Mismatched Decoding: Generalized Mutual Information under Small Mismatch

Priyanka Patel
University of Cambridge
pp490@cantab.ac.uk

Francesc Molina
University of Cambridge
Universitat Politècnica de Catalunya
fm585@cam.ac.uk

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

***Abstract*—This paper investigates achievable information rates in mismatched decoding when the channel is close to the decoding rule in terms of relative entropy. We derive an approximation of the worst-case generalized mutual information as a function of the radius of a small relative entropy ball centered at the decoding metric, allowing to characterize the loss incurred due to good, yet imperfect channel estimation.**

I. Introduction and Problem Setup

Mismatched decoding is the problem that studies reliable communication employing a fixed and possibly sub-optimal metric for decoding. Mismatched decoding encompasses a number of important problems such as channel uncertainty, bit-interleaved coded modulation, finite-precision arithmetic and zero-error communication [1]. The problem is described as follows. Consider reliable transmission of M messages over a discrete memoryless channel with input X and output Y, taking values from discrete alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively. The input distribution is denoted by $Q_X(x) = \Pr[X = x]$ for all $x \in \mathcal{X}$ and the channel transition distribution is defined as $W(y|x) = \Pr[Y = y|X = x]$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. For transmission, the encoder transmits the n-symbol codeword $\boldsymbol{x}^{(m)} = (x_1^{(m)}, \ldots, x_n^{(m)})$ corresponding to message $m = 1, \ldots, M$ from the codebook $\mathcal{C}_n = \{\boldsymbol{x}^{(i)}\}_{1\leq i\leq M}$. The decoder receives $\boldsymbol{y}$ and estimates the transmitted message as

$$\widehat{m} = \underset{1\leq \bar{m}\leq M}{\operatorname{argmax}} \prod_{j=1}^{n} q\big(x_j^{(\bar{m})}, y_j\big). \tag{1}$$

When $q(x, y) = W(y|x)$, the decoder is said to be matched and coincides with maximum-likelihood decoding; in any other case, the decoder is referred to as mismatched. An error is declared when $\widehat{m} \neq m$, and the probability of error for $\mathcal{C}_n$ is defined as $p_e(\mathcal{C}_n) = \Pr[\widehat{m} \neq m]$.

A number of achievable rates for mismatched decoding have been derived in the literature [1]. When standard i.i.d. random coding is employed, the corresponding rate is the generalized mutual information (GMI) [2] given by

$$I_{\text{GMI}}(Q_X) = \sup_{s\geq 0} \ \mathbb{E}_{Q_X\times W}\left[\log \frac{q(X,Y)^s}{\mathbb{E}_{Q_X}[q(\bar{X},Y)^s|Y]}\right]. \tag{2}$$

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 PID2020-116683GB-C22. Francesc Molina is also supported by the Spanish Ministry of Universities through Margarita Salas Fellowship.

The GMI is known to be tight with respect to the ensemble of i.i.d. codes [1]. In general, we have that $I_{\text{GMI}}(Q_X) \leq C_{\text{M}}$, where C_{M} is the mismatch capacity. Although the GMI is an achievable rate for arbitrary decoding metrics $q(x, y)$, we consider the case where the decoder metric is a channel estimate $\widehat{W}(y|x)$ corresponding to the output of a channel estimator. We analyze the GMI for a mismatched decoder that uses the channel estimate $\widehat{W}(y|x)$ as if it were perfect. We impose a constraint on the level of mismatch between estimated and true channels by defining an appropriate distance measure, and find the worst-case achievable rate for small mismatch. Similarly to [3], for small mismatch between the channel estimate $\widehat{W}$ and the true channel W we require that

$$W \in \mathcal{B}(Q_X, \widehat{W}, r) = \big\{W \colon \text{D}(\widehat{W}\|W|Q_X) \leq r\big\}, \tag{3}$$

where $\mathcal{B}(Q_X, \widehat{W}, r)$ is a relative entropy ball centered at $\widehat{W}$ of radius r, assumed to be small. This definition adopts a decoder-centric perspective in which the ball is centered around the known quantity, i.e., the channel estimate employed to decode.

One of the advantages of this formulation for sufficiently small r is that we can resort to [4, eq. (1)–(4)] to express the relative entropy as function of $\theta(y|x) \triangleq W(y|x) - \widehat{W}(y|x)$ minus a non-negative term of minor relevance, as

$$\text{D}(\widehat{W}\|W|Q_X) = \frac{1}{2}\sum_{x,y} Q_X(x)\frac{\theta^2(y|x)}{\widehat{W}(y|x)} - o\bigg(\sum_{x,y} Q_X(x)\frac{\theta^2(y|x)}{\widehat{W}(y|x)}\bigg). \tag{4}$$

Without loss of generality, we adopt throughout the paper natural logarithms and information units in nats.

II. Worst-Case GMI

In this section, we derive the worst-case GMI for small mismatch. We begin by defining the mismatched information density as

$$i_s(x, y) = \log \frac{\widehat{W}(y|x)^s}{\mathbb{E}_{Q_X}[\widehat{W}(y|X)^s]}, \tag{5}$$

where $s \geq 0$, for which the GMI can therefore be written as

$$I_{\text{GMI}}(Q_X) = \sup_{s\geq 0} \ \mathbb{E}_{Q_X\times W}[i_s(X, Y)]. \tag{6}$$

The worst-case GMI is defined as

$$\underline{I}_{\text{GMI}}(Q_X,\widehat{W},r)=\min_{W\in\mathcal{B}(Q_X,\widehat{W},r)}\sup_{s\geq 0}\ \mathbb{E}_{Q_X\times W}[i_s(X,Y)] \quad (7)$$

where the minimization is over all valid conditional probability distributions W in the relative entropy ball $\mathcal{B}(Q_X,\widehat{W},r)$. Since the true channel is unknown, the worst-case GMI problem (7) finds the channel that gives the worst possible GMI. This gives an indication of the loss incurred by good (but not perfect) channel estimation.

Theorem 1. *Consider a channel estimate $\widehat{W}$ and fixed input distribution Q_X. Then, for sufficiently small $r\geq 0$, the worst-case GMI is*

$$\begin{aligned}&\underline{I}_{\text{GMI}}(Q_X,\widehat{W},r)\\&\quad=\sup_{s\geq 0}\ I_s^{\text{ML}}(Q_X,\widehat{W})-\sqrt{2r\cdot V_s(Q_X,\widehat{W})}-o(r)\end{aligned} \quad (8)$$

where the term $o(r)$ is non-negative,

$$I_s^{\text{ML}}(Q_X,\widehat{W})=\mathbb{E}_{Q_X\times\widehat{W}}\big[i_s(X,Y)\big], \quad (9)$$

and

$$V_s(Q_X,\widehat{W})=\mathbb{E}_{Q_X}\big[\text{Var}_{\widehat{W}}[i_s(X,Y)|X]\big]. \quad (10)$$

Proof. The proof of Theorem 1 is provided in Appendix A; only the main steps are outlined here. We minimize the dual expression for GMI (7) dropping the $o(\cdot)$ term in (4) as [4], thus obtaining an accurate upper bound on $\underline{I}_{\text{GMI}}$ as $r\to 0$. The convex minimization problem is vectorized and then solved using the standard Lagrangian method. □

In addition, observe that for a fixed $\widehat{W}$ the worst-case GMI is upper bounded by the mutual information between input and output achieved through estimated channel $\widehat{W}$ with input Q_X:

$$\begin{aligned}\underline{I}_{\text{GMI}}(Q_X,\widehat{W},r)&\leq\sup_{s\geq 0}\ I_s^{\text{ML}}(Q_X,\widehat{W}) &(11)\\&=I_{\text{MI}}(Q_X,\widehat{W}). &(12)\end{aligned}$$

Rates above this cannot be achieved. The bound is tight at $r=0$, in which case $W=\widehat{W}$ and $I_{\text{MI}}(Q_X,W)=I_{\text{MI}}(Q_X,\widehat{W})$.

Corollary 1. *Let the approximate worst-case GMI be*

$$\underline{\tilde{I}}_{\text{GMI}}(Q_X,\widehat{W},r)=\sup_{s\geq 0}\ I_s^{\text{ML}}(Q_X,\widehat{W})-\sqrt{2r\cdot V_s(Q_X,\widehat{W})}. \quad (13)$$

Then, the minimizing channel transition distribution is

$$\tilde{W}^*_{\text{GMI}}(y|x)=\widehat{W}(y|x)\Big(1-\sqrt{2r}\cdot\varphi(x,y,i_s)\Big) \quad (14)$$

with

$$\varphi(x,y,i_s)\triangleq\frac{i_s(x,y)-\mathbb{E}_{\widehat{W}}[i_s(x,Y)]}{\sqrt{V_s(Q_X,\widehat{W})}}. \quad (15)$$

Observe that $\tilde{W}^*_{\text{GMI}}$ is only a valid conditional probability distribution provided it is non-negative, for which the following condition on the radius of the divergence ball must hold for all $(x,y)\in\mathcal{X}\times\mathcal{Y}$, everywhere $\widehat{W}(y|x)>0$:

$$r<\frac{1}{2\varphi^2(x,y,i_s)}. \quad (16)$$

The condition is not restrictive for sufficiently small r and values of s near the optimal one. Specific examples are shown in the simulations.

Corollary 2. *The approximate worst-case GMI can be lower-bounded by setting $s=1$ as*

$$\underline{\tilde{I}}_{\text{GMI}}\geq I_{\text{MI}}(Q_X,\widehat{W})-\sqrt{2r\cdot V_1(Q_X,\widehat{W})}\ ; \quad (17)$$

a tight approximation to the worst-case GMI for sufficiently small values of r. As $r\to 0$, the penalty term shrinks until the estimated channel mutual information $I_{\text{MI}}(Q_X,\widehat{W})$ is achieved, above which rates cannot be achieved.

A. Example: Symmetric $\widehat{W}$ and Equiprobable Q_X

We derive the worst-case GMI for discrete and symmetric estimated channels $\widehat{W}$ and an equiprobable input distribution $Q_X(x)=|\mathcal{X}|^{-1}$ (where $|\mathcal{X}|$ is the cardinality of the input set). Due to the symmetry of $\widehat{W}$, previous expressions can be further simplified and expressed using one of its rows that we denote $\widehat{W}_{\text{sym}}$. The approximate worst-case GMI is given by

$$\begin{aligned}\underline{\tilde{I}}_{\text{GMI}}(Q_X,\widehat{W},r)=\sup_{s\geq 0}\Bigg\{&\log\frac{|\mathcal{X}|}{\sum_y\widehat{W}_{\text{sym}}(y)^s}-sH(\widehat{W}_{\text{sym}})\\&+\sqrt{2r\cdot V_s(Q_X,\widehat{W})}\Bigg\}\end{aligned} \quad (18)$$

with

$$\begin{aligned}V_s(Q_X,\widehat{W})&=s^2\text{Var}_{\widehat{W}_{\text{sym}}}[\log\widehat{W}_{\text{sym}}]\\&=s^2\Big(\mathbb{E}_{\widehat{W}_{\text{sym}}}[\log^2\widehat{W}_{\text{sym}}]-H^2(\widehat{W}_{\text{sym}})\Big)\end{aligned} \quad (19)$$

and where

$$H(\widehat{W}_{\text{sym}})=-\sum_y\widehat{W}_{\text{sym}}(y)\log\widehat{W}_{\text{sym}}(y) \quad (20)$$

is the entropy of the probability mass function $\widehat{W}_{\text{sym}}$. Equation (18) can be lower bounded by setting $s=1$ to yield

$$\underline{\tilde{I}}_{\text{GMI}}(Q_X,\widehat{W},r)\geq C(\widehat{W})-\sqrt{2r\cdot\text{Var}_{\widehat{W}_{\text{sym}}}[\log\widehat{W}_{\text{sym}}]} \quad (21)$$

where $C(\widehat{W})\triangleq\log|\mathcal{X}|-H(\widehat{W}_{\text{sym}})$ is the matched capacity of (symmetric) DMC $\widehat{W}$.

B. Example: Ternary-Input Ternary-Output $\widehat{W}$

We compute the *approximate worst-case* GMI $\underline{\tilde{I}}_{\text{GMI}}$ from (13) for input distribution channel estimate $\widehat{W}$ given by

$$Q_X=\begin{bmatrix}0.3 & 0.3 & 0.4\end{bmatrix} \quad (22)$$

$$\widehat{W}=\begin{bmatrix}0.85 & 0.05 & 0.1\\0.15 & 0.825 & 0.025\\0.025 & 0.1 & 0.875\end{bmatrix}. \quad (23)$$

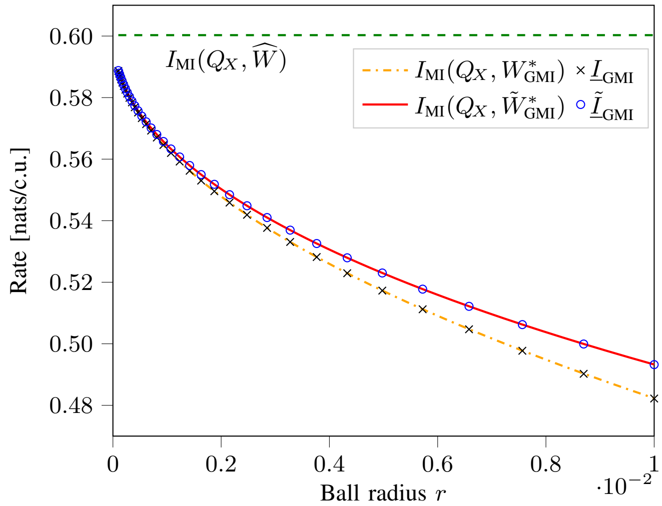

Fig. 1. Achievable rates and approximations computed for fixed Q_X in (22) and estimated channel $\widehat{W}$ in (23).

We plot the approximation in Figure 1 along with the worst-case GMI $\underline{I}_{\mathrm{GMI}}$, numerically computed from (7) using an off-the-shelf solver. To clarify notation, asterisks in superscripts indicate optimal variables.

$I_{\mathrm{MI}}(Q_X, \widehat{W})$ is shown by dashed line in Figure 1 for reference; it is achievable as $r \to 0$. We also plot the mutual information $I_{\mathrm{MI}}(Q_X, \tilde{W}^*_{\mathrm{GMI}})$ of the channel from Corollary 1, as well as that of the optimal channel W^*_{GMI} for the worst-case GMI. For $r > 0$, the curves decrease rapidly from the reference $I_{\mathrm{MI}}(Q_X, \widehat{W})$. In particular, both the true and approximated worst-case GMI decrease with an infinitely negative gradient at $r = 0$. The mutual information for the worst-case channels also exhibits similar behavior. This is because the most harmful channel in the relative entropy ball is such that it causes the GMI rate to decrease with an infinite slope. This shows that even a small mismatch can have a significant impact on the achievable transmission rates.

Our final comment is related to the validity of the approximation. In the example reported in Figure 1, $r = 0.52$ is the maximum radius limit at which $\tilde{W}^*_{\mathrm{GMI}}$ and the corresponding GMI remain positive. For all other channel estimates $\widehat{W}$ we considered, the limit is not restrictive for the range of validity of the approximation, i.e., $r < 0.01$.

Appendix A
Proof of Theorem 1

We formulate the problem based on the dual expression as

$$\underline{I}_{\mathrm{GMI}}(Q_X, \widehat{W}, r) = \min_{W \in \mathcal{B}(Q_X, \widehat{W}, r)} \sup_{s \geq 0} \mathbb{E}_{Q_X \times W}[i_s(X, Y)] \quad (24)$$

$$= \sup_{s \geq 0} \min_{W \in \mathcal{B}(Q_X, \widehat{W}, r)} \mathbb{E}_{Q_X \times W}[i_s(X, Y)] \quad (25)$$

The minimax theorem [5] is applied to switch the order of the optimizations from (24) to (25) since $\mathbb{E}_{Q_X \times W}[i_s(X, Y)]$ is convex with respect to W and concave with respect to s [1, Ch. 2.3], and constraints are convex in W.

The inner optimization problem can be vectorized and rewritten in terms of the auxiliary vector

$$\boldsymbol{\theta} = \left[\theta(y_1|x_1), \ldots, \theta(y_{|\mathcal{Y}|}|x_1), \theta(y_1|x_2), \ldots, \theta(y_{|\mathcal{Y}|}|x_{|\mathcal{X}|})\right]^T \quad (26)$$

where $\theta(y|x) = W(y|x) - \widehat{W}(y|x)$. It follows that for sufficiently small r

$$\underline{I}_s(Q_X, \widehat{W}, r) = \min_{\substack{\frac{1}{2}\boldsymbol{\theta}^T K(\widehat{W})\boldsymbol{\theta} - o(\boldsymbol{\theta}^T K(\widehat{W})\boldsymbol{\theta}) \leq r \\ \mathbf{1}_j^T \boldsymbol{\theta} = 0,\ 1 \leq j \leq |\mathcal{X}|}} \left\{ I_s^{\mathrm{ML}}(Q_X, \widehat{W}) + \boldsymbol{\theta}^T \nabla I_s \right\} \quad (27)$$

$$= \min_{\substack{\frac{1}{2}\boldsymbol{\theta}^T K(\widehat{W})\boldsymbol{\theta} \leq r \\ \mathbf{1}_j^T \boldsymbol{\theta} = 0,\ 1 \leq j \leq |\mathcal{X}|}} \left\{ I_s^{\mathrm{ML}}(Q_X, \widehat{W}) + \boldsymbol{\theta}^T \nabla I_s \right\} - o(r) \quad (28)$$

with

$$I_s^{\mathrm{ML}}(Q_X, \widehat{W}) = \mathbb{E}_{Q_X \times \widehat{W}}[i_s(X, Y)], \quad (29)$$

$$K(\widehat{W}) = \mathrm{diag}\left(\frac{Q_X(x_1)}{\widehat{W}(y_1|x_1)}, \ \ldots, \frac{Q_X(x_{|\mathcal{X}|})}{\widehat{W}(y_{|\mathcal{Y}|}|x_{|\mathcal{X}|})}\right), \quad (30)$$

$$\nabla I_s = [Q_X(x_1) i_s(x_1, y_1), \ \ldots, Q_X(x_{|\mathcal{X}|}) i_s(x_{|\mathcal{X}|}, y_{|\mathcal{Y}|})]^T, \quad (31)$$

$$\mathbf{1}_j = [0 \ \ldots \ 0 \ 1_{(1,j)} \ \ldots \ 1_{(|\mathcal{Y}|,j)} \ 0 \ \ldots \ 0]^T. \quad (32)$$

In the optimization problem, the $\mathbf{1}_j^T \boldsymbol{\theta} = 0$ constraints ensure that for every $x_j \in \mathcal{X}$, $\sum_y W(y|x_j) = 1$. To handle the error terms in the inequality constraint, it is easy to see that the constraint is dominated by the first term as $r \to 0$. Then, the problem can be equivalently written by translating the lowest-order term of the constraint to the cost function as $o(\boldsymbol{\theta}^T K(\widehat{W})\boldsymbol{\theta})$, which turns into $o(r)$ after applying the constraint. We do not explicitly impose a positivity constraint on W since a sufficiently small $r \geq 0$ exists such that the positivity of the resulting conditional distribution is guaranteed. The resulting optimization problem is convex, so the KKT conditions are necessary and sufficient [6]. The standard Lagrangian method is used to solve it.

References

[1] J. Scarlett, A. Guillén i Fàbregas, A. Somekh-Baruch, and A. Martinez, "Information-Theoretic Foundations of Mismatched Decoding," *Foundations and Trends® in Communications and Information Theory*, vol. 17, no. 2–3, pp. 149–401, 2020.

[2] G. Kaplan and S. Shamai, "Information rates and error exponents of compound channels with application to antipodal signaling in a fading environment," *AEU. Archiv für Elektronik und Übertragungstechnik*, vol. 47, no. 4, pp. 228–239, 1993.

[3] P. Boroumand and A. Guillén i Fàbregas, "Mismatched Hypothesis Testing: Error Exponent Sensitivity," *IEEE Transactions on Information Theory*, vol. 68, pp. 6738–6761, 10 2022.

[4] S. Borade and L. Zheng, "Euclidean information theory," in *2008 IEEE International Zurich Seminar on Communications*, 2008, pp. 14–17.

[5] K. Fan, "Minimax Theorems," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 39, no. 1, pp. 42–47, 1953.

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