Properties of a Recent Upper Bound to the Mismatch Capacity

Ehsan Asadi Kangarshahi
University of Cambridge
ea460@cam.ac.uk

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

***Abstract*—We study several properties of the upper bound on the mismatch capacity problem we recently proposed. In particular, we show that the bound can be cast as a convex-concave saddlepoint problem enabling efficient computation. Moreover, as opposed to multiple achievability bound in the literature, we show that the multiletter version of this bound does not yield any gain. In addition, we show a necessary condition for the mismatch capacity to be strictly smaller than the channel capacity for binary-input channels.**

I. Introduction and Preliminaries

We consider reliable communication over a discrete memoryless channel (DMC) W with a given decoding metric [1], [2]. This problem arises when the decoder uses a suboptimal decoding rule due to limited computational resources, or imperfect channel estimation. Moreover, it is shown in [2] that important problems in information theory, like zero-error capacity of a channel can be cast as instances of mismatched decoding problem. Multiple achievability results have been reported in the literature [1]–[4] (see also [5]). These results are derived by random-coding techniques, i.e. analyzing the average probability of error of mismatched decoder over an ensemble of codebooks. On the other hand, the only single-letter converse was given in [6], where it was claimed that for binary-input DMCs, the mismatch capacity was the achievable rate derived in [3], [4]. Reference [7] provided a counterexample to this converse invalidating its claim. Multiletter converse results were proposed in [8].

We assume input and output alphabets are $\mathcal{X} = \{1, 2, \cdots, J\}$ and $\mathcal{Y} = \{1, 2, \cdots, K\}$, respectively, with $J, K < \infty$. We denote the channel transition probability by $W(k|j), k \in \mathcal{Y}, j \in \mathcal{X}$. A codebook $\mathcal{C}_n$ is defined as a set of M sequences $\mathcal{C}_n = \big\{\boldsymbol{x}(1), \boldsymbol{x}(2), \cdots, \boldsymbol{x}(M)\big\}$, where $\boldsymbol{x}(m) = \big(x_1(m), x_2(m), \cdots, x_n(m)\big) \in \mathcal{X}^n$, for $m \in \{1, 2, \cdots, M\}$. A message $m \in \{1, 2, \cdots, M\}$ is chosen equiprobably and $\boldsymbol{x}(m)$ is sent over the channel. The channel produces a noisy observation $\boldsymbol{y} = (y_1, y_2, \cdots, y_n) \in \mathcal{Y}^n$ according to $W^n(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^n W(y_i|x_i)$. Upon observing $\boldsymbol{y} \in \mathcal{Y}^n$ the decoder produces an estimate of the transmitted message $\hat{m} \in \{1, 2, \cdots, M\}$. The decoder that minimizes the error probability is the maximum-likelihood (ML) decoder, that produces the message estimate $\hat{m}$ according to

$$\hat{m} = \underset{i \in \{1,2,\cdots,M\}}{\arg\max}\ W^n\big(\boldsymbol{y}|\boldsymbol{x}(i)\big). \tag{1}$$

Rate $R > 0$ is achievable if for any $\epsilon > 0$ there exists a sequence of length-n codebooks $\{\mathcal{C}_n\}_{n=1}^{\infty}$ such that $|\mathcal{C}_n| \geq 2^{n(R-\epsilon)}$, and $\liminf_{n\to\infty} P_e(\mathcal{C}_n) = 0$. The capacity of W, denoted by $C(W)$, is defined as the largest achievable rate.

In multiple practical scenarios, it is not possible to use a decoder based on W^n and instead, the decoder produces the message estimate $\hat{m}$ as

$$\hat{m} = \underset{i \in \{1,2,\cdots,M\}}{\arg\max}\ d\big(\boldsymbol{x}(i), \boldsymbol{y}\big), \tag{2}$$

where,

$$d\big(\boldsymbol{x}(i), \boldsymbol{y}\big) = \sum_{\ell=1}^{n} d\big(x_\ell(i), y_\ell\big) \tag{3}$$

The mismatch capacity $C_d(W)$ is defined as the largest achievable rate when the decoder is (2). Recently, we have shown that $C_d(W)$ is upper bounded by the following quantity,

$$\bar{R}_d(W) = \max_{P_X} \min_{\substack{P_{Y\hat{Y}|X} \in \mathcal{M}_{\max}(d) \\ P_{Y|X} = W}} I(P_X, P_{\hat{Y}|X}) \tag{4}$$

where $I(P_X, P_{\hat{Y}|X}) \triangleq I(X; \hat{Y})$ and the set $\mathcal{M}_{\max}(d)$ is given in the following definition.

Definition 1: Let $P_{Y\hat{Y}|X}$ be a joint conditional distribution and define the set $\mathcal{S}(k_1, k_2) \triangleq \big\{i \in \mathcal{X} | i = \arg\max_{i' \in \mathcal{X}} d(i', k_2) - d(i', k_1)\big\}$. We say that $P_{Y\hat{Y}|X}$ is a maximal joint conditional distribution if for all $(j, k_1, k_2) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$,

$$P_{Y\hat{Y}|X}(k_1, k_2|j) = 0 \ \text{ if } \ j \notin \mathcal{S}(k_1, k_2). \tag{5}$$

For a given decoding metric d, we define the set of maximal joint conditional distributions to be $\mathcal{M}_{\max}(d)$.

In this paper we study some properties of the upper bound. Specifically, in Section II we show that computing our upper bound is a convex-concave saddlepoint problem and we derive the optimality KKT conditions. In Section III, we show that the multiletter version of the upper bound coincides with the single-letter one. In Section IV we derive a sufficient condition for $C_d(W) < C(W)$ for binary-input channels.

This work was supported in part by the European Research Council under Grant 725411, and by the Spanish Ministry of Economy and Competitiveness under Grant TEC2016-78434-C3-1-R.

II. Convexity Analysis

In this section, we show that the optimization (4) is a convex-concave saddlepoint problem. First we argue that the constraints induce a convex set.

Lemma 1: For any channel W and metric d, the set of joint conditional distributions $P_{Y\hat{Y}|X}$ satisfying both $P_{Y\hat{Y}|X} \in \mathcal{M}_{\max}(d)$ and $P_{Y|X} = W$, is a convex set.

Proof: Let $P_{Y\hat{Y}|X}$ and $P'_{Y\hat{Y}|X}$ both satisfy the above constraints. Now for any $0 < \lambda < 1$ we have,

$$\lambda P_{Y|X} + (1-\lambda) P'_{Y|X} = W. \tag{6}$$

In addition, if for some k_1, k_2 we have $j \notin \mathcal{S}(k_1, k_2)$, both $P_{Y\hat{Y}|X}(k_1, k_2|j)$ and $P'_{Y\hat{Y}|X}(k_1, k_2|j)$ are equal to zero, and so is any linear combination of them. Therefore,

$$\lambda P_{Y\hat{Y}|X} + (1-\lambda) P'_{Y\hat{Y}|X} \in \mathcal{M}_{\max}(d). \tag{7}$$

■

Moreover, $I(P_X, P_{\hat{Y}|X})$ is convex in terms of $P_{\hat{Y}|X}$, and concave in terms of P_X. Since $P_{\hat{Y}|X}$ is a linear function of $P_{Y\hat{Y}|X}$, we get that $I(P_X, P_{\hat{Y}|X})$ is convex in terms of $P_{Y\hat{Y}|X}$. Therefore from the minimax theorem [9] we get,

$$\bar{R}_d(W) = \max_{P_X} \min_{\substack{P_{Y\hat{Y}|X} \in \mathcal{M}_{\max}(d) \\ P_{Y|X} = W}} I(P_X, P_{\hat{Y}|X}) \tag{8}$$

$$= \min_{\substack{P_{Y\hat{Y}|X} \in \mathcal{M}_{\max}(d) \\ P_{Y|X} = W}} \max_{P_X} I(P_X, P_{\hat{Y}|X}) \tag{9}$$

$$= \min_{\substack{P_{Y\hat{Y}|X} \in \mathcal{M}_{\max}(d) \\ P_{Y|X} = W}} C(P_{\hat{Y}|X}). \tag{10}$$

The rest of this section is devoted to deriving the KKT conditions for the optimization problem in (4). Given that $I(P_X, P_{\hat{Y}|X})$ is convex in $P_{Y\hat{Y}|X}$, and concave in P_X, then the KKT conditions are sufficient for global optimality. For convenience, we define $f(P_X, P_{Y\hat{Y}|X}) \triangleq I(P_X, P_{\hat{Y}|X})$ and rewrite the optimization problem in (4) as,

$$\bar{R}_d(W) = \max_{P_X} \min_{\substack{P_{Y\hat{Y}|X} \in \mathcal{M}_{\max}(d) \\ P_{Y|X} = W}} f(P_X, P_{Y\hat{Y}|X}). \tag{11}$$

Let $\widehat{P}_X, \widehat{P}_{Y\hat{Y}|X}$ be the optimal input and joint conditional distributions in (11) and $\hat{q}_{\hat{Y}}$ be the output distribution induced by $\widehat{P}_X$ and $\widehat{P}_{\hat{Y}|X}$. Then for $\widehat{P}_X$ we have the following constraints:

$$\widehat{P}_X(j) \geq 0, \quad \forall j \in \mathcal{X} \tag{12}$$

$$\sum_{j \in \mathcal{X}} \widehat{P}_X(j) = 1. \tag{13}$$

Let $\mu_j, j = 1, 2, \cdots, J$ be the Lagrange multipliers corresponding the inequalities in (12) and ρ be the Lagrange multiplier corresponding to (13). Therefore, from stationarity we have,

$$\frac{\partial}{\partial P_X(j)} f(P_X, \widehat{P}_{Y\hat{Y}|X}) \bigg|_{P_X = \widehat{P}_X} = \rho + \mu_j \tag{14}$$

and from the complementary slackness [10] we have $\mu_j \widehat{P}_X(j) = 0$ and from the dual feasibility we have $\mu_j \geq 0$ which leads to the separation of the equations of in two cases. If $\widehat{P}_X(j) > 0$

$$\frac{\partial}{\partial P_X(j)} f(P_X, \widehat{P}_{Y\hat{Y}|X}) \bigg|_{P_X = \widehat{P}_X} = \rho, \tag{15}$$

while when $\widehat{P}_X(j) = 0$ we have

$$\frac{\partial}{\partial P_X(j)} f(P_X, \widehat{P}_{Y\hat{Y}|X}) \bigg|_{P_X = \widehat{P}_X} \leq \rho. \tag{16}$$

Note that, because there is no other constraint on μ_j, all of the KKT conditions are summarized in (16) and (15). Moreover, computing the derivatives in (15) and (16) gives

$$\frac{\partial}{\partial P_X(j)} f(P_X, \widehat{P}_{Y\hat{Y}|X}) \bigg|_{P_X = \widehat{P}_X} = \sum_{k \in \mathcal{Y}} \widehat{P}_{\hat{Y}|X}(k|j) \log \frac{\widehat{P}_{\hat{Y}|X}(k|j)}{\hat{q}_{\hat{Y}}(k)} - 1. \tag{17}$$

As for $\widehat{P}_{Y\hat{Y}|X}$, we have the following constraints. For all $j, k_1, k_2 \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$,

$$\widehat{P}_{Y\hat{Y}|X}(k_1, k_2|j) \geq 0, \tag{18}$$

$$\widehat{P}_{Y\hat{Y}|X}(k_1, k_2|j) = 0, \text{ if } j \notin \mathcal{S}(k_1, k_2) \tag{19}$$

where (18) corresponds to $\widehat{P}_{Y\hat{Y}|X}(k_1, k_2|j)$ being a distribution and (19) corresponds to $\widehat{P}_{Y\hat{Y}|X}(k_1, k_2|j) \in \mathcal{M}_{max}(d)$. Moreover from the constraint $P_{Y|X} = W$ we get for all $j, k_1 \in \mathcal{X} \times \mathcal{Y}$

$$\sum_{k_2} \widehat{P}_{Y\hat{Y}|X}(k_1, k_2|j) = W(k_1|j). \tag{20}$$

For the ease of notation, we skip the step of explicitly considering a Lagrange multiplier for (18). However, after simplification, The following KKT conditions are equivalent to the full KKT conditions considering a Lagrange multiplier for (18). Details follow similarly to the above derivation. If we use a Lagrange multiplier λ_{j,k_1} for each of the conditions in (20), we have when $\widehat{P}_{Y\hat{Y}|X}(k_1, k_2|j) > 0$

$$\frac{\partial}{\partial P_{Y\hat{Y}|X}(k_1, k_2|j)} f(\widehat{P}_X, P_{Y\hat{Y}|X}) \bigg|_{P_{Y\hat{Y}|X} = \widehat{P}_{Y\hat{Y}|X}} = \lambda_{j,k_1} \tag{21}$$

and when $\widehat{P}_{Y\hat{Y}|X}(k_1, k_2|j) = 0$ and $j \in \mathcal{S}(k_1, k_2)$ we have

$$\frac{\partial}{\partial P_{Y\hat{Y}|X}(k_1, k_2|j)} f(\widehat{P}_X, P_{Y\hat{Y}|X}) \bigg|_{P_{Y\hat{Y}|X} = \widehat{P}_{Y\hat{Y}|X}} \geq \lambda_{j,k_1}. \tag{22}$$

Explicitly computing the derivative gives

$$\frac{\partial}{\partial P_{Y\hat{Y}|X}(k_1, k_2|j)} f(\widehat{P}_X, P_{Y\hat{Y}|X}) \bigg|_{P_{Y\hat{Y}|X} = \widehat{P}_{Y\hat{Y}|X}} \tag{23}$$

$$= \widehat{P}_X(j) \log \frac{\widehat{P}_{\hat{Y}|X}(k_2|j)}{\hat{q}_{\hat{Y}}(k_2)}. \tag{24}$$

Summarizing, for the KKT optimality conditions of we get the following inequalities

1) For $\widehat{P}_X(j) > 0$,

$$\sum_{k\in\mathcal{Y}} \widehat{P}_{\hat{Y}|X}(k|j) \log \frac{\widehat{P}_{\hat{Y}|X}(k|j)}{\hat{q}_{\hat{Y}}(k)} = 1 + \rho, \tag{25}$$

2) For $\widehat{P}_X(j) = 0$,

$$\sum_{k\in\mathcal{Y}} \widehat{P}_{\hat{Y}|X}(k|j) \log \frac{\widehat{P}_{\hat{Y}|X}(k|j)}{\hat{q}_{\hat{Y}}(k)} \leq 1 + \rho, \tag{26}$$

3) For $\widehat{P}_{Y\hat{Y}|X}(k_1, k_2|j) > 0$,

$$\widehat{P}_X(j) \log \frac{\widehat{P}_{\hat{Y}|X}(k_2|j)}{\hat{q}_{\hat{Y}}(k_2)} = \lambda_{j,k_1}, \tag{27}$$

4) For $\widehat{P}_{Y\hat{Y}|X}(k_1, k_2|j) = 0$ and $j \in \mathcal{S}(k_1, k_2)$,

$$\widehat{P}_X(j) \log \frac{\widehat{P}_{\hat{Y}|X}(k_2|j)}{\hat{q}_{\hat{Y}}(k_2)} \geq \lambda_{j,k_1}. \tag{28}$$

In the next section, we employ the above KKT conditions to analyze the multiletter version of our bound.

III. Multiletter Bound

In this section, we study the multiletter extension of the bound (4). In particular, we show that the multiletter version cannot improve on the single-letter bound. We define the ℓ-letter decoding metric $d^{(\ell)} : \mathcal{X}^\ell \times \mathcal{Y}^\ell \to \mathbb{R}$ as follows

$$d^{(\ell)}\big((x_1, x_2, \cdots, x_\ell), (y_1, y_2, \cdots, y_\ell)\big) = \sum_{i=1}^{\ell} d(x_i, y_i). \tag{29}$$

This decoding metric definition is consistent with the additive decoder we have defined in (3). We denote $\boldsymbol{j} \in \mathcal{X}^\ell$ and $\boldsymbol{k} \in \mathcal{Y}^\ell$ as the ℓ-letter inputs and outputs, respectively. Let $W^{(\ell)}$ denote a DMC over input alphabet $\mathcal{X}^\ell$ and output alphabet $\mathcal{Y}^\ell$ with the channel rule $W^{(\ell)}\big((y_1, y_2, \cdots, y_\ell)|(x_1, x_2, \cdots, x_\ell)\big) = \prod_{i=1}^{\ell} W(y_i|x_i)$. Additionally, we define $P_X^{(\ell)}$ and $P_{Y\hat{Y}|X}^{(\ell)}$ accordingly

$$P_X^{(\ell)}(x_1, \ldots x_\ell) = \prod_{i=1}^{\ell} P_X(x_i) \tag{30}$$

$$P_{Y\hat{Y}|X}^{(\ell)}\big((y_1, y_2, \cdots, y_\ell), (\hat{y}_1, \hat{y}_2, \cdots, \quad \hat{y}_\ell)|(x_1, x_2, \cdots, x_\ell)\big)$$
$$= \prod_{i=1}^{\ell} P_{Y\hat{Y}|X}(y_i, \hat{y}_i|x_i) \tag{31}$$

X^ℓ and $Y^\ell, \hat{Y}^\ell$ denote random variables defined on alphabets $\mathcal{X}^\ell$, $\mathcal{Y}^\ell$ and $\mathcal{Y}^\ell$, respectively. Moreover, $\mathcal{S}^{(\ell)}(\boldsymbol{k}_1, \boldsymbol{k}_2)$ is defined as

$$\mathcal{S}^{(\ell)}(\boldsymbol{k}_1,\boldsymbol{k}_2) \triangleq \big\{\boldsymbol{i} \in \mathcal{X}^\ell \,|\, \boldsymbol{i} = \arg\max_{\boldsymbol{i}'\in\mathcal{X}^\ell} d^{(\ell)}(\boldsymbol{i}', \boldsymbol{k}_2) - d^{(\ell)}(\boldsymbol{i}', \boldsymbol{k}_1)\big\}. \tag{32}$$

In the following lemma we characterize the sets $\mathcal{S}^{(\ell)}(\boldsymbol{k}_1, \boldsymbol{k}_2)$ and relate them to $\mathcal{S}(k_{1,i}, k_{2,i}), i = 1, 2, \cdots, \ell$.

Lemma 2:] For $\boldsymbol{j} \in \mathcal{X}^\ell, \boldsymbol{k}_1 \in \mathcal{Y}^\ell, \boldsymbol{k}_2 \in \mathcal{Y}^\ell$ we have that $\boldsymbol{j} \in \mathcal{S}^{(\ell)}(\boldsymbol{k}_1, \boldsymbol{k}_2)$ if and only if for all $1 \leq i \leq \ell$ we have

$$j_i \in \mathcal{S}(k_{1,i}, k_{2,i}). \tag{33}$$

Proof: We have

$$\arg\max_{\boldsymbol{j}\in\mathcal{X}^\ell} d^{(\ell)}(\boldsymbol{j}, \boldsymbol{k}_2) - d^{(\ell)}(\boldsymbol{j}, \boldsymbol{k}_1) \tag{34}$$

$$= \arg\max_{\boldsymbol{j}\in\mathcal{X}^\ell} \sum_{i=1}^{\ell} d(j_i, k_{2,i}) - d(j_i, k_{2,i}) \tag{35}$$

$$= \arg\max_{(j_1,j_2,\cdots,j_\ell)\in\mathcal{X}^\ell} \sum_{i=1}^{\ell} d(j_i, k_{2,i}) - d(j_i, k_{2,i}) \tag{36}$$

From (36) we get that if $(j_1, j_2, \cdots, j_\ell) \in \mathcal{S}(\boldsymbol{k}_1, \boldsymbol{k}_2)$ then for all $1 \leq i \leq \ell$ we should have $j_i \in \mathcal{S}(k_{1,i}, k_{2,i})$. Therefore,

$$\mathcal{S}^{(\ell)}(\boldsymbol{k}_1, \boldsymbol{k}_2) = \mathcal{S}(k_{1,1}, k_{2,1}) \times \mathcal{S}(k_{1,2}, k_{2,2}) \times \cdots \times \mathcal{S}(k_{1,\ell}, k_{2,\ell}). \tag{37}$$

■

For the above ℓ-letter alphabets and distributions, the construction and analysis of the bound remains unchanged. Therefore, (4) remains valid for its ℓ-letter extension, which can be written as

$$\bar{R}_d^{(\ell)}(W) \triangleq \frac{1}{\ell} \bar{R}_{d^{(\ell)}}(W^{(\ell)}) \tag{38}$$

$$= \frac{1}{\ell} \max_{p_{X^\ell}} \min_{\substack{P_{Y^\ell\hat{Y}^\ell|X^\ell}\in\mathcal{M}_{\max}(d^{(\ell)}) \\ P_{Y^\ell|X^\ell}=W^{(\ell)}}} I(p_{X^\ell}, P_{Y^\ell\hat{Y}^\ell|X^\ell}). \tag{39}$$

We have the following result.

Proposition 1:

$$\bar{R}_d^{(\ell)}(W) = \bar{R}_d(W). \tag{40}$$

Proof: Given that $I(P_X, P_{\hat{Y}|X})$ is convex in $P_{Y\hat{Y}|X}$, and concave in P_X, the KKT conditions are also sufficient for global optimality. Similarly, $f(p_{X^\ell}, P_{Y^\ell\hat{Y}^\ell|X^\ell})$ is convex in terms of p_{X^ℓ} and concave in terms of $P_{Y^\ell\hat{Y}^\ell|X^\ell}$. Here we use the optimality conditions derived in the past section to show that if $P_X^*, P_{Y\hat{Y}|X}^*$ are the optimal distributions for the single-letter bound then $\widehat{P}_X^{(\ell)}, \widehat{P}_{Y\hat{Y}|X}^{(\ell)}$ defined in (30) and (31) are optimal distributions for the multiletter version. As a result, if we find a feasible pair $P_{Y^\ell\hat{Y}^\ell|X^\ell}, P_{X^\ell}$ such that when fixing $P_{Y^\ell\hat{Y}^\ell|X^\ell}$, the input distribution P_{X^ℓ} is a maximizer of $f(\cdot, P_{Y^\ell\hat{Y}^\ell|X^\ell})$, and when fixing P_{X^ℓ}, the joint conditional distribution $P_{Y^\ell\hat{Y}^\ell|X^\ell}$ is a minimizer of $f(p_{X^\ell}, \cdot)$, then the pair $(P_{Y^\ell\hat{Y}^\ell|X^\ell}, P_{X^\ell})$ is a saddlepoint.

We need to show that if $\widehat{P}_X, \widehat{P}_{Y\hat{Y}|X}$ is a saddlepoint for the single-letter case, then, $\widehat{P}_X^{(\ell)}, \widehat{P}_{Y\hat{Y}|X}^{(\ell)}$ is a saddlepoint for the multiletter bound. Based on the aforementioned argument, it is sufficient to show that $\widehat{P}_{Y\hat{Y}|X}^{(\ell)}$ is a minimizer of (39) by fixing

$\widehat{P}_X^{(\ell)}$. This is because it is known that $\frac{1}{\ell}C(\widehat{P}_{\hat{Y}|X}^{(\ell)}) = C(P_{\hat{Y}|X})$, i.e., the product distribution $P_X^{*(\ell)}$ achieves $C(\widehat{P}_{\hat{Y}|X}^{(\ell)})$.

In the following lemma we prove that by fixing $\widehat{P}_X^{(\ell)}$, then $\widehat{P}_{Y\hat{Y}|X}^{(\ell)}$ satisfies the KKT conditions and hence, it is a minimizer of (39). Before stating the result we recall that the multiletter counterparts of the single-letter KKT conditions given in (27) and (28) hold. Moreover, as in the single-letter case, the multiletter KKT conditions are sufficient for global optimality, because the function $f(\widehat{P}_X^{(\ell)}, \cdot)$ is concave. Using Lemma 3 below completes the proof. ∎

Lemma 3: Let $\widehat{P}_X, \widehat{P}_{Y\hat{Y}|X}$ be a saddlepoint for optimization problem (4). Set $P_{X^\ell} = \widehat{P}_X^{(\ell)}$. Then, the joint conditional distribution $\widehat{P}_{Y\hat{Y}|X}^{(\ell)}$ is a minimizer of

$$\min_{\substack{P_{Y^\ell\hat{Y}^\ell|X^\ell} \in \mathcal{M}_{\max}(d^{(\ell)}) \\ P_{Y^\ell|X^\ell} = W^{(\ell)}}} f\big(\widehat{P}_X^{(\ell)}, P_{Y^\ell\hat{Y}^\ell|X^\ell}\big). \tag{41}$$

Proof: We should show that by setting $P_{X^\ell} = \widehat{P}_X^{(\ell)}$, the multiletter versions of the KKT conditions (27) and (28) hold for $\widehat{P}_{Y\hat{Y}|X}^{(\ell)}$. Generalizing the conditions of (27) and (28) to the multiletter case, and setting $P_{Y^\ell\hat{Y}^\ell|X^\ell} = \widehat{P}_{Y\hat{Y}|X}^{(\ell)}$, we should show that for all $\boldsymbol{j}, \boldsymbol{k}_1 \in \mathcal{X}^\ell \times \mathcal{Y}^\ell$ there exist $\lambda_{\boldsymbol{j},\boldsymbol{k}_1}$ such that the conditions below are fulfilled. If we show this, then the Lemma is proved because these are precisely the conditions for the minimizer of (41).

i) When $\widehat{P}_{Y\hat{Y}|X}^{(\ell)}(\boldsymbol{k}_1, \boldsymbol{k}_2|\boldsymbol{j}) > 0$ we must have,

$$\frac{\partial}{\partial P_{Y^\ell\hat{Y}^\ell|X^\ell}(\boldsymbol{k}_1, \boldsymbol{k}_2|\boldsymbol{j})} f(\widehat{P}_X^{(\ell)}, P_{Y^\ell\hat{Y}^\ell|X^\ell})\Big|_{P_{Y^\ell\hat{Y}^\ell|X^\ell} = \widehat{P}_{Y\hat{Y}|X}^{(\ell)}} = \lambda_{\boldsymbol{j},\boldsymbol{k}_1}. \tag{42}$$

ii) When $\widehat{P}_{Y\hat{Y}|X}^{(\ell)}(\boldsymbol{k}_1, \boldsymbol{k}_2|\boldsymbol{j}) = 0$ and $\boldsymbol{j} \in \mathcal{S}^{(\ell)}(\boldsymbol{k}_1, \boldsymbol{k}_2)$ we must have that,

$$\frac{\partial}{\partial P_{Y^\ell\hat{Y}^\ell|X^\ell}(\boldsymbol{k}_1, \boldsymbol{k}_2|\boldsymbol{j})} f(\widehat{P}_X^{(\ell)}, P_{Y^\ell\hat{Y}^\ell|X^\ell})\Big|_{P_{Y^\ell\hat{Y}^\ell|X^\ell} = \widehat{P}_{Y\hat{Y}|X}^{(\ell)}} \geq \lambda_{\boldsymbol{j},\boldsymbol{k}_1}. \tag{43}$$

Similarly to (23), the derivative in (42) and (43) is,

$$\frac{\partial}{\partial P_{Y^\ell\hat{Y}^\ell|X^\ell}(\boldsymbol{k}_1, \boldsymbol{k}_2|\boldsymbol{j})} f(\widehat{P}_X^{(\ell)}, P_{Y^\ell\hat{Y}^\ell|X^\ell})\Big|_{P_{Y^\ell\hat{Y}^\ell|X^\ell} = \widehat{P}_{Y\hat{Y}|X}^{(\ell)}} = \widehat{P}_X^{(\ell)}(\boldsymbol{j}) \log \frac{\widehat{P}_{\hat{Y}|X}^{(\ell)}(\boldsymbol{k}_1|\boldsymbol{j})}{\hat{q}_{\hat{Y}}^{(\ell)}(\boldsymbol{k}_1)} \tag{44}$$

which, by using that $P_{Y^\ell\hat{Y}^\ell|X^\ell} = \widehat{P}_{Y\hat{Y}|X}^{(\ell)}$, $\widehat{P}_X^{(\ell)}$ and $\hat{q}_{\hat{Y}}^{(\ell)}$ are product distributions, gives,

$$\widehat{P}_X^{(\ell)}(\boldsymbol{j}) \log \frac{\widehat{P}_{\hat{Y}|X}^{(\ell)}(\boldsymbol{k}_1|\boldsymbol{j})}{\hat{q}_{\hat{Y}}^{(\ell)}(\boldsymbol{k}_1)} = \widehat{P}_X(j_1)\widehat{P}_X(j_2)\cdots\widehat{P}_X(j_\ell)\left(\sum_{i=1}^{\ell} \log \frac{\widehat{P}_{\hat{Y}|X}(k_{2,i}|j_i)}{\hat{q}_{\hat{Y}}(k_{2,i})}\right) \tag{45}$$

In order to show that there exist some coefficients $\lambda_{\boldsymbol{j},\boldsymbol{k}_1}$ satisfying both (42) and (43), we make a particular choice and show that the choice satisfies both (42) and (43). To this end, define

$$\lambda_{\boldsymbol{j},\boldsymbol{k}_1} = \begin{cases} 0 & \widehat{P}_X(\boldsymbol{j}) = 0 \\ \prod_{i=1}^{\ell} \widehat{P}_X(j_i) \left(\sum_{i=1}^{\ell} \frac{\lambda_{j_i,k_{1,i}}}{\widehat{P}_X(j_i)}\right) & \widehat{P}_X(\boldsymbol{j}) \neq 0 \end{cases} \tag{46}$$

where $\lambda_{j_i,k_{1,i}}$ is the single-letter Lagrange multiplier corresponding to j_i and $k_{1,i}$.

Now, excluding the cases where $\widehat{P}_X(j_1)\widehat{P}_X(j_2)\cdots\widehat{P}_X(j_\ell) = 0$ where from (45), (42) and (43) the KKT conditions clearly hold, we have two cases i) When $\widehat{P}_{Y\hat{Y}|X}^{(\ell)}(\boldsymbol{j}, \boldsymbol{k}_1, \boldsymbol{k}_2) > 0$, then for all $1 \leq i \leq \ell$ we must have $\widehat{P}_{Y\hat{Y}|X}(k_{1,i}, k_{2,i}|j_i) > 0$ and therefore, (27) is valid. We have to verify that this implies that (42) is also valid. Thus,

$$\begin{aligned} &\frac{\partial}{\partial P_{Y^\ell\hat{Y}^\ell|X^\ell}(\boldsymbol{k}_1, \boldsymbol{k}_2|\boldsymbol{j})} f(\widehat{P}_X, P_{Y^\ell\hat{Y}^\ell|X^\ell})\Big|_{P_{Y^\ell\hat{Y}^\ell|X^\ell} = \widehat{P}_{Y\hat{Y}|X}^{(\ell)}} \\ &= \widehat{P}_X(j_1)\widehat{P}_X(j_2)\cdots\widehat{P}_X(j_\ell)\left(\sum_{i=1}^{\ell} \log \frac{\widehat{P}_{\hat{Y}|X}(k_{2,i}|j_i)}{\hat{q}_{\hat{Y}}(k_{2,i})}\right) && (47) \\ &= \widehat{P}_X(j_1)\widehat{P}_X(j_2)\cdots\widehat{P}_X(j_\ell)\left(\sum_{i=1}^{\ell} \frac{\lambda_{j_i,k_{1,i}}}{\widehat{P}_X(j_i)}\right) && (48) \\ &= \lambda_{\boldsymbol{j},\boldsymbol{k}_1} && (49) \end{aligned}$$

where (48) holds from the single-letter optimality in (27). ii) When $\widehat{P}_{Y\hat{Y}|X}^{(\ell)}(\boldsymbol{k}_1, \boldsymbol{k}_2|\boldsymbol{j}) = 0$ and $\boldsymbol{j} \in \mathcal{S}^{(\ell)}(\boldsymbol{k}_1, \boldsymbol{k}_2)$, as a result of Lemma 2, we have that $\mathcal{S}^{(\ell)}(\boldsymbol{k}_1, \boldsymbol{k}_2)$ is a product set, i.e., for all $1 \leq i \leq \ell$,

$$j_i \in \mathcal{S}(k_{1,i}, k_{2,i}). \tag{50}$$

Moreover, either $\widehat{P}_{Y\hat{Y}|X}(k_{1,i}, k_{2,i}|j_i) > 0$ where (27) is satisfied or $\widehat{P}_{Y\hat{Y}|X}(k_{1,i}, k_{2,i}|j_i) = 0$ where (28) is satisfied. Now, with these assumptions, we should verify that (43) is valid. We have,

$$\begin{aligned} &\frac{\partial}{\partial P_{Y^\ell\hat{Y}^\ell|X^\ell}(\boldsymbol{k}_1, \boldsymbol{k}_2|\boldsymbol{j})} f(\widehat{P}_X, P_{Y^\ell\hat{Y}^\ell|X^\ell})\Big|_{P_{Y^\ell\hat{Y}^\ell|X^\ell} = \widehat{P}_{Y\hat{Y}|X}^{(\ell)}} \\ &= \widehat{P}_X(j_1)\widehat{P}_X(j_2)\cdots\widehat{P}_X(j_\ell)\left(\sum_{i=1}^{\ell} \log \frac{\widehat{P}_{\hat{Y}|X}(k_{2,i}|j_i)}{\hat{q}_{\hat{Y}}(k_{2,i})}\right) && (51) \\ &\geq \widehat{P}_X(j_1)\widehat{P}_X(j_2)\cdots\widehat{P}_X(j_\ell)\left(\sum_{i=1}^{\ell} \frac{\lambda_{j_i,k_{1,i}}}{\widehat{P}_X(j_i)}\right) && (52) \\ &= \lambda_{\boldsymbol{j},\boldsymbol{k}_1} && (53) \end{aligned}$$

where (52) is true because of the single-letter optimality in (27) and (28).

∎

IV. Binary-input Channels

In [2], the authors state that for any DMC and decoding metric $d(x,y)$, the mismatch capacity $C_d(W)$ remains unaltered for a decoder with metric $\tilde{d}(x,y) = d(x,y)+a(x)+b(y)$, where $a(x), b(y)$ are functions of the input and output, respectively. This property suggests that for binary-input channels, the mismatch capacity $C_d(W)$ is only a function of the metric difference $d(1,y) - d(2,y)$. In this section, we show a necessary condition for $C_d(W) < C(W)$ for binary-input channels based on the above observation.

Definition 2: We say that two sequences $\{\alpha_i\}_{i=1}^K$ and $\{\beta_i\}_{i=1}^K$ have the same order if for all $1 \le i_1, i_2 \le K$

$$\alpha_{i_1} \ge \alpha_{i_2} \Rightarrow \beta_{i_1} \ge \beta_{i_2}. \tag{54}$$

We have the following result.

Theorem 1: Assume that $W(k|j) > 0$, for all $j = 1, 2$, $k = 1, \dots K$. If the sequences $\big\{\log W(k|1) - \log W(k|2)\big\}_{k=1}^K$ and $\big\{d(1,k) - d(2,k)\big\}_{k=1}^K$ do not have the same order, then $\bar{R}_d(W) < C(W)$.

Proof: Without loss of generality, we assume that the sequence $\big\{d(1,k) - d(2,k)\big\}_{k=1}^K$ is non-decreasing, i.e., for $k_1 \le k_2$,

$$d(1,k_1) - d(2,k_1) \le d(1,k_2) - d(2,k_2). \tag{55}$$

This assumption simplifies the evaluation of the sets $\mathcal{S}(\cdot,\cdot)$. For $k_1 = k_2$ we have $\mathcal{S}(k_1,k_2) = \{1,2\}$. Moreover, when $k_1 < k_2$ from (55) and Definition 1, we have that $1 \in \mathcal{S}(k_1,k_2)$ and $2 \in \mathcal{S}(k_2,k_1)$.

We prove a slightly stronger result. In particular, we prove that the condition $C_d(W) = C(W)$ implies that sequences

$$\Big\{\widehat{P}_X(1)\log\frac{W(k|1)}{\hat{q}_{\hat{Y}}(k)}\Big\}_{k=1}^K,\ \Big\{-\widehat{P}_X(2)\log\frac{W(k|2)}{\hat{q}_{\hat{Y}}(k)}\Big\}_{k=1}^K \tag{56}$$

both should have the same order as the decoding metric difference sequence $\{d(1,k) - d(2,k)\}_{k=1}^K$, where recall that the notation $\widehat{P}_X$ refers to the capacity-achieving distribution of W.

Now assume that $C_d(W) = C(W)$. Therefore, $\widehat{P}_X, P_{Y\hat{Y}|X} = P_{YY|X}$ must be a saddle point of (9). As a result, the KKT conditions in (27) (28) must hold. Observe that

$$P_{YY|X}(k_1,k_2|j) = \begin{cases} W(k_1|j) & k_1 = k_2 \\ 0 & k_1 \ne k_2. \end{cases} \tag{57}$$

Therefore, combining the KKT conditions in (27) (28) we have,

1) If $k_1 = k_2$, for both $j = 1, 2$ we have

$$\widehat{P}_X(j)\log\frac{W(k_1|j)}{\hat{q}_Y(k_1)} = \lambda_{j,k_1} \tag{58}$$

2) If $k_1 < k_2$ we know $1 \in \mathcal{S}(k_1,k_2)$ and $2 \in \mathcal{S}(k_2,k_1)$

$$\widehat{P}_X(1)\log\frac{W(k_2|1)}{\widehat{q}_Y(k_2)} \ge \lambda_{1,k_1} \tag{59}$$

$$\widehat{P}_X(2)\log\frac{W(k_1|2)}{\widehat{q}_Y(k_1)} \ge \lambda_{2,k_2} \tag{60}$$

Therefore we get if $k_1 < k_2$

$$\widehat{P}_X(1)\log\frac{W(k_2|1)}{\widehat{q}_Y(k_2)} \ge \lambda_{1,k_1} = \widehat{P}_X(1)\log\frac{W(k_1|1)}{\hat{q}_Y(k_1)} \tag{61}$$

$$\widehat{P}_X(2)\log\frac{W(k_1|2)}{\widehat{q}_Y(k_1)} \ge \lambda_{2,k_2} = \widehat{P}_X(2)\log\frac{W(k_2|2)}{\hat{q}_Y(k_2)}. \tag{62}$$

Therefore we get that $\Big\{\widehat{P}_X(1)\log\frac{W(k|1)}{\hat{q}_Y(k)}\Big\}_{k=1}^K$ and $-\Big\{\widehat{P}_X(2)\log\frac{W(k|2)}{\hat{q}_Y(k)}\Big\}_{k=1}^K$ are both non-decreasing sequences and so is any linear combination of them with positive coefficients. Therefore, since

$$\begin{aligned}\log W(k|1) - \log W(k|2) = &\frac{1}{\widehat{P}_X(1)}\Big(\widehat{P}_X(1)\log\frac{W(k|1)}{\hat{q}_Y(k)}\Big) \\ &- \frac{1}{\widehat{P}_X(2)}\Big(\widehat{P}_X(2)\log\frac{W(k|2)}{\hat{q}_Y(k)}\Big)\end{aligned} \tag{63}$$

we conclude that the sequence $\{\log W(k|1) - \log W(k|2)\}_{k=1}^K$ is a non-decreasing sequence.

■

References

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

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

[3] J. Y. N. Hui, "Fundamental issues of multiple accessing," Ph.D. dissertation, Massachusetts Institute of Technology, 1983.

[4] I. Csiszár and J. Körner, "Graph decomposition: A new key to coding theorems," *IEEE Trans. Inf. Theory*, vol. 27, pp. 5–12, Jan. 1981.

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

[6] 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, 1995.

[7] J. Scarlett, A. Somekh-Baruch, A. Martinez, 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, pp. 5387–5395, Oct. 2015.

[8] A. Somekh-Baruch, "Converse theorems for the DMC with mismatched decoding," *IEEE Trans. Inf. Theory*, vol. 64, pp. 6196–6207, Sept. 2018.

[9] J. von Neumann, "Zur Theorie der Gesellschaftsspiele," *Math. Ann.*, vol. 100, pp. 295–320, 1928.

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