# Error Exponents of Discrete Memoryless Channels Under Small Mismatch

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

Francesc Molina
University of Cambridge
Universitat Pompeu Fabra
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 error exponents of i.i.d. and constant-composition codes for a decoder whose decoding metric is close to the channel probability law in terms of relative entropy. We derive approximations of the worst-case achievable error exponents as functions of the radius of a small relative entropy ball centered at the decoding metric, and characterize the error terms of the underlying approximations.**

## I. Introduction and Problem Setup

The problem of mismatched decoding arises in situations when the decoder uses a fixed sub-optimal decoding rule. This problem becomes particularly relevant when channel uncertainty or system complexity constraints impede the implementation of the optimal maximum-likelihood (ML) decoder. Mismatched decoding is also connected to other important problems in information theory, such as zero-error communication and finite-precision arithmetic [1].

The problem is described as follows. Consider the reliable transmission of $M$ messages over a discrete memoryless channel taking values from discrete input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively. The channel transition distribution is defined from the single-letter distribution $W(y|x)$ for all pairs $(x,y) \in \mathcal{X} \times \mathcal{Y}$. The transmitter selects message $m$ from $\{1,\dots,M\}$ with equal probability and transmits codeword $\boldsymbol{x}^{(m)} = (x_1^{(m)},\dots,x_n^{(m)})$ from codebook $\mathcal{C}_n = \{\boldsymbol{x}^{(i)}\}_{1\le i\le M}$ over $n$ channel uses. Vector $\boldsymbol{y} = (y_1,\dots,y_n)$ is received with probability $\prod_{i=1}^n W(y_i|x_i^{(m)})$, and used to estimate the transmitted message by performing maximum metric decoding with fixed metric $q(x,y)$ as

$$\widehat{m} = \underset{1\le \bar{m}\le M}{\operatorname{argmax}} \prod_{i=1}^n q\big(x_i^{(\bar{m})}, y_i\big). \tag{1}$$

When $q(x,y) = W(y|x)$, the decoder is said to be matched and coincides with ML decoding. An error occur whenever $\widehat{m} \neq m$, and the probability of error for the chosen codebook $\mathcal{C}_n$ is defined as $p_e(\mathcal{C}_n) = \Pr[\widehat{m} \neq m]$. An error exponent $E(R)$ is said to be achievable if there exists a sequence of codes of rate $R = \frac{1}{n}\log M$ such that

$$\lim_{n\to\infty} -\frac{1}{n}\log p_e(\mathcal{C}_n) \ge E(R). \tag{2}$$

We consider random coding with two types of codebooks: i.i.d. codebooks, where codewords are independently drawn from a product distribution of i.i.d. symbols $Q_X(x)$, and constant-composition codebooks, where codewords are equiprobable on the set of sequences with a given empirical distribution. We denote by $\bar{p}_e$ the average error probability over all randomly-generated codes of a specific ensemble.

The ensemble-tight error exponents of i.i.d. and constant-composition codes are respectively defined as [1, Ch. 7.2]

$$E_{\mathrm{r}}^{\mathrm{iid}}(Q_X, R) = \max_{\rho\in[0,1]} E_0^{\mathrm{iid}}(Q_X,\rho) - \rho R \tag{3}$$

$$E_{\mathrm{r}}^{\mathrm{cc}}(Q_X, R) = \max_{\rho\in[0,1]} E_0^{\mathrm{cc}}(Q_X,\rho) - \rho R, \tag{4}$$

with respective Gallager functions

$$E_0^{\mathrm{iid}}(Q_X,\rho) = \sup_{s\ge 0} \; -\log \mathbb{E}_{Q_X\times W}[\varepsilon_{s,0,\rho}(X,Y)] \tag{5}$$

$$E_0^{\mathrm{cc}}(Q_X,\rho) = \sup_{s\ge 0, a(\cdot)} -\mathbb{E}_{Q_X}\big[\log \mathbb{E}_W[\varepsilon_{s,a,\rho}(X,Y)|X]\big] \tag{6}$$

where

$$\varepsilon_{s,a,\rho}(x,y) \triangleq \left(\frac{\mathbb{E}_{Q_X}[q(X,y)^s e^{a(X)}]}{q(x,y)^s e^{a(x)}}\right)^{\rho}. \tag{7}$$

When the optimal decoder $q(x,y) = W(y|x)$ is used instead, expressions for the matched Gallager functions are recovered by setting $s = \frac{1}{1+\rho}$: in particular, Gallager $E_0$ function for i.i.d. codes [2], and the $E_0$ function for constant-composition codes derived by Csiszár [3, Ch. 10] and expressed in dual form by Poltyrev [4]. Observe that by Jensen's inequality we have that $E_{\mathrm{r}}^{\mathrm{iid}}(Q_X,R) \le E_{\mathrm{r}}^{\mathrm{cc}}(Q_X,R)$ for a fixed input distribution $Q_X$ and any pair $(W,q)$ [5, Ch. 2].

In this work, we consider that a channel estimate $\widehat{W}(y|x)$ is available from the output of a channel estimator. The channel estimate is assumed to be close to the channel probability law and is used as decoding metric $q(x,y)$. We characterize the level of mismatch between estimated and true channels by the relative entropy, and find the worst-case error exponents of i.i.d. and constant-composition codes when the former relative entropy is small. We assume natural logarithms throughout.

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.

## II. Worst-Case Error Exponents

In this section, we derive the worst-case error exponents under i.i.d. and constant-composition random coding for small mismatch. Similarly to [6], for small mismatch between the channel estimate $\widehat{W}$ and the true channel $W$ we require that

$$W \in \mathcal{B}(Q_X, \widehat{W}, r) = \big\{W\colon D(\widehat{W}\|W|Q_X) \le r\big\}, \tag{8}$$

where $\mathcal{B}(Q_X, \widehat{W}, r)$ is a relative entropy ball centered at $\widehat{W}$ of small radius $r$. This definition adopts a decoder-centric perspective in which $\mathcal{B}$ is centered at the known quantity, the channel estimate and decoding metric $\widehat{W}$. Moreover, since we are interested in the small radius regime, we follow [7, eq. (1)–(4)] and expand the relative entropy as

$$D(\widehat{W}\|W|Q_X) = \frac{1}{2}\sum_{x,y} Q_X(x)\frac{\theta^2(y|x)}{\widehat{W}(y|x)} + o\bigg(\sum_{x,y} Q_X(x)\frac{\theta^2(y|x)}{\widehat{W}(y|x)}\bigg) \tag{9}$$

where $\theta(y|x) \triangleq W(y|x) - \widehat{W}(y|x)$.

We define the worst-case mismatched random coding error exponents as

$$\underline{E}_{\mathrm{r}}^{\mathrm{iid}}(Q_X, \widehat{W}, R, r) = \min_{W\in\mathcal{B}} \max_{\rho\in[0,1]} E_0^{\mathrm{iid}}(Q_X, \widehat{W}, W, \rho) - \rho R \tag{10}$$

$$\underline{E}_{\mathrm{r}}^{\mathrm{cc}}(Q_X, \widehat{W}, R, r) = \min_{W\in\mathcal{B}} \max_{\rho\in[0,1]} E_0^{\mathrm{cc}}(Q_X, \widehat{W}, W, \rho) - \rho R, \tag{11}$$

where $E_0^{\mathrm{iid}}, E_0^{\mathrm{cc}}$ are defined in (5)–(6) with $\varepsilon_{s,a,\rho}$ as in (7) with $q = \widehat{W}$.

**Lemma 1.** *Finding the worst-case mismatched random coding error exponent is equivalent to first finding its corresponding worst-case Gallager function, i.e.,*

$$\underline{E}_{\mathrm{r}}(Q_X, \widehat{W}, R, r) = \max_{\rho\in[0,1]} \underline{E}_0(Q_X, \widehat{W}, \rho, r) - \rho R \tag{12}$$

*where*

$$\underline{E}_0(Q_X, \widehat{W}, \rho, r) = \min_{W\in\mathcal{B}} E_0(Q_X, \widehat{W}, W, \rho). \tag{13}$$

*Proof.* The minimax theorem [8] is applied to swap the order of the optimizations. Fixing $(Q_X, R)$, and noticing that $\rho R$ is independent of $W$, the problem is then equivalent to minimizing $E_0$ over $W$ prior to maximizing over $\rho$. □

We thus focus on deriving the worst-case mismatched Gallager functions for i.i.d. and constant-composition codes. We present them separately as they require different techniques.

### A. I.i.d. Random Coding

We derive the worst-case mismatched Gallager function for i.i.d. random coding by solving

$$\underline{E}_0^{\mathrm{iid}}(Q_X, \widehat{W}, r) = \min_{W\in\mathcal{B}} \sup_{s\ge 0} -\log \mathbb{E}_{Q_X\times W}[\varepsilon_{s,0,\rho}(X,Y)]. \tag{14}$$

**Theorem 1.** *Consider a channel estimate $\widehat{W}$ and fixed input distribution $Q_X$. Then, for sufficiently small $r \ge 0$, the worst-case mismatched i.i.d. Gallager function can be expressed as*

$$\underline{E}_0^{\mathrm{iid}}(Q_X, \widehat{W}, \rho, r) = \sup_{s\ge 0}\bigg\{-\log\Big(e^{-E_{s,\rho}^{\mathrm{ML}}(Q_X,\widehat{W})} + \sqrt{2r\cdot V(Q_X, \widehat{W}, \varepsilon_{s,0,\rho})}\Big) + o(r)\bigg\} \tag{15}$$

*where*

$$E_{s,\rho}^{\mathrm{ML}}(Q_X, \widehat{W}) = -\log\sum_y \Big(\sum_x Q_X(x)\widehat{W}(y|x)^{1-s\rho}\Big) \cdot\Big(\sum_x Q_X(x)\widehat{W}(y|x)^{s}\Big)^{\rho}, \tag{16}$$

$$V(Q_X, \widehat{W}, \varepsilon_{s,0,\rho}) = \mathbb{E}_{Q_X}\big[\mathrm{Var}_{\widehat{W}}[\varepsilon_{s,0,\rho}(X,Y)|X]\big]. \tag{17}$$

*Proof.* See Appendix A for a sketch. □

**Corollary 1.1.** *For sufficiently small $r$, the worst-case mismatched i.i.d. Gallager function is expanded as*

$$\underline{E}_0^{\mathrm{iid}}(Q_X, \widehat{W}, \rho, r) = \sup_{s\ge 0}\bigg\{E_{s,\rho}^{\mathrm{ML}}(Q_X, \widehat{W}) - e^{E_{s,\rho}^{\mathrm{ML}}(Q_X,\widehat{W})}\sqrt{2r\cdot V(Q_X, \widehat{W}, \varepsilon_{s,0,\rho})} + O(r)\bigg\}. \tag{18}$$

The expansion incurs an error of order $O(r)$, which makes the error term $o(r)$ from the approximation of relative entropy negligible.

**Corollary 1.2.** *Let the approximate worst-case mismatched i.i.d. Gallager function be given by*

$$\underline{\tilde{E}}_0^{\mathrm{iid}}(Q_X, \widehat{W}, \rho, r) = \sup_{s\ge 0}\bigg\{-\log\Big(e^{-E_{s,\rho}^{\mathrm{ML}}(Q_X,\widehat{W})} + \sqrt{2r\cdot V(Q_X, \widehat{W}, \varepsilon_{s,0,\rho})}\Big)\bigg\}. \tag{19}$$

*The optimal (minimizing) conditional channel distribution is*

$$\tilde{W}_{\mathrm{iid}}^{*}(y|x) = \widehat{W}(y|x)\Big(1 + \sqrt{2r}\cdot\varphi_{\mathrm{iid}}(x,y,\varepsilon_{s,0,\rho})\Big) \tag{20}$$

*where we have defined*

$$\varphi_{\mathrm{iid}}(x,y,\varepsilon_{s,0,\rho}) \triangleq \frac{\varepsilon_{s,0,\rho}(x,y) - \mathbb{E}_{\widehat{W}}[\varepsilon_{s,0,\rho}(x,Y)]}{\sqrt{V(Q_X, \widehat{W}, \varepsilon_{s,0,\rho})}}. \tag{21}$$

The non-negativity of $\tilde{W}_{\mathrm{iid}}^{*}$ is guaranteed by satisfying the following condition on the ball radius for all $(x,y) \in \mathcal{X}\times\mathcal{Y}$, everywhere $\widehat{W}(y|x) > 0$:

$$r < \frac{1}{2\varphi_{\mathrm{iid}}^2(x,y,\varepsilon_{s,0,\rho})}. \tag{22}$$

### B. Constant-Composition Random Coding

We now derive the worst-case mismatched Gallager function for constant-composition random coding by solving

$$\begin{aligned}
&\underline{E}_0^{\mathrm{cc}}(Q_X, \widehat{W}, r)\\
&\quad = \min_{W\in\mathcal{B}} \sup_{s\geq 0, a(\cdot)} -\mathbb{E}_{Q_X}\big[\log \mathbb{E}_W[\varepsilon_{s,a,\rho}(X,Y)|X]\big].
\end{aligned} \tag{23}$$

This optimization problem is more intricate than the i.i.d. case, since the expectation over the input distribution is outside the logarithm. In fact, the solution has no explicit form. Instead, we perform a Taylor expansion of the objective function around $\widehat{W}$ and then address the resulting problem. This gives

$$\begin{aligned}
&\underline{E}_0^{\mathrm{cc}}(Q_X, \widehat{W}, r)\\
&\quad = \min_{W\in\mathcal{B}} \sup_{s\geq 0, a(\cdot)} \bigg\{ -\mathbb{E}_{Q_X}\big[\log \mathbb{E}_{\widehat{W}}[\varepsilon_{s,a,\rho}(X,Y)|X]\big]\\
&\qquad - \sum_{x,y}\theta(y|x)\mathcal{E}(x,y) + o\bigg(\sum_{x,y}\theta(y|x)\mathcal{E}(x,y)\bigg)\bigg\}
\end{aligned} \tag{24}$$

with

$$\mathcal{E}(x,y) = \frac{Q_X(x)\varepsilon_{s,a,\rho}(x,y)}{\mathbb{E}_{\widehat{W}}[\varepsilon_{s,a,\rho}(x,Y)]}. \tag{25}$$

**Theorem 2.** *Consider a channel estimate $\widehat{W}$ and fixed input distribution $Q_X$. Then, for sufficiently small $r \geq 0$, the worst-case mismatched constant-composition Gallager function is*

$$\begin{aligned}
\underline{E}_0^{\mathrm{cc}}(Q_X, \widehat{W}, \rho,\, r) = \sup_{s\geq 0, a(\cdot)} &\bigg\{ E^{\mathrm{ML}}_{s,a,\rho}(Q_X, \widehat{W})\\
&- \sqrt{2r\cdot V_{\mathrm{cc}}(Q_X, \widehat{W}, \varepsilon_{s,a,\rho})} + o(\sqrt{r})\bigg\}
\end{aligned} \tag{26}$$

*where*

$$\begin{aligned}
E^{\mathrm{ML}}_{s,a,\rho}(Q_X, \widehat{W}) = -\sum_x Q_X(x)\log\bigg(&\sum_y \widehat{W}(y|x)^{1-s\rho}\\
&\cdot e^{-\rho\cdot a(x)}\bigg(\sum_{\bar{x}} Q_X(\bar{x})\widehat{W}(y|\bar{x})^s e^{a(\bar{x})}\bigg)^{\rho}\bigg),
\end{aligned} \tag{27}$$

$$V_{\mathrm{cc}}(Q_X, \widehat{W}, \varepsilon_{s,a,\rho}) = \mathbb{E}_{Q_X}\left[\frac{\mathrm{Var}_{\widehat{W}}[\varepsilon_{s,a,\rho}(X,Y)|X]}{\mathbb{E}^2_{\widehat{W}}[\varepsilon_{s,a,\rho}(X,Y)|X]}\right]. \tag{28}$$

*Proof.* It follows a similar structure to that of Theorem 1 in Appendix A. Main differences are detailed in Appendix B. ☐

**Corollary 2.1.** *Let the approximate worst-case mismatched constant-composition Gallager function be given by*

$$\begin{aligned}
\underline{\tilde{E}}_0^{\mathrm{cc}}(Q_X, \widehat{W}, \rho, r) = \sup_{s\geq 0, a(\cdot)} &\bigg\{ E^{\mathrm{ML}}_{s,a,\rho}(Q_X, \widehat{W})\\
&- \sqrt{2r\cdot V_{\mathrm{cc}}(Q_X, \widehat{W}, \varepsilon_{s,a,\rho})}\bigg\}.
\end{aligned} \tag{29}$$

*The optimal (minimizing) conditional channel distribution for this approximation is*

$$\tilde{W}^*_{\mathrm{cc}}(y|x) = \widehat{W}(y|x)\left(1 + \sqrt{2r}\cdot\frac{\varphi_{\mathrm{cc}}(x,y,\varepsilon_{s,a,\rho})}{\mathbb{E}_{\widehat{W}}[\varepsilon_{s,a,\rho}(x,Y)]}\right) \tag{30}$$

*with*

$$\varphi_{\mathrm{cc}}(x,y,\varepsilon_{s,a,\rho}) = \frac{\varepsilon_{s,a,\rho}(x,y) - \mathbb{E}_{\widehat{W}}[\varepsilon_{s,a,\rho}(x,Y)]}{V_{\mathrm{cc}}(Q_X, \widehat{W}, \varepsilon_{s,a,\rho})}. \tag{31}$$

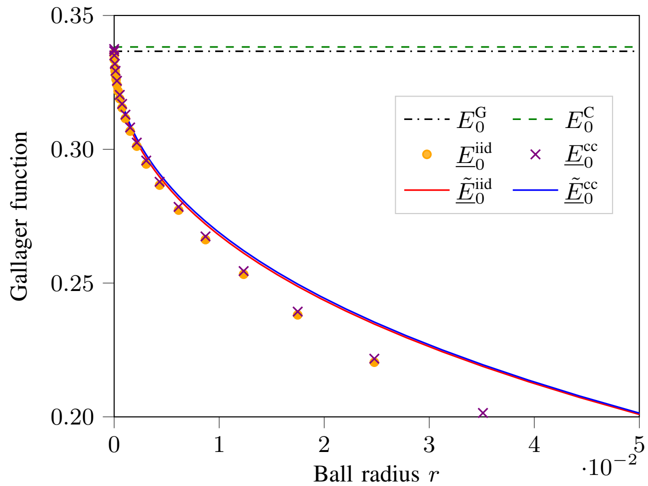


Fig. 1. Gallager functions computed for fixed $\rho = 0.7$, input distribution $Q_X$ in (32) and estimated channel $\widehat{W}$ in (33).

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

We compute $\underline{\tilde{E}}_0^{\mathrm{iid}}, \underline{\tilde{E}}_0^{\mathrm{cc}}$ from (19), (29) for input distribution $Q_X$ and channel estimate $\widehat{W}$ given by

$$Q_X = \begin{bmatrix} 0.3 & 0.3 & 0.4 \end{bmatrix} \tag{32}$$

$$\widehat{W} = \begin{bmatrix} 0.85 & 0.05 & 0.10\\ 0.03 & 0.945 & 0.025\\ 0.025 & 0.10 & 0.875 \end{bmatrix}. \tag{33}$$

We plot in Figure 1 the approximations and their true counterparts $\underline{E}_0^{\mathrm{iid}}, \underline{E}_0^{\mathrm{cc}}$, numerically computed from (14), (23) using an off-the-shelf solver. The matched Gallager $E_0$ functions for i.i.d. ($E_0^{\mathrm{G}}$) and constant-composition ($E_0^{\mathrm{C}}$) codes are shown for reference as dashed lines; they are achievable as $r \to 0$.

The figure illustrates that $\underline{E}_0^{\mathrm{iid}} \leq \underline{E}_0^{\mathrm{cc}}$ and $\underline{\tilde{E}}_0^{\mathrm{iid}} \leq \underline{\tilde{E}}_0^{\mathrm{cc}}$ for all $r$, despite the order of the approximations being substantially different. For $r > 0$, the worst-case $E_0$ functions decrease rapidly from their matched counterparts. Indeed, the worst-case functions drop with infinite slope at $r = 0$, showing even accurate estimation can have a significant impact on the achievable probability of error.

The curves have similar graphical shapes, with the approximations increasingly higher than true curves. Thus, with increasing $r$, we increasingly overestimate the worst-case Gallager function, hence also overestimate the resulting error exponent.

## Appendix A
## Proof of Theorem 1

The problem is formulated as

$$\underline{E}_0^{\mathrm{iid}} = \min_{W\in\mathcal{B}} \sup_{s\geq 0} -\log \mathbb{E}_{Q_X\times W}[\varepsilon_{s,0,\rho}(X,Y)] \tag{34}$$

$$= \sup_{s\geq 0} \min_{W\in\mathcal{B}} -\log \mathbb{E}_{Q_X\times W}[\varepsilon_{s,0,\rho}(X,Y)] \tag{35}$$

$$= \sup_{s\geq 0} -\log \max_{W\in\mathcal{B}} \sum_{x,y} Q_X(x)W(y|x)\varepsilon_{s,0,\rho}(x,y). \tag{36}$$

The expectation is convex with respect to $W$ and concave with respect to $s$ [1, Ch. 2.3], and the constraints are convex in $W$, hence we apply the minimax theorem [8] to flip the order of optimizations from (34) to (35). The minimization is moved inside the logarithm and the expectation written in full (36).

Using $\theta(y|x) = W(y|x) - \widehat{W}(y|x)$, the inner maximization can be rewritten as

$$\max_{W\in\mathcal{B}} \sum_{x,y} Q_X(x)\left(\widehat{W}(y|x) + \theta(y|x)\right)\varepsilon_{s,0,\rho}(x,y) \tag{37}$$

$$= e^{-E^{\mathrm{ML}}_{s,\rho}(Q_X,\widehat{W})} + \max_{W\in\mathcal{B}} \sum_{x,y} Q_X(x)\theta(y|x)\varepsilon_{s,0,\rho}(x,y) \tag{38}$$

with $E^{\mathrm{ML}}_{s,\rho}(Q_X,\widehat{W})$ as defined in (16).

The maximization problem is vectorized and then solved using the standard Lagrangian method. More specifically, we vectorize the constraint and cost function of the maximization in (38) using the auxiliary vector

$$\boldsymbol{\theta} = \left[\theta(y_1|x_1),\ldots,\theta(y_{|\mathcal{Y}|}|x_1),\theta(y_1|x_2),\ldots,\theta(y_{|\mathcal{Y}|}|x_{|\mathcal{X}|})\right]^T \tag{39}$$

and turn the maximization into

$$\underline{E}^{\mathrm{iid}}_{s,\rho} = \max_{\substack{d(\boldsymbol{\theta})+o(d(\boldsymbol{\theta}))\le r\\ \underline{\mathbf{1}}_j^T\boldsymbol{\theta}=0,\ 1\le j\le|\mathcal{X}|}} e^{-E^{\mathrm{ML}}_{s,\rho}(Q_X,\widehat{W})} + \boldsymbol{\theta}^T\mathcal{E} \tag{40}$$

$$= \max_{\substack{d(\boldsymbol{\theta})\le r\\ \underline{\mathbf{1}}_j^T\boldsymbol{\theta}=0,\ 1\le j\le|\mathcal{X}|}} e^{-E^{\mathrm{ML}}_{s,\rho}(Q_X,\widehat{W})} + \boldsymbol{\theta}^T\mathcal{E} + o(r) \tag{41}$$

under the following definitions

$$d(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^T K(\widehat{W})\boldsymbol{\theta} \tag{42}$$

$$K(\widehat{W}) = \mathrm{diag}\left(\frac{Q_X(x_1)}{\widehat{W}(y_1|x_1)},\ \ldots,\frac{Q_X(x_{|\mathcal{X}|})}{\widehat{W}(y_{|\mathcal{Y}|}|x_{|\mathcal{X}|})}\right) \tag{43}$$

$$\underline{\mathbf{1}}_j = \ [0\ \ldots\ 0\ 1_{(1,j)}\ \ldots\ 1_{(|\mathcal{Y}|,j)}\ 0\ \ldots\ 0]^T \tag{44}$$

$$\mathcal{E} = [Q_X(x_1)\varepsilon_{s,0,\rho}(x_1,y_1),\ \ldots,\ Q_X(x_{|\mathcal{X}|})\varepsilon_{s,0,\rho}(x_{|\mathcal{X}|},y_{|\mathcal{Y}|})]^T. \tag{45}$$

In the optimization problem, the $\underline{\mathbf{1}}_j^T\boldsymbol{\theta}=0$ constraints ensure that for every $x_j\in\mathcal{X}$, $\sum_y W(y|x_j) = 1$. The constraint $d(\boldsymbol{\theta}) + o(d(\boldsymbol{\theta})) \le r$ in (40) is a rewriting of (9) from $d(\boldsymbol{\theta})$ in (42). For $r$ sufficiently small, the approximation error term can be omitted from the constraint, as $d(\boldsymbol{\theta}) \le r$, translating the same error to the cost function, which will incur an error of order $o(d(\boldsymbol{\theta})) = o(r)$. The approximation becomes accurate as $r\to 0$, as shown numerically. We do not explicitly impose any positivity constraint on $W$ since a sufficiently small $r\ge 0$ exists such that the positivity of the resulting conditional distribution is guaranteed.

Problem (41) is linear in $\boldsymbol{\theta}$ with linear and quadratic constraints, so the KKT conditions are necessary and sufficient [9]. The standard Lagrangian method is used to solve it.

## Appendix B
## Proof of Theorem 2

The structure of the proof of Theorem 2 follows that of its i.i.d. counterpart in Appendix A closely. Recall that we performed a Taylor expansion of the cost function (23) and thus wish to solve (24). Upon applying the minimax theorem to (24), we cannot further move the minimization inside the logarithm. Therefore, in place of (40), we now solve

$$\underline{E}^{\mathrm{cc}}_0 = \min_{\substack{d(\boldsymbol{\theta})+o(d(\boldsymbol{\theta}))\le r\\ \underline{\mathbf{1}}_j^T\boldsymbol{\theta}=0,\ 1\le j\le|\mathcal{X}|}} E^{\mathrm{ML}}_{s,a,\rho}(Q_X,\widehat{W}) - \boldsymbol{\theta}^T\mathcal{E} + o(\boldsymbol{\theta}^T\mathcal{E}) \tag{46}$$

where $E^{\mathrm{ML}}_{s,a,\rho}(Q_X,\widehat{W})$ is defined in (27) and vector $\mathcal{E}$ in (45).

We deal with the order terms in the constraint and objective function as follows. As $r\to 0$, the constraint is dominated by the first term, and can be rewritten as $d(\boldsymbol{\theta})\le r$. This results in an additional error $o(d(\boldsymbol{\theta})) = o(r)$ in the objective function. Next, the error term in the objective function is $o(\boldsymbol{\theta}^T\mathcal{E}) = o(\|\boldsymbol{\theta}\|_\infty) = o(\sqrt{r})$. The result follows from writing $o(d(\boldsymbol{\theta})) = o(\|\boldsymbol{\theta}\|^2_\infty)$ and $o(d(\boldsymbol{\theta})) = o(r)$, and consequently $o(\|\boldsymbol{\theta}\|_\infty) = o(\sqrt{r})$. Therefore, we have

$$\underline{E}^{\mathrm{cc}}_0 = \min_{\substack{d(\boldsymbol{\theta})\le r\\ \underline{\mathbf{1}}_j^T\boldsymbol{\theta}=0,\ 1\le j\le|\mathcal{X}|}} E^{\mathrm{ML}}_{s,a,\rho}(Q_X,\widehat{W}) - \boldsymbol{\theta}^T\mathcal{E} + o(\sqrt{r}) \tag{47}$$

We have omitted the $o(r)$ term corresponding to approximating relative entropy as it is negligible in front of $o(\sqrt{r})$, the error committed to linearizing the error exponent.

From here, we mimic the derivation in Appendix A to obtain the final result.

## 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] R. G. Gallager, *Information Theory and Reliable Communication*. John Wiley & Sons Inc., 1968.

[3] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011.

[4] G. Poltyrev, "Random coding bounds for discrete memoryless channels," *Problems of Information Transmission*, vol. 18, no. 1, pp. 9–21, 1982.

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons Inc., 2006.

[6] 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.

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

[8] 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.

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