# On the Capacity Game of Private Fingerprinting Systems under Collusion Attacks


Anelia Somekh-Baruch[1] and Neri Merhav
Department of Electrical Engineering
Technion - I.I.T., Haifa 32000, Israel.
{anelia@tx, merhav@ee}.technion.ac.il



**Abstract**

The problem of fingerprinting in the presence of *collusive* attacks is considered. It is modelled as a game between a fingerprinter and a decoder, on the one hand, and a coalition of two or more attackers, on the other hand. The fingerprinter distributes, to different users, different fingerprinted copies of a host data (*covertext*) embedded with different fingerprints. The coalition members create a forgery of the data while aiming at erasing the fingerprints in order not to be detected. Their action is modelled by a multiple access channel (MAC). The decoder, who has access to the original covertext data, observes the forgery and decodes one of the messages in order to identify one of the members of the coalition. Motivated by a worst-case approach, we assume that the coalition of attackers is informed of the hiding strategy taken by the fingerprinter and the decoder, while they are uninformed of the attacking scheme. A single-letter expression for the capacity is derived under the assumption that the host data is drawn from a memoryless stationary source and some mild assumptions on the operation of the encoder. Also, a lower bound on the error exponent is derived as a by-product of the achievability part, and asymptotically optimum strategies of the parties involved are characterized.



**Index Terms**- Capacity, coding with side information, error exponents, information hiding, MMI decoder, private watermarking, randomized code, steganography, universal decoding, watermarking.


## 1 Introduction

Watermarking systems have several applications such as copyright protection, data authentication, steganography, data encryption and fingerprinting. In fingerprinting applications, the embedded information is used to identify one of many users. In other words, the same covertext is distributed to different users with different fingerprints (that designate, e.g., the different digital signatures or serial numbers of the copies they are provided with). In this situation, collusive attacks on the fingerprinting system are possible and have to be taken into account in the code design. In a collusion attack, two or more users form a coalition,

[1]This work is part of the Ph.D. research of the first author.

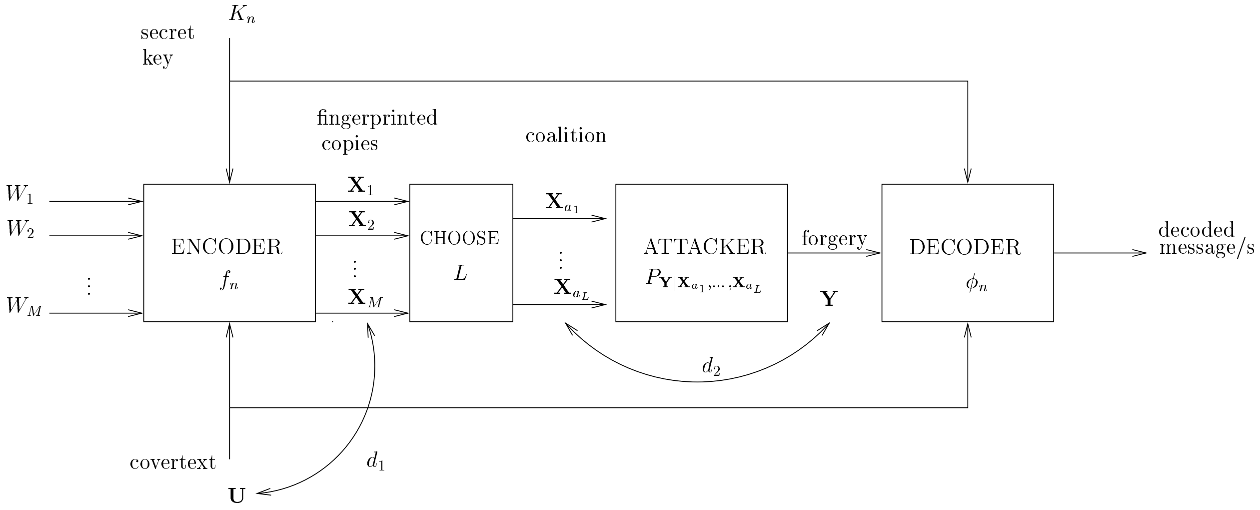


Figure 1: center

Figure 1. The private fingerprinting game.

and contribute their distinct fingerprinted copies in order to create a better forgery. The distortion between the forgery and the original data should not exceed a certain level. Two models of the game may be considered: the private game in which the covertext is available also at the decoder's side, and the public game, where it is only available to the encoder. In this paper, we consider the private game where the decoder is facing the rather complicated goal of reliably identifying coalition members based on the forgery and the covertext.

The problem of designing a private fingerprinting system in the presence of colluding attackers can be modelled as a game between an encoder-decoder and a few attackers. A realistic worst-case approach, taken in this paper, is based on the assumption that the attackers are informed of the covertext distribution and the coding-decoding strategy (up to a random secret key), whereas the encoder and decoder are not informed of the attack strategy. Thus, the encoder and decoder are assumed to adopt a random coding strategy as a means of protection against malicious attackers. Random coding is enabled by a secret key that specifies the particular codebook that has been drawn, which is shared by the encoder and decoder. The action of the attackers is allowed to be stochastic, and therefore one can model the attack by a MAC, whose output is the forgery and whose inputs are fingerprinted copies that are observed by the coalition members. This model of the game is depicted in Figure 1.

The fingerprinting game is closely related to (and is actually an extension of) the watermarking game, that in turn can be modelled as a coded communication system equipped with side information, for a single user as opposed to one of many users. Watermarking systems have been studied from the information theoretic point of view in several works (see e.g., [6], [7], [17], [14], [18], [23], [24] and [25]). In [23], the capacity and error exponents of private watermarking games are investigated. It is assumed that the attack channel is unknown to the encoder and decoder. Under this assumption, the capacity of the private game with expected distortion constraints is zero, and thus almost-sure and large deviations distortion constraints are investigated in [23] and single-letter expressions for the error exponent and capacity of the private game are obtained.

Several researchers (see, e.g., [1], [2], [3], [4], [5], [10], [12], [13], [19], [20], [21], [26], [27], and references therein) have proposed and analyzed fingerprinting systems that aimed at protection against collusion attacks under various conditions. The research in this problem area has largely focused on combinatorial analysis and algorithmic issues, whereas there has been much less work on information–theoretic aspects. As an exception to the last statement, we mention several papers, such as [15] and [26] where the Gaussian–additive MAC has been used to model the attack channel, and where the forgery is the sum (of possibly filtered versions) of the copies distributed to the members of the coalition plus independent, additive Gaussian noise. Another exception is [17] where a variant of the public fingerprinting game is considered. The attacker in [17] creates $L$ forgeries corresponding to the $L$ fingerprinted copies he observes. Each copy is fed into a single-user detector, and the capacity is defined w.r.t. the probability of error averaged over the miss detection probabilities of the $L$ messages. In [2] (see also references therein) another setting of the fingerprinting problem is stated and analyzed, where the members of the coalition are assigned fingerprints and produce a forgery of the fingerprints when they can only alter those coordinates of the fingerprint in which at least two of the fingerprints differ. This is opposed to the setting discussed in this paper where we require that the distortion between the forgery and at least one of the fingerprinted copies of the covertext is smaller than some threshold.

The goal of this paper is to extend the results of [23] to the setting of the private fingerprinting game, and thereby allow for more general encoding strategies and attack strategies than in [15] and [26]. We address the problem of the maximin game between the fingerprinter and the MAC applied by the users. We consider two types of decoders: a *single-output (SO) decoder* whose single output is a message index and a *multiple-output (MO) decoder* whose output is a list containing $L$ message indices (where $L$ is the size of the coalition). The two decoders aim at detecting *only one* member of the coalition. When the SO decoder is concerned, an error is declared if its output does not belong to the coalition, whereas when the MO decoder is concerned, an error is declared when none of its outputs belongs to the coalition. The reason for aiming at this goal (adopted also in, e.g., [2],[21] and references therein) is that it can be shown that the target of detecting the entire coalition is impossible since the attacker can decide to almost ignore one of its inputs, and thus prevent reliable decoding of the input that was ignored.

It should be pointed out that the model of the MAC in the present scenario differs (and hence, is harder to analyze) in many ways from the one in the classical setting:

1. It is not necessarily memoryless as the attacker can operate off-line.

2. It is an unknown channel except for meeting the distortion constraints.

3. Universal decoding does not necessarily exist.

4. There is an inherent constraint of identical codebooks for the two (or more) users.

5. The decoder aims at detecting only one member of the coalition. This dictates an unconventional error criterion, where an error is declared only when none of the coalition member is detected.

6. The covertext is available at the transmitter and the receiver and serves as side information.

In this paper, it is assumed that the encoder uses constant composition (CC) codes and under this assumption, a single-letter expression for the capacity of the private fingerprinting game with respect to (w.r.t.) the two types of decoders is found, and it is shown that their capacities are the same. Asymptotically optimal strategies, taken by the adversaries, are characterized. Also, lower bounds on the error exponents of the two types of decoders are derived.

The outline of the paper is as follows: In Section 2, some notation conventions are defined. A 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 capacity game and the random coding error exponent, and Section 5 is devoted to a discussion of the results. Finally, the proofs of the theorems appear in Section 6.

# 2 Notation and Definitions

Henceforth, we adopt the following notation conventions. Random variables (RV's) 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 the corresponding 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 a 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., $\lim_{n \to \infty} \frac{1}{n} \log \frac{c_n}{d_n} = 0$. Similarly, $c_n \stackrel{\cdot}{\geq} d_n$ will stand for $\liminf_{n \to \infty} \frac{1}{n} \log \frac{c_n}{d_n} \geq 0$, and so on.

The empirical pmf, induced by a vector $\mathbf{x} \in \mathcal{X}^n$, is the vector $\hat{P}_{\mathbf{x}} = \left\{ \hat{P}_{\mathbf{x}}(a), \; a \in \mathcal{X} \right\}$, where $\hat{P}_{\mathbf{x}}(a)$ is the relative frequency of the letter $a$ in the vector $\mathbf{x}$. The set of empirical 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 $T_{\mathbf{x}}$ (or $T(\hat{P}_{\mathbf{x}})$) 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 $\hat{P}_{\mathbf{xy}} = \left\{ \hat{P}_{\mathbf{xy}}(a,b), \; a \in \mathcal{X}^n, b \in \mathcal{Y}^n \right\}$, where $\hat{P}_{\mathbf{xy}}(a,b)$ is the relative frequency of $(x_i,y_i)=(a,b)$. The type class $T_{\mathbf{xy}}$ (or $T(\hat{P}_{\mathbf{xy}})$) 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 $T_{\mathbf{y}|\mathbf{x}}$ (or $T_{\mathbf{x}}(\hat{P}_{\mathbf{y}|\mathbf{x}})$), 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,

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

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 $H(P)$ interchangeably. 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.,

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

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

We shall have a particular interest in strongly exchangeable channels. A *strongly exchangeable channel* is defined as a conditional distribution $P_{\mathbf{Y}|\mathbf{X}}$ with input alphabet $\mathcal{X}^n$ and output alphabet $\mathcal{Y}^n$ for which, for every $\mathbf{x}\in\mathcal{X}^n$, $\mathbf{y}\in\mathcal{Y}^n$ and every permutation $\pi$ of $\{1,\ldots,n\}$,

$$P_{\mathbf{X}|\mathbf{Y}}(\mathbf{y}|\mathbf{x})=P_{\mathbf{X}|\mathbf{Y}}(\pi\mathbf{y}|\pi\mathbf{x}). \tag{6}$$

Obviously, every DMC is strongly exchangeable.

# 3 Statement of the Problem

Let $\mathcal{U},\mathcal{X}$ and $\mathcal{Y}$ be finite sets designating the alphabets of a covertext symbol, fingerprinted symbol and forgery symbol, respectively. Let $\mathbf{U}$ designate the covertext source within which the fingerprints 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 single-letter marginal pmf $P_U$, and we shall assume that $\min_{u\in\mathcal{U}}P_U(u)>0$. The fingerprinter creates, at random, $M=\lfloor 2^{nR}\rfloor$ fingerprinted versions of the covertext, denoted $\mathbf{X}_i$, $i = 1, \ldots, M$. The resulting collection of the fingerprinted copies, $\mathbf{X}_i$, $i = 1, \ldots, M$, will be referred to as a codebook.

A secret key, $K_n$, is an RV, independent of the fingerprints and the covertext, known to both the encoder and decoder, but unknown to the attacker. Let $\mathcal{K}_n$ stand for the alphabet of $K_n$, and let $P_{K_n}$ stand for the distribution of $K_n$.

**Definition 1.** *A rate-R fingerprinting encoder of block-length n is a function which maps the secret key realization $k_n$, the covertext data $\mathbf{u}$, and the watermark message $m \in \mathcal{M}_n$ into a stegotext vector (or, fingerprinted vector) $\mathbf{x}$, i.e.,*

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

*where*

$$\mathcal{M}_n = \{1, \ldots, M\}. \tag{8}$$

Having created $\mathbf{X}_i$, $i = 1, \ldots, M$, the encoder distributes them arbitrarily to $M$ different users. We assume that the distributed copies should meet an almost-sure distortion constraint w.r.t. a given distortion level $D_1$, i.e.,

$$\Pr\{d_1(\mathbf{U}, \mathbf{X}_i) > nD_1\} = 0 \quad \forall i \in \mathcal{M}_n, \tag{9}$$

where $d_1 : \mathcal{U} \times \mathcal{X} \to \mathbb{R}_+$ denotes a single-letter distortion measure and $d_1(\mathbf{u}, \mathbf{x}) = \sum_{j=1}^n d_1(u_j, x_j)$ for $\mathbf{u} \in \mathcal{U}^n$ and $\mathbf{x} \in \mathcal{X}^n$.

Let $W_a$ and $W_b$ be the two[2] users that take part in the coalition. It is assumed that $W_a, W_b$ are independent and both uniformly distributed over the set $\mathcal{M}_n$. Let $\mathbf{X}_a$ and $\mathbf{X}_b$ stand for the two corresponding sequences of fingerprinted data received by the users $W_a$ and $W_b$. Namely, $\mathbf{X}_a$ and $\mathbf{X}_b$ are available to the attacker who creates the forgery, $\mathbf{Y}$, as the output of an attack channel $P_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}$ from $\mathcal{X}^n \times \mathcal{X}^n$ to $\mathcal{Y}^n$.

Let $d_2 : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ denote another single-letter distortion measure, 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 $\mathbf{y} \in \mathcal{Y}^n$. The attacker is assumed to meet the following distortion constraint

$$\Pr\{d_2(\mathbf{X}_a, \mathbf{Y}) \le nD_2 \text{ or } d_2(\mathbf{X}_b, \mathbf{Y}) \le nD_2\} \;=\; 1. \tag{10}$$

Denote by $\mathcal{P}_n^{d_2}$ the set of channels satisfying (10).

The decoder observes the realization of the secret key (and thus, knows the particular codebook that has been drawn), which is unavailable to the attacker, the covertext $\mathbf{U}$ and the forgery $\mathbf{Y}$ and its aim is to detect one member of the coalition. As mentioned earlier, we consider two types of decoders: A single-output (SO) decoder, which is a function

$$\phi_n : \mathcal{K}_n \times \mathcal{U}^n \times \mathcal{Y}^n \to \mathcal{M}_n, \tag{11}$$

[2]In the general fingerprinting game, the coalition may have more than two members. However, for the sake of simplicity, we focus on the case of two coalition members. In Section 7, we demonstrate how the results can be extended to a general coalition size.

and a multiple-output (MO) decoder is a function

$$\psi_n:\ \mathcal{K}_n \times \mathcal{U}^n \times \mathcal{Y}^n \to \mathcal{M}_n^2. \tag{12}$$

The quadruple $F_n = (P_{K_n}, f_n, \mathcal{M}_n, \phi_n)$ ($F_n = (P_{K_n}, f_n, \mathcal{M}_n, \psi_n)$) will be referred to as *a rate $R$ randomized fingerprinting SO (MO) code.*
We shall focus on the following subclass of fingerprinting codes:

**Definition 2.** *A rate-$R$ constant composition (CC) fingerprinting code of block-length $n$ is a code with the following structure of a secret key and encoder: the fingerprinted copies $\mathbf{X}_i$, $i \in \mathcal{M}_n$ are drawn independently given $\mathbf{U}$, uniformly over a single conditional type class $T_{\mathbf{x}|\mathbf{u}}$ for all $\mathbf{u}$. The choice of the conditional type class is defined by a mapping*

$$\eta_n:\ \mathcal{U}^n \to \mathcal{P}(\mathcal{X}|\mathcal{U}),\ s.t.\ \eta_n(\mathbf{u}) \in \mathbb{P}_n(\mathcal{X}, \hat{P}_{\mathbf{u}}), \tag{13}$$

*i.e., it is given by $T_{\mathbf{u}}(\eta_n(\mathbf{u}))$.*

In order to consider conditional types such that constraint (9) is met, we consider the set $\mathcal{N}_n(D_1)$ of functions $\eta_n:\ \mathcal{U}^n \to \mathcal{P}(\mathcal{X}|\mathcal{U})$ such that (13) holds and

$$E_{\hat{P}_{\mathbf{u}}\eta_n(\mathbf{u})} d_1(U, X) \le D_1. \tag{14}$$

Denote by $\mathcal{F}_n^{d_1,SO}(R)$ and $\mathcal{F}_n^{d_1,MO}(R)$ the set of rate-$R$ randomized CC fingerprinting SO and MO codes satisfying (9), respectively. Denote

$$\ell(n) = \sqrt{(2\ln 2)(|\mathcal{U}|+1)\frac{\log(n+1)}{n}}. \tag{15}$$

We require that the constant composition code will have a certain "smoothness" property which guarantees that provided that the observed covertext sequence $\mathbf{u}$ is typical, small fluctuations in its empirical statistics do not alter the encoder's strategy dramatically.

**Definition 3.** *A smooth sequence of rate-$R$ constant composition (CC) fingerprinting code of block-length $n$ is one that abides by the following "smoothness" requirement: There exists a conditional pmf $P_{X|U} \in \mathcal{P}(\mathcal{X}|\mathcal{U})$ such that for all $\epsilon > 0$, $\exists n_0(\epsilon)$ such that for all $n \ge n_0(\epsilon)$, if $\|\hat{P}_{\mathbf{u}} - P_U\| \le \ell(n)$ then*

$$\|P_U \times P_{X|U} - \hat{P}_{\mathbf{u}} \times \eta_n(\mathbf{u})\| \le \epsilon. \tag{16}$$

This smoothness requirement is obviously consistent with the distortion constraint (9).

The fact that we focus on the wide class of CC fingerprinting encoders can be justified by practical considerations. Any practical randomized encoder should have some enumeration mechanism, where one first randomly selects a number under the uniform distribution in some range (in particular, an integer according to the key), and then this number is mapped to a codeword (given $\mathbf{U}$). It is desired then that to implement this mapping, one should not need (exponentially) large tables but can use a simple function. It is well known that there are indeed simple ways to enumerate sequences which belong to the same type class. See, for example, [8], where such an enumeration method is proposed, and the same idea can be easily extended to conditional type classes.

For a given realization of a secret key, $K_n$, let the output of the SO decoder be given by $\hat{W} = \phi_n(K_n, \mathbf{U}, \mathbf{Y})$, and let $(\hat{\nu}, \hat{\mu}) = \psi(K_n, \mathbf{U}, \mathbf{Y}) = (\psi_\nu(K_n, \mathbf{U}, \mathbf{Y}), \psi_\mu(K_n, \mathbf{U}, \mathbf{Y}))$ be the output of the MO decoder.

An error occurs if no coalition member is identified by the decoder (either SO or MO), and therefore, the average probabilities of error of the SO decoder and the MO decoder are defined as

$$\begin{aligned} P_e\left(F_n, P_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}\right) &\triangleq \Pr(\hat{W} \neq W_a \text{ and } \hat{W} \neq W_b) \\ P_e\left(F_n, P_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}\right) &\triangleq \Pr(\hat{\nu} \neq W_a \text{ and } \hat{\nu} \neq W_b \text{ and } \hat{\mu} \neq W_a \text{ and } \hat{\mu} \neq W_b), \end{aligned} \tag{17}$$

respectively, where the average is over the covertext, the members of the coalition, the ensemble of all possible codebooks, and the action of the attack channel $P_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}$, when the randomized code $F_n$ is employed.

**Definition 4.** *An achievable rate $R$ w.r.t. a SO decoder and CC codes is one for which there exists a smooth sequence $F_n \in \mathcal{F}_n^{d_1,SO}(R)$, $n \geq 1$ such that* $\lim_{n\to\infty} \sup_{P_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b} \in \mathcal{P}_n(D_2)} P_e\left(F_n, P_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}\right) = 0.$

**Definition 5.** *The CC capacity of the private fingerprinting game w.r.t. a SO decoder is defined as the supremum of all achievable rates.*

Similar definitions apply for the achievable rate and the CC capacity w.r.t. a MO decoder. In fact, the capacity is a function of the distortion levels $D_1, D_2$ and the covertext symbol distribution $P_U$.

The maxmin error exponent of the fingerprinting game w.r.t. a SO decoder is defined by

$$\begin{aligned} &e^{SO}(P_U, D_1, D_2, R) \triangleq \\ &\liminf_{n\to\infty} \left[ -\frac{1}{n} \log \min_{F_n \in \mathcal{F}_n^{d_1,SO}(R)} \max_{P_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b} \in \mathcal{P}_n^{d_2}} P_e(F_n, P_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}) \right]. \end{aligned} \tag{18}$$

The maxmin error exponent of the fingerprinting game w.r.t. a MO decoder $e^{MO}(P_U, D_1, D_2, R)$ is defined similarly.

Our goal, in this paper, is to provide single-letter expressions for the capacities of the fingerprinting game w.r.t. the two types of decoders and CC codes, and to characterize the asymptotically optimal encoder and decoders and a worst case attack channel. Moreover, the analysis of the average probability of error in the direct part enables us to obtain lower bounds on the error exponents of the two decoders.

# 4 Main Results

For given pmfs $P_{\tilde{U}}$ and $V$ defined on the sets $\mathcal{U}$ and $\mathcal{X}^2$ respectively, define the sets of conditional pmfs

$$\mathcal{D}_{d_1}(D_1, P_{\tilde{U}}) \triangleq \{P_{X|\tilde{U}} : E_{P_{\tilde{U}}P_{X|\tilde{U}}} d_1(\tilde{U}, X) \leq D_1\} \tag{19}$$

$$\mathcal{D}_{d_2}(D_2, V) \triangleq \left\{P_{Y|\tilde{X}X} : \min\{E_{VP_{Y|\tilde{X}X}} d_2(X, Y), E_{VP_{Y|\tilde{X}X}} d_2(\tilde{X}, Y)\} \leq D_2\right\} \tag{20}$$

The main result is the following coding theorem.

**Theorem 1.** *The CC capacity[3] of the private fingerprinting game w.r.t. a SO decoder coincides with that of a MO decoder and is given by*

$$C(P_U, D_1, D_2) \triangleq \max_{P_{X|U} \in \mathcal{D}_{d_1}(D_1, P_U)} \min_{P_{Y|X\tilde{X}} \in \mathcal{D}_{d_2}(D_2, P_{X\tilde{X}})} \frac{1}{2} I(X\tilde{X}; Y|U), \tag{21}$$

*where $U \sim P_U$, $\tilde{X}$ is an RV which is independent of $X$ given $U$ and satisfies $P_{\tilde{X}|U} = P_{X|U}$, and $U \leftrightarrow (X, \tilde{X}) \leftrightarrow Y$.*

The proof of the theorem consists of a direct part and a converse part. In the direct part, an upper bound on the probability of error is established using the method of types. This bound decays exponentially with $n$. Therefore, the direct part yields also a lower bound on the error exponent which is stated explicitly in the next two propositions. Proposition 1 refers to a SO decoder and Proposition 2 refers to a MO decoder, and their proofs appear in Sections 6.1 and 6.2, respectively. Proposition 1 can serve as a direct part for the MO decoder as well, because the MO decoder has two outputs while the SO decoder has only one output, so the capacity of a MO decoder cannot be lower than that of a SO decoder. In spite of this fact, we provide a separate proof of the direct theorem for the MO decoder for two reasons: (a) to characterize the asymptotically optimal MO decoder and (b) to establish a tighter expression for the error exponent. Although the converse part for the MO decoder could have been sufficient for proving Theorem 1 we also provide a separate proof for the converse parts w.r.t. the SO decoder in order to characterize the asymptotically worst-case attacker w.r.t. the SO decoder. The proofs of the converse parts w.r.t. the SO and MO decoders can be found in Sections 6.3 and 6.4, respectively.

For a given pmf $P_{\tilde{U}} \in \mathcal{P}(\mathcal{U})$ and a general real-valued $k$, define the following quantity

$$\begin{aligned} & \tilde{\xi}_k(P_{\tilde{U}}, D_1, D_2, R) \\ \triangleq\ & \max_{P_{X|\tilde{U}}} \min_{P_{\tilde{X}Y|\tilde{U}X}} \left[ I(X; \tilde{X}|\tilde{U}) + I(\tilde{U}; Y|X\tilde{X}) + k \cdot \varepsilon(R, P_{\tilde{U}X\tilde{X}Y}) \right], \end{aligned} \tag{22}$$

with

$$\varepsilon(R, P_{\tilde{U}X\tilde{X}Y}) \triangleq \left| \frac{1}{2} I(X; Y|\tilde{U}) + \frac{1}{2} I(X; \tilde{X}Y|\tilde{U}) - R \right|^+, \tag{23}$$

$\tilde{U} \sim P_{\tilde{U}}$, $|t|^+ = \max\{0, t\}$, the maximization is w.r.t. the set $\mathcal{D}_{d_1}(D_1, P_{\tilde{U}})$, and the minimization is over the set $\{P_{Y\tilde{X}|\tilde{U}X} :\ P_{\tilde{X}|\tilde{U}} = P_{X|\tilde{U}},\ \min\{Ed_2(X, Y), Ed_2(\tilde{X}, Y)\} \leq D_2\}$.

**Proposition 1.** *The error exponent of the private fingerprinting game w.r.t. a SO decoder satisfies*

$$e^{SO}(P_U, D_1, D_2, R) \ \geq\ \xi_1(P_U, D_1, D_2, R) \tag{24}$$

[3]Since the capacity w.r.t. general codes, i.e., codes that are not necessarily CC codes, is not lower than the capacity w.r.t. CC codes, this theorem provides also a lower bound on the capacity of the more general codes.

*where $\xi_k(P_U, D_1, D_2, R)$ for a general real-valued $k$ is defined as*

$$\xi_k(P_U, D_1, D_2, R) = \min_{P_{\tilde{U}}} \left\{ D(P_{\tilde{U}} \| P_U) + \tilde{\xi}_k(P_{\tilde{U}}, D_1, D_2, R) \right\}. \tag{25}$$

The smallest rate for which $\xi_1(P_U, D_1, D_2, R) = 0$ is obviously a lower bound on the capacity. It is easy to verify that this rate is indeed $C(P_U, D_1, D_2)$: For $\xi_1(P_U, D_1, D_2, R)$ to be zero, $D(P_{\tilde{U}}||P_U)$ must be zero which means that $P_{\tilde{U}} = P_U$. Moreover, $I(X; \tilde{X}|U)$ and $I(U; Y|X\tilde{X})$ are non-negative by definition, so for $\xi_1(P_U, D_1, D_2, R)$ to be zero, $P_{\tilde{X}Y|UX}$ must be such that $X \leftrightarrow U \leftrightarrow \tilde{X}$ and $U \leftrightarrow X\tilde{X} \leftrightarrow Y$ form Markov chains. Also, $I(X; Y|U) + I(X; \tilde{X}Y|U) = I(X\tilde{X}; Y|U)$. Therefore the smallest $R$ that solves the equation

$$\max_{P_{X|\tilde{U}} \in \mathcal{D}_{d_1}(D_1, P_{\tilde{U}})} \quad \min_{P_{Y|\tilde{X}X} \in \mathcal{D}_{d_2}(D_2, P_{X\tilde{X}})} \left| \frac{1}{2} I_{P_U P_{X|U} P_{\tilde{X}|\tilde{U}} P_{Y|\tilde{X}X}}(X\tilde{X}; Y|U) - R \right|^+ = 0, \tag{26}$$

with $P_{\tilde{X}|\tilde{U}} = P_{X|\tilde{U}}$, is exactly $C(P_U, D_1, D_2)$.

**Proposition 2.** *The error exponent of the private fingerprinting game w.r.t. a MO decoder satisfies*

$$e^{MO}(P_U, D_1, D_2, R) \quad \geq \quad \xi_2(P_U, D_1, D_2, R) \tag{27}$$

*where $\xi_2(P_U, D_1, D_2, R)$ is defined as in (25) for $k = 2$.*

The term $\tilde{\xi}_k(P_{\tilde{U}}, D_1, D_2, R)$ for $P_{\tilde{U}} \in \mathbb{P}_n(\mathcal{U})$, which appears in $\xi_k(P_U, D_1, D_2, R)$, can be interpreted as the error exponent conditioned on the event that $\mathbf{U}$ lies within the type-class corresponding to $P_{\tilde{U}}$, and the term $D(P_{\tilde{U}}||P_U)$ is a result of the averaging over the types.

# 5 Discussion

A few important comments are in order:

As mentioned in the introduction, the setting of the fingerprinting problem differs from that of conventional coded communication over the classical MAC in many aspects. The ordinary capacity region of the MAC is defined w.r.t. the probability of error in joint decoding of the two messages, whereas in our case, we content ourselves with correct decoding of only one of them (see the discussion in the introduction). As a result, there is no need for introducing an auxiliary RV into the scheme, as opposed to the ordinary MAC (see, e.g. [11]). In spite of these differences, the capacity $C(P_U, D_1, D_2)$ bears some resemblance to the rate-sum bound of the MAC. The capacity region of the MAC given input distributions $P_1(X), P_2(\tilde{X})$ is given by the set of rate pairs $(R_1, R_2)$, satisfying

$$0 \leq R_1 \leq I(X; Y|\tilde{X}) \; ; \; 0 \leq R_2 \leq I(\tilde{X}; Y|X) \; ; \; 0 \leq R_1 + R_2 \leq I(X, \tilde{X}; Y). \tag{28}$$

In the case of two users who (use the same codebook and hence) have the same rate, $R_1 = R_2 = R$, the corresponding upper limit is $R \leq \min\{I(X; Y|\tilde{X}), I(\tilde{X}; Y|X), \frac{1}{2} I(X, \tilde{X}; Y)\}$, the minimizer depending on which of the above three lines is crossed first by the 45-degree line $R_1 = R_2$. The parallel capacity expression we get is given by $\frac{1}{2}I(X\tilde{X};Y|U)$, which is obviously bigger than $\min\{I(X;Y|U\tilde{X}), I(\tilde{X};Y|UX), \frac{1}{2}I(X\tilde{X};Y|U)\}$, and this is primarily due to the fact that we account for correct decoding of only one of the fingerprints, whereas in the MAC a joint decoding of the two messages is performed.

Had we been interested in the following distortion constraint,

$$\Pr\left\{d_2(\mathbf{X}_a,\mathbf{Y}) \leq nD_2 \text{ and } d_2(\mathbf{X}_b,\mathbf{Y}) \leq nD_2\right\} \;=\; 1, \tag{29}$$

it would no longer have been an impossible mission to detect the members of the entire coalition at the decoder's side. Therefore, in this case, the setup is more similar to the ordinary MAC, and as a result, the expression for the capacity and the error exponent (see [11]) would resemble that of the classical MAC.

An asymptotically optimal MO decoder that is informed of $\mathbf{u}$ and the codebook, observes the forgery $\mathbf{y}$ and operates as follows:

$$(\hat{w}_1,\hat{w}_2) \;=\; \mathrm{argmin}_{w_1,\, w_2\neq w_1}|T_{\mathbf{x}_{w_1}\mathbf{x}_{w_2}|\mathbf{uy}}|. \tag{30}$$

The asymptotically optimal SO decoder operates in two stages. In the first stage, it chooses $(\hat{w}_1,\hat{w}_2)$ using decision rule (30), and then the eventual decision is given by

$$\hat{w} \;=\; \begin{cases} \hat{w}_1 & \text{if } |T_{\mathbf{x}_{\hat{w}_1}|\mathbf{uy}}| \leq |T_{\mathbf{x}_{\hat{w}_2}|\mathbf{uy}}| \\ \hat{w}_2 & \text{otherwise} \end{cases}. \tag{31}$$

Since these two decoders achieve capacity, they can be regarded as universal SO and MO decoders for the class of channels $\mathcal{P}_n(D_2)$, w.r.t. random CC coding. Since $|T_{\mathbf{x}|\mathbf{uy}}| \doteq e^{n\hat{H}_{\mathbf{x}|\mathbf{uy}}}$ and $|T_{\mathbf{xx}'|\mathbf{uy}}| \doteq e^{n\hat{H}_{\mathbf{xx}'|\mathbf{uy}}}$, the alternative decoders which are also asymptotically optimal are given by

$$(\hat{w}_1,\hat{w}_2) \;=\; \mathrm{argmin}_{w_1,\, w_2\neq w_1}\hat{H}_{\mathbf{x}_{w_1}\mathbf{x}_{w_2}|\mathbf{uy}}. \tag{32}$$

$$\hat{w} \;=\; \begin{cases} \hat{w}_1 & \text{if } \hat{H}_{\mathbf{x}_{\hat{w}_1}|\mathbf{uy}} \leq \hat{H}_{\mathbf{x}_{\hat{w}_2}|\mathbf{uy}} \\ \hat{w}_2 & \text{otherwise} \end{cases}. \tag{33}$$

The proof of the converse part of Theorem 1 establishes the fact that when the encoder uses CC codes of rates higher than $C(P_U,D_1,D_2)$, the attacker can prevent reliable detection of any member of the coalition. The attacker, who knows $F_n$, can calculate $P_e\left(F_n, P_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}\right)$ for every $P_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}$, and then choose the attack channel associated with the highest $P_e$. We show that in order to prevent reliable identification of one of the members of the coalition at rates higher than $C(P_U,D_1,D_2)$, it is sufficient to consider attack channels that are uniform over a single conditional type $T_{\mathbf{y}|\mathbf{x}\tilde{\mathbf{x}}}$ for every $\mathbf{x},\tilde{\mathbf{x}}\in\mathcal{X}^n$.

The main tool behind the proof of the converse part is a variation of Fano's inequality, stating that if $P_e\left(F_n, P_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}\right)$ can be made arbitrarily small for large $n$, then,

$$R \;\leq\; \frac{1}{2n}I(\mathbf{X}_a\mathbf{X}_b;\mathbf{Y}|K_n,\mathbf{U}) + \epsilon_n, \tag{34}$$

where $\epsilon_n$ vanishes as $n$ tends infinity. We further show that the r.h.s. of (34) can be upper bounded by the single-letter expression given by $C(P_U,D_1,D_2)$. This is done by restricting the attack channel to be uniform over a single conditional type class $T_{\mathbf{y}|\mathbf{x},\tilde{\mathbf{x}}}$ given every $\mathbf{x}, \tilde{\mathbf{x}} \in \mathcal{X}^n$ yielding the upper bound.

The results can be extended to the case of a coalition of $L$ members whenever $L$ is constant, i.e., $L$ does not grow with $n$. If the size of the coalition tends to infinity as $n$ tend infinity, the capacity is zero. In Section 7, we give an outline of the proof of how this extension can be carried out, proving that the CC capacity of the private fingerprinting game w.r.t. a SO decoder and w.r.t. a MO decoder[4] is given by

$$C^{(L)}(P_U, D_1, D_2) = \max_{P_{X|U} \in \mathcal{D}_{d_1}(D_1, P_U)} \min_{P_{Y|X_1,\ldots,X_L}} \frac{1}{L} I(X_1, \ldots, X_L; Y|U), \tag{35}$$

where $U \sim P_U$, the minimization is over the set $\{P_{Y|X_1,\ldots,X_L} : \min_{i\in\{1,\ldots,L\}} Ed_2(X_i, Y) \leq D_2\}$, $(X_1, \ldots, X_L)$ is a random vector whose components are i.i.d. given $U$ with marginal pmf $P_{X_i|U} = P_{X|U}$, and $U \leftrightarrow (X_1, \ldots, X_L) \leftrightarrow Y$.

The asymptotically optimal MO decoder is given by

$$\hat{w} = \operatorname{argmin} \left| T_{\mathbf{x}_{i_1},\ldots,\mathbf{x}_{i_L}|\mathbf{uy}} \right|, \tag{36}$$

where the minimization is over all vectors $\{i_1, \ldots, i_L\} \in \mathcal{M}_n^L$ whose components are all distinct.

In a previous paper [23], it has been shown that the capacity of the private watermarking game is given by

$$\max_{P_{X|U}} \min_{P_{Y|UX}} I(X; Y|U), \tag{37}$$

where $U \sim P_U$, the maximization is over the set $\mathcal{D}_{d_1}(D_1, P_U)$, the minimization is over the set $\{P_{Y|UX} : Ed_1(X, Y) \leq D_2\}$, and $U \leftrightarrow X \leftrightarrow Y$. This capacity is achieved by the asymptotically optimal decoder for the ordinary watermarking problem, i.e.,

$$\hat{w} = \operatorname{argmin}_m |T_{\mathbf{x}_m|\mathbf{uy}}|, \tag{38}$$

which resembles the maximum mutual information (MMI) decoder [9]. The watermarking game can be viewed as a degenerate case of the fingerprinting game with a coalition that is composed of one member only. Since in the fingerprinting game the attacker may always choose to ignore one of its inputs, the capacity of the watermarking game (37) is an upper bound on the capacity of the fingerprinting game. We have also shown that the asymptotically optimal strategies of the fingerprinting game, given in (30) and (31), differ from (38), and hence, the expressions for the error exponents are also different.

It would be interesting to see what happens whenever one overestimates the size of the coalition. If the actual size of the coalition $L_1$ is unknown but it is upper bounded by some number, say $K$, one can use the CC code that minimizes the quantity $\max_{P_{\mathbf{Y}|\mathbf{X}_1,\ldots,\mathbf{X}_K} : \min_i d_2(\mathbf{X}_i,\mathbf{Y}) \leq nD_2} P_e(F_n, P_{\mathbf{Y}|\mathbf{X}_1,\ldots,\mathbf{X}_K})$ with the decoder (36) for $L = K$ and achieve the rate $C^{(K)}(P_U, D_1, D_2)$. Obviously $C^{(K)}(P_U, D_1, D_2) \leq C^{(L_1)}(P_U, D_1, D_2)$, but it is yet unclear whether this inequality is strict.

In this paper, we have established a single-letter expression for the CC capacity of the fingerprinting game. We conjecture that $C(P_U, D_1, D_2)$ is indeed the capacity even if the encoder takes on the more general form of (7).

[4]Here, by MO decoder, we mean a decoder having $L$ outputs.

# 6 Proofs of the Theorems

## 6.1 Proof of Proposition 1: Direct Theorem for a SO Decoder

The encoder draws the codewords $\mathbf{X}_i$, $i \in \mathcal{M}_n$, using a rate-$R$ CC fingerprinting code of block-length $n$. First, the encoder chooses a mapping $\tilde{\eta}_n : \mathcal{P}(\mathcal{U}) \to \mathcal{P}(\mathcal{X}|\mathcal{U})$ which satisfies the following requirements[5]

- $E_{P\times\tilde{\eta}_n(P)} d_1(U, X) \le D_1 \quad \forall P \in \mathcal{P}(\mathcal{U})$
- $\forall P \in \mathbb{P}_n(\mathcal{U}),\ \tilde{\eta}_n(P) \in \mathbb{P}_n(\mathcal{X}, P)$
- $\exists P_{X|U} \in \mathcal{P}(\mathcal{X}|\mathcal{U})$ s.t. if $\|P - P_U\| \le \ell(n)$, then for every $P$, $\tilde{\eta}_n(P)$ minimizes $\|P \times \tilde{\eta}_n(P) - P_U P_{X|U}\|$, (39)

where $\ell(n)$ is defined in (15). The set of mappings $\tilde{\eta}_n : \mathcal{P}(\mathcal{U}) \to \mathcal{P}(\mathcal{X}|\mathcal{U})$ which satisfy the above requirements are denoted $\tilde{\mathcal{N}}_n(D_1)$. The first two requirements assure that only conditional types such that constraint (9) is met are considered, and the last requirement fulfills the smoothness condition (see Definition 3, eq. (16)).

The fingerprinted copies $\mathbf{X}_i$, $i \in \mathcal{M}_n$ are then drawn independently given $\mathbf{U}$, uniformly over a single conditional type class $T_{\mathbf{u}}(\tilde{\eta}_n(\hat{P}_{\mathbf{u}}))$ whose choice is induced by the mapping $\tilde{\eta}_n(\hat{P}_{\mathbf{u}})$. We denote for convenience $T_{\tilde{\eta}_n}(\mathbf{u}) = T_{\mathbf{u}}(\tilde{\eta}_n(\hat{P}_{\mathbf{u}}))$. The decoder is given by (31).

Denote by $\mathbf{P}^M$ the joint pmf induced by $\{\mathbf{X}_i\}_{i=1}^M$. Assume, without loss of generality[6], that $W_a = 1, W_b = 2$. The probability of error, given that $(\mathbf{X}_a, \mathbf{X}_b, \mathbf{U}, \mathbf{Y})$ take on the values $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{u}, \mathbf{y})$, is upper bounded by the sum of the probabilities of the following three events

$$\begin{aligned}
A_1(\mathbf{u}\mathbf{x}_1\mathbf{x}_2\mathbf{y}) &\triangleq \{\exists i \ge 3, j \ge 3, i \ne j : \ |T_{\mathbf{x}_i\mathbf{x}_j|\mathbf{u}\mathbf{y}}| \le |T_{\mathbf{x}_1\mathbf{x}_2|\mathbf{u}\mathbf{y}}|\} \\
A_2(\mathbf{u}\mathbf{x}_1\mathbf{x}_2\mathbf{y}) &\triangleq \{\exists i \ge 3 : \ |T_{\mathbf{x}_1\mathbf{x}_i|\mathbf{u}\mathbf{y}}| \le |T_{\mathbf{x}_1\mathbf{x}_2|\mathbf{u}\mathbf{y}}| \text{ and } |T_{\mathbf{x}_i|\mathbf{u}\mathbf{y}}| \le |T_{\mathbf{x}_1|\mathbf{u}\mathbf{y}}|\} \\
A_3(\mathbf{u}\mathbf{x}_1\mathbf{x}_2\mathbf{y}) &\triangleq \{\exists i \ge 3 : \ |T_{\mathbf{x}_i\mathbf{x}_2|\mathbf{u}\mathbf{y}}| \le |T_{\mathbf{x}_1\mathbf{x}_2|\mathbf{u}\mathbf{y}}| \text{ and } |T_{\mathbf{x}_i|\mathbf{u}\mathbf{y}}| \le |T_{\mathbf{x}_2|\mathbf{u}\mathbf{y}}|\}.
\end{aligned} \tag{40}$$

**Lemma 1.** *Let $\mathbf{u}, \mathbf{x}, \hat{\mathbf{x}}, \mathbf{y} \in \mathcal{U}^n \times \mathcal{X}^n \times \mathcal{X}^n \times \mathcal{Y}^n$ be given $n$-vectors such that $\hat{\mathbf{x}} \in T_{\mathbf{x}|\mathbf{u}}$. The quantity*

$$\begin{aligned}
P_M &\triangleq P_M(\mathbf{x}, \hat{\mathbf{x}}, \mathbf{u}, \mathbf{y}) \\
&= \mathbf{P}^M\left\{\exists (i,j) \in \{1, \ldots, M\}^2, i \ne j \ s.t.\ (\mathbf{X}_i, \mathbf{X}_j) \in T_{\mathbf{x},\hat{\mathbf{x}}|\mathbf{u}\mathbf{y}}\right\}
\end{aligned} \tag{41}$$

*satisfies*

$$P_M \le \min\left\{1, \binom{M}{2} \cdot \frac{|T_{\mathbf{x}\hat{\mathbf{x}}|\mathbf{u}\mathbf{y}}|}{|T_{\mathbf{x}|\mathbf{u}}|^2}\right\}. \tag{42}$$

*Proof.* The lemma is a direct consequence of the fact that $\binom{M}{2}$ is the number of possibilities of choosing a pair of vectors among the $M$ vectors, and $\frac{|T_{\mathbf{x}\hat{\mathbf{x}}|\mathbf{u}\mathbf{y}}|}{|T_{\mathbf{x}|\mathbf{u}}|^2}$ is the probability that two i.i.d. vectors, uniformly distributed over $T_{\mathbf{x}|\mathbf{u}}$, lie within $T_{\mathbf{x}\hat{\mathbf{x}}|\mathbf{u}\mathbf{y}}$. □

[5]Here, by definition if $\mathbf{u}' \in T_{\mathbf{u}}$ then the $\tilde{\eta}_n(\hat{P}_{\mathbf{u}'}) = \tilde{\eta}_n(\hat{P}_{\mathbf{u}})$, whereas in the definition of $\eta_n$ in (13), $\eta_n(\mathbf{u})$ is not necessarily equal to $\eta_n(\mathbf{u}')$.

[6]The case $W_a = W_b$ can be treated similarly.

With a little abuse of notation, denote by $\mathbf{P}^M_{\mathbf{x}_1\mathbf{x}_2}$ the joint pmf of $\{\mathbf{X}_i\}_{i=1}^M$, given that $\mathbf{X}_1=\mathbf{x}_1$ and $\mathbf{X}_2=\mathbf{x}_2$. We have by Lemma 1 and the fact that $\mathbf{x}_1,\mathbf{x}_2\in T_{\tilde{\eta}_n}(\mathbf{u})$

$$\begin{aligned}
& \mathbf{P}^M_{\mathbf{x}_1\mathbf{x}_2}\{A_1(\mathbf{u}\mathbf{x}_1\mathbf{x}_2\mathbf{y})|\mathbf{u}\mathbf{x}_1\mathbf{x}_2\mathbf{y}\} \\
\leq\ & \max\left\{1, \binom{M-2}{2}\frac{|T_{\mathbf{x}_1\mathbf{x}_2|\mathbf{u}\mathbf{y}}|}{|T_{\tilde{\eta}_n}(\mathbf{u})|^2}\right\} \\
\leq\ & (n+1)^{2|\mathcal{U}|\mathcal{X}|}e^{-n|\hat{I}_{\mathbf{x}_1;\mathbf{y}|\mathbf{u}}+\hat{I}_{\mathbf{x}_2;\mathbf{x}_1\mathbf{y}|\mathbf{u}}-2R|^+}. \qquad (43)
\end{aligned}$$

The probability of the second event which results in error can be upper bounded as follows

$$\begin{aligned}
& \mathbf{P}^M_{\mathbf{x}_1\mathbf{x}_2}\{A_2(\mathbf{u}\mathbf{x}_1\mathbf{x}_2\mathbf{y})|\mathbf{u}\mathbf{x}_1\mathbf{x}_2\mathbf{y}\} \\
=\ & \mathbf{P}^M_{\mathbf{x}_1\mathbf{x}_2}\{\exists i\geq 3:\ |T_{\mathbf{x}_1|\mathbf{u}\mathbf{y}}|\cdot|T_{\mathbf{x}_i|\mathbf{u}\mathbf{x}_1\mathbf{y}}|\leq|T_{\mathbf{x}_1\mathbf{x}_2|\mathbf{u}\mathbf{y}}| \text{ and } |T_{\mathbf{x}_i|\mathbf{u}\mathbf{y}}|\leq|T_{\mathbf{x}_1|\mathbf{u}\mathbf{y}}|\} \\
\leq\ & \mathbf{P}^M_{\mathbf{x}_1\mathbf{x}_2}\{\exists i\geq 3:\ |T_{\mathbf{x}_i|\mathbf{u}\mathbf{y}}|\cdot|T_{\mathbf{x}_i|\mathbf{u}\mathbf{x}_1\mathbf{y}}|\leq|T_{\mathbf{x}_1\mathbf{x}_2|\mathbf{u}\mathbf{y}}|\} \\
\leq\ & \mathbf{P}^M_{\mathbf{x}_1\mathbf{x}_2}\{\exists i\geq 3:\ |T_{\mathbf{x}_i|\mathbf{u}\mathbf{x}_1\mathbf{y}}|^2\leq|T_{\mathbf{x}_1\mathbf{x}_2|\mathbf{u}\mathbf{y}}|\} \\
=\ & 1-\left(1-\sum_{\mathbf{x}':\ |T_{\mathbf{x}_i|\mathbf{u}\mathbf{x}_1\mathbf{y}}|\leq|T_{\mathbf{x}_1\mathbf{x}_2|\mathbf{u}\mathbf{y}}|^{1/2}}\frac{1}{|T_{\tilde{\eta}_n}(\mathbf{u})|}\right)^{M-2} \\
\leq\ & \max\left\{1, \sum_{\mathbf{x}':\ |T_{\mathbf{x}_i|\mathbf{u}\mathbf{x}_1\mathbf{y}}|\leq|T_{\mathbf{x}_1\mathbf{x}_2|\mathbf{u}\mathbf{y}}|^{1/2}} M\cdot\frac{1}{|T_{\tilde{\eta}_n}(\mathbf{u})|}\right\} \qquad (44) \\
\leq\ & \max\left\{1,(n+1)^{|\mathcal{U}||\mathcal{X}|^2|\mathcal{Y}|}M\frac{|T_{\mathbf{x}_1\mathbf{x}_2|\mathbf{u}\mathbf{y}}|^{1/2}}{|T_{\tilde{\eta}_n}(\mathbf{u})|}\right\} \\
\doteq\ & e^{-n|\frac{1}{2}\hat{I}_{\mathbf{x}_1;\mathbf{y}|\mathbf{u}}+\frac{1}{2}\hat{I}_{\mathbf{x}_2;\mathbf{x}_1\mathbf{y}|\mathbf{u}}-R|^+}, \qquad (45)
\end{aligned}$$

where (44) follows from the fact that for every $a\in[0,1]$, $1-(1-a)^M\leq\min\{1,Ma\}$.

Similarly,

$$\mathbf{P}^M_{\mathbf{x}_1\mathbf{x}_2}\{A_3(\mathbf{u}\mathbf{x}_1\mathbf{x}_2\mathbf{y})|\mathbf{u}\mathbf{x}_1\mathbf{x}_2\mathbf{y}\} \ \dot{\leq}\ e^{-n\left(\frac{1}{2}\hat{I}_{\mathbf{x}_1;\mathbf{y}|\mathbf{u}}+|\frac{1}{2}\hat{I}_{\mathbf{x}_2;\mathbf{x}_1\mathbf{y}|\mathbf{u}}-R|^+\right)}. \qquad (46)$$

Hence,

$$\begin{aligned}
\mathbf{P}^M_{\mathbf{x}_1\mathbf{x}_2}\{\text{ error }|\mathbf{u}\mathbf{x}_1\mathbf{x}_2\mathbf{y}\} &\leq \sum_{i=1}^{3}\mathbf{P}^M_{\mathbf{x}_1\mathbf{x}_2}\{A_i(\mathbf{u}\mathbf{x}_1\mathbf{x}_2\mathbf{y})|\mathbf{u}\mathbf{x}_1\mathbf{x}_2\mathbf{y}\} \\
&\dot{\leq} e^{-n|\frac{1}{2}\hat{I}_{\mathbf{x}_1;\mathbf{y}|\mathbf{u}}+\frac{1}{2}\hat{I}_{\mathbf{x}_2;\mathbf{x}_1\mathbf{y}|\mathbf{u}}-R|^+}. \qquad (47)
\end{aligned}$$

Denote

$$e_R(P_{UX\tilde{X}Y}) \ \triangleq\ e^{-n|\frac{1}{2}I(X;Y|U)+\frac{1}{2}I(\tilde{X};XY|U)-R|^+}, \qquad (48)$$

and note that the r.h.s. of (47) is a equal to $e_R(\hat{P}_{\mathbf{u}\mathbf{x}_1\mathbf{x}_2\mathbf{y}})$, and thus, we obtain

$$\begin{aligned}
& \inf_{\tilde{\eta}_n\in\tilde{\mathcal{N}}_n(D_1),\,\phi_n}\ \sup_{P_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}}\in\mathcal{P}_n^{d_2}} P_e(\tilde{\eta}_n,\phi_n,P_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}}) \\
\dot{\leq}\ & \inf_{\tilde{\eta}_n\in\tilde{\mathcal{N}}_n(D_1)}\ \sup_{P_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}}\in\mathcal{P}_n^{d_2}}\ \sum_{\mathbf{u},\,(\mathbf{x},\tilde{\mathbf{x}})\in T_{\tilde{\eta}_n}(\mathbf{u}),\,\mathbf{y}} P_U^n(\mathbf{u})\frac{1}{|T_{\tilde{\eta}_n}(\mathbf{u})|^2}P_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}}(\mathbf{y}|\mathbf{x}\tilde{\mathbf{x}})e_R(\hat{P}_{\mathbf{u}\mathbf{x}\tilde{\mathbf{x}}\mathbf{y}}), \qquad (49)
\end{aligned}$$

where $P_e(\tilde{\eta}_n, \phi_n, P_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}})$ is the average probability of error when the encoder uses the mapping $\tilde{\eta}_n$, the attack channel is $P_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}}$ and the decoder is $\phi_n$.

Next, we show that one can consider only exchangeable attack channels without loss of generality. Let $P_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}}$ be a given attack channel, which is a member of $\mathcal{P}_n^{d_2}$, and let $\pi$ be any one of the $n!$ permutations of $\{1, \ldots, n\}$. Denote by $P^{\pi}_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}}$ the channel defined by

$$P^{\pi}_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}}(\mathbf{y}|\mathbf{x}\tilde{\mathbf{x}}) = P_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}}(\pi\mathbf{y}|\pi\mathbf{x}\pi\tilde{\mathbf{x}}), \tag{50}$$

where $\pi\mathbf{x}$ designates the sequence $\mathbf{x}$ permuted according to $\pi$.

For a given choice of $\tilde{\eta}_n$ denote

$$L(P_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}}) \triangleq \sum_{\mathbf{u}} P_U^n(\mathbf{u}) \sum_{(\mathbf{x},\tilde{\mathbf{x}})\in T_{\tilde{\eta}_n}(\mathbf{u})} \frac{1}{|T_{\tilde{\eta}_n}(\mathbf{u})|^2} \sum_{\mathbf{y}} P_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}}(\mathbf{y}|\mathbf{x}\tilde{\mathbf{x}}) e_R(\hat{P}_{\mathbf{u}\mathbf{x}\tilde{\mathbf{x}}\mathbf{y}}). \tag{51}$$

Since $e_R(\hat{P}_{\mathbf{u}\mathbf{x}\tilde{\mathbf{x}}\mathbf{y}})$ depends solely on the type class $T_{\mathbf{u}\mathbf{x}\tilde{\mathbf{x}}\mathbf{y}}$, since if $\mathbf{u}' \in T_{\mathbf{u}}$, then $\tilde{\eta}_n(\hat{P}_{\mathbf{u}}) = \tilde{\eta}_n(\hat{P}_{\mathbf{u}'})$ since the covertext distribution is memoryless, i.e., $P_U^n(\mathbf{u})$ depends only on the type class of $\mathbf{u}$, and since $L(P_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}})$ is an affine functional of $P_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}}$ we have,

$$L\left(P_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}}\right) = L\left(P^{\pi}_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}}\right) = L\left(\frac{1}{n}\sum_{\pi} P^{\pi}_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}}\right). \tag{52}$$

Note that $\frac{1}{n}\sum_{\pi} P^{\pi}_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}}$ is a strongly exchangeable channel, as if $T_{\mathbf{x}\tilde{\mathbf{x}}\mathbf{y}} = T_{\mathbf{x}'\tilde{\mathbf{x}}'\mathbf{y}'}$, there exists a permutation $\bar{\pi}$ such that $(\mathbf{x}'\tilde{\mathbf{x}}'\mathbf{y}') = (\bar{\pi}\mathbf{x}\bar{\pi}\tilde{\mathbf{x}}\bar{\pi}\tilde{\mathbf{y}})$, and thus $\sum_{\pi} P_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}}(\pi\mathbf{y}|\pi\mathbf{x}\pi\tilde{\mathbf{x}}) = \sum_{\pi} P_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}}(\pi\mathbf{y}'|\pi\mathbf{x}'\pi\tilde{\mathbf{x}'})$.

Thus, $\frac{1}{n}\sum_{\pi} P^{\pi}_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}}(\mathbf{y}|\mathbf{x}\tilde{\mathbf{x}}) = \frac{P_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}}(T_{\mathbf{y}|\mathbf{x}\tilde{\mathbf{x}}}|\mathbf{x}\tilde{\mathbf{x}})}{|T_{\mathbf{y}|\mathbf{x}\tilde{\mathbf{x}}}|}$. Now, observe that (10) implies

$$P_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}}(T_{\mathbf{y}|\mathbf{x}\tilde{\mathbf{x}}}|\mathbf{x}\tilde{\mathbf{x}}) \leq \mathbf{1}\left\{\min\{d_2(\mathbf{x},\mathbf{y}), d_2(\tilde{\mathbf{x}},\mathbf{y})\} \leq nD_2\right\}, \tag{53}$$

and since $L(P_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}})$ is affine in $P_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}}$ we have

$$L\left(P_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}}\right) = L\left(\frac{P_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}}\left(T_{\mathbf{y}|\mathbf{x}\tilde{\mathbf{x}}}|\mathbf{x}\tilde{\mathbf{x}}\right)}{|T_{\mathbf{y}|\mathbf{x}\tilde{\mathbf{x}}}|}\right) \leq L\left(\frac{\mathbf{1}\left\{\min\{d_2(\mathbf{x},\mathbf{y}), d_2(\tilde{\mathbf{x}},\mathbf{y})\} \leq nD_2\right\}}{|T_{\mathbf{y}|\mathbf{x}\tilde{\mathbf{x}}}|}\right). \tag{54}$$

Eqs. (49) and (54) yield for a given mapping $\tilde{\eta}_n \in \tilde{\mathcal{N}}_n(D_1)$,

$$\begin{aligned}
&\sup_{P_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}}\in\mathcal{P}_n^{d_2}} L(P_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}}). \\
&\dot{\leq} \sum_{\mathbf{u}} P_U^n(\mathbf{u}) \sum_{(\mathbf{x}',\tilde{\mathbf{x}})\in T_{\tilde{\eta}_n}(\mathbf{u})} \frac{1}{|T_{\tilde{\eta}_n}(\mathbf{u})|^2} \sum_{\mathbf{y}:\, \min\{d_2(\mathbf{x}',\mathbf{y}),d_2(\tilde{\mathbf{x}},\mathbf{y})\}\leq nD_2} \frac{1}{|T_{\mathbf{y}|\mathbf{x}'\tilde{\mathbf{x}}}|} e_R(\hat{P}_{\mathbf{u}\mathbf{x}'\tilde{\mathbf{x}}\mathbf{y}}).
\end{aligned} \tag{55}$$

The inner-most summation satisfies

$$\begin{aligned}
&\sum_{\mathbf{y}:\, \min\{d_2(\mathbf{x}',\mathbf{y}),d_2(\tilde{\mathbf{x}},\mathbf{y})\}\leq nD_2} \frac{1}{|T_{\mathbf{y}|\mathbf{x}'\tilde{\mathbf{x}}}|} e_R(\hat{P}_{\mathbf{u}\mathbf{x}'\tilde{\mathbf{x}}\mathbf{y}}) \\
&\doteq \max_{\{T_{\mathbf{y}|\mathbf{u}\mathbf{x}\tilde{\mathbf{x}}}\}:\, \min\{d_2(\mathbf{x},\mathbf{y}),d_2(\tilde{\mathbf{x}},\mathbf{y})\}\leq nD_2} e^{-n\hat{I}_{\mathbf{u};\mathbf{y}|\mathbf{x}\tilde{\mathbf{x}}}} e_R(\hat{P}_{\mathbf{u}\mathbf{x}\tilde{\mathbf{x}}\mathbf{y}}) \\
&\triangleq g_R(\hat{P}_{\mathbf{u}\mathbf{x}\tilde{\mathbf{x}}})
\end{aligned} \tag{56}$$

because $\frac{1}{|T_{\mathbf{y}|\mathbf{x}'\tilde{\mathbf{x}}}|}e_R(\hat{P}_{\mathbf{u}\mathbf{x}'\tilde{\mathbf{x}}\mathbf{y}})$ dependes only on $T_{\mathbf{u}\mathbf{x}\tilde{\mathbf{x}}\mathbf{y}}$ and $\frac{|T_{\mathbf{y}|\mathbf{u}\mathbf{x}'\tilde{\mathbf{x}}}|}{|T_{\mathbf{y}|\mathbf{x}'\tilde{\mathbf{x}}}|} \doteq e^{-n\hat{I}_{\mathbf{u};\mathbf{y}|\mathbf{x}\tilde{\mathbf{x}}}}$. For a pmf $P_{\tilde{U}X\tilde{X}Y} \in \mathcal{U} \times \mathcal{X}^2 \times \mathcal{Y}$, denote, for convenience,

$$G_R(P_{\tilde{U}X\tilde{X}Y}) = I(X;\tilde{X}|\tilde{U}) + I(\tilde{U};Y|X\tilde{X}) + \left|\frac{1}{2}I(X;Y|\tilde{U}) + \frac{1}{2}I(\tilde{X};XY|\tilde{U}) - R\right|^+ . \quad (57)$$

We have

$$\begin{aligned}
&\sum_{(\mathbf{x}',\tilde{\mathbf{x}})\in T_{\tilde{\eta}_n}(\mathbf{u})} \frac{1}{|T_{\tilde{\eta}_n}(\mathbf{u})|^2} g_R(\hat{P}_{\mathbf{u}\mathbf{x}\tilde{\mathbf{x}}}) \\
\doteq\ & \sum_{\mathbf{x}'\in T_{\tilde{\eta}_n}(\mathbf{u})} \frac{1}{|T_{\tilde{\eta}_n}(\mathbf{u})|} \sum_{T_{\tilde{\mathbf{x}}|\mathbf{u}\mathbf{x}'}\subseteq T_{\tilde{\eta}_n}(\mathbf{u})} \frac{|T_{\tilde{\mathbf{x}}|\mathbf{x}'\mathbf{u}}|}{|T_{\tilde{\eta}_n}(\mathbf{u})|} g_R(\hat{P}_{\mathbf{u}\mathbf{x}\tilde{\mathbf{x}}}) \\
\doteq\ & \max_{T_{\tilde{\mathbf{x}}|\mathbf{u}\mathbf{x}}\subseteq T_{\tilde{\eta}_n}(\mathbf{u})} e^{-n\hat{I}_{\mathbf{x};\tilde{\mathbf{x}}|\mathbf{u}}} g_R(\hat{P}_{\mathbf{u}\mathbf{x}\tilde{\mathbf{x}}}) \\
\le\ & \exp\left\{-n\left(\min_{P_{\tilde{X}Y|UX}\in\Phi(\hat{P}_{\mathbf{u}}\times\tilde{\eta}_n(\hat{P}_{\mathbf{u}}),D_2)} G_R(\hat{P}_{\mathbf{u}}\times\tilde{\eta}_n(\hat{P}_{\mathbf{u}})\times P_{\tilde{X}Y|UX})\right)\right\}, \qquad (58)
\end{aligned}$$

where the last step is by definition of $e_R(\hat{P}_{\mathbf{u}\mathbf{x}\tilde{\mathbf{x}}\mathbf{y}})$ (see (47)), and $\Phi(\hat{P}_{\mathbf{u}} \times \tilde{\eta}_n(\hat{P}_{\mathbf{u}}), D_2)$ is the set of conditional pmf's $P_{\tilde{X}Y|UX} \in \mathcal{P}(\mathcal{X}\times\mathcal{Y}|\mathcal{U}\times\mathcal{X})$ satisfying the following two conditions

$$\begin{aligned}
&\sum_x \tilde{\eta}_n(\hat{P}_{\mathbf{u}})(x|u)P_{\tilde{X}|UX}(\tilde{x}|ux) = \tilde{\eta}_n(\hat{P}_{\mathbf{u}})(\tilde{x}|u) \quad \forall u\tilde{x}\in\mathcal{U}\times\mathcal{X}, \\
&\min\{E_{\hat{P}_{\mathbf{u}}\times\tilde{\eta}_n(\hat{P}_{\mathbf{u}})P_{\tilde{X}Y|UX}} d_2(X,Y), E_{\hat{P}_{\mathbf{u}}\times\tilde{\eta}_n(\hat{P}_{\mathbf{u}})P_{\tilde{X}Y|UX}} d_2(\tilde{X},Y)\} \le D_2. \qquad (59)
\end{aligned}$$

Consequently, the r.h.s. of (55) is equal (in the exponential scale) to

$$\begin{aligned}
&\sum_{T_{\mathbf{u}}} \exp\left\{-n\left(D(\hat{P}_{\mathbf{u}}\|P_U) + \min_{P_{\tilde{X}Y|UX}\in\Phi(\hat{P}_{\mathbf{u}}\times\tilde{\eta}_n(\hat{P}_{\mathbf{u}}),D_2)} G_R(\hat{P}_{\mathbf{u}}\times\tilde{\eta}_n(\hat{P}_{\mathbf{u}})\times P_{\tilde{X}Y|UX})\right)\right\} \\
\triangleq\ & \sum_{T_{\mathbf{u}}} \Xi(\hat{P}_{\mathbf{u}}, \tilde{\eta}_n(\hat{P}_{\mathbf{u}}), D_2), \qquad (60)
\end{aligned}$$

where for a pmf $P \in \mathcal{P}(\mathcal{U})$

$$\begin{aligned}
&\Xi(P, \tilde{\eta}_n(P), D_2) \\
\triangleq\ & \exp\left\{-n\left(D(P\|P_U) + \min_{P_{\tilde{X}Y|UX}\in\Phi(P\times\tilde{\eta}_n(P),D_2)} G_R(P\times\tilde{\eta}_n(P)\times P_{\tilde{X}Y|UX})\right)\right\}. \qquad (61)
\end{aligned}$$

To upper bound the infimum over $\tilde{\eta}_n \in \tilde{\mathcal{N}}_n(D_1)$ of (60), we shall separate the summation over the type classes as follows

$$
\begin{aligned}
& \inf_{\tilde{\eta}_n \in \tilde{\mathcal{N}}_n(D_1)} \left( \sum_{T_{\mathbf{u}}:\, \|\hat{P}_{\mathbf{u}} - P_U\| \leq \ell(n)} \Xi(\hat{P}_{\mathbf{u}}, \tilde{\eta}_n(\hat{P}_{\mathbf{u}}), D_2) + \sum_{T_{\mathbf{u}}:\, \|\hat{P}_{\mathbf{u}} - P_U\| > \ell(n)} \Xi(\hat{P}_{\mathbf{u}}, \tilde{\eta}_n(\hat{P}_{\mathbf{u}}), D_2) \right) \\
= & \inf_{\tilde{\eta}_n \in \tilde{\mathcal{N}}_n(D_1)} \left( \sum_{T_{\mathbf{u}}:\, \|\hat{P}_{\mathbf{u}} - P_U\| \leq \ell(n)} \Xi(\hat{P}_{\mathbf{u}}, \tilde{\eta}_n(\hat{P}_{\mathbf{u}}), D_2) \right) \\
+ & \sum_{T_{\mathbf{u}}:\, \|\hat{P}_{\mathbf{u}} - P_U\| > \ell(n)} \inf_{\tilde{\eta}_n(\hat{P}_{\mathbf{u}})} \Xi(\hat{P}_{\mathbf{u}}, \tilde{\eta}_n(\hat{P}_{\mathbf{u}}), D_2) \qquad (62)
\end{aligned}
$$

where the infimum is inserted inside the summation in the right term since the third requirement in (39) applies only for types s.t. $\|\hat{P}_{\mathbf{u}} - P_U\| \leq \ell(n)$. From the continuity of $\Xi(P, \tilde{\eta}_n(P), D_2)$ in $P$ and from the third requirement in (39) as $\ell(n) \to 0$ it follows that there exists a sequence $\epsilon_n \geq 0$ which converges to zero such that the left term of the above equation is upper bounded as follows

$$
\begin{aligned}
& \inf_{\tilde{\eta}_n \in \tilde{\mathcal{N}}_n(D_1)} \left( \sum_{T_{\mathbf{u}}:\, \|\hat{P}_{\mathbf{u}} - P_U\| \leq \ell(n)} \Xi(\hat{P}_{\mathbf{u}}, \tilde{\eta}_n(\hat{P}_{\mathbf{u}}), D_2) \right) \\
\leq & \inf_{P_{X|U}: E_{P_U P_{X|U}} d_1(U,X) \leq D_1 - \epsilon_n} \Xi(P_U, P_{X|U}, D_2 + \epsilon_n) + \epsilon_n. \qquad (63)
\end{aligned}
$$

As for the right term of (62), we obtain similarly

$$
\begin{aligned}
& \sum_{T_{\mathbf{u}}:\, \|\hat{P}_{\mathbf{u}} - P_U\| > \ell(n)} \inf_{\tilde{\eta}_n(\hat{P}_{\mathbf{u}})} \Xi(\hat{P}_{\mathbf{u}}, \tilde{\eta}_n(\hat{P}_{\mathbf{u}}), D_2) \\
\leq & \sup_{P_{\tilde{U}} \in \mathcal{P}(\mathcal{U})} \inf_{P_{X|\tilde{U}}:\, E_{P_{\tilde{U}} P_{X|\tilde{U}}} d_1(\tilde{U},X) \leq D_1 - \epsilon_n} \Xi(P_{\tilde{U}}, P_{X|\tilde{U}}, D_2 + \epsilon_n) + \epsilon_n. \qquad (64)
\end{aligned}
$$

Gathering (49)-(64) yields,

$$
\begin{aligned}
& \inf_{\tilde{\eta}_n \in \tilde{\mathcal{N}}_n(D_1),\, \phi_n} \sup_{P_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}} \in \mathcal{P}_n^{d_2}} P_e(\tilde{\eta}_n, \phi_n, P_{\mathbf{Y}|\mathbf{X}\tilde{\mathbf{X}}}) \\
\stackrel{.}{\leq} & \sup_{P_{\tilde{U}} \in \mathcal{P}(\mathcal{U})} \inf_{P_{X|\tilde{U}}:\, E_{P_{\tilde{U}} P_{X|\tilde{U}}} d_1(\tilde{U},X) \leq D_1 - \epsilon_n} \Xi(P_{\tilde{U}}, P_{X|\tilde{U}}, D_2 + \epsilon_n) + \epsilon_n \\
= & \ \xi_1(P_U, D_1 - \epsilon_n, D_2 + \epsilon_n, R) + \epsilon_n. \qquad (65)
\end{aligned}
$$

$\square$

## 6.2 Proof of Proposition 2: Direct Theorem for a MO Decoder

As the decoder is allowed to have two outputs, we use the decoder

$$(\hat{w}_1, \hat{w}_2) = \mathrm{argmin}_{w_1, w_2 \neq w_1} |T_{\mathbf{x}_{w_1} \mathbf{x}_{w_2} | \mathbf{uy}}|, \qquad (66)$$

as in (30).

Assume, again, without loss of generality[7], that $W_a = 1$ and $W_b = 2$. Since the decoder errs only if there exist two codewords $\mathbf{x}_i, \mathbf{x}_j, i \geq 3, i \neq j \geq 3$ such that $|T_{\mathbf{x}_i\mathbf{x}_j|\mathbf{uy}}| \leq |T_{\mathbf{x}_1\mathbf{x}_2|\mathbf{uy}}|$, we have $P(\text{error}|\mathbf{ux}_1\mathbf{x}_2\mathbf{y}) = P_{M-2}(\mathbf{x}_1\mathbf{x}_2, \mathbf{u}, \mathbf{y})$ (see (41)), and by Lemma 1

$$\begin{aligned} P(\text{error}|\mathbf{ux}_1\mathbf{x}_2\mathbf{y}) &\leq \max\left\{1, M^2\frac{|T_{\mathbf{x}_1\mathbf{x}_2|\mathbf{uy}}|}{|T_{\tilde{\eta}_n}(\mathbf{u})|^2}\right\} \\ &\leq (n+1)^{2|\mathcal{U}||\mathcal{X}|}e^{-n|\hat{I}_{\mathbf{x}_1;\mathbf{y}|\mathbf{u}}+\hat{I}_{\mathbf{x}_2;\mathbf{x}_1\mathbf{y}|\mathbf{u}}-2R|^+} \\ &\doteq e_R^2(\hat{P}_{\mathbf{ux}_1\mathbf{x}_2\mathbf{y}}). \end{aligned} \tag{67}$$

where $e_R(\hat{P}_{\mathbf{ux}_1\mathbf{x}_2\mathbf{y}})$ is defined in (48). The rest of the derivation can be carried out exactly as in the proof of Theorem 1 (see (48)-(65)), with the only difference that $e_R^2(\hat{P}_{\mathbf{ux}_1\mathbf{x}_2\mathbf{y}})$ replaces $e_R(\hat{P}_{\mathbf{ux}_1\mathbf{x}_2\mathbf{y}})$.

## 6.3 Converse Theorem for the SO Decoder

For a given codebook (or, realization of the secret key $K_n$), recall that the estimate of the embedded message is given by $\hat{W} = \phi(K_n, \mathbf{U}, \mathbf{Y})$. An error occurs if the estimate is not a member of the coalition, and therefore, the probability of error is defined as

$$P_e = \Pr(\hat{W} \neq W_a, \hat{W} \neq W_b). \tag{68}$$

Define,

$$\begin{aligned} E_a &= \begin{cases} 1 & \text{if } \hat{W} \neq W_a \\ 0 & \text{if } \hat{W} = W_a \end{cases} \\ E_b &= \begin{cases} 1 & \text{if } \hat{W} \neq W_b \\ 0 & \text{if } \hat{W} = W_b \end{cases}. \end{aligned} \tag{69}$$

Using the chain rule for entropies to expand $H(E_a, E_b, W_a, W_b|K_n\mathbf{UY})$ in two different ways, we have,

$$\begin{aligned} H(E_a, E_b, W_a, W_b|K_n\mathbf{UY}) &= H(W_a, W_b|K_n, \mathbf{U}, \mathbf{Y}) + H(E_a, E_b|K_n, \mathbf{U}, \mathbf{Y}, W_a, W_b) \\ &= H(E_a, E_b|K_n, \mathbf{U}, \mathbf{Y}) + H(W_a, W_b|K_n, \mathbf{U}, \mathbf{Y}, E_a, E_b). \end{aligned} \tag{70}$$

Also, $H(E_a, E_b) \leq 2$ since $E_a, E_b$ are binary valued RVs, this yields,

$$\begin{aligned} & H(W_a, W_b|K_n, \mathbf{U}, \mathbf{Y}) \\ =\; & H(E_a, E_b|K_n, \mathbf{U}, \mathbf{Y}) + H(W_a, W_b|K_n, \mathbf{U}, \mathbf{Y}, E_a, E_b) - H(E_a, E_b|K_n, \mathbf{U}, \mathbf{Y}, W_a, W_b) \\ \leq\; & 2 + H(W_a, W_b|K_n, \mathbf{U}, \mathbf{Y}, E_a, E_b). \end{aligned} \tag{71}$$

[7]The case $W_a = W_b$ can be treated similarly.

The term $H(W_a, W_b|K_n, \mathbf{U}, \mathbf{Y}, E_a, E_b)$, can be bounded as follows

$$
\begin{aligned}
& H(W_a, W_b|K_n, \mathbf{U}, \mathbf{Y}, E_a, E_b) \\
= \; & P_e \cdot H(W_a, W_b|K_n, \mathbf{U}, \mathbf{Y}, \hat{W} \neq W_a, \hat{W} \neq W_b) \\
+ \; & \Pr(\hat{W} = W_b, W_a \neq W_b) \cdot H(W_a, W_b|K_n, \mathbf{U}, \mathbf{Y}, \hat{W} = W_b, W_a \neq W_b) \\
+ \; & \Pr(\hat{W} = W_a, W_a \neq W_b) \cdot H(W_a, W_b|K_n, \mathbf{U}, \mathbf{Y}, \hat{W} = W_a, W_a \neq W_b) \\
+ \; & \Pr(\hat{W} = W_a = W_b) \cdot H(W_a, W_b|K_n, \mathbf{U}, \mathbf{Y}, \hat{W} = W_a = W_b) \qquad (72) \\
\leq \; & P_e \cdot 2nR + \Pr(\hat{W} = W_b, W_a \neq W_b) \cdot H(W_a|K_n, \mathbf{U}, \mathbf{Y}, W_b) \\
+ \; & \Pr(\hat{W} = W_a, W_a \neq W_b) \cdot H(W_b|K_n, \mathbf{U}, \mathbf{Y}, W_a) \qquad (73)
\end{aligned}
$$

where (72) is because $\hat{W} = \phi(K_n, \mathbf{U}, \mathbf{Y})$, so $W_i$ is known given $\hat{W} = W_i, K_n, \mathbf{U}, \mathbf{Y}$, hence,

$$
H(W_a, W_b|K_n, \mathbf{U}, \mathbf{Y}, \hat{W} = W_a = W_b) = 0, \qquad (74)
$$

and

$$
\begin{aligned}
& H(W_a, W_b|K_n, \mathbf{U}, \mathbf{Y}, \hat{W} = W_b, W_a \neq W_b) \\
= \; & H(W_a, W_b|K_n, \mathbf{U}, \mathbf{Y}, \hat{W} = W_b, W_b, W_a \neq W_b) \\
\leq \; & H(W_a|K_n, \mathbf{U}, \mathbf{Y}, W_b). \qquad (75)
\end{aligned}
$$

Now, we shall constrain the attacker to use only channels that are symmetric w.r.t. its two inputs $\mathbf{X}_a$ and $\mathbf{X}_b$, i.e.,

$$
P_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}(\mathbf{y}|\mathbf{x}\tilde{\mathbf{x}}) \; = \; P_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}(\mathbf{y}|\tilde{\mathbf{x}}\mathbf{x}), \quad \forall \mathbf{x}, \tilde{\mathbf{x}} \in \mathcal{X}^n. \qquad (76)
$$

Under this assumption it can be argued that

$$
\Pr(\hat{W} = W_b, W_a \neq W_b) \; = \; \Pr(\hat{W} = W_a, W_a \neq W_b). \qquad (77)
$$

To see why this is true, note that given every pair of realizations $(k_n, \mathbf{u})$ of $(K_n, \mathbf{U})$,

$$
\begin{aligned}
& \Pr(\hat{W} = W_b, W_a \neq W_b|K_n = k_n, \mathbf{U} = \mathbf{u}) \\
= \; & \sum_{w_a, w_b \neq w_a, \mathbf{y}: \phi(k_n, \mathbf{u}, \mathbf{y}) = w_b} \Pr(w_a w_b \mathbf{y}|k_n \mathbf{u}) \\
= \; & \sum_{i=1}^{M} \sum_{j \neq i} \sum_{\mathbf{y}: \phi(k_n, \mathbf{u}, \mathbf{y}) = j} \frac{1}{M} \cdot \frac{1}{(M-1)} P_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}(\mathbf{y}|\mathbf{x}(\mathbf{u}, k_n, i), \mathbf{x}(\mathbf{u}, k_n, j)) \\
= \; & \sum_{i=1}^{M} \sum_{j \neq i} \sum_{\mathbf{y}: \phi(k_n, \mathbf{u}, \mathbf{y}) = i} \frac{1}{M} \cdot \frac{1}{(M-1)} P_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}(\mathbf{y}|\mathbf{x}(\mathbf{u}, k_n, j), \mathbf{x}(\mathbf{u}, k_n, i)) \\
= \; & \Pr(\hat{W} = W_a, W_a \neq W_b|K_n = k_n, \mathbf{U} = \mathbf{u}), \qquad (78)
\end{aligned}
$$

where $\mathbf{x}(\cdot)$ is the mapping from the specific codebook defined by $\mathbf{u}, k_n$ and the message index to a codeword.

So, we have

$$\begin{aligned} & H(W_a, W_b|K_n, \mathbf{U}, \mathbf{Y}, E_a, E_b) \\ \leq \;& P_e \cdot 2nR + \frac{1}{2}H(W_a|K_n, \mathbf{U}, \mathbf{Y}, W_b) + \frac{1}{2}H(W_b|K_n, \mathbf{U}, \mathbf{Y}, W_a). \end{aligned} \tag{79}$$

Now,

$$\begin{aligned} 2nR &= H(W_a, W_b|K_n, \mathbf{U}) = H(W_a, W_b|K_n, \mathbf{U}, \mathbf{Y}) + I(W_a, W_b; \mathbf{Y}|K_n, \mathbf{U}) \\ &\leq 2 + P_e \cdot 2nR \\ &+ \frac{1}{2}H(W_a|K_n, \mathbf{U}, \mathbf{Y}, W_b) + \frac{1}{2}H(W_b|K_n, \mathbf{U}, \mathbf{Y}, W_a) + I(W_a, W_b; \mathbf{Y}|K_n, \mathbf{U}). \end{aligned} \tag{80}$$

Subtracting the term $H(W_a|W_b, K_n, \mathbf{U})$ (or $H(W_b|W_a, K_n, \mathbf{U})$), which is equal to $nR$, from both sides of the inequality, we get

$$\begin{aligned} nR &\leq n\epsilon_{nR} - \frac{1}{2}I(W_a; \mathbf{Y}|W_b, K_n, \mathbf{U}) - \frac{1}{2}I(W_b; \mathbf{Y}|W_a, K_n, \mathbf{U}) + I(W_a, W_b; \mathbf{Y}|K_n, \mathbf{U}) \\ &= n\epsilon_{nR} + \frac{1}{2}I(W_a; \mathbf{Y}|K_n, \mathbf{U}) + \frac{1}{2}I(W_b; \mathbf{Y}|K_n, \mathbf{U}), \end{aligned} \tag{81}$$

where $\epsilon_{nR} = \frac{2}{n} + P_e \cdot 2R$ which vanishes if $R$ is below capacity, and the last equality is due to the chain rule for mutual information

$$I(W_a, W_b; \mathbf{Y}|K_n\mathbf{U}) \;=\; I(W_a; \mathbf{Y}|K_n\mathbf{U}) + I(W_b; \mathbf{Y}|W_aK_n\mathbf{U}). \tag{82}$$

Therefore, we have,

$$\begin{aligned} R &\leq \epsilon_{n,R} + \frac{1}{n}\left[\frac{1}{2}I(W_a; \mathbf{Y}|K_n, \mathbf{U}) + \frac{1}{2}I(W_b; \mathbf{Y}|K_n, \mathbf{U})\right] \\ &= \epsilon_{n,R} + \frac{1}{n}\left[\frac{1}{2}I(\mathbf{X}_a; \mathbf{Y}|K_n, \mathbf{U}) + \frac{1}{2}I(\mathbf{X}_b; \mathbf{Y}|K_n, \mathbf{U})\right] \\ &\leq \epsilon_{n,R} + \frac{1}{n}\left[\frac{1}{2}I(\mathbf{X}_a; \mathbf{Y}|K_n, \mathbf{U}) + \frac{1}{2}I(\mathbf{X}_b; \mathbf{Y}\mathbf{X}_a|K_n, \mathbf{U})\right] \\ &= \epsilon_{n,R} + \frac{1}{2n}I(\mathbf{X}_a\mathbf{X}_b; \mathbf{Y}|K_n, \mathbf{U}), \end{aligned} \tag{83}$$

where (83) is due to the chain rule for mutual information and the following equality

$$I(C; D|BA) \;=\; I(C; BD|A), \tag{84}$$

that holds whenever $A, B, C, D$ are RV's that are distributed $P_A \times P_{B|A} \times P_{C|A} \times P_{D|B,C}$ (applied for $(\mathbf{U}, K_n), \mathbf{X}_a, \mathbf{X}_b, \mathbf{Y}$ in the roles of $A, B, C, D$, respectively).

This implies that for a given random coding distribution

$$R \;\leq\; \min_{P_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b} \in \mathcal{P}_n^{d_2}:\, P_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b} = P_{\mathbf{Y}|\mathbf{X}_b\mathbf{X}_a}} \frac{1}{2n}I(\mathbf{X}_a\mathbf{X}_b; \mathbf{Y}|K_n, \mathbf{U}). \tag{85}$$

Next, we shall show that the symmetry constraint (76) does not cause any loss of generality.

**Lemma 2.**

$$
\begin{aligned}
& \min_{P_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b} \in \mathcal{P}_n^{d_2}:\, P_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b} = P_{\mathbf{Y}|\mathbf{X}_b\mathbf{X}_a}} \frac{1}{2n} I(\mathbf{X}_a\mathbf{X}_b;\mathbf{Y}|K_n,\mathbf{U}) \\
= & \min_{P_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b} \in \mathcal{P}_n^{d_2}} \frac{1}{2n} I(\mathbf{X}_a\mathbf{X}_b;\mathbf{Y}|K_n,\mathbf{U}). \qquad (86)
\end{aligned}
$$

The proof of Lemma 2 appears in Appendix A.

The rest of the proof is devoted to upper bounding $I(\mathbf{X}_a\mathbf{X}_b;\mathbf{Y}|K_n,\mathbf{U})$ by a single-letter expression. This will be done by constraining the attacker to use pmf's that are uniform over a single conditional type class given every pair $\mathbf{x},\mathbf{x}'$ in a strongly exchangeable manner. Since for every joint types of pairs of sequences, $\hat{P}_{\mathbf{x}\tilde{\mathbf{x}}}$, the attacker assigns a conditional types class $T_{\mathbf{y}|\mathbf{x}\tilde{\mathbf{x}}}$ within which lies the forgery, we can define the operation of the attacker as choosing a mapping from $\mathcal{P}(\mathcal{X}\times\mathcal{X})$ to $\mathcal{P}(\mathcal{Y}|\mathcal{X}\times\mathcal{X})$. Consider the following set

**Definition 6.** *Let $\mathcal{R}_n(D_2)$ be the set of of functions $\rho_n : \mathcal{P}(\mathcal{X}\times\mathcal{X}) \to \mathcal{P}(\mathcal{Y}|\mathcal{X}\times\mathcal{X})$ such that for all $P_{X\tilde{X}} \in \mathcal{P}(\mathcal{X}\times\mathcal{X})$, the conditional pmf $\rho_n(P_{X\tilde{X}})$ satisfies the following two conditions*

$$
\min\left\{E_{P_{X\tilde{X}}\rho_n(P_{X\tilde{X}})} d_2(X,Y), E_{P_{X\tilde{X}}\rho_n(P_{X\tilde{X}})} d_2(\tilde{X},Y)\right\} \le D_2, \ \text{and} \qquad (87)
$$

$$
\text{if } P_{X\tilde{X}} \in \mathbb{P}_n(\mathcal{X}^2) \text{ then } P_{Y|X\tilde{X}} \in \mathbb{P}_n(\mathcal{Y}, P_{X\tilde{X}}). \qquad (88)
$$

**Definition 7.** *For two sequences $\mathbf{x}\mathbf{x}'$ we shall denote by $T_{\rho_n}(\mathbf{x}\mathbf{x}')$ the conditional type class induced by the conditional empirical measure $\rho_n(\hat{P}_{\mathbf{x}\mathbf{x}'})$ given the pair $\mathbf{x}\mathbf{x}'$, i.e., the set of sequences $\mathbf{y}$ such that $\hat{P}_{\mathbf{x}\mathbf{x}'\mathbf{y}} = \hat{P}_{\mathbf{x}\mathbf{x}'} \times \rho_n(\hat{P}_{\mathbf{x}\mathbf{x}'})$.*

The attacker optimizes the choice of a function $\rho_n \in \mathcal{R}_n(D_2)$, and when s/he observes $\mathbf{x}_a, \mathbf{x}_b$ s/he draws $\mathbf{Y}$ uniformly over the conditional type class $T_{\rho_n}(\mathbf{x}\mathbf{x}')$. The requirement (87) ensures that distortion constraint (10) is met.

Under the constraint that the attacker uses pmf's that are uniform over a single conditional type class given every pair of observations, we have

$$
\begin{aligned}
I(\mathbf{X}_a\mathbf{X}_b;\mathbf{Y}|K_n\mathbf{U}) &= H(\mathbf{Y}|K_n\mathbf{U}) - H(\mathbf{Y}|K_n\mathbf{X}_a\mathbf{X}_b\mathbf{U}) \\
&\le H(\mathbf{Y}|\mathbf{U}) - H(\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b) \qquad (89) \\
&= H(\mathbf{Y}|\mathbf{U}) - E\log|T_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}|, \qquad (90)
\end{aligned}
$$

where (89) is because conditioning reduces entropy and $(\mathbf{U},K_n) \leftrightarrow \mathbf{X}_a\mathbf{X}_b \leftrightarrow \mathbf{Y}$ is a Markov chain, and (90) is a consequence of the fact that we constrain the attacker to use channels that are uniform over a single conditional type class given every $(\mathbf{x}_a,\mathbf{x}_b)$.

Moreover,

$$
\begin{aligned}
H(\mathbf{Y}|\mathbf{U}) &\le H(\mathbf{Y},\hat{P}_{\mathbf{UY}}|\mathbf{U}) = H(\hat{P}_{\mathbf{UY}}|\mathbf{U}) + H(\mathbf{Y}|\mathbf{U}\hat{P}_{\mathbf{UY}}) \\
&\le |\mathcal{U}||\mathcal{Y}|\log(n+1) + E\log|T_{\mathbf{Y}|\mathbf{U}}|, \qquad (91)
\end{aligned}
$$

where the last step follows from the fact that the number of types grows polynomially with $n$ and the entropy of a finite-alphabet random variable is upper bounded by the logarithm of the cardinality of the alphabet.

Since for all $\mathbf{u}, \mathbf{x}_a, \mathbf{x}_b, \mathbf{y}$

$$\begin{aligned}\log|T_{\mathbf{y}|\mathbf{u}}| - \log|T_{\mathbf{y}|\mathbf{x}_a\mathbf{x}_b}| &\leq \hat{H}_{\mathbf{y}|\mathbf{u}} - \hat{H}_{\mathbf{y}|\mathbf{x}_a\mathbf{x}_b} + |\mathcal{X}|^2|\mathcal{Y}|\log(n+1) \\ &\leq \hat{I}_{\mathbf{x}_a\mathbf{x}_b;\mathbf{y}|\mathbf{u}} + |\mathcal{X}|^2|\mathcal{Y}|\log(n+1), \end{aligned} \tag{92}$$

equations (85), (86), (89) and (91) yield

$$R \leq \epsilon'_{n,R} + \max_{\eta_n \in \mathcal{N}_n(D_1)} \min_{\rho_n \in \mathcal{R}_n(D_2)} \frac{1}{2} E\hat{I}_{\mathbf{X}_a\mathbf{X}_b;\mathbf{Y}|\mathbf{U}}, \tag{93}$$

where $\epsilon'_{n,R} = \epsilon_{n,R} + \frac{|\mathcal{X}|^2|\mathcal{Y}|\log(n+1)}{n}$ and $\mathcal{N}_n(D_1)$ is defined in eq. (14). The following Lemma concludes the proof of the converse part for SO decoder.

**Lemma 3.** *Under the smoothness assumption (see Definition 3)*

$$\limsup_{n\to\infty} \max_{\eta_n \in \mathcal{N}_n(D_1)} \min_{\rho_n \in \mathcal{R}_n(D_2)} \frac{1}{2} E\hat{I}_{\mathbf{X}_a\mathbf{X}_b;\mathbf{Y}|\mathbf{U}} \leq C(P_U, D_1, D_2). \tag{94}$$

The proof of Lemma 3 appears in Appendix A.

## 6.4 Converse Theorem for a MO Decoder

In order to prove the converse part, we shall establish inequality (83) also for the MO decoder (with $\epsilon_{n,R} = \frac{4}{n} + 2RP_e$). This implies that also the capacity of the MO decoder is upper bounded[8] by $C(P_U, D_1, D_2)$ (21).

Denote by $\hat{\nu} = \psi_\nu(K_n, \mathbf{U}, \mathbf{Y}), \hat{\mu} = \psi_\mu(K_n, \mathbf{U}, \mathbf{Y})$ the two outputs of the decoder, i.e., the estimates of $W_a, W_b$. Define the following 4 random variables

$$\begin{aligned} E_a &= \begin{cases} 1 & \text{if } \hat{\nu} \neq W_a \\ 0 & \text{if } \hat{\nu} = W_a \end{cases} \quad ; \quad E_b = \begin{cases} 1 & \text{if } \hat{\nu} \neq W_b \\ 0 & \text{if } \hat{\nu} = W_b \end{cases} \\ M_a &= \begin{cases} 1 & \text{if } \hat{\mu} \neq W_a \\ 0 & \text{if } \hat{\mu} = W_a \end{cases} \quad ; \quad M_b = \begin{cases} 1 & \text{if } \hat{\mu} \neq W_b \\ 0 & \text{if } \hat{\mu} = W_b \end{cases} . \end{aligned} \tag{95}$$

Table 1 summarizes the various possibilities

Table 1: Values of the conditional entropy $H(W_a, W_b|K_n, \mathbf{U}, \mathbf{Y}, e_a, e_b, m_a, m_b)$ where $H(W_a, W_b|K_n, \mathbf{U}, \mathbf{Y}, e_a, e_b, m_a, m_b)$ abbreviates $H(W_a, W_b|K_n, \mathbf{U}, \mathbf{Y}, E_a = e_a, E_b = e_b, M_a = m_a, M_b = m_b)$, and the sign "$-$" signifies that either this case is impossible or that it implies that $\hat{\nu} = \hat{\mu}$ and it can be assumed without loss of generality that always $\hat{\nu} \neq \hat{\mu}$. For example, the case $(e_a, e_b, m_a, m_b) = (0001)$ can never occur, because $e_a = e_b = 0$ implies $W_a = W_b$, whereas $m_a \neq m_b$ implies $W_a \neq W_b$, the case $(e_a, e_b, m_a, m_b) = (0011)$ implies that $\nu = W_a = W_b$ and since $\nu$ is a function of $(K_n, \mathbf{U}, \mathbf{Y})$, we have $H(W_a, W_b|K_n, \mathbf{U}, \mathbf{Y}, 0, 0, 1, 1) = 0$.

[8] In fact, this can also serve as a converse for the SO decoder, because the capacity of the SO decoder cannot exceed that of the MO decoder. But for reasons of simplicity, and in order to characterize the worst case attack channel, we use the proof above for the SO decoder.

| $e_a$ | $e_b$ | $m_a$ | $m_b$ | description | $H(W_a, W_b \mid K_n, \mathbf{U}, \mathbf{Y}, e_a, e_b, m_a, m_b)$ |
|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | $\hat{\nu} = \hat{\mu} = W_a = W_b$ | 0 |
| 0 | 0 | 0 | 1 | - | |
| 0 | 0 | 1 | 0 | - | |
| 0 | 0 | 1 | 1 | $\hat{\nu} = W_a = W_b,\ \hat{\mu} \neq \hat{\nu}$ | 0 |
| 0 | 1 | 0 | 0 | - | |
| 0 | 1 | 0 | 1 | - | |
| 0 | 1 | 1 | 0 | $\hat{\nu} = W_a, \hat{\mu} = W_b, \hat{\mu} \neq \hat{\nu}$ | 0 |
| 0 | 1 | 1 | 1 | $\hat{\nu} = W_a \neq W_b$,$\hat{\mu}$ errs | |
| 1 | 0 | 0 | 0 | - | |
| 1 | 0 | 0 | 1 | $\hat{\nu} = W_b, \hat{\mu} = W_a, \hat{\mu} \neq \hat{\nu}$ | 0 |
| 1 | 0 | 1 | 0 | - | |
| 1 | 0 | 1 | 1 | $\hat{\nu} = W_b \neq W_a$, $\hat{\mu}$ errs | |
| 1 | 1 | 0 | 0 | $\hat{\mu} = W_a = W_b \neq \hat{\nu}$ | 0 |
| 1 | 1 | 0 | 1 | $\hat{\mu} = W_a \neq W_b$,, $\hat{\nu}$ errs | |
| 1 | 1 | 1 | 0 | $\hat{\mu} = W_b \neq W_a$, $\hat{\nu}$ errs | |
| 1 | 1 | 1 | 1 | error | |

Using the chain rule for entropies to expand $H(E_a, E_b, M_a, M_b, W_a, W_b \mid K_n \mathbf{UY})$ in two different ways, we have,

$$\begin{aligned} & H(E_a, E_b, M_a, M_b, W_a, W_b \mid K_n \mathbf{UY}) \\ = \; & H(W_a, W_b \mid K_n, \mathbf{U}, \mathbf{Y}) + H(E_a, E_b, M_a, M_b, \mid K_n, \mathbf{U}, \mathbf{Y}, W_a, W_b) \\ = \; & H(E_a, E_b, M_a, M_b, \mid K_n, \mathbf{U}, \mathbf{Y}) + H(W_a, W_b \mid K_n, \mathbf{U}, \mathbf{Y}, E_a, E_b, M_a, M_b, ). \end{aligned} \tag{96}$$

Also, $H(E_a, E_b, M_a, M_b) \leq 4$, this yields,

$$\begin{aligned} & H(W_a, W_b \mid K_n, \mathbf{U}, \mathbf{Y}) \\ = \; & H(E_a, E_b, M_a, M_b \mid K_n, \mathbf{U}, \mathbf{Y}) + H(W_a, W_b \mid K_n, \mathbf{U}, \mathbf{Y}, E_a, E_b, M_a, M_b) \\ & - H(E_a, E_b, M_a, M_b, \mid K_n, \mathbf{U}, \mathbf{Y}, W_a, W_b) \\ \leq \; & 4 + H(W_a, W_b \mid K_n, \mathbf{U}, \mathbf{Y}, E_a, E_b, M_a, M_b). \end{aligned} \tag{97}$$

Using Table 1, the term $H(W_a, W_b \mid K_n, \mathbf{U}, \mathbf{Y}, E_a, E_b, M_a, M_b)$ can be bounded as follows

$$\begin{aligned} & H(W_a, W_b \mid K_n, \mathbf{U}, \mathbf{Y}, E_a, E_b, M_a, M_b) \\ = \; & P_e \cdot 2nR \\ + \; & \Pr(\hat{\nu} = W_a \neq W_b, \hat{\mu} \text{ errs }) \cdot H(W_a, W_b \mid K_n, \mathbf{U}, \mathbf{Y}, \hat{\nu} = W_a \neq W_b, \hat{\mu} \text{ errs }) \\ + \; & \Pr(\hat{\nu} = W_b \neq W_a, \hat{\mu} \text{ errs }) \cdot H(W_a, W_b \mid K_n, \mathbf{U}, \mathbf{Y}, \hat{\nu} = W_b \neq W_a, \hat{\mu} \text{ errs }) \\ + \; & \Pr(\hat{\mu} = W_a \neq W_b, \hat{\nu} \text{ errs }) \cdot H(W_a, W_b \mid K_n, \mathbf{U}, \mathbf{Y}, \hat{\mu} = W_a \neq W_b, \hat{\nu} \text{ errs }) \\ + \; & \Pr(\hat{\mu} = W_b \neq W_a, \hat{\nu} \text{ errs }) \cdot H(W_a, W_b \mid K_n, \mathbf{U}, \mathbf{Y}, \hat{\mu} = W_b \neq W_a, \hat{\nu} \text{ errs }) \end{aligned} \tag{98}$$

$$
\begin{aligned}
&\leq P_e \cdot 2nR \\
&+ \Pr(\hat{\nu} = W_a \neq W_b, \hat{\mu} \text{ errs }) \cdot H(W_b|K_n, \mathbf{U}, \mathbf{Y}, W_a) \\
&+ \Pr(\hat{\nu} = W_b \neq W_a, \hat{\mu} \text{ errs }) \cdot H(W_a|K_n, \mathbf{U}, \mathbf{Y}, W_b) \\
&+ \Pr(\hat{\mu} = W_a \neq W_b, \hat{\nu} \text{ errs }) \cdot H(W_b|K_n, \mathbf{U}, \mathbf{Y}, W_a) \\
&+ \Pr(\hat{\mu} = W_b \neq W_a, \hat{\nu} \text{ errs }) \cdot H(W_a|K_n, \mathbf{U}, \mathbf{Y}, W_b) \\
&= P_e \cdot 2nR \\
&+ p_1 \cdot H(W_b|K_n, \mathbf{U}, \mathbf{Y}, W_a) + p_2 \cdot H(W_a|K_n, \mathbf{U}, \mathbf{Y}, W_b) \qquad (99)
\end{aligned}
$$

where (98) is because $\hat{\nu} = \psi_\nu(K_n, \mathbf{U}, \mathbf{Y})$, $\hat{\mu} = \psi_\mu(K_n, \mathbf{U}, \mathbf{Y})$, so $W_a$, for example, is known given $\hat{\nu} = W_a, K_n, \mathbf{U}, \mathbf{Y}$, and

$$
\begin{aligned}
p_1 &\triangleq \Pr(\hat{\nu} = W_a \neq W_b, \hat{\mu} \text{ errs }) + \Pr(\hat{\mu} = W_a \neq W_b, \hat{\nu} \text{ errs }), \\
p_2 &\triangleq \Pr(\hat{\nu} = W_b \neq W_a, \hat{\mu} \text{ errs }) + \Pr(\hat{\mu} = W_b \neq W_a, \hat{\nu} \text{ errs }). \qquad (100)
\end{aligned}
$$

Now, similarly to eq. (77) derived for the SO decoder, it can be argued that $p_1 = p_2$ under assumption (76). The rest of the proof is identical to the proof of the converse for the SO decoder (from eq. (78) and further on).

# 7 Extension to a size $L$ coalition

In this section, we sketch the proof for the extension of the results to the case of a larger coalition.

The extension of the direct theorem for the MO decoder (35) is quite straightforward. Assuming that $W_{a_1} = 1, \ldots, W_{a_L} = L$, we have

$$
\begin{aligned}
& P(\text{error}|\mathbf{u}\mathbf{x}_1, \ldots, \mathbf{x}_L\mathbf{y}) \\
&= \mathbf{P}^M \left\{ \exists (i_1, \ldots, i_L), i_m \neq i_k,\ i_m \geq L \forall (k,m) \text{ s.t. } (\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_L}) \in T_{\mathbf{x}_1,\ldots,\mathbf{x}_L|\mathbf{u}\mathbf{y}} \right\} \\
&\leq \max\left\{1, M^L \frac{|T_{\mathbf{x}_1,\ldots,\mathbf{x}_L|\mathbf{u}\mathbf{y}}|}{|T_{\eta_n}(\mathbf{u})|^L}\right\} \qquad (101) \\
&\leq (n+1)^{2|\mathcal{U}|\mathcal{X}|} e^{-n|\hat{I}_{\mathbf{x}_1;\mathbf{y}|\mathbf{u}} + \sum_{i=2}^{L} \hat{I}_{\mathbf{x}_i;\mathbf{x}_1,\ldots,\mathbf{x}_{i-1}\mathbf{y}|\mathbf{u}} - LR|^+} \\
&\triangleq \tilde{e}_{R,L}(\hat{P}_{\mathbf{u}\mathbf{x}_1,\ldots,\mathbf{x}_L\mathbf{y}}), \qquad (102)
\end{aligned}
$$

where (101) is because there are $\begin{pmatrix} M \\ L \end{pmatrix} \leq M^L$ possibilities of choosing an $L$-tuple among the $M$ vectors and $\frac{|T_{\mathbf{x}_1,\ldots,\mathbf{x}_L|\mathbf{u}\mathbf{y}}|}{|T_{\eta_n}(\mathbf{u})|^L}$ is the probability that $L$ i.i.d. vectors uniformly distributed over $T_{\eta_n}(\mathbf{u})$ would lie within $T_{\mathbf{x}_1,\ldots,\mathbf{x}_L|\mathbf{u}\mathbf{y}}$.

The rest of the derivation of the direct part can be carried out similarly to the proof of Theorem 1 (see (48)-(65)), with $\tilde{e}_{R,L}(\hat{P}_{\mathbf{u}\mathbf{x}_1,\ldots,\mathbf{x}_L\mathbf{y}})$ replacing $e_R(\hat{P}_{\mathbf{u}\mathbf{x}_1\mathbf{x}_2\mathbf{y}})$, and it is easily verified that indeed, (35) is the rate at which the error exponent vanishes, and therefore it serves as a lower bound on the capacity.

As for the converse part, similarly to the derivation of (83), one derives the inequality,

$$
R \leq \epsilon_{n,R} + \frac{1}{nL} \sum_{i=1}^{L} I(\mathbf{X}_{a_i}; \mathbf{Y}|K_n\mathbf{U}), \qquad (103)
$$

where $\mathbf{X}_{a_1},\ldots,\mathbf{X}_{a_L}$ are the fingerprinted copies observed by the attacker, for the MO decoder (with $\epsilon_{n,R}=\frac{L^2}{n}+LRP_e$). Now, note that

$$\begin{aligned}\frac{1}{nL}\sum_{i=1}^{L} I(\mathbf{X}_{a_i};\mathbf{Y}|K_n\mathbf{U}) &\leq \frac{1}{nL}\sum_{i=1}^{L} I(\mathbf{X}_{a_i};\mathbf{X}_{a_1},\ldots,\mathbf{X}_{a_{i-1}}\mathbf{Y}|K_n\mathbf{U}) \\ &= \frac{1}{nL}\sum_{i=1}^{L} I(\mathbf{X}_{a_i};\mathbf{Y}|K_n\mathbf{U}\mathbf{X}_{a_1},\ldots,\mathbf{X}_{a_{i-1}}) \qquad (104) \\ &= \frac{1}{nL} I(\mathbf{X}_{a_1},\ldots,\mathbf{X}_{a_L};\mathbf{Y}|K_n\mathbf{U}) \qquad (105)\end{aligned}$$

where (104) follows from (84) applied to $(\mathbf{U},K_n),\mathbf{X}_{a_i},(\mathbf{X}_{a_1},\ldots,\mathbf{X}_{a_{i-1}}),\mathbf{Y}$ in the roles of $A,B,C,D$ respectively) and (105) follows from the chain rule for the mutual information. The rest of the proof is very similar to that of a coalition of two members.

# A Proofs of Lemmas

**Proof of Lemma 2**

Let $P'_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}$ be a given attack channel. Denote $P^{inv}_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}(\mathbf{y}|\tilde{\mathbf{x}}\mathbf{x})=P'_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}(\mathbf{y}|\mathbf{x}\tilde{\mathbf{x}})$. Since by definition $P_{\mathbf{X}_a\mathbf{X}_b|\mathbf{U}K_n}=P_{\mathbf{X}_b\mathbf{X}_a|\mathbf{U}K_n}$, by symmetry, we have

$$I_{P_{\mathbf{U}K_n\mathbf{X}_a\mathbf{X}_b}P'_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}}(\mathbf{X}_a\mathbf{X}_b;\mathbf{Y}|K_n,\mathbf{U})=I_{P_{\mathbf{U}K_n\mathbf{X}_a\mathbf{X}_b}P^{inv}_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}}(\mathbf{X}_a\mathbf{X}_b;\mathbf{Y}|K_n,\mathbf{U}). \qquad (106)$$

Now note that $I_{P_{\mathbf{U}K_n\mathbf{X}_a\mathbf{X}_b}P_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}}(\mathbf{X}_a\mathbf{X}_b;\mathbf{Y}|K_n,\mathbf{U})$ is convex in $P_{\mathbf{Y}|\mathbf{X}_b\mathbf{X}_a}$ for fixed $P_{\mathbf{X}_a\mathbf{X}_bK_n\mathbf{U}}$, thus if one defines $\bar{P}_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}=\frac{1}{2}P'_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}+\frac{1}{2}P^{inv}_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}$,

$$\begin{aligned}&I_{P_{\mathbf{U}K_n\mathbf{X}_a\mathbf{X}_b}\bar{P}_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}}(\mathbf{X}_a\mathbf{X}_b;\mathbf{Y}|K_n,\mathbf{U}) \\ \leq\ &\frac{1}{2}I_{P_{\mathbf{U}K_n\mathbf{X}_a\mathbf{X}_b}P'_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}}(\mathbf{X}_a\mathbf{X}_b;\mathbf{Y}|K_n,\mathbf{U})+\frac{1}{2}I_{P_{\mathbf{U}K_n\mathbf{X}_a\mathbf{X}_b}P^{inv}_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}}(\mathbf{X}_a\mathbf{X}_b;\mathbf{Y}|K_n,\mathbf{U}) \\ =\ &I_{P_{\mathbf{U}K_n\mathbf{X}_a\mathbf{X}_b}P'_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}}(\mathbf{X}_a\mathbf{X}_b;\mathbf{Y}|K_n,\mathbf{U}), \qquad (107)\end{aligned}$$

which proves that the attacker would always prefer to use the symmetric channel $\bar{P}_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}$ (which, also meets the distortion constraint), over $P'_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}$. Since this is true for every attack channel $P'_{\mathbf{Y}|\mathbf{X}_a\mathbf{X}_b}$, this concludes the proof of (86).

□

**Proof of Lemma 3**

Now we use the assumption that the encoder uses random coding with codewords that are i.i.d. given $\mathbf{U}$. We shall write the expectation $E\hat{I}_{\mathbf{X}_a\mathbf{X}_b;\mathbf{Y}|\mathbf{U}}$ under this assumption (denoting for convenience $T_{\eta_n}(\mathbf{u}) \triangleq T_{\mathbf{u}}(\eta_n(\mathbf{u})))$

$$\begin{aligned}
& E\hat{I}_{\mathbf{X}_a\mathbf{X}_b;\mathbf{Y}|\mathbf{U}} \\
&= \sum_{\mathbf{u}} P_U^n(\mathbf{u}) \sum_{\mathbf{x}\mathbf{x}'\in T_{\eta_n}(\mathbf{u})} \frac{1}{|T_{\eta_n}(\mathbf{u})|^2} \sum_{\mathbf{y}\in T_{\rho_n}(\mathbf{x}\mathbf{x}')} \frac{1}{|T_{\rho_n}(\mathbf{x}\mathbf{x}')|} \hat{I}_{\mathbf{x}\mathbf{x}';\mathbf{y}|\mathbf{u}} \\
&= \sum_{\mathbf{u}} P_U^n(\mathbf{u}) \sum_{\mathbf{x}\mathbf{x}'\in T_{\eta_n}(\mathbf{u})} \frac{1}{|T_{\eta_n}(\mathbf{u})|^2} \sum_{T_{\mathbf{y}|\mathbf{u}\mathbf{x}\mathbf{x}'}\subseteq T_{\rho_n}(\mathbf{x}\mathbf{x}')} \frac{|T_{\mathbf{y}|\mathbf{u}\mathbf{x}\mathbf{x}'}|}{|T_{\rho_n}(\mathbf{x}\mathbf{x}')|} \hat{I}_{\mathbf{x}\mathbf{x}';\mathbf{y}|\mathbf{u}}. \qquad (108)
\end{aligned}$$

Next, define $B = |\mathcal{U}||\mathcal{X}^2||\mathcal{Y}|$ and let $g(n) = \frac{n+1}{n^{B+2}}$ we have

$$\begin{aligned}
& \sum_{T_{\mathbf{y}|\mathbf{u}\mathbf{x}\mathbf{x}'}\subseteq T_{\rho_n}(\mathbf{x}\mathbf{x}')} \frac{|T_{\mathbf{y}|\mathbf{u}\mathbf{x}\mathbf{x}'}|}{|T_{\rho_n}(\mathbf{x}\mathbf{x}')|} \hat{I}_{\mathbf{x}\mathbf{x}';\mathbf{y}|\mathbf{u}} \\
&\le \sum_{T_{\mathbf{y}|\mathbf{u}\mathbf{x}\mathbf{x}'}\subseteq T_{\rho_n}(\mathbf{x}\mathbf{x}'):\frac{|T_{\mathbf{y}|\mathbf{u}\mathbf{x}\mathbf{x}'}|}{|T_{\rho_n}(\mathbf{x}\mathbf{x}')|}\ge g(n)} \frac{|T_{\mathbf{y}|\mathbf{u}\mathbf{x}\mathbf{x}'}|}{|T_{\rho_n}(\mathbf{x}\mathbf{x}')|} \hat{I}_{\mathbf{x}\mathbf{x}';\mathbf{y}|\mathbf{u}} + g(n)(n+1)^B \log|\mathcal{Y}| \\
&\le \max_{T_{\mathbf{y}|\mathbf{u}\mathbf{x}\mathbf{x}'}\subseteq T_{\rho_n}(\mathbf{x}\mathbf{x}'):\frac{|T_{\mathbf{y}|\mathbf{u}\mathbf{x}\mathbf{x}'}|}{|T_{\rho_n}(\mathbf{x}\mathbf{x}')|}\ge g(n)} \hat{I}_{\mathbf{x}\mathbf{x}';\mathbf{y}|\mathbf{u}} + g(n)(n+1)^B \log|\mathcal{Y}| \qquad (109) \\
&\le \max_{\left\{T_{\mathbf{y}|\mathbf{u}\mathbf{x}\mathbf{x}'}\subseteq T_{\rho_n}(\mathbf{x}\mathbf{x}'):\hat{I}_{\mathbf{u};\mathbf{y}|\mathbf{x}\mathbf{x}'}\le\frac{B'\log n}{n}\right\}} \hat{I}_{\mathbf{x}\mathbf{x}';\mathbf{y}|\mathbf{u}} + \frac{(n+1)^{B+1}}{n^{B+2}} \log|\mathcal{Y}|, \qquad (110)
\end{aligned}$$

where (109) is because $\sum_{T_{\mathbf{y}|\mathbf{u}\mathbf{x}\mathbf{x}'}\subseteq T_{\rho_n}(\mathbf{x}\mathbf{x}')} \frac{|T_{\mathbf{y}|\mathbf{u}\mathbf{x}\mathbf{x}'}|}{|T_{\rho_n}(\mathbf{x}\mathbf{x}')|} = 1$, and in the last step $B' \triangleq \frac{B(B+2)}{|\mathcal{U}|}$ and

$$\frac{|T_{\mathbf{y}|\mathbf{u}\mathbf{x}\mathbf{x}'}|}{|T_{\rho_n}(\mathbf{x}\mathbf{x}')|} \ge g(n) \Rightarrow \hat{I}_{\mathbf{u};\mathbf{y}|\mathbf{x}\mathbf{x}'} \le \frac{B}{n|\mathcal{U}|} \log\frac{n+1}{g(n)}. \qquad (111)$$

Since $\mathbb{P}_n(\mathcal{Y}, \hat{P}_{\mathbf{u}\mathbf{x}\mathbf{x}'}) \subset \mathcal{P}(\mathcal{Y}|\mathcal{U}\times\mathcal{X}^2)$, we have

$$\begin{aligned}
& \max_{\left\{T_{\mathbf{y}|\mathbf{u}\mathbf{x}\mathbf{x}'}\subseteq T_{\rho_n}(\mathbf{x}\mathbf{x}'):\hat{I}_{\mathbf{u};\mathbf{y}|\mathbf{x}\mathbf{x}'}\le\frac{B'\log n}{n}\right\}} \hat{I}_{\mathbf{x}\mathbf{x}';\mathbf{y}|\mathbf{u}} \\
&\le \max_{W\in\Psi_{\rho_n}(\hat{P}_{\mathbf{u}\mathbf{x}\mathbf{x}'}):\ I_{\hat{P}_{\mathbf{u}\mathbf{x}\mathbf{x}'}W}(U;Y|XX')\le\frac{B'\log n}{n}} I_{\hat{P}_{\mathbf{u}\mathbf{x}\mathbf{x}'}W}(XX';Y|U), \qquad (112)
\end{aligned}$$

where

$$\begin{aligned}
& \Psi_{\rho_n}(\hat{P}_{\mathbf{u}\mathbf{x}\mathbf{x}'}) \\
\triangleq\ & \left\{W\in\mathcal{P}(\mathcal{Y}|\mathcal{U}\times\mathcal{X}^2) : \sum_{u\in\mathcal{U}} \hat{P}_{\mathbf{u}|\mathbf{x}\mathbf{x}'}(u|xx')W(y|uxx') = \rho_n(\hat{P}_{\mathbf{x}\mathbf{x}'})(y|xx')\ \forall(y,x,x')\right\} \qquad (113)
\end{aligned}$$

Loosely speaking, the condition $I_{\hat{P}_{\mathbf{u}\mathbf{x}\mathbf{x}'}W}(U;Y|XX') \le \frac{B'\log n}{n}$ implies that $U \leftrightarrow XX' \leftrightarrow Y$ is "almost" a Markov chain. Using continuity arguments, we prove in the following lemma, that it can indeed be approximated by a Markov chain.

**Lemma 4.** *There exists a sequence* $\epsilon_n > 0$ *such that* $\lim_{n\to\infty}\epsilon_n = 0$ *and for all* $\hat{P}_{\mathbf{uxx}'} \in \mathcal{P}(\mathcal{U}\times\mathcal{X}^2)$

$$\begin{aligned}
&\max_{W\in\Psi_{\rho_n}(\hat{P}_{\mathbf{uxx}'}): I_{\hat{P}_{\mathbf{uxx}'}W}(U;Y|XX')\le\frac{B'\log n}{n}} I_{\hat{P}_{\mathbf{uxx}'}W}(U;Y|XX') \\
\le\ & \max_{W\in\Psi_{\rho_n}(\hat{P}_{\mathbf{uxx}'}): I_{\hat{P}_{\mathbf{uxx}'}W}(U;Y|XX')=0} I_{\hat{P}_{\mathbf{uxx}'}W}(U;Y|XX') + \epsilon_n && (114)\\
=\ & I_{\hat{P}_{\mathbf{uxx}'}\rho_n(\hat{P}_{\mathbf{xx}'})}(XX';Y|U) + \epsilon_n. && (115)
\end{aligned}$$

The proof of Lemma 4 appears in the sequel of Appendix A.

Gathering (108)-(115) yields that there exists a sequence $\delta_n \to 0$ such that

$$\begin{aligned}
& E\hat{I}_{\mathbf{X}_a\mathbf{X}_b;\mathbf{Y}|\mathbf{U}} \\
\le\ & \sum_{\mathbf{u}} P_U^n(\mathbf{u}) \sum_{\mathbf{x},\mathbf{x}'\in T_{\eta_n}(\mathbf{u})} \frac{1}{|T_{\eta_n}(\mathbf{u})|^2} I_{\hat{P}_{\mathbf{uxx}'}\rho_n(\hat{P}_{\mathbf{xx}'})}(XX';Y|U) + \delta_n \\
=\ & \sum_{\mathbf{u}} P_U^n(\mathbf{u}) \sum_{\mathbf{x}\in T_{\eta_n}(\mathbf{u})} \frac{1}{|T_{\eta_n}(\mathbf{u})|} \sum_{T_{\mathbf{x}'|\mathbf{xu}}\subseteq T_{\eta_n}(\mathbf{u})} \frac{|T_{\mathbf{x}'|\mathbf{xu}}|}{|T_{\eta_n}(\mathbf{u})|} I_{\hat{P}_{\mathbf{uxx}'}\rho_n(\hat{P}_{\mathbf{xx}'})}(XX';Y|U). && (116)
\end{aligned}$$

Now, fix $\mathbf{x}\in T_{\eta_n}(\mathbf{u})$, let $D = |\mathcal{U}||\mathcal{X}|^2$ and define $h(n) = \frac{n+1}{n^{D+2}}$, we have

$$\begin{aligned}
& \sum_{T_{\mathbf{x}'|\mathbf{xu}}\subseteq T_{\eta_n}(\mathbf{u})} \frac{|T_{\mathbf{x}'|\mathbf{xu}}|}{|T_{\eta_n}(\mathbf{u})|} I_{\hat{P}_{\mathbf{uxx}'}\rho_n(\hat{P}_{\mathbf{xx}'})}(XX';Y|U) \\
\le\ & \sum_{T_{\mathbf{x}'|\mathbf{xu}}\subseteq T_{\eta_n}(\mathbf{u}): \frac{|T_{\mathbf{x}'|\mathbf{xu}}|}{|T_{\eta_n}(\mathbf{u})|}\ge h(n)} \frac{|T_{\mathbf{x}'|\mathbf{xu}}|}{|T_{\eta_n}(\mathbf{u})|} I_{\hat{P}_{\mathbf{uxx}'}\rho_n(\hat{P}_{\mathbf{xx}'})}(XX';Y|U) + h(n)(n+1)^D\log|\mathcal{Y}| \\
\le\ & \max_{T_{\mathbf{x}'|\mathbf{xu}}\subseteq T_{\eta_n}(\mathbf{u}): \frac{|T_{\mathbf{x}'|\mathbf{xu}}|}{|T_{\eta_n}(\mathbf{u})|}\ge h(n)} I_{\hat{P}_{\mathbf{uxx}'}\rho_n(\hat{P}_{\mathbf{xx}'})}(XX';Y|U) + h(n)(n+1)^D\log|\mathcal{Y}| && (117)\\
\le\ & \max_{\left\{T_{\mathbf{x}'|\mathbf{ux}}\in T_{\eta_n}(\mathbf{u}): \hat{I}_{\mathbf{x};\mathbf{x}'|\mathbf{u}}\le\frac{D'\log n}{n}\right\}} I_{\hat{P}_{\mathbf{uxx}'}\rho_n(\hat{P}_{\mathbf{xx}'})}(XX';Y|U) + \frac{(n+1)^{D+1}}{n^{D+2}}\log|\mathcal{Y}|, && (118)
\end{aligned}$$

where (117) is because $\sum_{T_{\mathbf{x}'|\mathbf{ux}}\subseteq T_{\eta_n}(\mathbf{u})} \frac{|T_{\mathbf{x}'|\mathbf{ux}}|}{|T_{\eta_n}(\mathbf{u})|} = 1$, and in the last step $D' \triangleq \frac{D(D+2)}{|\mathcal{U}|}$ and

$$\frac{|T_{\mathbf{x}'|\mathbf{ux}}|}{|T_{\eta_n}(\mathbf{u})|} \ge h(n) \Rightarrow \hat{I}_{\mathbf{x};\mathbf{x}'|\mathbf{u}} \le \frac{D}{n|\mathcal{U}|}\log\frac{n+1}{h(n)}. \tag{119}$$

Note that, again, the condition $\hat{I}_{\mathbf{x}';\mathbf{x}|\mathbf{u}} = I_{\hat{P}_{\mathbf{uxx}'}}(X;X'|U) \le \frac{D'\log n}{n}$ implies that $X - U - X'$ is "almost" a Markov chain, and we can use continuity arguments that are similar to those used in the chain of inequalities (112)-(115) to obtain that there exists a sequence $\delta_n' \to 0$ such that for every $\mathbf{ux}$,

$$\begin{aligned}
& \max_{\left\{T_{\mathbf{x}'|\mathbf{ux}}\in T_{\eta_n}(\mathbf{u}): \hat{I}_{\mathbf{u};\mathbf{y}|\mathbf{xx}'}\le\frac{D'\log n}{n}\right\}} I_{\hat{P}_{\mathbf{uxx}'}\rho_n(\hat{P}_{\mathbf{xx}'})}(XX';Y|U) \\
\le\ & I(\mathbf{u},\eta_n,\rho_n) + \delta_n', && (120)
\end{aligned}$$

where $I(\mathbf{u}, \eta_n, \rho_n)$ designates the mutual information $I_P(XX'; Y|U)$ with $P \in \mathcal{P}(\mathcal{U}\mathcal{X}^2\mathcal{Y})$ defined as follows:

- $U \sim \hat{P}_{\mathbf{u}}$,
- $P_{X|U} = \eta_n(\mathbf{u}), X \leftrightarrow U \leftrightarrow \tilde{X}, P_{\tilde{X}|U} = P_{X|U}$,
- $P_{Y|X\tilde{X}} = \rho_n(P_{X\tilde{X}}), U \leftrightarrow (X, \tilde{X}) \leftrightarrow Y$.

So, the CC capacity is essentially upper bounded by

$$\frac{1}{2} \max_{\eta_n \in \mathcal{N}(D_1)} \min_{\rho_n \in \mathcal{R}_n(D_2)} \sum_{\mathbf{u}} P_U^n(\mathbf{u}) I(\mathbf{u}, \eta_n, \rho_n). \tag{121}$$

It is therefore left to upper bound $\sum_{\mathbf{u}} P_U^n(\mathbf{u}) I(\mathbf{u}, \eta_n, \rho_n)$. To do that, take the function $h_2(n) = \frac{(|\mathcal{U}|+1)\log(n+1)}{n}$, and note that

$$\begin{aligned}
&\sum_{\mathbf{u}} P_U^n(\mathbf{u}) I(\mathbf{u}, \eta_n, \rho_n) \\
\leq\ & \max_{\mathbf{u}:\, D(\hat{P}_{\mathbf{u}}\|P_U) \leq h_2(n)} I(\mathbf{u}, \eta_n, \rho_n) + \sum_{\mathbf{u}:\, D(\hat{P}_{\mathbf{u}}\|P_U) > h_2(n)} P_U^n(\mathbf{u}) \log|\mathcal{Y}| \\
\leq\ & \max_{\mathbf{u}:\, D(\hat{P}_{\mathbf{u}}\|P_U) \leq h_2(n)} I(\mathbf{u}, \eta_n, \rho_n) + \frac{1}{(n+1)} \log|\mathcal{Y}|.
\end{aligned} \tag{122}$$

Pinsker's inequality implies that

$$\|P_1 - P_2\| \leq \sqrt{(2\ln 2) D(P_1\|P_2)}, \tag{123}$$

thus, if $D(\hat{P}_{\mathbf{u}}\|P_U) \leq h_2(n)$, we have $\|\hat{P}_{\mathbf{u}} - P_U\| \leq \sqrt{(2\ln 2)(|\mathcal{U}|+1)\frac{\log(n+1)}{n}} \triangleq \ell(n)$. Therefore, the CC capacity is essentially upper bounded by

$$\frac{1}{2} \max_{\eta_n \in \mathcal{N}(D_1)} \min_{\rho_n \in \mathcal{R}_n(D_2)} \max_{\mathbf{u}:\, \|\hat{P}_{\mathbf{u}} - P_U\| \leq \ell(n)} I(\mathbf{u}, \eta_n, \rho_n). \tag{124}$$

Now, recall that the capacity is defined w.r.t. sequences of smooth codes (see Definition 3). This means that there exists a conditional measure $P_{X|U} \in \mathcal{P}(\mathcal{X}|\mathcal{U})$ such that for all $\epsilon > 0$, $\exists n_0(\epsilon)$ such that for all $n \geq n_0(\epsilon)$, whenever $\|\hat{P}_{\mathbf{u}} - P_U\| \leq \ell(n)$, one has $\|\hat{P}_{\mathbf{u}} \times \eta_n(\mathbf{u}) - P_U \times P_{X|U}\| \leq \epsilon$. Hence, from the uniform continuity of $I_P(XX'; Y|U)$ w.r.t. $P$, it follows that the CC capacity is given by $C(P_U, D_1, D_2)$.

□

**Proof of Lemma 4**

To see why the lemma holds, we apply Pinsker's inequality (123) and argue that of all $(u, x, x') \in \mathcal{U} \times \mathcal{X}^2$, and every pmf $\tilde{P} \in \mathcal{P}(\mathcal{U} \times \mathcal{X}^2 \times \mathcal{Y})$:

$$\sum_{y \in \mathcal{Y}} \left| \tilde{P}(y|uxx') - \tilde{P}(y|xx) \right| \leq \sqrt{(2\ln 2) D(\tilde{P}(\cdot|uxx')\|\tilde{P}(\cdot|xx'))} \triangleq \sqrt{(2\ln 2) I_{\tilde{P}}(u; Y|xx')}. \tag{125}$$

Multiplying by $\tilde{P}(uxx')$ and summing over $(u,x,x') \in \mathcal{U} \times \mathcal{X}^2$, we get,

$$\begin{aligned}
&\sum_{(u,x,x',y)\in\mathcal{U}\times\mathcal{X}^2\times\mathcal{Y}} \left|\tilde{P}(u,x,x')\tilde{P}(y|uxx') - \tilde{P}(u,x,x')\tilde{P}(y|xx')\right| \\
\leq\ & \sum_{(u,x,x')\in\mathcal{U}\times\mathcal{X}^2} \tilde{P}(u,x,x')\sqrt{(2\ln 2) I_{\tilde{P}}(u;Y|xx')} \\
&\leq \sqrt{(2\ln 2) I_{\tilde{P}}(U;Y|XX')}, \qquad (126)
\end{aligned}$$

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

$$\begin{aligned}
&I(U;Y|XX') \leq \frac{B'\log n}{n} \Longrightarrow \\
&\sum_{(u,x,x',y)\in\mathcal{U}\times\mathcal{X}^2\times\mathcal{Y}} \left|\tilde{P}(u,x,x')\tilde{P}(y|uxx') - \tilde{P}(u,x,x')\tilde{P}(y|xx')\right| \leq \sqrt{(2\ln 2)\frac{B'\log n}{n}} \qquad (127)
\end{aligned}$$

and the desired claim (114) follows from the uniform continuity of $I_{\tilde{P}}(XX';Y|U)$ w.r.t. $\tilde{P}$. $\square$

# References


[1] A. Barg, G. R. Blakley, and G. Kabatiansky, "Good digital fingerprinting codes," *Proc. ISIT 2001*, p. 161, Washington, D.C., June 2001.

[2] A. Barg, G. R. Blakley, and G. Kabatiansky, "Digital fingerprinting codes: problem statements, constructions, identification of traitors," *IEEE Trans. Inform. Theory*, vol. 49, no. 4, pp. 852-865, Apr. 2003.

[3] D. Boneh and J. Shaw, "Collusion–secure fingerprinting for digital data," *IEEE Trans. Inform. Theory*, vol. 44, no. 5, pp. 1897–1905, September 1998.

[4] B. Chor, A. Fiat, M. Naor, and B. Pinkas, "Tracing traitors," *IEEE Trans. Inform. Theory*, vol. 46, no. 3, pp. 893–910, May 2000.

[5] G. Cohen, S. Encheva, and G. Zémor, "Copyright protection for digital data," *IEEE Communication Letters*, vol. 4, no. 5, pp. 158–160, May 2000.

[6] A. Cohen and A. Lapidoth, "The Gaussian watermarking game," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1639-1667, June 2002.

[7] A. Cohen and A. Lapidoth, "The capacity of the vector Gaussian watermarking game," in *Proc. ISIT 2001*, p. 5, 2001.

[8] T. M. Cover, "Enumerative source coding," *IEEE Trans. Inform. Theory*, vol. IT-19, no. 1, pp. 73–77, Jan. 1973.

[9] I. Csiszár and J. Körner, *Coding theorems for discrete memoryless systems.* New York: Academic, 1981.

[10] J. Dittmann, A. Behr, M. Stabenau, P. Schmitt, J. Schwenk, and J. Ueberberg, Combining digital watermarks and collusion secure fingerprints for digital images," preprint, 2001.

[11] Y. S. Liu and B. L. Hughes, "A new universal random coding bound for hte multiple access channel," *IEEE Trans. Inform. Theory*, vol. 42, no. 2, pp. 376–386, March 1996.

[12] J. Löfvenberg, *Codes for digital fingerprints*, Ph.D. dissertation, Department of Electrical Engineering, Linköping University, Sweden, June 2001.

[13] J. Löfvenberg and N. Wilberg, "Random codes for digital fingerprinting," Report Reg nr:LiTH-ISY-R-2059, Department of Electrical Engineering, Linköping University, Sweden, June 2001.

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

[15] P. Moulin and A. Briassouli, "The Gaussian fingerprinting game," *Proc. 2002 Conference on Information Sciences and Systems*, Princeton University, March 20–22, 2002.

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

[17] P. Moulin and J. A. O'Sullivan, "Information-theoretic analysis of information hiding," *IEEE Trans. Inform. Theory*, vol. 49, no. 3, March 2003.

[18] J. A. O'Sullivan, P. Moulin, and J. M. Ettinger, "Information–theoretic analysis of steganography," in *Proc. ISIT '98*, M.I.T., Cambridge, MA, p. 297, August 1998.

[19] R. Safavi–Naini and Y. Wang, "Collusion secure $q$–ary fingerprinting for perceptual content," preprint 2001.

[20] R. Safavi–Naini and Y. Wang, "Sequential traitor tracing," *IEEE Trans. Inform. Theory*, vol. 49, no. 5, pp. 1319-1326, May 2003.

[21] A. Silverberg, J. Staddon, and J.L. Walker, "Application of list decoding to tracing traitors," *IEEE Trans. Inform. Theory*, vol. 49, no. 5, pp. 1312-1318, May 2003.

[22] A. Somekh-Baruch and N. Merhav, " On the error exponent and capacity games of private watermarking systems," CCIT 346 EE 1280, Technion-I.I.T., EE Dept., June 2001. Available at `http://www.ee.technion.ac.il/ merhav`

[23] A. Somekh-Baruch and N. Merhav, "On the error exponent and capacity games of private watermarking systems," *IEEE Trans. Inform. Theory*, vol. 49, no. 3, pp. 537-562, March 2003.

[24] A. Somekh-Baruch and N. Merhav, "On the capacity game of public watermarking systems," accepted to *IEEE Trans. Inform. Theory.*

[25] Y. Steinberg and N. Merhav, "Identification in the presence of side information with application to watermarking," *IEEE Trans. Inform. Theory*, vol. 47, no. 4, pp. 1410-1422, May 2001.

[26] J. K. Su, J. J. Eggers, and B. Girod, "Capacity of digital watermarks subjected to an optimal collusion attack," *Proc. EUSIPCO 2000.*

[27] K. Su, D. Kundur, and D. Hatzinakos, "A novel approach to collusion–resistant video watermarking," preprint, 2001.