Competitive Minimax Universal Decoding for Several Ensembles of Random Codes *

Yaniv Akirav and Neri Merhav
Department of Electrical Engineering
Technion - Israeli Institute of Technology
Technion City, Haifa 32000, Israel
Emails:[yaniva@tx,merhav@ee].technion.ac.il

July 22, 2007

Abstract

Universally achievable error exponents pertaining to certain families of channels (most notably, discrete memoryless channels (DMC's)), and various ensembles of random codes, are studied by combining the competitive minimax approach, proposed by Feder and Merhav, with Chernoff bound and Gallager's techniques for the analysis of error exponents. In particular, we derive a single–letter expression for the largest, universally achievable fraction ξ of the optimum error exponent pertaining to the optimum ML decoding. Moreover, a simpler single–letter expression for a lower bound to ξ is presented. To demonstrate the tightness of this lower bound, we use it to show that $\xi = 1$, for the binary symmetric channel (BSC), when the random coding distribution is uniform over: (i) all codes (of a given rate), and (ii) all linear codes, in agreement with well–known results. We also show that $\xi = 1$ for the uniform ensemble of systematic linear codes, and for that of time–varying convolutional codes in the bit-error–rate sense. For the latter case, we also show how the corresponding universal decoder can be efficiently implemented using a slightly modified version of the Viterbi algorithm which employs two trellises.

Index Terms: error exponent, universal decoding, generalized likelihood ratio test, channel uncertainty, competitive minimax, Viterbi algorithm, maximum mutual information decoding.

*This research was supported by the Israel Science Foundation (ISF), grant no. 223/05.

1 Introduction

In many real–life situations, encountered in digital coded communication systems, channel variability and uncertainty prohibit the use of the optimum maximum likelihood (ML) decoder, and so, universal decoders, independent of the unknown channel parameters, are sought.

The topic of universal coding and decoding for unknown channels has received considerable attention in the last three decades. In [5], Goppa offered the *maximum mutual information* (MMI) decoder, which decides in favor of the code vector with maximum empirical mutual information with the channel output. Goppa showed that for DMC's, MMI decoding achieves capacity. Csiszár and Körner [2] also explored the universal decoding problem for DMC's with finite input and output alphabet. They showed that the random coding error exponent associated with a uniform random coding distribution over a type class achieves the optimum error exponent. Csiszár [1] proved that for any channel within the class of DMC's with additive noise, and the uniform random coding distribution over linear codes, the optimum error exponent is achievable by a decoder minimizing the noise empirical entropy, universally for all the channels in the class. Ziv [12] explored the universal decoding problem for finite state channels with finite input and output alphabets, for which the next channel state is a deterministic (but unknown) function of the channel current state and current inputs and outputs. For codes governed by a uniform random coding over a given set, he proved that a decoder based on the Lempel-Ziv algorithm asymptotically achieves the error exponent associated with ML decoding. In [6], Ziv and Lapidoth proved that the latter decoder is universal for a wider class of finite–state channels. In [3], Feder and Lapidoth found sufficient conditions for families of channels, to have universal decoders that asymptotically achieve the random coding error exponent associated with ML decoding.

Universal coding and decoding were explored also with regard to the generalized likelihood ratio test (GLRT). In this approach, each message is scored according to the maximum likelihood (over the parameter space) of the channel output vector given the message, and a decision is made in favor of the message that attains the highest maximum likelihood. Although provably optimum in certain asymptotic situations [11], [2, p. 165, Theorem 5.2], there are cases where the GLRT is strictly suboptimum [6, Sect. III, pp. 1754–1755], [4,

Appendix].

The competitive minimax criterion, first presented in [4], is an attempt for a general methodological approach to the problem of universal decoding. According to this approach, the criterion is the minimum (over all decision rules) of the maximum (over all channels in the family) of the ratio between the error probability associated with a given channel and given decision rule, and the error probability of the ML decoder for that channel, raised to some power $\xi \in [0,1]$ (cf. eq. (2) below). The largest power $\xi = \xi^*$ such that the value of this minimax ratio does not grow exponentially with the block length, is the maximum universally achievable fraction of the ML error exponent.

The main contribution of this paper is in deriving a single–letter expression to ξ^*, in terms of the rate R and a general random coding distribution, for fairly general families of channels and ensembles of random codes. While in previous works the universality was proved for certain channel models (e.g. finite–state channels, etc.) and random coding distributions (e.g. uniform distribution over a given type class, etc.), this work deals with general families of DMC's (cf. Sect. II) and general random coding distributions (cf. eq. (7)). We should note that a similar technique can be used to broaden the result for ξ^* to other channel families, e.g. Markov channels, finite state channels, etc.
In addition, a single–letter expression for a lower bound to ξ^* is presented, which is simpler to work with, and is believed to be tight. This lower bound is true also for random coding distribution over ensembles of linear code and systematic linear codes. The tightness of this lower bound is demonstrated for the case of the BSC. For this model, we show that $\xi^* = 1$, when the random coding distribution is uniform over all codes and over all linear codes, in agreement with well–known results. We also show that $\xi^* = 1$ for the ensemble of systematic linear codes, and for that of time–varying convolutional codes in the bit-error–rate sense. Using the fact that in the case of the BSC, the minimax decoding metric degenerates to a simpler metric, we propose an efficient implementation based on a slightly modified version of the Viterbi algorithm.

The outline of the paper is as follows. In Section II, we establish the notation that will be used throughout the paper and provide a formal definition of the universal decoding problem. In Section III, the main results are stated and discussed. Section IV contains a detailed proof of the single–letter expression for ξ^* will be provided. In Section V, the tightness of the lower bound to ξ^* is demonstrated for the case of the BSC with an unknown

crossover probability. In Section VI, we prove that for the ensemble of time-varying convolutional codes and the BSC with an unknown crossover probability, the minimax decoder achieves the same bit error exponent as the ML decoder, which is used when the parameter is known.

2 Notation and Problem Definition

Throughout this paper, scalar random variables (RV's) will be denoted by capital letters, their sample values will be denoted by the respective lower case letters, and their alphabets will be denoted by the respective calligraphic letters. A similar convention will apply to random vectors of dimension N and their sample values, which will be denoted with same symbols in the bold face font. The set of all N–vectors with components taking values in a certain alphabet, will be denoted as the same alphabet superscripted by N.

Information theoretic quantities like entropies, conditional entropies, and mutual informations, will be denoted following the usual conventions of the information theory literature, e.g., $H(X)$, $H(X|Y)$, $I(X;Y)$, and so on. With a slight abuse of notation, when we wish to emphasize the dependence of the entropy on the underlying probability distribution P, we denote it by $H(P)$.

The mutual information between the input and the output of the channel $\{P_\theta(y|x), x \in \mathcal{X}, y \in \mathcal{Y}\}$, when the input is governed by Q, will be denoted by

$$I_\theta(Q) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} Q(x) P_\theta(y|x) \ln \frac{P_\theta(y|x)}{\sum_{x\in\mathcal{X}} Q(x) P_\theta(y|x)}, \tag{1}$$

and the capacity of the channel will be denoted by $C_\theta = \max_Q I_\theta(Q)$.

The number of occurrences of a letter $a \in \mathcal{X}$ in a vector $\boldsymbol{x} \in \mathcal{X}^N$ will be denoted by $N_{\boldsymbol{x}}(a)$. The empirical distribution of $\boldsymbol{x}$ will be denoted by $P_{\boldsymbol{x}} = \{P_{\boldsymbol{x}}(a) = N_{\boldsymbol{x}}(a)/N,\ a \in \mathcal{X}\}$. The type class of $\boldsymbol{x}$ is defined as $T_{\boldsymbol{x}} = \left\{\boldsymbol{x'} : P_{\boldsymbol{x'}} = P_{\boldsymbol{x}}\right\}$ and $H_{\boldsymbol{x}}(X) = -\sum_{a\in\mathcal{X}} P_{\boldsymbol{x}}(a) \ln P_{\boldsymbol{x}}(a)$ will denote the entropy of a random variable (RV) X, with distribution $P_{\boldsymbol{x}}$. Similarly, the number of occurrences of a letter pair $(a,b) \in \mathcal{X}\times\mathcal{Y}$ in the vector pair $(\boldsymbol{x},\boldsymbol{y})$ will be denoted by $N_{\boldsymbol{xy}}(a,b)$, $P_{\boldsymbol{xy}} = \{P_{\boldsymbol{xy}}(a,b) = N_{\boldsymbol{xy}}(a,b)/N,\ (a,b) \in \mathcal{X}\times\mathcal{Y}\}$ will denote the joint empirical distribution of $(\boldsymbol{x},\boldsymbol{y})$, $T_{\boldsymbol{xy}} = \left\{\boldsymbol{x'},\boldsymbol{y'} : P_{\boldsymbol{x'},\boldsymbol{y'}} = P_{\boldsymbol{xy}}\right\}$ will stand for the joint type class of $(\boldsymbol{x},\boldsymbol{y})$, and $H_{\boldsymbol{xy}}(X,Y) = -\sum_{a,b\in\mathcal{X}\times\mathcal{Y}} P_{\boldsymbol{xy}}(a,b) \ln P_{\boldsymbol{xy}}(a,b)$ will denote the joint entropy of RV's (X,Y) with joint distribution $P_{\boldsymbol{xy}}$. We will use $T_{\boldsymbol{x}|\boldsymbol{y}} = \left\{\boldsymbol{x'} : P_{\boldsymbol{x'y}} = P_{\boldsymbol{xy}}\right\}$ to denote the conditional type class of $\boldsymbol{x}$ given $\boldsymbol{y}$, $P_{\boldsymbol{x}|\boldsymbol{y}}(a|b) = N_{\boldsymbol{xy}}(a,b)/N_{\boldsymbol{y}}(b),\ (a,b) \in$

$\mathcal{X} \times \mathcal{Y}$, to denote the conditional empirical distribution related to $(a,b) \in \mathcal{X} \times \mathcal{Y}$, and $H_{\boldsymbol{xy}}(X|Y) = -\sum_{a,b \in \mathcal{X} \times \mathcal{Y}} P_{\boldsymbol{xy}}(a,b) \ln P_{\boldsymbol{x}|\boldsymbol{y}}(a|b)$ to denote the conditional entropy of X given Y, induced by the joint distribution $P_{\boldsymbol{xy}}$. The empirical mutual information between RV's X and Y with joint distribution $P_{\boldsymbol{xy}}$ will be denoted by $I_{\boldsymbol{xy}}(X;Y) = H_{\boldsymbol{x}}(X) - H_{\boldsymbol{xy}}(X|Y)$.

The expectation of a function $F(X,Y)$, where X and Y are RV's distributed according to the empirical distribution of $\boldsymbol{x}$ and $\boldsymbol{y}$, will be denoted by

$$\hat{E}_{\boldsymbol{xy}}\{F(X,Y)\} = \sum_{a \in \mathcal{X}} \sum_{b \in \mathcal{Y}} P_{\boldsymbol{xy}}(a,b) F(a,b).$$

The notation $E_Q\{F(\boldsymbol{X})\}$ will be used for the expectation of a function $F(\boldsymbol{X})$, where the random vector $\boldsymbol{X}$ is governed by Q.

The Hamming distance between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ will be denoted by $d(\boldsymbol{x},\boldsymbol{y})$, and its normalization by N will be denoted by $\delta(\boldsymbol{x},\boldsymbol{y})$. For a finite set $\mathcal{A}$, $|\mathcal{A}|$ will stand for its cardinality. The divergence between two probability measures P and Q over an alphabet $\mathcal{U}$ will be denoted by $D(P||Q) = \sum_{u \in \mathcal{U}} P(u) \ln \frac{P(u)}{Q(u)}$, where $0 \ln 0$ and $0 \ln \frac{0}{0}$ are defined as 0, and $P \ln \frac{P}{0}$ for $P > 0$ is defined as ∞. For two positive sequences $\{A_N\}_{N \geq 1}$ and $\{B_N\}_{N \geq 1}$, the notation $A_N \doteq B_N$ will express the fact that $\{A_N\}_{N \geq 1}$ and $\{B_N\}_{N \geq 1}$ are of the same exponential order, i.e.,

$$\lim_{N \to \infty} \frac{1}{N} \ln (A_N / B_N) = 0.$$

Consider a DMC with a finite input alphabet $\mathcal{X}$, a finite output alphabet $\mathcal{Y}$, and single letter transition probabilities $\{P_\theta(y|x), x \in \mathcal{X}, y \in \mathcal{Y}\}$, where θ is an unknown parameter vector, taking values in some set Θ. The channel is fed by an input vector of length N, $\boldsymbol{x} \in \mathcal{X}^N$, and generates an output vector $\boldsymbol{y} \in \mathcal{Y}^N$ according to $P_\theta(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^N P_\theta(y_i|x_i)$. A rate-$R$ block code of length N consists of $M = e^{NR}$ N–vectors $\boldsymbol{x_m} \in \mathcal{X}^N$, $0 \leq m \leq M-1$, representing M different messages. A decoder Ω is a partition of $\mathcal{Y}^N$ into M regions, $\Omega_0, \Omega_1, \ldots, \Omega_{M-1}$, such that if $\boldsymbol{y}$ falls into Ω_m, a decision is made in favor of message m.

Given a code $\mathcal{C}$, the competitive minimax criterion [4] is defined as

$$S_N \stackrel{\Delta}{=} \min_{\Omega} \max_{\theta \in \Theta} \left\{ \frac{P_E(\Omega|\theta)}{[P_E{}^*(\theta)]^\xi} \right\}, \quad 0 \leq \xi \leq 1, \tag{2}$$

where $P_E(\Omega|\theta) = \frac{1}{M} \sum_{m=0}^{M-1} \sum_{\boldsymbol{y} \in \Omega_m^c} P_\theta(\boldsymbol{y}|\boldsymbol{x_m})$ is the error probability related to a decoder Ω for a given value of θ, and $P_E^*(\theta) = \min_\Omega P_E(\Omega|\theta)$ is the ML decoding error probability when θ is known.

The ratio $P_E(\Omega|\theta)/[P_E{}^*(\theta)]^\xi$ designates the loss in error probability, caused by using a universal decoder which is ignorant of θ, relative to the optimal ML decoding for that θ. The parameter ξ can be interpreted as the fraction of the optimal error exponent to which the universal decoder error exponent is compared. In order to minimize this loss uniformly over all Θ, a decoder Ω which minimizes the worst case of that ratio (i.e., its maximum), is sought.

As S_N addresses the ratio between the error probabilities, it corresponds to the difference between the error exponents related to these errors. It is well known that for most channels, the decoding error decays exponentially with the block length N. Therefore, if the value of S_N, for a decision rule Ω achieved by (2), grows sub–exponentially with N, i.e., $\lim_{N\to\infty}\frac{1}{N}\ln S_N = 0$, it means that, uniformly over Θ, the error probability associated with Ω decays with an exponential rate which is at least a fraction ξ of the error exponent rate of $P_E{}^*(\theta)$.

In [4], the following decision rule has been shown to be asymptotically optimal in the minimax sense for a given ξ:

$$\Omega_m = \{\boldsymbol{y}|f(\boldsymbol{x_m},\boldsymbol{y}) \geq f(\boldsymbol{x_{m'}},\boldsymbol{y}),\ \forall m' \neq m\} \tag{3}$$

with ties broken arbitrarily, where

$$f(\boldsymbol{x},\boldsymbol{y}) \ \stackrel{\Delta}{=}\ \max_{\theta\in\Theta} f_\theta(\boldsymbol{x},\boldsymbol{y}), \tag{4}$$

$$f_\theta(\boldsymbol{x},\boldsymbol{y}) \ \stackrel{\Delta}{=}\ \frac{1}{N}\ln\ P_\theta(\boldsymbol{y}|\boldsymbol{x}) + \xi E^*(\theta), \tag{5}$$

and $E^*(\theta)$ stands for the asymptotic exponent associated with $P_E^*(\theta)$. A decoder Ω, defined by (3), will be called the *minimax decoder* hereafter.

A natural question that may arise, at this point, is with regard to the choice of the free parameter ξ. As mentioned above, the main guideline proposed in [4] is to seek the maximum value ξ^* of ξ such that S_N would still grow sub–exponentially with N.

In the random coding regime, the error probabilities at the numerator and the denominator of (2) are replaced by the corresponding average error probabilities, i.e.,

$$\overline{S}_N \stackrel{\Delta}{=} \min_{\Omega}\max_{\theta\in\Theta}\left\{\frac{\overline{P}_E(\Omega|\theta)}{[\overline{P}_E{}^*(\theta)]^\xi}\right\} \tag{6}$$

and the decoder (3) is used, with $E^*(\theta)$ being replaced by $E_r^*(\theta)$, the random coding error exponent associated with $\overline{P}_E^*(\theta)$.

The main purpose of this paper is to translate the above–mentioned guideline for the choice of ξ into a concrete single–letter formula for the random coding regime.

3 Statement of Results

In this section, by evaluating the exponential order of $\overline{S}_N$, we derive a formula for ξ^*, the largest value of ξ for which $\overline{S}_N$ is sub–exponential in N. Moreover, an expression for the lower bound to ξ^* is also derived, and its tightness is demonstrated for the BSC model and for several ensembles of random codes.

3.1 General codes

We begin with a few definitions. For every positive integer N, let Q_N be a random coding distribution for N–vectors, of the following form:

$$Q_N(\boldsymbol{x}) = \frac{Q_N(T_{\boldsymbol{x}})}{|T_{\boldsymbol{x}}|}, \tag{7}$$

i.e., uniform distribution for all the vectors within the same type class. Of course,

$$\sum_{T_{\boldsymbol{x}}} Q_N(T_{\boldsymbol{x}}) = 1.$$

Now, let

$$\Delta_N(P_{\boldsymbol{x}}) = -\frac{1}{N} \ln Q_N(T_{\boldsymbol{x}}),$$

and let $\Delta_N^*(P)$ be an extension of the function $\Delta_N(P_{\boldsymbol{x}})$ that is defined over the continuum of probability distributions over $\mathcal{X}$ (rather than just the set of rational probability distributions with denominator N). We next define the class $\mathcal{Q}$ of sequences of random coding distributions $\{Q_N\}$ as follows: A sequence of random coding distributions $\{Q_N\}_{N\geq 1}$ is said to belong to the class $\mathcal{Q}$ if there exists such an extension $\Delta_N^*(P)$ that converges, as $N \to \infty$, to a certain non–negative functional $\Delta^*(P)$, uniformly over all probability distributions $\{P\}$ over $\mathcal{X}$.

It is easy to see that the class $\mathcal{Q}$ essentially covers all random coding distributions that are customarily used (and much more). In particular, to approximate a random coding distribution which is uniform within a small neighborhood of one type class – corresponding to a probability distribution P_0, and which vanishes elsewhere, we set $\Delta^*(P) = 0$ for every P in that neighborhood of P_0, and $\Delta^*(P) = \infty$ elsewhere. For the case where

Q is i.i.d., $\Delta^*(P) = D(P\|Q)$. In particular, if $Q(\boldsymbol{x}) = 1/|\mathcal{X}|^N$ for all $\boldsymbol{x} \in \mathcal{X}^N$, then $\Delta^*(P) = \ln|\mathcal{X}| - H(P)$.

Given a joint distribution P_{XY}, a real α, and a value of $\theta \in \Theta$, let

$$A(\theta, \alpha, P_{XY}) \stackrel{\Delta}{=} I(X;Y) + \Delta^*\left(\sum_{b\in\mathcal{Y}} P_Y(b)P_{X|Y}(\cdot|b)\right) - \alpha E \ln P_\theta(Y|X), \tag{8}$$

where $E\{\cdot\}$ is the expectation and $I(X;Y)$ is the mutual information w.r.t. a generic joint distribution $P_{XY}(a,b) = P_Y(b)P_{X|Y}(a|b)$ of the RV's (X,Y).

Next, for distributions P_Y,$P_{X|Y}$ and $P_{X'|Y}$, two parameters $\theta, \theta' \in \Theta$, and reals $0 \le \rho \le 1$ and $s \ge 0$, define:

$$B(\theta, \theta', P_Y, P_{X|Y}, P_{X'|Y}, s, \rho) \stackrel{\Delta}{=} A(\theta, 1 - s\rho, P_{XY}) + \rho \cdot A(\theta', s, P_{X'Y}) - H(Y), \tag{9}$$

where $H(Y)$ is the entropy of Y induced by P_Y. Finally, let

$$\begin{aligned}\xi^*(R) \;=\; \min_{P_{XY}} \min_{\theta'\in\Theta} \max\Bigg\{ &\min_{\theta\in\Theta} \max_{\substack{0\le\rho\le 1\\ 0\le s\le 1/\rho}} \min_{P_{X'|Y}} \frac{B(\theta,\theta',P_Y,P_{X|Y},P_{X'|Y},s,\rho) - \rho R}{(1-\rho s)E_r^*(\theta) + \rho s E_r^*(\theta')}, \\ &\max_{\theta\in\Theta} \max_{\substack{0\le\rho\le 1\\ s\ge 1/\rho}} \min_{P_{X'|Y}} \frac{B(\theta,\theta',P_Y,P_{X|Y},P_{X'|Y},s,\rho) - \rho R}{(1-\rho s)E_r^*(\theta) + \rho s E_r^*(\theta')}\Bigg\}\end{aligned} \tag{10}$$

Our main result, in this section, is the following:

Theorem 1 *Consider a sequence of ensembles of codes, where each codeword is drawn independently, under a distribution Q_N, and the sequence $\{Q_N\}_{N\ge1}$ is a member of the class $\mathcal{Q}$. Then,*

1. *For every $\xi \le \xi^*(R)$, $\lim_{N\to\infty} \frac{1}{N} \ln \overline{S}_N \le 0$.*
2. *There exists a sequence of encoders $\{\mathcal{C}_N\}_{N\ge1}$ and minimax decoders $\{\Omega_N\}_{N\ge1}$ with $\xi = \xi^*(R)$, for which:*

$$\liminf_{N\to\infty} \left[-\frac{1}{N} \ln P_E(\Omega_N|\theta)\right] \ge \xi \cdot E^*(\theta)$$

   uniformly over $\theta \in \Theta$.
3. *For every $\xi > \xi^*(R)$, $\lim_{N\to\infty} \frac{1}{N} \ln \overline{S}_N > 0$.*

The proof of Theorem 1 appears in Section IV.

We now pause to discuss Theorem 1 and some of its aspects.

The theorem suggests a conceptually simple strategy for universal decoding: Given R and the sequence $\{Q_N\}_{N\geq 1}$, first, compute $\xi^*(R)$ using eq. (10). This may require some non-trivial optimization procedures, but it has to be done only once. It should be mentioned that if closed–form analytic expression does not seem available, the computation can be carried out at least numerically, since this is a single–letter expression. Once $\xi^*(R)$ has been computed, apply the minimax decoding rule with $\xi = \xi^*(R)$ and the theorem guarantees that the resulting random coding error exponent associated with the decoder is as specified in the second item of that theorem. Moreover, the third item of the theorem implies that in the random coding regime, $\xi^*(R)$ is the largest fraction of $E^*(\theta)$ that is uniformly achievable by a universal decoder.

As mentioned earlier, when Q is uniform i.i.d., $\Delta^*(P) = \ln|\mathcal{X}| - H(X)$ (where X is governed by P), and therefore

$$A(\theta, \alpha, P_{XY}) \quad = \quad \ln|\mathcal{X}| - H(X|Y) - \alpha E \ln P_\theta(Y|X). \tag{11}$$

This observation will be used in Section V which deals with the BSC model, as well as in Section A.1 of the Appendix (ensembles of linear and systematic linear codes), as they both assume a binary i.i.d. random coding distribution.

The theorem is interesting, of course, only when $\xi^*(R) > 0$, which is the case in many situations, at least as long as R is not too large. It should be pointed out that the exponential rate $\xi^*(R) \cdot E^*(\theta)$, guaranteed by Theorem 1, is only a lower bound to the real exponential rate (as the minimax criterion is aimed to consider all $\theta \in \Theta$), and that true exponential rate, at some points in Θ, might be larger.

As mentioned above, the exact formula for ξ^*, given in eq. (10), includes many optimizations and hence might be complicated for calculation. Therefore, we next present a simpler expression for a lower bound to ξ^*, denoted by $\xi^*_{LB}(R)$, which we believe is tight at least for several families of channels. Another motivation for presenting $\xi^*_{LB}(R)$ is that it holds also for ensembles of linear and systematic linear codes, as we will shall in the next subsection. The expression for $\xi^*_{LB}(R)$ will be derived from $\xi^*(R)$ by: (i) avoiding the inner maximization between two terms in (10) by choosing the left term, and (ii) interchanging

between the minimization over $P_{X|Y}$ and the maximization over λ and ρ, i.e:

$$\xi^*_{LB}(R) \triangleq \min_{P_Y} \min_{\theta\in\Theta} \min_{\theta'\in\Theta} \max_{\substack{0\le\rho\le 1\\ 0\le\lambda\le 1/\rho}} \min_{P_{X|Y}} \min_{P_{X'|Y}} \frac{B(\theta,\theta',P_Y,P_{X|Y},P_{X'|Y},\lambda,\rho)-\rho R}{(1-\lambda\rho)\cdot E^*_r(\theta)+\lambda\rho\cdot E^*_r(\theta')}. \tag{12}$$

As $\xi^*_{LB}(R)$ is a lower bound to ξ^*, it is obvious to see that parts 1 and 2 of Theorem 1 hold for it as well.

3.2 Linear codes

We next provide a variation of $\xi^*_{LB}(R)$ for ensembles of linear codes and systematic linear codes. Prior to that, we first define these ensembles. A linear code is defined by mapping each of the $M = 2^K$ binary information (row) vectors $\boldsymbol{u_m}$, $0 \le m \le M-1$, of length K, into its corresponding code (row) vector $\boldsymbol{v_m}$, of length N, in the following way:

$$\boldsymbol{v_m} = \boldsymbol{u_m}\boldsymbol{G} \oplus \boldsymbol{v_0}, \quad m = 0, 1, \ldots, M-1,$$

where $\boldsymbol{G}$ is a binary generator matrix of dimension $K \times N$ and $\boldsymbol{v_0}$ is an additive vector of length N. The $\oplus$ operation denotes a summation modulo 2 and the multiplication between $\boldsymbol{u_m}$ and $\boldsymbol{G}$ is conducted over the field $GF(2)$. A systematic linear code is defined in the same manner, with the restriction that the left $K \times K$ block of $\boldsymbol{G}$ (the systematic part of $\boldsymbol{G}$) forms the identity matrix (thus, the first K bits of each code vector, $\boldsymbol{v_m}$, form the corresponding information vector, $\boldsymbol{u_m}$).

We now consider a random coding distribution, which is i.i.d. over the ensemble of linear codes (or systematic linear codes), for which the elements of $\boldsymbol{G}$ (or $\tilde{\boldsymbol{G}}$, the non-systematic part of $\boldsymbol{G}$, in the case of systematic linear codes) and $\boldsymbol{v_0}$ are drawn independently using a uniform single–letter distribution $Q^* = \left\{\frac{1}{2}, \frac{1}{2}\right\}$ (fair coin tossing). We also define the family of the *binary-input, output-symmetric (BIOS) channels*, as channels with a binary input alphabet $\mathcal{X}$ ("0" and "1"), an output alphabet $\mathcal{Y}$ (possibly infinite), where the transition probabilities satisfy $P(y|0) = P(-y|1), \forall y \in \mathcal{Y}$, for a well defined operation "$-$" (note that the definition of symmetry can be used as long as each $y \in \mathcal{Y}$ satisfies that $-y \in \mathcal{Y}$ as well). For example, the BSC, when mapping "0" $\rightarrow +1$ and "1" $\rightarrow -1$, is a BIOS channel. The additive Gaussian channel with two antipodal input letters, x_1 and x_2, is also a BIOS channel.

The following theorem is stated with regard to codes governed by the above mentioned ensembles and transmitted via a BIOS channel:

Theorem 2 *Consider the sequence of ensembles of linear or systematic linear codes, where the elements of* $\boldsymbol{G}$ *(or* $\tilde{\boldsymbol{G}}$*) and* $\boldsymbol{v_0}$ *are drawn independently by fair coin tossing. Let* $\{P_\theta, \theta \in \Theta\}$ *be a family of BIOS DMC's. Then, the lower bound* $\xi^*_{LB}(R)$ *of eq. (12), continues to hold, with* $\Delta^*(P) = \ln 2 - H(P)$.

Theorem 2 is proved in Section A.1 of the Appendix.

The single–letter expression derivation for $\xi^*_{LB}(R)$ is carried out (see Section A.1 of the Appendix) using the same techniques as in Gallager's classical work, which are tight in the random coding sense. We therefore believe that the achievable lower bounds to the real exponential rates are tight as well. To demonstrate the tightness of the lower bounds suggested in (12) (for general codes) and in Theorem 2 (for linear and systematic linear codes), we have the following lemma:

Lemma 1 *Consider the family of BSC's parameterized by the crossover probability* θ*. Then,* $\xi^*_{LB}(R) = 1$ *and hence* $\xi^*(R) = 1$*, in the following cases:*
(i) The ensemble of all codes with $Q_N(\boldsymbol{x}) = 2^{-N}$ *for all* $\boldsymbol{x}$*.*
(ii) The ensemble of linear codes and systematic linear codes, as in Theorem 2, with $\Delta^*(P) = \ln 2 - H(P)$.

Lemma 1 is proved in Section V.

It should be mentioned that proving that under the BSC model $\xi^* = 1$ is universally achievable by random coding over general codes and linear codes is by no means new, as it was already proved and discussed in [1]. Nevertheless, it demonstrates the tightness of $\xi^*_{LB}(R)$. However, to the best of our knowledge, the same result regarding ensembles of systematic linear codes has not been proved yet and is first shown here.

3.3 Convolutional codes

For the special case of the BSC mentioned above, we now introduce the following result, related to ensembles of time-varying convolutional codes, when the minimax decoding is used. Prior to that, we first define this ensemble and the bit error exponent related to it.

A convolutional code of rate b/n (b, n – positive integers) and constraint length Kb is defined as one for which at each time instant $t \geq 0$, the code vector of length n, $\boldsymbol{v_t}$, is obtained by

$$\boldsymbol{v_t} = \sum_{j=0}^{\min\{t,K-1\}} \boldsymbol{u_{t-j}} \boldsymbol{G_j} \oplus \boldsymbol{v_0}, \tag{13}$$

where $\boldsymbol{u_{t-j}}$ is a binary information row vector of length b at time $t-j$, $\boldsymbol{G_j}, 0 \leq j \leq K-1$, are binary matrices with b rows and n columns each, and $\boldsymbol{v_0}$ is a vector of length n.

Let us now consider a code $\mathcal{C}$, governed by i.i.d. random coding over the ensemble of time-varying convolutional codes, whose code vector of time instant $t \geq 0$, $\boldsymbol{v_t}$, is obtained by

$$\boldsymbol{v_t} = \sum_{j=0}^{\min\{t,K-1\}} \boldsymbol{u_{t-j}} \boldsymbol{G_j^t} \oplus \boldsymbol{v_0^t}, \tag{14}$$

where at each time instant t, the elements of $\boldsymbol{G_j^t}, 0 \leq j \leq K-1$ and $\boldsymbol{v_0^t}$ are drawn independently using the uniform single–letter distribution $\left\{\frac{1}{2}, \frac{1}{2}\right\}$.

The average bit error probability, $\overline{P_b(\Omega_K)}$, associated with a sequence of decoders $\Omega_K = \{\Omega_{K,N}\}_{N=1}^{\infty}$ of block length N and constraint length K, and averaged over the ensemble of time-varying convolutional codes, is defined as the expected relative frequency of bit errors in the decoded information stream, i.e.

$$\overline{P_b(\Omega_K)} = \limsup_{N\to\infty} \overline{P_b(\Omega_{K,N})}. \tag{15}$$

The bit error exponent associated with a sequence of decoders $\Omega = \{\Omega_K\}_{K=1}^{\infty}$ is defined as

$$\overline{E_b(\Omega)} = -\limsup_{K\to\infty} \frac{1}{K} \ln \overline{P_b(\Omega_K)}. \tag{16}$$

Theorem 3 *Consider the sequence of ensembles of time–varying convolutional codes of rate b/n and constraint length Kb (with $K \to \infty$), described as in the previous paragraph, and assume a family of BSC's parameterized by the crossover probability θ.*
The achievable bit error exponent (as defined in (16)) using the minimax decoder is equal to the one when θ is known and the ML decoder is used.

The proof of this theorem is based on the following observation:
Under the BSC model with an unknown crossover probability θ, the minimax decision rule (as defined in (3)) is equivalent to a decision rule, denoted by Λ, and defined as:

$$\Lambda_m = \left\{\boldsymbol{y} | \rho(\boldsymbol{x_m}, \boldsymbol{y}) \leq \rho(\boldsymbol{x_{m'}}, \boldsymbol{y}),\ \forall m' \neq m\right\}, \tag{17}$$

with ties broken arbitrarily, where

$$\rho(\boldsymbol{x}, \boldsymbol{y}) = \min\left\{\delta(\boldsymbol{x}, \boldsymbol{y}), 1 - \delta(\boldsymbol{x}, \boldsymbol{y})\right\}. \tag{18}$$

As mentioned in Section II, $\delta(\boldsymbol{x}, \boldsymbol{y})$ denotes the normalized Hamming distance between $\boldsymbol{x}$ and $\boldsymbol{y}$. This equivalence is proved in Section A.7 of the Appendix. We should note that

for this case, the minimax decoder coincides with the MMI decoder as well. Based on this equivalence, the full proof of Theorem 3 is given in Section VII. We also introduce an efficient implementation of minimax decoding, based on a slightly modified version of the Viterbi algorithm. This is done by applying the Viterbi algorithm twice: first for minimum Hamming distance, and then for maximum Hamming distance. This process results in two survivors and the selection between them is done in favor of the one whose normalized Hamming metric is more distant from $\frac{1}{2}$ (the one with the minimal ρ).

4 Proof of Theorem 1

We first observe that for a DMC, $\{P_\theta(y|x), x \in \mathcal{X}, y \in \mathcal{Y}\}$, and for each vector pair $(\boldsymbol{x}, \boldsymbol{y})$, the minimax metric for a given θ, $f_\theta(\boldsymbol{x}, \boldsymbol{y})$, depends on $\boldsymbol{x}$ and $\boldsymbol{y}$ only via their joint empirical distribution:

$$f_\theta(\boldsymbol{x}, \boldsymbol{y}) = \hat{E}_{\boldsymbol{xy}} \ln P_\theta(Y|X) + \xi E_r^*(\theta). \tag{19}$$

We, therefore, conclude that the value of θ maximizing $f_\theta(\boldsymbol{x}, \boldsymbol{y})$ also depends on $\boldsymbol{x}$ and $\boldsymbol{y}$ only via their joint empirical distribution. Let Θ_N denote the subset of Θ with values of θ that achieve $\max_\theta f_\theta(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}, \boldsymbol{y})$ as $(\boldsymbol{x}, \boldsymbol{y})$ exhaust $\mathcal{X}^N \times \mathcal{Y}^N$. In the decoding process, maximization over θ can be achieved only by points in Θ_N. Since the number of joint empirical distributions of $(\boldsymbol{x}, \boldsymbol{y})$ is upper bounded by $(N+1)^{|X||Y|}$, then $|\Theta_N| \leq (N+1)^{|X||Y|}$ as well.

As a first step, we assume given channel input and output vectors, $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. Considering a random coding distribution, Q_N, we exponentially evaluate the probability of having another codeword $\boldsymbol{x'}$ that is preferred by the minimax decoder over $\boldsymbol{x}$. This probability will be denoted by $a(\boldsymbol{x}, \boldsymbol{y})$.

$$\begin{aligned}
a(\boldsymbol{x}, \boldsymbol{y}) &= Q_N\{f(\boldsymbol{X'}, \boldsymbol{y}) \geq f(\boldsymbol{x}, \boldsymbol{y})\} \\
&= Q_N\left\{\max_{\theta' \in \Theta_N} f_{\theta'}(\boldsymbol{X'}, \boldsymbol{y}) \geq f(\boldsymbol{x}, \boldsymbol{y})\right\} \\
&\overset{(a)}{\doteq} \max_{\theta' \in \Theta_N} Q_N\{f_{\theta'}(\boldsymbol{X'}, \boldsymbol{y}) \geq f(\boldsymbol{x}, \boldsymbol{y})\} \\
&= \max_{\theta' \in \Theta_N} Q_N\left\{\sum_{i=1}^N \ln P_{\theta'}(y_i|X_i') \geq -N\xi E_r^*(\theta') + N \cdot f(\boldsymbol{x}, \boldsymbol{y})\right\} \\
&\overset{(b)}{\doteq} \max_{\theta' \in \Theta_N} \min_{s \geq 0} E_{Q_N}\left[\exp\left\{s[\sum_{i=1}^N \ln P_{\theta'}(y_i|X_i') + N\xi E_r^*(\theta')]\right\}\right] \cdot \exp\{-sNf(\boldsymbol{x}, \boldsymbol{y})\}
\end{aligned}$$

$$= \max_{\theta' \in \Theta_N} \min_{s \geq 0} E_{Q_N} e^{sNf_{\theta'}(\boldsymbol{X'},\boldsymbol{y})} \cdot e^{-sNf(\boldsymbol{x},\boldsymbol{y})}, \tag{20}$$

where (a) is true since

$$\begin{aligned}
\max_{\theta' \in \Theta_N} Q_N \{f_{\theta'}(\boldsymbol{X'},\boldsymbol{y}) \geq f(\boldsymbol{x},\boldsymbol{y})\} &\leq Q_N \left\{ \max_{\theta' \in \Theta_N} f_{\theta'}(\boldsymbol{X'},\boldsymbol{y}) \geq f(\boldsymbol{x},\boldsymbol{y}) \right\} \\
&= Q_N \left\{ \bigcup_{\theta' \in \Theta_N} f_{\theta'}(\boldsymbol{X'},\boldsymbol{y}) \geq f(\boldsymbol{x},\boldsymbol{y}) \right\} \\
&\leq \sum_{\theta' \in \Theta_N} Q_N \{f_{\theta'}(\boldsymbol{X'},\boldsymbol{y}) \geq f(\boldsymbol{x},\boldsymbol{y})\} \\
&= |\Theta_N| \cdot \max_{\theta' \in \Theta_N} Q_N \{f_{\theta'}(\boldsymbol{X'},\boldsymbol{y}) \geq f(\boldsymbol{x},\boldsymbol{y})\}, \qquad (21)
\end{aligned}$$

and in (b) we used the Cheroff bound, which is tight in the exponential sense.

By using the method of types, it is proved in Section A.3 of the Appendix that for any real α,

$$E_{Q_N}[e^{N\alpha f_\theta(\boldsymbol{X},\boldsymbol{y})}] \doteq e^{N\left[\alpha\xi E_r^*(\theta) - \min_{P_{\boldsymbol{x}|\boldsymbol{y}}} A(\theta,\alpha,P_{\boldsymbol{x}\boldsymbol{y}})\right]}, \tag{22}$$

where the function $A(\theta, \alpha, P_{xy})$ is defined as in (8).
Using this observation, we can continue to evaluate $a(\boldsymbol{x},\boldsymbol{y})$ as follows:

$$\begin{aligned}
a(\boldsymbol{x},\boldsymbol{y}) &\doteq \max_{\theta' \in \Theta_N} \min_{s \geq 0} \exp\Big\{N[s\xi E_r^*(\theta') - \min_{P_{\boldsymbol{x'}|\boldsymbol{y}}} A(\theta', s, P_{\boldsymbol{x'}\boldsymbol{y}})]\Big\} \cdot \exp\{-sNf(\boldsymbol{x},\boldsymbol{y})\} \\
&= \max_{\theta' \in \Theta_N} \min_{s \geq 0} \exp\Big\{-N[-s\xi E_r^*(\theta') + \min_{P_{\boldsymbol{x'}|\boldsymbol{y}}} A(\theta', s, P_{\boldsymbol{x'}\boldsymbol{y}}) + sf(\boldsymbol{x},\boldsymbol{y})]\Big\} \\
&\triangleq \max_{\theta' \in \Theta_N} \min_{s \geq 0} \exp\Big\{-N[G(\theta', s, \xi, P_{\boldsymbol{x}\boldsymbol{y}})]\Big\}. \qquad (23)
\end{aligned}$$

Therefore, the probability that the decoder will prefer any of the other $M-1$ codevectors rather than the transmitted codevector $\boldsymbol{x}$ can be evaluated as follows:

$$\begin{aligned}
1 - (1 - a(\boldsymbol{x},\boldsymbol{y}))^{M-1} &\overset{(a)}{\doteq} \min\{1, e^{NR} \cdot a(\boldsymbol{x},\boldsymbol{y})\} \\
&\doteq \min\left\{1, \max_{\theta' \in \Theta_N} \min_{s \geq 0} \exp\{-N[G(\theta', s, \xi, P_{\boldsymbol{x}\boldsymbol{y}}) - R]\}\right\} \\
&= \max_{\theta' \in \Theta_N} \min_{s \geq 0} \exp\Big\{-N \cdot \max_{0 \leq \rho \leq 1} \rho[G(\theta', s, \xi, P_{\boldsymbol{x}\boldsymbol{y}}) - R]\Big\} \\
&= \max_{\theta' \in \Theta_N} \min_{\substack{0 \leq \rho \leq 1 \\ s \geq 0}} \exp\Big\{-N \cdot [\rho G(\theta', s, \xi, P_{\boldsymbol{x}\boldsymbol{y}}) - \rho R]\Big\}, \qquad (24)
\end{aligned}$$

where the equivalence in (a) (see [9], Section V, and [8], Section A.2 p. 109-110) implies that the union bound in the random coding error exponent is tight.

Now, we will evaluate $\overline{S}_N$, the average of the minimax criterion over the ensemble of codebooks governed by a random coding distribution, for the minimax decoder defined in (3):

$$
\begin{aligned}
\overline{S}_N &\doteq \max_{\theta\in\Theta}\left\{\frac{\overline{P}_E\left(\Omega|\theta\right)}{e^{-N\xi E_r^*(\theta)}}\right\} \\
&= \max_{\theta\in\Theta}\left\{e^{N\xi E_r^*(\theta)}\sum_{\boldsymbol{x}\in\mathcal{X}^N}Q_N(\boldsymbol{x})\sum_{\boldsymbol{y}\in\mathcal{Y}^N}P_\theta(\boldsymbol{y}|\boldsymbol{x})\left[1-(1-a(\boldsymbol{x},\boldsymbol{y}))^{M-1}\right]\right\} \\
&\doteq \max_{\theta\in\Theta}\left\{\sum_{\boldsymbol{x}\in\mathcal{X}^N}Q_N(\boldsymbol{x})\sum_{\boldsymbol{y}\in\mathcal{Y}^N}e^{Nf_\theta(\boldsymbol{x},\boldsymbol{y})}\max_{\theta'\in\Theta_N}\min_{\substack{0\le\rho\le 1\\ s\ge 0}}\exp\Big\{-N[\rho\ G(\theta',s,\xi,P_{\boldsymbol{xy}})-\rho R]\Big\}\right\} \\
&\overset{(a)}{\doteq} \max_{\theta\in\Theta}\Bigg\{\sum_{T_{\boldsymbol{xy}}\subset\mathcal{X}^N\times\mathcal{Y}^N}Q_N(T_{\boldsymbol{x}})\left|T_{\boldsymbol{y}|\boldsymbol{x}}\right|e^{Nf_\theta(\boldsymbol{x},\boldsymbol{y})}\max_{\theta'\in\Theta_N}\min_{\substack{0\le\rho\le 1\\ s\ge 0}} \\
&\qquad e^{N\rho s\xi E_r^*(\theta')}\cdot e^{-N\rho\min_{P_{\boldsymbol{x'}|\boldsymbol{y}}}A(\theta',s,P_{\boldsymbol{x'y}})}\cdot e^{-N\rho sf(\boldsymbol{x},\boldsymbol{y})}\cdot e^{N\rho R}\Bigg\} \\
&\doteq \max_{P_{\boldsymbol{xy}}}\Bigg\{e^{-N\Delta_N^*(P_{\boldsymbol{x}})}\cdot e^{NH_{\boldsymbol{xy}}(Y|X)}\left[\max_{\theta\in\Theta_N}e^{Nf_\theta(\boldsymbol{x},\boldsymbol{y})}\right]\max_{\theta'\in\Theta_N}\min_{\substack{0\le\rho\le 1\\ s\ge 0}}\max_{P_{\boldsymbol{x'}|\boldsymbol{y}}} \\
&\qquad e^{N\rho s\xi E_r^*(\theta')}\cdot e^{-N\rho A(\theta',s,P_{\boldsymbol{x'y}})}\cdot\left[\max_{\theta''\in\Theta_N}e^{Nf_{\theta''}(\boldsymbol{x},\boldsymbol{y})}\right]^{-\rho s}\cdot e^{N\rho R}\Bigg\} \\
&\overset{(b)}{\doteq} \max_{P_{\boldsymbol{xy}}}\max_{\theta'\in\Theta_N}\min_{\substack{0\le\rho\le 1\\ s\ge 0}}\max_{P_{\boldsymbol{x'}|\boldsymbol{y}}}\Bigg\{e^{-N\Delta^*(P_{\boldsymbol{x}})}e^{NH_{\boldsymbol{xy}}(Y|X)}e^{N\rho s\xi E_r^*(\theta')}e^{-N\rho A(\theta',s,P_{\boldsymbol{x'y}})} \\
&\qquad\left[\max_{\theta\in\Theta_N}e^{Nf_\theta(\boldsymbol{x},\boldsymbol{y})}\right]^{1-\rho s}e^{N\rho R}\Bigg\} \\
&= \max_{P_{\boldsymbol{xy}}}\max_{\theta'\in\Theta_N}\min_{\substack{0\le\rho\le 1\\ s\ge 0}}\max_{P_{\boldsymbol{x'}|\boldsymbol{y}}}\Bigg\{\exp\Big\{N[-\Delta^*(P_{\boldsymbol{x}})+H_{\boldsymbol{xy}}(Y|X)+\rho s\xi E_r^*(\theta') \\
&\qquad -\rho A(\theta',s,P_{\boldsymbol{x'y}})+\rho R]\Big\}\left[\max_{\theta\in\Theta_N}e^{Nf_\theta(\boldsymbol{x},\boldsymbol{y})}\right]^{1-\rho s}\Bigg\} \\
&\triangleq \max_{P_{\boldsymbol{xy}}}\max_{\theta'\in\Theta_N}\min_{\substack{0\le\rho\le 1\\ s\ge 0}}\max_{P_{\boldsymbol{x'}|\boldsymbol{y}}}\Bigg\{\exp\Big\{N\cdot T(\theta',P_{\boldsymbol{y}},P_{\boldsymbol{x}|\boldsymbol{y}},P_{\boldsymbol{x'}|\boldsymbol{y}},s,\rho,\xi,R)\Big\} \\
&\qquad\left[\max_{\theta\in\Theta_N}e^{Nf_\theta(\boldsymbol{x},\boldsymbol{y})}\right]^{1-\rho s}\Bigg\}, \qquad (25)
\end{aligned}
$$

where in (a) we switched to a summation over the joint empirical types of $\boldsymbol{x}$ and $\boldsymbol{y}$ (which is legitimate since both $f_\theta(\boldsymbol{x},\boldsymbol{y})$ and $G(\theta',s,\xi,P_{\boldsymbol{xy}})$ depend on $\boldsymbol{x}$ and $\boldsymbol{y}$ via their joint empirical distribution), and in (b), we used the convergence assumption of the random coding distributions within the class $\mathcal{Q}$ to claim that $\Delta_N^*(P_{\boldsymbol{x}})\to\Delta^*(P_{\boldsymbol{x}})$ as $N\to\infty$ independently of $P_{\boldsymbol{x}}$, and also united the optimizations over θ and θ''.

We should observe that:

$$\min_{\substack{0\le\rho\le 1\\ s\ge 0}} \max_{P_{\boldsymbol{x'}|\boldsymbol{y}}} \left\{ \exp\left\{ N\cdot T(\theta', P_{\boldsymbol{y}}, P_{\boldsymbol{x}|\boldsymbol{y}}, P_{\boldsymbol{x'}|\boldsymbol{y}}, s, \rho, \xi, R)\right\} \left[\max_{\theta\in\Theta_N} e^{Nf_\theta(\boldsymbol{x},\boldsymbol{y})}\right]^{1-\rho s}\right\} =$$

$$\begin{aligned}
&= \min\Bigg\{ \min_{\substack{0\le\rho\le 1\\ 0\le s\le 1/\rho}} \max_{P_{\boldsymbol{x'}|\boldsymbol{y}}} \max_{\theta\in\Theta_N} \exp\Big\{ N\left[T(\theta', P_{\boldsymbol{y}}, P_{\boldsymbol{x}|\boldsymbol{y}}, P_{\boldsymbol{x'}|\boldsymbol{y}}, s, \rho, \xi, R) + f_\theta(\boldsymbol{x},\boldsymbol{y})(1-\rho s)\right]\Big\},\\
&\qquad \min_{\substack{0\le\rho\le 1\\ s\ge 1/\rho}} \max_{P_{\boldsymbol{x'}|\boldsymbol{y}}} \min_{\theta\in\Theta_N} \exp\Big\{ N\left[T(\theta', P_{\boldsymbol{y}}, P_{\boldsymbol{x}|\boldsymbol{y}}, P_{\boldsymbol{x'}|\boldsymbol{y}}, s, \rho, \xi, R) + f_\theta(\boldsymbol{x},\boldsymbol{y})(1-\rho s)\right]\Big\}\Bigg\}\\
&\overset{(a)}{=} \min\Bigg\{ \max_{\theta\in\Theta_N} \min_{\substack{0\le\rho\le 1\\ 0\le s\le 1/\rho}} \max_{P_{\boldsymbol{x'}|\boldsymbol{y}}} \exp\Big\{ N\left[T(\theta', P_{\boldsymbol{y}}, P_{\boldsymbol{x}|\boldsymbol{y}}, P_{\boldsymbol{x'}|\boldsymbol{y}}, s, \rho, \xi, R) + f_\theta(\boldsymbol{x},\boldsymbol{y})(1-\rho s)\right]\Big\},\\
&\qquad \min_{\theta\in\Theta_N} \min_{\substack{0\le\rho\le 1\\ s\ge 1/\rho}} \max_{P_{\boldsymbol{x'}|\boldsymbol{y}}} \exp\Big\{ N\left[T(\theta', P_{\boldsymbol{y}}, P_{\boldsymbol{x}|\boldsymbol{y}}, P_{\boldsymbol{x'}|\boldsymbol{y}}, s, \rho, \xi, R) + f_\theta(\boldsymbol{x},\boldsymbol{y})(1-\rho s)\right]\Big\}\Bigg\}
\end{aligned} \tag{26}$$

$$\begin{aligned}
&\overset{\Delta}{=} \min\Bigg\{ \max_{\theta\in\Theta_N} \min_{\substack{0\le\rho\le 1\\ 0\le s\le 1/\rho}} \max_{P_{\boldsymbol{x'}|\boldsymbol{y}}} \exp\left\{ N\cdot \tilde{T}(\theta,\theta', P_{\boldsymbol{y}}, P_{\boldsymbol{x}|\boldsymbol{y}}, P_{\boldsymbol{x'}|\boldsymbol{y}}, s, \rho, \xi, R)\right\},\\
&\qquad \min_{\theta\in\Theta_N} \min_{\substack{0\le\rho\le 1\\ s\ge 1/\rho}} \max_{P_{\boldsymbol{x'}|\boldsymbol{y}}} \exp\left\{ N\cdot \tilde{T}(\theta,\theta', P_{\boldsymbol{y}}, P_{\boldsymbol{x}|\boldsymbol{y}}, P_{\boldsymbol{x'}|\boldsymbol{y}}, s, \rho, \xi, R)\right\}\Bigg\},
\end{aligned} \tag{27}$$

where in (a), two interchanges are made: one between the minimization over ρ and s and the maximization over θ in the left term of the outer minimization, and one between the maximization over $P_{\boldsymbol{x'}|\boldsymbol{y}}$ and the minimization over θ in the right term of the outer minimization. The first interchange is justified in the Appendix, Section A.2. The second interchange is possible since the term to be optimized is a product of two exponential terms, one depends on $P_{\boldsymbol{x'}|\boldsymbol{y}}$ and one depends on θ, therefore the optimizations can be done independently.

Consequently, we conclude that:

$$\begin{aligned}
\overline{S}_N &\doteq \max_{P_{\boldsymbol{xy}}} \max_{\theta'\in\Theta_N} \min\Bigg\{ \max_{\theta\in\Theta_N} \min_{\substack{0\le\rho\le 1\\ 0\le s\le 1/\rho}} \max_{P_{\boldsymbol{x'}|\boldsymbol{y}}} \exp\left\{ N\cdot \tilde{T}(\theta,\theta', P_{\boldsymbol{y}}, P_{\boldsymbol{x}|\boldsymbol{y}}, P_{\boldsymbol{x'}|\boldsymbol{y}}, s, \rho, \xi, R)\right\},\\
&\qquad \min_{\theta\in\Theta_N} \min_{\substack{0\le\rho\le 1\\ s\ge 1/\rho}} \max_{P_{\boldsymbol{x'}|\boldsymbol{y}}} \exp\left\{ N\cdot \tilde{T}(\theta,\theta', P_{\boldsymbol{y}}, P_{\boldsymbol{x}|\boldsymbol{y}}, P_{\boldsymbol{x'}|\boldsymbol{y}}, s, \rho, \xi, R)\right\}\Bigg\}\\
&= \max_{P_{\boldsymbol{xy}}} \max_{\theta'\in\Theta_N} \min\Bigg\{ \exp\Big\{ N\cdot \max_{\theta\in\Theta_N} \min_{\substack{0\le\rho\le 1\\ 0\le s\le 1/\rho}} \max_{P_{\boldsymbol{x'}|\boldsymbol{y}}} \tilde{T}(\theta,\theta', P_{\boldsymbol{y}}, P_{\boldsymbol{x}|\boldsymbol{y}}, P_{\boldsymbol{x'}|\boldsymbol{y}}, s, \rho, \xi, R)\Big\},\\
&\qquad \exp\Big\{ N\cdot \min_{\theta\in\Theta_N} \min_{\substack{0\le\rho\le 1\\ s\ge 1/\rho}} \max_{P_{\boldsymbol{x'}|\boldsymbol{y}}} \tilde{T}(\theta,\theta', P_{\boldsymbol{y}}, P_{\boldsymbol{x}|\boldsymbol{y}}, P_{\boldsymbol{x'}|\boldsymbol{y}}, s, \rho, \xi, R)\Big\}\Bigg\}
\end{aligned}$$

$$= \exp\Bigg\{ N \cdot \max_{P_{\boldsymbol{xy}}} \max_{\theta' \in \Theta_N} \min\Bigg\{ \max_{\theta \in \Theta_N} \min_{\substack{0 \le \rho \le 1 \\ 0 \le s \le 1/\rho}} \max_{P_{\boldsymbol{x'}|\boldsymbol{y}}} \tilde{T}(\theta, \theta', P_{\boldsymbol{y}}, P_{\boldsymbol{x}|\boldsymbol{y}}, P_{\boldsymbol{x'}|\boldsymbol{y}}, s, \rho, \xi, R), \\ \min_{\theta \in \Theta_N} \min_{\substack{0 \le \rho \le 1 \\ s \ge 1/\rho}} \max_{P_{\boldsymbol{x'}|\boldsymbol{y}}} \tilde{T}(\theta, \theta', P_{\boldsymbol{y}}, P_{\boldsymbol{x}|\boldsymbol{y}}, P_{\boldsymbol{x'}|\boldsymbol{y}}, s, \rho, \xi, R) \Bigg\} \Bigg\}. \tag{28}$$

Now,

$\tilde{T}(\theta, \theta', P_{\boldsymbol{y}}, P_{\boldsymbol{x}|\boldsymbol{y}}, P_{\boldsymbol{x'}|\boldsymbol{y}}, s, \rho, \xi, R) =$

$$\begin{aligned} &= -\Delta^*(P_{\boldsymbol{x}}) + H_{\boldsymbol{y}}(Y) - I_{\boldsymbol{xy}}(X;Y) + \rho s \xi E_r^*(\theta') - \rho A(\theta', s, P_{\boldsymbol{x'y}}) \\ &\quad + (1-\rho s)\hat{E}_{\boldsymbol{xy}} \ln P_\theta(Y|X) + (1-\rho s)\xi E_r^*(\theta) + \rho R \\ &= -A(\theta, 1-\rho s, P_{\boldsymbol{xy}}) - \rho A(\theta', s, P_{\boldsymbol{x'y}}) + H_{\boldsymbol{y}}(Y) + \rho s \xi E_r^*(\theta') \\ &\quad + (1-\rho s)\xi E_r^*(\theta) + \rho R \\ &= -B(\theta, \theta', P_{\boldsymbol{y}}, P_{\boldsymbol{x}|\boldsymbol{y}}, P_{\boldsymbol{x'}|\boldsymbol{y}}, s, \rho) + \rho s \xi E_r^*(\theta') + (1-\rho s)\xi E_r^*(\theta) + \rho R, \end{aligned} \tag{29}$$

where the function $B(\theta, \theta', P_{\boldsymbol{y}}, P_{\boldsymbol{x}|\boldsymbol{y}}, P_{\boldsymbol{x'}|\boldsymbol{y}}, s, \rho)$ is defined as in (9).

Therefore, in order for $\overline{S}_N$ to grow sub–exponentially with N, we seek the maximal ξ such that:

$$\max_{P_{\boldsymbol{xy}}} \max_{\theta' \in \Theta_N} \min \Bigg\{ \max_{\theta \in \Theta_N} \min_{\substack{0 \le \rho \le 1 \\ 0 \le s \le 1/\rho}} \max_{P_{\boldsymbol{x'}|\boldsymbol{y}}} \tilde{T}(\theta, \theta', P_{\boldsymbol{y}}, P_{\boldsymbol{x}|\boldsymbol{y}}, P_{\boldsymbol{x'}|\boldsymbol{y}}, s, \rho, \xi, R), \\ \min_{\theta \in \Theta_N} \min_{\substack{0 \le \rho \le 1 \\ s \ge 1/\rho}} \max_{P_{\boldsymbol{x'}|\boldsymbol{y}}} \tilde{T}(\theta, \theta', P_{\boldsymbol{y}}, P_{\boldsymbol{x}|\boldsymbol{y}}, P_{\boldsymbol{x'}|\boldsymbol{y}}, s, \rho, \xi, R) \Bigg\} \le 0 \tag{30}$$

As the empirical distributions become dense in continuum of probability distributions as $N \to \infty$, and since the function $\tilde{T}(\theta, \theta', P_{\boldsymbol{y}}, P_{\boldsymbol{x}|\boldsymbol{y}}, P_{\boldsymbol{x'}|\boldsymbol{y}}, s, \rho, \xi, R)$ is continuous in $P_{\boldsymbol{y}}$, $P_{\boldsymbol{x}|\boldsymbol{y}}$ and $P_{\boldsymbol{x'}|\boldsymbol{y}}$, it is equivalent to perform the above optimizations over continuous distributions rather than empirical distributions. The same token can be used in order to broaden the maximization space for θ and θ' from Θ_N to Θ. Thus, the condition becomes:

$$\max_{P_{X_y}} \max_{\theta' \in \Theta} \min \Bigg\{ \max_{\theta \in \Theta} \min_{\substack{0 \le \rho \le 1 \\ 0 \le s \le 1/\rho}} \max_{P_{X'|y}} \tilde{T}(\theta, \theta', P_y, P_{X|y}, P_{X'|y}, s, \rho, \xi, R), \\ \min_{\theta \in \Theta} \min_{\substack{0 \le \rho \le 1 \\ s \ge 1/\rho}} \max_{P_{X'|y}} \tilde{T}(\theta, \theta', P_y, P_{X|y}, P_{X'|y}, s, \rho, \xi, R) \Bigg\} \le 0 \tag{31}$$

In other words, a maximal ξ is sought such that:

$\forall P_{Xy}, \forall \theta' \in \Theta$

$$\max_{\theta\in\Theta} \min_{\substack{0\le\rho\le 1\\ 0\le s\le 1/\rho}} \max_{P_{X'|y}} \tilde{T}(\theta,\theta',P_y,P_{X|y},P_{X'|y},s,\rho,\xi,R) \le 0 \tag{32}$$

or

$$\min_{\theta\in\Theta} \min_{\substack{0\le\rho\le 1\\ s\ge 1/\rho}} \max_{P_{X'|y}} \tilde{T}(\theta,\theta',P_y,P_{X|y},P_{X'|y},s,\rho,\xi,R) \le 0 \tag{33}$$

An equivalent condition is:

$\forall P_{Xy}, \forall \theta' \in \Theta$

$$\xi \le \min_{\theta\in\Theta} \max_{\substack{0\le\rho\le 1\\ 0\le s\le 1/\rho}} \min_{P_{X'|y}} \frac{B(\theta,\theta',P_y,P_{X|y},P_{X'|y},s,\rho)-\rho R}{(1-\rho s)E_r^*(\theta)+\rho s E_r^*(\theta')} \tag{34}$$

or

$$\xi \le \max_{\theta\in\Theta} \max_{\substack{0\le\rho\le 1\\ s\ge 1/\rho}} \min_{P_{X'|y}} \frac{B(\theta,\theta',P_y,P_{X|y},P_{X'|y},s,\rho)-\rho R}{(1-\rho s)E_r^*(\theta)+\rho s E_r^*(\theta')} \tag{35}$$

Therefore,

$$\begin{aligned}\xi^*(R) \;=\; \min_{P_{Xy}} \min_{\theta'\in\Theta} \max\Big\{ & \min_{\theta\in\Theta} \max_{\substack{0\le\rho\le 1\\ 0\le s\le 1/\rho}} \min_{P_{X'|y}} \frac{B(\theta,\theta',P_y,P_{X|y},P_{X'|y},s,\rho)-\rho R}{(1-\rho s)E_r^*(\theta)+\rho s E_r^*(\theta')},\\ & \max_{\theta\in\Theta} \max_{\substack{0\le\rho\le 1\\ s\ge 1/\rho}} \min_{P_{X'|y}} \frac{B(\theta,\theta',P_y,P_{X|y},P_{X'|y},s,\rho)-\rho R}{(1-\rho s)E_r^*(\theta)+\rho s E_r^*(\theta')}\Big\}.\end{aligned} \tag{36}$$

5 Example - the BSC

In this section, we demonstrate that for the special case of BSC with an unknown crossover probability, and a uniform random coding distribution, $\xi_{LB}^*(R) = 1$ and hence $\xi^*(R) = 1$, in agreement with well known results [1].

Consider the lower bound (12) and choose the uniform single–letter random coding distribution $Q^* = \{\frac{1}{2}, \frac{1}{2}\}$.

Now, the value of $A(\theta, \alpha, P_{XY})$ is (see (11)):

$$A(\theta,\alpha,P_{XY}) \;=\; \ln 2 - H(X|Y) - \alpha E \ln P_\theta(Y|X) \tag{37}$$

Therefore,

$$\min_{P_{X|Y}} A(\theta,\alpha,P_{XY}) \;=\; \ln 2 - \max_{P_{X|Y}} \{H(X|Y) + \alpha E \ln P_\theta(Y|X)\} \tag{38}$$

In addition, for the case of BSC with an unknown crossover probability, θ, we have (see [7], Section VI):

$$\begin{aligned} \max_{P_{X|Y}} \Big\{ H(X|Y) + \alpha E \ln P_\theta(Y|X) \Big\} &= \ln[(1-\theta)^\alpha + \theta^\alpha] \\ &\triangleq \mathcal{V}(\theta,\alpha) \end{aligned} \tag{39}$$

From these two observations, we conclude that:

$$\min_{P_{X|Y}} A(\theta,\alpha,P_{XY}) \;=\; \ln 2 - \mathcal{V}(\theta,\alpha) \tag{40}$$

Using (9), we get:

$$\begin{aligned} \xi^*_{LB}(R) &= \min_{P_Y} \min_{\theta,\theta'\in\Theta} \max_{\substack{0\le\rho\le 1 \\ 0\le\lambda\le 1/\rho}} \min_{P_{X|Y}} \min_{P_{X'|Y}} \frac{A(\theta,1-\lambda\rho,P_{XY}) + \rho A(\theta',\lambda,P_{X'Y}) - H(Y) - \rho R}{(1-\lambda\rho)\cdot E_r^*(\theta) + \lambda\rho \cdot E_r^*(\theta')} \\ &= \min_{P_Y} \min_{\theta,\theta'\in\Theta} \max_{\substack{0\le\rho\le 1 \\ 0\le\lambda\le 1/\rho}} \frac{(1+\rho)\ln 2 - \mathcal{V}(\theta,1-\lambda\rho) - \rho\mathcal{V}(\theta',\lambda) - H(Y) - \rho R}{(1-\lambda\rho)\cdot E_r^*(\theta) + \lambda\rho \cdot E_r^*(\theta')} \\ &\ge \min_{\theta,\theta'\in\Theta} \max_{\substack{0\le\rho\le 1 \\ 0\le\lambda\le 1/\rho}} \min_{P_Y} \frac{(1+\rho)\ln 2 - \mathcal{V}(\theta,1-\lambda\rho) - \rho\mathcal{V}(\theta',\lambda) - H(Y) - \rho R}{(1-\lambda\rho)\cdot E_r^*(\theta) + \lambda\rho \cdot E_r^*(\theta')} \\ &= \min_{\theta,\theta'\in\Theta} \max_{\substack{0\le\rho\le 1 \\ 0\le\lambda\le 1/\rho}} \frac{\rho\ln 2 - \mathcal{V}(\theta,1-\lambda\rho) - \rho\mathcal{V}(\theta',\lambda) - \rho R}{(1-\lambda\rho)\cdot E_r^*(\theta) + \lambda\rho \cdot E_r^*(\theta')}. \end{aligned} \tag{41}$$

Now, the random coding error exponent associated with ML decoding, $E_r^*(\theta)$, to which the minimax decoding error exponent is compared, is achieved for the BSC model by the following optimization (see [10, Sect. 3.1, 3.2 and 3.4]):

$$\begin{aligned} E_r^*(\theta) &= \max_{0\le\rho\le 1} \max_{Q} \Big\{ -\ln \sum_{y\in\{0,1\}} \Big[\sum_{x\in\{0,1\}} Q(x)\cdot P_\theta(y|x)^{\frac{1}{1+\rho}} \Big]^{1+\rho} - \rho R \Big\} \\ &\overset{(a)}{=} \max_{0\le\rho\le 1} \Big\{ \rho\ln 2 - (1+\rho)\ln\Big[(1-\theta)^{\frac{1}{1+\rho}} + \theta^{\frac{1}{1+\rho}} \Big] - \rho R \Big\} \\ &= \max_{0\le\rho\le 1} \Big\{ \rho\ln 2 - (1+\rho)\,\mathcal{V}\left(\theta, \frac{1}{1+\rho}\right) - \rho R \Big\} \\ &\triangleq \max_{0\le\rho\le 1} E_r(\theta,\rho), \end{aligned} \tag{42}$$

where in (a), the inner maximization is achieved by taking $Q^* = \{\frac{1}{2},\frac{1}{2}\}$ ([10, Sect. 3.4]). Let us now define $\rho' = \frac{\lambda\rho}{1-\lambda\rho}$ and $\rho'' = \frac{1}{\lambda} - 1$, and rewrite the numerator of (41) as follows:

$\rho \ln 2 - \mathcal{V}(\theta, 1-\lambda\rho) - \rho\mathcal{V}\left(\theta', \lambda\right) - \rho R =$

$$
\begin{aligned}
&= \rho \ln 2 - \mathcal{V}\left(\theta, \frac{1}{1+\rho'}\right) - \rho\mathcal{V}\left(\theta', \frac{1}{1+\rho''}\right) - \rho R \\
&= (1-\lambda\rho)\Big[\rho' \ln 2 - (1+\rho')\mathcal{V}\left(\theta, \frac{1}{1+\rho'}\right) - \rho' R\Big] + \\
&\quad \lambda\rho\Big[\rho'' \ln 2 - (1+\rho'')\mathcal{V}\left(\theta', \frac{1}{1+\rho''}\right) - \rho'' R\Big] \\
&= (1-\lambda\rho) \cdot E_r\left(\theta, \rho'\right) + \lambda\rho \cdot E_r\left(\theta', \rho''\right) \\
&= (1-\lambda\rho) \cdot E_r\left(\theta, \frac{\lambda\rho}{1-\lambda\rho}\right) + \lambda\rho \cdot E_r\left(\theta', \frac{1}{\lambda} - 1\right). \qquad (43)
\end{aligned}
$$

Finally, we get that

$$
\xi^*_{LB}\left(R\right) \geq \min_{\theta,\theta' \in \Theta} \max_{\substack{0 \leq \rho \leq 1 \\ 0 \leq \lambda \leq 1/\rho}} \frac{(1-\lambda\rho) \cdot E_r(\theta, \frac{\lambda\rho}{1-\lambda\rho}) + \lambda\rho \cdot E_r(\theta', \frac{1}{\lambda} - 1)}{(1-\lambda\rho) \cdot E^*_r(\theta) + \lambda\rho \cdot E^*_r(\theta')}. \qquad (44)
$$

Now, by choosing $\lambda = \frac{1}{1+\tilde{\rho}}$, where $\tilde{\rho}$ is the achiever of $E^*_r(\theta') = \max_{0 \leq \rho \leq 1} E_r(\theta', \rho)$, and $\rho = \frac{\hat{\rho}}{1+\hat{\rho}}(1+\tilde{\rho})$, where $\hat{\rho}$ is the achiever of $E^*_r(\theta) = \max_{0 \leq \rho \leq 1} E_r(\theta, \rho)$ (observing that $\frac{\hat{\rho}}{1+\hat{\rho}}(1+\tilde{\rho}) \leq 1$, therefore this choice is feasible), we get that both the numerator and the denominator of (44) equal to $(1-\lambda\rho) \cdot E_r(\theta, \hat{\rho}) + \lambda\rho \cdot E_r(\theta', \tilde{\rho})$, and so, $\xi^*_{LB}\left(R\right) = 1$.

We should note that for the BSC model, the same conclusion (i.e., $\xi^* = 1$) holds also for linear codes and systematic linear codes (as the optimal random coding distribution that was used is $Q^* = \{\frac{1}{2}, \frac{1}{2}\}$ (see (42)).

6 Proof of Theorem 3

First, consider a given channel output related to the entire transmitted sequence of information. Without loss of generality, the all-zero message will be assumed to be transmitted. Let us now consider a segment of length $K+l$, $l \geq 0$, of the transmitted information vector, and any other incorrect path diverging from it at node j and emerging at node $j+K+l$ (note that the minimum length of a diverging path is K since after a non-zero vector is inserted to the encoder, $K-1$ zero vectors are needed in order to return to the all-zero state).

We observe that the information sequence related to such an incorrect path has the following structure (we ignore the values of the information sequence outside the range

$(j, j+K+l-1)$):

$$\boldsymbol{u_j}, \boldsymbol{u_{j+1}}, \ldots, \boldsymbol{u_{j+l}}, \underbrace{\boldsymbol{0}, \ldots, \boldsymbol{0}}_{K-1}$$

where all of the vectors are of length b.

In order for the incorrect path to diverge exactly from node j to node $j+K+l$, $\boldsymbol{u_j}$ and $\boldsymbol{u_{j+l}}$ can be any of the 2^b-1 non-zero vectors (thus, there are $(2^b-1)^2$ possibilities for their values), and each of the $l-1$ information vectors $\boldsymbol{u_{j+1}}, \ldots, \boldsymbol{u_{j+l-1}}$ can be any binary vector of length b, with the restriction of no more than $K-2$ consecutive all-zero vectors (thus, there are less than $2^{b(l-1)}$ possibilities for their values). Therefore, the number of such incorrect paths, denoted by M, is upper-bounded by

$$M \leq \left(2^b-1\right)^2 2^{b(l-1)} \leq \left(2^b-1\right) 2^{bl} \tag{45}$$

We next upper bound the probability that an incorrect path is preferred by the minimax decoder (minimizing the metric ρ) over the correct path, and then average this probability over the ensemble of time–varying convolutional codes.

We will use $\boldsymbol{V_j} = [\boldsymbol{v_j}, \boldsymbol{v_{j+1}}, \ldots, \boldsymbol{v_{j+K+l-1}}]$ to denote the code vector of length $N = n(K+l)$ that corresponds to the correct all-zeros path, while $\boldsymbol{V_j'}$ and $\boldsymbol{V_j''}$ will be used to denote code vectors that correspond to other incorrect paths. The notation $\overline{\boldsymbol{V}}_{\boldsymbol{j}}$ will be used for the complement vector of $\boldsymbol{V_j}$. A segment of length N of the corresponding channel output will be denoted by $\boldsymbol{Y_j}$, and Q^* will be used to denote the random coding distribution.

$\overline{\Pr\left\{\rho(\boldsymbol{V_j'}, \boldsymbol{Y_j}) \leq \rho(\boldsymbol{V_j}, \boldsymbol{Y_j})|\theta\right\}} =$

$$\begin{aligned}
&= \sum_{\boldsymbol{V_j}, \boldsymbol{V_j'}} Q^*(\boldsymbol{V_j}, \boldsymbol{V_j'}) \Pr\left\{\rho(\boldsymbol{V_j'}, \boldsymbol{Y_j}) \leq \rho(\boldsymbol{V_j}, \boldsymbol{Y_j})|\theta\right\} \\
&\overset{(a)}{=} 2^{-2N} \sum_{\boldsymbol{V_j}, \boldsymbol{V_j'}} \Pr\left\{\rho(\boldsymbol{V_j'}, \boldsymbol{Y_j}) \leq \rho(\boldsymbol{V_j}, \boldsymbol{Y_j})\right\} \\
&= 2^{-2N} \sum_{\boldsymbol{V_j}, \boldsymbol{V_j'}} \Pr\Big\{\min\left\{\delta(\boldsymbol{V_j'}, \boldsymbol{Y_j}), 1-\delta(\boldsymbol{V_j'}, \boldsymbol{Y_j})\right\} \leq \min\left\{\delta(\boldsymbol{V_j}, \boldsymbol{Y_j}), 1-\delta(\boldsymbol{V_j}, \boldsymbol{Y_j})\right\}\Big\} \\
&= 2^{-2N} \sum_{\boldsymbol{V_j}, \boldsymbol{V_j'}} \Pr\Big\{\left[\delta(\boldsymbol{V_j'}, \boldsymbol{Y_j}) \leq \min\left\{\delta(\boldsymbol{V_j}, \boldsymbol{Y_j}), 1-\delta(\boldsymbol{V_j}, \boldsymbol{Y_j})\right\}\right] \\
&\qquad\qquad \bigcup \left[\delta(\overline{\boldsymbol{V}}_{\boldsymbol{j}}', \boldsymbol{Y_j}) \leq \min\left\{\delta(\boldsymbol{V_j}, \boldsymbol{Y_j}), 1-\delta(\boldsymbol{V_j}, \boldsymbol{Y_j})\right\}\right]\Big\} \\
&\overset{(b)}{\leq} 2^{-2N} \sum_{\boldsymbol{V_j}, \boldsymbol{V_j'}} \Pr\Big\{\delta(\boldsymbol{V_j'}, \boldsymbol{Y_j}) \leq \delta(\boldsymbol{V_j}, \boldsymbol{Y_j}) \bigcup \delta(\overline{\boldsymbol{V}}_{\boldsymbol{j}}', \boldsymbol{Y_j}) \leq \delta(\boldsymbol{V_j}, \boldsymbol{Y_j})\Big\}
\end{aligned}$$

$$
\begin{aligned}
&\overset{(c)}{\leq} \quad 2^{-2N}\sum_{\boldsymbol{V_j}}\sum_{\boldsymbol{V_j'}}\Pr\Big\{\delta(\boldsymbol{V_j'},\boldsymbol{Y_j})\leq\delta(\boldsymbol{V_j},\boldsymbol{Y_j})\Big\}+2^{-2N}\sum_{\boldsymbol{V_j}}\sum_{\boldsymbol{V_j'}}\Pr\Big\{\delta(\overline{\boldsymbol{V}}_{\boldsymbol{j}}',\boldsymbol{Y_j})\leq\delta(\boldsymbol{V_j},\boldsymbol{Y_j})\Big\}\\
&\overset{(d)}{=} \quad 2^{-2N}\sum_{\boldsymbol{V_j}}\sum_{\boldsymbol{V_j'}}\Pr\Big\{\delta(\boldsymbol{V_j'},\boldsymbol{Y_j})\leq\delta(\boldsymbol{V_j},\boldsymbol{Y_j})\Big\}+2^{-2N}\sum_{\boldsymbol{V_j}}\sum_{\boldsymbol{V_j''}}\Pr\Big\{\delta(\boldsymbol{V_j''},\boldsymbol{Y_j})\leq\delta(\boldsymbol{V_j},\boldsymbol{Y_j})\Big\}\\
&= \quad 2\cdot 2^{-2N}\sum_{\boldsymbol{V_j}}\sum_{\boldsymbol{V_j'}}\Pr\Big\{\delta(\boldsymbol{V_j'},\boldsymbol{Y_j})\leq\delta(\boldsymbol{V_j},\boldsymbol{Y_j})\Big\}\\
&\overset{(e)}{\leq} \quad 2\cdot 2^{-2N}\sum_{\boldsymbol{V_j}}\sum_{\boldsymbol{V_j'}}\sum_{\boldsymbol{Y_j}}\sqrt{P_\theta(\boldsymbol{Y_j}|\boldsymbol{V_j})P_\theta(\boldsymbol{Y_j}|\boldsymbol{V_j'})}\\
&\overset{(f)}{=} \quad 2\left\{\sum_y\left[\sum_v\frac{1}{2}\sqrt{P_\theta(y|v)}\right]^2\right\}^N\\
&\doteq \quad e^{-NR_{\theta,0}(Q^*)},
\end{aligned}
\tag{46}
$$

where

$$
R_{\theta,0}\Big(Q^*\Big)=-\ln\sum_y\left[\sum_v\frac{1}{2}\sqrt{P_\theta(y|v)}\right]^2.
$$

In (a) we used the fact that both $\boldsymbol{V_j}$ and $\boldsymbol{V_j'}$ can attain each of their 2^N possible values equiprobably and independently. This claim for $\boldsymbol{V_j}$ (which corresponds to the all–zero path) can be justified due to the fact that the elements of $\boldsymbol{G_j^t}, 0\leq j\leq K-1$ and $\boldsymbol{v_0^t}$ are repeatedly randomized at each time instant (see (14)). Therefore, $\forall\ 0\leq i\leq K+l, \boldsymbol{v_{j+i}}=\boldsymbol{v_0^{j+i}}$, thus each one of these vectors is likely to attain each of its 2^n values equiprobably. This claim for $\boldsymbol{V_j'}$ (which correspond to the incorrect path) can be justified since $\boldsymbol{u_j}$ and $\boldsymbol{u_{j+l}}$ are non–zero and $\boldsymbol{u_{j+1}},\ldots,\boldsymbol{u_{j+l-1}}$ cannot include more than $K-2$ consecutive all–zero vectors. Thus, each code vector of $\boldsymbol{v_{j+i}'}$, $0\leq i\leq K+l$ is formed by the modulo-2 sum of $\boldsymbol{v_0^{j+i}}$ with at least one of the rows of $\boldsymbol{G_0^{j+i}},\boldsymbol{G_1^{j+i}},\ldots,\boldsymbol{G_{K-1}^{j+i}}$ and is therefore likely to attain each of its 2^n values with equal probability as well and independently with the other code vectors (this fact is dealt in details in [10, Sect. 5.1]). (b) is true since we switched into looser conditions inside each event in the probability term. In (c) we used the union bound. In (d) we used the fact that observing $\delta(\overline{\boldsymbol{V}}_{\boldsymbol{j}},\boldsymbol{Y_j})$, when summing up over all of $\boldsymbol{V_j}$'s possible values, is equivalent to observing $\delta(\boldsymbol{V_j},\boldsymbol{Y_j})$ (since in both cases, each of the 2^N values of the vector is covered by the summation). In (e) we used the Bhattacharyya bound for the pairwise error probability when using ML decision rule, and (f) is true since the channel is memoryless.

We proved that the probability that other code segment would be preferred by the minimax decoder over the correct segment, when averaged over the ensemble of time-varying

convolutional codes, is upper bounded by twice the bound achieved for ML decoder in [10]. Thus, it is exponentially of the same order. The subsequent steps in deriving an upper bound to the bit error exponent for rates $R \leq R_{\theta,0}\left(Q^*\right)$ are identical to that of ML decoder (see [10, Sect. 5.1]) and the final result is the same.

Therefore, it was proved that when using the minimax decoder, the achievable exponent for bit error probability is no less than when the channel parameter is known and the ML decoder is used. The same error exponent was proved to be achievable for rates up to $R_{\theta,0}\left(Q^*\right)$.

In order to extend the average upper bound for the bit error probability to rates higher than $R_{\theta,0}\left(Q^*\right)$, we will use a slightly different technique.

First, we upper bound $\pi_{l,\theta}(j)$, the probability that a branch in the minimax based decoding path will occur by any one of the other possible paths, starting at node j and reemerging after $K+l$ branches. We should observe, as mentioned in (45), that the number of such diverging paths satisfies $M \leq \left(2^b-1\right)2^{bl}$. The code segments associated with these M incorrect paths will be denoted by $\boldsymbol{V_j^{(1)}},\ldots,\boldsymbol{V_j^{(M)}}$, respectively.

$\pi_{l,\theta}(j) =$

$$
\begin{aligned}
&= \Pr\left\{\exists\ 1\leq i\leq M : \rho(\boldsymbol{V_j^{(i)}},\boldsymbol{Y_j}) \leq \rho(\boldsymbol{V_j},\boldsymbol{Y_j})\right\} \\
&= \Pr\Big\{\exists\ 1\leq i\leq M : \min\left\{\delta(\boldsymbol{V_j^{(i)}},\boldsymbol{Y_j}), 1-\delta(\boldsymbol{V_j^{(i)}},\boldsymbol{Y_j})\right\} \leq \min\left\{\delta(\boldsymbol{V_j},\boldsymbol{Y_j}), 1-\delta(\boldsymbol{V_j},\boldsymbol{Y_j})\right\}\Big\} \\
&= \Pr\Big\{\exists\ 1\leq i\leq M : \delta(\boldsymbol{V_j^{(i)}},\boldsymbol{Y_j}) \leq \min\left\{\delta(\boldsymbol{V_j},\boldsymbol{Y_j}), 1-\delta(\boldsymbol{V_j},\boldsymbol{Y_j})\right\} \bigcup \\
&\qquad\qquad \delta(\overline{\boldsymbol{V}}_{\boldsymbol{j}}^{\boldsymbol{(i)}},\boldsymbol{Y_j}) \leq \min\left\{\delta(\boldsymbol{V_j},\boldsymbol{Y_j}), 1-\delta(\boldsymbol{V_j},\boldsymbol{Y_j})\right\}\Big\} \\
&\overset{(a)}{\leq} \Pr\Big\{\exists\ 1\leq i\leq M : \delta(\boldsymbol{V_j^{(i)}},\boldsymbol{Y_j}) \leq \delta(\boldsymbol{V_j},\boldsymbol{Y_j}) \bigcup \delta(\overline{\boldsymbol{V}}_{\boldsymbol{j}}^{\boldsymbol{(i)}},\boldsymbol{Y_j}) \leq \delta(\boldsymbol{V_j},\boldsymbol{Y_j})\Big\} \\
&\overset{(b)}{\leq} \Pr\Big\{\exists\ 1\leq i\leq M : \delta(\boldsymbol{V_j^{(i)}},\boldsymbol{Y_j}) \leq \delta(\boldsymbol{V_j},\boldsymbol{Y_j})\} + \\
&\qquad \Pr\Big\{\exists\ 1\leq i\leq M : \delta(\overline{\boldsymbol{V}}_{\boldsymbol{j}}^{\boldsymbol{(i)}},\boldsymbol{Y_j}) \leq \delta(\boldsymbol{V_j},\boldsymbol{Y_j})\Big\} \\
&\overset{(c)}{\leq} \sum_{\boldsymbol{Y_j}} P_\theta(\boldsymbol{Y_j}|\boldsymbol{V_j})^{\frac{1}{1+\rho}}\Big[\sum_{i=1}^{M} P_\theta(\boldsymbol{Y_j}|\boldsymbol{V_j^{(i)}})^{\frac{1}{1+\rho}}\Big]^\rho + \sum_{\boldsymbol{Y_j}} P_\theta(\boldsymbol{Y_j}|\boldsymbol{V_j})^{\frac{1}{1+\rho}}\Big[\sum_{i=1}^{M} P_\theta(\boldsymbol{Y_j}|\overline{\boldsymbol{V}}_{\boldsymbol{j}}^{\boldsymbol{(i)}})^{\frac{1}{1+\rho}}\Big]^\rho, \quad (47)
\end{aligned}
$$

where (a) is true since we increased the right terms of the two inequalities, and thus increased the probability for union of these two events, in (b), the union bound was used, and in (c), we used the Gallager bound for the error probability when using the ML decision rule. This error was used for each of the two error probabilities.

We now move to upper bound the average of $\pi_{l,\theta}(j)$ over the ensemble of time-varying colvolutional codes:

$\overline{\pi_{l,\theta}(j)} =$

$$
\begin{aligned}
&\overset{(a)}{=} \sum_{\boldsymbol{V_j}} 2^{-N} \sum_{\boldsymbol{V_j^{(1)}}} \cdots \sum_{\boldsymbol{V_j^{(M)}}} 2^{-NM} \pi_{l,\theta}(j) \\
&\overset{(b)}{\leq} \sum_{\boldsymbol{Y_j}} \sum_{\boldsymbol{V_j}} 2^{-N} P_\theta(\boldsymbol{Y_j}|\boldsymbol{V_j})^{\frac{1}{1+\rho}} \sum_{\boldsymbol{V_j^{(1)}}} \cdots \sum_{\boldsymbol{V_j^{(M)}}} 2^{-NM} \\
&\quad \Big[\Big[\sum_{i=1}^{M} P_\theta(\boldsymbol{Y_j}|\boldsymbol{V_j^{(i)}})^{\frac{1}{1+\rho}} \Big]^\rho + \Big[\sum_{i=1}^{M} P_\theta(\boldsymbol{Y_j}|\overline{\boldsymbol{V}}_{\boldsymbol{j}}^{(\boldsymbol{i})})^{\frac{1}{1+\rho}} \Big]^\rho \Big] \\
&\overset{(c)}{=} 2 \cdot \sum_{\boldsymbol{Y_j}} \sum_{\boldsymbol{V_j}} 2^{-N} P_\theta(\boldsymbol{Y_j}|\boldsymbol{V_j})^{\frac{1}{1+\rho}} \sum_{\boldsymbol{V_j^{(1)}}} \cdots \sum_{\boldsymbol{V_j^{(M)}}} 2^{-NM} \Big[\sum_{i=1}^{M} P_\theta(\boldsymbol{Y_j}|\boldsymbol{V_j^{(i)}})^{\frac{1}{1+\rho}} \Big]^\rho \\
&\overset{(d)}{\leq} 2 \cdot \sum_{\boldsymbol{Y_j}} \sum_{\boldsymbol{V_j}} 2^{-N} P_\theta(\boldsymbol{Y_j}|\boldsymbol{V_j})^{\frac{1}{1+\rho}} \Big[\sum_{i=1}^{M} \sum_{\boldsymbol{V_j^{(1)}}} \cdots \sum_{\boldsymbol{V_j^{(M)}}} 2^{-NM} P_\theta(\boldsymbol{Y_j}|\boldsymbol{V_j^{(i)}})^{\frac{1}{1+\rho}} \Big]^\rho, \ 0 \leq \rho \leq 1 \\
&\overset{(e)}{=} 2 \cdot \sum_{\boldsymbol{Y_j}} \sum_{\boldsymbol{V_j}} 2^{-N} P_\theta(\boldsymbol{Y_j}|\boldsymbol{V_j})^{\frac{1}{1+\rho}} \Big[\sum_{i=1}^{M} \sum_{\boldsymbol{V_j^{(i)}}} 2^{-N} P_\theta(\boldsymbol{Y_j}|\boldsymbol{V_j^{(i)}})^{\frac{1}{1+\rho}} \Big]^\rho, \qquad 0 \leq \rho \leq 1 \\
&\overset{(f)}{\leq} 2 \cdot \Big(2^b - 1\Big) 2^{bl\rho} \sum_{\boldsymbol{Y_j}} \sum_{\boldsymbol{V_j}} 2^{-N} P_\theta(\boldsymbol{Y_j}|\boldsymbol{V_j})^{\frac{1}{1+\rho}} \Big[\sum_{\boldsymbol{V_j^{(i)}}} 2^{-N} P_\theta(\boldsymbol{Y_j}|\boldsymbol{V_j^{(i)}})^{\frac{1}{1+\rho}} \Big]^\rho, \qquad 0 \leq \rho \leq 1 \\
&= 2 \cdot \Big(2^b - 1\Big) 2^{bl\rho} \sum_{\boldsymbol{Y_j}} \Big[\sum_{\boldsymbol{V_j}} 2^{-N} P_\theta(\boldsymbol{Y_j}|\boldsymbol{V_j})^{\frac{1}{1+\rho}} \Big]^{1+\rho}, \qquad 0 \leq \rho \leq 1 \\
&\overset{(g)}{=} 2 \cdot \Big(2^b - 1\Big) 2^{bl\rho} \left\{ \sum_{y} \Big[\sum_{v} 2^{-N} P_\theta\left(y|v\right)^{\frac{1}{1+\rho}} \Big]^{1+\rho} \right\}^N, \qquad 0 \leq \rho \leq 1 \\
&\doteq \Big(2^b - 1\Big) 2^{bl\rho} e^{-(K+l)nE_{\theta,0}(\rho, \{\frac{1}{2}, \frac{1}{2}\})}, \qquad 0 \leq \rho \leq 1, \qquad (48)
\end{aligned}
$$

where

$$E_{\theta,0}\left(\rho, Q^*\right) = -\ln \sum_{y} \Big[\sum_{v} 2^{-N} P_\theta\left(y|v\right)^{\frac{1}{1+\rho}} \Big]^{1+\rho}.$$

In (a), we sum over all possible code vectors associated with the different paths in the trellis. As explained earlier, each code vector can attain all of its 2^N values equiprobably and independently with the other code vectors. In (b), we used the result from (47). (c) is true since examining $P_\theta(\boldsymbol{Y_j}|\overline{\boldsymbol{V}}_{\boldsymbol{j}}^{(\boldsymbol{i})}), 1 \leq i \leq M$, when summing up over all of $\boldsymbol{V_j^{(1)}}, \ldots, \boldsymbol{V_j^{(M)}}$ possible values, is equivalent to the examination of $P_\theta(\boldsymbol{Y_j}|\boldsymbol{V_j^{(i)}}), 1 \leq i \leq M$. In (d), we

bound ourselves to $0 \leq \rho \leq 1$ and use Jensen's inequality. (e) is true since for a fixed i, $P_\theta(\boldsymbol{Y_j}|\boldsymbol{V_j^{(i)}})$ depends only on $\boldsymbol{V_j^{(i)}}$, and is enumerated for the $2^{N(M-1)}$ possibilities of $\boldsymbol{V_j^{(1)}}, \ldots, \boldsymbol{V_j^{(i-1)}}, \boldsymbol{V_j^{(i+1)}}, \ldots, \boldsymbol{V_j^{(M)}}$. In (f), we upper bound M by $\left(2^b - 1\right) 2^{bl}$, and (g) is true since the BSC is memoryless.

As in the above proof for rates up to $R_{\theta,0}\left(Q^*\right)$, the subsequent steps in deriving an upper bound to the bit error exponent for rates $R_{\theta,0}\left(Q^*\right) \leq R \leq C_\theta$ for the minimax decoder are identical to that of ML decoder (see [10, Sect. 5.1]) and the final result is the same. This completes the proof that the achievable exponent for bit error probability of the minimax decoder is equal to that of the ML decoder, for all rates up to capacity.

A. Appendix

A.1 Proof of eq. (12) for ensembles of Linear and Systematic Linear Codes

In this section, we examine the performance of the minimax decoding rule with respect to uniform i.i.d. random coding over ensembles of linear codes and systematic linear codes. We will prove that for a family of BIOS channels, the same single–letter formula for the lower bound to the achievable fraction ξ^* is obtained, with uniform i.i.d. random coding distribution $Q^* = \left\{\frac{1}{2}, \frac{1}{2}\right\}$ (i.e. $\Delta^*(P) = \ln 2 - H(P)$).

Using Gallager's techniques, we first upper bound the decoding error probability given that the m-th message was sent for a given θ in the following way:

$$
\begin{aligned}
P_{E_m}\left(\Omega|\theta\right) &= \sum_{\boldsymbol{y}\in\mathcal{Y}^N} P_\theta(\boldsymbol{y}|\boldsymbol{v_m}) 1\Big\{\exists m' \neq m : \max_{\theta'\in\Theta_N} f_{\theta'}(\boldsymbol{v_{m'}}, \boldsymbol{y}) \geq \max_{\theta''\in\Theta_N} f_{\theta''}(\boldsymbol{v_m}, \boldsymbol{y})\Big\} \\
&= \sum_{\boldsymbol{y}\in\mathcal{Y}^N} P_\theta(\boldsymbol{y}|\boldsymbol{v_m}) 1\Big\{\exists \theta', \exists m' \neq m : f_{\theta'}(\boldsymbol{v_{m'}}, \boldsymbol{y}) \geq \max_{\theta''\in\Theta_N} f_{\theta''}(\boldsymbol{v_m}, \boldsymbol{y})\Big\} \\
&\overset{(a)}{=} \sum_{\boldsymbol{y}\in\mathcal{Y}^N} P_\theta(\boldsymbol{y}|\boldsymbol{v_m}) \max_{\theta'\in\Theta_N} 1\Big\{\exists m' \neq m : f_{\theta'}(\boldsymbol{v_{m'}}, \boldsymbol{y}) \geq \max_{\theta''\in\Theta_N} f_{\theta''}(\boldsymbol{v_m}, \boldsymbol{y})\Big\} \\
&= \sum_{\boldsymbol{y}\in\mathcal{Y}^N} P_\theta(\boldsymbol{y}|\boldsymbol{v_m}) \max_{\theta'\in\Theta_N} 1\Big\{\exists m' \neq m : \frac{f_{\theta'}(\boldsymbol{v_{m'}}, \boldsymbol{y})}{\max_{\theta''\in\Theta_N} f_{\theta''}(\boldsymbol{v_m}, \boldsymbol{y})} \geq 1\Big\} \\
&\overset{(b)}{\leq} \sum_{\boldsymbol{y}\in\mathcal{Y}^N} P_\theta(\boldsymbol{y}|\boldsymbol{v_m}) \max_{\theta'\in\Theta_N} \min_{\substack{\lambda\geq 0 \\ \rho\geq 0}} \left[\sum_{m'\neq m} \Big(\frac{e^{N f_{\theta'}(\boldsymbol{v_{m'}}, \boldsymbol{y})}}{\max_{\theta''\in\Theta_N} e^{N f_{\theta''}(\boldsymbol{v_m}, \boldsymbol{y})}}\Big)^\lambda\right]^\rho, \qquad (49)
\end{aligned}
$$

where $\rho \geq 0$ and $\lambda \geq 0$ are free parameters.

(a) is true since if we denote with $A(\theta)$ an event dependent on $\theta \in \Theta_N$, and denote with C a constant, then

$$
1\left\{\exists \theta \in \Theta_N : A(\theta) > C\right\} = \max_{\theta\in\Theta_N} 1\left\{A(\theta) > C\right\}.
$$

(b) is true since if we denote with $f_1(m')$ and $f_2(m)$ two non–negative functions of m' and m respectively, then (using Gallager's technique)

$$1\left\{\exists m' \neq m : \frac{f_1(m')}{f_2(m)} \geq 1\right\} \leq \min_{\substack{\lambda \geq 0 \\ \rho \geq 0}} \left[\sum_{m' \neq m} \left(\frac{e^{f_1(m')}}{e^{f_2(m)}}\right)^{\lambda}\right]^{\rho}.$$

Based on (49), we now develop an upper bound to the minimax criterion related to a specific linear code (i.e., specific values of $\boldsymbol{G}$ and $\boldsymbol{v_0}$, thus denoted by $S_N(\boldsymbol{v_0}, \boldsymbol{G})$):

$$\begin{aligned}
S_N(\boldsymbol{v_0}, \boldsymbol{G}) &= \max_{\theta \in \Theta}\left\{\frac{P_E(\Omega|\theta)}{[\overline{P}_E{}^*(\theta)]^{\xi}}\right\} \doteq \max_{\theta \in \Theta}\left\{\frac{P_E(\Omega|\theta)}{e^{-N\xi E_r^*(\theta)}}\right\} \\
&\leq \max_{\theta \in \Theta}\left\{\frac{1}{M}\sum_{m=0}^{M-1}\sum_{\boldsymbol{y} \in \mathcal{Y}^N} e^{N\xi E_r^*(\theta)} P_\theta(\boldsymbol{y}|\boldsymbol{v_m}) \max_{\theta' \in \Theta_N} \min_{\substack{\lambda \geq 0 \\ \rho \geq 0}} \right. \\
&\qquad \left.\left(\sum_{m' \neq m}\left[\frac{e^{Nf_{\theta'}(\boldsymbol{v_{m'}}, \boldsymbol{y})}}{\max_{\theta'' \in \Theta_N} e^{Nf_{\theta''}(\boldsymbol{v_m}, \boldsymbol{y})}}\right]^{\lambda}\right)^{\rho}\right\} \\
&= \max_{\theta \in \Theta}\left\{\frac{1}{M}\sum_{m=0}^{M-1}\sum_{\boldsymbol{y} \in \mathcal{Y}^N} e^{N \cdot f_\theta(\boldsymbol{v_m}, \boldsymbol{y})} \max_{\theta' \in \Theta_N} \min_{\substack{\lambda \geq 0 \\ \rho \geq 0}} \right. \\
&\qquad \left.\left(\sum_{m' \neq m}\left[\frac{e^{Nf_{\theta'}(\boldsymbol{v_{m'}}, \boldsymbol{y})}}{\max_{\theta'' \in \Theta_N} e^{Nf_{\theta''}(\boldsymbol{v_m}, \boldsymbol{y})}}\right]^{\lambda}\right)^{\rho}\right\} \\
&\overset{(a)}{\leq} \frac{1}{M}\sum_{m=0}^{M-1}\sum_{\boldsymbol{y} \in \mathcal{Y}^N}\left(\max_{\theta \in \Theta_N} e^{Nf_\theta(\boldsymbol{v_m}, \boldsymbol{y})}\right) \max_{\theta' \in \Theta_N} \min_{\substack{\lambda \geq 0 \\ \rho \geq 0}} \left\{\frac{\left[\sum_{m' \neq m} e^{N\lambda f_{\theta'}(\boldsymbol{v_{m'}}, \boldsymbol{y})}\right]^{\rho}}{\left(\max_{\theta'' \in \Theta_N} e^{Nf_{\theta''}(\boldsymbol{v_m}, \boldsymbol{y})}\right)^{\lambda\rho}}\right\} \\
&\overset{(b)}{\leq} \frac{1}{M}\sum_{m=0}^{M-1}\sum_{\boldsymbol{y} \in \mathcal{Y}^N} \max_{\theta' \in \Theta_N} \min_{\substack{\lambda \geq 0 \\ \rho \geq 0}}\left\{\left[\max_{\theta \in \Theta_N} e^{Nf_\theta(\boldsymbol{v_m}, \boldsymbol{y})}\right]^{1-\lambda\rho}\left[\sum_{m' \neq m} e^{N\lambda f_{\theta'}(\boldsymbol{v_{m'}}, \boldsymbol{y})}\right]^{\rho}\right\} \\
&\overset{(c)}{\leq} \frac{1}{M}\sum_{m=0}^{M-1}\sum_{\boldsymbol{y} \in \mathcal{Y}^N} \max_{\theta' \in \Theta_N} \min_{\substack{\rho \geq 0 \\ 0 \leq \lambda \leq 1/\rho}} \max_{\theta \in \Theta_N}\left\{e^{N(1-\lambda\rho)f_\theta(\boldsymbol{v_m}, \boldsymbol{y})}\left[\sum_{m' \neq m} e^{N\lambda f_{\theta'}(\boldsymbol{v_{m'}}, \boldsymbol{y})}\right]^{\rho}\right\} \\
&\overset{(d)}{=} \frac{1}{M}\sum_{m=0}^{M-1}\sum_{\boldsymbol{y} \in \mathcal{Y}^N} \max_{\theta \in \Theta_N} \max_{\theta' \in \Theta_N} \min_{\substack{\rho \geq 0 \\ 0 \leq \lambda \leq 1/\rho}}\left\{e^{N(1-\lambda\rho)f_\theta(\boldsymbol{v_m}, \boldsymbol{y})}\left[\sum_{m' \neq m} e^{N\lambda f_{\theta'}(\boldsymbol{v_{m'}}, \boldsymbol{y})}\right]^{\rho}\right\}.
\end{aligned} \tag{50}$$

The passages (a)–(d) are explained as follows: In (a) we used the fact that the maximum of an expectation is no greater than the expectation of the maximum and changed the maximization of θ to be over Θ_N. (b) is true since θ and θ'' maximize two identical expressions, and therefore can be united. In (c) we restricted the range of the optimization to $1 - \lambda\rho \geq 0$

($\Rightarrow \lambda \leq 1/\rho$). In (d) we used the fact that for given $\boldsymbol{v_m}$, $\boldsymbol{y}$ and θ'

$$\min_{\substack{\rho\geq 0\\ 0\leq\lambda\leq 1/\rho}} \max_{\theta\in\Theta_N} \left\{ e^{N(1-\lambda\rho)f_\theta(\boldsymbol{v_m},\boldsymbol{y})} \Big[\sum_{m'\neq m} e^{N\lambda f_{\theta'}(\boldsymbol{v_{m'}},\boldsymbol{y})} \Big]^\rho \right\} \;=$$

$$\max_{\theta\in\Theta_N} \min_{\substack{\rho\geq 0\\ 0\leq\lambda\leq 1/\rho}} \left\{ e^{N(1-\lambda\rho)f_\theta(\boldsymbol{v_m},\boldsymbol{y})} \Big[\sum_{m'\neq m} e^{N\lambda f_{\theta'}(\boldsymbol{v_{m'}},\boldsymbol{y})} \Big]^\rho \right\}. \tag{51}$$

This interchange between the minimization over λ and ρ and the maximization over θ is justified in the Appendix, Section A.2.

Prior to deriving the single–letter formula for the lower bound to ξ^*, we first present the following claim:

Lemma 2 *When a linear code is used for a BIOS channel and minimax decoding is used, the error probability for the m-th message is equal for all* $0 \leq m \leq M-1$.

This lemma is proved in Section A.4 of the Appendix.

Based on this observation, we can assume, without loss of generality, that $\boldsymbol{u_0} = \boldsymbol{0}$ was transmitted, and then the upper bound to S_N can be expressed as:

$$S_N(\boldsymbol{v_0}, \boldsymbol{G}) \;\leq\; \sum_{\boldsymbol{y}\in Y^N} \max_{\theta\in\Theta_N} \max_{\theta'\in\Theta_N} \min_{\substack{\rho\geq 0\\ 0\leq\lambda\leq 1/\rho}} \left\{ e^{N(1-\lambda\rho)f_\theta(\boldsymbol{v_0},\boldsymbol{y})} \Big[\sum_{m=1}^{M-1} e^{N\lambda f_{\theta'}(\boldsymbol{v_m},\boldsymbol{y})} \Big]^\rho \right\} \tag{52}$$

In the following subsections, we will use the same technique to derive two upper bounds on the minimax criterion, one for the ensemble of linear codes and one for the ensemble of systematic linear codes.

Linear Codes

By averaging S_N over the ensemble of linear codes:

$$\begin{aligned}
\overline{S}_N &\overset{(a)}{=} 2^{-(K+1)N} \sum_{\boldsymbol{v_0},\boldsymbol{G}} S_N(\boldsymbol{v_0}, \boldsymbol{G}) \\
&\leq 2^{-(K+1)N} \sum_{\boldsymbol{v_0},\boldsymbol{G}} \sum_{\boldsymbol{y}\in Y^N} \max_{\theta\in\Theta_N} \max_{\theta'\in\Theta_N} \min_{\substack{\rho\geq 0\\ 0\leq\lambda\leq 1/\rho}} \left\{ e^{N(1-\lambda\rho)f_\theta(\boldsymbol{v_0},\boldsymbol{y})} \left[\sum_{m=1}^{M-1} e^{N\lambda f_{\theta'}(\boldsymbol{v_m},\boldsymbol{y})} \right]^\rho \right\} \qquad (53) \\
&\overset{(b)}{\leq} 2^{-(K+1)N} \sum_{\boldsymbol{v_0},\boldsymbol{G}} \sum_{\boldsymbol{y}\in Y^N} \sum_{\theta\in\Theta_N} \sum_{\theta'\in\Theta_N} \min_{\substack{\rho\geq 0\\ 0\leq\lambda\leq 1/\rho}} \left\{ e^{N(1-\lambda\rho)f_\theta(\boldsymbol{v_0},\boldsymbol{y})} \left[\sum_{m=1}^{M-1} e^{N\lambda f_{\theta'}(\boldsymbol{v_m},\boldsymbol{y})} \right]^\rho \right\}
\end{aligned}$$

$$
\begin{aligned}
&\overset{(c)}{\leq} \sum_{\boldsymbol{y}\in Y^N} \sum_{\theta\in\Theta_N} \sum_{\theta'\in\Theta_N} \min_{\substack{\rho\geq 0\\ 0\leq\lambda\leq 1/\rho}} \\
&\quad \left\{ 2^{-(K+1)N} \sum_{\boldsymbol{v_0},\boldsymbol{G}} \left(e^{N(1-\lambda\rho) f_\theta(\boldsymbol{v_0},\boldsymbol{y})} \left[\sum_{m=1}^{M-1} e^{N\lambda f_{\theta'}(\boldsymbol{v_m},\boldsymbol{y})} \right]^\rho \right) \right\} \\
&\overset{(d)}{\leq} |\Theta_N|^2 \sum_{\boldsymbol{y}\in Y^N} \max_{\theta\in\Theta_N} \max_{\theta'\in\Theta_N} \min_{\substack{\rho\geq 0\\ 0\leq\lambda\leq 1/\rho}} \\
&\quad \left\{ 2^{-N} \sum_{\boldsymbol{v_0}} e^{N(1-\lambda\rho) f_\theta(\boldsymbol{v_0},\boldsymbol{y})} 2^{-KN} \sum_{\boldsymbol{G}} \left[\sum_{m=1}^{M-1} e^{N\lambda f_{\theta'}(\boldsymbol{v_m},\boldsymbol{y})} \right]^\rho \right\} \qquad (54) \\
&\overset{(e)}{\leq} |\Theta_N|^2 \sum_{\boldsymbol{y}\in Y^N} \max_{\theta\in\Theta_N} \max_{\theta'\in\Theta_N} \min_{\substack{0\leq\rho\leq 1\\ 0\leq\lambda\leq 1/\rho}} \\
&\quad \left\{ 2^{-N} \sum_{\boldsymbol{v_0}} e^{N(1-\lambda\rho) f_\theta(\boldsymbol{v_0},\boldsymbol{y})} \left[2^{-KN} \sum_{\boldsymbol{G}} \sum_{m=1}^{M-1} e^{N\lambda f_{\theta'}(\boldsymbol{v_m},\boldsymbol{y})} \right]^\rho \right\} \\
&= |\Theta_N|^2 \sum_{\boldsymbol{y}\in Y^N} \max_{\theta\in\Theta_N} \max_{\theta'\in\Theta_N} \min_{\substack{0\leq\rho\leq 1\\ 0\leq\lambda\leq 1/\rho}} \\
&\quad \left\{ 2^{-N} \sum_{\boldsymbol{v_0}} e^{N(1-\lambda\rho) f_\theta(\boldsymbol{v_0},\boldsymbol{y})} \left[2^{-KN} \sum_{m=1}^{M-1} \sum_{\boldsymbol{G}} e^{N\lambda f_{\theta'}(\boldsymbol{v_m},\boldsymbol{y})} \right]^\rho \right\} \\
&\overset{(f)}{=} |\Theta_N|^2 \sum_{\boldsymbol{y}\in Y^N} \max_{\theta\in\Theta_N} \max_{\theta'\in\Theta_N} \min_{\substack{0\leq\rho\leq 1\\ 0\leq\lambda\leq 1/\rho}} \\
&\quad \left\{ 2^{-N} \sum_{\boldsymbol{v_0}} e^{N(1-\lambda\rho) f_\theta(\boldsymbol{v_0},\boldsymbol{y})} \left[(M-1)\, 2^{-N} \sum_{\boldsymbol{v}} e^{N\lambda f_{\theta'}(\boldsymbol{v},\boldsymbol{y})} \right]^\rho \right\} \\
&\leq |\Theta_N|^2 \sum_{\boldsymbol{y}\in Y^N} \max_{\theta\in\Theta_N} \max_{\theta'\in\Theta_N} \min_{\substack{0\leq\rho\leq 1\\ 0\leq\lambda\leq 1/\rho}} \\
&\quad \left\{ M^\rho \left[2^{-N} \sum_{\boldsymbol{v}} e^{N(1-\lambda\rho) f_\theta(\boldsymbol{v},\boldsymbol{y})} \right] \cdot \left[2^{-N} \sum_{\boldsymbol{v'}} e^{N\lambda f_{\theta'}(\boldsymbol{v'},\boldsymbol{y})} \right]^\rho \right\}, \qquad (55)
\end{aligned}
$$

where the steps (a)–(f) are as follows: The equality in (a) is obtained by averaging over $2^{(K+1)N}$ equiprobable values of $\boldsymbol{v_0}$ and $\boldsymbol{G}$. (b) and (d) follow from the fact that for a non-negative function $f(\theta)$, non-negative function $f(\theta)$,

$$
\max_{\theta\in\Theta_N} f(\theta) \leq \sum_{\theta\in\Theta_N} f(\theta) \leq |\Theta_N| \cdot \max_{\theta\in\Theta_N} f(\theta). \qquad (56)
$$

(c) is true since an expectation of a minimum is upper-bounded by the minimum of the expectation. In (e), we limit the optimization over ρ to $0 \leq \rho \leq 1$ and use Jensen's

inequality. In (f), we used the following equivalence for the two inner summations:

$$\sum_{m=1}^{M-1}\sum_{\boldsymbol{G}} e^{N\lambda f_{\theta'}(\boldsymbol{v_m},\boldsymbol{y})} = (M-1)2^{(K-1)N}\sum_{\boldsymbol{v}} e^{N\lambda f_{\theta'}(\boldsymbol{v},\boldsymbol{y})}. \tag{57}$$

This equivalence is proved in Section A.5 of the Appendix.

From (19), we conclude that the term inside the summation in (55) is identical for all $\boldsymbol{y}$'s of the same type class. Thus, the summation can be conducted over types. Using (22), we continue to upper bound $\overline{S}_N$ in the following way (note that the function $A(\theta,\alpha,P_{\boldsymbol{xy}})$ used here corresponds to a binary i.i.d. random coding distribution, as specified in (11)):

$$\begin{aligned}
\overline{S}_N &\overset{(a)}{\leq} |\Theta_N|^2 \sum_{T_{\boldsymbol{y}}} \max_{\theta\in\Theta_N}\max_{\theta'\in\Theta_N}\min_{\substack{0\leq\rho\leq 1\\ 0\leq\lambda\leq 1/\rho}} \\
&\quad \Bigg\{ e^{N\rho R} e^{NH_{\boldsymbol{y}}(Y)} e^{N\left[(1-\lambda\rho)\xi E_r^*(\theta)-\min_{P_{\boldsymbol{x}|\boldsymbol{y}}} A(\theta,1-\lambda\rho,P_{\boldsymbol{xy}})\right]} \\
&\quad e^{N\left[\lambda\rho\xi E_r^*(\theta')-\rho\cdot\min_{P_{\boldsymbol{x'}|\boldsymbol{y}}} A(\theta',\lambda,P_{\boldsymbol{x'y}})\right]} \Bigg\} \\
&= |\Theta_N|^2 \sum_{T_{\boldsymbol{y}}} \max_{\theta\in\Theta_N}\max_{\theta'\in\Theta_N}\min_{\substack{0\leq\rho\leq 1\\ 0\leq\lambda\leq 1/\rho}} \\
&\quad \Bigg\{ \exp\Big\{ N[\rho R + H_{\boldsymbol{y}}(Y) + (1-\lambda\rho)\xi E_r^*(\theta) - \min_{P_{\boldsymbol{x}|\boldsymbol{y}}} A(\theta,1-\lambda\rho,P_{\boldsymbol{xy}}) \\
&\quad +\lambda\rho\xi E_r^*(\theta') - \rho\cdot\min_{P_{\boldsymbol{x'}|\boldsymbol{y}}} A(\theta',\lambda,P_{\boldsymbol{x'y}})]\Big\}\Bigg\} \\
&\overset{(b)}{\leq} |\Theta_N|^2 (N+1)^{|\mathcal{Y}|} \max_{P_{\boldsymbol{y}}}\max_{\theta\in\Theta_N}\max_{\theta'\in\Theta_N}\min_{\substack{0\leq\rho\leq 1\\ 0\leq\lambda\leq 1/\rho}} \\
&\quad \Bigg\{ \exp\Big\{ N[\rho R + H_{\boldsymbol{y}}(Y) + (1-\lambda\rho)\xi E_r^*(\theta) - \min_{P_{\boldsymbol{x}|\boldsymbol{y}}} A(\theta,1-\lambda\rho,P_{\boldsymbol{xy}}) \\
&\quad +\lambda\rho\xi E_r^*(\theta') - \rho\cdot\min_{P_{\boldsymbol{x'}|\boldsymbol{y}}} A(\theta',\lambda,P_{\boldsymbol{x'y}})]\Big\}\Bigg\} \\
&= |\Theta_N|^2 (N+1)^{|\mathcal{Y}|}\cdot\exp\Bigg\{ N\cdot\max_{P_{\boldsymbol{y}}}\max_{\theta\in\Theta_N}\max_{\theta'\in\Theta_N}\min_{\substack{0\leq\rho\leq 1\\ 0\leq\lambda\leq 1/\rho}}\max_{P_{\boldsymbol{x}|\boldsymbol{y}}}\max_{P_{\boldsymbol{x'}|\boldsymbol{y}}} \\
&\quad [\rho R + H_{\boldsymbol{y}}(Y) + (1-\lambda\rho)\xi E_r^*(\theta) - A(\theta,1-\lambda\rho,P_{\boldsymbol{xy}}) \\
&\quad +\lambda\rho\xi E_r^*(\theta') - \rho\cdot A(\theta',\lambda,P_{\boldsymbol{x'y}})]\Bigg\},
\end{aligned} \tag{58}$$

where in (a) we upper bound $|T_{\boldsymbol{y}}|$ by $e^{N\cdot H_{\boldsymbol{y}}(Y)}$, and in (b) we upper bound the summation of the functional over $T_{\boldsymbol{y}}$ by the product of the maximal value (achieved by a specific distribution $P_{\boldsymbol{y}}$) with $(N+1)^{|\mathcal{Y}|}$, which is an upper bound to the number of type classes $\{T_{\boldsymbol{y}}\}$.

As explained earlier, we seek the maximal ξ such that $\overline{S}_N$ grows sub–exponentially with N. To this end, we can ignore the factor $|\Theta_N|^2 (N+1)^{|\mathcal{Y}|}$ in (58), as it grows polynomially with N. Moreover, as mentioned in Section V, the optimizations can be conducted over continuous distributions and over the entire parameter space, Θ. Thus, a maximal ξ is sought, such that (using (9)):

$$\max_{P_Y} \max_{\theta,\theta' \in \Theta} \min_{\substack{0 \le \rho \le 1 \\ 0 \le \lambda \le 1/\rho}} \max_{P_{X|Y}} \max_{P_{X'|Y}} \Big[\rho R + (1-\lambda\rho)\xi E_r^*(\theta) + \lambda\rho\xi E_r^*(\theta') - B(\theta,\theta',P_Y,P_{X|Y},P_{X'|Y},\lambda,\rho) \Big] \le 0. \tag{59}$$

An equivalent condition to (59) is

$$\forall\, P_Y,\ \forall\, \theta,\theta' \in \Theta,\ \exists\, 0 \le \rho \le 1, 0 \le \lambda \le 1/\rho\text{:} \quad \forall\, P_{X|Y},\ \forall\, P_{X'|Y}$$

$$\rho R + (1-\lambda\rho)\xi E_r^*(\theta) + \lambda\rho\xi E_r^*(\theta') - B(\theta,\theta',P_Y,P_{X|Y},P_{X'|Y},\lambda,\rho) \le 0$$

or,

$$\forall\, P_Y,\ \forall\, \theta,\theta' \in \Theta,\ \exists\, 0 \le \rho \le 1, 0 \le \lambda \le 1/\rho\text{:} \quad \forall\, P_{X|Y},\ \forall\, P_{X'|Y}$$

$$\xi \le \frac{B(\theta,\theta',P_Y,P_{X|Y},P_{X'|Y},\lambda,\rho) - \rho R}{(1-\lambda\rho)\cdot E_r^*(\theta) + \lambda\rho \cdot E_r^*(\theta')}.$$

Consequently, for ensembles of linear codes and BIOS channels, the lower bound to ξ^* is the same as in (12), with a uniform i.i.d. random coding distribution, $Q^* = \left\{\frac{1}{2}, \frac{1}{2}\right\}$.

Systematic Linear Codes

A similar technique will be used now to achieve identical results for the ensemble of systematic linear codes.

By averaging S_N over this ensemble:

$$\begin{aligned}
\overline{S}_N &\overset{(a)}{=} 2^{-K(N-K)} 2^{-N} \sum_{\tilde{\boldsymbol{G}}} \sum_{\boldsymbol{v_0}} S_N(\boldsymbol{v_0}, \boldsymbol{G}) \\
&\le 2^{-K(N-K)} 2^{-N} \sum_{\tilde{\boldsymbol{G}}} \sum_{\boldsymbol{v_0}} \sum_{\boldsymbol{y} \in Y^N} \max_{\theta \in \Theta_N} \max_{\theta' \in \Theta_N} \min_{\substack{\rho \ge 0 \\ 0 \le \lambda \le 1/\rho}} \\
&\quad \left\{ e^{N(1-\lambda\rho) f_\theta(\boldsymbol{v_0},\boldsymbol{y})} \left[\sum_{m=1}^{M-1} e^{N\lambda f_{\theta'}(\boldsymbol{v_m},\boldsymbol{y})} \right]^\rho \right\} \\
&\overset{(b)}{\le} |\Theta_N|^2 \sum_{\boldsymbol{y} \in Y^N} \max_{\theta \in \Theta_N} \max_{\theta' \in \Theta_N} \min_{\substack{0 \le \rho \le 1 \\ 0 \le \lambda \le 1/\rho}} \\
&\quad \left\{ 2^{-N} \sum_{\boldsymbol{v_0}} e^{N(1-\lambda\rho) f_\theta(\boldsymbol{v_0},\boldsymbol{y})} \left[2^{-K(N-K)} \sum_{m=1}^{M-1} \sum_{\tilde{\boldsymbol{G}}} e^{N\lambda f_{\theta'}(\boldsymbol{v_m},\boldsymbol{y})} \right]^\rho \right\}
\end{aligned}$$

$$
\begin{aligned}
&\overset{(c)}{=} |\Theta_N|^2 \sum_{\boldsymbol{y}\in Y^N} \max_{\theta\in\Theta_N} \max_{\theta'\in\Theta_N} \min_{\substack{0\le\rho\le 1\\ 0\le\lambda\le 1/\rho}} \left\{ 2^{-N} \sum_{\boldsymbol{v_0}} e^{N(1-\lambda\rho) f_\theta(\boldsymbol{v_0},\boldsymbol{y})} \left[2^{-(N-K)} \sum_{\boldsymbol{v}} e^{N\lambda f_{\theta'}(\boldsymbol{v},\boldsymbol{y})} \right]^\rho \right\} \\
&\overset{(d)}{=} |\Theta_N|^2 \sum_{\boldsymbol{y}\in Y^N} \max_{\theta\in\Theta_N} \max_{\theta'\in\Theta_N} \min_{\substack{0\le\rho\le 1\\ 0\le\lambda\le 1/\rho}} \left\{ 2^{-N} \sum_{\boldsymbol{v_0}} e^{N(1-\lambda\rho) f_\theta(\boldsymbol{v_0},\boldsymbol{y})} \left[M 2^{-N} \sum_{\boldsymbol{v}} e^{N\lambda f_{\theta'}(\boldsymbol{v},\boldsymbol{y})} \right]^\rho \right\} \\
&= |\Theta_N|^2 \sum_{\boldsymbol{y}\in Y^N} \max_{\theta\in\Theta_N} \max_{\theta'\in\Theta_N} \min_{\substack{0\le\rho\le 1\\ 0\le\lambda\le 1/\rho}} \Bigg\{ M^\rho \\
&\quad \left[2^{-N} \sum_{\boldsymbol{v}} e^{N(1-\lambda\rho) f_\theta(\boldsymbol{v},\boldsymbol{y})} \right] \left[2^{-N} \sum_{\boldsymbol{v'}} e^{N\lambda f_{\theta'}(\boldsymbol{v'},\boldsymbol{y})} \right]^\rho . \Bigg\}
\end{aligned}
\tag{60}
$$

The equality in (a) is obtained by averaging over $2^{K(N-K)}$ and 2^N equiprobable values of $\boldsymbol{v_0}$ and $\tilde{\boldsymbol{G}}$ (the non–systematic part of $\boldsymbol{G}$), respectively. (b) is obtained by taking identical steps as done for ensemble of linear codes in the previous subsection (see the inequalities between (53) and (54)). In (c), we used the following equivalence for the two inner summations:

$$
\sum_{m=1}^{M-1} \sum_{\tilde{\boldsymbol{G}}} e^{N\lambda f_{\theta'}(\boldsymbol{v_m},\boldsymbol{y})} = 2^{(K-1)(N-K)} \sum_{\boldsymbol{v}} e^{N\lambda f_{\theta'}(\boldsymbol{v},\boldsymbol{y})}. \tag{61}
$$

This equivalence is proved in Section A.6 of the Appendix. In (d), we used the equality $M = 2^K$.

Finally, the upper bound to $\overline{S}_N$ achieved in (60) is identical to the one related to ensembles of linear codes (see (55)), and therefore the final lower bound to ξ^* for the case of systematic linear codes is also identical to (12) with uniform i.i.d. random coding distribution, $Q^* = \left\{\frac{1}{2}, \frac{1}{2}\right\}$.

A.2 Proof of eq. (26) and eq. (51)

Let θ^* maximize $f_\theta(\boldsymbol{v_m}, \boldsymbol{y})$, and let $F(\lambda, \rho)$ be a nonnegative function. Then,

$$
\min_{\substack{\rho\ge 0\\ 0\le\lambda\le 1/\rho}} \max_{\theta\in\Theta_N} \left\{ e^{N(1-\lambda\rho) f_\theta(\boldsymbol{v_m},\boldsymbol{y})} \cdot F(\lambda,\rho) \right\} =
$$

$$
\begin{aligned}
&= \min_{\substack{\rho\ge 0\\ 0\le\lambda\le 1/\rho}} \left\{ e^{N(1-\lambda\rho) f_{\theta^*}(\boldsymbol{v_m},\boldsymbol{y})} \cdot F(\lambda,\rho) \right\} \\
&\overset{(a)}{\le} \max_{\theta\in\Theta_N} \min_{\substack{\rho\ge 0\\ 0\le\lambda\le 1/\rho}} \left\{ e^{N(1-\lambda\rho) f_\theta(\boldsymbol{v_m},\boldsymbol{y})} \cdot F(\lambda,\rho) \right\} \\
&\le \min_{\substack{\rho\ge 0\\ 0\le\lambda\le 1/\rho}} \max_{\theta\in\Theta_N} \left\{ e^{N(1-\lambda\rho) f_\theta(\boldsymbol{v_m},\boldsymbol{y})} \cdot F(\lambda,\rho) \right\},
\end{aligned}
\tag{62}
$$

where (a) is true since the value of the function for a specific θ^* in Θ_N is always upper–bounded by the maximization of the function over $\theta \in \Theta_N$. Thus, all inequalities must be achieved with equalities.

A.3 Proof of eq. (22)

For $\alpha \in \Re$ and $\boldsymbol{y} \in \mathcal{Y}^N$, we exponentially evaluate $E[e^{N\alpha f_\theta(\boldsymbol{x},\boldsymbol{y})}]$, where the average is calculated over the ensemble of random coding distribution of the form: $Q_N(\boldsymbol{x}) = \frac{Q_N(T_{\boldsymbol{x}})}{|T_{\boldsymbol{x}}|}$, (as described in (22)):

$$\begin{aligned} E[e^{N\alpha f_\theta(\boldsymbol{X},\boldsymbol{y})}] &= \sum_{\boldsymbol{x}\in\mathcal{X}^N} Q_N(\boldsymbol{x}) e^{N\alpha f_\theta(\boldsymbol{x},\boldsymbol{y})} \\ &\overset{(a)}{=} \sum_{T_{\boldsymbol{x}|\boldsymbol{y}}\subset\mathcal{X}^N} \left|T_{\boldsymbol{x}|\boldsymbol{y}}\right| Q_N(\boldsymbol{x}) e^{N\alpha f_\theta(\boldsymbol{x},\boldsymbol{y})} \\ &= \sum_{T_{\boldsymbol{x}|\boldsymbol{y}}\subset\mathcal{X}^N} \left|T_{\boldsymbol{x}|\boldsymbol{y}}\right| \frac{e^{-N\Delta_N^*(P_{\boldsymbol{x}})}}{|T_{\boldsymbol{x}}|} e^{N\alpha f_\theta(\boldsymbol{x},\boldsymbol{y})} \\ &\overset{(b)}{=} \sum_{T_{\boldsymbol{x}|\boldsymbol{y}}\subset\mathcal{X}^N} \left|T_{\boldsymbol{x}|\boldsymbol{y}}\right| \frac{e^{-N(\Delta^*(P_{\boldsymbol{x}})-\tilde{\epsilon}_N)}}{|T_{\boldsymbol{x}}|} e^{N\alpha f_\theta(\boldsymbol{x},\boldsymbol{y})}, \end{aligned} \tag{63}$$

where $\tilde{\epsilon}_N \to 0$ as $N \to \infty$ independently of $P_{\boldsymbol{x}}$.

We should note that (a) is true since $f_\theta(\boldsymbol{x},\boldsymbol{y})$ depends on $\boldsymbol{x}$ and $\boldsymbol{y}$ only via their joint empirical distribution and the summation can be conducted over types instead, and since the average is calculated for a given $\boldsymbol{y}$, we sum over $T_{\boldsymbol{x}|\boldsymbol{y}}$. In (b) we used the convergence assumption for the random coding distributions withing the class $\mathcal{Q}$.

Thus, we continue to evaluate $E[e^{N\alpha f_\theta(\boldsymbol{x},\boldsymbol{y})}]$ as follows:

$$\begin{aligned} E[e^{N\alpha f_\theta(\boldsymbol{x},\boldsymbol{y})}] &\overset{(a)}{\doteq} \sum_{T_{\boldsymbol{x}|\boldsymbol{y}}\subset\mathcal{X}^N} \exp\Big\{N[-I_{\boldsymbol{x}\boldsymbol{y}}(X;Y) - \Delta^*(P_{\boldsymbol{x}}) + \alpha f_\theta(\boldsymbol{x},\boldsymbol{y})]\Big\} \\ &\overset{(b)}{\doteq} \exp\Big\{N \cdot \max_{P_{\boldsymbol{x}|\boldsymbol{y}}}[-I_{\boldsymbol{x}\boldsymbol{y}}(X;Y) - \Delta^*(P_{\boldsymbol{x}}) + \alpha f_\theta(\boldsymbol{x},\boldsymbol{y})]\Big\} \\ &\overset{(c)}{=} \exp\Big\{N \cdot \max_{P_{\boldsymbol{x}|\boldsymbol{y}}}[-I_{\boldsymbol{x}\boldsymbol{y}}(X;Y) - \Delta^*(P_{\boldsymbol{x}}) \\ &\qquad +\alpha\hat{E}_{\boldsymbol{x}\boldsymbol{y}} \ln P_\theta(Y|X) + \alpha\xi E_r^*(\theta)]\Big\} \\ &= \exp\Big\{N[\alpha\xi E_r^*(\theta) - \min_{P_{\boldsymbol{x}|\boldsymbol{y}}}\{I_{\boldsymbol{x}\boldsymbol{y}}(X;Y) \\ &\qquad +\Delta^*(P_{\boldsymbol{x}}) - \alpha\hat{E}_{\boldsymbol{x}\boldsymbol{y}} \ln P_\theta(Y|X)\}]\Big\} \\ &= e^{N\left[\alpha\xi E_r^*(\theta) - \min_{P_{\boldsymbol{x}|\boldsymbol{y}}} A(\theta,\alpha,P_{\boldsymbol{x}\boldsymbol{y}})\right]}, \end{aligned} \tag{64}$$

where in (a), we used the facts that $\left|T_{\boldsymbol{x}|\boldsymbol{y}}\right| \doteq e^{N\cdot H_{\boldsymbol{xy}}(X|Y)}$ and $|T_{\boldsymbol{x}}| \doteq e^{N\cdot H_{\boldsymbol{x}}(X)}$. (b) is true since the summation of the functional over $T_{\boldsymbol{x}|\boldsymbol{y}} \subset \mathcal{X}^N$ is lower bounded by its maximal value (achieved by a specific distribution $P_{\boldsymbol{x}|\boldsymbol{y}}$), and upper bounded by the product of its maximal value with $(N+1)^{|\mathcal{X}||\mathcal{Y}|}$. In (c), we expressed the minimax metric in terms of the joint empirical distribution as described in (19).

A.4 Proof of Lemma 2

In this section, we prove that when a linear code is used for a BIOS channel and the minimax decision rule is used (denoted by Ω), the error probability for the m-th message (of length N), $\boldsymbol{v_m} = (v_{m0}, \ldots, v_{m(N-1)})$, is the same for all m, that is,

$$P_{E_m}(\Omega|\theta) = P_E(\Omega|\theta) \qquad for \;\; 0 \le m \le M-1. \tag{65}$$

Considering a binary input channel, we denote the channel crossover probabilities for a single letter as $\quad P_\theta(y|v=0) \triangleq P_{\theta,0}(y) \quad$ and $\quad P_\theta(y|v=1) \triangleq P_{\theta,1}(y)$.
If the channel is also output symmetric then,

$$P_{\theta,1}(y) = P_{\theta,0}(-y), \quad \forall y \in \mathcal{Y}$$

The error probability for the m-th message using minimax decoding is:

$$\begin{aligned} P_{E_m}(\Omega|\theta) &= \sum_{\boldsymbol{y}\in{\Lambda_m}^c} P_\theta(\boldsymbol{y}|\boldsymbol{v_m}) \\ &= \sum_{\boldsymbol{y}\in{\Lambda_m}^c} \prod_{n:v_{mn}=0} P_{\theta,0}(y_n) \prod_{n:v_{mn}=1} P_{\theta,1}(y_n) \\ &= \sum_{\boldsymbol{y}\in{\Lambda_m}^c} \prod_{n:v_{mn}=0} P_{\theta,0}(y_n) \prod_{n:v_{mn}=1} P_{\theta,0}(-y_n), \end{aligned} \tag{66}$$

where

$$\begin{aligned} {\Lambda_m}^c &= \left\{ \boldsymbol{y} : \max_{\theta'} \left\{ \frac{1}{N} \ln P_{\theta'}(\boldsymbol{y}|\boldsymbol{v_{m'}}) + \xi E_r^*(\theta') \right\} \ge \max_{\theta''} \left\{ \frac{1}{N} \ln P_{\theta''}(\boldsymbol{y}|\boldsymbol{v_m}) + \xi E_r^*(\theta'') \right\}, \right. \\ &\qquad \left. for \;\; some \;\; m' \ne m \right\} \\ &= \left\{ \boldsymbol{y} : \max_{\theta'} \left\{ \sum_{n=0}^{N-1} \ln P_{\theta'}(y_n|v_{m'n}) + N\xi E_r^*(\theta') \right\} \ge \right. \\ &\qquad \left. \max_{\theta''} \left\{ \sum_{n=0}^{N-1} \ln P_{\theta''}(y_n|v_{mn}) + N\xi E_r^*(\theta'') \right\}, \quad for \;\; some \;\; m' \ne m \right\} \end{aligned}$$

$$
\begin{aligned}
= \quad & \Bigg\{ \boldsymbol{y} : \max_{\theta'} \Big\{ \sum_{\substack{t:\ v_{mt}=0 \\ v_{m't}=0}} \ln P_{\theta',0}(y_t) + \sum_{\substack{t:\ v_{mt}=0 \\ v_{m't}=1}} \ln P_{\theta',1}(y_t) + \sum_{\substack{t:\ v_{mt}=1 \\ v_{m't}=0}} \ln P_{\theta',0}(y_t) + \\
& \qquad \sum_{\substack{t:\ v_{mt}=1 \\ v_{m't}=1}} \ln P_{\theta',1}(y_t) + N\xi E_r^*(\theta') \Big\} \geq \\
& \quad \max_{\theta''} \Big\{ \sum_{\substack{t:\ v_{mt}=0 \\ v_{m't}=0}} \ln P_{\theta'',0}(y_t) + \sum_{\substack{t:\ v_{mt}=0 \\ v_{m't}=1}} \ln P_{\theta'',0}(y_t) + \sum_{\substack{t:\ v_{mt}=1 \\ v_{m't}=0}} \ln P_{\theta'',1}(y_t) + \\
& \qquad \sum_{\substack{t:\ v_{mt}=1 \\ v_{m't}=1}} \ln P_{\theta'',1}(y_t) + N\xi E_r^*(\theta'') \Big\}, \qquad for\ some\ m' \neq m \Bigg\} \\
= \quad & \Bigg\{ \boldsymbol{y} : \max_{\theta'} \Big\{ \sum_{\substack{t:\ v_{mt}=0 \\ v_{m't}=0}} \ln P_{\theta',0}(y_t) + \sum_{\substack{t:\ v_{mt}=0 \\ v_{m't}=1}} \ln P_{\theta',0}(-y_t) + \sum_{\substack{t:\ v_{mt}=1 \\ v_{m't}=0}} \ln P_{\theta',0}(y_t) + \\
& \qquad \sum_{\substack{t:\ v_{mt}=1 \\ v_{m't}=1}} \ln P_{\theta',0}(-y_t) + N\xi E_r^*(\theta') \Big\} \geq \\
& \quad \max_{\theta''} \Big\{ \sum_{\substack{t:\ v_{mt}=0 \\ v_{m't}=0}} \ln P_{\theta'',0}(y_t) + \sum_{\substack{t:\ v_{mt}=0 \\ v_{m't}=1}} \ln P_{\theta'',0}(y_t) + \sum_{\substack{t:\ v_{mt}=1 \\ v_{m't}=0}} \ln P_{\theta'',0}(-y_t) + \\
& \qquad \sum_{\substack{t:\ v_{mt}=1 \\ v_{m't}=1}} \ln P_{\theta'',0}(-y_t) + N\xi E_r^*(\theta'') \Big\}, \qquad for\ some\ m' \neq m \Bigg\}. \qquad (67)
\end{aligned}
$$

Using the following transformation to dummy variables

$$
z_n = \begin{cases} y_n, & \forall n\ :v_{mn}=0 \\ -y_n, & \forall n\ : v_{mn}=1 \end{cases}
$$

we get that

$$
\begin{aligned}
P_{E_m}(f|\theta) &= \sum_{\boldsymbol{z}\in{\Lambda_m}^c} \prod_{n:v_{mn}=0} P_{\theta,0}(z_n) \prod_{n:v_{mn}=1} P_{\theta,0}(z_n) \\
&= \sum_{\boldsymbol{z}\in{\Lambda_m}^c} \prod_{n=0}^{N-1} P_{\theta,0}(z_n), \qquad (68)
\end{aligned}
$$

where

$$
\begin{aligned}
{\Lambda_m}^c = \quad & \Bigg\{ \boldsymbol{z} : \max_{\theta'} \Big\{ \sum_{\substack{t:\ v_{mt}=0 \\ v_{m't}=0}} \ln P_{\theta',0}(z_t) + \sum_{\substack{t:\ v_{mt}=0 \\ v_{m't}=1}} \ln P_{\theta',0}(-z_t) + \sum_{\substack{t:\ v_{mt}=1 \\ v_{m't}=0}} \ln P_{\theta',0}(-z_t) + \\
& \qquad \sum_{\substack{t:\ v_{mt}=1 \\ v_{m't}=1}} \ln P_{\theta',0}(z_t) + N\xi E_r^*(\theta') \Big\} \geq \\
& \quad \max_{\theta''} \Big\{ \sum_{\substack{t:\ v_{mt}=0 \\ v_{m't}=0}} \ln P_{\theta'',0}(z_t) + \sum_{\substack{t:\ v_{mt}=0 \\ v_{m't}=1}} \ln P_{\theta'',0}(z_t) + \sum_{\substack{t:\ v_{mt}=1 \\ v_{m't}=0}} \ln P_{\theta'',0}(z_t) +
\end{aligned}
$$

$$\sum_{\substack{t:\ v_{mt}=1\\ v_{m't}=1}} \ln P_{\theta'',0}(z_t) + N\xi E_r^*(\theta'')\Big\}, \qquad for\ some\ m' \neq m\Bigg\}$$

$$= \Bigg\{\boldsymbol{z} : \max_{\theta'}\Big\{\sum_{p:v_{mp}=v_{m'p}} \ln P_{\theta',0}(z_p) + \sum_{q:v_{mq}\neq v_{m'q}} \ln P_{\theta',0}(-z_q) + N\xi E_r^*(\theta')\Big\} \geq$$
$$\max_{\theta''}\Big\{\sum_{p:v_{mp}=v_{m'p}} \ln P_{\theta'',0}(z_p) + \sum_{q:v_{mq}\neq v_{m'q}} \ln P_{\theta'',0}(z_q) + N\xi E_r^*(\theta'')\Big\},$$
$$for\ some\ m' \neq m\Bigg\}. \tag{69}$$

Now, on the one hand, (68) and (69) describe $P_{E_m}(f|\theta)$ and ${\Lambda_m}^c$, respectively, for each $0 \leq m \leq M-1$. On the other hand, we should note that the terms for $P_{E_0}(f|\theta)$ and ${\Lambda_0}^c$ (describing the case where $\boldsymbol{v_0} = \boldsymbol{0}$ is transmitted) are obtained by assigning $m = 0$ in (66) and (67). By doing that, the result terms coincide with (68) and (69), respectively (which, as mentioned before, correspond to the m-th message). This observation completes the proof.

A.5 Proof of eq. (57)

First, by the way of constructing the linear code, we know that:

$$\boldsymbol{v_{m'}} = \boldsymbol{u_{m'}}\boldsymbol{G} \oplus \boldsymbol{v_0}, \qquad \forall\ 0 \leq m' \leq M-1 \tag{70}$$

Since $1 \leq m' \leq M-1$ implies $\boldsymbol{u_{m'}} \neq \boldsymbol{0}$, then for each information vector in this set there is at least one index i for which $u_{m'i} = 1$. Consequently, the construction of each code vector $\boldsymbol{v_{m'}}$, $1 \leq m' \leq M-1$, can be written in the following way:

$$\boldsymbol{v_{m'}} = \boldsymbol{u_{m'}}\boldsymbol{G} \oplus \boldsymbol{v_0} = \boldsymbol{g_i} \oplus \Big[\sum_{j\neq i} u_{m'j}\boldsymbol{g_j}\Big] \oplus \boldsymbol{v_0},$$

where $\boldsymbol{g_i}$ stands for the i-th row in $\boldsymbol{G}$.

Therefore:

$$\begin{aligned}
\sum_{m'=1}^{M-1}\sum_{\boldsymbol{G}} e^{N\lambda f_{\theta'}(\boldsymbol{v_{m'}},\boldsymbol{y})} &= \sum_{m'=1}^{M-1}\sum_{\boldsymbol{G}\setminus\boldsymbol{g_i}}\sum_{\boldsymbol{g_i}} \exp\Big\{N f_{\theta'}(\boldsymbol{g_i} \oplus \Big[\sum_{j\neq i} u_{m'j}\boldsymbol{g_j}\Big] \oplus \boldsymbol{v_0}, \boldsymbol{y})\lambda\Big\} \\
&\overset{(a)}{=} \sum_{m'=1}^{M-1}\sum_{\boldsymbol{G}\setminus\boldsymbol{g_i}}\sum_{\boldsymbol{v}} e^{N f_{\theta'}(\boldsymbol{v},\boldsymbol{y})\lambda} \\
&= \sum_{m'=1}^{M-1} 2^{(K-1)N}\sum_{\boldsymbol{v}} e^{N f_{\theta'}(\boldsymbol{v},\boldsymbol{y})\lambda}
\end{aligned}$$

$$= \quad (M-1)2^{(K-1)N}\sum_{\boldsymbol{v}} e^{Nf_{\theta'}(\boldsymbol{v},\boldsymbol{y})\lambda}, \tag{71}$$

where (a) is true since for fixed values of m', $\boldsymbol{G}\backslash\boldsymbol{g_i}$ (in the outer summations) and $\boldsymbol{v_0}$, the row vector, which is denoted by $\boldsymbol{v}$, is fixed, causing $\boldsymbol{g_i}$ to sum up over all the binary vectors of length N.

A.6 Proof of eq. (61)

In this section, we prove the equality, which is given in (61), and used in (60).
First, by the way of constructing a systematic linear code:

$$\begin{aligned}
\boldsymbol{v_{m'}} &= \boldsymbol{u_{m'}}\boldsymbol{G} \oplus \boldsymbol{v_0} \\
&= \Big[\boldsymbol{u_{m'}}; \sum_{i=1}^{K} u_{m'i}\tilde{\boldsymbol{g}}_{\boldsymbol{i}}\Big] \oplus \boldsymbol{v_0} \\
&= \Big[\boldsymbol{u_{m'}}; \overbrace{0\dots0}^{N-K}\Big] \oplus \Big[\overbrace{0\dots0}^{K}; \sum_{i=1}^{K} u_{m'i}\tilde{\boldsymbol{g}}_{\boldsymbol{i}}\Big] \oplus \boldsymbol{v_0}, \qquad \forall\ 0 \le m' \le M-1,
\end{aligned} \tag{72}$$

where $\tilde{\boldsymbol{g}}_{\boldsymbol{i}}$ stands for the i'th row in $\tilde{\boldsymbol{G}}$ (the non-systematic part of $\boldsymbol{G}$).
We observe that for $1 \le m' \le M-1$, $\boldsymbol{u_{m'}} \neq \boldsymbol{0}$. Thus, for each information vector in this set there's at least one index i for which $u_{m'i} = 1$. Consequently, the construction of each code vector $\boldsymbol{v_{m'}}$, $1 \le m' \le M-1$, can be written in the following way:

$$\boldsymbol{v_{m'}} = \Big[\boldsymbol{u_{m'}}; \tilde{\boldsymbol{g}}_{\boldsymbol{i}}\Big] \oplus \Big[\overbrace{0\dots0}^{K}; \sum_{j\neq i} u_{m'j}\tilde{\boldsymbol{g}}_{\boldsymbol{j}}\Big] \oplus \boldsymbol{v_0}.$$

Therefore:

$$\begin{aligned}
\sum_{m'=1}^{M-1}\sum_{\tilde{\boldsymbol{G}}} e^{Nf_{\theta'}(\boldsymbol{v_{m'}},\boldsymbol{y})\lambda} &= \sum_{m'=1}^{M-1}\sum_{\tilde{\boldsymbol{G}}\backslash\tilde{\boldsymbol{g}}_i}\sum_{\tilde{\boldsymbol{g}}_i} \exp\Big\{Nf_{\theta'}\Big([\boldsymbol{u_{m'}};\tilde{\boldsymbol{g}}_{\boldsymbol{i}}] \oplus \Big[0\dots0; \sum_{j\neq i} u_{m'j}\tilde{\boldsymbol{g}}_{\boldsymbol{j}}\Big] \oplus \boldsymbol{v_0}, \boldsymbol{y}\Big)\lambda\Big\} \\
&\overset{(a)}{=} \sum_{\tilde{\boldsymbol{G}}\backslash\tilde{\boldsymbol{g}}_i}\sum_{m'=1}^{M-1}\sum_{\tilde{\boldsymbol{g}}_i} e^{Nf_{\theta'}\left([\boldsymbol{u_{m'}};\tilde{\boldsymbol{g}}_i]\oplus\boldsymbol{v},\boldsymbol{y}\right)\lambda} \\
&\overset{(b)}{\le} \sum_{\tilde{\boldsymbol{G}}\backslash\tilde{\boldsymbol{g}}_i}\sum_{m'=0}^{M-1}\sum_{\tilde{\boldsymbol{g}}_i} e^{Nf_{\theta'}\left([\boldsymbol{u_{m'}};\tilde{\boldsymbol{g}}_i]\oplus\boldsymbol{v},\boldsymbol{y}\right)\lambda} \\
&\overset{(c)}{=} \sum_{\tilde{\boldsymbol{G}}\backslash\tilde{\boldsymbol{g}}_i}\sum_{\boldsymbol{v}} e^{Nf_{\theta'}(\boldsymbol{v},\boldsymbol{y})\lambda} \\
&= 2^{(K-1)(N-K)}\sum_{\boldsymbol{v}} e^{Nf_{\theta'}(\boldsymbol{v},\boldsymbol{y})\lambda},
\end{aligned} \tag{73}$$

where (a) is true since for fixed values of m', $\tilde{\boldsymbol{G}}\backslash\tilde{\boldsymbol{g}}_{\boldsymbol{i}}$ (in the outer summations) and $\boldsymbol{v_0}$, the row vector, which is denoted by $\boldsymbol{v}$, is fixed. In (b), $m'=0$ was added to the summation, and since the inner term in the summation is always non-negative the result cannot get smaller. (c) is true since for a fixed $\boldsymbol{v}$, summing up over $0 \le m' \le M-1$ and $\tilde{\boldsymbol{g}}_{\boldsymbol{i}}$ is equivalent to the summation over all the possibilities for a vector of length N.

A.7 Equivalence between decision rules - Ω and Λ

In this section, we prove the equivalence between the minimax decision rule, Ω, maximizing the metric $f(\boldsymbol{x},\boldsymbol{y})$ (as defined in (3)), and a decision rule Λ, minimizing $\rho(\boldsymbol{x},\boldsymbol{y})$ (as defined in (17)). We will prove that for a given output $\boldsymbol{y}\in\mathcal{Y}$, each $\boldsymbol{x_1},\boldsymbol{x_2}\in\mathcal{X}$ satisfy:

$$f(\boldsymbol{x_1},\boldsymbol{y}) \ge f(\boldsymbol{x_2},\boldsymbol{y}) \Longleftrightarrow \rho(\boldsymbol{x_1},\boldsymbol{y}) \le \rho(\boldsymbol{x_2},\boldsymbol{y}). \tag{74}$$

First, we should note that $f(\boldsymbol{x},\boldsymbol{y})$ satisfies:

$$\begin{aligned}
f(\boldsymbol{x},\boldsymbol{y}) &= \max_{0\le\theta\le1} f_\theta(\boldsymbol{x},\boldsymbol{y}) \\
&= \max_{0\le\theta\le1}\Big\{\frac{1}{N}\left[\ln P_\theta(\boldsymbol{y}|\boldsymbol{x}) + N\xi E_r^*(\theta)\right]\Big\} \\
&\overset{(a)}{=} \max_{0\le\theta\le1}\Big\{\frac{1}{N}\Big[d(\boldsymbol{x},\boldsymbol{y})\ln\theta + (N-d(\boldsymbol{x},\boldsymbol{y}))\ln(1-\theta) + N\xi E_r^*(\theta)\Big]\Big\} \\
&= \max_{0\le\theta\le1}\Big\{\delta(\boldsymbol{x},\boldsymbol{y})\ln\theta + (1-\delta(\boldsymbol{x},\boldsymbol{y}))\ln(1-\theta) + \xi E_r^*(\theta)\Big\} \\
&= \max_{0\le\theta\le1} f_\theta\left(\delta(\boldsymbol{x},\boldsymbol{y})\right), \quad 0\le\delta(\boldsymbol{x},\boldsymbol{y})\le1 \\
&= f\left(\delta(\boldsymbol{x},\boldsymbol{y})\right), \quad 0\le\delta(\boldsymbol{x},\boldsymbol{y})\le1.
\end{aligned} \tag{75}$$

In (a), we used the following representation for the BSC transition probability:

$$P_\theta(\boldsymbol{y}|\boldsymbol{x}) = \theta^{d(\boldsymbol{x},\boldsymbol{y})}(1-\theta)^{N-d(\boldsymbol{x},\boldsymbol{y})}.$$

We conclude that the value of $f(\boldsymbol{x},\boldsymbol{y})$ is equal for all code vectors with the same (normalized) Hamming distance from $\boldsymbol{y}$, and therefore can be defined as $f(\delta(\boldsymbol{x},\boldsymbol{y}))$, $\quad 0\le\delta(\boldsymbol{x},\boldsymbol{y})\le1$.

Next, we now prove that $f(\boldsymbol{x},\boldsymbol{y})$ has the same value for a code vector $\boldsymbol{x}$ and its complement, $\overline{\boldsymbol{x}}$:

$$\begin{aligned}
f(\overline{\boldsymbol{x}},\boldsymbol{y}) &= f(\delta(\overline{\boldsymbol{x}},\boldsymbol{y})) \\
&= f(1-\delta(\boldsymbol{x},\boldsymbol{y})) \\
&= \max_{0\le\theta\le1} f_\theta(1-\delta(\boldsymbol{x},\boldsymbol{y}))
\end{aligned}$$

$$
\begin{aligned}
&= \max_{0\le\theta\le 1}\Big\{(1-\delta(\boldsymbol{x},\boldsymbol{y}))\ln\theta+\delta(\boldsymbol{x},\boldsymbol{y})\ln(1-\theta)+\xi E_r^*(\theta)\Big\}\\
&\overset{(a)}{=} \max_{0\le\tilde{\theta}\le 1}\Big\{(1-\delta(\boldsymbol{x},\boldsymbol{y}))\ln\left(1-\tilde{\theta}\right)+\delta(\boldsymbol{x},\boldsymbol{y})\ln\tilde{\theta}+\xi E_r^*(1-\tilde{\theta})\Big\}\\
&\overset{(b)}{=} \max_{0\le\tilde{\theta}\le 1}\Big\{(1-\delta(\boldsymbol{x},\boldsymbol{y}))\ln\left(1-\tilde{\theta}\right)+\delta(\boldsymbol{x},\boldsymbol{y})\ln\tilde{\theta}+\xi E_r^*(\tilde{\theta})\Big\}\\
&= \max_{0\le\tilde{\theta}\le 1} f_{\tilde{\theta}}(\delta(\boldsymbol{x},\boldsymbol{y}))\\
&= f(\delta(\boldsymbol{x},\boldsymbol{y}))\\
&= f(\boldsymbol{x},\boldsymbol{y}). \qquad (76)
\end{aligned}
$$

In (a), we changed the variable in the maximization, $\tilde{\theta}=1-\theta$, and (b) is true since for the BSC model the ML error exponent, $E_r^*(\theta)$, is symmetric around $\theta=\frac{1}{2}$ (see (42)).

Using the fact that both $f(\delta(\boldsymbol{x},\boldsymbol{y}))$ and $\rho(\delta(\boldsymbol{x},\boldsymbol{y}))$ are equal for $\delta(\boldsymbol{x},\boldsymbol{y})$ and $1-\delta(\boldsymbol{x},\boldsymbol{y})$, it is sufficient to prove (74) for $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ satisfying $\delta(\boldsymbol{x_1},\boldsymbol{y})\le\frac{1}{2}$ and $\delta(\boldsymbol{x_2},\boldsymbol{y})\le\frac{1}{2}$ (and thus $\rho(\boldsymbol{x_1},\boldsymbol{y})=\delta(\boldsymbol{x_1},\boldsymbol{y})$, $\rho(\boldsymbol{x_2},\boldsymbol{y})=\delta(\boldsymbol{x_2},\boldsymbol{y})$).

In the rest of the proof, we will denote $\delta(\boldsymbol{x_1},\boldsymbol{y})\overset{\Delta}{=}\delta_1$, $\delta(\boldsymbol{x_2},\boldsymbol{y})\overset{\Delta}{=}\delta_2$. It is therefore sufficient to show that

$$
f(\delta_1)\ge f(\delta_2)\Longleftrightarrow 0\le\delta_1\le\delta_2\le\frac{1}{2}. \qquad (77)
$$

This equivalence will be shown in two steps:

First, we note that $0\le\delta_1\le\delta_2\le\frac{1}{2}$ satisfy that $\forall\, 0\le\theta\le\frac{1}{2}$:

$$
\delta_1\ln\left(\frac{\theta}{1-\theta}\right)\ge\delta_2\ln\left(\frac{\theta}{1-\theta}\right). \qquad (78)
$$

By adding $\ln(1-\theta)+\xi E_r^*(\theta)$ to both sides of (78) we get:

$$
\delta_1\ln\left(\frac{\theta}{1-\theta}\right)+\ln(1-\theta)+\xi E_r^*(\theta)\ge\delta_2\ln\left(\frac{\theta}{1-\theta}\right)+\ln(1-\theta)+\xi E_r^*(\theta) \qquad (79)
$$

or

$$
\delta_1\ln\theta+(1-\delta_1)\ln(1-\theta)+\xi E_r^*(\theta)\ge\delta_2\ln\theta+(1-\delta_2)\ln(1-\theta)+\xi E_r^*(\theta). \qquad (80)
$$

This inequality is true for the values of θ, which maximize the both sides of (80). i.e.:

$$
\begin{aligned}
&\max_{0\le\theta\le\frac{1}{2}}\{\delta_1\ln\theta+(1-\delta_1)\ln(1-\theta)+\xi E_r^*(\theta)\}\ge\\
&\quad\max_{0\le\theta\le\frac{1}{2}}\{\delta_2\ln\theta+(1-\delta_2)\ln(1-\theta)+\xi E_r^*(\theta)\} \qquad (81)
\end{aligned}
$$

or

$$
\max_{0\le\theta\le\frac{1}{2}} f_\theta(\delta_1)\ge\max_{0\le\theta\le\frac{1}{2}} f_\theta(\delta_2). \qquad (82)
$$

In order to complete the proof, one must broaden the maximization ranges over θ in (82) into $0 \leq \theta \leq 1$. In order to justify that this broadening is possible, we present the following observation:

Each $0 \leq \delta \leq \frac{1}{2}$ satisfy that $\forall \frac{1}{2} \leq \theta \leq 1$:

$$\delta \ln\left(\frac{\theta}{1-\theta}\right) \leq (1-\delta) \ln\left(\frac{\theta}{1-\theta}\right). \tag{83}$$

By adding $\ln(1-\theta) + \xi E_r^*(\theta)$ to both sides of (83) we get:

$$\delta \ln\left(\frac{\theta}{1-\theta}\right) + \ln(1-\theta) + \xi E_r^*(\theta) \leq (1-\delta) \ln\left(\frac{\theta}{1-\theta}\right) + \ln(1-\theta) + \xi E_r^*(\theta) \tag{84}$$

or

$$\delta \ln\theta + (1-\delta)\ln(1-\theta) + \xi E_r^*(\theta) \leq \delta \ln(1-\theta) + (1-\delta)\ln\theta + \xi E_r^*(\theta). \tag{85}$$

Using the fact that for the BSC model the ML error exponent, $E_r^*(\theta)$, is symmetric around $\theta = \frac{1}{2}$ (see (42)), we can rewrite (85) as:

$$\delta \ln\theta + (1-\delta)\ln(1-\theta) + \xi E_r^*(\theta) \leq \delta \ln(1-\theta) + (1-\delta)\ln\theta + \xi E_r^*(1-\theta) \tag{86}$$

or

$$f_\theta(\delta(\boldsymbol{x}, \boldsymbol{y})) \leq f_{1-\theta}(\delta(\boldsymbol{x}, \boldsymbol{y})). \tag{87}$$

The meaning of (87) is that when $0 \leq \delta \leq \frac{1}{2}$, for each $\frac{1}{2} \leq \theta \leq 1$, $f_\theta(\delta)$ is always upper bounded by $f_{1-\theta}(\delta)$ where $0 \leq 1-\theta \leq \frac{1}{2}$. Thus, maximization of $f_\theta(\delta)$ over $0 \leq \theta \leq 1$ is obviously accomplished by θ in $\left[0, \frac{1}{2}\right]$.

Therefore, (82) finally becomes:

$$\max_{0 \leq \theta \leq 1} f_\theta(\delta_1) \geq \max_{0 \leq \theta \leq 1} f_\theta(\delta_2) \tag{88}$$

thus,

$$0 \leq \delta_1 \leq \delta_2 \leq \frac{1}{2} \quad \Leftrightarrow \quad f(\delta_1) \geq f(\delta_2), \tag{89}$$

and the proof is complete.

References

[1] I. Csiszár, "Linear Codes for Sources and Source Networks: Error Exponents, Universal Coding," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 585-592, July 1982.

[2] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press 1981.

[3] M. Feder and A. Lapidoth, "Universal Decoders for Channels with Memory," *IEEE Trans. Inform. Theory*, vol. IT–44, no. 5, pp. 1726–1745, September 1998.

[4] M. Feder and N. Merhav, "Universal Composite Hypothesis Testing - A Competitive Minimax Approach," *IEEE Trans. Inform. Theory*, vol. IT-48, pp. 1504-1517, June 2002.

[5] V. D. Goppa, "Nonprobabilistic Mutual Information Without Memory," *Probl. Cont. Information Theory*, Vol. 4, pp. 97-102, 1975.

[6] A. Lapidoth and J. Ziv, "On the Universality of the LZ-based Decoding Algorithm," *IEEE Trans. Inform. Theory*, vol. IT–44, no. 5, pp. 1746–1755, September 1998.

[7] N. Merhav and M. Feder, "Minimax Universal Decoding with an Erasure Option," *IEEE Trans. Inform. Theory*, vol. IT–53, no. 5, pp. 1664–1675, May 2007.

[8] N. Shulman, "Communication over an Unknown Channel via Common Broadcasting," Ph.D. dissertation, Tel Aviv University, July 2003.

[9] A. Somekh-Baruch and N. Merhav, "Achievable error exponents for the private fingerprinting game," *IEEE Trans. Inform. Theory*, vol. IT–53, no. 5, pp. 1827–1838, May 2007.

[10] A. J. Viterbi and J. K. Omura, *Principles of Digital Communication and Coding.* McGraw–Hill, 1979.

[11] O. Zeitouni, J. Ziv, and N. Merhav, "When is the Generalized Likelihood Ratio Test Optimal?" *IEEE Trans. Inform. Theory*, vol. IT–38, no. 5, pp. 1597–1602, September 1992.

[12] J. Ziv, "Universal Decoding for Finite-State Channels," *IEEE Trans. Inform. Theory*, vol. IT–31, no. 4, pp. 453–460, July 1985.