tag{39} \]

\[ E_{q,p}^q(R, P, W) \leq E_{q,s}^{q,sym}(R, P, W) \leq E_{q,sym}^q(R, P, W). \tag{40} \]

The inequalities of (39)-(40) follow from (29)-(30). The following corollary follows from (6), and Theorems 2 and 3.

Corollary 1: For any $|Z| < \infty$, 

\[ E^q(R, W) \leq E_{q,p}^q(R, W) \leq E_{q,s}^{q,sym}(R, W) \leq E_{q,sym}^q(R, W), \tag{41} \]

\[ E^q(R, W) \leq E_{q,p}^q(R, W) \leq E_{q,s}^{q,sym}(R, W) \leq E_{q,sym}^q(R, W). \tag{42} \]

The following corollary, which is easily verified, states that $C_q(W)$ is a rate above which the average error probability in $q$ decoding cannot vanish exponentially fast.

Corollary 2: For all $R > C_q(W)$, $E_{q,p}^q(R, P, W) = 0$.

4) Upper Bound on the Reliability Function for Type-Dependent Metrics: Our next corollary extends the results of Theorems 2 and 3 to the case of type-dependent metrics, similar to the extension in [21]. The class of type-dependent metrics generalizes additive metrics in the following manner. It is assumed that the decoding metric $q(x, y)$ depends on $x, y$ solely via their joint empirical distribution; i.e., $q(x, y) = q(P_{x,y})$, so $q$ can be viewed as a mapping from the empirical distributions to the reals $q: \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$. More generally, in order not to restrict attention to a specific block-length $n$, we assume that it maps the simplex to a real number; i.e.,

\[ q: \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}. \tag{43} \]

We refer to this class of metrics as type-dependent (formerly referred to as $\alpha$-decoders by Csiszár and Körner [3]). In the case of type-dependent metrics, (1) becomes:

\[ \tilde{n} = \arg \max_{i \in [1, \ldots, m_n]} q(P_{x,y}). \tag{44} \]

The equivalent of the set of two-output channels $\mathcal{W}_q^{sym}(P_X)$ in (18) for type-dependent metrics is given by

\[ \mathcal{W}_q^{sym}(P_X) = \left\{ P_{Y|X}: \min_{V_{X,Y,Z}: V_{X,Y,Z} \in \mathcal{P}_{sym}(X^2 \times Z)} q(V_{X,Y}) \geq q(P_{XY}) \right\}, \tag{45} \]

yielding $\bar{E}_{q,s}^{q,sym}(R, P, W)$ as defined in (38).

The result pertaining to type-dependent metrics is the following.

Corollary 3: Let $|Z| < \infty$, and let $q(P_{XY})$ be convex in $P_{Y|X}$ for fixed $P_X$, then for all $n$, and any $P \in \mathcal{P}_n(\mathcal{X})$

\[ e^q_n(R, P, W) \leq \bar{E}_{q,s}^{q,sym}(R - \epsilon_{n,c}, P, W) + \epsilon_{n,d}, \tag{46} \]

where $\epsilon_{n,c} = O\left(\log n \over n\right)$ and $\epsilon_{n,d} = O\left(\log n \over n\right)$, and consequently

\[ E^q(R, W) \leq \bar{E}_{q,s}^{q,sym}(R, W). \tag{47} \]

The corollary is proved in Appendix D.

V. BINARY-INPUT CHANNELS

As mentioned above, in [17], a single-letter expression for the mismatch capacity in the case of the binary input binary output channel was derived. The single-letter converse result reported in [32] for binary-input DMCs (with $2 < |\gamma| < \infty$) was disproved in [15]. Specifically, a rate based on superposition coding was shown to exceed the claimed mismatch capacity of [32].

The following lemma specifies simplified explicit expressions for $C_q^{sym}(W), C_q^{sp}(W), and C_q^{psd}(W)$ for binary input DMCs, by explicitly solving the minimization problems in the definitions of $W_q^{sym}(P_X), W_q^{sp}(P_X),$ and $W_q^{psd}(P_X)$. Let $d_q(P_{YZ|x,z}) = \sum_y [P(y|0, z) - P(y|1, z)][q(1, y) - q(0, y)]$.

\[ \text{Lemma 1: Let } W_{Y|X} \text{ be a DMC with a binary input alphabet } \mathcal{X} = \{0, 1\}, \text{ and let } |Z| < \infty, \text{ then}^4 \]

\[ C_q^{sym}(W) = \max_{P_X} \min_{P_{YZ|X}} I(X; Z) \tag{48} \]

\[ \leq \max_{P_X} \min_{P_{YZ|X}: P_{Y|X} = W} I(X; Z), \tag{49} \]

Consequently

\[ C_q^{sym}(W) \leq C_q^{sp}(W) \leq C_q^{psd}(W). \tag{50} \]

Further,\n
\[ C_q^{sym}(W) = C_q^{sym}(W). \tag{51} \]

\[ C_q^{psd}(W) = \max_{P_X} \min_{P_{YZ|X}: P_{Y|X} = W} I(X; Z). \tag{52} \]

\[ \text{Lemma 1 is proved in Appendix B. For given } P_X, \text{ it is easy to check whether the condition in the minimization in (49) is satisfied for a suggested channel } P_{YZ|X}, \text{ as well as the conditions in (50)-(53). Further, it is evident that in the binary-input case } C_q^{sym}(W) \leq C_q^{psd}(W). \]

Example 1: Consider the example $\mathcal{X} = \{0, 1\}, \mathcal{Y} = \mathcal{Z} = \{0, 1, 2\}$ (also studied in [15] and [24]):

\[ W_{Y|X} = \begin{pmatrix} 0.97 & 0.03 & 0 \\ 0.1 & 0.9 & 0.8 \end{pmatrix}, \tag{53} \]

\[ q(x, y) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \log(0.5) & \log(1.36) \end{pmatrix}. \tag{54} \]

\[ \text{Since } P_{X|Z}(y|x, z) \text{ is not defined for } (x, z) \text{ such that } P_{XZ}(x, z) = 0, \text{ the condition } P_{Z}(z)P_{X|Z}(0|z)P_{X|Z}(1|z)d_q(P_{YZ|x,z}) \geq 0 \text{ should be understood as } P_{Z}(z)P_{X|Z}(0|z)P_{X|Z}(1|z) > 0 = d_q(P_{YZ|x,z}) \geq 0. \tag{55} \]
It is easy to verify that the following two-output channel
\[
\tilde{W}_{YZ|X}(y, z|x) = \begin{cases} 
0.77 & (x, y, z) = (0, 0, 1) \\
0.6 & (x, y, z) = (1, 2, 2) \\
0.2 & (x, y, z) \in \{(0, 0, 0), (1, 2, 1)\} \\
0.1 & (x, y, z) \in \{(1, 0, 0), (1, 1, 1)\} \\
0.03 & (x, y, z) = (0, 1, 1) \\
0 & \text{otherwise}
\end{cases}
\]
satisfies
\[
\tilde{W}_{Y|X} = W_{Y|X}, \tilde{W}_{Z|X} = \begin{pmatrix} 0.2 & 0.8 & 0 \\
0.1 & 0.3 & 0.6 \end{pmatrix},
\]
Assume w.l.o.g. that the maximizing \(P_X\) is non-degenerate (as otherwise the mismatch capacity equals zero). It is easy to verify that \(d_q(\tilde{W}_{Y|X}, z) \geq 0\) for all \(z\) such that \(\tilde{W}_{Z|X}(z|1) > 0\), i.e., \(z \in \{0, 1\}\). Thus,
\[
C_{q,\text{sym}}(W_{Y|X}) \leq \max_{P_X} I(P_X \times \tilde{W}_{Z|X}) \\
\approx 0.4081 \text{ [bits/channel use]},
\]
where \(P_X \approx (0.59, 0.41)\). For a comparison with previous results pertaining to this example and other cases, see the next Section VI-3.

VI. A COMPARISON TO PREVIOUS RESULTS

We next discuss the relationship between our new results and some relevant previous converse results. We begin by describing the major differences in the proof technique. Further, in Sections VI-3 and VI-2 we describe the differences between the mismatch capacity and reliability function expression, respectively.

1) Main Differences Compared to Proof Techniques of Previous Works: The Z-receiver of our previous works \([21, 23]\) was either genie-aided or not (depending on the specific theorem), but the major difference is that it did not serve as a genie to the Y-receiver in these works. These works relied on an argument that the broadcast channel \(P_{YZ|X}\) is such that if the Y-receiver successfully decodes the message, then so does the Z receiver (in the case of \([23]\) this was relaxed to Z-receiver that is very likely to succeed).

As for \([20, 24, 25]\), these works did not consider multicast transmission nor genies, the technique therein relied on a construction of a graph in the output space. Our bounding technique using a two-output channel, in which the Z-receiver serves as a genie to the Y-receiver is significantly simpler and does not involve the construction of a graph and graph-theoretic tools.

Another major difference in the proof technique between the current work and \([24, 25]\) is that in the current work we exploit the symmetric requirement from the codewords: For the reliability function, we analyze \(P_{\tilde{E}_{ij}|x_i, z}\), the probability of the error event \(\tilde{E}_{ij}\) that \(q(x_i, Y)\) exceeds \(q(x_i, \hat{Y})\) given that \(i\) is transmitted and \(z\) received, and lower bound the average of \(\frac{1}{2}[P_{\tilde{E}_{ij}|x_i, z} + P_{\tilde{E}_{ij}|x_i, z}]\); (see Eq. (84)). For the mismatch capacity upper bound we state a condition such that for any pair of codewords \(x_t, x_k\) in the list, the joint empirical distribution is such that either given \((x_t, z)\) the event \(E_{tk}\) is very likely to occur or given \((x_k, z)\) the event \(E_{kt}\) is very likely to occur (see Lemma 6 Eqs. (126)-(127)).

The upper bound of \([24]\) and \([25]\) implies that the symmetry between the possible events (of each of the messages being the transmitted one) was not exploited in the derivation.

2) Reliability Function: The general upper bounds which hold for ML decoding are applicable for mismatched decoding as well, and in particular, the classical sphere-packing bound \([27]\), which is given by:
\[
E_{sp}(R, P, W) = \min_{P_{Y|X} : I(P_X \times Y) \leq R} D(P_{Y|X} \parallel W|P),
\]
as it holds for all metrics \(q\) including the ML metric.

Consider \((39)-(40),\) yielding that \(E_{sp}(R, P, W)|_{q=\log W}\) essentially upper bounds \(e_q^*(R, P, W)|_{q=\log W}\). Taking \(Z = Y\) (instead of minimizing) in our bound \((33),\) \(E_{sp}(R, P, W)|_{q=\log W}\), and noting that \(P_{Y|X} = I(Y | Z) \in W_q(P_X)\) (since \(P_{XZ} = P_{XZ}\) implies that \(P_{XY} = P_{XY}\)) and thus \(E_q(X|Y) = E_q(X, Y) = 0\) yields \(E_{sp}(R, P, W)|_{q=\log W} \leq E_{sp}(R, P, W)\).

In \([24]\), an upper bound on the mismatched reliability function was derived, which we denote (to avoid confusion) as
\[
f_q^{sp}(R, P, W) \triangleq \min_{P_{Y|X} : I(P_X \times Y) \leq R} \min_{P_{X|Y}, \hat{Y} \in \mathcal{X}} E_{Y}(\hat{Y} | X, \hat{Y}) \leq \min_{P_{X|Y} : P_{XZ} = P_{XZ}, \hat{Y} \in \mathcal{X}, (X, \hat{Y}) \rightarrow Y} \min_{P_{X|Z} : P_{XZ} \in \mathcal{P}_{\text{sym}}(\mathcal{A} \times \mathcal{Z})} E_{P_{XZ}}(\tilde{Y} | X, Z) . \tag{59}
\]
Note that we have (treating \(\hat{Y}\) in the role of \(Z\))
\[
\min_{P_{X|Z} : P_{XZ} \in \mathcal{P}_{\text{sym}}(\mathcal{A} \times \mathcal{Z})} E_{P_{XZ}}(\tilde{Y} | X, Z) \leq \min_{P_{X|Z} : P_{XZ} \in \mathcal{P}_{\text{sym}}(\mathcal{A} \times \mathcal{Z})} E_{P_{XZ}}(\tilde{Y} | X, Z) . \tag{59}
\]
In addition to being at least as tight, the bound \(E_q^{sp}(R, P, W)\) has the advantage that due to the symmetry, solving the associated minimization problem on the r.h.s. of \((59)\) involves fewer degrees of freedom to optimize over and narrower parameters range compared to the l.h.s. of \((59)\). A similar comment holds even more so when \(E_{sp}^{sym}(R, P, W)\) is concerned, due to the constraint \(\forall (x, z), P_{X|Z}(x | z) \geq P_{X|Z}(x | z) \text{ in } (19)\).

3) Mismatch Capacity: As mentioned above, \([20]\), a single-letter bound was derived by forming a transformation of the channel into another translated channel from \(\mathcal{X} \rightarrow \tilde{Y}\) such that \(q\)-decoding error at the latter implies \(q\)-decoding error at the original channel. The idea was to connect the two channels by means of a graph in the output space \(\tilde{Y}\).

In \([21]\), the multicast transmission proof technique was proposed, which yielded a few tighter bounds. The main idea of the proof is that same message is transmitted simultaneously
to two decoders over a two-output (broadcast) channel $W_{YZ|X}$ with two outputs $Y$ and $Z$. The $Z$-decoder employs an additive decoding metric $\rho$ that can be optimized, and the two-output channel belongs to the set

$$\Gamma(q, \rho) \triangleq \left\{ P_{YZ|X}: \forall (x,y,z): \rho(x,y) - q(x,y) < \max_{x' \in X} |\rho(x',y) - q(x',y)| \right\}. \tag{60}$$

It is easily verified (see details in [21]), that any two-output channel in this class has the property that an error occurs at the $Z$-decoder, only if the $Y$-receiver makes an error. Thus, for any codebook $C_n$, we have

$$P_e(W_{YZ|X}, C_n, q) \geq P_e(W_{Z|X}, C_n, \rho) \tag{61}$$

and the following bound holds for any stationary memoryless channel.

$$C_q(W) \leq \max_{P_X} \min_{P_{YZ|X} \in \Gamma(q, \rho)} I(X; Z), \tag{62}$$

It can be shown that the bound of [20] can be expressed as a special case for the suboptimal choice of $\rho = q$ and $Z = Y$. It is very easy to verify whether channel $P_{YZ|X}$ lies in the set of the minimization since calculating the marginal $P_{Y|X}$ involves a simple summation, and since $\Gamma(q, \rho)$ only dictates constraints on necessary zero values for $P_{YZ|X}$. Furthermore, this max-min problem is easily solvable numerically (see [21, Section IV-C]).

Another bound that was derived in [21] and is applicable to DMCs, enlarges the set of channels to one that depends also on the input distribution $P_X$. This bound was obtained by including a genie that helps the $Z$-receiver by providing it with the actual value of the joint empirical distribution of the transmitted codeword and the output sequence $Z$. In [23] the bound was improved to include a minimization over the larger class of channels

$$\Theta^*(q, P_X) \triangleq \left\{ P_{YZ|X}: \forall (x,y,z): \exists \bar{X}(\bar{x}, \bar{y}) \geq \bar{X}(X, Y) \right\}. \tag{63}$$

These channels satisfy the condition that if the $Y$-decoder successfully decodes the message, then with high probability (approaching 1 as the block length tends to infinity), also does the genie-aided $Z$-decoder.

A further improved bound was derived in [24] where the set of channels is given by

$$\mathcal{M}_{max}(q, P_X) \triangleq \left\{ P_{YZ|X}: \forall (x,y,z): \exists \bar{X}(\bar{x}, \bar{y}) \geq \bar{X}(X, Y) \right\}, \tag{63}$$

which includes the additional constraint $\bar{X} = (X, Z) - Y$. The resulting bound is denoted $R(W, q)$.

Note that

$$\left\{ V_{\bar{X}Z}\in \mathcal{P}_{sym}(\mathcal{X}^2 \times \mathcal{Z}): V_{XZ} = V_{\bar{X}Z} = P_{XZ} \right\} \tag{64}$$

Therefore, from (11) we obtain

$$C_q^{sym}(W) \leq \bar{R}(W, q). \tag{65}$$

In the next lemma we show that there are many cases for which (64) is obtained in strict inclusion (for $|\mathcal{X}| > 2$). It is likely that there exist such cases where strict inclusion results in a strict inequality in (65). This is reinforced by the fact that $C_q(W) \leq C_q^{sym}(W) \leq C_q^{sym}(W)$, and since the additional constraint which appears in $C_q^{sym}(W)$; i.e., $P_{X\bar{X}}(x, x|z) \geq P_{XZ}(x|z)$ may be active. But, even more importantly, the above mentioned previous bounds except (150) are rather complicated to compute, compared to the bounds $C_q^{ed}(W)$ and $C_q^{sym}(W)$.

**Lemma 2:** For $|\mathcal{X}| > 2$ the following inclusion is strict:

$$\left\{ V_{\bar{X}ZZ}\in \mathcal{P}_{sym}(\mathcal{X}^2 \times \mathcal{Z}): V_{XZ} = V_{\bar{X}Z} = P_{XZ} \right\} \subset \left\{ V_{X\bar{X}Z}: V_{XZ} = V_{\bar{X}Z} = P_{XZ} \right\}. \tag{66}$$

**Proof:** It is sufficient to establish the result for $|\mathcal{X}| = 3$. Consider the following $3 \times 3$ symmetric matrix:

$$A(z) = \left( \begin{array}{ccc} a_1(z) & a_2(z) & a_3(z) \\ a_2(z) & a_3(z) & a_4(z) \\ a_3(z) & a_4(z) & a_5(z) \end{array} \right), \tag{67}$$

where $a_i(z) \in (0,1)$. Next, consider the matrix $B(z)$ which is obtained by subtracting or adding a positive quantity which satisfies the condition $\Delta \leq \min(a_2(z), a_3(z), a_5(z), 1 - a_2(z), 1 - a_3(z), 1 - a_5(z))$ from $A(z)$’s entries as follows:

$$B(z) = \left( \begin{array}{ccc} a_1(z) & a_2(z) + \Delta & a_3(z) - \Delta \\ a_2(z) - \Delta & a_4(z) & a_5(z) + \Delta \\ a_3(z) + \Delta & a_5(z) & a_6(z) \end{array} \right). \tag{68}$$

Note that the condition guarantees that $0 \leq B_{i,j}(z) \leq 1$. A more general choice is also applicable with the appropriate conditions on $k_1, k_2 \neq k_3, k_4$ (which are not all necessarily positive) ensuring that $0 \leq K_{i,j}(z) \leq 1$ where:

$$K(z) = \left( \begin{array}{ccc} a_1(z) + k_1 & a_2(z) + k_2 & a_3(z) - k_1 - k_2 \\ a_2(z) + k_3 & a_4(z) + k_4 & a_5(z) - k_3 - k_4 \\ a_3(z) - k_1 - k_3 & a_5(z) - k_2 - k_4 & a_6(z) + \sum_{i=1}^{4} k_i \end{array} \right). \tag{69}$$

Clearly, the sums of corresponding columns and rows of $A(z)$ and $K(z)$ are identical (e.g. $\sum_{i} A_{i,j}(z) = \sum_{j} K_{i,j}(z)$), but while $A(z)$ is symmetric, $K(z)$ is not (since $k_2 \neq k_3$). Now, let $V_{X\bar{X}} \in \mathcal{P}_{sym}(\mathcal{X}^2 \times \mathcal{Z})$ which satisfies $V_{XZ} = V_{\bar{X}Z} = P_{XZ}$ be given, and think of $A_{i,j}(z)$ as $V_{X\bar{X}|Z=i,j}(i,j|z)$ and define $V_{XZ|Z=i,j}(i,j|z) = K_{i,j}(z)$ which is clearly non symmetric. Thus $P_{Z|X}V_{XZ|Z=i,j}$ belongs to the set on the r.h.s. of (66) but does not belong to the set on the l.h.s. of (66), and the lemma follows.

It is easy to realize that for $|\mathcal{X}| = 2$ the inclusion (64) is not strict; i.e., (64) is obtained with equality. Combining this with (52), we obtain

If $|\mathcal{X}| = 2$,

then $\bar{R}(W, q) = C_q^{sym}(W) = C_q^{sym}(W) \geq C_q(W). \tag{70}$
Further, unlike our Lemma 1, the optimization problem over $V_{X|Y}^{\epsilon}$ in (63) associated with binary input channels was not solved explicitly in [24], and therefore our expression (50) is significantly simpler, and easier to compute.

Consequently, for the binary input ternary output Example 1) (see (54)) it was established numerically in [24] that

$$R(W,q) \leq 0.4999,$$  \hspace{1cm} (71)

while we rigorously established (58); i.e.,

$$C_q(W) \leq \max_{P_x} I(P_X \times \hat{W}_{Z|X}) \approx 0.4081.$$  \hspace{1cm} (72)

Note that it is not evident whether the bound 0.4081 may be improved by calculating $C_q(W)$ rather than $C_q(W)$, or by another choice of channel instead of $\hat{W}_{YZ|X}$.

VII. PROOF OF THEOREM 2

In this section we prove Theorem 2 based on the outline described in Section IV-A.

Let a DMC $W = W_{Y|X}$ be given. Fix $P_n \in \mathcal{P}_n(\mathcal{X})$, and let $C_n = \{\vec{P}_{i}\}_{i=1}^{\lfloor \frac{b}{n}\rfloor}$ be a $P_n$-constant composition codebook of rate $R$ for the channel $W_{Y|X}$. Consider another channel from $\mathcal{X} \times \mathcal{Y}$ to a finite set $Z$ denoted by $W_{Z|X,Y}$, which along with $W_{Y|X}$ constitutes a two-output channel $W_{Y,Z|X}$. For technical reasons, we assume that $W_{Z|X,Y}$ takes on the following form for all $(x, y, z)$:

$$W_{Z|X,Y}(z|x, y) = (1 - \tau_n) \cdot W_{Z|X,Y}(z|x, y) + \frac{\tau_n}{|Z|},$$  \hspace{1cm} (73)

where $\tau_n = \frac{1}{n}$, and $W_{Z|X,Y}(z|x, y)$ is some conditional distribution from $\mathcal{X} \times \mathcal{Y}$ to $Z$. Note that this implies that for all $(x, y, z)$

$$W_{Y|Z}(y|x) = \frac{W(y|x)W(z|x, y)}{W(z|x)} \geq W_{Y|X}(y|x) \cdot \frac{\tau_n}{|Z|}.$$  \hspace{1cm} (74)

Denote

$$w_{\min} \triangleq \min_{(x,y)}: W(y|x) > 0 \hspace{0.5cm} W_{Y|X}(y|x),$$  \hspace{1cm} (75)

$$t_{n,\min} \triangleq \frac{\tau_n}{|Z|} \cdot w_{\min}.$$  \hspace{1cm} (76)

Given the channel input $X \in T_n(P_n)$, and the joint type-class of $(X, Z)$, $P_{XZ} = \hat{P}_{XZ}$, we clearly have that $Z$ is distributed uniformly over $T_n(\hat{P}_{Z|X|X})$; i.e.,

$$\Pr(Z = z|X = x, \hat{P}_{XZ} = \hat{P}_{XZ}) = \frac{1}{|T_n(\hat{P}_{Z|X|X})|} \cdot \mathbb{1}_{z \in T_n(\hat{P}_{Z|X|X})}.$$  \hspace{1cm} (77)

Recall the definition of $\mathcal{L}(z, \hat{P}_{XZ})$, and $\mathcal{L}(z, \hat{P}_{XZ})$ in (7), and note that by definition, $|\mathcal{L}(z, \hat{P}_{XZ})| \geq 1$ for any $\hat{P}_{XZ}$ which is a possible joint empirical distribution of a channel input-output sequences pair $x$, $z$. Further recall the assumption that the Y-decoder is informed of the list $\mathcal{L}(z, \hat{P}_{XZ}) = \{x_i(Z, \hat{P}_{XZ})\}$ and employs the decoding rule (8).

It is easily verified that for any possible channel output $z$ such that $\hat{P}_Z = \hat{P}_Z$, it holds that $\{x_i(z, \hat{P}_{XZ})\}$ are equiprobable given $(Z = z, \hat{P}_{XZ} = \hat{P}_{XZ})$; that is,

$$P(X = x_i(z, \hat{P}_{XZ})|Z = z, \hat{P}_{XZ} = \hat{P}_{XZ}) = \frac{1}{|\mathcal{L}(z, \hat{P}_{XZ})|}.$$  \hspace{1cm} (78)

To see this, note that by applying Bayes’ law twice we have

$$\Pr(X = x_i(z, \hat{P}_{XZ}), Z = z|\hat{P}_{XZ} = \hat{P}_{XZ})$$

$$= \Pr(Z = z|\hat{P}_{XZ} = \hat{P}_{XZ}) \cdot \Pr(X = x_i(z, \hat{P}_{XZ})|Z = z, \hat{P}_{XZ} = \hat{P}_{XZ})$$

$$= \Pr(X = x_i(z, \hat{P}_{XZ})|\hat{P}_{XZ} = \hat{P}_{XZ}) \cdot \Pr(Z = z|X = x_i(z, \hat{P}_{XZ}), \hat{P}_{XZ} = \hat{P}_{XZ}).$$  \hspace{1cm} (79)

Now, since the code is constant composition, the actual joint type-class $\hat{P}_{XZ}$ does not depend on the codeword $X$, and hence $\Pr(X = x_i(z, \hat{P}_{XZ})|\hat{P}_{XZ} = \hat{P}_{XZ}) = \frac{1}{|T_n(\hat{P}_{XZ})|}$, and further we have $\Pr(Z = z|X = x_i(z, \hat{P}_{XZ}), \hat{P}_{XZ} = \hat{P}_{XZ}) = \frac{1}{|T_n(\hat{P}_{Z|X|X})|}$, this yields

$$\Pr(X = x_i(z, \hat{P}_{XZ})|Z = z, \hat{P}_{XZ} = \hat{P}_{XZ})$$

$$= \frac{1}{|T_n(\hat{P}_{Z|X|X})|} \cdot M_n \cdot \Pr(Z = z|\hat{P}_{XZ} = \hat{P}_{XZ}).$$  \hspace{1cm} (80)

Since the r.h.s. does not depend on $i$ we obtain the desired result (78).

Next, let

$$E_{ij} \triangleq E_{ij}(z, \hat{P}_{XZ}) \triangleq \{y : q(x_i(z, \hat{P}_{XZ}), y) \geq q(x_i(z, \hat{P}_{XZ}), y)\},$$  \hspace{1cm} (81)

and adopt the shorthand notation

$$x_i \triangleq x_i(z, \hat{P}_{XZ}), \mathcal{L} \triangleq \mathcal{L}(z, \hat{P}_{XZ}).$$  \hspace{1cm} (82)

Since $\Pr(error|x, z) = \Pr(\cup_{j \neq i} E_{ij}|x, z)$, we have the following lower bound on the average error probability in $q$-mismatched decoding at the Y-receiver given that $Z = z$ and $\hat{P}_{XZ} = \hat{P}_{XZ}$,

$$\Pr(error|z, \hat{P}_{XZ} = \hat{P}_{XZ})$$

$$\geq \left\{ \frac{1}{|\mathcal{L}|} \sum_{i,j \in \mathcal{L}} \Pr(E_{ij}|x, z) \right\} \mathcal{L} \geq 2 \hspace{0.5cm}|\mathcal{L}| \geq 1.$$  \hspace{1cm} (83)

Evidently, we shall focus on the case where $(z, \hat{P}_{XZ})$ are such that $|\mathcal{L}| \geq 2$ to obtain a meaningful lower bound on the error probability. Denote $h_n \triangleq \frac{1}{(n+1)[|\mathcal{X}|^2|\mathcal{Y}|]}$ and note that

$$\Pr(E_{ij}|x, z) = \sum_{y : q(x_i, y) \geq q(x_i, y)} W_{Y|X}(y|x, z),$$

$$\geq h_n \cdot V_{X|Z} \in \mathcal{P}_n(|\mathcal{X}^2 \times |\mathcal{Y}|$$

$$\geq V_{XZ} = \hat{P}_{XZ}; q(V_{X|Y}) \geq q(V_{X|Z})$$

$$\geq h_n \cdot e^{-nD(V_{Y|X}||W_{Y|X}|\hat{P}_{XZ}|)}.$$  \hspace{1cm} (84)
where for an empirical distribution $P_{\tilde{X}Z} \in \mathcal{P}_n(\mathbb{X}^2 \times \mathbb{Z})$ satisfying $P_{\tilde{X}Z} = P_{XZ}$ we define

$$\Omega_n(P_{\tilde{X}Z}, W_{Y|XZ}) \triangleq \min_{V_{\tilde{X}XY} \in \mathcal{S}(P_{\tilde{X}Z})} D(V_{Y|XZ} || W_{Y|XZ}|P_{\tilde{X}Z})$$

We present the following approximation lemma whose proof appears in Appendix C.

**Lemma 3:** For all $P_{\tilde{X}Z} \in \mathcal{P}_n(\mathbb{X}^2 \times \mathbb{Z})$,

$$\Omega_n(P_{\tilde{X}Z}, W_{Y|XZ}) \leq \Omega(P_{\tilde{X}Z}, W_{Y|XZ}) + \delta_n,$$

where $\delta_n = 2 \cdot \frac{|\mathbb{X}^2||\mathbb{Z}|}{n} \log \frac{n^2 |\mathbb{Z}|}{w_{\text{min}}}$, with $w_{\text{min}}$ defined in (76).

Thus,

$$\Pr(C_{ij}|x_i, z) \geq h_n \cdot e^{-n[\Omega(P_{\tilde{X}Z}, W_{Y|XZ})+\delta_n]}.$$  

Further,

$$\sum_{x_i, z_i \in \mathcal{L}, j \neq i} e^{-n[\Omega(P_{\tilde{X}Z}, W_{Y|XZ})]} = \sum_{x_i, z_i \in \mathcal{L}, j \neq i} \left[ e^{-n[\Omega(P_{\tilde{X}Z}, W_{Y|XZ})]} + e^{-n[\Omega(P_{\tilde{X}Z}, W_{Y|XZ})]} \right] \geq \sum_{x_i, z_i \in \mathcal{L}, j \neq i} \frac{1}{2} e^{-n \min\{\Omega(P_{\tilde{X}Z}, W_{Y|XZ}), \Omega(P_{\tilde{X}Z}, W_{Y|XZ})\}}$$

Next, observe that since $\Omega(\tilde{P}_{XZX}, W_{Y|XZ}) = \min_{V_{Y|XZ} \in \mathcal{S}(\tilde{P}_{XZ})} D(V_{Y|XZ} || W_{Y|XZ}|\tilde{P}_{XZ})$,

$$\begin{align*}
&\min \left\{ \Omega(\tilde{P}_{xzz}, W_{Y|XZ}), \Omega(P_{xzz}, W_{Y|XZ}) \right\} \\
&\leq \min \left\{ \min_{V_{Y|XZ} \in \mathcal{S}(\tilde{P}_{xzz})} D(V_{Y|XZ} || W_{Y|XZ}|\tilde{P}_{xzz}), \min_{I_V(X;Y|XZ)=0} D(V_{Y|XZ} || W_{Y|XZ}|\tilde{P}_{xzz}) \right\} \\
&= \min_{V_{Y|XZ} \in \mathcal{S}(\tilde{P}_{xzz})} D(V_{Y|XZ} || W_{Y|XZ}|\tilde{P}_{xzz}),
\end{align*}$$

where (96) holds since $D(V_{Y|XZ} || W_{Y|XZ}|\tilde{P}_{xzz}) = D(V_{Y|XZ} || W_{Y|XZ}|\tilde{P}_{xzz}) + I_V(X;Y|XZ)$, and since $\tilde{P}_{xzz} = \tilde{P}_{xzz} = \hat{P}_{xzz}$, (96) holds since $\min\{\text{min}_{S_i} f(t), \text{min}_{S_2} f(t)\} = \min_{S_i \cup S_2} f(t)$, (98) follows since $S^d(\frac{1}{2}P_{xzz} + \hat{P}_{xzz}) \subseteq S^d(\tilde{P}_{xzz}) \cup S^d(\hat{P}_{xzz})$ which follows from the linearity of the expectation; that is,

$$S^d(\tilde{P}_{xzz}) \cup S^d(\hat{P}_{xzz}) = \left\{ V_{Y|XZ} \in \mathcal{Y} \left| \mathcal{L}(x_i, \tilde{P}_{XZ}) \geq 2 \right. \right\}$$

where (93) follows by switching the roles of the summation indices and multiplying and dividing by 2, (94) follows since for positive $A, B$, we have, $A + B \geq \max\{A, B\}$, and since $f(t) = e^{-t}$ is a monotonically decreasing function.

Now, in analogy to the definition of $S^d(P_{\tilde{X}Z})$ in (89), define the following set of conditional distributions rather than joint distributions

$$S^d(P_{XZ}) \triangleq \left\{ V_{Y|XZ} : P_{XZ} \times V_{Y|XZ} \in S^d(P_{\tilde{X}Z}) \right\}.$$
We next establish a stronger result with $W_q(\hat{P}_X)$ in lieu of $W_q^{sym}(\hat{P}_X)$, but note that the proof can be shortened if we wish to prove (34) with $E_q^{sym}$ replaced by $E_q^{sym}$.  

Claim 1: For any $x_i, x_j \in L$:

$$\min_{\nu_{XYZ} \in S(\{1\}|x_i=x_j+\hat{P}_{xyz})} D(V_{X|Z}||W_{X|Z}|\hat{P}_{XZ}) \leq \min_{\nu_{XYZ} \notin W_q(\hat{P}_X): \nu_{XZ}=\hat{P}_{XZ}} D(V_{Y|XZ}||W_{Y|XZ}|\hat{P}_{XZ}),$$

(103)

Claim 1 is proved in Appendix E.

Hence, from (101) and (103), we obtain that whenever $|L(z, \hat{P}_X)| \geq 2$,

$$\Pr(error|z, \hat{P}_{XZ}) \geq \hat{P}_{XZ} \geq e^{-n} \left[ \min_{\nu_{Y|X} \in W_q(\hat{P}_X): \nu_{XZ}=\hat{P}_{XZ}} D(V_{Y|XZ}||W_{Y|XZ}|\hat{P}_{XZ}) + k_n \right],$$

(104)

Define

$$G_n \triangleq {(n+1)}^{-1} $$

(105)

We next present a lemma which enables to assess the size of the list $L(Z, \hat{P}_X)$ defined in (7).

Lemma 4: Let a codebook $C_n = \{x_i\}_{i=1}^{M_n}$ be given, let $X$ denote the random codeword (distributed uniformly over $C_n$), and let $Z$ denote the output of the channel $W_n^{n}|X$ when fed by $X$. For any $\tau \geq 0$,

$$\Pr \left( |L(Z, \hat{P}_X)| \geq e^{n\tau} | \hat{P}_{XZ} \right) \geq 1 - G_n \cdot e^{-n[R-I(\hat{P}_{XZ})-\tau]},$$

(106)

Lemma 4 is proved in Appendix F.

Note that Lemma 4, implies that for any $\hat{P}_{XZ}, \tau_n > 0$, and $\epsilon_n > 0$, such that

$$R \geq I(\hat{P}_{XZ}) + \epsilon_n + \frac{|X||Z|-1}{n} \log(n+1) + \tau_n,$$

(107)

it holds that

$$\Pr \left( |L(Z, \hat{P}_X)| < e^{n\epsilon_n} | \hat{P}_{XZ} = \hat{P}_{XZ} \right) \leq e^{-n\epsilon_n}.$$  

(108)

Consequently, denoting $b_n = (1 - e^{-n\epsilon_n})$ if $\hat{P}_{XZ}$ is a possible joint empirical distribution of a codeword and a channel output $Z$ such that (107) holds, we have for $\epsilon_n > 1/n$ that

$$\Pr(error|\hat{P}_{XZ} = \hat{P}_{XZ}) \geq \Pr(error, L(Z, \hat{P}_X) \geq e^{n\epsilon_n} | \hat{P}_{XZ} = \hat{P}_{XZ}) \geq b_n \cdot \Pr(error|\hat{P}_{XZ} = \hat{P}_{XZ}, L(Z, \hat{P}_X) \geq e^{n\epsilon_n}) \geq b_n \cdot \min_{z \in T_n(\hat{P}_X): |z| \geq e^{n\epsilon_n}} \Pr(error|\hat{P}_{XZ} = \hat{P}_{XZ}, Z = z)$$

(111)

$$\geq b_n \cdot e^{-n} \left[ \min_{\nu_{Y|X} \in W_q(\hat{P}_X): \nu_{XZ}=\hat{P}_{XZ}} D(V_{Y|XZ}||W_{Y|XZ}|\hat{P}_{XZ}) + k_n \right],$$

(112)

where the last step follows from (104).

Now, let

$$\Psi \left( P_{XZ}, W_{Y|XZ} \right) \triangleq \min_{\nu_{XYZ} \in W_q(\hat{P}_X): \nu_{XZ}=\hat{P}_{XZ}} D(V_{Y|XZ}||W_{Y|XZ}|\hat{P}_{XZ}).$$

(113)

Since (112) holds for every $\hat{P}_{XZ} \in P_n(X \times Z)$ such that $\hat{P}_X = P_n$, which is a possible empirical distribution of $X$, $Z$, and which satisfies (107), and since

$$\Pr(error) = \sum \Pr(\hat{P}_{XZ} = \hat{P}_{XZ}) \cdot \Pr(error|\hat{P}_{XZ}) \geq \sum \frac{1}{n+1} e^{-nD(\hat{P}_{XZ}||W_{X|Z}|\hat{P}_X)} \cdot \Pr(error|\hat{P}_{XZ} = \hat{P}_{XZ}),$$

(114)

where $e_n^{\epsilon} = e_n + \frac{|X||Z|-1}{n} \log(n+1) + \epsilon_n$. Denoting $\delta_n'' = \delta_n - \frac{1}{n} \log(1 - e^{-n\epsilon_n}) + \frac{|X||Z|}{n} \log(n+1)$, we get

$$-\frac{1}{n} \log \Pr(error) \leq \min_{\hat{P}_{XZ} \in P_n(X \times Z): \hat{P}_X = P_n} D(\hat{P}_{XZ}||W_{X|Z}|\hat{P}_X)$$

$$+ \Psi \left( \hat{P}_{XZ}, W_{Y|XZ} \right) + \delta_n''.$$  

(115)

The following lemma shows that the minimization over empirical conditional distributions $P_n(X \times Z)$ can be approximated by minimization over the simplex $\mathcal{P}(X \times Z)$.

Lemma 5: For any $P_n \in \mathcal{P}_n(X)$, and any $P_{Y|X} \in W_q(P_n)$, there exists $P_{XZ} \in P_n(X \times Z)$ such that $P_X = P_n$, $P_{Z|X} = P_{Y|XZ} \in W_q(P_n)$,

$$\|P_n \times P_{Z|X} - P_{XZ}\| \leq \frac{|X||Z|}{n},$$

and $P_{Z|X}(x|z) = 0 \Rightarrow P_{Z|X}(z|x) = 0$.  

(116)

Next, let $P_{XZ}^*$ be the minimizer of (113). Note that $P_{XZ} - P_{XZ}^* \| \leq \frac{|X||Z|}{n}$ implies $|D(P_{XZ}^*||W_{Y|XZ}|P_X) - D(P_{XZ}||W_{Y|XZ}|P_X)| \leq 2\frac{|X||Z|}{n} \log n + \frac{|X||Z|}{n} \log \frac{n}{w_{min}} \leq \delta_n$, where $w_{min}$ is defined in (76) and further, $|I(P_{XZ}^*) - I(P_n \times P_{Z|X})| \leq 2\frac{|X||Z|}{n} \log n + \delta_n \leq \epsilon_n$. This implies that for any $P_n \in \mathcal{P}_n(X)$,

$$\min_{P_{XZ} \in P_n(X \times Z): \hat{P}_X = P_n} D(\hat{P}_{XZ}||W_{X|Z}|\hat{P}_X) + \Psi \left( \hat{P}_{XZ}, W_{Y|XZ} \right)$$

(117)
Thus, we obtained
\[
- \frac{1}{n} \log \Pr(\text{error}) \leq \min_{P_{Y|X} \in \mathcal{W}_n(P_{1n})} D(P_{Y|X} \| W_{Y|X}|P_n) + \tau_n + \delta''.
\]
(118)

Now, since $W_{Z|XY}$ can be optimized, with the exception that (73) must hold, we can choose in (73) $W'_{Z|XY} = P_{Z|XY}$, yielding $\|W_{Z|XY} - P_{Z|XY}\| \leq \frac{\epsilon_n}{2}$. Since we also have $P_{Z|XY} \ll W_{Z|XY}$, from Pinsker's inverse inequality (see e.g. [33, Lemma 4.1]), we get $D(P_{Z|XY} \| W_{Z|XY}|P_n \times P_{X|Z}) \leq 2\epsilon_n$, and consequently
\[
- \frac{1}{n} \log \Pr(\text{error}) \leq \min_{P_{Y|X} \in \mathcal{W}_n(P_{1n})} D(P_{Y|X} \| W_{Y|X}|P_n) + \tau_n + \frac{2\epsilon_n}{|Z|}.
\]
(119)

Now, taking for example $\epsilon_n = \tilde{\epsilon}_n = \frac{\log n}{n}$, $\tau_n = \frac{1}{n}$, and recalling that $t_{n, \min} = w_{\min} / n|m|$, we obtain
\[
- \frac{1}{n} \log \Pr(\text{error}) \leq \min_{P_{Y|X} \in \mathcal{W}_n(P_{1n})} D(P_{Y|X} \| W_{Y|X}|P_n) + \epsilon_{n,b},
\]
(120)
where $\epsilon_{n,b} = 2\frac{|X|^2 |Y|^2}{n} \log n + \frac{|X||Y||Y|}{w_{\min}} \log n |Z| + 2 \cdot \frac{|X|^2 |Y|^2}{n} \log n |Z| + \frac{|X||Y||Y|+1}{n} \log (n+1) + \frac{\ln(2)}{n} - \frac{1}{n} \log (1 - \frac{1}{n}) + \frac{2}{n} |Z| + \frac{2}{n} \frac{1}{n} |Z| - 1 \log (n+1).

**VIII. PROOF OF THEOREM 1**

Let a DMC $W = W_{Y|X}$ be given. As in the proof of Theorem 2, consider another channel from $X \times Y$ to a finite set $Z$ denoted by $W_{Z|XY}$, which along with $W_{Y|X}$ constitutes a two-output channel $W_{Y,Z|X}$. Fix $P_n \in \mathcal{P}_n(X')$, and let $C_n = \{ \mathcal{L}_i \}_{i=1}^{M_n}$ be a $P_n$-constant composition codebook of rate $R$ for the channel $W_{Y|X}$. Consider the genie-aided-genie setup depicted in Fig. 1.

First, we treat the case of $q = -\infty$: note that if there exists a pair of symbols $(x_0, y_0)$ such that $q(x_0, y_0) = -\infty$ and $W(y_0|x_0) > 0$, then $x_0$ must be absent from $C_n$ for all $n$ sufficiently large, as otherwise the maximal error probability would not vanish as $n$ tends to infinity. Hence assume w.l.o.g.
\[
\forall n, P_n(x) = 0 \text{ if } y: q(x,y) = -\infty.
\]
(121)

Recall the definition of the list $L(z, \widehat{P}_{XZ})$ (see (7)). We use the shorthand notation as in (83); i.e., $L = L(z, \widehat{P}_{XZ})$, and $x_i = x_i(z, P_{XZ})$. Further recall the notion of $\Delta^2(P_{X|ZU}, P_{Y|XZ})$ and $W_n(P_X)$ defined in (12), (9), respectively, (see also (13)), and let
\[
A_q(W_{Y|XZ}, P_n) \equiv \min_{P_X: P_X = P_n} \Delta^2(P_{X|ZU}, W_{Y|XZ}) \geq 0.
\]
(122)

For any collection of codewords $\mathcal{L}' \subseteq \mathcal{L}$ such that $|\mathcal{L}'| \geq 2$, let
\[
P^{avg,\mathcal{L}'}_{XZ\bar{X}}(x, z, \bar{x}) = \frac{1}{|\mathcal{L}'|(|\mathcal{L}'| - 1)} \sum_{i,j \in \mathcal{L}' \setminus j \neq i} \widehat{P}_{x_i, x_j}(x, z, \bar{x}).
\]
(124)

**Lemma 6:** If $P_{XZ} \in A_q(W_{Y|XZ}, P_n) \cap \mathcal{P}_n(X \times Z)$, then for any subset $\mathcal{L}' \subseteq \mathcal{L}(z, \widehat{P}_{XZ})$ such that $|\mathcal{L}'| \geq 2$,
\[
\sum_{x,z,\bar{x}} \widehat{P}_{x,z,\bar{x}}(x, z, x) \cdot W(y|x, z) [q(\bar{x}, y) - q(x, y)] \geq 0,
\]
(125)
and consequently, for any pair of codewords $x_i, x_k$ in $\mathcal{L}$, either
\[
\sum_{x,z,\bar{x}} \widehat{P}_{x,z,\bar{x}}(x, z, x) \cdot W(y|x, z) [q(\bar{x}, y) - q(x, y)] \geq 0,
\]
(126)

or
\[
\sum_{x,z,\bar{x}} \widehat{P}_{x,z,\bar{x}}(x, z, x) \cdot W(y|x, z) [q(\bar{x}, y) - q(x, y)] \geq 0.
\]
(127)

**Proof:** Recall the definition of $\mathcal{P}^{avg}_{XZ\bar{X}}$ in (212), hence similar to (226),
\[
\sum_{i,j \in \mathcal{L}' \setminus j \neq i} \widehat{P}_{x_i, x_j}(x, z, \bar{x}) \cdot W(y|x, z) [q(\bar{x}, y) - q(x, y)]
\]
(128)
\[
= \sum_{x,z,\bar{x}} P^{avg,\mathcal{L}'}_{XZ\bar{X}}(x, z, \bar{x}) \cdot W(y|x, z) [q(\bar{x}, y) - q(x, y)]
\]
(129)
\[
= \frac{|\mathcal{L}'|}{|\mathcal{L}'| - 1} \sum_{x,z,\bar{x},y} P_{XZ\bar{X}}(x, z, \bar{x}) \cdot W(y|x, z) [q(\bar{x}, y) - q(x, y)]
\]
(130)
\[
\geq \min_{P_{XZ}: P_{XZ} = \widehat{P}_{XZ}} \frac{|\mathcal{L}'|}{|\mathcal{L}'| - 1} \sum_{u,z,\bar{u},\bar{z}} P_{UZ}(u,z) P_{X|UZ}(x|u, z) \times W(y|x, z) [q(\bar{x}, y) - q(x, y)]
\]
(131)
\[
\geq 0,
\]
(132)
where we abbreviate $W = W_{Y|XZ}$, (130) follows (213), (131) follows by definition of $\mathcal{P}^{avg}_{XZ\bar{X}}$ (see (212)), and the last step is by definition of $A_q(P_n, W_{Y|XZ})$. \(\square\)
The next lemma shows that for at least half of the pairs of distinct codewords, the pairwise error probability is bounded away from zero. It is based on [17, Lemma 3], which was used by Csiszár and Narayan to establish a necessary and sufficient condition for the positivity of the mismatch capacity.

**Lemma 7:** Assume (121) holds. There exists $\rho > 0$, such that for all $n$ sufficiently large, for any $P_n \in \mathcal{P}_n(X)$, any $P_{\mathcal{X}Z} \in \mathcal{A}_q(W_{|XZ}, P_n)$, and any $z \in \mathcal{T}_n(P_{\mathcal{X}Z})$ which satisfies $|\mathcal{L}(z, P_{\mathcal{X}Z})| \geq 2$, any pair of codewords $x_\ell, x_k$ in $\mathcal{L}(z, P_{\mathcal{X}Z})$ such that $z$ is a possible output of $W_{|X}$ when fed by $x_\ell$ or $x_k$ satisfies either

$$W^n_{Y|XZ} \left(q(x_k, Y) \geq q(x_\ell, Y) \mid X = x_\ell, Z = z \right) > \nu,$$

or

$$W^n_{Y|XZ} \left(q(x_k, Y) \geq q(x_\ell, Y) \mid X = x_\ell, Z = z \right) > \nu.$$  

(133)

(134)

**Proof:** From Lemma 6, we know that for any pair of codewords $x_\ell, x_k$ in $\mathcal{L}(z, P_{\mathcal{X}Z})$ we have either (126) or (127). Assume w.l.o.g. that (126) holds, then clearly,

$$W^n_{Y|XZ} \left(q(x_k, Y) \geq q(x_\ell, Y) \mid X = x_\ell, Z = z \right)$$

$$\geq W^n_{Y|XZ} \left(q(x_k, Y) \geq q(x_\ell, Y) + \sum_{x, \bar{x}, y} \hat{P}_{x_{\ell}z_{x\bar{x}}}(x, \bar{x}, y) \times W(y|x, z) \mid x = x_\ell, Z = z \right)$$

$$\triangleq W^n_{Y|XZ} \left(\sum_{i=1}^n S_i \geq 0 \mid X = x_\ell, Z = z \right),$$

where $\{S_i\}$ are the following RVs satisfying $E(S_i \mid x_\ell, z) = 0$,

$$S_{i, \ell, k} = [q(x_{i, \ell}, Y) - q(x_{i, k}, Y)]$$

$$- E[q(x_{i, \ell}, Y) - q(x_{i, k}, Y)] |Z| = z_i, X_i = x_{i, \ell}.$$  

(135)

(136)

Note also that the distribution of $S_i$ given $(x_\ell, z)$ depends only on $(x_{\ell, i}, x_i, z_i)$, and therefore, there is a finite set, denoted $\mathcal{P}$ (of size not exceeding $|Z|^2$) of conditional distributions that $S_i$ can have given $(x_\ell, z)$. Hence, it follows from [17, Lemma 3] that if the $q(\cdot)$ values are bounded, there exists a constant $\nu > 0$ such that for all $n$,

$$W^n_{Y|XZ} \left(\sum_{i=1}^n S_i \geq 0 \mid x_\ell, z \right) > \nu.$$  

(137)

(138)

It remains to treat infinite values of $E(q(x_k, Y) \mid x_\ell, z)$ and $E(q(x_\ell, Y) \mid x_\ell, z)$: Since we assume metric values $q \in \mathbb{R} \cup \{-\infty\}$, the cases of interest for their values are limited to $(c, -\infty)$, $(-\infty, c)$, $(-\infty, -\infty)$, where $c$ represents a finite constant, respectively. The case $(-\infty, c)$ yields the inequality (138) trivially.

Since we assume that $z$ is a possible output of $W^n_{|X}$ when fed by $x_\ell$ (or $x_k$), the cases $(c, -\infty)$ and $(-\infty, -\infty)$; i.e., $E(q(x_\ell, Y) \mid x_\ell, z) = -\infty$, imply that there exists a symbol $x_0$ in $x_\ell$ and a pair $(x_0, y_0)$ such that $W(y_0|x_0) > 0$ and $q(x_0, y_0) = -\infty$, in contradiction to assumption (121).

Now, since the members of $\mathcal{L}(z, P_{\mathcal{X}Z})$ are equiprobable given $(z, P_{\mathcal{X}Z})$ (see (78)), we obtain from Lemma 7 that if $|\mathcal{L}(z, P_{\mathcal{X}Z})| \geq 2$, there exists $\nu > 0$ such that

$$P(\text{error} \mid \hat{P}_{\mathcal{X}Z} = P_{\mathcal{X}Z}, Z = z)$$

$$\geq \frac{1}{2} \nu \cdot 1 \{\hat{P}_{\mathcal{X}Z} \in \mathcal{A}_q(W_{|Y|XZ}, P_n)\}.$$  

(139)

Next recall (107)-(111), stating that for any $\hat{P}_{\mathcal{X}Z}$, $\epsilon_n > 0$, and $\bar{c}_n > 0$ such that

$$R_n \geq f(\hat{P}_{\mathcal{X}Z}) + \epsilon_n + \left[|X| |Z| - 1 \right] \log(n + 1) + \bar{c}_n,$$

(140)

it holds that $Pr(\{ \mathcal{L}(Z, \hat{P}_{\mathcal{X}Z}) < e^{-\epsilon_n} \mid \hat{P}_{\mathcal{X}Z} = \hat{P}_{\mathcal{X}Z} \} \leq e^{-\epsilon_n}$, and also

$$Pr(\text{error} \mid \hat{P}_{\mathcal{X}Z} = \hat{P}_{\mathcal{X}Z})$$

$$\geq b_n \cdot \min_{z \in \mathcal{T}_n(\hat{P}_{\mathcal{X}})} Pr(\text{error} \mid \hat{P}_{\mathcal{X}Z} = \hat{P}_{\mathcal{X}Z}, Z = z),$$

(139)

(141)

where $b_n = (1 - e^{-\epsilon_n})$, and thus, we get from (139) and (141) that if (140) holds,

$$Pr(\text{error} \mid \hat{P}_{\mathcal{X}Z} = \hat{P}_{\mathcal{X}Z})$$

$$\geq b_n \cdot \frac{1}{2} \nu \cdot 1 \{\hat{P}_{\mathcal{X}Z} \in \mathcal{A}_q(W_{|Y|XZ}, P_n)\}.$$  

(142)

Next, take a vanishing sequence $c_n$ that satisfies $\lim_{n \to \infty} n[c_n - \frac{1}{n} \log(n + 1)] = \infty$; we have

$$Pr(D(\hat{P}_{\mathcal{X}Z}, \|W_{|XZ}|P_n\)) > c_n \mid X = x$$

$$= \min_{z: D(\hat{P}_{\mathcal{X}Z} \| W_{|XZ}|P_n) > c_n} \max_{z: D(\hat{P}_{\mathcal{X}Z} \| W_{|XZ}|P_n) > c_n} \left[ e^{-nD(\hat{P}_{\mathcal{X}Z} \| W_{|XZ}|P_n)\mid X = x} \right]$$

$$\leq e^{-n[c_n - \frac{1}{n} \log(n + 1)]},$$

(143)

(144)

(145)

where $G_n$ is defined in (105). Denote $f_n = c_n - \frac{1}{n} \log(n + 1)$, and $d_n = c_n + \frac{|X| |Z| - 1}{n} \log(n + 1) + \bar{c}_n$, this gives

$$Pr(\text{error}) = \sum_{\hat{P}_{\mathcal{X}Z} \in \mathcal{P}_n(X \times Z)} Pr(\text{error} \mid \hat{P}_{\mathcal{X}Z} = \hat{P}_{\mathcal{X}Z})$$

$$\geq \sum_{\hat{P}_{\mathcal{X}Z} \in \mathcal{A}_q(W_{|Y|XZ}, P_n) \cap \mathcal{P}_n(X \times Z)} Pr(\hat{P}_{\mathcal{X}Z} = \hat{P}_{\mathcal{X}Z})$$

$$\times Pr(\text{error} \mid \hat{P}_{\mathcal{X}Z} = \hat{P}_{\mathcal{X}Z})$$

$$\geq (1 - e^{-d_n}) \cdot b_n \cdot \frac{1}{2} \nu \cdot 1 \left\{ \exists \hat{P}_{\mathcal{X}Z} \in K_q(R, P_n, W_{|Y|X}) \right\}.$$  

(146)

(147)
superiority relation is denoted by $(\hat{P}_{Z|X}, \hat{q}_{ML})$, where $\hat{q}_{ML} = \log \hat{P}_{Z|X}$ is the maximum likelihood metric w.r.t. $\hat{P}_{Z|X}$ such that $(W, q) \rightarrow (\hat{P}_{Z|X}, \hat{q}_{ML})$ then
$$E^q(R, W) = E^q(R, \hat{P}_{Z|X}).$$

Note that the theorem implies that for the range of rates such that $E(R, \hat{P}_{Z|X})$ is known; e.g., at $R = 0^+$ or above the critical rate where the tangential straight line bound meets the sphere packing bound, the $E^q(R, W)$ is known as well.

Theorem 4 can be extended to yield tighter bounds for larger classes of channels that depend also on the codebook composition $\hat{P}$ using a similar approach. For example, using the set (see [21])
$$\Gamma(q, \rho, P) = \left\{ P_{Y|X} : \min_{P_{XZ}\in\Gamma(q|\rho)} \left[ g(V_{XY}) - q(V_{XY}) \right] \geq 0 \right\}.$$

X. CONCLUDING REMARKS

The new technique presented in the paper yields the tightest bounds known to date on the mismatch capacity and the reliability function with mismatched decoding for $R > 0$. One of the main contributions of our work is the derivation of bounds that are easier to compute compared to previous bounds, either by reducing the number of degrees of freedom of the parameters that are optimized in the calculation of the bounds (such as in $C_q^{sym}(W)$), or by providing a looser bound that is considerably easier to compute, $C_q^{psd}(W)$ in addition to our previous bound (62) of [21].

It would be interesting to see whether there are cases for which $C_\theta(q)$ can be strictly tighter than $C_q^{sym}(W)$ or not; another interesting question is what are the exact relations between the bounds $C_q(W)$, $C_q^{sym}(W)$, $C_q^{psd}(W)$ and $C_q^{sym}(W)$. A partial answer to this question was given by analyzing the binary input channels.

Similar questions apply also to the bounds on the reliability function.

APPENDIX

A. Proof of Proposition 1

1) Proof of $C_\theta(q) \leq C_q^{sym}(W)$: Clearly, the marginal $P_{\hat{X}ZX}$ distribution of $P_{\hat{X}ZX}$, where $\hat{X} - (U, Z) - X$ and $P_{XZU} = P_{\hat{X}ZX}$ is symmetric, and thus satisfies $P_{XZU} \in P_{sym}(X^2 \times Z)$.

Secondly, note that any distribution $P_{\hat{X}ZX}$ where $\hat{X} - (U, Z) - X$ and $P_{XZU} = P_{\hat{X}ZX}$ satisfies for any $(x, z) \in X \times Z$:
$$P_{\hat{X}|Z}(x|z) = \sum_u P_{U|Z}(u|z) \left[ P_{X|UZ}(x|u, z) \right]^2$$
where (155) follows from Jensen’s inequality. Hence, 
\[
P_{X|Z}(x|z) \geq P_{X}(z) \]
holds for any \(z\) and this yields that
\[
P_{Y|XZ}(y|x,z) \geq \frac{1}{P_{X}(z)} \sum_{x,z} P_{X}(x,z)V(x|z)P_{Y|XZ}(y|x,z)[q(\bar{x}, y) − q(x, y)] \geq 0,
\]
because the requirement
\[
\sum_{x,z} V(x|z)V(\bar{x}|z)D^q(P_{Y|XZ|x,z})_{x,\bar{x}} \geq 0
\]
for any \((V(1|z), \ldots, V(|\mathcal{X}||z))\) which is a probability vector (that has non-negative entries) is looser than the same requirement for any real vector \(V \in \mathbb{R}\), which is nothing but the definition of a p.s.d. matrix.

### B. Proof of Lemma 1

Consider a binary input DMC with \(\mathcal{X} = \{0, 1\}\), and \(|\mathcal{Y}| < \infty\), whose transition probability distribution is given by
\[
\{W(y|x)\}_{x \in \mathcal{X}, y \in \mathcal{Y}}.
\]
Let \(Z\) be a finite set, and consider the bound \(C_q^{sym}(W)\) in (25), where \(\psi_q^{sym}(P_{X})\) is defined in (19), and can be expressed as \(\psi_q^{sym}(P_{X})\) defined as
\[
\psi_q^{sym}(P_{X}) \triangleq \min_{V_{XZ} \in P_{sym}(X^2 \times Z)} \mathbb{E}_{V_{XZ}}[q(\bar{X}, Y) − q(X, Y)] \geq 0.
\]

We next solve the minimization problem of (165) explicitly. We have from symmetry of \(V\)
\[
\mathbb{E}_{V_{XZ}}[q(\bar{X}, Y) − q(X, Y)] = \sum_{x,z} P_{Z}(z)V_{X|Z}(x, \bar{x}|z)P_{Y|(X,z)}[q(\bar{x}, y) − q(x, y)]
\]
\[
= \sum_{z, (x, \bar{x}) \in \{0, 1\}^2, \bar{x} \neq x} P_{Z}(z)V_{X|Z}(x, \bar{x}|z)
\]
\[
\times \sum_{y} P(y|x, z)[q(\bar{x}, y) − q(x, y)]
\]
\[
= \frac{1}{2} \sum_{z, (x, \bar{x}) \in \{0, 1\}^2, \bar{x} \neq x} P_{Z}(z)V_{X|Z}(x, \bar{x}|z)d_q(P_{Y|X,Z}|x, z)
\]
\[
= \sum_{z} P_{Z}(z)d_q(a(z))d_q(P_{Y|X,Z}|x, z),
\]
where \(d_q(P_{Y|X,Z}|x, z)\) is defined in (48), and \(a(z) \triangleq V_{X|Z}(0, 1|z) = V_{X|Z}(1, 0|z)\). Note that the constraint \(V_{X,Z|Z}(0,0|z) \geq P_{X}(0|z)\) combined with \(V_{X,Z|Z}(0,0|z) + V_{X,Z|Z}(1,1|z) = P_{X}(0|z)\) becomes \(a(z) \leq P_{X}(0|z)P_{X}(1|z)\). Therefore, the minimizing \(a(z)\) equals
\[
d_q^{sym}(z) = \min_{P_{X|Z}, P_{X}(0|z)P_{X}(1|z)} d_q(P_{Y|X,Z}|x, z) \geq 0.
\]
\[
d_q^{opt}(z) = \left\{ \begin{array}{ll}
0 & \text{if } P_{X}(0|z)P_{X}(1|z) < 0, \\
p_{X|Z}(0|z)P_{X}(1|z) & \text{if } d_q(P_{Y|X,Z}|x, z) \geq 0.
\end{array} \right.
\]
This yields
\[
\min_{V_{X|Z} \in \mathcal{P}_{sym}(X^2 \times Z)} \mathbb{E}_{V_{X|Z}} P_{Y|X,Z} [q(\tilde{X}, Y) - q(X,Y)]
\]
\[
= \sum_z P_z(z) a_{sym}^\alpha (z) d_q(P_{Y|Z|X}, z) = \sum_{z: d_q(P_{Y|Z|X}, z) < 0} P_z(z) P_{X|Z}(0|z) P_{X|Z}(1|z) d_q(P_{Y|Z|X}, z),
\]
(171)

Now, this quantity is non-negative iff for all \(z\)
\[
d_q(P_{Y|Z|X}, z) < 0 \Rightarrow P_z(z) P_{X|Z}(0|z) P_{X|Z}(1|z) = 0.
\]
(173)

Note that the condition (173) can also be rewritten as:
\[
\forall z, P_z(z) P_{X|Z}(0|z) P_{X|Z}(1|z) d_q(P_{Y|Z|X}, z) \geq 0,
\]
(174)

which yields (50), and since
\[
P_{X|Z}(0|z) P_{X|Z}(1|z) d_q(P_{Y|Z|X}, z)
\]
\[
= \frac{1}{2} \sum_{(x, \bar{x}) \in \{(0,1)\}} P_{X|Z}(x|z) P_{X|Z}(\bar{x}|z) d_q(P_{Y|Z|X}, z)
\]
\[
= \sum_{(x, \bar{x}) \in \{(0,1)\}} P_{X|Z}(x|z) P_{X|Z}(\bar{x}|z)
\]
\[
\times \sum_y P_y|X|Z(y|x,z)\left[q(\bar{x}, y) - q(x,y)\right]
\]
\[
= \sum_{x,y} P_{X|Z}(x,y) |q(\bar{x}, y) - q(x,y)| - \sum_{x,y} P_{X|Y|Z}(x,y) |q(x,y)|,
\]
(176)

we get the equivalent expression (49).

To obtain (51), assume w.l.o.g. that the maximizing \(P_X\) is non-degenerate; i.e., \(P_X(0) \in (0, 1)\), as otherwise \(C_q(W) = 0\). In this case, the coupling between \(P_X\) and the set of the minimization is removed since
\[
P_z(z) P_{X|Z}(0|z) P_{X|Z}(1|z) d_q(P_{Y|Z|X}, z) \geq 0
\]
is equivalent to
\[
P_{X|Z}(z|0) P_{X|Z}(1|z) d_q(P_{Y|Z|X}, z) \geq 0,
\]
and thus
\[
\max_{P_X} \min_{P_{Y|X}=W} \mathbb{I}(X; Z) = \max_{P_X} \\min_{P_{Y|X}=W} \mathbb{I}(X; Z)
\]
\[
\leq \min_{P_{Y|X}=W} \mathbb{I}(X; Z)
\]
(178)

where (179) follows since \(\max_{a \in A} \min_{b \in B} f(a, b) \leq \min_{b \in B} \max_{a \in A} f(a, b)\), and the last step follows from Shannon's capacity formula.

The bound (52) follows similarly to (50), but without the constraint \(V_{X|Z}(x,z) \geq P_X^+(z|x)\). The optimal \(\alpha(z)\) of (169) in this case is given by
\[
a_{opt}^{-\alpha}(z) = \left\{ \begin{array}{ll}
0 & d_q(P_{Y|Z|X}, z) \geq 0 \\
\min\{P_{X|Z}(0|z), P_{X|Z}(1|z)\} & d_q(P_{Y|Z|X}, z) < 0 ,
\end{array} \right.
\]
(181)

yielding the condition
\[
\sum_z P_z(z) a_{opt}^{-\alpha}(z) d_q(P_{Y|Z|X}, z) = \sum_{z: d_q(P_{Y|Z|X}, z) < 0} \min\{P_{X|Z}(0|z), P_{X|Z}(1|z)\} d_q(P_{Y|Z|X}, z) \geq 0,
\]
(182)

which is equivalent to (174), proving (52).

As for (53), note that
\[
\mathcal{D}^q(P_{Y|X,Z=x=}) = \begin{pmatrix} 0 & d_q(P_{Y|Z|X}, z) \\
0 & 0 \end{pmatrix},
\]
(183)

hence in this case, \(\mathcal{D}^q(P_{Y|X,Z=x=}) \geq 0\) is equivalent to
\[
d_q(P_{Y|Z|X}, z) = 0\), which yields (53) so in this case we have \(C_{sym}^q(W) \leq C_{opt}^q(W)\).

C. Proof of Lemma 3

Consider the argument of the minimization in (90); i.e.,
\[
\begin{align*}
D(V_{Y|X,Z|X} || W_{Y|X|Z}) &= \sum_{x, \bar{x}, z} P_{X|Z}(\bar{x}, x, z) \times \sum_y V_{Y|X,Z}(y|x, z, \bar{x}) \log \frac{V_{Y|X,Z}(y|x, z, \bar{x})}{W_{Y|X|Z}(y|x, z)}
\end{align*}
\]
(184)

and recall the definition of \(S^d(P_{X|Z})\) in (95).

Clearly, if \(P_{X|Z} = \alpha\) such that \(\Omega(P_{X|Z}, W_{Y|X|Z}) = \infty\), the inequality (91) holds trivially, thus assume \(\Omega(P_{X|Z}, W_{Y|X|Z}) < \infty\), which implies that \(S^d(P_{X|Z})\) is non-empty, and that a minimizer satisfies \(V_{Y|X,Z}(y|x, z, \bar{x}) = 0\) whenever \(W(y|x, z) = 0\). Let \(V_{Y|X,Z|X} \in S^d(P_{X|Z})\) be such a minimizer.

We next show that if \(P_{X|Z} \in \mathcal{P}_n(X \times Z)\), then \(V_{Y|X,Z|X}\) can be approximated by an empirical \(V_{Y|X,Z|X}^{(n)} \in S^d(P_{X|Z})\); that is, such that \(P_{X|Z} \times V_{Y|X,Z|X}^{(n)} \in \mathcal{P}_d(X \times Z \times X)\).

To this end, we introduce two additional technical lemmas. The first lemma is obtained as a special case of Krein-Milman Theorem. It asserts that any distribution in \(\mathcal{P}(A)\) where \(A\) is finite, can be expressed as a convex combination of no more than \(2^{|A|}\) empirical distributions in \(\mathcal{P}(A)\), which are all \(1/\ell\) close to it in the \(L_{\infty}\)-sense:

**Lemma 9:** Let \(\xi = (\xi_1, \xi_2, \ldots, \xi_{|A|}) \in \mathcal{P}(A)\) be a given distribution, let \(\ell \geq 1\) be an integer, and consider the following

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
convex subset of $\mathcal{P}(A)$

$$\Pi(\xi, \ell) \triangleq \left\{ V \in \mathcal{P}(A) : \forall j \in A, \frac{[\xi_j]}{\ell} \leq V(j) \leq \frac{[\xi_j]}{\ell} \right\}.$$  

(186)

There exist $K \leq 2^{[A]}$ empirical distributions $\{P^{(i)}\}_{i=1}^K$ in $\Pi(\xi, \ell) \cap \mathcal{P}(A)$ such that any $P \in \Pi(\xi, \ell)$ (and in particular, $\xi$) can be expressed as:

$$P = \sum_{i=1}^K \alpha_i \cdot P^{(i)}$$  

(187)

for some $\{\alpha_i\}$, such that $\alpha_i \in [0, 1]$ and $\sum_{i=1}^K \alpha_i = 1$.

**Proof**: By Krein-Milman’s Theorem, $\Pi(\xi, \ell)$ is the closed convex hull of its extreme points. The extreme points of $\Pi(\xi, \ell)$ are empirical distributions of order $\ell$. There are no more than $2^{[A]}$ empirical distributions in $\Pi(\xi, \ell) \cap \mathcal{P}(A)$, because each entry $V(j)$ can only take 2 values.

The next lemma is a straightforward consequence of Lemma 9.

**Lemma 10**: Let $A, B$ be finite sets, and $\ell$ an integer. For any $Q_A \in \mathcal{P}(A)$ and $V_{B|A} \in \mathcal{P}(B|A)$, there exist empirical distributions $P_{AB}^{(j)} \in \mathcal{P}(A \times B)$, $j = 1, 2, \ldots, K$, with $K \leq |A| \cdot 2^{|B|}$, such that

$$Q_A \times V_{B|A} = \sum_{j=1}^K \beta_j P_{AB}^{(j)}$$  

(188)

and

$$\forall j, P_{AB}^{(j)} = Q_A, \|Q_A \times V_{B|A} - P_{AB}^{(j)}\| \leq \frac{|A| \cdot |B|}{\ell},$$

(189)

where $\beta_j = \sum_{i=1}^K \alpha_i V_{B|A}^{(i)}(b|a)$.

**Proof**: For each $a \in A$, we use Lemma 9 to express $V_{B|A=a}$ as a convex combination of empirical distributions,

$$V_{B|A=a}(b|a) = \sum_{i=1}^{K_a} \alpha_i V_{B|A}^{(i)}(b|a),$$  

(190)

where $K_a \leq 2^{|B|}$, and naturally, if $V_{B|A=a} = 0$, we take $V_{B|A}^{(i)}(b|a) = 0$ for all $i$, and by definition we have $V_{B|A=a} \in \mathcal{P}_{\ell}(A)$, where $\ell = \ell \cdot Q(a)$, and for all $i$, $\|V_{B|A=a} - V_{B|A}^{(i)}\| \leq \frac{|A| \cdot |B|}{\ell}$.

Now, let $j = (i_1, \ldots, i_{|A|})$ denote the index that takes $K = \sum_{x=1}^{|A|} K_a$ values, denote further

$$P_{AB}^{(j)}(a, b) = Q(a) \cdot V_{B|A}^{(i)}(b|a),$$  

(191)

thus, we obtain (188), where for $j = (i_1, \ldots, i_{|A|})$ we have $\beta_j = \alpha_j$, and clearly (189) holds.

Next, we invoke Lemma 10 with $(n, X^2 \times Z, P_{XZ\bar{X}}, Y, V_{Y|XZ\bar{X}})$ in the roles of $(\ell, A, Q_A, B, P_{B|A})$, respectively, and we let $V_{Y|XZ\bar{X}}^{(j,n)}$ denote $P_{B|A}^{(j)}$.

By this construction we have for all $j$, $\|P_{XZ\bar{X}} \times V_{Y|XZ\bar{X}}^{(n)} - P_{XZ\bar{X}} \times Y_{Y|XZ\bar{X}}\| \leq \frac{|X|^2 |Z|^{|Y|}}{n}$, and by affinity of $q(\cdot)$, we have that there must exist at least one index $j_0$ such that

$$q(V_{Y|XZ\bar{X}}^{(j_0,n)}) - q(V_{Y|XZ\bar{X}}^{(j_0,n)}) \geq q(V_{XZ\bar{X}}^{*}) - q(V_{XZ\bar{X}}^{*}) \geq 0,$$

and hence $V_{XZ\bar{X}Y}^{*} \in \mathcal{S}(P_{XZ\bar{X}})$. Recall again that the argument of the minimization is $D(Y|XZ\bar{X}) - \mathbb{E}P_{XZ\bar{X}} \times V_{Y|XZ\bar{X}} \log W_{Y|XZ}$.

Hence, denoting $A = |X|^2 |Z|^{|Y|}$, and $s_n = \frac{|X|^2 |Z|^{|Y|}}{n}$ from [27, Lemma 2.7], if follows that

$$D(Y|XZ\bar{X}) \log W_{Y|XZ} \leq -2 \log s_n + \log \frac{1}{t_{n,min}},$$  

(192)

where $t_{n,min}$ is defined (76) and this concludes the proof of (91).

**D. Proof of Corollary 3**

The proof is similar to that of Theorem 2, so it is advisable to read the latter up to (101), before reading this proof. For convenience, we introduce the following notation for a joint distribution $V_{XZ\bar{X}Y}$:

$$r_Y(q_{XZ\bar{X}Y}) = q(V_{XZ\bar{X}Y}).$$  

(193)

Denote also

$$A^* = \{V_{Y|XZ\bar{X}} : r_Y(V_{Y|XZ\bar{X}}) = 0\}.$$  

(194)

We follow the steps of the proof of Theorem 2 up to (101) where step (98) follows since

$$\mathcal{S}^{d}\left(\frac{1}{2}[\hat{P}_{xz\bar{z}z} + \hat{P}_{xz\bar{z}z}], A^*\right) \subseteq \left(\mathcal{S}^{d}(\hat{P}_{xz\bar{z}z}) \cap A^*\right) \cup \left(\mathcal{S}^{d}(\hat{P}_{xz\bar{z}z}) \cap A^*\right)$$  

(195)

is also valid for convex type-dependent metrics. To realize this, note that since $\hat{P}_{xz\bar{z}z} = \hat{P}_{xz\bar{z}z}$,

$$S^{d}(\hat{P}_{xz\bar{z}z}) \cap A^* \subseteq \left\{ V_{Y|XZ\bar{X}} : r_Y(V_{Y|XZ\bar{X}}) \geq q(V_{XZ\bar{X}}) \right\}$$  

(196)

or $r_Y(\hat{P}_{xz\bar{z}z} \times Y_{Y|XZ\bar{X}}) \geq q(V_{XZ\bar{X}})$

(197)

$$\geq \left\{ V_{Y|XZ} : r_Y(\hat{P}_{xz\bar{z}z} \times Y_{Y|XZ\bar{X}}) \geq q(V_{XZ\bar{X}}) \right\},$$  

(198)

where the last step follows from convexity of $q(P_{XZ\bar{X}})$ in $P_{Y|X}$ for fixed $P_X$.

Thus, we have (101), i.e.,

$$\Pr(error|z, \hat{P}_{XZ} = \hat{P}_{XZ}) \geq \exp \left\{ -n \max_{x_i \in \mathcal{L}, z \in \mathcal{L} \cup \mathcal{L} \setminus i} D(Y|XZ) \log \frac{1}{t_{n,min}} \right\}$$

(199)
\[
\geq \exp\left\{-n\left(\max_{P_{XZ} \in \mathcal{P}_{XZ}: P_{XZ} = \tilde{P}_{XZ}} F(P_{XZ}, W_Y | XZ) + k_n\right)\right\},
\]
where
\[
F(P_{XZ}, W_Y | XZ) \triangleq \min_{V_{XYZ} \in \mathcal{S}(P_{XZ} + P_{XZ}^*)): \mathbb{I}_V(X: Y | XZ) = 0} D(V_{XYZ} \parallel W_{Y | XZ} | \tilde{P}_{XZ}).
\]

Next, we use (109)-(111) and this gives for \(\tilde{P}_{XZ}\) which is a possible joint empirical distribution of a codeword and a channel output \(Z\) such that (107) holds, for \(\bar{\epsilon}_n > 1/n\)
\[
\Pr(\text{error} | \tilde{P}_{XZ}) \geq b_n e^{-n\max_{e \in E, x \in \mathcal{X}, z \in \mathcal{Z}} F(\tilde{P}_{XZ}, W_Y | XZ) + k_n} \geq b_n e^{-n\max_{P_{XZ} \in \mathcal{P}_{XZ}: P_{XZ} = \tilde{P}_{XZ}} F(P_{XZ}, W_Y | XZ) + k_n}.
\]

And, similar to (114)-(115) this gives
\[
-\frac{1}{n} \log \Pr(\text{error}) \leq \min_{P_{XZ} \in \mathcal{P}_{XZ}(X \times Z): P_{XZ} = P_n, \mathbb{I}(\tilde{P}_{XZ}) \leq R - \bar{\epsilon}_n''} \left(D(\tilde{P}_{XZ} | X \parallel W_{Y | X}) + \delta_n''\right),
\]
where \(\bar{\epsilon}_n'' \triangleq \epsilon_n' + \frac{|X| - 1}{n} \log(n + 1) + \tilde{\epsilon}_n\). It remains to show that the minimization over empirical conditional distributions \(\mathcal{P}_n(X \times Z)\) can be approximated by a minimization over the simplex \(\mathcal{P}(X \times Z)\), as asserted in the following lemma whose proof appears in Appendix I.

**Lemma 11:**
\[
\min_{P_{XZ} \in \mathcal{P}(X \times Z): P_{XZ} = P_n, \mathbb{I}(P_{XZ}) \leq R - \bar{\epsilon}_n''} \left(D(\tilde{P}_{XZ} | X \parallel W_{Y | X}) + \delta_n''\right) \leq \min_{P_{XZ} \in \mathcal{P}(X \times Z): P_{XZ} = P_n, \mathbb{I}(P_{XZ}) \leq R - \epsilon_n} \left(D(P_{XZ} | X \parallel W_{Y | X}) + \delta_n + \epsilon_n, n\right),
\]
where \(\epsilon_n = 2\frac{|X||Z|}{n} \log n + \frac{2|X||Y||Z|}{n} \log n + \frac{e^{t \mathbb{L}}}{n}\) and \(\bar{\epsilon}_n = 2\frac{|X||Y||Z|}{n} \log n + \frac{e^{t \mathbb{L}}}{n}\), where \(w_{\text{min}}\) is defined in (76).
Further, for any given $P_{Y|XZX}$, letting $\mathcal{T}_{XZX|Y}^C = \mathcal{T}_{XZX|Y}^C \times P_{Y|XZX}$, and $\mathcal{P}_{XZX|Y}^\text{avg} = \mathcal{P}_{XZX|Y}^\text{avg} \times P_{Y|XZX}$, one has
\[
q(\mathcal{P}_{XZX|Y}^\text{avg}) - q(\mathcal{P}_{X|Y}^\text{avg}) \geq 0 \iff q(\mathcal{T}_{XZX|Y}^C) - q(\mathcal{T}_{X|Y}^C) \geq 0
\]
(214)

Proof: Recall that all members of $\mathcal{L}'$ lie in $\mathcal{T}_n(P_{X|Z|Z})$. Next, based on Plotkin’s counting idea (similar to (28)), we obtain that for all $(x, \bar{x}, z) \in \mathcal{X}^2 \times \mathcal{Z}$ such that $x \neq \bar{x}$,
\[
|\mathcal{L}'|(|\mathcal{L}'| - 1) \cdot P_{XZ|XZ}(x, x, \bar{x}) = \sum_{i, j \neq i} \tilde{P}_{x_Xx}(x, x, \bar{x})
\]
(215)
\[
= \sum_{j \in \mathcal{L}' \cup \mathcal{L}'} \tilde{P}_{x_Xx}(x, z, \bar{x}) - \sum_{j \in \mathcal{L}'} \tilde{P}_{x_Xx}(x, z, \bar{x})
\]
(216)
\[
= \frac{1}{n} \sum_{i=1}^{n} \left[ \sum_{j \in \mathcal{L}' \cup \mathcal{L}'} \mathbb{1}_{\{x_j(t) = x, z(t) = z, \bar{x}(t) = \bar{x}\}} \right]
\]
(217)
\[
= \frac{1}{n} \sum_{i=1}^{n} \left[ \sum_{j \in \mathcal{L}'} \mathbb{1}_{\{x_j(t) = x\}} \cdot \mathbb{1}_{\{z(t) = z, \bar{x}(t) = \bar{x}\}} \right]
\]
(218)
\[
= \frac{1}{n} \sum_{i=1}^{n} \left[ \sum_{j \in \mathcal{L}'} \mathbb{1}_{\{x_j(t) = x, z(t) = z, \bar{x}(t) = \bar{x}\}} \right]
\]
(219)
\[
\Delta = |\mathcal{L}'| |\mathcal{T}_{X|Y}(t, z) \cdot \mathcal{T}_{X|Y}^C(t, x) \cdot \mathcal{T}_{X|Y}^C(t, z)|
\]
(220)
where (217) follows since $\sum_{j \in \mathcal{L}'} \tilde{P}_{x_Xx}(x, z, \bar{x}) = 0$ whenever $x \neq \bar{x}$, and (219) follows since the indicator $\mathbb{1}_{\{x_i(t) = x\}}$ does not depend on $i$. This yields (213). To prove (214), for given $P_{Y|XZX}$ denote $\mathcal{T}_{XZX|Y}^C = \mathcal{T}_{XZX|Y}^C \times P_{Y|XZX}$, and $\mathcal{P}_{XZX|Y}^\text{avg} = \mathcal{P}_{XZX|Y}^\text{avg} \times P_{Y|XZX}$, and note that from (213) it follows that
\[
q(\mathcal{P}_{X|Y}^\text{avg}) - q(\mathcal{P}_{XZX|Y}^\text{avg}) = \frac{|\mathcal{L}'|}{|\mathcal{L}'| - 1} \cdot \left( q(\mathcal{T}_{X|Y}^C) - q(\mathcal{T}_{XZX|Y}^C) \right),
\]
(221)
and thus (214) follows. \hfill \square

The following lemma shows that the random variable $T$ (uniformly distributed over $\{1, \ldots, n\}$) can be replaced by another random variable $U$ of finite alphabet cardinality that does not increase with $n$.

Lemma 13: There exists a random variable $U$ whose alphabet size is $|U| \leq |\mathcal{X}|^2 |\mathcal{Z}|$, and a joint distribution $P_{UZ} \in \mathcal{P}(U \times \mathcal{X} \times \mathcal{Z})$ such that for any $(x, \bar{x}, z)$,
\[
\sum_u P_{UZ}(u, z) P_{X|UZ}(x|u, z) P_{X|UZ}(\bar{x}|u, z) = \mathcal{T}_{XZX}(x, \bar{x}, z).
\]
(222)

There also exists a joint distribution $P_{UXZ} \in \mathcal{P}(U \times \mathcal{X} \times \mathcal{Z})$ such that for any $(x, \bar{x}, z)$ (222) holds, $Z$ is a deterministic function of $U$, and $|U| \leq |\mathcal{X}|^2 |\mathcal{Z}| + 1$.

Lemma 13 is proved in Appendix G. It remains to show that (208) holds for any $V_{Y|XZ}$ such that $V_{Y|XZ} \in \mathcal{W}_q(\hat{P}_X)$ and $V_{XZ} = \hat{P}_{XZ}$. Let $\mathcal{L} = \{x_i, z_i\}$. Obviously, $\mathcal{L} \subseteq \mathcal{L}$, and thus we can invoke (214) to obtain
\[
\frac{|\mathcal{L}|}{|\mathcal{L}| - 1} \sum_{x, \bar{x}, z, y} P_{XZ}(x, \bar{x}) V(y|x, z) \left[ q(\bar{x}, y) - q(x, y) \right]
\]
(223)
\[
\geq \min_{P_{XZ} : P_{XZ} \in \mathcal{W}_q(\hat{P}_X)} \frac{|\mathcal{L}|}{|\mathcal{L}| - 1} \sum_{u, x, z, y} P_{UZ}(u, z) P_{X|UZ}(x|u, z)
\times P_{X|UZ}(x|u, z) V(y|x, z) \left[ q(\bar{x}, y) - q(x, y) \right]
\]
(224)
\[
\geq 0,
\]
(225)
where (223) follows by definition of $P_{XZX|Y}^\text{avg}$ (224) follows from (213), (225) follows by definition of $\mathcal{T}_{XZX|Y}^C$ and from Lemma 13, and (226) follows since $V_{Y|XZ} \in \mathcal{W}_q(\hat{P}_X)$ and $V_{XZ} = \hat{P}_{XZ}$. Thus, we have shown that (208) holds and consequently also (103).

F. Proof of Lemma 4

From the law of total probability
\[
\Pr \left( |\mathcal{L}(Z, \hat{P}_{XZ})| < e^{n^r} | \hat{P}_{XZ} = \hat{P}_{XZ} \right)
\]
(227)
\[
= \frac{1}{M_n} \sum_{i=1}^{M_n} \Pr \left( |\mathcal{L}(Z, \hat{P}_{XZ})| < e^{n^r} | X = \mathfrak{z}_i, \hat{P}_{XZ} = \hat{P}_{XZ} \right)
\]
(228)
\[
= \frac{1}{M_n} \sum_{i=1}^{M_n} \left| \{ z \in T_n(\hat{P}_{Z|X}(\mathfrak{z}_i)) : |\mathcal{L}| < e^{n^r} \} \right| \frac{1}{|T_n(\hat{P}_{Z|X}(\mathfrak{z}_i))|}
\]
(229)
\[
= \frac{1}{M_n} \sum_{i=1}^{M_n} \left| \{ z \in T_n(\hat{P}_{Z|X}(\mathfrak{z}_i)) : |\mathcal{L}| < e^{n^r} \} \right| \frac{1}{|T_n(\hat{P}_{Z|X}(\mathfrak{z}_i))|}
\]
(230)
\[
\leq \frac{1}{M_n} e^{n^r} \left| T_n(\hat{P}_{Z|X}(\mathfrak{z}_i)) \right| \frac{1}{|T_n(\hat{P}_{Z|X}(\mathfrak{z}_i))|}
\]
(231)
\[
\leq (n + 1)^{|\mathcal{X}|^2 |\mathcal{Z}| - 1} \cdot e^{-n(R - l(\hat{P}_{XZ}) - l)}
\]
(232)
\[
= e^{-n(R - l(\hat{P}_{XZ}) - l) \log(n + 1)},
\]
(233)
where (229) follows since $Z$ is uniform over $T_n(\hat{P}_{Z|X}(\mathfrak{z}_i))$, (231) holds by replacing the count over codewords by a count over sequences $z$, and (232) follows by a standard bound on the size of a type-class.

G. Proof of Lemma 13

We prove that while $T$ is a RV uniformly distributed over $\{1, \ldots, n\}$, indeed the cardinality of the alphabet of the random variable $U$ can be limited without loss of generality by $|\mathcal{X}|^2 |\mathcal{Z}|$. This is done by an application of the support lemma (Caratheodory’s Theorem). Note that by Bayes’
Law $\mathcal{P}_{XZ}(x, x, z) = \sum_i \mathcal{P}_T(t) \frac{\mathcal{P}_{XZ \mid T}(z, x, t) \mathcal{P}_{Z \mid T}(z, t)}{\mathcal{P}_{Z \mid T}(z, t)}$, and thus there exists a distribution $\mathcal{V}_{U \mid XZ}$ such that the expectations (w.r.t. $U$) of the following $|U| \leq |X|^2 \cdot |Z|$ functions of $\mathcal{V}_{U \mid XZ}$:

$$\left\{ \mathcal{V}_{Z \mid U}(z, x, u) \mathcal{V}_{Z \mid U}(z, x, u) \mathcal{V}_{Z \mid U}(z, x, u) \right\}_{(z, x, u) \in \mathbb{Z} \times \mathbb{Z}^2}$$

(234)

preserve those of $\mathcal{P}_{T \mid XZ}$, i.e.,

$$\forall (z, x, \bar{x}), \sum_u \mathcal{V}_{U}(u) \mathcal{V}_{XZ \mid U}(z, x, u) \mathcal{V}_{Z \mid U}(z, x, u) = \sum_i \mathcal{P}_T(t) \mathcal{P}_{XZ \mid T}(z, x, t) \mathcal{P}_{Z \mid T}(z, t)$$

(235)

because there are in fact $|Z| \cdot |X|^2 - 1$ degrees of freedom in $\mathcal{P}_{XZ \mid \bar{x}}(z, x, \bar{x})$, it suffices to preserve only $|X|^2 \cdot |Z| - 1$ of the functions.

As for the last assertion of the lemma, note that preserving the expectation of one additional functional $H_{\beta}(Z \mid U = u) = -\sum_{z, x} \mathcal{V}_{XZ \mid U}(z, x, u) \log \sum_{z, x} \mathcal{V}_{XZ \mid U}(z, x, u)$ yields

$$\sum_u \mathcal{V}_{U}(u) H_{\beta}(Z \mid U = u) = \sum_i \mathcal{P}_T(t) H_{\beta}(Z \mid T = t),$$

(236)

since $H_{\beta}(Z \mid T) = 0$, $Z$ is also deterministic function of $U$, where the alphabet cardinality increase is 1.

H. Proof of Lemma 5

Let $P_{XZ} = P_n \times P_{Z \mid X}$, recall the definition of $\Delta^\beta(P_{XZU}, P_{Y \mid Z})$ in (12), and let

$$Q_{\alpha i}(P_{XZ}, P_{Y \mid Z}) \triangleq \min_{P_{U \mid XZ}} \Delta^\beta(P_{XZU}, P_{Y \mid Z}).$$

(237)

We use Lemma 10 to express $P_{XZ} = P_n \times P_{Z \mid X}$ as a convex combination of empirical distributions, with $(X, Z, P_{XZ}, n)$ in the roles of $(A, B, P_{AB}, t)$, respectively, to obtain

$$P_{XZ} = \sum_i \beta_i \cdot P_{XZ}^{(i)}$$

(238)

where

$$\forall i, P_{XZ}^{(i)} = P_n, \|P_{XZ} - P_{XZ}^{(i)}\| \leq \frac{|X| \cdot |Z|}{n}, \quad \text{and}$$

$$P_{Z \mid X}(z \mid x) = 0 \Rightarrow P_{XZ}^{(i)}(z \mid x) = 0.$$  

(239)

Denote by $\{P_{XZ}^{(i)}\}_{U \mid XZ}$ the minimizers corresponding to $\{P_{XZ}^{(i)}\}$ in (237), respectively; that is, those $\{P_{U \mid XZ}^{(i)}\}$ which satisfy

$$Q_{\alpha i}(P_{XZ}^{(i)}; P_{Y \mid Z}) \triangleq \Delta^\beta(P_{XZ}^{(i)} \times P_{U \mid XZ}, P_{Y \mid Z}).$$

(240)

Let $\{P_{XZ}^{(i)}\}$ be the corresponding induced probabilities (by Bayes’ Law); i.e.,

$$P_{XZ}^{(i)}(x, z) = P_{XZ}^{(i)}(u \mid x, z) \cdot P_{XZ}^{(i)}(\bar{x}, z) \cdot P_{U \mid XZ}^{(i)}(u, z),$$

(241)

Let $T$ be the RV whose distribution is $[\beta_1, \ldots, \beta_K]$, and signifies the value of “$u$” in (241), we have $P_{XZ \mid T}(x, x, z, i) = \beta_i P_{XZ}^{(i)}(x, x, z)$, and therefore

$$\max_i Q_{\alpha i}(P_{XZ}^{(i)}; P_{Y \mid Z}) \geq \sum_i \beta_i Q_{\alpha i}(P_{XZ}^{(i)}; P_{Y \mid Z}) \quad \geq \sum_i \beta_i E_{P_{XZ}^{(i)}} \left[ P_{Y \mid Z}(q(X, Y) - q(X, Y)) \right] \triangleq \Delta^\beta(P_n \times P_{Z \mid X} \times P_{U \mid XZ}, P_{Y \mid Z}) \triangleq \Delta^\beta(P_n \times P_{Z \mid X} \times P_{U \mid XZ}, P_{Y \mid Z}) \geq \Delta^\beta(P_n \times P_{Z \mid X} \times P_{U \mid XZ}, P_{Y \mid Z}) \geq 0,$$

(242)

(243)

(244)

(245)

(246)

(247)

(248)

Finally, let $i^*$ be the maximizer of (242), thus $P_{XZ}^{(i^*)}$ serves as the empirical distribution whose existence is claimed in Lemma 5.

I. Proof of Lemma 8

We continue the proof of Lemma 5 (in Appendix H) applied to $W_{Y \mid Z \mid X}$ in the role of $P_{Y \mid Z}$.

To conclude the proof of Lemma 8, we verify that $P_{XZ}^{(i^*)} \in K_q(R, P_n, W_{Y \mid Z \mid X})$ by checking that the 3 requirements in its definition (148) are met.

1) From (242)-(248) we have that $P_{XZ}^{(i^*)} \in A_q(W_{Y \mid Z \mid X}, P_n)$.

2) From (239) we have $D(P_{XZ}^{(i^*)} \mid \mid W_{Y \mid Z \mid X} \mid P_n) \leq 2^{-|X|^2 \cdot n \cdot \min_{x, z} W_{XZ}(x, z) \log n + 1}$, where $c_n \gg \frac{1}{n}$ (for example $c_n = n^{-1/2}$), for $n$ sufficiently large, we clearly have $D(P_{XZ}^{(i^*)} \mid \mid W_{Y \mid Z \mid X} \mid P_n) \leq f_n$.

3) By continuity of the mutual information, inequality (239) also implies that $\|I(P_n \times W_{Z \mid X}) - I(P_{XZ}^{(i^*)})\| \leq \frac{|X|^2 \cdot |Z| \cdot \log n}{n}$, see [27, Lemma 2.7]). Since $I(P_n \times W_{Z \mid X}) \leq R - \epsilon$, this gives $I(P_{XZ}^{(i^*)}) \leq R - \epsilon + \frac{|X|^2 \cdot |Z| \cdot \log n}{n}$. Since $d_n = \epsilon_n + \frac{|X|^2 \cdot |Z| \cdot \log n}{n} + \epsilon_n$, where both $\epsilon_n, \bar{\epsilon}_n$ are vanishing sequences, for $n$ sufficiently large we thus have $I(P_{XZ}^{(i^*)}) \leq R - d_n$.

Hence, $P_{XZ}^{(i^*)} \in K_q(R, P_n, W_{Y \mid Z \mid X})$, and this concludes the proof of Lemma 8.

J. Proof of Lemma 11

Let $P_{XZ}$ denote the minimizer of the following function

$$\min_{P_{XZ} \leq P_{XZ} \leq P_n, I(P_{XZ}) \leq R} D(P_{XZ} \mid \mid W_{Z \mid X} \mid P_X)$$

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Next, we use Lemma 10 to express $P_{XZ}^*$ as a convex combination of empirical distributions $P_{XZ}^* = \sum_i \beta_i P_{XZ}^{(i)}$, with $(n, X', P_n, Z, P_{Z|X}^*)$ in the roles of $(e, A, Q_A, B, P_{B|A})$, respectively, and we let $P_{XZ}^{(i)}$ denote $P_{B|A}^{(i)}$.

By this construction we have for all $i\leq n$, $P_{X|Z}^* - P_n \times P_{Z|X}^{(i)} \leq \frac{|X|\|Z|}{n}$. Now, for each $i$, denote by $P_{X|Z}^{(i)}$ the maximizing $\mu_{X|Z}^{(i)}$ of:

$$\max_{\mu_{X|Z}: \mu_{XZ}=\mu_{XZ}^{(i)}}, \mu_{XZ} = P_{XZ}^{(i)}$$

and since

$$P_{XZ}^* \leq \frac{|X|\|Z|}{n}$$

(250) follows since by definition, the marginal $(X,Z)$ distribution of $P_{XZ}^*$ is equal to $P_{XZ}^{(i)}$, and thus $\sum_i \beta_i P_{XZ}^{(i)}$ yields an $F(\cdot)$ value that cannot exceed the maximal possible value. Step (251) is by definition of $F(\cdot)$. Now, by affinity of the expectation, there must exist $\mu_{X|Z}$ such that the minimizing $V_{Y|XZ}$ belongs to $S^d(\frac{1}{2} P_{XZ}^{(i)} + P_{XZ}^{(i)} \frac{XZ}{XZ})$, and hence

$$\min_{V_{Y|XZ} \in S^d(\frac{1}{2} P_{XZ}^{(i)} + P_{XZ}^{(i)} \frac{XZ}{XZ})} D(V_{Y|XZ} \| W_{Y|XZ}) \sum_i \beta_i P_{XZ}^{(i)}$$

(252) holds since $D(P_{Y|XZ} \| W_{Y|XZ}) \geq 0$.

where (253) follows since by definition $\sum_i \beta_i P_{XZ}^{(i)} = P_{XZ}^*$, and since $\|P_{XZ}^* - P_{XZ}^{(i)}\| \leq \frac{|X|\|Z|}{n}$, which implies that for any $P_{Y|XZ}$, $|D(P_{Y|XZ} \| W_{Y|XZ})| - D(P_{Y|XZ} \| W_{Y|XZ})| \leq \frac{|X|\|Z|}{n}$, where $t_{n, min}$ is defined in (76). The last step follows by definition of $P_{XZ}^{(i)}$ as the maximizing $\mu_{X|Z}^{(i)}$ of:

$$\max_{\mu_{X|Z}: \mu_{XZ}=\mu_{XZ}^{(i)}}, \mu_{XZ} = P_{XZ}^{(i)}$$

Now, $P_{XZ}^* + \max_{\mu_{X|Z}: \mu_{XZ}=\mu_{XZ}^{(i)}}, \mu_{XZ} = P_{XZ}^{(i)}$

(255) holds since $D(P_{Y|XZ} \| W_{Y|XZ}) \leq 2 \frac{|X|\|Z|}{n}$, which implies $|I(P_{XZ}) - I(P_{XZ}^*)| \leq 2 \frac{|X|\|Z|}{n}$, and hence (256) follows since $P_{XZ}^{(i)}$ is an empirical distribution of order $n$, and since $I(P_{XZ}^*) \leq R$.
[13] J. Scarlett, A. Martínez, and A. Guillén i Fàbregas, “Multiuser random coding techniques for mismatched decoding,” IEEE Trans. Inf. Theory, vol. 62, no. 7, pp. 3950–3970, Jul. 2016.

[14] J. Scarlett, L. Peng, N. Merhav, A. Martínez, and A. Guillén i Fàbregas, “Expurgated random-coding ensembles: Exponents, refinements, and connections,” IEEE Trans. Inf. Theory, vol. 60, no. 8, pp. 4449–4462, Aug. 2014.

[15] J. Scarlett, A. Somekh-Baruch, A. Martínez, and A. Guillén i Fàbregas, “A counter-example to the mismatched decoding converse for binary-input discrete memoryless channels,” IEEE Trans. Inf. Theory, vol. 61, no. 10, pp. 5387–5395, Oct. 2015.

[16] J. Scarlett, A. Guillén i Fàbregas, A. Somekh-Baruch, and A. Martínez, “Information-theoretic foundations of mismatched decoding,” Found. Trends Commun. Inf. Theory, vol. 17, nos. 2–3, pp. 149–401, 2020, doi: 10.1561/0100000101.

[17] I. Csiszár and P. Narayan, “Channel capacity for a given decoding metric,” IEEE Trans. Inf. Theory, vol. 41, no. 1, pp. 35–43, Jan. 1995.

[18] A. Somekh-Baruch, “A general formula for the mismatch capacity,” IEEE Trans. Inf. Theory, vol. 61, no. 9, pp. 4554–4568, Sep. 2015.

[19] A. Somekh-Baruch, “Converse theorems for the DMC with mismatched decoding,” IEEE Trans. Inf. Theory, vol. 64, no. 9, pp. 6196–6207, Sep. 2018.

[20] E. Asadi Kangarshahi and A. Guillén i Fàbregas, “A single-letter upper bound to the mismatch capacity,” IEEE Trans. Inf. Theory, vol. 67, no. 4, pp. 2013–2033, Apr. 2021.

[21] A. Somekh-Baruch, “A single-letter upper bound on the mismatch capacity via multicast transmission,” IEEE Trans. Inf. Theory, vol. 68, no. 5, pp. 2801–2812, May 2022.

[22] A. Somekh-Baruch, “A single-letter upper bound on the mismatch capacity via a multicasting approach,” in Proc. IEEE Int. Theory Workshop (ITW), Apr. 2021, pp. 1–5.

[23] A. Somekh-Baruch, “Robust multicasting and an upper bound on the mismatch capacity of the DMC,” in Proc. IEEE Int. Symp. Inf. Theory (ISIT), Jul. 2021, pp. 1130–1135.

[24] E. Asadi Kangarshahi and A. Guillén i Fàbregas, “A sphere-packing error exponent for mismatched decoding,” 2021, arXiv:2111.04797.

[25] E. A. Kangarshahi and A. Guillén i Fàbregas, “A sphere-packing exponent for mismatched decoding,” in Proc. IEEE Int. Symp. Inf. Theory (ISIT), Jul. 2021, pp. 1136–1141.

[26] R. G. Gallager, Information Theory and Reliable Communication. New York, NY, USA: Wiley, 1968.

[27] I. Csiszár and J. Körner, Information Theory: Coding Theorems for Discrete Memoryless Systems. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[28] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, “Lower bounds to error probability for coding on discrete memoryless channels. I,” Inf. Control, vol. 10, no. 5, pp. 522–552, May 1967. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0019995867912004

[29] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, “Lower bounds to error probability for coding on discrete memoryless channels. II,” Inf. Control, vol. 10, no. 1, pp. 65–103, Feb. 1967.

[30] M. Bondaschi, A. Guillén i Fàbregas, and M. Dalai, “Mismatched decoding reliability function at zero rate,” IEEE Trans. Inf. Theory, vol. 68, no. 3, pp. 1482–1495, Mar. 2022.

[31] J. E. Prussing, “The principal minor test for semidefinite matrices,” J. Guid., Control, Dyn., vol. 9, no. 1, pp. 121–122, Jan. 1986.

[32] V. B. Balakirsky, “A converse coding theorem for mismatched decoding at the output of binary-input memoryless channels,” IEEE Trans. Inf. Theory, vol. 41, no. 6, pp. 1889–1902, Nov. 1995.

[33] F. Götze, H. Sambale, and A. Sinulis, “Higher order concentration for functions of weakly dependent random variables,” Electron. J. Probab., vol. 24, pp. 1–19, Jan. 2019, doi: 10.1214/19-EJP338.

Anelia Somekh-Baruch (Member, IEEE) received the B.Sc. degree from Tel-Aviv University, Tel-Aviv-Yafo, Israel, in 1996, and the M.Sc. and Ph.D. degrees from the Technion–Israel Institute of Technology, Haifa, Israel, in 1999 and 2003, respectively, all in electrical engineering. From 2004 to 2009, she has held a Visiting Research Associate appointments with the Electrical Engineering Department, Princeton University, Princeton, NJ, USA, and Technion. In 2009, she joined the Faculty of Engineering, Bar-Ilan University, Ramat Gan, Israel, where she is currently an Associate Professor. Her research interests include information theory, in particular Shannon theory and channel coding and topics in communication theory. Since 2020, she has been an Area Editor of the Foundations and Trends in Communications and Information Theory. She received the Tel-Aviv University Program for outstanding B.Sc. students scholarship, the Viterbi scholarship, the Rothschild Foundation scholarship for post-doctoral studies, and the Marie Curie Outgoing International Fellowship.