On the Error Exponent and Capacity Games of Private Watermarking Systems

Anelia Somekh-Baruch and Neri Merhav[1]

Abstract

Watermarking systems are analyzed as a game between an information hider, a decoder and an attacker. The information hider is allowed to cause some tolerable level of distortion to the original data within which the message is hidden, and the resulting distorted data can suffer some additional amount of distortion caused by an attacker who aims at erasing the message. Two games are investigated: the error exponent game and the coding capacity game. Motivated by a worst-case approach, we assume that the attacker is informed of the hiding strategy taken by the information hider and the decoder, which are uninformed of the attacking scheme. This approach leads to the maximin error exponent and maximin coding capacity as objective functions. It is assumed that the host data is drawn from a finite-alphabet memoryless stationary source, and its realization (side information) is available at the encoder and the decoder. A single-letter expression for the maximin error exponent is found under large deviations distortion constraints. Moreover, we find an asymptotically optimal random coding distribution, a universal decoder, and a worst-case attack channel. It is proved that there is a saddle-point in the asymptotic exponent and that the minimax and the maximin error exponents are equal. Finally, a single letter expression for the coding capacity, i.e., the maximin reliable information rate, is found.

1 Introduction

Due to the dramatic increase in the use of digital data, its availability in the Internet and the ease at which data can be copied and stored, there is increased interest in watermarking codes in their various applications, as reflected in the growing number of papers describing watermarking schemes and the emerging standards for practical systems (see, e.g., [3], [18]). Watermarking is a form of hiding or embedding information in a host digital data, referred to as *covertext*, such as video, image, or an audio signal, so as to enable copyright protection or data encryption.

There are some conflicting requirements for a good watermarking code, the relative importance of which may depend on the application. The first is *perceptual transparency*, that is, the watermark should be invisible to the naked eye, or when audio signals are concerned,

[1]The authors are with the Department of EE, Technion - I.I.T., Haifa, Israel. Emails: anelia@tx.technion.ac.il, merhav@ee.technion.ac.il .

the original and watermarked signals should be indistinguishable by the human auditory system. Another, very important requirement is *robustness* to attacks, caused by either innocent data processing, e.g., data compression or filtering, or malicious parties who wish to invalidate the watermark by introducing a little amount of distortion to the original data. Usually, the *information rate* is also required to be high, i.e., the embedded message should convey as much information as possible. In some applications, the very existence of a watermark in the data should remain obscure, let alone the removal of the watermarked message.

Recently, a few information theorists have begun to devote research efforts to investigate watermarking systems. In [17] watermarking was characterized from the information theoretic point of view as a capacity game between an information hider and an attacker. In that work, the goal of the information hider (encoder) was to embed a watermark in the host data (*covertext*) at the highest possible information rate, while the distortion between the watermarked data (*stegotext*) and the original covertext is kept smaller than some threshold. An attacker may seek to destroy the hidden message by creating a distorted version of the stegotext (*forgery*) while the additional distortion it may cause is also limited.

As the transmitted information is the watermark and the covertext serves only as a medium in which the transmitted data is embedded, the covertext data may be considered as side information (SI) at the transmitter. Two extreme scenarios of the game may be considered. In the first, the side information is available to the encoder only, while in the second, it is available also to the decoder. These two scenarios, are referred to as the *public game* and the *private game* correspondingly [1].

Several papers introduce various models for the watermarking game and study these models, referring usually to either capacity or error exponent as a performance criterion. There are a few common assumptions in some of the models of the watermarking game that have been studied thus far:

1. The covertext is modeled by a stochastic process, usually memoryless and stationary, that is unavailable to the attacker, hence, the covertext $\mathbf{U}$, the stegotext $\mathbf{X}$ and the forgery $\mathbf{Y}$, (in this order) form a Markov chain.

2. Two single-letter distortion measures $d_1 : \mathcal{U} \times \mathcal{X} \to \mathbb{R}_+$ and $d_2 : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ are defined, where $\mathcal{U}, \mathcal{X}, \mathcal{Y}$ are the alphabets of the covertext, stegotext, and the forgery, respectively. The encoder is constrained to meet a certain distortion constraint w.r.t. distortion measure d_1, and the attacker is constrained to meet a certain distortion constraint w.r.t. distortion measure d_2.

3. It is assumed that the attacker is allowed to apply a stochastic strategy, and thus is modeled as an *attack channel*, that is, a sequence of conditional probability mass functions (pmf's) given the stegotext sequence.

4. In some papers, a worst-case approach is taken, assuming that the attacker is informed of the information hider's strategy (encoder), which in turn, is allowed to act stochastically, i.e., to apply a random coding scheme using a secret key shared with the decoder. This assumption creates a communication model that is closely related to the research area of coding and decoding for unknown channels (see [11] for a survey on this subject). Hence, unlike in classical information theory, where random coding is used as a

proof technique for existence of good codes and achievability, the role of randomized coding in the setup of the watermarking game, is for security and resilience to attacks. Moreover, in this setup, random coding cannot be used as a means of proving the existence of a deterministic code which is good simultaneously for all the family of attack channels of interest (unless this family of channels is small).

5. The value of the game is usually the Shannon capacity, or the exponent of erroneously decoding the message. The Shannon capacity is relevant as a performance criterion when the maximum number of distinct messages that can be hidden in the covertext is of interest. The number of different messages M may correspond to different possible owners. The error exponent is an interesting target function as in many practical watermarking applications it is more desirable to obtain the lowest possible probability of error at a given information rate, rather than to maximize the information rate.

In [17] the capacity of the private game is analyzed. The covertext is assumed to be a discrete memoryless source (DMS). It is assumed that the decoder *is informed* of the attack channel, and thus, can apply maximum likelihood decoding. The Shannon coding capacity is found under the very restrictive assumption that the attacker can use only random memoryless (or block-wise memoryless) strategies which meet some expected distortion constraint.

Cohen and Lapidoth [1] characterized the Gaussian watermarking game of the Shannon capacity. In both private and public versions of the game, the covertext is assumed to be a zero-mean, Gaussian white noise process. The coding capacity of the private and public versions of the game with almost sure constraints are shown to be equal under these conditions. Specifically, it is given an explicit expression which is a function of the allowable distortion levels D_3, D_2 and the variance σ_u^2 of the Gaussian covertext source (see also [2] for an extension of this model).

The setting of the game introduced in [13] refers to the private game, where the covertext is assumed to be a DMS, the attacker has perfect knowledge of the watermarking channel, and is allowed to apply onln *causal* strategies which satisfy a certain distortion constraint. The encoder is uninformed of the attack strategy, but the decoder is. The target function of the game is the Gallager upper bound to the average probability of error. An exact analysis of the Gallager bound was performed and a single-letter expression for it was provided. This single-letter expression is a result of the fact that if the information hider utilizes a discrete memoryless channel (DMC) to generate random codewords for every covertext message, and that the worst case causal attack channel, for this specific game, is memoryless as well.

In many cases, watermarking applications do not require *full* decoding of the message, but merely to examine whether a watermark exists in the data, and if so, whether it matches a specified hypothesized message i out of M messages. In other words, only a binary output is required: either a "yes" or "no" answer to the question whether message number i has been hidden in the data. The selected message is not known in advance to either the encoder and decoder. Motivated by this fact, Steinberg and Merhav [19] analyzed watermarking codes as identification (ID) codes with side information. The attacker is modeled as a DMC and full knowledge of the attack channel is available to both the encoder and the decoder. For the private version of the game, formulas are derived in [19] for the ID capacity and for achievable error exponents. Lower and upper bounds on the ID capacity are found for the public version of the game.

Merhav [14] considered the following problem of hypothesis testing: Given a sequence observed at the output of an unknown channel, which is only known to belong to a class $\mathcal{W}$ of finite-state channels with a given number of states, one should decide which one of two given sequences was fed into the channel. As stated in [14], such a situation may arise in watermark detection, or more accurately, in resolving rightful ownerships of watermarked data [4]. A universal asymptotically optimum decision rule, based on Ziv's universal decoding metric for finite-state channels [23], is introduced. Its optimality is in the sense of a Neyman-Pearson-like criterion.

In this work, we focus on the private version of the game where the covertext is drawn from a finite-alphabet DMS. We investigate the error exponent game and the capacity game. Since expected distortion constraints on the attack channel allow for non-ergodic, zero-capacity channels, more restrictive constraints are required to obtain meaningful results. One option is to impose an "in-probability" distortion constraint on the attacker, namely, to assume a vanishingly small probability to the event that the distortion $d_2(\mathbf{X}, \mathbf{Y})$, caused by the attacker, exceeds a certain per-symbol distortion level D_2. While this suffices for obtaining a meaningful result on the maximin capacity, when it comes to error exponents, such constraints no longer suffice as they allow for arbitrarily slow convergence rates, which in turn allow for zero-error-exponent attack channels. Thus, there is a need to quantify the allowed *rate* of probability of large distortion, which leads us to a large deviations distortion constraint on the probability of the distortion caused by the attacker to exceed a certain level.

A single-letter expression for the maximin random coding error exponent is found under the assumption that the random coding distribution belongs to the wide class of exchangeable distributions[2] which are invariant to permutations (to be defined later). Moreover, we find an asymptotically optimal random coding distribution, a universal decoder, and a worst-case attack channel that is independent of the random coding distribution, and thus, it is proved that there is a saddle-point of the asymptotic error exponent. Thus, the minimax and the maximin error exponents are equal. We further show that the single-letter expression of the error exponent remains valid in a certain range of high rates, even for the wider class of encoders (see also [1]) which allows for *any*[3] randomized coding scheme with a secret key shared by the encoder and decoder. In this setup, we also establish a converse theorem and find the capacity of the watermarking game.

Our work extends the results of [13], where, as mentioned earlier, an *exact* analysis of a Gallager-type [8] lower bound on the error exponent was performed. It should be emphasized that by relaxing the aspiration of the exact analysis as in [13] and resorting to asymptotic analysis, we open the door to the derivation of considerably stronger results in three aspects: (a) we provide an asymptotic expression for the exact error exponent rather than merely a Gallager bound, (b) in contrast to [13], our analysis includes also non-causal attack channels[4],

[2]This class includes memoryless and constant composition distributions.

[3]In this wide class of randomized codes, the codewords may not necessarily be drawn independently and with identical distribution given the covertext, as is done in ordinary random coding.

[4]This extension is very important as in watermarking applications, real-time processing is usually not

and (c) we relax the unnatural implicit assumption, taken in [13], that the decoder knows the channel.

This work also extends the private game results of [1] in several aspects: (a) we analyze the general finite alphabet DMS case, as opposed to [1] where the covertext is assumed to be Gaussian, (b) we investigate the error exponent as well as the coding capacity, while in [1] the capacity is investigated, (c) the almost sure distortion constraint imposed in [1] can be viewed as special case of the large deviations distortion constraint we define, and (d) the distortion function in [1] is the squared error, while we address general distortion functions.

Our results can also be viewed as a generalization of those of Stiglitz [20],[21] regarding error exponents for the jamming game as we provide an asymptotic expression for the exact error exponent rather than a lower bound, we do not impose a-priori restrictions on the decoder (as in [20],[21]), and, as mentioned earlier, we do not assume causality of the attack channels.

The outline of the paper is as follows: In Section 2, some notation conventions are defined. A preliminary statement of the problem, which is relevant to the entire paper is given in Section 3. Section 4 is dedicated to the results concerning the random coding error exponent, and in Section 5 the sphere packing bound is presented. Finally, Section 6 is devoted to a derivation of a direct and converse theorems yielding the capacity formula.

2 Notation and Definitions

Henceforth, we adopt the following notation conventions. Random variables will be denoted by capital letters, while their realizations will be denoted by the respective lower case letters. Random vectors of dimension n will be denoted by boldface letters. Thus, for example, if $\mathbf{X}$ denotes a random vector $(X_1, \ldots, X_n)$, then $\mathbf{x} = (x_1, \ldots, x_n)$ will designate a specific sample value of $\mathbf{X}$. The alphabet of a scalar RV X will be designated by a caligraphic letter $\mathcal{X}$. The n-fold Cartesian power of a generic alphabet $\mathcal{A}$, that is, the set of all n-vectors over $\mathcal{A}$, will be denoted $\mathcal{A}^n$.

The set of probability mass functions (pmf's), defined on a alphabet $\mathcal{X}$, will be denoted by $\mathcal{P}(\mathcal{X})$, and the set of conditional pmf's from $\mathcal{U}$ to $\mathcal{X}$ will be denoted $\mathcal{P}(\mathcal{X}|\mathcal{U})$, i.e.,

$$\mathcal{P}(\mathcal{X}|\mathcal{U}) = \left\{ P(X|U): \; P(x|u) \geq 0, \; \sum_{x' \in \mathcal{X}} P(x'|u) = 1, \; \forall (u,x) \in \mathcal{U} \times \mathcal{X} \right\}. \tag{1}$$

The notation $\mathbf{1}\{A\}$, where A is an event, will designate the indicator function of A, i.e., $\mathbf{1}\{A\} = 1$ if A occurs, and $\mathbf{1}\{A\} = 0$ otherwise. We adopt the convention that if the set T is empty, then $\min_{t \in T} f(t) = \infty$, and similarly, $\max_{t \in T} f(t) = -\infty$. The notation $c_n \doteq d_n$, for two sequences $\{c_n\}_{n \geq 1}$ and $\{d_n\}_{n \geq 1}$, will express asymptotic equality in the exponential scale, i.e., $\liminf_{n \to \infty} \frac{1}{n} \log \frac{c_n}{d_n} = 0$. Similarly, $c_n \stackrel{\cdot}{\geq} d_n$ will stand for $\lim_{n \to \infty} \frac{1}{n} \log \frac{c_n}{d_n} \geq 0$, and so on.

The empirical pmf induced by a vector $\mathbf{x} \in \mathcal{X}^n$ is the vector $\hat{P}_{\mathbf{x}} = \{n_{\mathbf{x}}(a)/n, \; a \in \mathcal{X}\}$, where $n_{\mathbf{x}}(a)/n$ is the relative frequency of the letter a in the vector $\mathbf{x}$. The set of empirical

necessary.

pmf's induced by all n-vectors in $\mathcal{X}^n$ will be denoted by $\mathbb{P}_n(\mathcal{X})$, i.e.,

$$\mathbb{P}_n(\mathcal{X}) = \left\{\hat{P}_{\mathbf{x}}\right\}_{\mathbf{x}\in\mathcal{X}^n}. \tag{2}$$

The type class $T_{\mathbf{x}}$ is the set of n-vectors $\tilde{\mathbf{x}}$ such that $\hat{P}_{\tilde{\mathbf{x}}} = \hat{P}_{\mathbf{x}}$. Similarly, the joint empirical pmf induced by two n-vectors, $\mathbf{x}, \mathbf{y}$, is the vector $\hat{P}_{\mathbf{xy}} = \{n_{\mathbf{xy}}(a,b)/n,\ a \in \mathcal{X}^n, b \in \mathcal{Y}^n\}$ where $n_{\mathbf{xy}}(a,b)$ is the number of occurrences of $(x_i, y_i) = (a,b)$. The type class $T_{\mathbf{xy}}$ is the set of all pairs of n-vectors $\tilde{\mathbf{x}} \in \mathcal{X}^n, \tilde{\mathbf{y}} \in \mathcal{Y}^n$, such that $\hat{P}_{\tilde{\mathbf{x}}\tilde{\mathbf{y}}} = \hat{P}_{\mathbf{xy}}$. The conditional type class $T_{\mathbf{y}|\mathbf{x}}$, for a given $\mathbf{x}$, is the set of all n-vectors $\tilde{\mathbf{y}} \in \mathcal{Y}^n$ such that $T_{\mathbf{x}\tilde{\mathbf{y}}} = T_{\mathbf{xy}}$, and the conditional empirical pmf $\hat{P}_{\mathbf{y}|\mathbf{x}}$ is defined by $\hat{P}_{\mathbf{y}|\mathbf{x}}(y|x) = \frac{\hat{P}_{\mathbf{xy}}(x,y)}{\hat{P}_{\mathbf{x}}(x)}$, $\forall x \in \mathcal{X} : \hat{P}_{\mathbf{x}}(x) > 0$.

For a given empirical pmf $\hat{P} \in \mathbb{P}_n(\mathcal{A})$, define the set of conditional empirical pmf's,

$$\mathbb{P}_n(\mathcal{B}, \hat{P}) = \left\{P \in \mathcal{P}(\mathcal{B}|\mathcal{A}) :\ n\hat{P}(a)P(b|a) \text{ is an integer for all } (a,b) \in \mathcal{A}\times\mathcal{B}\right\}. \tag{3}$$

Conditional pmf's defined on n-vectors will be denoted by a subscript n, e.g., Q_n, as opposed to conditional pmf's defined on scalar alphabets which will be denoted without a subscript, e.g., Q.

The variational distance between two pmfs P and P', defined on the same set $\mathcal{A}$, will be denoted by $||P - P'||$, i.e.,

$$||P - P'|| = \sum_{a\in\mathcal{A}} |P(a) - P'(a)|. \tag{4}$$

Information-theoretic quantities, such as the entropy of the random variable X whose pmf is P will be denoted by either $H_P(X)$ or $H(P)$. Similarly, the mutual information between X and Y given U, with joint pmf P will be denoted by $I_P(X;Y|U)$, etc. The divergence or Kullback-Leibler distance between two pmf's P and Q on $\mathcal{A}$, where $|\mathcal{A}| < \infty$, is defined as $D(P\|Q) = \sum_{a\in\mathcal{A}} P(a)\log\frac{P(a)}{Q(a)}$, where we use the convention that $0\log 0 = 0$ and $p\log\frac{p}{0} = \infty$.

Information-theoretic quantities governed by empirical measures induced by the n-vectors $\mathbf{u}, \mathbf{x}, \mathbf{y}$, will have a special notation, e.g.,

$$\begin{aligned}\hat{H}_{\mathbf{x}} &= H_{\hat{P}_{\mathbf{x}}}(X) \\ \hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}} &\stackrel{\triangle}{=} I_{\hat{P}_{\mathbf{uxy}}}(X;Y|U).\end{aligned} \tag{5}$$

The notation $U \leftrightarrow X \leftrightarrow Y$ will signify that the RV's U, X, Y, in this order, form a Markov chain.

3 Preliminaries

Let $\mathbf{U}$ designate the covertext data within which the watermark will be hidden. The n-vector $\mathbf{U}$ is composed of n i.i.d. RV's whose joint pmf is denoted by P_U^n with marginal pmf P_U, and we shall assume, without loss of generality, that $\min_{u\in\mathcal{U}} P_U(u) > 0$. Let $\mathbf{X}$ stand for the watermarked data, and let $\mathbf{Y}$ denote the forgery, i.e., the output of the attack channel. It will be assumed that scalar RV's all take on values in the same alphabet, i.e., $\mathcal{X} = \mathcal{Y} = \mathcal{U}$, which will be denoted sometimes generically by $\mathcal{A}$. Let $d_1 : \mathcal{U} \times \mathcal{X} \to \mathbb{R}_+$ denote a single-letter distortion measure. For the sake of simplicity[5], we shall assume that $d_1(a, a) = 0$ for all $a \in \mathcal{A}$. For two vectors $\mathbf{u}$ and $\mathbf{x}$, let $d_1(\mathbf{u}, \mathbf{x}) = \sum_{i=1}^{n} d_1(u_i, x_i)$.

A secret key Θ is a random variable, independent the watermark message, known to both the encoder and decoder, and unknown to the attacker[6]

Let Γ stand for the alphabet of the random variable Θ, and let P_Θ stand for the distribution of Θ.

Definition 1. *A rate-R watermarking code of size n is a triple $(\mathcal{M}_n, f_n, g_n)$, where $\mathcal{M}_n$ is the watermarking message set, f_n is the encoder, and g_n is the decoder. The message set is*

$$\mathcal{M}_n = \left\{1, ..., \lfloor 2^{nR} \rfloor\right\}. \tag{6}$$

The encoder is a function

$$f_n : \Gamma \times \mathcal{U}^n \times \mathcal{M}_n \to \mathcal{X}^n, \tag{7}$$

which maps the secret key realization θ, the covertext data $\mathbf{u}$ and a watermark message $m \in \mathcal{M}_n$ into a stegotext vector (or, watermarked vector) $\mathbf{x} \in \mathcal{X}^n$. The decoder is a function

$$g_n : \ \Gamma \times \mathcal{U}^n \times \mathcal{Y}^n \to \mathcal{M}_n, \tag{8}$$

which maps the secret key realization θ, the covertext message $\mathbf{u}$ and the forgery $\mathbf{y}$ to an estimated watermark message $\hat{m}$.

The set of all n-th order decoders, i.e., the set of all functions $g_n : \Gamma \times \mathcal{U}^n \times \mathcal{Y}^n \to \mathcal{M}_n$, will be denoted by $\mathcal{G}_n$. For the sake of simplicity, the decoder will be denoted by $g_n : \ \mathcal{U}^n \times \mathcal{Y}^n \to \mathcal{M}_n$, i.e., we shall omit Γ from the notation, where it should be understood that the decoder depends also on the specific codebook corresponding to $\mathbf{u}$.

Let $d_2 : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ denote another single-letter distortion measure, where again, we shall assume that for all $a \in \mathcal{A}$, $d_2(a, a) = 0$, and denote $d_2(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} d_2(x_i, y_i)$ for $\mathbf{x} \in \mathcal{X}^n$ and $\mathbf{y} \in \mathcal{Y}^n$. An attack subject to distortion level D_2 with respect to d_2 and a given exponential rate $\lambda > 0$ is a (possibly randomized) map $h_n : \mathcal{X}^n \to \mathcal{Y}^n$ where,

$$\Pr\left\{d_2(\mathbf{X}, \mathbf{Y}) > nD_2 | \mathbf{X} = \mathbf{x}\right\} \le e^{-\lambda n} \quad \forall \mathbf{x} \in \mathcal{A}^n. \tag{9}$$

We henceforth refer to (9) as a *large deviations distortion constraint.* As we allow for randomized attacks, we model the action of the attacker by a channel W from $\mathbf{X}$ to $\mathbf{Y}$. The set of attack channels which satisfy (9) will be denoted by $\mathcal{W}_n = \mathcal{W}_n(\lambda, D_2)$. The large deviations distortion constraint investigated here is less strict than the almost-sure constraint presented in [1], which is equivalent to a special case of (9) where $\lambda = \infty$.

[5]It is natural to make this assumption as all scalar RV's take on values in the same alphabet.

[6]As opposed to the public watermarking game, when the private watermarking game is concerned, one can assume, without loss of generality, that Θ is independent of $\mathbf{U}$. This is because any scheme which allows for dependence can be outperformed by one which does not with Θ taking values in a bigger alphabet. To see this, consider arbitrary $(P_{\Theta|\mathbf{U}}, f_n, g_n)$. Define a one to one mapping from $\mathcal{U}^n$ to $1, ..., |\mathcal{U}|^n$, let Θ_i be the RV distributed according to $P_{\Theta|\mathbf{u}_i}$ independent of $\mathbf{U}$, and regard the alternative secret key given by the random vector $\bar{\Theta} = (\Theta_1, ..., \Theta_{|\mathcal{U}|^n})$. Since $\mathbf{u}$ is available also at the decoder, the decoder can use the realization θ_i corresponding to the actual $\mathbf{u}$ to perform the decoding using g_n. By definition, this scheme is identical to the original one, and thus it attains the same performance.

4 The Random Coding Error Exponent

In this section, it is assumed that the encoder adopts the following random coding mechanism: For each $\mathbf{u} \in \mathcal{U}^n$, the encoder generates independently, at random, $\lfloor 2^{nR} \rfloor$ codewords by using a certain channel $Q_n : \mathcal{U}^n \to \mathcal{X}^n$ fed by $\mathbf{u}$. The channel Q_n will be referred to as the *watermarking channel* [17]. It is assumed that the decoder is informed of the randomly selected codebook for every $\mathbf{u} \in \mathcal{U}^n$.

It should be emphasized, that in this section, the structure of the randomized encoder is limited, that is, as opposed to the general definition of the encoder which allows for an arbitrary joint distribution of the entire codebook (e.g., deterministic codes), here, the codewords corresponding to $\mathbf{u}$ are assumed i.i.d. given $\mathbf{u}$.

Our results, in this section, refer to two kinds of distortion constraints imposed on the encoder. The first one is an expected distortion constraint:

$$E\left\{d_1(\mathbf{U},\mathbf{X})|\mathbf{U}=\mathbf{u}\right\} \le nD_1 \quad \forall \mathbf{u} \in \mathcal{U}^n, \tag{10}$$

and the second one is a large deviations constraint, similar to (9):

$$\Pr\left\{d_1(\mathbf{U},\mathbf{X}) > nD_1|\mathbf{U}=\mathbf{u}\right\} \le e^{-\nu n} \quad \forall \mathbf{u} \in \mathcal{U}^n \tag{11}$$

where $\nu > 0$. Let $\mathcal{Q}_n^{(1)}(D_1)$ and $\mathcal{Q}_n^{(2)}(\nu, D_1)$ denote the classes of channels $Q_n : \mathcal{U}^n \to \mathcal{X}^n$ that satisfy (10) and (11), respectively. In the sequel, we will use the generic notation $\mathcal{Q}_n$ whenever we refer to either $\mathcal{Q}_n^{(1)}(D_1)$ or $\mathcal{Q}_n^{(2)}(\nu, D_1)$.

Definition 2. *An exchangeable channel Q_n is one for which, for every $\mathbf{u} \in \mathcal{U}^n$,*

$$Q_n(\mathbf{x}'|\mathbf{u}) = Q_n(\mathbf{x}|\mathbf{u}), \ \forall \mathbf{u} \in \mathcal{U}^n \ \ \forall \mathbf{x}' \in T_{\mathbf{x}|\mathbf{u}}. \tag{12}$$

Denote by $\mathcal{Q}_n^{ex}$ the class of all exchangeable channels operating on n-tuples. Obviously, every DMC is in $\mathcal{Q}_n^{ex}$.

We next define the average probability of error investigated in this section.

Definition 3. *Let $P_e(Q_n, g_n, W_n)$ designate the average probability of error attained when the watermarking channel is Q_n, the attack channel is W_n and the decoder is g_n, i.e.,*

$$P_e(Q_n, g_n, W_n) = \sum_{\mathbf{u}} P_n(\mathbf{u}) \sum_{\mathbf{x}_1} \cdots \sum_{\mathbf{x}_M} \prod_{j=1}^{M} Q_n(\mathbf{x}_j|\mathbf{u}) \frac{1}{M} \sum_{i=1}^{M} \sum_{\mathbf{y}:\, g_n(\mathbf{u},\mathbf{y}) \ne i} W_n(\mathbf{y}|\mathbf{x}_i). \tag{13}$$

Our goal is to analyze the asymptotic exponent of the minimax error probability, namely, the maximin random-coding error exponent:

$$-\frac{1}{n} \log \left[\min_{\substack{Q_n \in \mathcal{Q}_n \cap \mathcal{Q}_n^{ex} \\ g_n \in \mathcal{G}_n}} \max_{W_n \in \mathcal{W}_n} P_e(Q_n, g_n, W_n) \right], \tag{14}$$

as $n \to \infty$, and to characterize an asymptotically minimizing sequence $\{(Q_n, g_n)\}_{n \ge 1}$ and an asymptotically maximizing sequence $\{W_n\}_{n \ge 1}$.

In Theorem 1 to follow, we present a single-letter expression for the maximin random-coding error exponent. For a given pmf P on $\mathcal{U}$, define:

$$\mathcal{D}_1(D_1, P) = \left\{ Q \in \mathcal{P}(\mathcal{X}|\mathcal{U}) : \sum_{(u,x)\in\mathcal{U}\times\mathcal{X}} P(u)Q(x|u)d_1(u,x) \leq D_1 \right\}. \tag{15}$$

Similarly, for a given pmf V on $\mathcal{U} \times \mathcal{X}$, define:

$$\mathcal{D}_2(D_2, V) = \left\{ W \in \mathcal{P}(\mathcal{Y}|\mathcal{X}\times\mathcal{U}) : \sum_{(u,x)\in\mathcal{U}\times\mathcal{X}} V(u,x) \sum_{y\in\mathcal{Y}} W(y|x,u)d_2(x,y) \leq D_2 \right\}. \tag{16}$$

Define the quantity:

$$\bar{E}(R, P, D_1, D_2, \lambda) = \min\left\{\lambda, \tilde{E}(R, P, D_1, D_2)\right\} \tag{17}$$

with

$$\tilde{E}(R, P, D_1, D_2) = \max_{Q\in\mathcal{D}_1(D_1,P)} \min_{W\in\mathcal{D}_2(D_2,P\times Q)} \left[I_{PQW}(U;Y|X) + |I_{PQW}(X;Y|U) - R|^+\right]$$

where $|a|^+$ denotes $\max\{0, a\}$ and $I_{PQW}(X;Y|U)$ stands for the conditional mutual information between the dummy RV's X and Y given U whose joint pmf is given by $P \times Q \times W$.

Next, define

$$E(R, D_1, D_2, \lambda) = \min_{P\in\mathcal{P}(\mathcal{U})} \left[D(P||P_U) + \bar{E}(R, P, D_1, D_2, \lambda)\right]. \tag{18}$$

Theorem 1. *The maximin random coding error exponent is given by $E(R, D_1, D_2, \lambda)$, i.e.,*

$$\lim_{n\to\infty} -\frac{1}{n} \log\left[\min_{\substack{Q_n\in\mathcal{Q}_n\cap\mathcal{Q}_n^{ex}\\ g_n\in\mathcal{G}_n}} \max_{W_n\in\mathcal{W}_n(\lambda,D_2)} P_e(Q_n, g_n, W_n)\right] = E(R, D_1, D_2, \lambda), \tag{19}$$

where $\mathcal{Q}_n$ is either $\mathcal{Q}_n^{(1)}(D_1)$ or $\mathcal{Q}_n^{(2)}(D_1, \nu)$.

A brief discussion and some comments on the proof of this theorem can be found in Sections 4.3 and 4.4. The proof of Theorem 1 is divided into two parts: The proof of the lower bound on the probability of error can be found in Section 7.1, and the proof of the upper bound appears in Section 7.2. The fact that these two bounds are of the same exponential order establishes the theorem.

4.1 Properties of the Maximin Error Exponent

Following are a few properties of the error exponent function, the proofs of which appear in Appendix A.

For a given pmf V on $\mathcal{U} \times \mathcal{X}$, define

$$A(R, D_2, V) \stackrel{\triangle}{=} \{W \in \mathcal{D}_2(D_2, V) : \; I_{VW}(X; Y|U) \geq R\} . \tag{20}$$

$$\bar{A}(R, D_2, V) \stackrel{\triangle}{=} \{W \in \mathcal{D}_2(D_2, V) : \; I_{VW}(X; Y|U) \leq R\} . \tag{21}$$

Lemma 1. *The function $\tilde{E}(\cdot, \cdot, \cdot, \cdot)$ has the following properties:*
(a) The function $\tilde{E}(R, P, D_1, D_2)$ has an alternative expression:

$$\begin{aligned} &\tilde{E}(R, P, D_1, D_2) \\ =& \max_{Q \in \mathcal{D}_1(D_1, P)} \min_{W \in A(R, D_2, P \times Q)} (I_{PQW}(U; Y|X) + I_{PQW}(X; Y|U) - R) . \end{aligned} \tag{22}$$

(b) The function $\tilde{E}(R, P, D_1, D_2)$ is convex in the pair (R, D_2).
(c) The function $\tilde{E}(R, P, D_1, D_2)$ can be rewritten as:

$$\tilde{E}(R, P, D_1, D_2) = \begin{cases} \tilde{E}(0, P, D_1, D_2) - R & 0 \leq R \leq R_{crit}(P) \\ \tilde{E}_{sp}(R, P, D_1, D_2) & R_{crit}(P) \leq R \leq \underline{C}(P) \\ 0 & R \geq \underline{C}(P) \end{cases} , \tag{23}$$

where,

$$\tilde{E}_{sp}(R, P, D_1, D_2) = \max_{Q \in \mathcal{D}_1(D_1, P)} \min_{W \in \bar{A}(R, D_2, P \times Q)} I_{PQW}(U; Y|X) \tag{24}$$

$$C(P) = \max_{Q \in \mathcal{D}_1(D_1, P)} \min_{\{W \in \mathcal{D}_2(D_2, P \times Q) :\, U \leftrightarrow X \leftrightarrow Y\}} I_{P_U QW}(X; Y|U)$$

$$R_{crit}(P) = I_{PQ^* W^*}(U; Y|X), \tag{25}$$

and where Q^*, W^* are the appropriate maximizer and minimizer of

$$\begin{aligned} \tilde{E}(0, P, D_1, D_2) &= \max_{Q \in \mathcal{D}_1(D_1, P)} \min_{W \in \mathcal{D}_2(D_2, P \times Q)} [I_{PQW}(U; Y|X) + I_{PQW}(X; Y|U)] \\ &= I_{PQ^* W^*}(U; Y|X) + I_{PQ^* W^*}(X; Y|U). \end{aligned} \tag{26}$$

The term $D(P||P_U)$, which appears in $E(R, D_1, D_2, \lambda)$, expresses the averaging over the types classes of covertext sequences. The function $\bar{E}(R, P, D_1, D_2, \lambda)$, for P that is an empirical measure of order n, can be interpreted as the error exponent conditioned on the event that $\mathbf{u}$ belongs to the type class corresponding to P. The function $\bar{E}(R, P, D_1, D_2, \lambda)$ is depicted as the solid line in Figure 1.

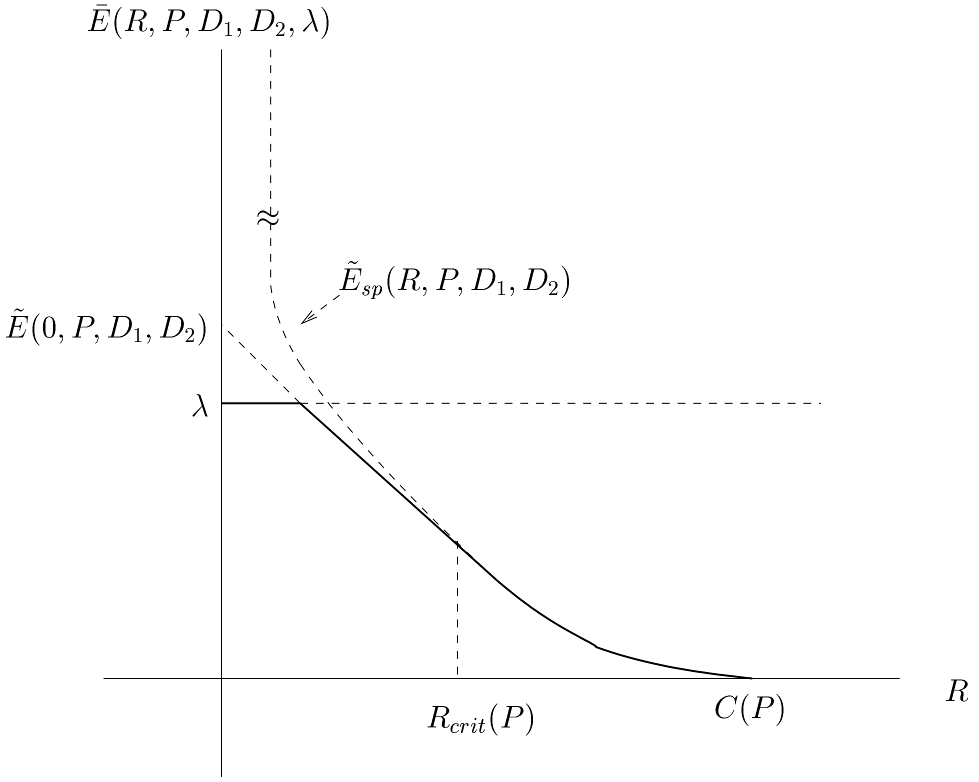

Figure 1. Conditional error exponent $\tilde{E}(R, P, D_1, D_2)$.

It is worth noting that although $\mathcal{W}_n$ is not a class of DMC's, the behavior of $\tilde{E}(R, P, D_1, D_2)$ resembles that of the classical Gallager bound for the DMC and the equivalent expression in [5] and [7], which is given by

$$E_r(R, W) \quad = \quad \max_P \min_{W'} \left[D(W' \| W | P) + | I_{PW'}(X; Y) - R|^+ \right], \tag{27}$$

where $D(W'||W|P) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} P(x) W'(y|x) \log \frac{W'(y|x)}{W(y|x)}$. This bound is also linear with slope -1 up to some rate and strictly convex in higher rates. This resemblance is due to the fact that in watermarking, as in the DMC case, it is the dominant type class $T_{\mathbf{uxy}}$ that determines the behaviour of the error exponent. As in the DMC case, the linear part of $E(R, P, D_1, D_2, \lambda)$ results from the union bound, which is tight in the random coding sense, but if we relax the assumption of i.i.d. codewords given $\mathbf{u}$, better performance can be obtained. To see this, consider the extreme case of a clean channel, i.e. $D_2 = 0$, $d_2(x, y) > 0 \; \forall x \neq y$ and $\lambda = \infty$, with $D_1 = \infty$, for which the random coding error exponent is finite, while clearly, zero probability of error can be attained using deterministic codes.

Note also that there is a value of λ above which the maximin error exponent is no longer dependent of λ, because when λ is large enough, the probability that the distortion exceeds D_2 has a negligible contribution to error events in the exponential scale.

4.2 Saddle Point of the Error Exponent

In Theorem 2 to follow, we identify the worst-case attack channel W_n^*, the optimal watermarking channel Q_n^*, and the optimal decoder g_n^*. This triple forms a saddle point of the asymptotic value of the watermarking game.

For a given $\mathbf{u} \in \mathcal{U}^n$, let $T_{\mathbf{x}^*|\mathbf{u}}$ be the conditional type-class induced by a maximizer $Q \in \mathbb{P}_n(\mathcal{X}, \hat{P}_{\mathbf{u}})$ of the expression:

$$\max_{Q \in \mathbb{P}_n(\mathcal{X}, \hat{P}_{\mathbf{u}}) \cap \mathcal{D}_1(D_1, \hat{P}_{\mathbf{u}})} \min_{W \in \mathcal{D}_2(D_2, \hat{P}_{\mathbf{u}} \times Q)} \left(I_{\hat{P}_{\mathbf{u}}QW}(U;Y|X) + \left| I_{\hat{P}_{\mathbf{u}}QW}(X;Y|U) - R \right|^+ \right). \tag{28}$$

Consider the watermarking channel Q_n^* defined by

$$Q_n^*(\mathbf{x}|\mathbf{u}) = \frac{1}{|T_{\mathbf{x}^*|\mathbf{u}}|} \mathbf{1}\left\{\mathbf{x} \in T_{\mathbf{x}^*|\mathbf{u}}\right\}. \tag{29}$$

Let $\{\mathbf{x}_m(\mathbf{u})\}_{m=1}^M$ designate the codewords corresponding to $\mathbf{u}$ and define the decoder:

$$g_n^*(\mathbf{u}, \mathbf{y}) = \mathrm{argmin}_m \left| T_{\mathbf{x}_m(\mathbf{u})|\mathbf{u},\mathbf{y}} \right|, \tag{30}$$

where ties are broken arbitrarily. Finally, define the attack channel:

$$W_n^* = (1 - e^{-\lambda n}) W_n^{(1)} + e^{-\lambda n} W_n^{(2)}, \tag{31}$$

where

$$W_n^{(1)}(\mathbf{y}|\mathbf{x}) = \frac{c_{n,\mathbf{x}}}{|T_{\mathbf{y}|\mathbf{x}}|} \mathbf{1}\left\{d_2(\mathbf{x},\mathbf{y}) \le nD_2\right\} \;; \quad W_n^{(2)}(\mathbf{y}|\mathbf{x}) = \frac{1}{|\mathcal{Y}|^n} \quad \forall \mathbf{x}, \mathbf{y} \tag{32}$$

$c_{n,\mathbf{x}}$ begin a normalization factor, which, is equal[7] to the reciprocal of the number of conditional types $\left\{T_{\mathbf{y}|\mathbf{x}}\right\}$ such that $d_2(\mathbf{x},\mathbf{y}) \le nD_2$.

Definition 4. *A conditionally fixed composition watermarking channel, Q_n, is one that assigns a uniform distribution to one conditional type $T_{\mathbf{x}|\mathbf{u}}$ for every $\mathbf{u}$.*

Denote by $\mathcal{Q}_n^{cfc}$ the set of conditionally fixed composition channels which operate on n-tuples.

Definition 5. *A strongly exchangeable channel, Q_n, is one that satisfies for all $\mathbf{u}, \mathbf{x}$,*

$$Q_n(\mathbf{x}'|\mathbf{u}') = Q_n(\mathbf{x}|\mathbf{u}), \; \forall (\mathbf{u}', \mathbf{x}') \in T_{\mathbf{ux}}. \tag{33}$$

While for an exchangeable channel the above equality is required to hold for all $\mathbf{x}$ in the same conditional type class w.r.t. $\mathbf{u}$, here, there is also symmetry between different vectors $\mathbf{u}$ of the same type class. Denote by $\mathcal{C}_n^{ex}$ the set of strongly exchangeable channels which operate on n-tuples.

Theorem 2. *(a) The sequence of attack channels $\{W_n^*\}_{n \ge 1}$ satisfies,*

$$\max_{W_n \in \mathcal{W}_n} P_e(Q_n, g_n^*, W_n) \;\; \doteq \;\; P_e(Q_n, g_n^*, W_n^*) \; \forall Q_n \in \mathcal{Q}_n \cap \mathcal{C}_n^{ex} \cap \mathcal{Q}_n^{cfc}. \tag{34}$$

(b) The sequence of pairs of watermarking channels and decoders $\{(Q_n^*, g_n^*)\}_{n \ge 1}$ satisfies,

$$\min_{\substack{Q_n \in \mathcal{Q}_n \cap \mathcal{Q}_{ex} \\ g_n \in \mathcal{G}_n}} P_e(Q_n, g_n, W_n^*) \;\; \doteq \;\; P_e(Q_n^*, g_n^*, W_n^*). \tag{35}$$

[7]Note that since d_2 is additive, $d_2(\mathbf{x},\mathbf{y}) = d_2(\mathbf{x},\mathbf{y}')$ for every $\mathbf{y}' \in T_{\mathbf{y}|\mathbf{x}}$.

Theorem 2 is proved in Appendix 7.3.

Note that since $\mathcal{C}_n^{ex} \cap \mathcal{Q}_n^{cfc} \subset \mathcal{Q}_n^{ex}$, parts ($a$) and ($b$) imply that $(Q_n,^* g_n^*, W_n^*)$ is an asymptotic saddle point of the game, W_n^* is asymptotically worst-case and (Q_n^*, g_n^*) are asymptotically optimal. Combining this with Theorem 1, we have:

$$\begin{aligned} & \min_{\substack{Q_n \in \mathcal{Q}_n \cap \mathcal{Q}_{ex} \\ g_n \in \mathcal{G}_n}} \max_{W_n \in \mathcal{W}_n(\lambda, D_2)} P_e(Q_n, g_n, W_n) \\ \doteq\ & \max_{W_n \in \mathcal{W}_n(\lambda, D_2)} \min_{\substack{Q_n \in \mathcal{Q}_n \cap \mathcal{Q}_n^{ex} \\ g_n \in \mathcal{G}_n}} P_e(Q_n, g_n, W_n) \\ \doteq\ & P_e(Q_n^*, g_n^*, W_n^*) \doteq e^{-nE(R, D_1, D_2, \lambda)}. \end{aligned} \tag{36}$$

4.3 Discussion

A few interesting observations are now in order.

- The asymptotically worst-case attack channel, W_n^*, depends solely on the distortion function $d_2(\cdot, \cdot)$, the distortion level D_2, and the required large deviations exponent λ. It is independent of the rate, R, the watermarking channel, and the covertext statistics P_U. Therefore, there is an asymptotic saddle point and the minimax and the maximin error exponents are equal. Moreover, this also means that the encoder and the decoder may always be tuned to W_n^*. The asymptotically optimal watermarking channel, Q_n^*, on the other hand, depends on $d_1(\cdot, \cdot), D_1, d_2(\cdot, \cdot), D_2$ and on the rate R.

- Note that the error exponent is the same for $\mathcal{Q}_n^{(1)}(D_1)$ and $\mathcal{Q}_n^{(2)}(D_1, \lambda)$. This is because, when it is assumed that Q_n is exchangeable, the watermarking channel that asymptotically maximizes the error exponent assigns probability one to a single conditional type.

- It is easily verified (see Section 7.5 for a proof) that the smallest rate for which $E(R, D_1, D_2, \lambda)$ vanishes, is given by:

$$C = \max_{Q \in \mathcal{D}_1(D_1, P_U)} \min_{\{W \in \mathcal{D}_2(D_2, P_U \times Q) : \, U \leftrightarrow X \leftrightarrow Y\}} I_{P_U Q W}(X; Y | U), \tag{37}$$

  and so, C serves as a lower bound to the capacity of the watermarking system under our assumptions. Note also, that C is equal to the capacity of memoryless attack channel under expected distortion constraint as appeared in earlier papers [15], [16], [17]. In Section 6, it will be proved that C is indeed the highest achievable rate, that is, the capacity, even if we allow for a wider class of encoders, namely, general encoders as defined in (7). Note that, as one may expect, for large enough D_2, $C = 0$.

- It is well known that when the channel is known to belong to a certain parametric family of channels (such as the memoryless channels and finite state channels), there exist universal decoders which guarantee reliable communication. In [12], the question of how many bits a decoder needs to know about the statistics of a channel in order to guarantee reliable communication in the random coding sense is investigated. Roughly speaking, it is proved there that in order to guarantee reliable transmission for *all* the channels which satisfy some regularity conditions, it is necessary and sufficient to provide the decoder with $2^{H(Y^n|X^n)+nR}$ bits of information regarding the channel. Therefore, the existence of a universal decoder g_n^* for the class of channels $\mathcal{W}_n$, which is not parametric and seemingly very rich, is somewhat surprising. This phenomenon can be explained by the fact that the class $\mathcal{W}_n$ is not as wide as one may think. In fact, it can be shown that when using g_n^* and applying a watermarking channel $Q_n \in \mathcal{Q}_n^{cfc} \cap \mathcal{C}_n^{ex}$ the attacker can attain the same performance by choosing the attack channel from the subclass of channels $\mathcal{W}_n \cap \mathcal{C}_n^{ex}$ (see Lemma 5, Section 7.2). The class $\mathcal{W}_n \cap \mathcal{C}_n^{ex}$, by definition, can be represented by a polynomial (with n) number of parameters[8] for every $\mathbf{x}$.

4.4 Comments on the Proofs of Theorems 1 and 2

In Sections 7.1 and 7.2, a lower bound and an upper bound on the minimax probability of error are derived using the method of types, and these bounds are shown to be of the same exponential order. The minimax probability of error is upper bounded by letting the watermarking channel be a conditionally fixed composition channel, and letting the decoder be g_n^* (defined in (30)) independently of the channel. This bounding technique has been used extensively in derivations of coding theorems for unknown channels, see e.g., [5]. A very crucial point in the proof of the upper bound is the fact that for all $W_n \in \mathcal{W}_n(D_2, \lambda)$, and for every $\mathbf{x} \in \mathcal{X}^n, \mathbf{y} \in \mathcal{Y}^n$

$$\sum_{\mathbf{y}' \in T_{\mathbf{y}|\mathbf{x}}} W_n(\mathbf{y}'|\mathbf{x}) \leq e^{-n\lambda \cdot \mathbf{1}\{d_2(\mathbf{x},\mathbf{y}) \leq nD_2)\}}. \tag{38}$$

This bound is due to the fact that if $d_2(\mathbf{x},\mathbf{y}) \leq nD_2$, the probability of the entire conditional type $T_{\mathbf{y}|\mathbf{x}}$ is trivially upper bounded by 1, and if $d_2(\mathbf{x},\mathbf{y}) > nD_2$, the probability of the conditional type $T_{\mathbf{y}|\mathbf{x}}$ cannot exceed $e^{-\lambda n}$ as otherwise the distortion constraint (9) is violated. The r.h.s. of eq. (38) can be normalized by a polynomial factor. In other words, the r.h.s. of (38) is also a conditional pmf (a channel) up to a polynomial factor that does not affect asymptotic analysis in the exponential scale. Therefore, every affine functional of an exchangeable channel W_n, like $P_e(Q_n, g_n, W_n)$, can be upper bounded by the same functional with W_n being replaced by the r.h.s. of (38). In other words, it is essentially always a worst case channel. The lower bound, on the other hand, is attained by assuming that a "genie-aided" strategy is used, which enables maximum likelihood decoding. Further, it is shown that these bounds are exponentially equivalent. The following chain of inequalities

[8] If $W_n \in \mathcal{W}_n \cap \mathcal{C}_n^{ex}$, then $W_n(\mathbf{y}|\mathbf{x}) = \frac{W_n(T_{\mathbf{y}|\mathbf{x}}|\mathbf{x})}{|T_{\mathbf{y}|\mathbf{x}}|}$, thus W_n is defined by $\{W_n(T_{\mathbf{y}|\mathbf{x}}|\mathbf{x})\}_{\mathbf{x}\in\mathcal{X}^n}$

summarizes the above statements:

$$
\begin{aligned}
& e^{-nE(R,D_1,D_2,\lambda)} \stackrel{\cdot}{\leq} \min_{Q_n\in\mathcal{Q}_n\cap\mathcal{Q}_n^{ex}} \max_{W_n\in\mathcal{W}_n(\lambda,D_2)} P_e(Q_n, g^{ML}_{W_n}, W_n) \\
\leq & \min_{\substack{Q_n\in\mathcal{Q}_n\cap\mathcal{Q}_n^{ex}\\ g_n\in\mathcal{G}_n}} \max_{W_n\in\mathcal{W}_n(\lambda,D_2)} P_e(Q_n, g_n, W_n) \\
\leq & \min_{Q_n\in\mathcal{Q}_n\cap\mathcal{Q}_n^{cfc}\cap\mathcal{C}_n^{ex}} \max_{W_n\in\mathcal{W}_n(\lambda,D_2)} P_e(Q_n, g_n^*, W_n) \stackrel{\cdot}{\leq} e^{-nE(R,D_1,D_2,\lambda)},
\end{aligned}
\tag{39}
$$

where $g^{ML}_{W_n}$ is the maximum likelihood decoder corresponding to the channel W_n. The right-most and the left most sides imply that these are, in fact, equalities in exponential sense.

As a conclusion from this observation, it may be inferred that Theorem 1 remains valid for yet another setting of the watermarking game that may be of interest, where Q_n is fixed, W_n is chosen based on the knowledge of Q_n, and the decoder knows both Q_n and $\mathbf{u}$ and has either partial or complete knowledge of W_n. At first glance, it seems surprising that the error exponent is equal for the two very different settings, but this is explained by the fact that the decoder g_n^* which performs well on W_n^* does well also on every exchangeable attack channel, which makes it universal in this sense. In fact, there is a remarkable resemblance between g_n^* and the MMI decoder [5], although the working assumption here is of a general unknown channel (that is only known to be keeping the distortion constraint) while in [5] the channel is known in advance to be a DMC. Since the code is conditionally fixed composition,

$$
\text{argmax}_m |T_{\mathbf{x}_m(\mathbf{u})|\mathbf{u},\mathbf{y}}| = \text{argmax}_m |T_{\mathbf{x}_m(\mathbf{u})\mathbf{y}|\mathbf{u}}| \tag{40}
$$

and also,

$$
\frac{|T_{\mathbf{xy}|\mathbf{u}}|}{|T_{\mathbf{x}|\mathbf{u}}|} \doteq e^{-n\hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}}} \tag{41}
$$

where $\hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}}$ is defined in (5), so, loosely speaking, the smaller $|T_{\mathbf{x},\mathbf{y}|\mathbf{u}}|$ is, the bigger is $\hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}}$.

The bounding technique is similar to that of D'yachkov [7], (see also [6]) who investigated the error exponent of fixed composition random codes for DMC's. There are, however, some major differences between [7] and our results:

1. In [7] the channel is a specified DMC.
2. We assume that the watermarking channel Q_n belongs to the class of exchangeable channels, which is a much wider class than the class of fixed-composition channels discussed in [7].
3. The distortion constraints that are imposed in the watermarking game.
4. The presence of side information in the watermarking game.

5 A Sphere-Packing Bound

The results thus far refer to a game where the encoder is restricted to the use of a random coding scheme, in which for every $\mathbf{u} \in \mathcal{U}^n$, the codewords in the codebook corresponding to $\mathbf{u}$ are drawn *independently*, and all according to the same conditional pmf $Q_n(\cdot|\mathbf{u})$. Henceforth, we relax this restriction, and investigate a wider class of encoders (as in [1]) which allow for *any* randomized coding scheme with a secret key Θ shared by the encoder and decoder. This means full generality of the joint pmf of every codebook given $\mathbf{u}$.

Allowing for a wider class of encoders should, in general, increase the error exponent. Surprisingly, it is shown that for a certain range of high rates (and possibly also in a certain range of low rates, depending on the value of λ), the random coding error exponent $E(R, D_1, D_2, \lambda)$ (defined in (18)) remains the best achievable error exponent.

Let $F_{n,M} = F_{n,M}(f_n, \Theta)$ stand for the (randomized) codebook created by the encoder f_n (see eq. (7)) and the secret key Θ, i.e., the codebook whose m-th codeword corresponding to $\mathbf{u}$ is given by:

$$\mathbf{x}_m = f_n(\theta, \mathbf{u}, m), \tag{42}$$

where θ is a realization of Θ.

Let $\mathcal{F}_{n,M} = \mathcal{F}_{n,M}(D_1)$ be the set of randomized codebooks, i.e., the set of pmf's defined on Γ (the alphabet of Θ) and functions f_n, such that

$$\Pr\left\{d_1(\mathbf{U}, \mathbf{X}) > nD_1 | \mathbf{U} = \mathbf{u}\right\} = 0, \quad \forall \mathbf{u} \in \mathcal{U}^n, \tag{43}$$

namely, $\nu = \infty$. Note that this constraint is equivalent to

$$Pr\left\{d_1(\mathbf{U}, \mathbf{X}) > nD_1\right\} = 0, \tag{44}$$

and is more restrictive than the constraints (10) and (11) posed in Section 4. Denote by $\mathcal{F}^d_{n,M}$ the set of all deterministic codebooks $F^d_{n,M} : \mathcal{U}^n \times \mathcal{M}_n \to \mathcal{X}^n$.

We shall denote by $P_e(\mathcal{F}_{n,M}, g_n, W_n)$ the probability of error averaged over the covertext, the watermark, the secret key and the action of the attack channel W_n, when the encoder f_n, and the decoder g_n are employed, i.e.,

$$P_e(\mathcal{F}_{n,M}, g_n, W_n) = \sum_{m=1}^{M} \frac{1}{M} E_{\mathbf{U},\Theta,\mathbf{Y}}\left(\mathbf{1}\left\{g_n(\Theta, \mathbf{U}, \mathbf{Y}) \neq m\right\} | \text{ the hidden message is } m\right). \tag{45}$$

Define,

$$E_{sp}(R, D_1, D_2, \lambda) = \min_{P \in \mathcal{P}(\mathcal{U})} \left[D(P||P_U) + \bar{E}_{sp}(R, P, D_1, D_2, \lambda)\right], \tag{46}$$

where

$$\bar{E}_{sp}(R, P, D_1, D_2, \lambda) = \min\left\{\lambda, \tilde{E}_{sp}(R, P, D_1, D_2)\right\}. \tag{47}$$

and $\tilde{E}_{sp}(R, P, D_1, D_2)$ is defined in (24). Recalling (23), one can see that $\tilde{E}(R, P, D_1, D_2) = \tilde{E}_{sp}(R, P, D_1, D_2)$ for all $R > R_{crit}(P)$, and thus,

$$\bar{E}(R, P, D_1, D_2, \lambda) = \bar{E}_{sp}(R, P, D_1, D_2, \lambda), \quad \forall R > R_{crit}(P) \text{ or } R \in [0, R_{c_1}(P)], \tag{48}$$

where $R_{c_1}(P) = \max\left\{0, \tilde{E}(0, P, D_1, D_2) - \lambda\right\}$.

Let $R_\infty(D_1, D_2)$ be the infimum of those rates for which[9] $\min_{P\in\mathcal{P}(\mathcal{U})}\left[D(P\|P_U) + \tilde{E}_{sp}(R, P, D_1, D_2)\right] = \infty$.

Theorem 3. *(Sphere-Packing Bound) For every (D_1, D_2), and for all $R \neq R_\infty(D_1, D_2)$,*

$$\min_{\substack{F_{n,M}\in\mathcal{F}_{n,M}\\ g_n\in\mathcal{G}_n}} \max_{W_n\in\mathcal{W}_n} P_e(F_{n,M}, g_n, W_n) \stackrel{\cdot}{\geq} e^{-nE_{sp}(R,D_1,D_2,\lambda)}. \tag{49}$$

Note that the minimization over $\{F_{n,M} \in \mathcal{F}_{n,M}\}$ is equivalent to minimizing over the set of all possible distributions of the secret key, Θ, and over all encoders $f_n : \Gamma \times \mathcal{U}^n \times \mathcal{Y}^n \to \mathcal{M}_n$.

The proof of Theorem 3 can be found in Section 7.4. The proof is based on the following observations:

- The minimax average probability of error attained by a randomized code $F_{n,M} \in \mathcal{F}_{n,M}$ can be lower bounded by the minimum average probability of error when the attack channel is W_n^* and the encoder uses the best deterministic code $F^d_{n,M} \in \mathcal{F}^d_{n,M}$ tuned to W_n^*.
- The maximum probability of error and the average probability of error are asymptotically equal in the exponential scale, for the channel W_n^* and any deterministic encoder.
- When the attack channel chosen by the attacker is W_n^*, one can assume that the encoder uses deterministic constant composition codes. This assumption does not reduce the information rate on the exponential scale.
- One can obtain a converse lemma w.r.t. the maximum probability of error for a specified class of attack channels (referred to as *invariantly uniform in type* (IUT) channels[10]) which are implementable by a genie-aided attacker that observes $\mathbf{u}$ as well as $\mathbf{x}$.
- It is shown that the maximum probability of error attained when the channel is W_n^* cannot be very small compared to the maximum probability of error attained for an IUT channel.

6 The Capacity Game

Thus far, we have investigated the maximin error exponent. We shall next present a coding theorem on the maximin capacity. The smallest rate at which the error exponent vanishes is a lower bound on the capacity, but this fact does not preclude the possibility that reliable information embedding can be made possible in higher rates with probability of error that vanishes in a sub-exponential manner. Therefore, we shall present a converse theorem that states that the smallest rate at which the error exponent vanishes, is in fact the capacity, in

[9]Equality to ∞ occurs if for every $P \in \mathcal{P}(\mathcal{U})$ there exists $Q \in \mathcal{D}_1(D_1, P)$, such that the set $\bar{A}(R, D_2, P\times Q)$ is empty.

[10]See Definition 7 in the beginning of Section 7.4.

the sense that there exist attack channels in $\mathcal{W}_n$, for which, any randomized encoder working in a higher rate will fail to guarantee vanishing probability of error. The converse theorem established here is very general as it accounts for the same class of encoders as in Section 5, i.e., encoders that allow for *any* randomized coding scheme with a secret key Θ shared by the encoder and decoder. Therefore, this converse holds trivially also for the smaller class of randomized encoders of independent codewords.

Definition 6. *An achievable rate R for a watermarking system with parameters D_1, D_2, λ is one for which there exists a sequence of rate-R watermarking codes $\{(\mathcal{M}_n, f_n, g_n)\}_{n\geq 1}$, and a secret key Θ, such that*[11]

$$Pr(d_1(\mathbf{U},\mathbf{X}) > nD_1) = 0, \tag{50}$$

and $\lim_{n\to\infty} \sup_{W_n \in \mathcal{W}_n(\lambda,D_2)} P_e\left(F_{n,M}, g_n, W_n\right) = 0.$

The *coding capacity* of this watermarking game is defined as the supremum of all achievable rates.

Theorem 4. *The coding capacity of the watermarking game is given by*

$$C = \max_{Q\in\mathcal{D}_1(D_1,P_U)} \min_{\{W\in\mathcal{D}_2(D_2,P_U\times Q):\, U\leftrightarrow X\leftrightarrow Y\}} I_{P_U Q W}(X;Y|U). \tag{51}$$

Theorem 4 extends the private game results of [1] as it provides a single-letter expression for the capacity of the private watermarking game where the covertext is assumed to be a general DMS, while in [1] it is assumed to be Gaussian. Moreover, the distortion functions investigated in [1] are $d_1(\mathbf{u},\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n (x_i - u_i)^2$ and $d_2(\mathbf{x},\mathbf{y}) = \frac{1}{n}\sum_{i=1}^n (y_i - x_i)^2$, while we address general additive distortion functions. The almost sure distortion constraint imposed in [1] can be viewed as a special case of (9) with $\lambda = \infty$.

As usual, we divide the proof into two parts, corresponding to the direct part and the converse part, the proofs of which appear in Appendices 7.5 and 7.6 respectively.

[11] This distortion constraint is equivalent to (11) with $\nu = \infty$. It is easily verified that our results in this section continue to hold also for the class of encoders satisfying (11) for every $\nu > 0$.

7 Proofs of Theorems

7.1 Proof of the Converse Part of Theorem 1

In this section, we prove that:

$$\min_{\substack{Q_n \in \mathcal{Q}_n \cap \mathcal{Q}_n^{ex} \\ g_n \in \mathcal{G}_n}} \max_{W_n \in \mathcal{W}_n} P_e(Q_n, g_n, W_n) \stackrel{\cdot}{\geq} e^{-nE(R,\lambda,D_1,D_2)}. \tag{52}$$

Recall that $g_{W_n}^{ML}$ stands for the maximum likelihood decoder which corresponds to an attack channel W_n, i.e., if $\mathbf{x}_m(\mathbf{u})$ denotes the m-th codeword in the codebook corresponding to $\mathbf{u}$, the maximum likelihood decoder decides according to:

$$\hat{m} = g_{W_n}^{ML}(\mathbf{u}, \mathbf{y}) = \text{argmax}_m W_n(\mathbf{y}|\mathbf{x}_m(\mathbf{u})), \tag{53}$$

where ties are broken arbitrarily.

For the sake of brevity, denote $\bar{\mathcal{Q}}_n \stackrel{\triangle}{=} \mathcal{Q}_n \cap \mathcal{Q}_n^{ex}$. The minimax average probability of error can be lower bounded by assuming that the decoder is informed of the attack channel, i.e.,

$$\begin{aligned}
\min_{\substack{Q_n \in \bar{\mathcal{Q}}_n \\ g_n \in \mathcal{G}_n}} \max_{W_n \in \mathcal{W}_n} P_e(Q_n, g_n, W_n) &\geq \min_{Q_n \in \bar{\mathcal{Q}}_n} \max_{W \in \mathcal{W}_n} \min_{g_n \in \mathcal{G}_n} P_e(Q_n, g_n, W_n) \\
&= \min_{Q_n \in \bar{\mathcal{Q}}_n} \max_{W_n \in \mathcal{W}_n} P_e(Q_n, g_{W_n}^{ML}, W_n) \\
&\geq \min_{Q_n \in \bar{\mathcal{Q}}_n} P_e(Q_n, g_{W_n^*}^{ML}, W_n^*),
\end{aligned} \tag{54}$$

where W_n^* (defined in (31)).

A simple operative argument shows that $P_e(Q_n, g_{W_n}^{ML}, W_n)$ is concave in W_n, i.e., for all $0 \leq \alpha \leq 1$, if $W_n^\alpha \triangleq \alpha W_n^{(1)} + (1-\alpha) W_n^{(2)}$, where $W_n^{(1)}$ and $W_n^{(2)}$ are two given channels, then:

$$P_e(Q_n, g_{W_n^\alpha}^{ML}, W_n^\alpha) \geq \alpha P_e(Q_n, g_{W_n^{(1)}}^{ML}, W_n^{(1)}) + (1-\alpha) P_e(Q_n, g_{W_n^{(2)}}^{ML}, W_n^{(2)}). \tag{55}$$

This is true because the left hand side (l.h.s.) corresponds to a situation where the decoder is uninformed of the actual mode of attack channel, $W_n^{(1)}$ or $W_n^{(2)}$, while the r.h.s. corresponds to an informed decoder.

Thus, recalling the definition of W_n^* in (31), as a convex combination of $W_n^{(1)}$ and $W_n^{(2)}$, yields,

$$\begin{aligned}
&\min_{Q_n \in \bar{\mathcal{Q}}_n} P_e(Q_n, g_{W_n^*}^{ML}, W_n^*) \\
\geq\ & (1 - e^{-\lambda n}) \min_{Q_n \in \bar{\mathcal{Q}}_n} P_e(Q_n, g_{W_n^{(1)}}^{ML}, W_n^{(1)}) + e^{-\lambda n} \min_{Q_n \in \bar{\mathcal{Q}}_n} P_e(Q_n, g_{W_n^{(2)}}^{ML}, W_n^{(2)}) \\
\geq\ & (1 - e^{-\lambda n}) \min_{Q_n \in \bar{\mathcal{Q}}_n} P_e(Q_n, g_{W_n^{(1)}}^{ML}, W_n^{(1)}) + e^{-\lambda n} \left(1 - \frac{1}{M}\right),
\end{aligned} \tag{56}$$

where the last step is due to the fact that when $W_n^{(2)}$ is active, $\mathbf{Y}$ is independent of $\mathbf{X}$.

Since $P_e(Q_n, g^{ML}_{W_n^{(1)}}, W_n^{(1)}) = \sum_{T_{\mathbf{u}}} P_U^n(T_{\mathbf{u}}) P_e(Q_n, g^{ML}_{W_n^{(1)}}, W_n^{(1)} | \mathbf{U} \in T_{\mathbf{u}})$, where $P_U^n(T_{\mathbf{u}}) = \sum_{\mathbf{u}' \in T_{\mathbf{u}}} P_U^n(\mathbf{u}')$, and the number of types is polynomial,

$$
\begin{aligned}
\min_{Q_n \in \bar{\mathcal{Q}}_n} P_e(Q_n, g^{ML}_{W_n^{(1)}}, W_n^{(1)}) &\doteq \max_{T_{\mathbf{u}}} P_U^n(T_{\mathbf{u}}) \min_{Q_n \in \bar{\mathcal{Q}}_n} P_e(Q_n, g^{ML}_{W_n^{(1)}}, W_n^{(1)} | \mathbf{U} \in T_{\mathbf{u}}) \\
&= \max_{T_{\mathbf{u}}} P_U^n(T_{\mathbf{u}}) \min_{Q_n \in \bar{\mathcal{Q}}_n} P_e(Q_n, g^{ML}_{W_n^{(1)}}, W_n^{(1)} | \mathbf{u}),
\end{aligned}
\tag{57}
$$

where $\mathbf{u}$ is an arbitrary member of $T_{\mathbf{u}}$ and $P_e(Q_n, g^{ML}_{W_n^{(1)}}, W_n^{(1)} | \mathbf{u})$ is the probability of error given $\mathbf{u}$, $W_n^{(1)}$ is the attack channel and Q_n and $g^{ML}_{W_n^{(1)}}$ are the watermarking channel and the decoder, respectively. Inequality (57) holds due to the fact that the minimization is done for each $Q_n(\cdot|\mathbf{u})$ separately[12], and there is a perfect symmetry in all $\mathbf{u}$ of the same type due to the symmetric nature of W_n^*.

Also, since it is assumed that the M words in a codebook are equiprobable,

$$
P_e(Q_n, g^{ML}_{W_n^{(1)}}, W_n^{(1)} | \mathbf{u}) = P_e(Q_n, g^{ML}_{W_n^{(1)}}, W_n^{(1)} | \mathbf{u}, m \text{ transmitted}),
\tag{58}
$$

so we get:

$$
\begin{aligned}
& P_e(Q_n, g^{ML}_{W_n^{(1)}}, W_n^{(1)} | \mathbf{u}) \\
= & \sum_{\mathbf{x}_m} Q_n(\mathbf{x}_m|\mathbf{u}) \sum_{\mathbf{y}} W_n^{(1)}(\mathbf{y}|\mathbf{x}_m) \left[1 - \left(1 - \sum_{\left\{\mathbf{x}' :\, W_n^{(1)}(\mathbf{y}|\mathbf{x}') \geq W_n^{(1)}(\mathbf{y}|\mathbf{x}_m)\right\}} Q_n(\mathbf{x}'|\mathbf{u}) \right)^{M-1} \right] \\
\geq & \sum_{\mathbf{x}_m} Q_n(\mathbf{x}_m|\mathbf{u}) \sum_{\mathbf{y}} W_n^{(1)}(\mathbf{y}|\mathbf{x}_m) \left[1 - \left(1 - \sum_{\mathbf{x}' \in T_{\mathbf{x}_m|\mathbf{y},\mathbf{u}}} Q_n(\mathbf{x}'|\mathbf{u}) \right)^{M-1} \right]
\end{aligned}
\tag{59}
$$

where the first equality follows from the assumption that all M codewords are chosen independently, and thus the $M-1$ words (other than $\mathbf{x}_m$) are independent of the received signal $\mathbf{Y}$. Inequality (59) follows from the fact that by definition of $W_n^{(1)}$,

$$
T_{\mathbf{x}_m|\mathbf{y},\mathbf{u}} \subseteq \left\{ \mathbf{x}' :\ W_n^{(1)}(\mathbf{y}|\mathbf{x}') \geq W_n^{(1)}(\mathbf{y}|\mathbf{x}_m) \right\}.
\tag{60}
$$

Henceforth, we shall omit the index m from the notation, and for the sake of brevity, we shall denote

$$
Q_n(T_{\mathbf{x}|\mathbf{u}}) \triangleq Q_n(T_{\mathbf{x}|\mathbf{u}}|\mathbf{u}) = \sum_{\mathbf{x}' \in T_{\mathbf{x}|\mathbf{u}}} Q_n(\mathbf{x}'|\mathbf{u}),
\tag{61}
$$

where we omit the conditioning on $\mathbf{u}$ for brevity.

When the watermarking channel Q_n is assumed exchangeable, we have $Q_n(\mathbf{x}|\mathbf{u}) = \frac{Q_n(T_{\mathbf{x}|\mathbf{u}})}{|T_{\mathbf{x}|\mathbf{u}}|}$, thus,

$$
\sum_{\mathbf{x}' \in T_{\mathbf{x}|\mathbf{y},\mathbf{u}}} Q_n(\mathbf{x}'|\mathbf{u}) = |T_{\mathbf{x}|\mathbf{u},\mathbf{y}}| \frac{Q_n(T_{\mathbf{x}|\mathbf{u}})}{|T_{\mathbf{x}|\mathbf{u}}|}.
\tag{62}
$$

[12]By definition of both $\mathcal{Q}_n^{(1)}(D_1)$ and $\mathcal{Q}_n^{(2)}(D_1, \nu)$, the distortion constraints are imposed for every $\mathbf{u} \in \mathcal{U}^n$.

So, denoting $a(T_{\mathbf{x}|\mathbf{u},\mathbf{y}}) \triangleq |T_{\mathbf{x}|\mathbf{u},\mathbf{y}}|\frac{Q_n(T_{\mathbf{x}|\mathbf{u}})}{|T_{\mathbf{x}|\mathbf{u}}|}$, we obtain from (58)-(62):

$$\begin{aligned}
&\min_{Q_n\in\bar{\mathcal{Q}}_n} P_e(Q_n, g^{ML}_{W_n^{(1)}}, W_n^{(1)}|\mathbf{u}) \\
\geq\ & \min_{Q_n\in\bar{\mathcal{Q}}_n} \sum_{\mathbf{x}} \frac{Q_n(T_{\mathbf{x}|\mathbf{u}})}{|T_{\mathbf{x}|\mathbf{u}}|} \sum_{\mathbf{y}} W_n^{(1)}(\mathbf{y}|\mathbf{x}) \left[1-\left(1-|T_{\mathbf{x}|\mathbf{u},\mathbf{y}}|\frac{Q_n(T_{\mathbf{x}|\mathbf{u}})}{|T_{\mathbf{x}|\mathbf{u}}|}\right)^{M-1}\right] \\
=\ & \min_{Q_n\in\bar{\mathcal{Q}}_n} \sum_{\mathbf{x}} \frac{Q_n(T_{\mathbf{x}|\mathbf{u}})}{|T_{\mathbf{x}|\mathbf{u}}|} \sum_{\mathbf{y}} W_n^{(1)}(\mathbf{y}|\mathbf{x}) \left[1-\left(1-a(T_{\mathbf{x}|\mathbf{u},\mathbf{y}})\right)^{M-1}\right]. \qquad (63)
\end{aligned}$$

Using the Newton binomial formula, we get:

$$1-(1-a)^{M-1} = \sum_{i=1}^{M-1} \binom{M-1}{i} a^i(1-a)^{M-1-i} \geq (M-1)a(1-a)^{M-2}. \qquad (64)$$

Recalling definition (5) and the fact that $\frac{|T_{\mathbf{x}_1|\mathbf{x}_2,\mathbf{x}_3}|}{|T_{\mathbf{x}_1|\mathbf{x}_3}|} \doteq e^{-n\hat{I}_{\mathbf{x}_1;\mathbf{x}_2|\mathbf{x}_3}}$ and using (64) with $a = a(T_{\mathbf{x}|\mathbf{u},\mathbf{y}})$ yields the following bound for every exchangeable Q_n:

$$\begin{aligned}
&P_e(Q_n, g^{ML}_{W_n^{(1)}}, W_n^{(1)}|\mathbf{u}) \\
\geq\ & \sum_{\mathbf{x}} \frac{Q_n(T_{\mathbf{x}|\mathbf{u}})}{|T_{\mathbf{x}|\mathbf{u}}|} \sum_{\mathbf{y}} W_n^{(1)}(\mathbf{y}|\mathbf{x})(M-1)a(T_{\mathbf{x}|\mathbf{u},\mathbf{y}})\left(1-a(T_{\mathbf{x}|\mathbf{u},\mathbf{y}})\right)^{M-2} \\
=\ & \sum_{\mathbf{x}} \frac{Q_n(T_{\mathbf{x}|\mathbf{u}})}{|T_{\mathbf{x}|\mathbf{u}}|} \sum_{\mathbf{y}} \frac{c_{n,\mathbf{x}}}{|T_{\mathbf{y}|\mathbf{x}}|}\mathbf{1}\left\{d_2(\mathbf{x},\mathbf{y})\leq nD_2\right\}(M-1)a(T_{\mathbf{x}|\mathbf{u},\mathbf{y}})\left(1-a(T_{\mathbf{x}|\mathbf{u},\mathbf{y}})\right)^{M-2} \qquad (65) \\
\geq\ & \sum_{\mathbf{x}} \frac{Q_n(T_{\mathbf{x}|\mathbf{u}})}{|T_{\mathbf{x}|\mathbf{u}}|} \max_{T_{\mathbf{y}|\mathbf{x},\mathbf{u}}} \frac{c_{n,\mathbf{x}}}{|T_{\mathbf{y}|\mathbf{x}}|}\mathbf{1}\left\{d_2(\mathbf{x},\mathbf{y})\leq nD_2\right\}|T_{\mathbf{y}|\mathbf{x},\mathbf{u}}|(M-1)a(T_{\mathbf{x}|\mathbf{u},\mathbf{y}})\left(1-a(T_{\mathbf{x}|\mathbf{u},\mathbf{y}})\right)^{M-2} \\
\stackrel{\cdot}{\geq}\ & \sum_{\mathbf{x}} \frac{Q_n(T_{\mathbf{x}|\mathbf{u}})}{|T_{\mathbf{x}|\mathbf{u}}|} \max_{T_{\mathbf{y}|\mathbf{x},\mathbf{u}}:d_2(\mathbf{x},\mathbf{y})\leq nD_2} e^{-n(\hat{I}_{\mathbf{u};\mathbf{y}|\mathbf{x}}+\hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}}-R)}Q_n(T_{\mathbf{x}|\mathbf{u}})\left(1-Q_n(T_{\mathbf{x}|\mathbf{u}})e^{-n\hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}}}\right)^M, \qquad (66)
\end{aligned}$$

where (65) follows from the definition of $W_n^{(1)}$ (32), and (66) holds since

$$c_{n,\mathbf{x}} = \left[\sum_{\mathbf{y}:\, d_2(\mathbf{x},\mathbf{y})\leq nD_2} \frac{1}{|T_{\mathbf{y}|\mathbf{x}}|}\right]^{-1} = \left[\sum_{T_{\mathbf{y}|\mathbf{x}}:d_2(\mathbf{x},\mathbf{y})\leq nD_2}\ \sum_{\mathbf{y}\in T_{\mathbf{y}|\mathbf{x}}} \frac{1}{|T_{\mathbf{y}|\mathbf{x}}|}\right]^{-1} \geq \frac{1}{(n+1)^{|\mathcal{X}||\mathcal{Y}|}}. \qquad (67)$$

Now,

$$
\begin{aligned}
&\sum_{\mathbf{x}} \frac{Q_n(T_{\mathbf{x}|\mathbf{u}})}{|T_{\mathbf{x}|\mathbf{u}}|} \max_{T_{\mathbf{y}|\mathbf{x},\mathbf{u}}:\, d_2(\mathbf{x},\mathbf{y})\le nD_2} e^{-n(\hat{I}_{\mathbf{u};\mathbf{y}|\mathbf{x}}+\hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}}-R)} Q_n(T_{\mathbf{x}|\mathbf{u}}) \left[1 - Q_n(T_{\mathbf{x}|\mathbf{u}}) e^{-n\hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}}}\right]^M \\
\dot{\ge}\ &\sum_{\mathbf{x}} \frac{Q_n^2(T_{\mathbf{x}|\mathbf{u}})}{|T_{\mathbf{x}|\mathbf{u}}|} \max_{\substack{T_{\mathbf{y}|\mathbf{x},\mathbf{u}}:\ d_2(\mathbf{x},\mathbf{y})\le nD_2,\\ Q_n(T_{\mathbf{x}|\mathbf{u}}) e^{-n\hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}}} \le \frac{1}{M}}} e^{-n(\hat{I}_{\mathbf{u};\mathbf{y}|\mathbf{x}}+\hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}}-R)} \left(1-\frac{1}{M}\right)^M \\
\dot{\ge}\ &\sum_{\mathbf{x}} \frac{Q_n^2(T_{\mathbf{x}|\mathbf{u}})}{|T_{\mathbf{x}|\mathbf{u}}|} \max_{\substack{T_{\mathbf{y}|\mathbf{x},\mathbf{u}}:\ d_2(\mathbf{x},\mathbf{y})\le nD_2,\\ e^{-n\hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}}} \le \frac{1}{M}}} e^{-n(\hat{I}_{\mathbf{u};\mathbf{y}|\mathbf{x}}+\hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}}-R)} \qquad (68) \\
=\ &\sum_{\mathbf{x}} \frac{Q_n^2(T_{\mathbf{x}|\mathbf{u}})}{|T_{\mathbf{x}|\mathbf{u}}|} \max_{T_{\mathbf{y}|\mathbf{x},\mathbf{u}}:\, d_2(\mathbf{x},\mathbf{y})\le nD_2, \hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}}\ge R} e^{-n(\hat{I}_{\mathbf{u};\mathbf{y}|\mathbf{x}}+\hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}}-R)} \qquad (69)
\end{aligned}
$$

where (68) holds since $\left(1-\frac{1}{M}\right)^M \to e^{-1}$ as $M\to\infty$, and $Q_n(T_{\mathbf{x}|\mathbf{u}}) \le 1$.

For a given pmf P of (U,X,Y), define:

$$J(P) \;=\; I_P(U;Y|X) + I_P(X;Y|U). \qquad (70)$$

For a given pmf V on $\mathcal{U}\times\mathcal{X}$, define the set

$$A_n(R,D_2,V) \stackrel{\triangle}{=} A(R,D_2,V) \cap \mathbb{P}_n(\mathcal{Y},V), \qquad (71)$$

where $A(R,D_2,V)$ is defined in (20) and $\mathbb{P}_n(\mathcal{Y},V)$ is defined in (3).

Note that, due to the additivity of $d_2(\mathbf{x},\mathbf{y}) = \sum_{i=1}^n d_2(x_i,y_i)$, it can be regarded as the expectation of $d_2(X,Y)$ where X,Y are RV's whose joint pmf is given by $\hat{P}_{\mathbf{xy}}$. Thus,

$$
\begin{aligned}
&\sum_{\mathbf{x}} \frac{Q_n^2(T_{\mathbf{x}|\mathbf{u}})}{|T_{\mathbf{x}|\mathbf{u}}|} \max_{T_{\mathbf{y}|\mathbf{x},\mathbf{u}}:\, d_2(\mathbf{x},\mathbf{y})\le nD_2, \hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}}\ge R} e^{-n(\hat{I}_{\mathbf{u};\mathbf{y}|\mathbf{x}}+\hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}}-R)} \\
\doteq\ &\max_{T_{\mathbf{x}|\mathbf{u}}} \max_{T_{\mathbf{y}|\mathbf{x},\mathbf{u}}:\, d_2(\mathbf{x},\mathbf{y})\le nD_2, \hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}}\ge R} e^{-n(\hat{I}_{\mathbf{u};\mathbf{y}|\mathbf{x}}+\hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}}-R)} Q_n^2(T_{\mathbf{x}|\mathbf{u}}) \\
=\ &\max_{T_{\mathbf{x}|\mathbf{u}}} \max_{W\in A_n(R,D_2,\hat{P}_{\mathbf{ux}})} e^{-n(J(\hat{P}_{\mathbf{ux}}\times W)-R)} Q_n^2(T_{\mathbf{x}|\mathbf{u}}), \qquad (72)
\end{aligned}
$$

Recall the definition of $\mathcal{Q}_n^{cfc}$ as the set of channels that assign a uniform distribution to one conditional type $T_{\mathbf{x}|\mathbf{u}}$ for every $\mathbf{u}$. We shall next show that there is no loss in asymptotic optimality if attention is confined to $\mathcal{Q}_n^{cfc}$ instead of Q_n^{ex}.

Let Q_n^{**} be the minimizer of

$$\min_{Q_n\in\bar{\mathcal{Q}}_n} \max_{T_{\mathbf{x}|\mathbf{u}}} \max_{W\in A_n(R,D_2,\hat{P}_{\mathbf{ux}})} e^{-n(J(\hat{P}_{\mathbf{ux}}\times W)-R)} Q_n^2(T_{\mathbf{x}|\mathbf{u}}) \qquad (73)$$

and recall that it must satisfy the distortion constraint imposed on the encoder ($Q_n^{**} \in \mathcal{Q}_n$). Let $\mathbf{x}'$ be such that $T_{\mathbf{x}'|\mathbf{u}} = \mathrm{argmax}_{T_{\mathbf{x}|\mathbf{u}}: d_1(\mathbf{u},\mathbf{x})\le nD_1} Q_n^{**}(T_{\mathbf{x}|\mathbf{u}})$ and a take a competing watermarking channel Q_n'' to be uniform in $T_{\mathbf{x}'|\mathbf{u}}$, i.e., $Q_n''(\mathbf{x}|\mathbf{u}) = \frac{1}{|T_{\mathbf{x}'|\mathbf{u}}|}\mathbf{1}\left\{\mathbf{x}\in T_{\mathbf{x}'|\mathbf{u}}\right\}$.

Define $D_{\min} \triangleq \min_{(u,x)\in\mathcal{U}\times\mathcal{X}: d_1(u,x)>0} d_1(u,x)$. In the following Lemma we prove that if $\mathcal{Q}_n = \mathcal{Q}_n^{(1)}(D_1)$, then

$$Q_n^{**}(T_{\mathbf{x}'|\mathbf{u}}) \geq \frac{D_{\min}}{2(n+1)^{|\mathcal{U}||\mathcal{X}|}(D_1 + \frac{D_{\min}}{2n})} \triangleq h^{(1)}(n). \tag{74}$$

Lemma 2. *Let $Q_n \in \mathcal{Q}_n^{(1)}(D_1)$ be an arbitrary watermarking channel, then, there exists a conditional type $T_{\mathbf{x}''|\mathbf{u}}$ s.t.*

$$Q_n(T_{\mathbf{x}''|\mathbf{u}}) \geq \frac{D_{\min}}{2(n+1)^{|\mathcal{U}||\mathcal{X}|}(D_1 + \frac{D_{\min}}{2n})} \tag{75}$$

and $d(\mathbf{u},\mathbf{x}'') \leq nD_1$.

Lemma 2 is proved in Appendix B.

In the next lemma we show that if $\mathcal{Q}_n = \mathcal{Q}_n^{(2)}(D_1,\nu)$ then

$$Q_n^{**}(T_{\mathbf{x}'|\mathbf{u}}) \geq \frac{1-e^{-n\nu}}{(n+1)^{|\mathcal{U}||\mathcal{X}|}} \triangleq h^{(2)}(n). \tag{76}$$

Lemma 3. *Given $Q_n \in \mathcal{Q}^{(2)}(D_1,\nu)$, there exists a conditional type $T_{\mathbf{x}''|\mathbf{u}}$ s.t. $Q_n(T_{\mathbf{x}''|\mathbf{u}}) \geq \frac{1-e^{-n\nu}}{(n+1)^{|\mathcal{U}||\mathcal{X}|}}$ and $d(\mathbf{u},\mathbf{x}'') \leq nD_1$.*

The proof of Lemma 3 can be found in Appendix B.

Let $h(n)$ stand for either $h^{(1)}(n)$ or $h^{(2)}(n)$ (depending on whether the set $\mathcal{Q}_n$ designates $\mathcal{Q}_n^{(1)}(D_1)$ or $\mathcal{Q}_n^{(2)}(D_1,\nu)$, respectively). Note that either way, $\log h(n) = O(\log n)$. Next, we shall show that substituting Q_n^{**} with Q_n'' results in a maximin error exponent that is no more than $O(\log n/n)$ smaller than that of Q_n^{**}. Let X' designate the RV whose pmf is $\hat{P}_{\mathbf{x}'}$. Then,

$$\begin{aligned}
&\max_{T_{\mathbf{x}|\mathbf{u}}} \max_{W\in A_n(R,D_2,\hat{P}_{\mathbf{ux}})} e^{-n\left[J(\hat{P}_{\mathbf{ux}}\times W)-R\right]}\left[Q_n^{**}(T_{\mathbf{x}|\mathbf{u}})\right]^2 \\
\geq& \max_{W\in A_n(R,D_2,\hat{P}_{\mathbf{ux}'})} e^{-n\left[J(\hat{P}_{\mathbf{ux}'}\times W)-R\right]}\left[Q_n^{**}(T_{\mathbf{x}'|\mathbf{u}})\right]^2 \\
\geq& \max_{W\in A_n(R,D_2,\hat{P}_{\mathbf{ux}'})} e^{-n\left[J(\hat{P}_{\mathbf{ux}'}\times W)-R\right]} h^2(n) \\
=& \max_{T_{\mathbf{x}|\mathbf{u}}} \max_{W\in A_n(R,D_2,\hat{P}_{\mathbf{ux}})} e^{-n\left[J(\hat{P}_{\mathbf{ux}}\times W)-R\right]}\left[Q_n''(T_{\mathbf{x}|\mathbf{u}})\right]^2 h^2(n) \\
\doteq& \max_{T_{\mathbf{x}|\mathbf{u}}} \max_{W\in A_n(R,D_2,\hat{P}_{\mathbf{ux}})} e^{-n\left[J(\hat{P}_{\mathbf{ux}}\times W)-R\right]}\left[Q_n''(T_{\mathbf{x}|\mathbf{u}})\right]^2
\end{aligned} \tag{77}$$

which demonstrates that asymptotically, Q_n'' is as good as Q_n^{**}.

Next,

$$\begin{aligned}
&\max_{T_{\mathbf{x}|\mathbf{u}}} \max_{W\in A_n(R,D_2,\hat{P}_{\mathbf{ux}})} e^{-n\left[J(\hat{P}_{\mathbf{ux}}\times W)-R\right]}\left[Q_n''(T_{\mathbf{x}|\mathbf{u}})\right]^2 \\
\stackrel{\cdot}{\geq}& \min_{Q_n\in\mathcal{Q}_n\cap\mathcal{Q}_n^{cfc}} \max_{T_{\mathbf{x}|\mathbf{u}}} \max_{W\in A_n(R,D_2,\hat{P}_{\mathbf{ux}})} e^{-n\left[J(\hat{P}_{\mathbf{ux}}\times W)-R\right]}\left[Q_n(T_{\mathbf{x}|\mathbf{u}})\right]^2 \\
=& \min_{T_{\mathbf{x}|\mathbf{u}}:\, d_1(\mathbf{u},\mathbf{x})\leq nD_1} \max_{W\in A_n(R,D_2,\hat{P}_{\mathbf{ux}})} e^{-n\left[J(\hat{P}_{\mathbf{ux}}\times W)-R\right]}.
\end{aligned} \tag{78}$$

The last inequality, together with (54), (54), (56), (65), (73), and (77) yields:

$$
\begin{aligned}
&\min_{Q_n\in\bar{\mathcal{Q}}_n, g_n\in\mathcal{G}_n} \max_{W_n\in\mathcal{W}_n} P_e(Q_n,g_n,W_n)\\
\dot{\geq}\;& (1-e^{-\lambda n})\max_{\mathbf{u}} P_U^n(T_{\mathbf{u}}) \min_{T_{\mathbf{x}|\mathbf{u}}:\, d_1(\mathbf{u},\mathbf{x})\le nD_1} \max_{W\in A_n(R,D_2,\hat{P}_{\mathbf{u}}\times Q)} e^{-n(J(\hat{P}_{\mathbf{u}}\times Q\times W)-R)}\\
&+e^{-\lambda n}\left(1-\frac{1}{M}\right)\\
\doteq\;& \max_{T_{\mathbf{u}}} \min_{T_{\mathbf{x}|\mathbf{u}}:\, d_1(\mathbf{u},\mathbf{x})\le nD_1} \max_{W\in A_n(R,D_2,\hat{P}_{\mathbf{ux}})} e^{-n\min\left\{\lambda, D(\hat{P}_{\mathbf{u}}||P_U)+J(\hat{P}_{\mathbf{ux}}\times W)-R\right\}} \qquad (79)
\end{aligned}
$$

where the last asymptotic inequality is because $\mathbf{U}$ is assumed memoryless and thus $P_U^n(T_{\mathbf{u}}) \doteq e^{-nD(\hat{P}_{\mathbf{u}}||P_U)}$.

Next, we shall be interested in replacing the maximization over conditional types (empirical pmf's) with maximization over continuous pmf's. We shall complete the derivation assuming that $D_2>0$. The case $D_2=0$ can be handled similarly and thus omitted.

Lemma 4. *Denote* $\Delta_n=\frac{|\mathcal{U}||\mathcal{X}||\mathcal{Y}|^2}{n}$. *For every* $V\in\mathcal{P}(\mathcal{U}\times\mathcal{X})$,

$$
\min_{W\in A_n(R,D_2,V)} J(V\times W) \le \min_{W\in A(R+\epsilon_{1,n},D_2-\epsilon_{2,n},V)} J(V\times W)+\epsilon_{1,n}. \qquad (80)
$$

where $\epsilon_{1,n}=(-8\Delta_n\log(\Delta_n/|\mathcal{U}||\mathcal{X}||\mathcal{Y}|))=8\frac{\log(n|\mathcal{Y}|)}{n}$ *and* $\epsilon_{2,n}=D_2^{\max}\Delta_n$.

As $\epsilon_{1,n},\epsilon_{2,n}$ are independent of V, Lemma 4 implies that

$$
\min_{W\in A_n(R,D_2,\hat{P}_{\mathbf{ux}})} J(\hat{P}_{\mathbf{ux}}\times W) \le \min_{W\in A(R+\epsilon_{1,n},D_2-\epsilon_{2,n},\hat{P}_{\mathbf{ux}})} J(\hat{P}_{\mathbf{ux}}\times W)+2\epsilon_{1,n}. \qquad (81)
$$

Recall the definition of $\mathbb{P}_n(\mathcal{A})$ in eq. (2). Equations (79) and (81) yield

$$
\begin{aligned}
&\min_{Q_n\in\bar{\mathcal{Q}}_n, g_n\in\mathcal{G}_n} \max_{W_n\in\mathcal{W}_n} P_e(Q_n,g_n,W_n)\\
\dot{\geq}\;& \max_{P\in\mathbb{P}_n(\mathcal{U})} \min_{Q\in\mathcal{D}_1(D_1,P)\cap\mathbb{P}_n(\mathcal{X},P)} \max_{W\in A(R+\epsilon_{1,n},D_2-\epsilon_{2,n},P\times Q)} e^{-n\min\{\lambda,D(P||P_U)+J(P\times Q\times W)-R\}}\\
\geq\;& \max_{P\in\mathbb{P}_n(\mathcal{U})} \min_{Q\in\mathcal{D}_1(D_1,P)} \max_{W\in A(R+\epsilon_{1,n},D_2-\epsilon_{2,n},P\times Q)} e^{-n\min\{\lambda,D(P||P_U)+J(P\times Q\times W)-R\}} \qquad (82)\\
\doteq\;& \max_{P\in\mathbb{P}_n(\mathcal{U})} e^{-n\min\left\{\lambda,D(P||P_U)+\tilde{E}(R+\epsilon_{1,n},P,D_1,D_2-\epsilon_{2,n})\right\}} \qquad (83)
\end{aligned}
$$

where (82) follows since $\mathcal{D}_1(D_1,P)\cap\mathbb{P}_n(\mathcal{X},P)\subseteq\mathcal{D}_1(D_1,P)$, and the last step is due to (22).

In order to complete the derivation, we should replace the minimization over empirical pmfs $P\in\mathbb{P}_n(\mathcal{U})$ with minimization over $\mathcal{P}(\mathcal{U})$. It can be easily shown that for every (D_1,D_2,R),

$$
\begin{aligned}
&\min_{P\in\mathbb{P}_n(\mathcal{U})}\left[D(P||P_U)+\tilde{E}(R,P,D_1,D_2)\right]\\
\le\;& \min_{P\in\mathcal{P}(\mathcal{U})}\left[D(P||P_U)+\tilde{E}(R+\epsilon_n,P,D_1+\tilde{\epsilon}_n,D_2-\epsilon'_n)\right]+\delta_n, \qquad (84)
\end{aligned}
$$

where $\epsilon_n = O\left(\frac{\log n}{n}\right)$, $\tilde{\epsilon}_n = O\left(\frac{1}{n}\right)$, $\epsilon'_n = O\left(\frac{1}{n}\right)$, and $\delta_n = O\left(\frac{\log n}{n}\right)$. The proof of this claim is very similar to the proof of Lemma 8, and thus is omitted.

Next, note that

$$\begin{aligned}
&\min\left\{\lambda, \min_{P\in\mathcal{P}(\mathcal{U})}\left(D(P||P_U) + \tilde{E}(R,P,D_1,D_2)\right)\right\} \\
= \;& \min\left\{\min_{P\in\mathcal{P}(\mathcal{U})}\left[D(P||P_U)+\lambda\right], \min_{P\in\mathcal{P}(\mathcal{U})}\left[D(P||P_U) + \tilde{E}(R,P,D_1,D_2)\right]\right\} \\
= \;& \min_{P\in\mathcal{P}(\mathcal{U})}\left[D(P||P_U) + \min\left\{\lambda, \tilde{E}(R,P,D_1,D_2)\right\}\right]. \qquad (85)
\end{aligned}$$

Thus,

$$\begin{aligned}
&\min_{Q_n\in\bar{\mathcal{Q}}_n, g_n\in\mathcal{G}_n}\;\max_{W_n\in\mathcal{W}_n} P_e(Q_n,g_n,W_n) \\
\dot{\geq}\;& e^{-nE(R+\epsilon_n, D_1+\tilde{\epsilon}_n, D_2-\epsilon'_n,\lambda)}. \qquad (86)
\end{aligned}$$

Finally, the proof of the continuity of $E(R,D_1,D_2,\lambda)$ w.r.t. (R,D_1,D_2) is omitted as it can be established similarly to that of $E_{sp}(R,D_1,D_2,\lambda)$ (to be defined later), which is proved in Lemma 9, Appendix B. □

7.2 Proof of the Direct Part of Theorem 1

In this section, we prove that

$$\min_{Q_n\in\mathcal{Q}_n\cap\mathcal{Q}_n^{ex},\, g_n\in\mathcal{G}_n}\;\max_{W_n\in\mathcal{W}_n} P_e(Q_n,g_n,W_n) \;\dot{\leq}\; e^{-nE(R,\lambda,D_1,D_2)}. \qquad (87)$$

Recall the definition of the $\mathcal{C}_n^{ex}$ in eq. (33) as the set of strongly exchangeable channels. The minimax probability of error, $\min_{Q_n\in\mathcal{Q}_n\cap\mathcal{Q}_n^{ex},\, g_n\in\mathcal{G}_n}\max_{W_n\in\mathcal{W}_n} P_e(Q_n,g_n,W_n)$, can be upper bounded by limiting Q_n to belong to any subset of $\mathcal{Q}_n\cap\mathcal{Q}_n^{ex}$ e.g., $\mathcal{Q}_n\cap\mathcal{Q}_n^{cfc}\cap\mathcal{C}_n^{ex}$, and by fixing the decoder to be a sub-optimal one, for instance, g_n^* (defined in (30)), i.e.,

$$\min_{\substack{Q_n\in\mathcal{Q}_n\cap\mathcal{Q}_n^{ex}\\ g_n\in\mathcal{G}_n}}\;\max_{W_n\in\mathcal{W}_n} P_e(Q_n,g,W_n) \leq \min_{Q_n\in\mathcal{Q}_n\cap\mathcal{Q}_n^{cfc}\cap\mathcal{C}_n^{ex}}\;\max_{W_n\in\mathcal{W}_n} P_e(Q_n,g_n^*,W_n). \qquad (88)$$

Our first claim, on the way to prove (87), is that when the decoder g_n^* is applied and the watermarking channel belongs to the set of strongly exchangeable, conditionally fixed composition channels, $\mathcal{Q}_n^{cfc}\cap\mathcal{Q}_n^{ex}$, the asymptotically worst case attack channel is strongly exchangeable as well. This claim is stated in the following lemma:

Lemma 5.

$$\min_{Q_n\in\mathcal{Q}_n\cap\mathcal{Q}_n^{cfc}\cap\mathcal{C}_n^{ex}}\;\max_{W_n\in\mathcal{W}_n} P_e(Q_n,g_n^*,W_n) = \min_{Q_n\in\mathcal{Q}_n\cap\mathcal{Q}_n^{cfc}\cap\mathcal{C}_n^{ex}}\;\max_{W_n\in\mathcal{W}_n\cap\mathcal{C}_n^{ex}} P_e(Q_n,g_n^*,W_n). \qquad (89)$$

Lemma 5 is proved in Appendix B.
Recall definition (61) of $Q_n(T_{\mathbf{x}|\mathbf{u}})$ and define $W_n(T_{\mathbf{y}|\mathbf{x}}) \triangleq W_n(T_{\mathbf{y}|\mathbf{x}}|\mathbf{u})= \sum_{\mathbf{y}'\in T_{\mathbf{y}|\mathbf{x}}} W_n(\mathbf{y}'|\mathbf{x})$. For the sake of brevity, denote $\tilde{\mathcal{Q}}_n \triangleq \mathcal{Q}_n \cap \mathcal{Q}_n^{cfc} \cap \mathcal{C}_n^{ex}$ and $\tilde{\mathcal{W}}_n \triangleq \mathcal{W}_n \cap \mathcal{C}_n^{ex}$. Hence, denoting

$$d(T_{\mathbf{x}|\mathbf{u},\mathbf{y}}) = \sum_{\mathbf{x}'\in T_{\mathbf{x}|\mathbf{u}}:\ |T_{\mathbf{x}'|\mathbf{u},\mathbf{y}}|\le|T_{\mathbf{x}|\mathbf{u},\mathbf{y}}|} \frac{1}{|T_{\mathbf{x}'|\mathbf{u}}|}, \tag{90}$$

we have by (88) and Lemma 5

$$\begin{aligned}
&\min_{\substack{Q_n\in\tilde{\mathcal{Q}}_n\\ g_n\in\mathcal{G}_n}} \max_{W_n\in\mathcal{W}_n} P_e(Q_n,g_n,W_n) \le \min_{Q_n\in\tilde{\mathcal{Q}}_n} \max_{W_n\in\tilde{\mathcal{W}}_n} P_e(Q_n,g_n^*,W_n)\\
= &\min_{Q_n\in\tilde{\mathcal{Q}}_n} \max_{W_n\in\tilde{\mathcal{W}}_n} \sum_{\mathbf{u}} P_U^n(\mathbf{u}) \sum_{\mathbf{x}} Q_n(\mathbf{x}|\mathbf{u}) \sum_{\mathbf{y}} W_n(\mathbf{y}|\mathbf{x})\\
&\left[1-\left(1-\sum_{\mathbf{x}'\in T_{\mathbf{x}|\mathbf{u}}:\ |T_{\mathbf{x}'|\mathbf{u},\mathbf{y}}|\le|T_{\mathbf{x}|\mathbf{u},\mathbf{y}}|} Q_n(\mathbf{x}'|\mathbf{u})\right)^{M-1}\right]\\
= &\min_{\left\{T_{\mathbf{x}|\mathbf{u}}:\ d_1(\mathbf{u},\mathbf{x})\le nD_1\right\}_{\mathbf{u}\in\mathcal{U}^n}} \max_{W_n\in\tilde{\mathcal{W}}_n} \sum_{\mathbf{u}} P_U^n(\mathbf{u}) \sum_{\mathbf{x}\in T_{\mathbf{x}|\mathbf{u}}} \frac{1}{|T_{\mathbf{x}|\mathbf{u}}|} \sum_{\mathbf{y}} W_n(\mathbf{y}|\mathbf{x}) \left(1-\left(1-d(T_{\mathbf{x}|\mathbf{u},\mathbf{y}})\right)^{M-1}\right)
\end{aligned} \tag{91}$$

where the last equality follows since a fixed composition encoder must assign probability 1 to some conditional type $T_{\mathbf{x}|\mathbf{u}}$ for every $\mathbf{u}$.

Since W_n is exchangeable, $W_n(\mathbf{y}|\mathbf{x}) = \frac{W_n(T_{\mathbf{y}|\mathbf{x}})}{|T_{\mathbf{y}|\mathbf{x}}|}$. Observe that $W_n \in \mathcal{W}_n$ implies that:

$$W_n(T_{\mathbf{y}|\mathbf{x}}) \le e^{-\lambda n\cdot\mathbf{1}\{d_2(\mathbf{x},\mathbf{y})>nD_2\}}. \tag{92}$$

Thus, for a given $T_{\mathbf{x}|\mathbf{u}}$,

$$\begin{aligned}
&\max_{W_n\in\tilde{\mathcal{W}}_n} \sum_{\mathbf{u}} P_U^n(\mathbf{u}) \sum_{\mathbf{x}\in T_{\mathbf{x}|\mathbf{u}}} \frac{1}{|T_{\mathbf{x}|\mathbf{u}}|} \sum_{\mathbf{y}} W_n(\mathbf{y}|\mathbf{x}) \left[1-\left(1-d(T_{\mathbf{x}|\mathbf{u}\mathbf{y}})\right)^{M-1}\right]\\
\le\ & \sum_{\mathbf{u}} P_U^n(\mathbf{u}) \sum_{\mathbf{x}\in T_{\mathbf{x}|\mathbf{u}}} \frac{1}{|T_{\mathbf{x}|\mathbf{u}}|} \sum_{\mathbf{y}} \frac{e^{-n\lambda\mathbf{1}\{d_2(\mathbf{x},\mathbf{y})>nD_2\}}}{|T_{\mathbf{y}|\mathbf{x}}|} \left[1-\left(1-d(T_{\mathbf{x}|\mathbf{u}\mathbf{y}})\right)^{M-1}\right].
\end{aligned} \tag{93}$$

Note that, $\left|\left\{\mathbf{x}'\in T_{\mathbf{x}|\mathbf{u}}:\ |T_{\mathbf{x}'|\mathbf{u},\mathbf{y}}|\le|T_{\mathbf{x}|\mathbf{u},\mathbf{y}}|\right\}\right| \le (n+1)^{|\mathcal{U}||\mathcal{X}||\mathcal{Y}|}|T_{\mathbf{x}|\mathbf{u},\mathbf{y}}|$, so we have:

$$d(T_{\mathbf{x}|\mathbf{u},\mathbf{y}}) = \sum_{\mathbf{x}'\in T_{\mathbf{x}|\mathbf{u}}:\ |T_{\mathbf{x}'|\mathbf{u},\mathbf{y}}|\le|T_{\mathbf{x}|\mathbf{u},\mathbf{y}}|} \frac{1}{|T_{\mathbf{x}|\mathbf{u}}|} \doteq \frac{|T_{\mathbf{x}|\mathbf{u},\mathbf{y}}|}{|T_{\mathbf{x}|\mathbf{u}}|} \tag{94}$$

Also, for $0\le d\le 1$,

$$1-(1-d)^{M-1} \le \min\left\{1, d(M-1)\right\}, \tag{95}$$

where the r.h.s. can be thought of as a union bound for independent events.

Fix $\mathbf{u}$ and $\mathbf{x}$, hence,

$$\begin{aligned}
&\sum_{\mathbf{y}} \frac{e^{-n\lambda \mathbf{1}\{d_2(\mathbf{x},\mathbf{y})>nD_2\}}}{|T_{\mathbf{y}|\mathbf{x}}|}\left[1-\left(1-d(T_{\mathbf{x}|\mathbf{uy}})\right)^{M-1}\right] \\
\leq\ &\sum_{\mathbf{y}} \frac{e^{-n\lambda \mathbf{1}\{d_2(\mathbf{x},\mathbf{y})>nD_2\}}}{|T_{\mathbf{y}|\mathbf{x}}|}\min\left\{1,(M-1)\frac{|T_{\mathbf{x}|\mathbf{u},\mathbf{y}}|}{|T_{\mathbf{x}|\mathbf{u}}|}\right\}. \qquad (96)
\end{aligned}$$

Recall that $\frac{|T_{\mathbf{x}|\mathbf{u},\mathbf{y}}|}{|T_{\mathbf{x}|\mathbf{u}}|} \doteq e^{-n\hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}}}$, (where $\hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}}$ is defined in (5)) this yields the asymptotic equality:

$$\begin{aligned}
&\sum_{\mathbf{y}} \frac{e^{-n\lambda \mathbf{1}\{d_2(\mathbf{x},\mathbf{y})>nD_2\}}}{|T_{\mathbf{y}|\mathbf{x}}|}\min\left\{1,e^{nR}\frac{|T_{\mathbf{x}|\mathbf{u},\mathbf{y}}|}{|T_{\mathbf{x}|\mathbf{u}}|}\right\} \\
\doteq\ &\max_{T_{\mathbf{y}|\mathbf{x},\mathbf{u}}} e^{-n\hat{I}_{\mathbf{u},\mathbf{y}|\mathbf{x}}}e^{-n\lambda \mathbf{1}\{d_2(\mathbf{x},\mathbf{y})>nD_2\}}\min\left\{1,e^{nR}e^{-n\hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}}}\right\} \\
=\ &\max_{T_{\mathbf{y}|\mathbf{x},\mathbf{u}}} e^{-n\hat{I}_{\mathbf{u},\mathbf{y}|\mathbf{x}}}e^{-n\lambda \mathbf{1}\{d_2(\mathbf{x},\mathbf{y})>nD_2\}}e^{-n\left|\hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}}-R\right|^{+}}, \qquad (97)
\end{aligned}$$

where $t^+ = \max\{0,t\}$. Recall the definitions of $\mathbb{P}_n(\mathcal{Y},V)$, $\mathbb{P}_n(\mathcal{X},P)$ (See (3)) and $\mathcal{D}_2(D_2,V)$ (defined in (16)). Let U,X be the RV's defined on alphabet $\mathcal{U}\times\mathcal{X}$ whose joint pmf is $\hat{P}_{\mathbf{ux}}$. Consider the following chain of inequalities:

$$\begin{aligned}
&\max_{T_{\mathbf{y}|\mathbf{x},\mathbf{u}}} e^{-n\left[\hat{I}_{\mathbf{u},\mathbf{y}|\mathbf{x}}+\lambda\mathbf{1}\{d_2(\mathbf{x},\mathbf{y})>nD_2\}+\left|\hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}}-R\right|^{+}\right]} \\
\leq\ &\max\left\{e^{-\lambda n}, \max_{T_{\mathbf{y}|\mathbf{x},\mathbf{u}}:d_2(\mathbf{x},\mathbf{y})\leq nD_2} e^{-n\left[\hat{I}_{\mathbf{u},\mathbf{y}|\mathbf{x}}+\left|\hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}}-R\right|^{+}\right]}\right\} \\
=\ &\max\left\{e^{-\lambda n}, \max_{W\in\mathbb{P}_n(\mathcal{Y},\hat{P}_{\mathbf{ux}})\cap\mathcal{D}_2(D_2,\hat{P}_{\mathbf{ux}})} e^{-n\left[I_{\hat{P}_{\mathbf{ux}}W}(U;Y|X)+\left|I_{\hat{P}_{\mathbf{ux}}W}(X;Y|U)-R\right|^{+}\right]}\right\} \\
\leq\ &\max\left\{e^{-\lambda n}, \max_{W\in\mathcal{D}_2(D_2,\hat{P}_{\mathbf{ux}})} e^{-n\left[I_{\hat{P}_{\mathbf{ux}}W}(U;Y|X)+\left|I_{\hat{P}_{\mathbf{ux}}W}(X;Y|U)-R\right|^{+}\right]}\right\}. \qquad (98)
\end{aligned}$$

Recall the definition of $\mathcal{D}_1(D_1,P)$ (15). Equations (91) - (98) yield:

$$\begin{aligned}
&\min_{\substack{Q_n\in\tilde{\mathcal{Q}}_n\\ g_n\in\mathcal{G}_n}}\max_{W_n\in\mathcal{W}_n} P_e(Q_n,g_n,W_n) \\
&\quad\stackrel{\cdot}{\leq} e^{-\lambda n}+\sum_{\mathbf{u}} P_U^n(\mathbf{u}) \min_{T_{\mathbf{x}|\mathbf{u}}:\, d_1(\mathbf{u},\mathbf{x})\leq nD_1}\ \max_{W\in\mathcal{D}_2(D_2,\hat{P}_{\mathbf{ux}})} e^{-n\left[I_{\hat{P}_{\mathbf{ux}}W}(U;Y|X)+\left|I_{\hat{P}_{\mathbf{ux}}W}(X;Y|U)-R\right|^{+}\right]} \\
&= e^{-\lambda n}+\sum_{\mathbf{u}} P_U^n(\mathbf{u}) \min_{Q\in\mathbb{P}_n(\mathcal{X},\hat{P}_{\mathbf{u}})\cap\mathcal{D}_1(D_1,\hat{P}_{\mathbf{u}})}\ \max_{W\in\mathcal{D}_2(D_2,\hat{P}_{\mathbf{u}}\times Q)} e^{-n\left[I_{\hat{P}_{\mathbf{u}}QW}(U;Y|X)+\left|I_{\hat{P}_{\mathbf{u}}QW}(X;Y|U)-R\right|^{+}\right]}. \qquad (99)
\end{aligned}$$

Define $B_n(D_1,\hat{P}_{\mathbf{u}}) = \mathbb{P}_n(\mathcal{X},\hat{P}_{\mathbf{u}})\cap\mathcal{D}_1(D_1,\hat{P}_{\mathbf{u}})$. The following lemma implies that the minimization over empirical pmf's $Q\in B_n(D_1,P)$ is asymptotically equivalent to the minimization over continuous pmf's $Q\in\mathcal{D}_1(D_1,\hat{P}_{\mathbf{u}})$. For a given pmf $V\in\mathcal{P}(\mathcal{U}\times\mathcal{X})$, define

$L(R, D_2, V)$:

$$L(R, D_2, V) = \min_{W \in \mathcal{D}_2(D_2, V)} \left(I_{VW}(U;Y|X) + |I_{VW}(X;Y|U) - R|^+ \right). \tag{100}$$

Lemma 6. *Define* $\Delta_n \stackrel{\triangle}{=} \frac{|\mathcal{U}||\mathcal{X}|^2}{n}$, $D_1^{\max} = \max_{u \in \mathcal{U}, x \in \mathcal{X}} d_1(u, x)$, $D_2^{\max} = \max_{x \in \mathcal{X}, y \in \mathcal{Y}} d_2(x, y)$, *thus,*

$$\max_{Q \in B_n(D_1 + \epsilon_{3,n}, P)} L(R, D_2 - \epsilon_{4,n}, P \times Q) \geq \max_{Q \in \mathcal{D}_1(D_1, P)} L(R, D_2, P \times Q) - \epsilon_{5,n}, \tag{101}$$

where $\epsilon_{3,n} = D_1^{\max} \Delta_n$, $\epsilon_{4,n} = D_2^{\max} \Delta_n$, $\epsilon_{5,n} = -8\Delta_n \log \left(\frac{\Delta_n}{|\mathcal{U}||\mathcal{X}||\mathcal{Y}|} \right)$.

Lemma 6 is proved in Appendix B.

Thus, eq. (99) and Lemma 6 imply that,

$$\begin{aligned}
& \min_{Q_n \in \tilde{\mathcal{Q}}_n,\, g_n \in \mathcal{G}_n} \max_{W_n \in \mathcal{W}_n} P_e(Q_n, g_n, W_n) \\
\stackrel{\cdot}{\leq} \; & \sum_{\mathbf{u}} P_U^n(\mathbf{u}) e^{-n \min\left\{\lambda, \tilde{E}(R, \hat{P}_{\mathbf{u}}, D_1 + \epsilon_{3,n}, D_2 - \epsilon_{4,n})\right\}} \\
\doteq \; & \max_{T_{\mathbf{u}}} P_U^n(T_{\mathbf{u}}) e^{-n \min\left\{\lambda, \tilde{E}(R, \hat{P}_{\mathbf{u}}, D_1 + \epsilon_{3,n}, D_2 - \epsilon_{4,n})\right\}} \\
\doteq \; & \max_{P \in \mathbb{P}_n(\mathcal{U})} e^{-n\left(D(P||P_U) + \tilde{E}(R, P, D_1 + \epsilon_{3,n}, D_2 - \epsilon_{4,n}, \lambda)\right)} \\
\leq \; & \max_{P \in \mathcal{P}(\mathcal{U})} e^{-n\left(D(P||P_U) + \tilde{E}(R, P, D_1 + \epsilon_{3,n}, D_2 - \epsilon_{4,n}, \lambda)\right)},
\end{aligned} \tag{102}$$

where the last asymptotic equality follows from the fact that the distribution of the covertext is memoryless. Finally, the Theorem follows from the continuity of $E(R, D_1, D_2, \lambda)$ in the pair (D_1, D_2). We do not prove this continuity as it can be established similarly to the continuity of the function $E_{sp}(R, D_1, D_2, \lambda)$ (to be defined later) w.r.t. (R, D_1, D_2) which is proved in Lemma 9, Appendix B.

□

7.3 Proof of Theorem 2

Proof of part (a):

Consider the attack channel

$$\tilde{W}_n^* = e^{-n\lambda} \tilde{W}_2^* + (1 - e^{-n\lambda}) W_n^{(1)} \tag{103}$$

where $W_n^{(1)}$ is defined in (32),

$$\tilde{W}_2^*(\mathbf{y}|\mathbf{x}) = b_{n,\mathbf{x}} \frac{1}{|T_{\mathbf{y}|\mathbf{x}}|} \mathbf{1}\left\{d_2(\mathbf{x}, \mathbf{y}) > nD_2\right\}, \tag{104}$$

and $b_{n,\mathbf{x}}$ is the appropriate polynomial normalization factor. Note that,

$$\tilde{W}_n^*(T_{\mathbf{y}|\mathbf{x}}) \doteq e^{-n\lambda \mathbf{1}\{d_2(\mathbf{x}, \mathbf{y}) > nD_2\}}. \tag{105}$$

Part (a) of Theorem 2 is proved by the two following claims.

Claim 1. *The channel $\tilde{W}_n^*$ is an asymptotically worst-case channel for $Q_n \in \mathcal{C}_n^{ex} \cap \mathcal{Q}_n^{cfc}$ and g_n^*.*

Proof of Claim 1:

In Lemma 5 (see Section 7.2), it is proved that for every $Q_n \in \mathcal{C}_n^{ex} \cap \mathcal{Q}_n^{cfc}$, and when g_n^* is applied, the worst case attack channel is exchangeable.

Denote $\tilde{a}(T_{\mathbf{uxy}}) = \sum_{\mathbf{x}' \in T_{\mathbf{x}|\mathbf{u}}:\, |T_{\mathbf{x}'|\mathbf{u},\mathbf{y}}| \le |T_{\mathbf{x}|\mathbf{u},\mathbf{y}}|} Q_n(\mathbf{x}'|\mathbf{u})$, thus, for all $Q_n \in \mathcal{C}_n^{ex} \cap \mathcal{Q}_n^{cfc}$,

$$\begin{aligned}
& \max_{W_n \in \mathcal{W}_n} P_e(Q_n, g_n^*, W_n) \\
= & \max_{W_n \in \mathcal{W}_n \cap \mathcal{C}_n^{ex}} P_e(Q_n, g_n^*, W_n) \\
= & \max_{W_n \in \mathcal{W}_n \cap \mathcal{C}_n^{ex}} \sum_{\mathbf{u}} P_U^n(\mathbf{u}) \sum_{\mathbf{x}} Q_n(\mathbf{x}|\mathbf{u}) \sum_{\mathbf{y}} W_n(\mathbf{y}|\mathbf{x}) \left[1 - (1 - \tilde{a}(T_{\mathbf{uxy}}))^{M-1}\right] \\
\le & \max_{W_n \in \mathcal{W}_n \cap \mathcal{C}_n^{ex}} \sum_{\mathbf{u}} P_U^n(\mathbf{u}) \sum_{\mathbf{x}} Q_n(\mathbf{x}|\mathbf{u}) \sum_{\mathbf{y}} \frac{e^{-n\lambda \mathbf{1}\{d_2(\mathbf{x},\mathbf{y}) > nD_2\}}}{|T_{\mathbf{y}|\mathbf{x}}|} \left[1 - (1 - \tilde{a}(T_{\mathbf{uxy}}))^{M-1}\right] \\
\doteq & \max_{W_n \in \mathcal{W}_n \cap \mathcal{C}_n^{ex}} \sum_{\mathbf{u}} P_U^n(\mathbf{u}) \sum_{\mathbf{x}} Q_n(\mathbf{x}|\mathbf{u}) \sum_{\mathbf{y}} \tilde{W}_n^*(\mathbf{y}|\mathbf{x}) \left[1 - (1 - \tilde{a}(T_{\mathbf{uxy}}))^{M-1}\right] && (106) \\
= & P_e(Q_n, g_n^*, \tilde{W}_n^*), && (107)
\end{aligned}$$

where the inequality follows from (92), and (106) follows from (105) . □

Claim 2. *For every watermarking channel Q_n,*

$$P_e(Q_n, g_n^*, \tilde{W}_n^*) \doteq P_e(Q_n, g_n^*, W_n^*). \quad (108)$$

Proof of Claim 2:

$$\begin{aligned}
P_e(Q_n, g_n^*, \tilde{W}_n^*) &= e^{-\lambda n} P_e(Q_n, g_n^*, \tilde{W}_2^*) + (1 - e^{-\lambda n}) P_e(Q_n, g_n^*, W_1^{(1)}) \\
&\overset{\cdot}{\le} e^{-\lambda n} P_e(Q_n, g_n^*, W_n^{(2)}) + (1 - e^{-\lambda n}) P_e(Q_n, g_n^*, W_1^{(1)}) \\
&= P_e(Q_n, g_n^*, W_n^*) && (109)
\end{aligned}$$

where the asymptotic inequality follows since $W_n^{(2)}$ is a useless channel, and thus, independently of the encoder-decoder (Q_n, g_n), the average probability of error $P_e(Q_n, g_n, W_n^{(2)})$ is greater than $1 - \frac{1}{M} \to 1$. □

Proof of part (b):
Recall the definition of $\{T_{\mathbf{x}^*|\mathbf{u}}\}$ (preceding eq. (28)),

$$
\begin{aligned}
& P_e(Q_n^*, g_n^*, W_n^*) \\
= & \sum_{\mathbf{u}} \sum_{\mathbf{x}\in T_{\mathbf{x}^*|\mathbf{u}}} \frac{1}{|T_{\mathbf{x}^*|\mathbf{u}}|} \sum_{\mathbf{y}} W_n^*(\mathbf{y}|\mathbf{x}) \left[1 - \left(1 - \sum_{\mathbf{x}'\in T_{\mathbf{x}^*|\mathbf{u}}:|T_{\mathbf{x}'|\mathbf{u},\mathbf{y}}|\le|T_{\mathbf{x}|\mathbf{u},\mathbf{y}}|} \frac{1}{|T_{\mathbf{x}^*|\mathbf{u}}|}\right)^{M-1}\right] \\
\le & \sum_{\mathbf{u}} \sum_{\mathbf{x}\in T_{\mathbf{x}^*|\mathbf{u}}} \frac{1}{|T_{\mathbf{x}^*|\mathbf{u}}|} \sum_{\mathbf{y}} \frac{e^{-\lambda n \mathbf{1}\{d_2(\mathbf{x},\mathbf{y})>nD_2\}}}{|T_{\mathbf{y}|\mathbf{x}}|} \left[1 - \left(1 - \sum_{\mathbf{x}'\in T_{\mathbf{x}^*|\mathbf{u}}:|T_{\mathbf{x}'|\mathbf{u},\mathbf{y}}|\le|T_{\mathbf{x}|\mathbf{u},\mathbf{y}}|} \frac{1}{|T_{\mathbf{x}^*|\mathbf{u}}|}\right)^{M-1}\right] \\
\le & \sum_{\mathbf{u}} \sum_{\mathbf{x}\in T_{\mathbf{x}^*|\mathbf{u}}} \frac{1}{|T_{\mathbf{x}^*|\mathbf{u}}|} \max\left\{e^{-\lambda n}, \max_{W\in\mathcal{D}_2(D_2,\hat{P}_{\mathbf{ux}})} e^{-n\left(I_{\hat{P}_{\mathbf{ux}}W}(U;Y|X)+\left|I_{\hat{P}_{\mathbf{ux}}W}(X;Y|U)-R\right|^+\right)}\right\} \qquad (110) \\
= & \sum_{\mathbf{u}} \min_{T_{\mathbf{x}|\mathbf{u}}:d_1(\mathbf{u},\mathbf{x})\le nD_1} \sum_{\mathbf{x}\in T_{\mathbf{x}|\mathbf{u}}} \frac{1}{|T_{\mathbf{x}|\mathbf{u}}|} \times \\
& \quad \max\left\{e^{-\lambda n}, \max_{W\in\mathcal{D}_2(D_2,\hat{P}_{\mathbf{ux}})} e^{-n\left(I_{\hat{P}_{\mathbf{ux}}W}(U;Y|X)+\left|I_{\hat{P}_{\mathbf{ux}}W}(X;Y|U)-R\right|^+\right)}\right\} \qquad (111) \\
\stackrel{\cdot}{\le} & \; e^{-nE(R,D_1,D_2,\lambda)} \qquad (112) \\
\stackrel{\cdot}{\le} & \min_{Q_n\in\mathcal{Q}_n\cap\mathcal{Q}_{ex},\, g\in\mathcal{G}_n} P_e(Q_n, g_n, W_n^*), \qquad (113)
\end{aligned}
$$

where the first inequality follows from (92), (110) follows from steps (96)-(98), equation (111) is by definition of $T_{\mathbf{x}^*|\mathbf{u}}$, (112) follows from (99)-(102), and the last asymptotic inequality is proved in Section 7.1.

□

7.4 Proof of Theorem 3

First, we define a class of channels that are implementable by a *genie-aided attacker* that observes the covertext sequence as well as the stegotext sequence. Then, we will establish a converse lemma for this class of channels which will be used to prove the theorem. This line of proof is based on the proof of the sphere packing bound for constant composition codes of the DMC [5, , Chapter 2, Theorem 5.3].

Definition 7. *A channel $P_n' : \mathcal{U}^n \times \mathcal{X}^n \to \mathcal{Y}^n$ is said to be invariantly uniform in type (IUT) if for every $(\mathbf{u}, \mathbf{x}, \mathbf{y}) \in \mathcal{U}^n \times \mathcal{X}^n \times \mathcal{Y}^n$,*

$$P_n'(\mathbf{y}|\mathbf{x},\mathbf{u}) = P_n'(\tilde{\mathbf{y}}|\tilde{\mathbf{x}},\mathbf{u}) \;\forall(\tilde{\mathbf{x}},\tilde{\mathbf{y}}) \in T_{\mathbf{xy}|\mathbf{u}}, \qquad (114)$$

and for all $(\mathbf{u}, \mathbf{x})$ there exists one conditional type $T_{\mathbf{y}|\mathbf{u},\mathbf{x}}$ such that

$$P_n'(\mathbf{y}'|\mathbf{x},\mathbf{u}) = \frac{1}{|T_{\mathbf{y}|\mathbf{x},\mathbf{u}}|} \mathbf{1}\left\{\mathbf{y}' \in T_{\mathbf{y}|\mathbf{x},\mathbf{u}}\right\}. \qquad (115)$$

For a sequence $\mathbf{u}$, a decoding rule g_n, and a deterministic codebook $F^d_{n,M}$, let $\Lambda^{\mathbf{u}}_m$ designate the set of sequences $\mathbf{y}$ such that $g_n(\mathbf{u},\mathbf{y}) = m$.

Definition 8. *We say that a code $\left(F^d_{n,M}, g_n\right)$ is an (n,ϵ)-code w.r.t. the sequence $\mathbf{u}$ for the channel $P'_n(\cdot|\cdot,\mathbf{u})$, if*

$$P'_n(\bar{\Lambda}^{\mathbf{u}}_m|\mathbf{x}_m(\mathbf{u}),\mathbf{u}) \leq \epsilon,\ \forall m \in \{1,...,M\}, \tag{116}$$

where $\bar{\Lambda}^{\mathbf{u}}_m = (\Lambda^{\mathbf{u}}_m)^c$.

Definition 9. *A deterministic code whose codewords, corresponding to $\mathbf{u}$, are members of the same conditional type $T_{\mathbf{x}|\mathbf{u}}$ will be referred to as a deterministic conditionally fixed composition*[13] *(DCCC) code.*

The following Lemma is a converse Theorem for IUT channels, and it forms an extension of [5, chapter 2, Corollary 1.4] which refers to DMC's.

Lemma 7. *(A converse for IUT channels) Fix a sequence $\mathbf{u}$. For every $\tau > 0, \epsilon > 0$ such that $\tau + \epsilon \leq 1$, if $\left(F^d_{n,M}, g_n\right)$ is a DCCC code whose codewords are in $T_{\mathbf{x}^*|\mathbf{u}}$ and is an (n,ϵ)-code w.r.t. $\mathbf{u}$ for the IUT channel $P'_n(\cdot|\cdot,\mathbf{u})$ then,*

$$\frac{1}{n}\log M \leq I_{\mathbf{x}^*;\mathbf{y}^*|\mathbf{u}} + a_{n,\tau}, \tag{117}$$

where $a_{n,\tau} \triangleq -\frac{\log\tau}{n} + |\mathcal{U}||\mathcal{X}||\mathcal{Y}|\frac{\log(n+1)}{n}$, and $\mathbf{y}^*$ is such that $\sum_{\mathbf{y}'\in T_{\mathbf{y}^*|\mathbf{u},\mathbf{x}^*}} P'_n(\mathbf{y}'|\mathbf{u},\mathbf{x}^*) = 1$.

The proof of Lemma 7 appears in Appendix B.

For a deterministic code $(F^d_{n,M}, g_n)$ and a channel W_n, define the maximum probability of error

$$P_{\max}(F^d_{n,M}, g_n, W_n) = \sum_{\mathbf{u}} P^n_U(\mathbf{u}) \max_{m\in\{1,\dots,M\}} \sum_{\mathbf{y}\in\bar{\Lambda}^{\mathbf{u}}_m} W(\mathbf{y}|\mathbf{x}_m(\mathbf{u})). \tag{118}$$

Obviously, $P_{\max}(F^d_{n,M}, g_n, W_n) \geq P_e(F^d_{n,M}, g_n, W_n)$. Also, a simple argument (see, e.g., [22, eq. 3.6.26]) establishes the following inequality between the maximum probability of error associated with a size M codebook and the average probability of error associated with a size $M/2$ codebook,

$$P_{\max}(F^d_{n,M}, g_n, W_n) \geq \frac{1}{2}P_e(F^d_{n,M/2}, g_n, W_n). \tag{119}$$

[13] The difference between this definition and the definition of a conditionally fixed composition channel (see definition 4), is that here the codebook is assumed deterministic.

Thus, we can establish the following chain of inequalities,

$$
\begin{aligned}
& \min_{\substack{F_{n,M}\in\mathcal{F}_{n,M}\\ g_n\in\mathcal{G}_n}} \max_{W_n\in\mathcal{W}_n} P_e(F_{n,M},g_n,W_n) \\
\geq & \min_{\substack{F_{n,M}\in\mathcal{F}_{n,M}\\ g_n\in\mathcal{G}_n}} P_e(F_{n,M},g_n,W_n^*) \\
= & \min_{\substack{F^d_{n,M}\in\mathcal{F}^d_{n,M}\\ g_n\in\mathcal{G}_n}} P_e(F^d_{n,M},g_n,W_n^*) && (120)\\
\geq & \frac{1}{2}\min_{\substack{F^d_{n,M/2}\in\mathcal{F}^d_{n,M/2}\\ g_n\in\mathcal{G}_n}} P_{\max}(F^d_{n,M/2},g_n,W_n^*) && (121)\\
\geq & (1-e^{-\lambda n})\min_{\substack{F^d_{n,M/2}\in\mathcal{F}^d_{n,M/2}\\ g_n\in\mathcal{G}_n}} P_{\max}(F^d_{n,M/2},g_n,W_n^{(1)}) \\
& +e^{-\lambda n}\min_{\substack{F^d_{n,M/2}\in\mathcal{F}^d_{n,M/2}\\ g_n\in\mathcal{G}_n}} P_{\max}(F^d_{n,M/2},g_n,W_n^{(2)}) && (122)\\
\geq & (1-e^{-\lambda n})\min_{\substack{F^d_{n,M/2}\in\mathcal{F}^d_{n}\\ g_n\in\mathcal{G}_n}} P_{\max}(F^d_{n,M/2},g_n,W_n^{(1)})+e^{-\lambda n}\left(1-\frac{1}{M}\right). && (123)
\end{aligned}
$$

where W_n^* is defined in (31), eq. (120) is due to the fact that randomization cannot improve performance relatively to the best code tuned to W_n^*, inequality (121) follows from (119), inequality (122) is because the r.h.s. expression corresponds to a situation where the encoder and decoder are informed of the channel mode while the l.h.s. corresponds to a situation where they are not informed, and (123) holds because $W_n^{(2)}$ is a useless channel.

Since for every $\mathbf{u}$ there are at the most $(n+1)^{|\mathcal{U}||\mathcal{X}|}$ conditional types $T_{\mathbf{x}|\mathbf{u}}$, for every codebook $F^d_{n,M/2}$, there must exist at least one conditional type such that

$$|F^d_{n,M/2}\cap T_{\mathbf{x}|\mathbf{u}}| \geq \frac{M/2}{(n+1)^{|\mathcal{U}||\mathcal{X}|}} \triangleq M' \doteq M, \tag{124}$$

and due to (43), this conditional type satisfies $d_1(\mathbf{u},\mathbf{x}')\leq nD_2, \forall\mathbf{x}'\in T_{\mathbf{x}|\mathbf{u}}$.

The above statement implies that:

$$
\begin{aligned}
& \min_{\substack{F^d_{n,M/2}\in\mathcal{F}^d_{n,M/2}\\ g_n\in\mathcal{G}_n}} P_{\max}(F^d_{n,M/2},g_n,W_n^{(1)}) \\
\geq & \sum_{\mathbf{u}} P_U^n(\mathbf{u}) \min_{T_{\mathbf{x}|\mathbf{u}}:d_1(\mathbf{u},\mathbf{x})\leq nD_1} \min_{\{\mathbf{x}_m(\mathbf{u})\}_{m=1}^{M'}\subseteq T_{\mathbf{x}|\mathbf{u}}} \max_m W_n^{(1)}(\bar{\Lambda}^{\mathbf{u}}_m|\mathbf{x}_m(\mathbf{u})), && (125)
\end{aligned}
$$

where $W_n^{(1)}(\bar{\Lambda}^{\mathbf{u}}_m|\mathbf{x}_m(\mathbf{u}))=\sum_{\mathbf{y}\in\bar{\Lambda}^{\mathbf{u}}_m} W_n^{(1)}(\mathbf{y}|\mathbf{x}_m(\mathbf{u}))$.

Fix $\mathbf{u}$ and let $\mathbf{x}^*$ be a member of the minimizer $T_{\mathbf{x}|\mathbf{u}}$ of the above expression. Take an arbitrary IUT channel P'_n which assigns probability one to a conditional type $T_{\mathbf{y}|\mathbf{u},\mathbf{x}^*}$ such that $d_2(\mathbf{x}^*,\mathbf{y})\leq nD_2$. Let $\mathbf{y}^*$ be an arbitrary member of this conditional type.

Fix a sequence $\{\tau_n\}_{n\geq 1}$ such that $\lim_{n\to\infty}\tau_n = 0$, to be specified later. According to Lemma 7, for every DCCC code whose M' codewords are in $T_{\mathbf{x}^*|\mathbf{u}}$, and which satisfies

$$M' > e^{n(I_{\mathbf{x}^*;\mathbf{y}^*|\mathbf{u}}+a_{n,\tau_n})}, \tag{126}$$

there exists m such that

$$P_n'(\bar{\Lambda}_m^{\mathbf{u}}|\mathbf{x}_m(\mathbf{u}),\mathbf{u}) \geq 1-\tau_n. \tag{127}$$

We now argue that (127) implies that $W_n^{(1)}(\bar{\Lambda}_m^{\mathbf{u}}|\mathbf{x}_m(\mathbf{u}))$ cannot be too small, and then we further lower bound the r.h.s. of (125).

Similarly to the derivation of the sphere packing bound in [5, Theorem 5.3], we use the log-sum inequality to assert that for two given pmf's P_1,P_2 on $\mathcal{Y}^n$ and any set $S\subseteq\mathcal{Y}^n$,

$$P_1(S)\log\frac{P_1(S)}{P_2(S)}+P_1(\bar{S})\log\frac{P_1(\bar{S})}{P_2(\bar{S})}\leq D(P_1||P_2), \tag{128}$$

and thus,

$$P_2(S)\geq\exp\left\{-\frac{D(P_1||P_2)+h(P_1(S))}{P_1(S)}\right\}, \tag{129}$$

where $h(x)\triangleq -x\log x-(1-x)\log(1-x)$, $0\leq x\leq 1$. Applying this to $P_n'(\cdot|\mathbf{u},\mathbf{x}_m(\mathbf{u}))$, $W_n^{(1)}(\cdot|\mathbf{x}_m(\mathbf{u}))$ and $\bar{\Lambda}_m^{\mathbf{u}}$ in the roles of P_1, P_2, S respectively, and defining

$$D(P_n'||W_n^{(1)}|\mathbf{u},\mathbf{x}_m(\mathbf{u}))=\sum_{\mathbf{y}\in\mathcal{Y}^n}P_n'(\mathbf{y}|\mathbf{x}_m(\mathbf{u}),\mathbf{u})\log\left(\frac{P_n'(\mathbf{y}|\mathbf{x}_m(\mathbf{u}),\mathbf{u})}{W_n^{(1)}(\mathbf{y}|\mathbf{x}_m(\mathbf{u}))}\right), \tag{130}$$

we get by (127),

$$\begin{aligned} W_n^{(1)}(\bar{\Lambda}_m^{\mathbf{u}}|\mathbf{x}_m(\mathbf{u})) &\geq \exp\left\{-\frac{D(P_n'||W_n^{(1)}|\mathbf{u},\mathbf{x}_m(\mathbf{u}))+h(1-\tau_n)}{1-\tau_n}\right\} \\ &\geq \frac{1}{2}\exp\left\{-D(P_n'||W_n^{(1)}|\mathbf{u},\mathbf{x}_m(\mathbf{u}))(1+2\tau_n)\right\}, \end{aligned} \tag{131}$$

if τ_n satisfies $\frac{h(1-\tau_n)}{1-\tau_n}<1$ (as we have $\tau_n<\frac{1}{2}$ and thus $\frac{1}{1-\tau_n}\leq 1+2\tau_n$). Next, note that

$$D(P_n'||W_n^{(1)}|\mathbf{u},\mathbf{x}_m(\mathbf{u}))\doteq\sum_{\mathbf{y}\in T_{\mathbf{y}^*|\mathbf{u},\mathbf{x}_m(\mathbf{u})}}\frac{1}{|T_{\mathbf{y}^*|\mathbf{u},\mathbf{x}_m(\mathbf{u})}|}\log\frac{|T_{\mathbf{y}^*|\mathbf{x}_m(\mathbf{u})}|}{|T_{\mathbf{y}^*|\mathbf{u},\mathbf{x}_m(\mathbf{u})}|}\doteq n\hat{I}_{\mathbf{u};\mathbf{y}^*|\mathbf{x}^*}, \tag{132}$$

where $\hat{I}_{\mathbf{u};\mathbf{y}^*|\mathbf{x}^*}$ is defined similarly to (5), hence,

$$W_n^{(1)}(\bar{\Lambda}_m^{\mathbf{u}}|\mathbf{x}_m(\mathbf{u}))\stackrel{\cdot}{\geq}\exp\left\{-n(1+2\tau_n)I_{\mathbf{u};\mathbf{y}^*|\mathbf{x}^*}\right\}. \tag{133}$$

As the above holds for every IUT channel which assigns probability 1 to conditional types which satisfy (126) and for which $d_2(\mathbf{x},\mathbf{y}) \le nD_2$, we have

$$
\begin{aligned}
& W_n^{(1)}(\bar{\Lambda}_m^{\mathbf{u}}|\mathbf{u},\mathbf{x}_m(\mathbf{u})) \\
\dot{\geq}\; & \exp\left\{-n(1+2\tau_n) \min_{T_{\mathbf{y}|\mathbf{x},\mathbf{u}}: d_2(\mathbf{x},\mathbf{y})\le nD_2, I_{\mathbf{x};\mathbf{y}|\mathbf{u}}+a_{\tau_n}\le R'} I_{\mathbf{u};\mathbf{y}|\mathbf{x}}\right\} \\
= \; & \exp\left\{-n(1+2\tau_n) \min_{W\in\bar{A}(R'-a_{n,\tau_n},D_2,\hat{P}_{\mathbf{ux}})\cap\mathbb{P}_n(\mathcal{Y},\hat{P}_{\mathbf{ux}})} I_{\hat{P}_{\mathbf{ux}}W}(U;Y|X)\right\}, \qquad (134)
\end{aligned}
$$

where $R' = \frac{\log M'}{n}$, $\mathbb{P}_n(\mathcal{Y},\hat{P}_{\mathbf{ux}})$ is defined in (3), and $\bar{A}(R,D_2,V)$ is defined in (21). Now, choose, for example, $\tau_n = \frac{\log n}{n}$. Gathering (120), (125) and (134), we get

$$
\begin{aligned}
& \min_{\substack{F_{n,M}\in\mathcal{F}_{n,M}\\ g_n\in\mathcal{G}_n}} \max_{W_n\in\mathcal{W}_n} P_e(F_{n,M}^d, g_n, W_n) \\
\dot{\geq}\; & \sum_{\mathbf{u}} P_U^n(\mathbf{u}) \min_{T_{\mathbf{x}|\mathbf{u}}: d_1(\mathbf{u},\mathbf{x})\le nD_1} \min_{\{\mathbf{x}_m(\mathbf{u})\}_{m=1}^{M'}, \mathbf{x}_m(\mathbf{u})\in T_{\mathbf{x}|\mathbf{u}}} \\
& \exp\left\{-n \min_{W\in\bar{A}(R'-a_{n,\tau_n},D_2,\hat{P}_{\mathbf{ux}})\cap\mathbb{P}_n(\mathcal{Y},\hat{P}_{\mathbf{ux}})} I_{\hat{P}_{\mathbf{ux}}W}(U;Y|X)\right\} + e^{-\lambda n} \\
= \; & \sum_{\mathbf{u}} P_U^n(\mathbf{u}) \min_{Q\in\mathcal{D}_1(D_1,\hat{P}_{\mathbf{u}})\cap\mathbb{P}_n(\mathcal{X},\hat{P}_{\mathbf{u}})} \\
& \exp\left\{-n \min_{W\in\bar{A}(R'-a_{n,\tau_n},D_2,\hat{P}_{\mathbf{u}}\times Q)\cap\mathbb{P}_n(\mathcal{Y},\hat{P}_{\mathbf{u}}\times Q)} I_{\hat{P}_{\mathbf{u}}QW}(U;Y|X)\right\} + e^{-\lambda n}. \qquad (135)
\end{aligned}
$$

The rest of the derivation is devoted to replacing optimizations over sets of empirical measures by optimizations over sets of continuous measures that are independent of n.

For every $V \in \mathcal{P}(\mathcal{U}\times\mathcal{X})$, define,

$$
\bar{A}_n(R,D_2,V) = \bar{A}(R,D_2,V)\cap\mathbb{P}_n(\mathcal{Y},V). \qquad (136)
$$

Denote $\Delta_n = \frac{|\mathcal{U}||\mathcal{X}||\mathcal{Y}|^2}{n}$, $D_1^{\max} = \max_{u,x} d_1(u,x)$ and $D_2^{\max} = \max_{x,y} d_2(x,y)$. Similarly to the proof of Lemma 4 (see Appendix B), one can show that

$$
\min_{W\in\bar{A}_n(R-\epsilon_{1,n},D_2-\epsilon_{2,n},V)} I_{VW}(U;Y|X) \le \min_{W\in\bar{A}(R,D_2,V)} I_{VW}(U;Y|X) + 2\epsilon_{1,n}. \qquad (137)
$$

where $\epsilon_{1,n} = (-8\Delta_n \log(\Delta_n/|\mathcal{U}||\mathcal{X}||\mathcal{Y}|)) = 8\frac{\log(n|\mathcal{Y}|)}{n}$ and $\epsilon_{2,n} = D_2^{\max}\Delta_n$.

Thus, denoting $P(T_{\mathbf{u}}) = \sum_{\mathbf{u}} P_U^n(\mathbf{u})$, we have

$$
\begin{aligned}
& \sum_{\mathbf{u}} P_U^n(\mathbf{u}) \min_{Q\in\mathcal{D}_1(D_1,\hat{P}_{\mathbf{u}})\cap\mathbb{P}_n(\mathcal{X},\hat{P}_{\mathbf{u}})} \exp\left\{-n \min_{W\in\bar{A}(R'-a_{n,\tau_n},D_2,\hat{P}_{\mathbf{u}}\times Q)\cap\mathbb{P}_n(\mathcal{Y},\hat{P}_{\mathbf{u}}\times Q)} I_{\hat{P}_{\mathbf{u}}QW}(U;Y|X)\right\} \\
\ge \; & \sum_{T_{\mathbf{u}}} P(T_{\mathbf{u}}) \min_{Q\in\mathcal{D}_1(D_1,\hat{P}_{\mathbf{u}})} \exp\left\{-n \min_{W\in\bar{A}(R'-a_{n,\tau_n}-\epsilon_{1,n},D_2-\epsilon_{2,n},\hat{P}_{\mathbf{u}}\times Q)} I_{\hat{P}_{\mathbf{u}}QW}(U;Y|X)\right\} \\
\doteq \; & \exp\left\{-n \min_{\hat{P}_{\mathbf{u}}\in\mathcal{P}_n(\mathcal{U})} \left[D(\hat{P}_{\mathbf{u}}||P_U) + \max_{Q\in\mathcal{D}_1(D_1,\hat{P}_{\mathbf{u}})} \min_{W\in\bar{A}(R'-a_{n,\tau_n}-\epsilon_{1,n},D_2-\epsilon_{2,n},\hat{P}_{\mathbf{u}}\times Q)} I_{\hat{P}_{\mathbf{u}}QW}(U;Y|X)\right]\right\} \\
= \; & \exp\left\{-n \min_{\hat{P}_{\mathbf{u}}\in\mathbb{P}_n(\mathcal{U})} \left[D(\hat{P}_{\mathbf{u}}\|P_U) + \tilde{E}_{sp}(R-a_{n,\tau_n}-\epsilon_{1,n},\hat{P}_{\mathbf{u}},D_1,D_2-\epsilon_{2,n})\right]\right\}. \qquad (138)
\end{aligned}
$$

where $\tilde{E}_{sp}(\cdot,\cdot,\cdot,\cdot)$ is defined in (24) and the last step follows since $\mathbf{U}$ is memoryless and thus $P(T_{\mathbf{u}}) \doteq e^{-nD(\hat{P}_{\mathbf{u}}\|P_U)}$.

The following Lemma is used to replace the minimization over $\mathbb{P}_n(\mathcal{U})$ with minimization over continuous measures $\mathcal{P}(\mathcal{U})$.

Lemma 8. *For every* (R, D_1, D_2),

$$\begin{aligned} & \min_{\hat{P}_{\mathbf{u}}\in\mathbb{P}_n(\mathcal{U})} \left[D(\hat{P}_{\mathbf{u}}\|P_U) + \tilde{E}_{sp}(R+\epsilon_{6,n}, \hat{P}_{\mathbf{u}}, D_1+\epsilon_{7,n}, D_2+\epsilon_{8,n})\right] \\ \leq & \min_{P\in\mathcal{P}(\mathcal{U})} \left[D(P\|P_U) + \tilde{E}_{sp}(R, P, D_1, D_2)\right] + \delta_{1,n}. \end{aligned} \tag{139}$$

where $\Delta_n = \frac{|\mathcal{U}|^2}{n}$, $\epsilon_{6,n} = -4\Delta_n \log\left(\frac{\Delta_n}{|\mathcal{U}||\mathcal{X}||\mathcal{Y}|}\right)$, $\epsilon_{7,n} = D_1^{\max}\Delta_n$, $\epsilon_{8,n} = D_2^{\max}\Delta_n$, *and* $\delta_{1,n} = \epsilon_{6,n} - \Delta_n \log\left(\frac{\Delta_n \min_{u\in\mathcal{U}} P_U(u)}{|\mathcal{U}|}\right)$.

The proof of Lemma 8 appears in Appendix B.

Denoting $\epsilon_n = \epsilon_{1,n} + \epsilon_{6,n} + a_{n,\tau_n}$, $\epsilon'_n = \epsilon_{2,n} + \epsilon_{8,n}$, and gathering (135), (138) and (139), yields,

$$\begin{aligned} & \min_{\substack{F_{n,M}\in\mathcal{F}_{n,M} \\ g_n\in\mathcal{G}_n}} \max_{W_n\in\mathcal{W}_n} P_e(F^d_{n,M}, g_n, W_n) \\ \stackrel{.}{\geq} & \exp\left(-n\min\left\{\lambda, \min_{P\in\mathcal{P}(\mathcal{U})}\left[D(P\|P_U) + \tilde{E}_{sp}(R-\epsilon_n, P, D_1+\epsilon_{7,n}, D_2-\epsilon'_n)\right]\right\}\right) \\ = & e^{-nE_{sp}(R-\epsilon_n, D_1+\epsilon_{7,n}, D_2-\epsilon'_n, \lambda)}, \end{aligned} \tag{140}$$

where the last equality follows since the order of the two outmost minimizations can be replaced. Finally, the continuity of $E_{sp}(R, D_1, D_2, \lambda)$ for all $R \neq R_\infty(D_1, D_2)$ is proved in the follwing Lemma (see Appendix B for the proof), which establishes the desired result.

Lemma 9. *If* $R \neq R_\infty(D_1, D_2)$, *and* $\epsilon_n, \beta_n, \delta_n$ *are vanishing as* $n \to \infty$, *then,*

$$\lim_{n\to\infty} E_{sp}(R+\epsilon_n, D_1+\beta_n, D_2+\delta_n, \lambda) = E_{sp}(R, D_1, D_2, \lambda). \tag{141}$$

7.5 Proof of the Direct Part of Theorem 4

The proof of the direct theorem is an immediate consequence of Theorem 1, as the smallest rate at which $E(R, D_1, D_2, \lambda)$ vanishes, denoted by $\underline{C}$, serves as a lower bound on the capacity. It is left to be shown that C is that rate. Recall that,

$$E(R, D_1, D_2, \lambda) = \min_{P\in\mathcal{P}(\mathcal{A})} \left[D(P||P_U) + \tilde{E}(R, P, D_1, D_2, \lambda)\right]. \tag{142}$$

For $E(R, D_1, D_2, \lambda)$ to be zero, $D(P||P_U)$ must be zero, and thus, the minimizing P must be equal to P_U, and therefore $\underline{C}$ is the smallest rate for which $\tilde{E}(R, P_U, D_1, D_2, \lambda) = 0$. Thus, recalling (17), we look for the smallest R satisfying

$$\max_{Q\in\mathcal{D}_1(D_1,P_U)} \min_{W\in\mathcal{D}_2(D_2,P_U\times Q)} \left(I_{P_UQW}(U;Y|X) + |I_{P_UQW}(X;Y|U) - R|^+\right) = 0. \tag{143}$$

The two summands in the above expression are non-negative by definition, so we have $I_{P_U QW}(U;Y|X) = 0$, which implies that $U \leftrightarrow X \leftrightarrow Y$ form a Markov chain. Consequently, $\underline{C}$ is the smallest R solving the equation,

$$\max_{Q \in \mathcal{D}_1(D_1, P_U)} \min_{W \in \mathcal{D}_2(D_2, P_U \times Q):\, U \leftrightarrow X \leftrightarrow Y} |I_{P_U QW}(X;Y|U) - R|^+ = 0, \tag{144}$$

which is exactly C as defined in (51).

7.6 Proof of the Converse Part of Theorem 4

As stated in [2], a simple application of Fano's Theorem yields that a reliable rate R code must satisfy,

$$R \leq \liminf_{n \to \infty} \frac{1}{n} I(\mathbf{X};\mathbf{Y}|\mathbf{U},\Theta), \tag{145}$$

where $(\Theta, \mathbf{U}, \mathbf{X}, \mathbf{Y})$ is drawn according to $P_U^n \times P_\Theta \times P_{\mathbf{X}|\mathbf{U},\Theta} \times W_n$, P_Θ being the distribution of the secret key used by the encoder and decoder, and $P_{\mathbf{X}|\mathbf{U},\Theta}$ is the distribution of the stegotext given the covertext and the secret key used by the encoder.

Thus, recalling that $W_n \in \mathcal{W}_n$ and $P_{\mathbf{X}|\mathbf{U},\Theta}$ satisfies the constraint (50) gives us the following upper bound on the capacity

$$\frac{1}{n} \max_{P_\Theta} \max_{P_{\mathbf{X}|\mathbf{U},\Theta}:\, Pr(d_1(\mathbf{U},\mathbf{X}) > nD_1) = 0} \min_{W_n \in \mathcal{W}_n} I(\mathbf{X};\mathbf{Y}|\mathbf{U},\Theta). \tag{146}$$

Note that

$$\begin{aligned} I(\mathbf{X};\mathbf{Y}|\mathbf{U},\Theta) &= H(\mathbf{Y}|\mathbf{U},\Theta) - H(\mathbf{Y}|\mathbf{X},\mathbf{U},\Theta) \\ &= H(\mathbf{Y}|\mathbf{U},\Theta) - H(\mathbf{Y}|\mathbf{X}), \end{aligned} \tag{147}$$

where the last equality follows since $(\mathbf{U},\Theta) \leftrightarrow \mathbf{X} \leftrightarrow \mathbf{Y}$ form a Markov chain.

In order to establish a converse theorem, it is sufficient to find an attack channel for which reliable transmission is made impossible at rates above C. Thus, we shall consider conditionally fixed composition channels, i.e., channels which assign probability one to a single conditional type class $T_{\mathbf{y}|\mathbf{x}}$ given every sequence $\mathbf{x}$, and that assign equal probability to sequences within this particular conditional type class[14]. For such channels, we have,

$$H(\mathbf{Y}|\mathbf{X}) = E \log |T_{\mathbf{Y}|\mathbf{X}}|, \tag{148}$$

as given every $\mathbf{x}$, $\mathbf{Y}$ is distributed *uniformly* in some (specified) conditional type $T_{\mathbf{y}|\mathbf{x}}$. Further,

$$\begin{aligned} H(\mathbf{Y}|\mathbf{U},\Theta) &= H(\mathbf{Y}, \hat{P}_{\mathbf{YU}}|\mathbf{U},\Theta) \\ &= H(\hat{P}_{\mathbf{YU}}|\mathbf{U},\Theta) + H(\mathbf{Y}|\hat{P}_{\mathbf{YU}},\mathbf{U},\Theta) \\ &\leq O(\log n) + E \log |T_{\mathbf{Y}|\mathbf{U}}|, \end{aligned} \tag{149}$$

[14]In order to maintain the distortion constraint, the attacker will be restricted to choose conditional types $\{T_{\mathbf{y}|\mathbf{x}}\}$ which satisfy $d_2(\mathbf{x},\mathbf{y}) \leq nD_2$, and thus the distortion will be lower than nD_2 with probability one, i.e., the attacker lies in the class $\mathcal{W}_n(\infty, D_2)$.

where the last inequality is due to the fact that the entropy of a discrete RV is upper bounded by the logarithm of the cardinality of its alphabet.

Recall that for the sequences $\mathbf{u}, \mathbf{x}, \mathbf{y}$, we denote for convenience, $\hat{H}_{\mathbf{x}} = H_{\hat{P}_{\mathbf{x}}}(X)$, $\hat{H}_{\mathbf{y}|\mathbf{x}} = H_{\hat{P}_{\mathbf{xy}}}(Y|X)$, $\hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}} = I_{\hat{P}_{\mathbf{uxy}}}(X;Y|U)$. Let $\hat{H}_{\mathbf{X}}$ be the RV designating the entropy defined by the (random) empirical measure induced by $\mathbf{X}$ and define $\hat{H}_{\mathbf{Y}|\mathbf{X}}$ and $\hat{I}_{\mathbf{X};\mathbf{Y}|\mathbf{U}}$ analogously.

Therefore, since $|T_{\mathbf{y}|\mathbf{u}}| \doteq e^{n\hat{H}_{\mathbf{y}|\mathbf{u}}}$

$$
\begin{aligned}
I(\mathbf{X};\mathbf{Y}|\mathbf{U},\Theta) &\leq O(\log n) + nE\log|T_{\mathbf{Y}|\mathbf{U}}| - nE\log|T_{\mathbf{Y}|\mathbf{X}}| \\
&= O(\log n) + nE\left(\hat{H}_{\mathbf{Y}|\mathbf{U}} - \hat{H}_{\mathbf{Y}|\mathbf{X}}\right) \\
&\leq nE\left(\hat{I}_{\mathbf{X};\mathbf{Y}|\mathbf{U}}\right) + O(\log n),
\end{aligned} \tag{150}
$$

where the expectation is w.r.t. the joint distribution of $\mathbf{U}, \mathbf{X}, \mathbf{Y}$.

The rest of the proof will be devoted to upper bounding $E\left(\hat{I}_{\mathbf{X};\mathbf{Y}|\mathbf{U}}\right)$. Due to the fact that the marginal pmf of $(\mathbf{U}, \mathbf{X}, \mathbf{Y})$ induced by $P_U^n \times P_\Theta \times P_{\mathbf{X}|\mathbf{U},\Theta} \times W_n$ defines this expectation, the only degree of freedom left for the encoder, in order to maximize $E\left(\hat{I}_{\mathbf{X};\mathbf{Y}|\mathbf{U}}\right)$, is the conditional pmf of $\mathbf{X}$ given $\mathbf{U}$ which will be referred to as a channel, and will henceforth be denoted by $Q_n(\cdot|\cdot)$.

Recall the distortion constraint (50) imposed on the encoder. Denote by $\tilde{\mathcal{Q}}_n = \mathcal{Q}_n^{(2)}(\infty, D_1)$ and $\mathcal{W}_n$ the sets of conditional pmfs satisfying (50) and (9), respectively.

Recall the definition of the set of strongly exchangeable channels (33) $\mathcal{C}_{ex}$ as the channels Q_n satisfying $Q_n(\mathbf{x}|\mathbf{u}) = Q_n(\pi(\mathbf{x})|\pi(\mathbf{u}))$ for every permutation π and every $\mathbf{u}, \mathbf{x}$.

The following lemma implies that the maximizer of $\max_{Q_n\in\tilde{\mathcal{Q}}_n} \min_{W_n\in\mathcal{W}_n} E\left(\hat{I}_{\mathbf{X};\mathbf{Y}|\mathbf{U}}\right)$ is a strongly exchangeable channel.

Lemma 10.

$$
\max_{Q_n\in\tilde{\mathcal{Q}}_n} \min_{W_n\in\mathcal{W}_n} E\left(\hat{I}_{\mathbf{X};\mathbf{Y}|\mathbf{U}}\right) = \max_{Q_n\in\tilde{\mathcal{Q}}_n\cap\mathcal{C}_{ex}} \min_{W_n\in\mathcal{W}_n} E\left(\hat{I}_{\mathbf{X};\mathbf{Y}|\mathbf{U}}\right). \tag{151}
$$

Proof. Let Q_n be a given channel, and let Q_n^π be the π-permuted channel, i.e., $Q_n^\pi(\mathbf{x}|\mathbf{u}) = Q_n(\pi(\mathbf{x})|\pi(\mathbf{u}))$. Letting $\bar{Q}_n(\mathbf{x}|\mathbf{u}) = \frac{1}{n!}\sum_\pi Q_n^\pi(\mathbf{x}|\mathbf{u})$, we get

$$
\begin{aligned}
&\min_{W_n\in\mathcal{W}_n} E_{P_U^n\bar{Q}_nW_n}\left(\hat{I}_{\mathbf{X};\mathbf{Y}|\mathbf{U}}\right) \\
&= \min_{W_n\in\mathcal{W}_n} \frac{1}{n!}\sum_\pi E_{P_U^nQ_n^\pi W_n}\left(\hat{I}_{\mathbf{X};\mathbf{Y}|\mathbf{U}}\right) && (152) \\
&\geq \frac{1}{n!}\sum_\pi \min_{W_n\in\mathcal{W}_n} E_{P_U^nQ_n^\pi W_n}\left(\hat{I}_{\mathbf{X};\mathbf{Y}|\mathbf{U}}\right) && (153) \\
&= \min_{W_n\in\mathcal{W}_n} E_{P_U^nQ_nW_n}\left(\hat{I}_{\mathbf{X};\mathbf{Y}|\mathbf{U}}\right), && (154)
\end{aligned}
$$

where (152) follows from the linearity of the expectation and (154) is due to a symmetry argument: Since P_U^n is a stationary memoryless source, the summands in (153) are equal for all π. □

Note that this Lemma remains valid for every exchangeable source, i.e., a pmf $P_n \in \mathcal{P}(\mathcal{U}^n)$ that satisfies $P_n(\mathbf{u}) = P_n(\pi(\mathbf{u}))$, for every vector $\mathbf{u}$ and for every permutation π.

As a result of Lemma 10, and the fact that $P_U^n(\mathbf{U})$ is memoryless, we have,

$$\begin{aligned} E\left(\hat{I}_{\mathbf{X};\mathbf{Y}|\mathbf{U}}\right) &= \sum_{\mathbf{u},\mathbf{x}} \frac{P(T_{\mathbf{ux}})}{|T_{\mathbf{ux}}|} \sum_{\mathbf{y}\in T_{\mathbf{y}|\mathbf{x}}} \frac{1}{|T_{\mathbf{y}|\mathbf{x}}|} \hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}} \\ &= \sum_{\mathbf{u},\mathbf{x}} \frac{P(T_{\mathbf{ux}})}{|T_{\mathbf{ux}}|} \sum_{T_{\mathbf{y}|\mathbf{x},\mathbf{u}} \subseteq T_{\mathbf{y}|\mathbf{x}}} \frac{|T_{\mathbf{y}|\mathbf{x},\mathbf{u}}|}{|T_{\mathbf{y}|\mathbf{x}}|} \hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}}, \end{aligned} \tag{155}$$

where $P(T_{\mathbf{ux}}) = \sum_{(\mathbf{u}',\mathbf{x}')\in T_{\mathbf{ux}}} P_U^n(\mathbf{u}')Q_n(\mathbf{x}'|\mathbf{u}')$, and $T_{\mathbf{y}|\mathbf{x}}$ is the conditional type class given $\mathbf{x}$ to which the attacker assigns probability 1.

Define $B \triangleq |\mathcal{U}||\mathcal{X}||\mathcal{Y}|$. As the number of joint type classes $T_{\mathbf{uxy}}$ is upper bounded by $(n+1)^B$, for every positive function $g(n)$,

$$\begin{aligned} & \frac{1}{n} \sum_{\mathbf{u},\mathbf{x}} \frac{P(T_{\mathbf{ux}})}{|T_{\mathbf{ux}}|} \sum_{T_{\mathbf{y}|\mathbf{x},\mathbf{u}} \subseteq T_{\mathbf{y}|\mathbf{x}}} \frac{|T_{\mathbf{y}|\mathbf{x},\mathbf{u}}|}{|T_{\mathbf{y}|\mathbf{x}}|} \hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}} \\ \leq\; & \frac{1}{n} \sum_{\mathbf{u},\mathbf{x}} \frac{P(T_{\mathbf{ux}})}{|T_{\mathbf{ux}}|} \sum_{\left\{T_{\mathbf{y}|\mathbf{x},\mathbf{u}} \subseteq T_{\mathbf{y}|\mathbf{x}} : \frac{|T_{\mathbf{y}|\mathbf{x},\mathbf{u}}|}{|T_{\mathbf{y}|\mathbf{x}}|} \geq g(n)\right\}} \frac{|T_{\mathbf{y}|\mathbf{x},\mathbf{u}}|}{|T_{\mathbf{y}|\mathbf{x}}|} \hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}} + \frac{g(n)(n+1)^B}{n} \log|\mathcal{X}|. \end{aligned} \tag{156}$$

Next, note that

$$\frac{|T_{\mathbf{y}|\mathbf{x},\mathbf{u}}|}{|T_{\mathbf{y}|\mathbf{x}}|} \geq g(n) \Rightarrow \hat{I}_{\mathbf{u};\mathbf{y}|\mathbf{x}} \leq \frac{B}{n|\mathcal{U}|} \log \frac{n+1}{g(n)}. \tag{157}$$

Choose $g(n) = \frac{n+1}{n^{B+1}}$, thus the right most term in (156) vanishes, and we get

$$\begin{aligned} & \frac{1}{n} \sum_{\mathbf{u},\mathbf{x}} \frac{P(T_{\mathbf{ux}})}{|T_{\mathbf{ux}}|} \sum_{\left\{T_{\mathbf{y}|\mathbf{x},\mathbf{u}} \subseteq T_{\mathbf{y}|\mathbf{x}} : \frac{|T_{\mathbf{y}|\mathbf{x},\mathbf{u}}|}{|T_{\mathbf{y}|\mathbf{x}}|} \geq g(n)\right\}} \frac{|T_{\mathbf{y}|\mathbf{x},\mathbf{u}}|}{|T_{\mathbf{y}|\mathbf{x}}|} \hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}} \\ \leq\; & \frac{1}{n} \sum_{\mathbf{u},\mathbf{x}} \frac{P(T_{\mathbf{ux}})}{|T_{\mathbf{ux}}|} \sum_{\left\{T_{\mathbf{y}|\mathbf{x},\mathbf{u}} \subseteq T_{\mathbf{y}|\mathbf{x}} : \hat{I}_{\mathbf{u};\mathbf{y}|\mathbf{x}} \leq \frac{B' \log n}{n}\right\}} \frac{|T_{\mathbf{y}|\mathbf{x},\mathbf{u}}|}{|T_{\mathbf{y}|\mathbf{x}}|} \hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}} \\ \leq\; & \frac{1}{n} \sum_{\mathbf{u},\mathbf{x}} \frac{P(T_{\mathbf{ux}})}{|T_{\mathbf{ux}}|} \max_{\left\{T_{\mathbf{y}|\mathbf{x},\mathbf{u}} \subseteq T_{\mathbf{y}|\mathbf{x}} : \hat{I}_{\mathbf{u};\mathbf{y}|\mathbf{x}} \leq \frac{B' \log n}{n}\right\}} \hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}} \end{aligned} \tag{158}$$

where $B' \triangleq \frac{B}{|\mathcal{U}|}(\frac{B}{|\mathcal{U}|}+1)$ and the last inequality follows since $\sum_{T_{\mathbf{y}|\mathbf{x},\mathbf{u}} \subseteq T_{\mathbf{y}|\mathbf{x}}} \frac{|T_{\mathbf{y}|\mathbf{x},\mathbf{u}}|}{|T_{\mathbf{y}|\mathbf{x}}|} = 1$. If the set $\left\{T_{\mathbf{y}|\mathbf{x},\mathbf{u}} \subseteq T_{\mathbf{y}|\mathbf{x}} : \hat{I}_{\mathbf{u};\mathbf{y}|\mathbf{x}} \leq \frac{B' \log n}{n}\right\}$ is empty, then the appropriate summand is equal to zero.

Since $\mathbb{P}_n(\mathcal{Y}|\mathcal{U}\times\mathcal{X}) \subset \mathcal{P}(\mathcal{Y}|\mathcal{U}\times\mathcal{X})$, we have

$$\begin{aligned} & \max_{T_{\mathbf{y}|\mathbf{x},\mathbf{u}} \subseteq T_{\mathbf{y}|\mathbf{x}} : \hat{I}_{\mathbf{u};\mathbf{y}|\mathbf{x}} \leq \frac{B' \log n}{n}} \hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}} \\ \leq\; & \max_{W \in \Psi(\hat{P}_{\mathbf{y}|\mathbf{x}}) :\, I_{\hat{P}_{\mathbf{ux}}W}(U;Y|X) \leq \frac{B' \log n}{n}} \hat{I}_{\hat{P}_{\mathbf{ux}}W}(X;Y|U), \end{aligned} \tag{159}$$

where

$$\Psi(\hat{P}_{\mathbf{y}|\mathbf{x}}) \triangleq \left\{ W \in \mathcal{P}(\mathcal{Y}|\mathcal{U}\times\mathcal{X}) : \sum_{u\in\mathcal{U}} \hat{P}_{\mathbf{u}|\mathbf{x}}(u|x)W(y|u,x) = \hat{P}_{\mathbf{y}|\mathbf{x}}(y|x)\ \forall (y,x) \right\}. \tag{160}$$

Loosely speaking, the condition $I_{\hat{P}_{\mathbf{ux}}W}(U;Y|X) \le \frac{B'\log n}{n}$ implies that $U - X - Y$ is "almost" a Markov chain. Using continuity arguments, we prove that there exists a sequence $\epsilon_n > 0$ such that $\lim_{n\to\infty}\epsilon_n = 0$ and for all $\hat{P}_{\mathbf{ux}} \in \mathcal{P}(\mathcal{U}\times\mathcal{X})$

$$\begin{aligned}
& \max_{W\in\Psi(\hat{P}_{\mathbf{y}|\mathbf{x}}): I_{\hat{P}_{\mathbf{ux}}W}(U;Y|X)\le\frac{B'\log n}{n}} \hat{I}_{\hat{P}_{\mathbf{ux}}W}(U;Y|X) \\
\le\ & \max_{W\in\Psi(\hat{P}_{\mathbf{y}|\mathbf{x}}): I_{\hat{P}_{\mathbf{ux}}W}(U;Y|X)=0} \hat{I}_{\hat{P}_{\mathbf{ux}}W}(U;Y|X) + \epsilon_n \qquad (161)\\
=\ & I_{\hat{P}_{\mathbf{ux}}\hat{P}_{\mathbf{y}|\mathbf{x}}}(X;Y|U) + \epsilon_n. \qquad (162)
\end{aligned}$$

To see why this is true, we apply Pinsker's inequality [5] and argue that of all $(u,x)\in\mathcal{U}\times\mathcal{X}$, and every p.m.f. $\tilde{P}\in\mathcal{P}(\mathcal{U}\times\mathcal{X}\times\mathcal{Y})$:

$$\sum_{y\in\mathcal{Y}} \left|\tilde{P}(y|u,x) - \tilde{P}(y|x)\right| \le \sqrt{2\ln 2 D(\tilde{P}(\cdot|u,x)||\tilde{P}(\cdot|x))} \triangleq \sqrt{2\ln 2 I_{\tilde{P}}(u;Y|x)}. \tag{163}$$

Multiplying by $\tilde{P}(u,x)$ and summing over $(u,x)\in\mathcal{U}\times\mathcal{X}$, we get,

$$\begin{aligned}
\sum_{(u,x,y)\in\mathcal{U}\times\mathcal{X}\times\mathcal{Y}} \left|\tilde{P}(u,x)\tilde{P}(y|u,x) - \tilde{P}(u,x)\tilde{P}(y|x)\right| &\le \sum_{(u,x)\in\mathcal{U}\times\mathcal{X}} \tilde{P}(u,x)\sqrt{2\ln 2 I_{\tilde{P}}(u;Y|x)} \\
&\le \sqrt{2\ln 2 I_{\tilde{P}}(U;Y|X)}, \qquad (164)
\end{aligned}$$

where the last step is due to Jensen's inequality and the concavity of the square root function. Thus,

$$\begin{aligned}
& I(U;Y|X) \le \frac{B'\log n}{n} \Longrightarrow \\
& \sum_{(u,x,y)\in\mathcal{U}\times\mathcal{X}\times\mathcal{Y}} \left|\tilde{P}(u,x)\tilde{P}(y|u,x) - \tilde{P}(u,x)\tilde{P}(y|x)\right| \le \sqrt{2\ln 2\frac{B'\log n}{n}}, \qquad (165)
\end{aligned}$$

and the desired claim (161) follows from the uniform continuity of $I_{\tilde{P}}(X;Y|U)$ w.r.t. $\tilde{P}$.

Combining eqs. (146)-(161), we get

$$\begin{aligned}
& \frac{1}{n}\max_{P_\Theta}\ \max_{P_{\mathbf{X}|\mathbf{U},\Theta}:\, Pr(d_1(\mathbf{U},\mathbf{X})>nD_1)=0}\ \min_{W_n\in\mathcal{W}_n} I(\mathbf{X};\mathbf{Y}|\mathbf{U},\Theta) \\
\le\ & \frac{1}{n}\max_{Q_n\in\tilde{\mathcal{Q}}_n}\ \min_{\left\{T_{\mathbf{y}|\mathbf{x}}:d_2(\mathbf{x},\mathbf{y})\le nD_2\right\}_{\mathbf{x}}} \sum_{\mathbf{u},\mathbf{x}} \frac{P(T_{\mathbf{ux}})}{|T_{\mathbf{ux}}|} I_{\hat{P}_{\mathbf{ux}}\hat{P}_{\mathbf{y}|\mathbf{x}}}(X;Y|U) + \frac{\epsilon_n}{n}. \qquad (166)
\end{aligned}$$

Next, we shall further restrict the attacker to operate in a strongly-exchangeable manner, i.e., if $\mathbf{x},\mathbf{y}$ are such that $W_n(\mathbf{y}|\mathbf{x}) = \frac{1}{|T_{\mathbf{y}|\mathbf{x}}|}$, then for every permutation π, $W_n(\pi(\mathbf{y})|\pi(\mathbf{x})) = \frac{1}{|T_{\pi(\mathbf{y})|\pi(\mathbf{x})}|}$. Thus, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{\mathbf{u},\mathbf{x}} \frac{P(T_{\mathbf{ux}})}{|T_{\mathbf{ux}}|} I_{\hat{P}_{\mathbf{ux}}\hat{P}_{\mathbf{y}|\mathbf{x}}}(X;Y|U) \\
= \; &\frac{1}{n}\sum_{T_{\mathbf{x}}} P(T_{\mathbf{x}}) \sum_{T_{\mathbf{u}|\mathbf{x}}:d_1(\mathbf{u},\mathbf{x})\leq nD_1} P(T_{\mathbf{u}|\mathbf{x}}|\mathbf{X}\in T_{\mathbf{x}}) I_{\hat{P}_{\mathbf{ux}}\hat{P}_{\mathbf{y}|\mathbf{x}}}(X;Y|U), \qquad (167)
\end{aligned}$$

where $P(T_{\mathbf{x}}) = \sum_{\mathbf{u}\in\mathcal{U}^n,\mathbf{x}'\in T_{\mathbf{x}}} P_U^n(\mathbf{u})Q_n(\mathbf{x}'|\mathbf{u})$, $P(T_{\mathbf{u}|\mathbf{x}}|\mathbf{X}\in T_{\mathbf{x}}) = \frac{P(T_{\mathbf{ux}})}{P(T_{\mathbf{x}})}$, and (167) is well defined, as the empirical measure $\hat{P}_{\mathbf{y}|\mathbf{x}}$ induced by the conditional type class $T_{\mathbf{y}|\mathbf{x}}$ chosen by the attacker depends only on the type class of $\mathbf{x}$.

Next, fix a positive function $h(n)$. Obviously,

$$\begin{aligned}
&\frac{1}{n}\sum_{T_{\mathbf{x}}} P(T_{\mathbf{x}}) \sum_{T_{\mathbf{u}|\mathbf{x}}:d_1(\mathbf{u},\mathbf{x})\leq nD_1} P(T_{\mathbf{u}|\mathbf{x}}|\mathbf{X}\in T_{\mathbf{x}}) I_{\hat{P}_{\mathbf{ux}}\hat{P}_{\mathbf{y}|\mathbf{x}}}(X;Y|U) \\
\leq \; &\frac{1}{n}\sum_{T_{\mathbf{x}}:P(T_{\mathbf{x}})\geq h(n)} P(T_{\mathbf{x}}) \sum_{T_{\mathbf{u}|\mathbf{x}}:d_1(\mathbf{u},\mathbf{x})\leq nD_1,\, P(T_{\mathbf{u}|\mathbf{x}}|\mathbf{X}\in T_{\mathbf{x}})\geq h(n)} \\
&P(T_{\mathbf{u}|\mathbf{x}}|\mathbf{X}\in T_{\mathbf{x}}) I_{\hat{P}_{\mathbf{ux}}\hat{P}_{\mathbf{y}|\mathbf{x}}}(X;Y|U) + h^2(n)\frac{(n+1)^B}{n}\log|\mathcal{X}|, \qquad (168)
\end{aligned}$$

since the term $P(T_{\mathbf{u}|\mathbf{x}}|\mathbf{X}\in T_{\mathbf{x}})$ is equal for all pairs of sequences $(\mathbf{u}',\mathbf{x}')$ of the same joint type, as Q_n is *strongly* exchangeable (by Lemma 10) and $\mathbf{U}$ is memoryless.

Now, note that whenever $P(T_{\mathbf{x}}) > h(n)$,

$$P(T_{\mathbf{u}|\mathbf{x}}|\mathbf{X}\in T_{\mathbf{x}}) \leq P(T_{\mathbf{u}}|\mathbf{X}\in T_{\mathbf{x}}) = \frac{P(T_{\mathbf{u}})Pr(\mathbf{X}\in T_{\mathbf{x}}|T_{\mathbf{u}})}{P(T_{\mathbf{x}})} \leq \frac{P(T_{\mathbf{u}})}{h(n)}. \qquad (169)$$

Thus, whenever $P(T_{\mathbf{x}}) > h(n)$ and $P(T_{\mathbf{u}|\mathbf{x}}|\mathbf{X}\in T_{\mathbf{x}}) \geq h(n)$ we get that $P(T_{\mathbf{u}}) \geq h^2(n)$, and hence, $D(\hat{P}_{\mathbf{u}}||P_U) \leq -\frac{1}{n}\log h^2(n)$. Choosing now $h(n) = (n+1)^{-B/2}$, the right-most term of (168) vanishes and also we get,

$$\begin{aligned}
&\sum_{T_{\mathbf{u}|\mathbf{x}}:d_1(\mathbf{u},\mathbf{x})\leq nD_1,\, P(T_{\mathbf{u}|\mathbf{x}}|\mathbf{X}\in T_{\mathbf{x}})\geq h_2(n)} P(T_{\mathbf{u}|\mathbf{x}}|\mathbf{X}\in T_{\mathbf{x}}) I_{\hat{P}_{\mathbf{ux}}\hat{P}_{\mathbf{y}|\mathbf{x}}}(X;Y|U) \\
\leq &\sum_{\hat{P}_{\mathbf{u}|\mathbf{x}}:d_1(\mathbf{u},\mathbf{x})\leq nD_1,\, D(\hat{P}_{\mathbf{u}}||P_U)\leq\frac{B\log n}{n}} P(T_{\mathbf{u}|\mathbf{x}}|\mathbf{X}\in T_{\mathbf{x}}) I_{\hat{P}_{\mathbf{ux}}\hat{P}_{\mathbf{y}|\mathbf{x}}}(X;Y|U) \\
\leq &\max_{\hat{P}_{\mathbf{u}|\mathbf{x}}:d_1(\mathbf{u},\mathbf{x})\leq nD_1,\, D(\hat{P}_{\mathbf{u}}||P_U)\leq\frac{B\log n}{n}} I_{\hat{P}_{\mathbf{ux}}\hat{P}_{\mathbf{y}|\mathbf{x}}}(X;Y|U) \\
\leq &\max_{\tilde{P}_{U|X}\in\mathcal{P}(\mathcal{U}|\mathcal{X}):E_{\hat{P}_{\mathbf{x}}\tilde{P}_{U|X}} d_1(U,X)\leq nD_1,\, D(\tilde{P}_U||P_U)\leq\frac{B\log n}{n}} I_{\hat{P}_{\mathbf{x}}\tilde{P}_{U|X}\hat{P}_{\mathbf{y}|\mathbf{x}}}(X;Y|U)., \qquad (170)
\end{aligned}$$

where $\tilde{P}_U$ is the marginal p.m.f. induced by $\hat{P}_{\mathbf{x}}\times\tilde{P}_{U|X}$. Next, similarly to the argument used in (162), it is easy to be convinced that there exists a sequence $\epsilon'_n > 0$, $n \geq 1$ such that

$\lim_{n\to\infty}\epsilon'_n = 0$, and for all $\hat{P}_{\mathbf{u}} \in \mathcal{P}(\mathcal{U})$ and $\hat{P}_{\mathbf{xy}} \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$,

$$\begin{aligned}
&\max_{\tilde{P}_{U|X}\in\mathcal{P}(\mathcal{U}|\mathcal{X}):E_{\hat{P}_{\mathbf{x}}\tilde{P}_{U|X}}d_1(U,X)\le nD_1,\ D(\tilde{P}_U\|P_U)\le\frac{B\log n}{n}} I_{\hat{P}_{\mathbf{x}}\tilde{P}_{U|X}\hat{P}_{\mathbf{y}|\mathbf{x}}}(X;Y|U)\\
\le\ &\max_{\tilde{P}_{U|X}\in\mathcal{P}(\mathcal{U}|\mathcal{X}):E_{\hat{P}_{\mathbf{x}}\tilde{P}_{U|X}}d_1(U,X)\le nD_1,\ D(\tilde{P}_U\|P_U)=0} I_{\hat{P}_{\mathbf{x}}\tilde{P}_{U|X}\hat{P}_{\mathbf{y}|\mathbf{x}}}(X;Y|U)+\epsilon'_n. \qquad (171)
\end{aligned}$$

Recall the definition of $\mathcal{D}_1(D_1,P)$ in (15), and for the sake of brevity, denote $\tilde{\mathcal{D}}_1^{\mathbf{x}}(D_1,P_U) = \left\{Q\in\mathcal{D}_1(D_1,P_U) : \sum_{u\in\mathcal{U}}P_U(u)Q(x|u) = \hat{P}_{\mathbf{x}}(x)\forall x\in\mathcal{X}\right\}$. Note that,

$$\begin{aligned}
&\max_{\tilde{P}_{U|X}\in\mathcal{P}(\mathcal{U}|\mathcal{X}):E_{\hat{P}_{\mathbf{x}}\tilde{P}_{U|X}}d_1(U,X)\le nD_1,\ D(\tilde{P}_U\|P_U)=0} I_{\hat{P}_{\mathbf{x}}\tilde{P}_{U|X}\hat{P}_{\mathbf{y}|\mathbf{x}}}(X;Y|U)\\
=\ &\max_{Q\in\tilde{\mathcal{D}}_1^{\mathbf{x}}(D_1,P_U)} I_{P_UQ\hat{P}_{\mathbf{y}|\mathbf{x}}}(X;Y|U). \qquad (172)
\end{aligned}$$

Thus, neglecting the term $\frac{\epsilon_n+\epsilon'_n}{n}$, (166) is upper bounded by

$$\begin{aligned}
&\frac{1}{n}\max_{Q_n\in\mathcal{Q}_n}\sum_{T_{\mathbf{x}}:P(T_{\mathbf{x}})>h(n)}P(T_{\mathbf{x}})\min_{T_{\mathbf{y}|\mathbf{x}}:d_2(\mathbf{x},\mathbf{y})\le nD_2}\max_{Q\in\tilde{\mathcal{D}}_1^{\mathbf{x}}(D_1,P_U)} I_{P_UQ\hat{P}_{\mathbf{y}|\mathbf{x}}}(X;Y|U)\\
\le\ &\frac{1}{n}\max_{T_{\mathbf{x}}}\min_{T_{\mathbf{y}|\mathbf{x}}:d_2(\mathbf{x},\mathbf{y})\le nD_2}\max_{Q\in\tilde{\mathcal{D}}_1^{\mathbf{x}}(D_1,P_U)} I_{P_UQ\hat{P}_{\mathbf{y}|\mathbf{x}}}(X;Y|U)\\
=\ &\frac{1}{n}\max_{T_{\mathbf{x}}}\min_{T_{\mathbf{y}|\mathbf{x}}:d_2(\mathbf{x},\mathbf{y})\le nD_2}\max_{Q\in\tilde{\mathcal{D}}_1^{\mathbf{x}}(D_1,P_U)} I_{P_UQ\hat{P}_{\mathbf{y}|\mathbf{x}}}(X;Y|U)\\
=\ &\frac{1}{n}\max_{\hat{P}_{\mathbf{x}}}\min_{W\in\mathbb{P}_n(\mathcal{Y},\hat{P}_{\mathbf{x}}):\ E_{\hat{P}_{\mathbf{x}}W}d_2(X,Y)\le D_2}\max_{Q\in\tilde{\mathcal{D}}_1^{\mathbf{x}}(D_1,P_U)} I_{P_UQ\hat{P}_{\mathbf{y}|\mathbf{x}}}(X;Y|U). \qquad (173)
\end{aligned}$$

Next, we shall be interested in replacing the minimization over empirical measures $\left\{W\in\mathbb{P}_n(\mathcal{Y},\hat{P}_{\mathbf{x}}):\ E_{\hat{P}_{\mathbf{x}}W}d_2(X,Y)\le D_2\right\}$, with a minimization over continuous ones $\left\{W\in\mathcal{P}(\mathcal{Y}|\mathcal{X}):E_{\hat{P}_{\mathbf{x}}W}d_2(X,Y)\le D_2\right\}$. Let $\bar{Q},\bar{W}$ be the appropriate maximizer and minimizer of $\max_{\hat{P}_{\mathbf{x}}}\min_{W\in\mathcal{P}(\mathcal{Y}|\mathcal{X}):\ E_{\hat{P}_{\mathbf{x}}W}d_2(X,Y)\le D_2}\max_{Q\in\tilde{\mathcal{D}}_1^{\mathbf{x}}(D_1,P_U)} I_{P_UQ\hat{P}_{\mathbf{y}|\mathbf{x}}}(X;Y|U)$ respectively. According to Claim 6 (Appendix C), there exists $W_q^{(n)}\in\mathbb{P}_n(\mathcal{Y},\hat{P}_{\mathbf{x}})$ such that $||\hat{P}_{\mathbf{x}}\times\bar{W}-\hat{P}_{\mathbf{x}}\times W_q^{(n)}||\le\frac{|\mathcal{X}||\mathcal{Y}^2|}{n}\stackrel{\triangle}{=}\Delta_n$. Claim 7 (Appendix C, equations (C.17), (C.18)) implies that $E_{\hat{P}_{\mathbf{x}}W_q^{(n)}}d_2(X,Y)\le D_2+D_2^{\max}\Delta_n$ (where $D_2^{\max}=\max_{x,y}d_2(x,y)$) and for every $V\in\mathcal{P}(\mathcal{U}\times\mathcal{X})$, $|I_{VW_q^{(n)}}(X;Y|U)-I_{V\bar{W}}(X;Y|U)|\le -4\Delta_n\log\left(\frac{\Delta_n}{B}\right)$, where $B=|\mathcal{U}||\mathcal{X}||\mathcal{Y}|$, thus,

$$\begin{aligned}
&\frac{1}{n}\max_{\hat{P}_{\mathbf{x}}}\min_{W\in\mathbb{P}_n(\mathcal{Y},\hat{P}_{\mathbf{x}}):\ E_{\hat{P}_{\mathbf{x}}W}d_2(X,Y)\le D_2}\max_{Q\in\tilde{\mathcal{D}}_1^{\mathbf{x}}(D_1,P_U)} I_{P_UQ\hat{P}_{\mathbf{y}|\mathbf{x}}}(X;Y|U)\\
\le\ &\frac{1}{n}\max_{\hat{P}_{\mathbf{x}}}\min_{W\in\mathcal{P}(\mathcal{Y}|\mathcal{X}):\ E_{\hat{P}_{\mathbf{x}}W}d_2(X,Y)\le D_2-D_2^{\max}\Delta_n}\max_{Q\in\tilde{\mathcal{D}}_1^{\mathbf{x}}(D_1,P_U)} I_{P_UQW}(X;Y|U)\\
&-4\Delta_n\log\left(\frac{\Delta_n}{B}\right). \qquad (174)
\end{aligned}$$

Now we shall replace the order of the two inner most optimizations. Note that the sets over which these optimizations are performed are rectangular, closed, convex sets, thus, using the minimax Theorem (see, for example, [10]), we can switch the order of the optimizations, that is,

$$\begin{aligned}
&\max_{\hat{P}_{\mathbf{x}}} \min_{W\in\mathcal{P}(\mathcal{Y}|\mathcal{X}):\, E_{\hat{P}_{\mathbf{x}}W} d_2(X,Y)\leq D_2 - D_2^{\max}\Delta_n} \max_{Q\in\tilde{\mathcal{D}}_1^{\mathbf{x}}(D_1,P_U)} I_{P_U QW}(X;Y|U)\\
=\;& \max_{\hat{P}_{\mathbf{x}}} \max_{Q\in\tilde{\mathcal{D}}_1^{\mathbf{x}}(D_1,P_U)} \min_{W\in\mathcal{D}_2(D_2-D_2^{\max}\Delta_n,P_U\times Q)} I_{P_U QW}(X;Y|U)\\
=\;& \max_{Q\in\mathcal{D}_1(D_1,P_U)} \min_{W\in\mathcal{D}_2(D_2-D_2^{\max}\Delta_n,P_U\times Q)} I_{P_U QW}(X;Y|U). \qquad (175)
\end{aligned}$$

We shall next show that the function $\max_{Q\in\mathcal{D}_1(D_1,P_U)}\min_{W\in\mathcal{D}_2(D_2,P_U\times Q)} I_{P_UQW}(X;Y|U)$ is continuous in D_2, and as $\lim_{n\to\infty}\Delta_n = 0$, the desired result will follow.

Denote $\delta_n = D_2^{\max}\Delta_n$. For all $Q\in\mathcal{D}_1(D_1,P_U)$ and $W\in\mathcal{D}_2(D_2,P_U\times Q)$, define $W'\in\mathcal{P}(\mathcal{Y}|\mathcal{U}\times\mathcal{X})$ as follows:

$$W'(y|u,x) = \begin{cases} \frac{D_2-\delta_n}{D_2} W(y|u,x) & y\neq x \\ 1-\frac{D_2-\delta_n}{D_2}(1-W(x|u,x)) & y = x \end{cases}. \qquad (176)$$

By definition, $W'\in\mathcal{D}_2(D_2-\delta_n,P_U\times Q)$ and it can easily be verified that $\|P_U\times Q\times W - P_U\times Q\times W'\|\leq \frac{2\delta_n}{D_2-\delta_n}$, which implies that $|I_{P_UQW}(X;Y|U) - I_{P_UQW'}(X;Y|U)|\leq -4\frac{2\delta_n}{D_2-\delta_n}\log\left(\frac{2\delta_n}{(D_2-\delta_n)|\mathcal{U}||\mathcal{X}||\mathcal{Y}|}\right)$ (see Claim 7, Appendix C for the proof). This establishes the continuity of $\max_{Q\in\mathcal{D}_1(D_1,P_U)}\min_{W\in\mathcal{D}_2(D_2,P_U\times Q)} I_{P_UQW}(X;Y|U)$ in D_2, and yields Theorem 4.

A Proof of Lemma 1

Following are the proofs of the properties:

Proof of part (a)**:**

Recall the definitions of $\mathcal{D}_2(D_2,V)$ in eq. (16), $A(R,D_2,V)$ in eq. (20) and $\bar{A}(R,D_2,V)$ in eq. (21). Define,

$$\begin{aligned}
E_1(R,D_2,V) &= \min_{W\in\bar{A}(R,D_2,V)} I_{VW}(U;Y|X) &(A.1)\\
E_2(R,D_2,V) &= \min_{W\in A(R,D_2,V)} I_{VW}(U;Y|X) + I_{VW}(X;Y|U) - R &(A.2)\\
L(R,D_2,V) &= \min_{W\in\mathcal{D}_2(D_2,V)} I_{VW}(U;Y|X) + |I_{VW}(X;Y|U)-R|^+. &(A.3)
\end{aligned}$$

By definition,

$$L(R,D_2,V) = \min\left\{E_1(R,D_2,V),E_2(R,D_2,V)\right\}. \qquad (A.4)$$

Next we shall prove that

$$E_1(R,D_2,V)\geq E_2(R,D_2,V), \qquad (A.5)$$

and thus, by (A.4) we have,

$$L(R, D_2, V) = E_2(R, D_2, V), \tag{A.6}$$

which yields the desired result.

Proof of (A.5)

Fix V. Define

$$C(V) \quad = \min_{W \in \mathcal{D}_2(D_2,V): U \leftrightarrow X \leftrightarrow Y} I_{VW}(X;Y|U). \tag{A.7}$$

It is easy to see that

$$E_1(R, D_2, V) > 0 \text{ for all } R < C(V), \tag{A.8}$$

because for $E_1(R, D_2, V)$ to be zero, $I_{VW}(U;Y|X)$ must be zero, which means that $U \leftrightarrow X \leftrightarrow Y$ is a Markov chain that satisfies the distortion constraint. So $E_1(R, D_2, V)$ vanishes at the smallest R for which there exists W such that $E_{VW}d_2(X,Y) \le D_2$, $U \leftrightarrow X \leftrightarrow Y$ is a Markov chain and $I_{VW}(X;Y|U) \le R$ which in turn, is exactly the minimum of $I_{VW}(X;Y|U)$ under these constraints.

Notice that $E_1(R, D_2, V)$ is infinite if the set $\bar{A}(R, D_2, V)$ is empty. Let $R^\infty(D_2, V)$ be the infimum of those R for which $E_1(R, D_2, V) < \infty$. Thus, the inequality (A.5) holds trivially for $R \le R^\infty(D_2, V)$.

Claim 3. *The function $E_1(R, D_2, V)$ is convex in R for $R > R^\infty(D_2, V)$ and fixed V.*

Claim 3 is proved in Appendix C.

The convexity of $E_1(R, D_2, V)$ w.r.t. R, together with (A.8) and the fact that by definition, $E_1(R, D_2, V)$ is non-increasing with R, imply that $E_1(R, D_2, V)$ is strictly decreasing in R within the range of rates $R^\infty(D_2, V) < R \le C(V)$. It is therefore concluded that for this range of rates, the minimum in $E_1(R, D_2, V) = \min_{W \in \bar{A}(R,D_2,V)} I_{VW}(U;Y|X)$ is attained at $I_{VW}(X;Y|U) = R$. Thus,

$$\begin{aligned}
E_1(R, D_2, V) \quad &= \min_{W \in \mathcal{D}_2(D_2,V):\, I_{VW}(X;Y|U)=R} I_{VW}(U;Y|X) \\
&\ge \min_{W \in \mathcal{D}_2(D_2,V):\, I_{VW}(X;Y|U)\ge R} I_{VW}(U;Y|X) + I_{VW}(X;Y|U) - R \\
&= E_2(R, D_2, V),
\end{aligned} \tag{A.9}$$

which concludes the proof of (A.5).

Proof of part (b)**:**

Fix $Q \in \mathcal{P}(\mathcal{X}|\mathcal{U})$ and denote $V = P \times Q$. We shall first show that $L(R, D_2, V)$ (defined in (A.3)) is convex in (R, D_2) for fixed V, and then, the convexity of

$$\tilde{E}(R, P, D_1, D_2) = \max_{Q \in \mathcal{D}_1(D_1,P)} L(R, D_2, P \times Q),$$

being a maximum of convex functions will follow.

Recall that

$$L(R, D_2, V) = \min_{W \in \mathcal{D}_2(D_2,V)} I_{VW}(U;Y|X) + |I_{VW}(X;Y|U) - R|^+ . \tag{A.10}$$

Take two pairs (D_2^a, R^a) and (D_2^b, R^b), and let W^a, W^b be the minimizers corresponding to these pairs. Now, take $\alpha \in [0,1]$ and denote $\underline{W} = \alpha W^a + (1-\alpha)W^b$, $\underline{R} = \alpha R^a + (1-\alpha)R^b$, $\underline{D}_2 = \alpha D_2^a + (1-\alpha)D_2^b$. Note that if $W^a \in \mathcal{D}_2(D_2^a, V)$ and $W^b \in \mathcal{D}_2(D_2^b, V)$, then $\underline{W} \in \mathcal{D}_2(\underline{D}_2, V)$. Since the function $|x|^+$ is convex in x, and $I_{VW}(U;Y|X)$, $I_{VW}(U;Y|X)$ are convex in W, we have

$$\begin{aligned}
& \alpha \left(I_{VW^a}(U;Y|X) + |I_{VW^a}(X;Y|U) - R^a|^+\right) \\
& +(1-\alpha)\left(I_{VW^b}(U;Y|X) + \left|I_{VW^b}(X;Y|U) - R^b\right|^+\right) \\
\geq\ & I_{V\underline{W}}(U;Y|X) + |\alpha I_{VW^a}(X;Y|U) + (1-\alpha) I_{VW^b}(X;Y|U) - \underline{R}|^+ \\
=\ & I_{V\underline{W}}(U;Y|X) + \max\{0, \alpha I_{VW^a}(X;Y|U) + (1-\alpha) I_{VW^b}(X;Y|U) - \underline{R}\} \\
\geq\ & I_{V\underline{W}}(U;Y|X) + \max\{0, I_{V\underline{W}}(X;Y|U) - \underline{R}\} \\
\geq\ & \min_{W \in \mathcal{D}_2(\underline{D}_2,V)} I_{VW}(U;Y|X) + |I_{VW}(X;Y|U) - \underline{R}|^+ .
\end{aligned} \tag{A.11}$$

Proof of part (c):

Let V be a pmf defined on $\mathcal{U} \times \mathcal{X}$ and recall the definitions of $E_1(R, D_2, V)$, $E_2(R, D_2, V)$ and $C(V)$ in eqs. (A.1), (A.2) and (A.7). We shall next show that there exists a rate, denoted $R_{crit}(V) \in [0, C(V)]$, such that:

$$E_2(R, D_2, V) = \begin{cases} E_2(0, D_2, V) - R & R \in [0, R_{crit}(V))] \\ E_1(R, D_2, V) & R \in [R_{crit}(V), C(V)] \\ 0 & R \geq C(V) \end{cases} \tag{A.12}$$

In other words, $E_2(R, D_2, V)$ is affine with slope -1 in the interval $[0, R_{crit}(V)]$, and is equal to $E_1(R, D_2, V)$ in the interval $[R_{crit}(V), C(V)]$.

Let $\mathcal{P}(D_2, V)$ be the set of conditional distributions W, such that

$$\min_{\tilde{W} \in \mathcal{D}_2(D_2,V)} I_{V\tilde{W}}(U;Y|X) + I_{V\tilde{W}}(X;Y|U) = I_{VW}(U;Y|X) + I_{VW}(X;Y|U) \tag{A.13}$$

and let W^{**} be such that

$$I_{VW^{**}}(X;Y|U) = \max_{W \in \mathcal{P}(D_2,V)} I_{VW}(X;Y|U). \tag{A.14}$$

Define

$$R_{crit}(V) = I_{VW^{**}}(X;Y|U). \tag{A.15}$$

Thus, by definition of $E_2(R, D_2, V)$, it follows, that for all $0 \leq R \leq R_{crit}(V)$,

$$\begin{aligned}
E_2(R, D_2, V) &= \min_{W \in A(R,D_2,V)} I_{VW}(U;Y|X) + I_{VW}(X;Y|U) - R \\
&= I_{VW^{**}}(U;Y|X) + I_{VW^{**}}(X;Y|U) - R,
\end{aligned} \tag{A.16}$$

and in the range of rates $R_{crit}(V) < R \leq C(V)$,

$$E_2(R, D_2, V) > I_{VW^{**}}(U;Y|X) + I_{VW^{**}}(X;Y|U) - R. \tag{A.17}$$

Note that, by definition of $E_2(R, D_2, V)$ and $C(V)$ as $U \leftrightarrow X \leftrightarrow Y$ form a Markov chain iff $I_{VW}(U;Y|X) = 0$,

$$E_2(R, D_2, V) = 0 \ \ \forall R > C(V). \tag{A.18}$$

It is left to be shown that for all $R \in [R_{crit}(V), C(V)]$:

$$E_2(R, D_2, V) = E_1(R, D_2, V). \tag{A.19}$$

By eq. (A.5), we know that $E_2(R, D_2, V) \leq E_1(R, D_2, V)$, so in order to prove (A.19), assume, conversely, that there exists $R'' > R_{crit}(V)$ for which the minimum in

$$E_2(R, D_2, V) = \min_{W \in A(R, D_2, V)} I_{VW}(U;Y|X) + I_{VW}(X;Y|U) - R \tag{A.20}$$

is attained at $\bar{W}$, with $I_{V\bar{W}}(X;Y|U) > R''$. Now, define $R' = I_{V\bar{W}}(X;Y|U)$. Then, for all $R \in [R'', R']$:

$$\begin{aligned} & \min_{W \in A(R, D_2, V)} I_{VW}(U;Y|X) + I_{VW}(X;Y|U) \\ \leq \ & I_{V\bar{W}}(U;Y|X) + I_{V\bar{W}}(X;Y|U) \\ = \ & \min_{W \in A(R'', D_2, V)} I_{VW}(U;Y|X) + I_{VW}(X;Y|U), \end{aligned} \tag{A.21}$$

and since $\min_{W \in A(R, D_2, V)} [I_{VW}(U;Y|X) + I_{VW}(X;Y|U)]$ is obviously non-increasing in R, the above inequality is attained with equality, that is, for all $R \in [R'', R']$:

$$\begin{aligned} & \min_{W \in A(R, D_2, V)} [I_{VW}(U;Y|X) + I_{VW}(X;Y|U)] \\ = \ & \min_{W \in A(R'', D_2, V)} [I_{VW}(U;Y|X) + I_{VW}(X;Y|U)], \end{aligned} \tag{A.22}$$

and thus, for all $R \in [R'', R']$:

$$E_2(R, D_2, V) + R = E_2(R'', D_2, V) + R'', \tag{A.23}$$

which yields, that $E_2(R, D_2, V)$ is linear in the interval $[R', R'']$ with a slope of -1.

Since $E_2(R, D_2, V)$ is convex, and its slope at $[0, I_{VW^{**}}(X;Y|U)]$ is -1 (see (A.16)), one concludes that its slope at the entire interval $[0, R'']$ is -1, and thus, $E_2(R'', D_2, V) = I_{VW^{**}}(U;Y|X) + I_{VW^{**}}(X;Y|U) - R''$, which contradicts the assumption $R'' > R_{crit}(V)$.

This concludes the proof of (A.12). Next, note that,

$$\tilde{E}_{sp}(R, P, D_1, D_2) = \max_{Q \in \mathcal{D}_1(D_1, P)} E_1(R, D_2, P \times Q), \tag{A.24}$$

where $\tilde{E}_{sp}(R, P, D_1, D_2)$ is defined in (24).

Equation (22) implies that

$$\tilde{E}(R, P, D_1, D_2) = \max_{Q \in \mathcal{D}_1(D_1, P)} E_2(R, D_2, P \times Q), \tag{A.25}$$

and thus the proof of part (c) follows by (A.12), (A.24), (A.25 and by the convexity of $E_2(R, D_2, V)$ w.r.t. R.

B Proofs of Auxiliary Lemmas

Proof of Lemma 2:

Let ϵ_n be a function of n to be specified later on, so

$$\begin{aligned}\Pr\left\{d(\mathbf{U},\mathbf{X}) > n(D_1+\epsilon_n)|\mathbf{U}=\mathbf{u}\right\} &\leq \frac{E_{Q_n}(d(\mathbf{U},\mathbf{X})|\mathbf{U}=\mathbf{u})}{n(D_1+\epsilon_n)} \\ &\leq \frac{D_1}{D_1+\epsilon_n},\end{aligned} \tag{B.1}$$

where the first inequality follows from Markov's inequality and the second one follows from the fact that Q_n satisfies the distortion constraint.

Therefore,

$$\Pr\left\{d(\mathbf{U},\mathbf{X}) \leq n(D_1+\epsilon_n)|\mathbf{U}=\mathbf{u}\right\} \geq \frac{\epsilon_n}{(\epsilon_n+D_1)}.$$

Let $\mathbf{x}''$ be such that $T_{\mathbf{x}''|\mathbf{u}} = \text{argmax}\left\{Q_n(T_{\mathbf{x}|\mathbf{u}}) : d(\mathbf{u},\mathbf{x}) \leq n(D_1+\epsilon_n)\right\}$. Since the number of conditional types for which $d(\mathbf{u},\mathbf{x}) \leq n(D_1+\epsilon_n)$ is, polynomial, we have,

$$Q_n(T_{\mathbf{x}''|\mathbf{u}}) \geq \frac{\epsilon_n}{(n+1)^{|\mathcal{U}||\mathcal{X}|}(\epsilon_n+D_1)}.$$

Choosing for example, $\epsilon_n = \frac{D_{\min}}{2n}$ yields, $Q_n(T_{\mathbf{x}''|\mathbf{u}}) \geq \frac{D_{\min}}{2(n+1)^{|\mathcal{U}||\mathcal{X}|}(D_1+\frac{D_{\min}}{2n})}$ and $d(\mathbf{u},\mathbf{x}'') \leq nD_1 + \frac{D_{\min}}{2}$ which implies that $d(\mathbf{u},\mathbf{x}'') \leq nD_1$ by definition of $D_{\min}$.

Proof of Lemma 3:

Proof.

$$Pr\left\{d(\mathbf{U},\mathbf{X}) \leq nD_1|\mathbf{U}=\mathbf{u}\right\} \geq 1-e^{-n\nu},$$

therefore, since the number of types is polynomial,

$$\max_{T_{\mathbf{x}|\mathbf{u}}:\, d(\mathbf{u},x)\leq nD_1} Q_n(T_{\mathbf{x}|\mathbf{u}}) \geq \frac{1-e^{-n\nu}}{(n+1)^{|\mathcal{U}||\mathcal{X}|}}.$$

Now, take $T_{\mathbf{x}''|\mathbf{u}} = \text{argmax}_{T_{\mathbf{x}|\mathbf{u}}:\, d(\mathbf{u},x)\leq nD_1} Q_n(T_{\mathbf{x}|\mathbf{u}})$. □

Proof of Lemma 4:

Let $\bar{W}$ be such that $J(V\times\bar{W}) = \min_{W\in A(R+\epsilon_{1,n},D_2-\epsilon_{2,n},V)} J(V\times W)$.

According to Claim 6 (see Appendix C), there exists $W_q^{(n)} \in \mathbb{P}_n(\mathcal{Y},V)$ such that $V\times\bar{W}$ can be approximated by $V\times W_q^{(n)}$, i.e.,

$$||V\times W_q^{(n)} - V\times\bar{W}|| \leq \Delta_n. \tag{B.2}$$

Thus, recalling that $\bar{W} \in A(R+\epsilon_{1,n}, D_2-\epsilon_{2,n}, V)$ and applying Claim 7 (see Appendix C) to $V \times W_q^{(n)}$, $V \times \bar{W}$ in the roles of P,Q, respectively, yields,

$$\begin{aligned} &E_{VW_q^{(n)}} d_2(X;Y) \leq D_2 \\ &I_{VW_q^{(n)}}(X;Y|U) \geq R, \end{aligned} \tag{B.3}$$

and thus, $W_q^{(n)} \in A_n(R, D_2, V)$. Claim 7 also implies that

$$\left| I_{VW_q^{(n)}}(X;Y|U) + I_{VW_q^{(n)}}(U;Y|X) - I_{V\bar{W}}(X;Y|U) - I_{V\bar{W}}(U;Y|X) \right| \leq \epsilon_{1,n}, \tag{B.4}$$

hence,

$$\begin{aligned} & \min_{W \in A_n(R,D_2,V)} J(V \times W) \\ \leq\ & J(V \times W_q^{(n)}) = I_{VW_q^{(n)}}(X;Y|U) + I_{VW_q^{(n)}}(U;Y|X) \\ \leq\ & I_{V\bar{W}}(X;Y|U) + I_{V\bar{W}}(U;Y|X) + \epsilon_{1,n} \\ =\ & \min_{W \in A(R+\epsilon_{1,n}, D_2-\epsilon_{2,n}, V)} J(V \times W) + \epsilon_{1,n} \end{aligned} \tag{B.5}$$

$\square$

Proof of Lemma 6:

Let $\bar{Q}$ be such that $L(R, D_2, P \times \bar{Q}) = \max_{Q \in \mathcal{D}_1(D_1,P)} L(R, D_2, P)$. A simple modification of Claim 6 (see Appendix C), implies that there exists $Q_q^{(n)} \in \mathbb{P}_n(\mathcal{X}, P)$ such that $P \times \bar{Q}$ can be approximated by $P \times Q_q^{(n)}$, i.e.,

$$||P \times Q_q^{(n)} - P \times \bar{Q}|| \leq \Delta_n. \tag{B.6}$$

Consider the following claim:

Claim 4. *If V_1 and V_2 are two distributions on alphabet $\mathcal{B} \stackrel{\triangle}{=} \mathcal{U} \times \mathcal{X}$ such that,*

$$\|V_1 - V_2\| = \Delta \leq \frac{1}{2}, \tag{B.7}$$

then,

$$L(R, D_2 + \Delta D_2^{\max}, V_1) \leq L(R, D_2, V_2) + \epsilon(\Delta), \tag{B.8}$$

where, $D_2^{\max} = \max_{(x,y)\in\mathcal{X}\times\mathcal{Y}} d_2(x,y)$, and $\epsilon(\Delta) = -8\Delta \log\left(\frac{\Delta}{|\mathcal{U}||\mathcal{X}||\mathcal{Y}|}\right)$.

Thus, applying Claim 4 (see the proof in Appendix C) to $\left(P \times \bar{Q}\right)$, $\left(P \times Q_q^{(n)}\right)$ in the roles of V_1,V_2 respectively, yields,

$$L(R, D_2 - \epsilon_{4,n}, Q_q^{(n)}) \leq L(R, D_2, \bar{Q}) + \epsilon_{5,n}. \tag{B.9}$$

Also, recalling that $\bar{Q} \in \mathcal{D}_1(D_1, P)$ and applying Claim 7, we get that $E_{PQ_q^{(n)}} d_1(U,X) \leq D_1 + D_1^{\max}\Delta_n$ and thus, $Q_q^{(n)} \in B_n(D_1 + \epsilon_{3,n}, P)$. Hence,

$$
\begin{aligned}
& \max_{Q \in B_n(D_1+\epsilon_{3,n},P)} L(R, D_2 - \Delta_n D_2^{\max}, Q) \\
\geq \ & L(R, D_2 - \epsilon_{4,n}, P \times Q_q^{(n)}) \\
\geq \ & L(R, D_2, P \times \bar{Q}) - \epsilon_{5,n} \\
= \ & \max_{Q \in \mathcal{D}_1(D_1,P)} L(R, D_2, P \times Q) - \epsilon_{5,n}.
\end{aligned} \tag{B.10}
$$

$\square$

Proof of Lemma 7:

Let $\{\mathbf{x}_m(\mathbf{u})\}_{m=1}^M$ be the codewords corresponding to $\mathbf{u}$. Further, choose arbitrary representatives $\{\mathbf{y}_{m,\mathbf{u}}\}_{m,\mathbf{u}}$ such that $\sum_{\mathbf{y}' \in T_{\mathbf{y}_{m,\mathbf{u}}|\mathbf{u},\mathbf{x}_m(\mathbf{u})}} P_n'(\mathbf{y}'|\mathbf{x}_m(\mathbf{u}), \mathbf{u}) = 1$. Note that, due to the fact that the codewords $\{\mathbf{x}_m(\mathbf{u})\}_{m=1}^M$ are in the same conditional type given $\mathbf{u}$ and P_n' is an IUT channel (and thus satisfies (114)), we have

$$
\hat{P}_{\mathbf{u},\mathbf{x}_m(\mathbf{u}),\mathbf{y}_{m,\mathbf{u}}} = \hat{P}_{\mathbf{u},\mathbf{x}_{m'}(\mathbf{u}),\mathbf{y}_{m',\mathbf{u}}} \ \ \forall (m, m'), \tag{B.11}
$$

and thus,

$$
T_{\mathbf{y}_{m,\mathbf{u}}|\mathbf{u}} = T_{\mathbf{y}_{m',\mathbf{u}}|\mathbf{u}} \ \ \forall (m, m'). \tag{B.12}
$$

For the sake of brevity, assume that $\mathbf{x}^* = \mathbf{x}_1$ and $\mathbf{y}^* = \mathbf{y}_{1,\mathbf{u}}$. Thus, for all m,

$$
\sum_{\mathbf{y}' \in T_{\mathbf{y}^*|\mathbf{u}}} P_n'(\mathbf{y}'|\mathbf{u}, \mathbf{x}_m(\mathbf{u})) = \sum_{\mathbf{y}' \in T_{\mathbf{y}_{m,\mathbf{u}}|\mathbf{u}}} P_n'(\mathbf{y}'|\mathbf{u}, \mathbf{x}_m(\mathbf{u})) = \sum_{\mathbf{y}' \in T_{\mathbf{y}_{m,\mathbf{u}}|\mathbf{u},\mathbf{x}_m(\mathbf{u})}} P_n'(\mathbf{y}'|\mathbf{u}, \mathbf{x}_m(\mathbf{u})) = 1 \tag{B.13}
$$

The fact that $(C_{n,M}, g_n)$ is an (n, ϵ)-code implies that

$$
1 - P_n'(\Lambda_m^{\mathbf{u}}|\mathbf{u}, \mathbf{x}_m(\mathbf{u})) \leq \epsilon \ \forall m \in \{1, ..., M\}. \tag{B.14}
$$

Combining (B.13) and (B.14) we get

$$
P_n'(T_{\mathbf{y}^*|\mathbf{u}} \cap \Lambda_m^{\mathbf{u}}|\mathbf{u}, \mathbf{x}_m(\mathbf{u})) = \sum_{\mathbf{y}' \in T_{\mathbf{y}^*|\mathbf{u}} \cap \Lambda_m^{\mathbf{u}}} P_n'(\mathbf{y}'|\mathbf{u}, \mathbf{x}_m(\mathbf{u})) \geq \tau \ \ \forall m. \tag{B.15}
$$

Since $P_n'(T_{\mathbf{y}_{m,\mathbf{u}}|\mathbf{u},\mathbf{x}_m(\mathbf{u})}|\mathbf{u}, \mathbf{x}_m(\mathbf{u})) = 1$, one can assume without loss of generality that $\Lambda_m^{\mathbf{u}} \subseteq T_{\mathbf{y}_{m,\mathbf{u}}|\mathbf{u},\mathbf{x}_m(\mathbf{u})}$, and one gets

$$
\frac{|T_{\mathbf{y}^*|\mathbf{u}} \cap \Lambda_m^{\mathbf{u}}|}{|T_{\mathbf{y}_{m,\mathbf{u}}|\mathbf{u},\mathbf{x}_m(\mathbf{u})}|} \geq \tau \ \ \forall m. \tag{B.16}
$$

Due to (B.11), we have $|T_{\mathbf{y}^*|\mathbf{u},\mathbf{x}^*}| = |T_{\mathbf{y}_{m,\mathbf{u}}|\mathbf{u},\mathbf{x}_m(\mathbf{u})}|$ for all m and thus,

$$
|T_{\mathbf{y}^*|\mathbf{u}}| \geq \sum_m |T_{\mathbf{y}^*|\mathbf{u}} \cap \Lambda_m^{\mathbf{u}}| \geq M |T_{\mathbf{y}^*|\mathbf{u},\mathbf{x}^*}| \tau. \tag{B.17}
$$

Hence,

$$M \leq \frac{1}{\tau}\frac{|T_{\mathbf{y}^*|\mathbf{u}}|}{|T_{\mathbf{y}^*|\mathbf{u},\mathbf{x}_m(\mathbf{u})}|} \leq \frac{1}{\tau}(n+1)^{|\mathcal{U}||\mathcal{X}||\mathcal{Y}|}e^{nI_{\mathbf{x}^*\mathbf{y}^*|\mathbf{u}}}. \tag{B.18}$$

which concludes the proof of Lemma 7.

Proof of Lemma 8:

Fix $P \in \mathcal{P}(\mathcal{U})$. A simple modification of Claim 6 (see Appendix C), implies that there exists $P_q^{(n)} \in \mathcal{P}_n(\mathcal{U})$ such that,

$$\left\|P - P_q^{(n)}\right\| \leq \frac{|\mathcal{U}|^2}{n} \triangleq \Delta_n. \tag{B.19}$$

According to Claim 7, eq. (C.20), we have,

$$\begin{aligned} &\left|D(P\|P_U) - D(P_q^{(n)}\|P_U)\right| \\ \leq\ & -\Delta_n \log\left(\frac{\Delta_n \min_{u\in\mathcal{U}} P_U(u)}{|\mathcal{U}|}\right). \end{aligned} \tag{B.20}$$

As a result of Claim 7, eq. (C.16),

$$\mathcal{D}_1(D_1, P) \subseteq \mathcal{D}_1(D_1 + D_1^{\max}\Delta_n, P_q^{(n)}). \tag{B.21}$$

Also, for a given $Q \in \mathcal{P}(\mathcal{X}|\mathcal{U})$, equations (C.17), (C.18) imply that,

$$\bar{A}\left(R - 4\Delta_n \log\left(\frac{\Delta_n}{|\mathcal{U}||\mathcal{X}||\mathcal{Y}|}\right), D_2 + D_2^{\max}\Delta_n, P_q^{(n)} \times Q\right) \subseteq \bar{A}(R, D_2, P \times Q). \tag{B.22}$$

Finally,

$$\begin{aligned} &\max_{Q\in\mathcal{D}_1(D_1,P)} \min_{W\in\bar{A}(R,D_2,P\times Q)} I_{PQW}(U;Y|X) \\ \geq\ & \max_{Q\in\mathcal{D}_1(D_1,P)} \min_{W\in\bar{A}(R,D_2,P\times Q)} I_{P_q^{(n)}QW}(U;Y|X) - \epsilon_{6,n} \\ \geq\ & \max_{Q\in\mathcal{D}_1(D_1-\epsilon_{7,n},P_q^{(n)})} \min_{W\in\bar{A}(R+\epsilon_{6,n},D_2+\epsilon_{8,n},P_q^{(n)}\times Q)} I_{P_q^{(n)}QW}(U;Y|X) - \epsilon_{6,n} \end{aligned} \tag{B.23}$$

where the first inequality follows from (C.19), and the last inequality follows from (B.21)-(B.22).

Gathering (B.20)-(B.23) we obtain the desired result. □

Proof of Lemma 9

Since $E_{sp}(R, D_1, D_2, \lambda)$ is increasing with D_1 for fixed (R, D_2, λ), decreasing with R for fixed (D_1, D_2, λ), and decreasing with D_2 for fixed (R, D_1, λ), it is sufficient to prove that,

$$\lim_{n\to\infty} E_{sp}(R-\epsilon_n, D_1+\beta_n, D_2-\delta_n, \lambda) \le E_{sp}(R, D_1, D_2, \lambda), \tag{B.24}$$

for $\epsilon_n, \beta_n, \delta_n$ that are non-negative. Note that

$$E_{sp}(R, D_1, D_2, \lambda) = \min\left\{\lambda, \min_{P\in\mathcal{P}(\mathcal{U})}\left[D(P\|P_U) + \tilde{E}_{sp}(R, P, D_1, D_2)\right]\right\}, \tag{B.25}$$

and let $P^* \in \mathcal{P}(\mathcal{U})$ be the minimizer of the above expression.

It is therefore sufficient to prove that for all $R > R_\infty(D_1, D_2)$,

$$\lim_{n\to\infty} \tilde{E}_{sp}(R-\epsilon_n, P^*, D_1+\beta_n, D_2-\delta_n) \le \tilde{E}_{sp}(R, P^*, D_1, D_2). \tag{B.26}$$

Let Q^* be the maximizer of,

$$\begin{aligned} & \tilde{E}_{sp}(R-\epsilon_n, P^*, D_1+\beta_n, D_2-\delta_n) \\ = & \max_{Q\in\mathcal{D}_1(D_1+\beta_n, P^*)} \min_{W\in\bar{A}(R-\epsilon_n, D_2-\delta_n, P^*\times Q)} I_{P^*QW}(U;Y|X). \end{aligned} \tag{B.27}$$

Define

$$\tilde{Q}(x|u) = \begin{cases} \frac{D_1}{D_1+\beta_n} Q^*(x|u) & x \ne u \\ 1 - \frac{D_1}{D_1+\beta_n}(1 - Q^*(u|u)) & x = u \end{cases}. \tag{B.28}$$

Obviously, $\tilde{Q} \in \mathcal{D}_1(D_1, P^*)$, and it easy to verify that $\|P^*\times Q^* - P^*\times\tilde{Q}\| \le \frac{2\beta_n}{D_1+\beta_n} \overset{n\to\infty}{\to} 0$. Thus, similarly to Claim 4, it can be easily shown that,

$$\begin{aligned} & \min_{W\in\bar{A}(R-\epsilon_n, D_2-\delta_n, P^*\times Q^*)} I_{P^*Q^*W}(U;Y|X) \\ \le & \min_{W\in\bar{A}(R-2\epsilon_n, D_2-2\delta_n, P^*\times\tilde{Q})} I_{P^*\tilde{Q}W}(U;Y|X) + \tilde{\epsilon}_n, \end{aligned} \tag{B.29}$$

where $\lim_{n\to\infty}\tilde{\epsilon}_n = 0$.

Hence, we get,

$$\tilde{E}_{sp}(R-\epsilon_n, P^*, D_1+\beta_n, D_2-\delta_n) \le \tilde{E}_{sp}(R-2\epsilon_n, P^*, D_1, D_2-2\delta_n) + \tilde{\epsilon}_n. \tag{B.30}$$

It is left to be shown that

$$\lim_{n\to\infty} \tilde{E}_{sp}(R-2\epsilon_n, P^*, D_1, D_2-2\delta_n) \le \tilde{E}_{sp}(R, P^*, D_1, D_2). \tag{B.31}$$

It is easily verified that if $R > R_\infty(D_1, D_2)$, then, for n sufficiently large, for all $Q \in \mathcal{D}_1(D_1, P^*)$,

$$\bar{A}(R-2\epsilon_n, D_2-2\delta_n, P^*\times Q) \ne \emptyset \tag{B.32}$$

Proof of (B.32):

Fix $Q \in \mathcal{D}_1(D_1, P^*)$. Denote $\Delta_R = R - R_\infty(D_1, D_2)$. By definition of $R_\infty(D_1, D_2)$, the sets $\bar{A}(R - x_n, D_2, P^* \times Q)$ for $0 \le x_n < \Delta_R$ are all non-empty.

Let $\dot{W}$ be a member of $\bar{A}(R - \frac{\Delta_R}{2}, D_2, P^* \times Q)$, and define $\ddot{W}$ as follows

$$\ddot{W}(y|x) = \begin{cases} \frac{D_2 - 2\delta_n}{D_2} \dot{W}(y|x) & y \ne x \\ 1 - \frac{D_2 - 2\delta_n}{D_2} \left(1 - \dot{W}(x|x)\right) & y = x \end{cases}. \tag{B.33}$$

By definition, $E_{P^* \times Q \times \ddot{W}} d_2(X, Y) \le D_2 - 2\delta_n$, and $\|P^* \times Q \times \ddot{W} - P^* \times Q \times \dot{W}\| \le \frac{4\delta_n}{D_2 - \delta_n}$. Thus, according to Claim 7, eq. (C.18) (see Appendix C) we have, $\left|I_{P^* \times Q \times \ddot{W}}(X; Y|U) - I_{P^* \times Q \times \dot{W}}(X; Y|U)\right| \le -4\frac{4\delta_n}{D_2 - \delta_n} \log\left(\frac{4\delta_n}{(D_2 - \delta_n)|\mathcal{U}||\mathcal{X}|\mathcal{Y}|}\right) \triangleq \ddot{\epsilon}_n$. Therefore, since $\dot{W} \in \bar{A}(R - \frac{\Delta_R}{2}, D_2, P^* \times Q)$, we have $\ddot{W} \in \bar{A}(R - \frac{\Delta_R}{2} + \ddot{\epsilon}_n, D_2 - 2\delta_n, P^* \times Q)$, and for n sufficiently large $R - \frac{\Delta_R}{2} + \ddot{\epsilon}_n \le R - 2\epsilon_n$, and $\ddot{W} \in \bar{A}(R - 2\epsilon_n, D_2 - 2\delta_n, P^* \times Q)$ which concludes the proof of (B.32).

Knowing that for n sufficiently large and for all $Q \in \mathcal{D}_1(D_1, P^*)$ the set $\bar{A}(R - 2\epsilon_n, D_2 - 2\delta_n, P^* \times Q)$ is non-empty, we can use the following claim:

Claim 5. *Let $R > 0, D_2 > 0$ and $V \in \mathcal{P}(\mathcal{U} \times \mathcal{X})$ be given, and let $\bar{W}$ be some arbitrary point in the set $\bar{A}(R, D_2, V)$. If the set $\bar{A}(R - \epsilon_n, D_2 - \delta_n, V)$ is non empty, then, it contains a point W' such that for all $(u, x) \in \mathcal{U} \times \mathcal{X}$, $\sum_{y \in \mathcal{Y}} \left|\bar{W}(y|u, x) - W'(y|u, x)\right| \le 2\alpha_n$, where, $\alpha_n = \min\left\{\frac{\epsilon_n}{R - \epsilon_n}, \frac{\delta_n}{D_2 - \delta_n}\right\}$.*

the proof of Claim 5 appears in Appendix C.

Thus, for all $\bar{W} \in \bar{A}(R, D_2, P^* \times Q)$ there exists $W' \in \bar{A}(R - 2\epsilon_n, D_2 - 2\delta_n, P^* \times Q)$ such that, $\left\|P^* \times Q \times \bar{W} - P^* \times Q \times W'\right\| \le \alpha_n$ where $\lim_{n \to \infty} \alpha_n = 0$. Due to Claim 7 eq. (C.19), $|I_{P^* Q \bar{W}}(U; Y|X) - I_{P^* Q W'}(U; Y|X)| \le -4\alpha_n \log\left(\frac{\alpha_n}{|\mathcal{U}||\mathcal{X}||\mathcal{Y}|}\right)$, thus for all $Q \in \mathcal{D}_1(D_1, P^*)$,

$$\begin{aligned} &\min_{W \in \bar{A}(R - 2\epsilon_n, D_2 - 2\delta_n, P^* \times Q)} I_{P^* Q W}(U; Y|X) \\ \le &\min_{W \in \bar{A}(R, D_2, P^* \times Q)} I_{P^* Q W}(U; Y|X) - 4\alpha_n \log\left(\frac{\alpha_n}{|\mathcal{U}||\mathcal{X}||\mathcal{Y}|}\right), \end{aligned} \tag{B.34}$$

which yields (B.31) and thus concludes the proof of Lemma 9.

C Appendix: Proofs of Auxiliary Claims

Proof of Claim 3:

First, note that $I_{VW}(U; Y|X)$ and $I_{VW}(X; Y|U)$ are convex functions of W for fixed V.

Let $R_1, R_2 > R^\infty(D_2, V)$ be given, let W_1 and W_2 be minimizers of $E_1(R_1, D_2, V)$ and $E_1(R_2, D_2, V)$, respectively, and denote their convex combination by $\bar{W} = \alpha W_1 + (1 - \alpha) W_2$.

Note that by linearity of $E_{VW}d_2(X,Y)$ in W and by the convexity of $I_{VW}(X;Y|U)$ in W, we have

$$E_{V\bar{W}}d_2(X,Y) = \sum_{u,x} V(u,x) \sum_{y} \bar{W}(y|u,x)d_2(x,y) \le D_2, \text{ and} \tag{C.1}$$

$$I_{V\bar{W}}(X;Y|U) \le \alpha I_{VW_1}(X;Y|U) + (1-\alpha)I_{VW_2}(X;Y|U) \le \alpha R_1 + (1-\alpha)R_2 \tag{C.2}$$

So we get,

$$\begin{aligned} \alpha E_1(R_1,D_2,V) + (1-\alpha)E_1(R_2,D_2,V) &= \alpha I_{VW_1}(U;Y|X) + (1-\alpha)I_{VW_2}(U;Y|X) \\ &\ge I_{V\bar{W}}(U;Y|X) \\ &\ge \min_{W\in\bar{A}(\alpha R_1+(1-\alpha)R_2,D_2,V)} I(U;Y|X) \\ &= E_1(\alpha R_1 + (1-\alpha)R_2, D_2, V) \end{aligned} \tag{C.3}$$

where the first inequality follows from the convexity of $I_{VW}(U;Y|X)$ w.r.t. W for fixed V, and the last equality follows from (C.1) and (C.2).

□

Proof of Claim 4:

Let $\bar{W}$ be such that

$$\begin{aligned} L(R,D_2,V_2) &= \min_{W\in\mathcal{D}_2(D_2,V_2)} \left[I_{V_2W}(U;Y|X) + |I_{V_2W}(X;Y|U) - R|^+\right] \\ &= I_{V_2\bar{W}}(U;Y|X) + |I_{V_2\bar{W}}(X;Y|U) - R|^+ . \end{aligned} \tag{C.4}$$

According to Claim 7, we have $|E_{V_1}d_2(X;Y) - E_{V_2}d_2(X;Y)| \le D_2^{\max}\Delta$, and thus,

$$\mathcal{D}_2(D_2,V_2) \subseteq \mathcal{D}_2(D_2 + D_2^{\max}\Delta, V_1). \tag{C.5}$$

Claim 7 also implies that

$$\begin{aligned} &\left|I_{V_2\bar{W}}(U;Y|X) + |I_{V_2\bar{W}}(X;Y|U) - R|^+ - I_{V_1\bar{W}}(U;Y|X) - |I_{V_1\bar{W}}(X;Y|U) - R|^+\right| \\ \le\ & \epsilon(\Delta). \end{aligned} \tag{C.6}$$

Thus,

$$\begin{aligned} & L(R,D_2,V_2) \\ =\ & \left[I_{V_2\bar{W}}(U;Y|X) + |I_{V_2\bar{W}}(X;Y|U) - R|^+\right] \\ \ge\ & \left[I_{V_1\bar{W}}(U;Y|X) + |I_{V_1\bar{W}}(X;Y|U) - R|^+\right] - \epsilon(\Delta) \\ \ge\ & \min_{W\in\mathcal{D}_2(D_2+D_2^{\max}\Delta,V_1)} \left[I_{V_2W}(U;Y|X) + |I_{V_2W}(X;Y|U) - R|^+\right] - \epsilon(\Delta) \\ =\ & L(R, D_2 + D_2^{\max}\Delta, V_1) - \epsilon(\Delta) \end{aligned} \tag{C.7}$$

Proof of Claim 5:

Fix a point $W_0 \in \bar{A}(R-\epsilon_n, D_2-\delta_n, V)$. Since the set $\bar{A}(R,D_2,V)$ is convex and since, obviously, $\bar{A}(R-\epsilon_n, D_2-\delta_n, V) \subseteq \bar{A}(R,D_2,V)$, we have,

$$\left\{W:\ \exists 0\le\alpha\le 1:\ W = \alpha W_0 + (1-\alpha)\bar{W}\right\} \subseteq \bar{A}(R,D_2,V), \tag{C.8}$$

i.e., the entire line segment between W_0 and $\bar{W}$ is in $\bar{A}(R, D_2, V)$.

Next, we show that for n sufficiently large, there exists a point W' on this line that it is also in $\bar{A}(R-\epsilon_n, D_2-\delta_n, V)$, and satisfies $\sum_{y\in\mathcal{Y}} \left|\bar{W}(y|u,x) - W'(y|u,x)\right| \le \alpha_n$ for all $(u,x) \in \mathcal{U}\times\mathcal{X}$.

Denote

$$W' = (1-\alpha_n)\bar{W} + \alpha_n W_0. \tag{C.9}$$

Due to the convexity of $I_{VW}(X;Y|U)$ w.r.t. W, we have,

$$\begin{aligned} I_{VW'}(X;Y|U) &\le (1-\alpha_n) I_{V\bar{W}}(X;Y|U) + \alpha_n I_{VW_0}(X;Y|U) \\ &\le (1-\alpha_n)R + \alpha_n(R-\epsilon_n) \\ &= R - \alpha_n\left(R-\epsilon_n\right) \le R - \epsilon_n. \end{aligned} \tag{C.10}$$

Moreover,

$$\begin{aligned} E_{VW'} d_2(X,Y) &= (1-\alpha_n) E_{V\bar{W}} d_2(X,Y) + \alpha_n E_{VW_0} d_2(X,Y) \\ &\le (1-\alpha_n) D_2 + \alpha_n (D_2 - \delta_n) \\ &= D_2 - \alpha_n(D_2-\delta_n) \le D_2 - \delta_n. \end{aligned} \tag{C.11}$$

Thus, $W' \in \bar{A}(R-\epsilon_n, D_2-\delta_n, V)$.

Furthermore,

$$\begin{aligned} |W'(y|u,x) - \bar{W}(y|u,x)| &= |(1-\alpha_n)\bar{W}(y|u,x) + \alpha_n W_0(y|u,x) - \bar{W}(y|u,x)| \\ &\le \left|\alpha_n \left(\bar{W}(y|u,x) + W_0(y|u,x)\right)\right| \le 2\alpha_n. \end{aligned} \tag{C.12}$$

Claim 6. *For all $n \ge 1$, every pmf V on some finite set $\mathcal{B}$ and every conditional pmf $\bar{W} \in \mathcal{P}(\mathcal{A}|\mathcal{B})$, there exists $W_q^{(n)} \in \mathbb{P}_n(\mathcal{A}, V)$ such that*

$$||V \times W_q^{(n)} - V \times \bar{W}|| = \sum_{(b,a)\in\mathcal{B}\times\mathcal{A}} V(b)\left|W_q^{(n)}(a|b) - \bar{W}(a|b)\right| \le \frac{|\mathcal{B}||\mathcal{A}|^2}{n}. \tag{C.13}$$

Proof. Define $m(b) = \lceil nV(b) \rceil$. By definition of $\mathbb{P}_n(\mathcal{A}, V)$, if $m(b) = 0$, $W_q^{(n)}(\cdot|b)$ can be chosen arbitrarily. For every $(b) \in \mathcal{B}$ such that $V(b) > 0$, choose $a_{\max}(b)$ such that $\bar{W}(a_{\max}(b)|b) \ge \bar{W}(a|b), \forall a \in \mathcal{A}$. For all $a \ne a_{\max}(b)$ let $W_q^{(n)}(a|b) = \frac{\lfloor m(b)\bar{W}(a|b)\rfloor}{m(b)}$, and take $W_q^{(n)}(a_{\max}(b)|b) = 1 - \sum_{a\in\mathcal{A}, a\ne a_{\max}(b)} W_q^{(n)}(a|b)$. This construction assures us that $|W_q^{(n)}(a|b) - \bar{W}(a|b)| \le \frac{|\mathcal{A}|-1}{m(b)}$, thus,

$$\begin{aligned} ||V \times W_q^{(n)} - V \times \bar{W}|| &\le \sum_{(b,a)\in\mathcal{B}\times\mathcal{A}} V(b)\frac{1}{m(b)} \le \sum_{(b,a)\in\mathcal{B}\times\mathcal{A}} \frac{|\mathcal{A}|-1}{n} \\ &= \frac{|\mathcal{B}||\mathcal{A}|(|\mathcal{A}|-1)}{n}. \end{aligned} \tag{C.14}$$

□

Claim 7. *If P and Q are two distributions on alphabet $\mathcal{B} \triangleq \mathcal{U} \times \mathcal{X} \times \mathcal{Y}$ such that,*

$$\|P - Q\| = \Delta \leq \frac{1}{2}, \tag{C.15}$$

then,

$$\begin{aligned}
|E_P d_1(U,X) - E_Q d_1(U,X)| &\leq D_1^{\max}\Delta &\text{(C.16)}\\
|E_P d_2(X;Y) - E_Q d_2(X;Y)| &\leq D_2^{\max}\Delta &\text{(C.17)}\\
|I_P(X;Y|U) - I_Q(X;Y|U)| &\leq -4\Delta\log\left(\frac{\Delta}{|\mathcal{B}|}\right) &\text{(C.18)}\\
|I_P(U;Y|X) - I_Q(U;Y|X)| &\leq -4\Delta\log\left(\frac{\Delta}{|\mathcal{B}|}\right), &\text{(C.19)}
\end{aligned}$$

where $D_1^{\max} = \max_{(u,x)\in\mathcal{U}\times\mathcal{X}} d_2(u,x)$, $D_2^{\max} = \max_{(x,y)\in\mathcal{X}\times\mathcal{Y}} d_2(x,y)$.
Also, for every given pmf $P^* \in \mathcal{P}(\mathcal{B})$ which satisfies $P^*(b_{\min}) \triangleq \min_{b\in\mathcal{B}} P^*(b) > 0$,

$$|D(P\|P^*) - D(Q\|P^*)| \leq -\Delta\log\left(\frac{\Delta P^*(b_{\min})}{|\mathcal{B}|}\right). \tag{C.20}$$

Proof. Inequality (C.16) follows since,

$$|E_P d_1(U,X) - E_Q d_1(U,X)| \leq \sum_{x,y} |P(x,y) - Q(x,y)|\, d_1(u,x), \tag{C.21}$$

and inequality (C.17) follows similarly.

Inequalities (C.18) and (C.19) follow as a straightforward extension of Lemma 2.7 in [5], which implies that if (C.15) holds, then:

$$|H_P(U,X,Y) - H_Q(U,X,Y)| \leq -\Delta\log\left(\frac{\Delta}{|\mathcal{B}|}\right), \tag{C.22}$$

and since $I(X,Y|U) = H(X,U) - H(U) - H(X,Y,U) + H(U,Y)$, thus,

$$\begin{aligned}
&|I_P(X;Y|U) - I_Q(X;Y|U)|\\
\leq\ &|H_P(X,U) - H_Q(X,U) - H_P(U) + H_Q(U) - H_P(U,X,Y)\\
&\quad + H_Q(U,X,Y) + H_P(U,Y) - H_Q(U,Y)|\\
\leq\ &4|H_P(U,X,Y) - H_Q(U,X,Y)|, \qquad\qquad \text{(C.23)}
\end{aligned}$$

and a similarly, $|I_P(U;Y|X) - I_Q(U;Y|X)| \leq 4|H_P(U,X,Y) - H_Q(U,X,Y)|$.

Finally, eq. (C.20) follows since, for every $P^* \in \mathcal{P}(\mathcal{B})$ which satisfies $P^*(b_{\min}) > 0$,

$$\begin{aligned}
&|D(P\|P^*) - D(Q\|P^*)|\\
\leq\ &|H(P) - H(Q)| + |E_P(\log P^*) - E_Q(\log P^*)|\\
\leq\ &-\Delta\log\left(\frac{\Delta}{|\mathcal{B}|}\right) - \Delta\log\left(P^*(b_{\min})\right), \qquad \text{(C.24)}
\end{aligned}$$

where $E_P(\log P^*) \triangleq \sum_{b\in\mathcal{B}} P(b)\log P^*(b)$. □

References

[1] A. Cohen and A. Lapidoth, "On the Gaussian watermarking game," in *Proc. ISIT 2000*, p. 48, 2000.

[2] A. Cohen and A. Lapidoth, "The capacity of the vector Gaussian watermarking game," preprint, 2000.

[3] I. J. Cox, M. L. Miller, and A. L. McKellips, "Watermarking as communications with side information," *Proc. IEEE*, vol. 87, no. 7, pp. 1127–1141, July 1999.

[4] S. Craver, N. Memon, B.-L. Yeo, and M. M. Yeung, "Resolving rightful ownerships with invisible watermarking techniques: limitations, attacks, and implications," *IEEE J. Select. Areas in Commun.,* vol. 16, no. 4, pp. 573–586, May 1998.

[5] I. Csiszár and J. Körner, *Coding Theorems for Discrete Memoryless Systems.* New York: Academic, 1981.

[6] A. G. D'yachkov, "Lower bound for ensemble-average error probability for a discrete memoryless channel," *Problems of Information Transmission (IPPI)*, pp. 93-98, April-June, 1980.

[7] A. G. D'yachkov, "Bounds on the average error probability for code ensemble with fixed composition," *Problems of Information Transmission (IPPI)*, pp. 255-259, July 1981.

[8] R. G. Gallager, *Information Theory and Reliable Communication.* New York: Wiley, 1968.

[9] R. G. Gallager, "The random coding bound is tight for the average code," *IEEE Trans. Inform. Theory*, vol. IT-19, no. 2, pp. 244-246, March 1973.

[10] S. Karlin, *Mathematical Methods and Theory in Game, Programming and Economics,* Addison-Wesley, Reading, Mass., 1959.

[11] A. Lapidoth and N. Narayan, "Reliable communication under channel uncerainty," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2148-2177, October 1998.

[12] N. Merhav, N. Merhav, "How many information bits does a decoder need about the channel statistics?" *IEEE Trans. Inform. Theory*, vol. 43, no. 5, pp. 1707-1714, September 1997.

[13] N. Merhav, "On random coding error exponents of watermarking systems," *IEEE Trans. Inform. Theory*, vol. 46, no. 2, pp. 420-430, March 2000.

[14] N. Merhav, "Universal detection of messages via finite-state channels," *IEEE Trans. Inform. Theory*, vol. 46, no. 6, pp. 2242-2246, September 2000.

[15] P. Moulin and J. A. O'Sullivan, "Information-theoretic analysis of information hiding," preprint, September 1999. Available at `www.ifp.uiuc.edu/~moulin/Papers/hiding.ps`

[16] P. Moulin and J. A. O'Sullivan, "Information-theoretic analysis of information hiding," *Proc. ISIT 20000*, p. 19, 2000.

[17] J. A. O'Sullivan, P. Moulin, and J. M. Ettinger, "Information theoretic analysis of steganography," *Proc. ISIT '98*, p. 297, 1998.

[18] F. A. P. Petitculas, R. J. Anderson, and M. G. Kuhn, "Information hiding - a survey," *Proc. IEEE*, vol. 87 no. 7, pp. 1062-1078, July 1999.

[19] Y. Steinberg and N. Merhav, "Identification in the presence of side information with application to watermarking," to appear in *IEEE Trans. Inform. Theory*, May 2001.

[20] I. G. Stiglitz, "Coding for a class of unknown channels," *IEEE Trans. Inform. Theory*, vol. IT–13, no. 2, pp. 189-195, April 1966.

[21] I. G. Stiglitz, "A coding theorem for a class of unknown channels," *IEEE Trans. Inform. Theory*, vol. IT–13, no. 2, pp. 217–220, April 1967.

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

[23] J. Ziv, "Universal decoding for finite-state channels," *IEEE Trans. Inform. Theory*, vol. 33, pp. 453-460, July 1985.