On the Capacity Game of Public Watermarking Systems

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

Watermarking codes are analyzed as a game between two players: an information hider, and a decoder, on the one hand, and an attacker on the other hand. It is assumed that the covertext (the original data within which the message is hidden) is drawn from a memoryless stationary source and its realization is available at the information hider only. The information hider is allowed to cause some tolerable level of distortion to the covertext, and the resulting distorted data can suffer some additional amount of distortion caused by an attacker who aims at erasing the message. Motivated by a worst-case approach, we assume that the attacker is informed of the hiding strategy taken by the information hider and the decoder, while they are uninformed of the attacking scheme. The capacity is expressed as the limit of a sequence of single-letter expressions under the assumption that the encoder uses constant composition codes.

Index Terms- Capacity, coding with side information, information hiding, MMI decoder, public watermarking, randomized code, steganography, universal decoding, watermarking.

1 Introduction

In the recent years, due to the rapid development of the Internet and simple devices for data storage and duplication, watermarking codes have been attracting a lot of attention. Watermarking codes are used to hide a message in digital data, referred to as *covertext*, such as image, text, or an audio signal, so as to enable copyright protection or data encryption. A good watermarking code is one that is robust, namely, the coded message is decodable even if the data suffers a certain degree of distortion caused by an attacker. The attack can be either intentional or a result of standard data processing, e.g., data compression or filtering. On the other hand, the amount of distortion caused by the encoder should not exceed a certain distortion level. Hence, the problem of designing a watermarking code can be modelled as a game between an encoder-decoder and an attacker. A worst-case approach leads to an assumption that the attacker is 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 (implemented by the key) as a means of protection against malicious attackers. Note that here, unlike in the classical theory of coded communication, the notion

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

of random coding not only serves as tool in a proof technique, but is an inherent part of the model. 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.

O'Sullivan, Moulin and Ettinger [15] and Moulin and O'Sullivan [13] characterized the watermarking codes from the information-theoretic point of view as a capacity game between the information hider and the attacker. They presented two models of watermarking systems: The *private* game, where the covertext is available to both the encoder and decoder, and the *public* game, where the covertext is available to the encoder only. The private and the public watermarking games are considered in [13] under the following assumptions: (a) The encoder and the attacker should meet some expected distortion constraints. (b) The attacker is allowed to use only discrete memoryless channels (DMCs) or block-wise memoryless channels. (c) The DMC chosen by the attacker is known to the decoder which can then apply maximum likelihood decoding.

In [17], the capacity and error exponents games of private watermarking games are investigated, where a more realistic assumption is considered, that the attack channel is completely general and 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. Single-letter expressions for the error exponent and capacity of the private game are obtained.

Cohen and Lapidoth [3] derived a capacity formula of the watermarking game for Gaussian covertext and an almost-sure squared-error distortion constraint. It was shown in [3] that for this model, the capacities of the public and private watermarking games are the same. This is in some analogy to Costa's result [5] on channel coding with side information under the Gaussian - quadratic regime. For other related work, see, e.g., [1], [2], [4], [7], [8], [14], [16] and references therein.

This paper extends the public game model of [13] in the sense that we drop the questionable assumptions that the decoder knows the attack channel and that this channel is memoryless or blockwise memoryless. In our model (as in [17],[18]), *any* attack channel that satisfies a certain maximum distortion constraint is allowed. We also extend the public game results of [3] in the sense that our results apply to any discrete memoryless source of finite alphabet, and any distortion measure. From the technical point of view, this extension of [3] is non-trivial since the converse part cannot rely on techniques that are valid for the private setting because, in general, the public and the private game capacities may differ (in contrast to [3] where they coincide).

The capacity C of the public game w.r.t. constant composition (CC) codes, is defined as the highest achievable rate under the assumption that the encoder uses constant composition codes. The CC capacity is expressed as the limit of a non-decreasing sequence of single-letter expressions $\{C_\ell\}_{\ell\geq 1}$ as ℓ tends infinity. It is shown that $C - \frac{h\log\ell}{\ell} \leq C_\ell \leq C$, where h is a constant that depends only on the distortion measure and the alphabet size of the output of the attack. A strategy taken by the encoder that guarantees achievability of the rate C_ℓ is characterized, and the optimal strategies taken by the attacker and the decoder are also characterized. As $C_\ell \geq C - \frac{h\log\ell}{\ell}$, the term $\frac{h\log\ell}{\ell}$ quantifies the difference between the capacity and the reliable embedding rate that is guaranteed by the encoder's strategy. Since $\frac{\log\ell}{\ell}$ can be made arbitrarily small (by working with sufficiently large ℓ), in this sense, the derived formula is 'almost' a single-letter expression.

The outline of the paper is as follows: In Section 2, some notation conventions are defined. A statement of the problem is given in Section 3. Section 4 devoted to a presentation of the results, and the proofs of the lower and upper bounds appear in Sections 5 and 6 respectively.

2 Notation

Henceforth, we adopt the following notation conventions. Random variables will be denoted by capital letters while their realizations will be denoted by the respective lower case letters. Random vectors of dimension n will be denoted by boldface letters. Thus, for example, if $\mathbf{X}$ denotes a random vector $(X_1, \ldots, X_n)$, then $\mathbf{x} = (x_1, \ldots, x_n)$ will designate a specific sample value of $\mathbf{X}$. The alphabet of a scalar random variable (RV) X will be designated by a caligraphic letter $\mathcal{X}$. For integers m, k such that $m \geq k$, the notation X_k^m will abbreviate $(X_k, X_{k+1}, ..., X_m)$. 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 complement of a set A will be denoted $\bar{A}$.

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., the transition matrices P such that $P(x|u) \geq 0$, $\sum_{x' \in \mathcal{X}} P(x'|u) = 1$, $\forall (u, x) \in \mathcal{U} \times \mathcal{X}$.

The pmf of a random variable X will be denoted with an appropriate subscript, e.g., P_X, similarly, the joint pmf of the RV's X and Y will be denoted P_{XY}, the pmf of a random vector $\mathbf{X}$ will be denoted by $P_{\mathbf{X}}$, and the conditional pmf of $\mathbf{Y}$ given $\mathbf{X}$ will be denoted by $P_{\mathbf{Y}|\mathbf{X}}$. If $\mathbf{X}$ is composed of n i.i.d. RV's with marginal P_X will shall denote $P_{\mathbf{X}} = P_X^n$

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 the minimum of a function over an empty set is ∞ and the maximum is $-\infty$. 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. For two functions $f_i : \mathbb{R} \to \mathbb{R}$, $i = 1, 2$, the notation $f_1(n) = O(f_2(n))$ will express the fact that the functions are of the same asymptotic order, i.e., $0 < \lim_{n\to\infty} \frac{f_1(n)}{f_2(n)} < \infty$. The notation $|t|^+$, where $t \in \mathbb{R}$, will designate $\max\{0, t\}$.

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 random empirical pmf induced by the random vector $\mathbf{X}$ will be denoted by $\hat{P}_{\mathbf{X}}$.

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.,

$$\mathbb{P}_n(\mathcal{X}) = \left\{\hat{P}_{\mathbf{x}}\right\}_{\mathbf{x}\in\mathcal{X}^n}. \tag{1}$$

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

For a given empirical pmf $\hat{P} \in \mathbb{P}_n(\mathcal{A})$, define the set $\mathbb{P}_n(\mathcal{B}, \hat{P})$ of conditional empirical pmf's, $P \in \mathcal{P}(\mathcal{B}|\mathcal{A})$ such that $n\hat{P}(a)P(b|a)$ is an integer for all $(a, b) \in \mathcal{A} \times \mathcal{B}$. For $\tilde{P} \in \mathbb{P}_n(\mathcal{B}, \hat{P}_{\mathbf{x}})$, we shall also use the notation $T_{\tilde{P}}(\mathbf{x})$ for the conditional type class containing the vectors $\mathbf{y}$ such that $\hat{P}_{\mathbf{xy}} = \hat{P}_{\mathbf{x}} \times \tilde{P}$.

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

$$\|P - P'\| = \sum_{a \in \mathcal{A}} |P(a) - P'(a)| \, . \tag{2}$$

Information-theoretic quantities, such as the entropy of the random variable X whose pmf is P will be denoted by either $H_P(X)$ or $H(P)$. Similarly, the mutual information between X and Y, with joint pmf P, will be denoted by $I_P(X;Y)$, 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 shorthand notation, e.g.,

$$\begin{aligned} \hat{H}_{\mathbf{x}} &\triangleq H_{\hat{P}_{\mathbf{x}}}(X) \\ \hat{I}_{\mathbf{x};\mathbf{y}|\mathbf{u}} &\triangleq I_{\hat{P}_{\mathbf{uxy}}}(X;Y|U). \end{aligned} \tag{3}$$

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

An *exchangeable pmf* $P_{\mathbf{X}}$ is a pmf that is permutation invariant in the sense that for all $\mathbf{x} \in \mathcal{X}^n$,

$$P_{\mathbf{X}}(\mathbf{x}') = P_{\mathbf{X}}(\mathbf{x}), \ \forall \mathbf{x}' \in T_{\mathbf{x}}, \tag{4}$$

or, in other words, for every permutation π of the $n!$ permutations of $\{1, ..., n\}$,

$$P_{\mathbf{X}}(\mathbf{x}) = P_{\mathbf{X}}(\pi\mathbf{x}), \ \forall \mathbf{x}, \tag{5}$$

where $\pi\mathbf{x}$ denotes the sequence $\mathbf{x}$ permuted according to π. Similarly, a *strongly exchangeable conditional pmf*, $P_{\mathbf{Y}|\mathbf{X}}$, is one that satisfies for all $\mathbf{x}, \mathbf{y}$,

$$P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}'|\mathbf{x}') = P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}), \ \forall (\mathbf{x}', \mathbf{y}') \in T_{\mathbf{xy}}, \tag{6}$$

or, in other words, for every permutation π of $\{1, ..., n\}$,

$$P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = P_{\mathbf{Y}|\mathbf{X}}(\pi\mathbf{y}|\pi\mathbf{x}), \ \forall \mathbf{x}, \mathbf{y}. \tag{7}$$

Clearly, every stationary discrete memoryless source (DMS), or a mixture of DMS's, is exchangeable, and every stationary DMC, or a mixture of DMC's, is a strongly exchangeable conditional pmf.

3 Statement of the Problem

Let $\mathbf{U}$ designate the covertext data within which the watermark will be hidden with marginal pmf P_U, and we shall assume, without loss of generality, that $\min_{u \in \mathcal{U}} P_U(u) > 0$. Let W denote the random hidden watermark message independent of $\mathbf{U}$, let $\mathbf{X}$ stand for the watermarked data, and let $\mathbf{Y}$ denote the forgery, i.e., the output of the attack channel. Let $d_1 : \mathcal{U} \times \mathcal{X} \to \mathbb{R}_+$ denote a single-letter distortion measure, and assume for convenience, that $\mathcal{U} = \mathcal{X}$ and that $d_1(x, x) = 0$, $\forall x$. For two vectors $\mathbf{u}$ and $\mathbf{x}$, let $d_1(\mathbf{u}, \mathbf{x}) = \sum_{i=1}^{n} d_1(u_i, x_i)$.

A secret key K_n is a random variable, independent of the watermark message and of $\mathbf{U}$, known to both the encoder and decoder, and unknown to the attacker. Let $\mathcal{K}_n$ stand for the alphabet of the random variable K_n, and let P_{K_n} stand for the distribution of K_n.

Definition 1. *A rate-R watermarking code of size n w.r.t. distortion level D_1 is a triple $(\mathcal{M}_n, f_n, \phi_n)$, where $\mathcal{M}_n = \{1, ..., M\}$, with $M \triangleq \lfloor 2^{nR} \rfloor$, is the watermarking message set, f_n is the encoder, and ϕ_n is the decoder. The encoder is a function*

$$f_n : \mathcal{K}_n \times \mathcal{U}^n \times \mathcal{M}_n \to \mathcal{X}^n, \tag{8}$$

which maps the secret key realization k, the covertext data $\mathbf{u}$ and the watermark message $m \in \mathcal{M}_n$ into a stegotext vector (or, watermarked vector) $\mathbf{x} \in \mathcal{X}^n$. The function f_n meets the distortion constraint[2]

$$Pr\{d_1(\mathbf{U}, \mathbf{X}) > nD_1\} = 0. \tag{9}$$

The decoder is a function

$$\phi_n : \mathcal{K}_n \times \mathcal{Y}^n \to \mathcal{M}_n, \tag{10}$$

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

We shall have a particular interest in the following sub-class of watermarking encoders, referred to as constant composition (CC) encoders.

Definition 2. *A CC encoder is defined by a mapping $\theta : \mathcal{P}(\mathcal{U}) \to \mathcal{P}(\mathcal{X}|\mathcal{U})$ such that*
1. for all $(k_n, \mathbf{u}, w)$, the stegotext sequence $\mathbf{x} = f_n(k_n, \mathbf{u}, w)$ lies within the conditional type class

$$T_\theta(\mathbf{u}) \triangleq \{\mathbf{x} : \hat{P}_{\mathbf{ux}} = \hat{P}_{\mathbf{u}} \times \theta(\hat{P}_{\mathbf{u}})\} \tag{11}$$

2. the mapping θ is continuous (at least) at the point P_U, i.e., for all $\epsilon > 0$ there exists $\delta(\epsilon) > 0$ such that for every $u \in \mathcal{U}$

$$\|P - P_U\| \le \delta(\epsilon) \Rightarrow \|\theta(P)(\cdot|u) - \theta(P_U)(\cdot|u)\| \le \epsilon. \tag{12}$$

Denote by $\mathcal{F}_n(D_1, \mathcal{U}^n \times \mathcal{M}_n \times \mathcal{K}_n)$ the set of CC encoders satisfying $d_1(\mathbf{u}, f_n(k_n, \mathbf{u}, w)) \le nD_1$, $\forall (k_n, \mathbf{u}, w) \in \mathcal{K}_n \times \mathcal{U}^n \times \mathcal{M}_n$.

[2]Note that the condition $\Pr\{d_1(\mathbf{U}, \mathbf{X}) > nD_1\} = 0$ is equivalent to $\Pr\{d_1(\mathbf{U}, \mathbf{X}) > nD_1 | \mathbf{U} = \mathbf{u}\} = 0 \ \forall \mathbf{u} \in \mathcal{U}^n$.

Whenever f_n is a CC encoder, the quadruple $\mathcal{C}_n = (P_{K_n}, \mathcal{M}_n, f_n, \phi_n)$ will be referred to as *a rate-R CC randomized watermarking code.*

The fact that we focus on the wide class of CC codes 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, [6], where such an enumeration method is proposed, and the same idea can be easily extended to conditional type classes.

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 $\mathbf{y} \in \mathcal{Y}^n$. Again, assume for convenience, that $\mathcal{X} = \mathcal{Y}$ and that $d_2(y, y) = 0$, $\forall y$. An attack subject to distortion level D_2 with respect to d_2 is a (possibly randomized) map $h_n : \mathcal{X}^n \to \mathcal{Y}^n$ where,

$$\Pr\{d_2(\mathbf{X}, \mathbf{Y}) > nD_2\} = 0. \tag{13}$$

We henceforth refer to (13) as a *maximum distortion constraint*. As we allow for randomized attacks, we model the action of the attacker by a channel $P_{\mathbf{Y}|\mathbf{X}}$ from $\mathbf{X}$ to $\mathbf{Y}$. The set of attack channels satisfying (13) will be denoted by $\mathcal{P}_n^{d_2}(D_2)$.

We shall denote by $P_e = P_e\left(\mathcal{C}_n, P_{\mathbf{Y}|\mathbf{X}}\right)$ the probability of error averaged over the ensemble of covertexts, watermarks, secret keys, and actions of the attack channel $P_{\mathbf{Y}|\mathbf{X}}$, when the randomized watermarking code $\mathcal{C}_n$ is employed.

An *achievable rate* R is one for which there exists a sequence $\{\mathcal{C}_n\}$ of rate-R randomized CC watermarking codes such that $\lim_{n\to\infty} \sup_{P_{\mathbf{Y}|\mathbf{X}} \in \mathcal{P}_n^{d_2}(D_2)} P_e\left(\mathcal{C}_n, P_{\mathbf{Y}|\mathbf{X}}\right) = 0$. The capacity $C = C(P_U, D_1, D_2)$ of the public watermarking game w.r.t. CC codes is defined as the supremum of all achievable rates. Our goal, in this paper, is to provide an expression of this capacity and to characterize good coding strategies and attack channels.

4 Main Result

In the following theorems, the capacity of the public watermarking game w.r.t. CC codes is expressed as the limit of a sequence of single-letter expressions. Define $D_{\min} = \min_{x,y: d_2(x,y)>0} d_2(x, y)$ and $D_{\max} = \max_{x,y} d_2(x, y)$ and define the following quantity

$$C_\ell = C_\ell(P_U, D_1, D_2) = \max\min\left[I(V; Y) - I(V; U)\right], \tag{14}$$

where V is some auxiliary random variable taking values in a finite alphabet of size $(\ell + 1)^{|\mathcal{X}||\mathcal{Y}|}$, $U \sim P_U$, the maximum is over $P_{V,X|U}$ s.t. $Ed_1(U, X) \le D_1$ and the minimum is over $P_{Y|X} = P_{Y|U,V,X}$ s.t. $Ed_2(X, Y) \le D_2$ and $(U, V) \leftrightarrow X \leftrightarrow Y$ is a Markov chain.

Theorem 1. *(Lower bound) For every $\ell < \infty$, $C \ge C_\ell$.*

Theorem 2. *(Upper bound) For $\ell > \max\{|\mathcal{Y}|^2, \frac{4D_{\max}|\mathcal{Y}|}{|D_{\min}|}\}$, $C \le C_\ell + h \cdot \frac{\log \ell}{\ell}$, where h is a function of $|\mathcal{Y}|, D_{\max}$ and $D_{\min}$.*

Corollary 1. *The capacity of the public watermarking game w.r.t. CC codes is given by* $C = \lim_{l\to\infty} C_\ell$.

Discussion:

The proof of Theorem 1 asserts that although the attacker knows the strategies taken by the encoder and decoder, they can guarantee achievability of C_ℓ for all $\ell < \infty$. It should be emphasized that the derivation of the theorem is not a straightforward extension of the Gel'fand-Pinsker analysis [11] because in the watermarking game, the channel is not fixed in advance and is not necessarily memoryless, so the encoder and decoder should choose their strategies regardless of the attack channel. The proof of the direct theorem is based on the observation that the attack channel can be chosen, without loss of optimality, as uniform over conditional type classes. Since the average probability of error is a linear function of $P_{\mathbf{Y}|\mathbf{X}}$, we can then use the trivial bound which applies to channels that are uniform over conditional types, $P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) \le \frac{\mathbf{1}\{d_2(\mathbf{x},\mathbf{y})\le nD_2\}}{|T_{\mathbf{y}|\mathbf{x}}|}$.

The proof of Theorem 2 asserts that at rates above $C_\ell + O(\log \ell/\ell)$, the attacker can prevent reliable transmission. The main idea behind the proof of the theorem is the following. A class of attack channels that are stationary and memoryless *given* the type class of $\mathbf{X}$ is introduced. It is proved that including attack channels belonging to this class does not help the attacker asymptotically.

Combining Theorems 1 and 2, we realize that $C_\ell \ge C - O(\log \ell/\ell)$, and since C_ℓ is achievable, this inequality quantifies the loss, in terms of rate, when the encoder employs an auxiliary RV having alphabet cardinality $(\ell+1)^{|\mathcal{X}||\mathcal{Y}|}$. It should be pointed out that our requirement for a *sequence* of auxiliary RV's with growing alphabet size, is sufficient for proving Corollary 1 (using Carathéodory's Theorem). It remains an open question whether it is also necessary.

It is interesting to compare the capacity formula of the Gel'fand-Pinsker channel [11] to the expression we obtain. The Gel'fand-Pinsker channel is given by $P_{\mathbf{Y}|\mathbf{S},\mathbf{X}}(\mathbf{y}|\mathbf{s},\mathbf{x}) = \prod_{i=1}^n P_{Y|S,X}(y_i|s_i,x_i)$ where $\mathbf{s} = (s_1, ..., s_n)$ is the channel state sequence, known to the transmitter but not the receiver, which is an i.i.d. process emitted from a finite alphabet source P_S. The capacity formula established in [11] is

$$C_{GP} = \max\left[I(V;Y) - I(V;S)\right], \tag{15}$$

where V is some auxiliary random variable taking values in a finite alphabet of size $|\mathcal{S}| + |\mathcal{X}|$, $S \sim P_S$, the maximum is over $P_{V,X|S}$ and $V \leftrightarrow (S,X) \leftrightarrow Y$ is a Markov chain. The resemblance to C_ℓ is obvious: in our setting the covertext $\mathbf{U}$ plays the role of the state sequence $\mathbf{S}$. There are, however, several differences between the formulas: the minimization over $P_{Y|X}$ in C_ℓ stems from the fact that in our setting there is a game between the attack channel and the encoder-decoder whereas the Gel'fand-Pinsker channel is a known fixed one. This is also the reason why here, unlike in [11], there is no upper bound (at least, not an obvious one) on the cardinality of the alphabet of V, and one cannot apply Carathéodory's Theorem straightforwardly, but rather, to a finite set of conditional pmf's $P_{Y|X}$ (whose size depends on ℓ) that approximate the set $\{P_{Y|X} : Ed_2(X,Y) \le D_2\}$. Due to the fact that the attacker does not observe the original covertext sequence, we have a Markov chain $(U,V) \leftrightarrow X \leftrightarrow Y$ (rather than $V \leftrightarrow (U,X) \leftrightarrow Y$). In contrast to the Gel'fand

Pinkser channel, where the conditional distribution $P_{V,X|U}$ in (15) can be broken up into a deterministic part and a random part, in our case it cannot be done. This is because while the functional $I(V;Y) - I(V;U)$ is convex in $P_{X|U,V}$ for fixed $P_{U,V}, P_{Y|U,V,X}$, the functional $\min_{P_{Y|U,V,X}}[I(V;Y) - I(V;U)]$ is not necessarily convex in $P_{X|U,V}$ for fixed $P_{U,V}$.

There is a discrepancy between our results and the public game results of [13], where a single-letter expression is attained for the capacity, which is given by $C_{|\mathcal{X}||\mathcal{U}|+1}$ (in terms of our notation). In other words, in [13] it is claimed that it is sufficient to consider auxiliary RV's having finite alphabet of cardinality $|\mathcal{X}||\mathcal{U}|+1$, while in this paper, we do not establish an upper bound on the alphabet cardinality (although we do not prove that it is necessarily infinite). The contradiction is explained by a mistake in the proofs of Propositions 4.1.-iv) and 4.1.-vii) in [13].

5 Proof of Theorem 1

In this section, we prove that C_ℓ, where ℓ is a given positive integer, is an achievable rate. The following encoder (which resembles the one introduced in [11]) and decoder guarantee achievability of C_ℓ for every attack channel that lies in $\mathcal{P}_n^{d_2}(D_2)$.

1. Let $\mathcal{V}$ be a set whose cardinality is given by $(\ell+1)^{|\mathcal{X}||\mathcal{Y}|}$ and let $P_{V|U} \in \mathcal{P}(\mathcal{V}|\mathcal{U})$ be a given conditional pmf (to be optimized in the sequel). Denote

$$J \triangleq 2^{n[I(V;U)+\epsilon_n]} \tag{16}$$

   with a proper choice of ϵ_n. A codebook is composed of $M \cdot J$ n-vectors $\{\mathbf{V}_{m,j}\}$, $m = 1,..,M$, $j = 1,...,J$, with i.i.d. components that are drawn according to the marginal pmf of V induced by $P_{UV} = P_U \times P_{V|U}$.

2. Upon observing a covertext sequence $\mathbf{u}$ and the message index m, the encoder chooses an *auxiliary word*, $\mathbf{v}_{m,j^*}$, using the decision rule:

$$j^* = \mathrm{argmin}_j \left\| \hat{P}_{\mathbf{uv}_{m,j}} - P_{UV} \right\|, \tag{17}$$

   where if the minimum is not unique, the choice of j^* is arbitrary among the minimizers.

3. The encoder then draws $\mathbf{X}$ using a pmf that is uniform over a single conditional type given $(\mathbf{u}, \mathbf{v}_{m,j^*})$. The choice of the conditional type class is defined by the mapping $\xi_n : \mathcal{P}(\mathcal{U}\times\mathcal{V}) \to \mathcal{P}(\mathcal{X}|\mathcal{U}\times\mathcal{V})$, i.e., $\mathbf{X}$ in distributed uniformly over $T_{\xi_n}(\mathbf{uv}_{m,j^*}) = \{\mathbf{x} \in \mathcal{X}^n : \hat{P}_{\mathbf{uv}_{m,j^*}\mathbf{x}} = \hat{P}_{\mathbf{uv}_{m,j^*}} \times \xi_n(\hat{P}_{\mathbf{uv}_{m,j^*}})\}$. In order to achieve capacity, it is sufficient to

assume that ξ_n satisfies the following requirements

$$
\begin{aligned}
&\bullet\, E_{P\times\xi_n(P)} d_1(U,X) \le D_1 \quad \forall P\in\mathcal{P}(\mathcal{U}\times\mathcal{V})\\
&\bullet\, \forall P\in\mathbb{P}_n(\mathcal{U}\times\mathcal{V}),\ \xi_n(P)\in\mathbb{P}_n(\mathcal{X},P)\\
&\bullet\, \xi_n(P) \text{ is continuous at the point } P=P_{UV}\\
&\bullet\, \forall P\in\mathcal{P}(\mathcal{U}), P_1,P_2\in\mathcal{P}(\mathcal{V}|\mathcal{U}), u\in\mathcal{U}, x\in\mathcal{X}\\
&\quad \sum_{v\in\mathcal{V}} P_1(v|u)\xi_n(P\times P_1)(x|uv)\\
&= \quad \sum_{v\in\mathcal{V}} P_2(v|u)\xi_n(P\times P_2)(x|uv).
\end{aligned} \tag{18}
$$

The first two requirements assure that only conditional types such that constraint (9) is met are considered, and the third requirement is adopted to make sure that (12) is met and for technical reasons. The fourth requirement is set in order to be compliant with (11), where a CC code is defined by a mapping from $\mathcal{P}(\mathcal{U})$ to $\mathcal{P}(\mathcal{X}|\mathcal{U})$, which implies that regardless of the value of $\mathbf{v}_{j,m}$, the stegotext $\mathbf{x}$ should lie in the same type class given $\mathbf{u}$.

4. The decoder that observes the forgery $\mathbf{y}$ and the codebook $\{\mathbf{v}_{m,j}\}$ operates as follows:

$$
\hat{m} = \text{argmax}_m \max_j \hat{I}_{\mathbf{v}_{m,j};\mathbf{y}} \tag{19}
$$

if the maximizing m is unique and arbitrary among the maximizers otherwise. This maximum mutual information (MMI) decoder (see [9]) can be regarded as a universal decoder for the class of channels $\mathcal{P}_n^{d_2}(D_2)$ as it enables reliable communications at rates below C_ℓ for all the channels in this class.

Assume, without loss of generality, that the transmitted message index is $m=1$. Let $\tilde{\mathbf{V}}$ stand for the random variable designating the auxiliary word chosen by the encoder as described in (17). Denote by $P_e\left(P_{V|U},\xi_n,P_{\mathbf{Y}|\mathbf{X}}\right)$ the average probability of error associated with $P_{V|U}$, ξ_n and the attack channel $P_{\mathbf{Y}|\mathbf{X}}$. The optimization problem faced by the encoder is that of choosing the optimal $P_{V|U}$ and ξ_n such that the resulting average probability of error is minimized. The average probability of error is therefore given by

$$
\begin{aligned}
&P_e\left(P_{V|U},\xi_n,P_{\mathbf{Y}|\mathbf{X}}\right)\\
&= \sum_{\mathbf{u},\mathbf{v}} P_{\mathbf{U}\tilde{\mathbf{V}}}(\mathbf{uv}) \sum_{\mathbf{x}\in T_{\xi_n}(\mathbf{uv})} \frac{1}{|T_{\xi_n}(\mathbf{uv})|} \sum_{\mathbf{y}} P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})\\
&\times \left[1-\left(\sum_{\mathbf{v}':\,\hat{I}_{\mathbf{v}';\mathbf{y}}<\hat{I}_{\mathbf{v};\mathbf{y}}} P_V^n(\mathbf{v}')\right)^{(M-1)J}\right],
\end{aligned} \tag{20}
$$

where $P_V^n(\mathbf{v}') \triangleq \prod_{i=1}^n P_V(v_i')$.

The term $\left[1-\left(\sum_{\mathbf{v}':\,\hat{I}_{\mathbf{v}';\mathbf{y}}<\hat{I}_{\mathbf{v};\mathbf{y}}} P_V^n(\mathbf{v}')\right)^{(M-1)J}\right]$ can be upper bounded as follows:

- For all $0 \le d \le 1$,

$$1-(1-d)^{(M-1)J} \le \min\left\{1,(M-1)Jd\right\}. \tag{21}$$

- Since P_V^n is memoryless, $P_V^n(\mathbf{v}') \le \frac{1}{|T_{\mathbf{v}'}|}$, $\forall \mathbf{v}'$.

This yields,

$$\begin{aligned}
& 1-\left(\sum_{\mathbf{v}':\, \hat{I}_{\mathbf{v}';\mathbf{y}}<\hat{I}_{\mathbf{v};\mathbf{y}}} P_V^n(\mathbf{v}')\right)^{(M-1)J} \\
\le\ & \min\left\{1, MJ \sum_{\mathbf{v}':\, \hat{I}_{\mathbf{v}';\mathbf{y}}\ge\hat{I}_{\mathbf{v};\mathbf{y}}} \frac{1}{|T_{\mathbf{v}'}|}\right\} \\
=\ & \min\left\{1, MJ \sum_{T_{\mathbf{v}'|\mathbf{y}}:\, \hat{I}_{\mathbf{v}';\mathbf{y}}\ge\hat{I}_{\mathbf{v};\mathbf{y}}} \frac{|T_{\mathbf{v}'|\mathbf{y}}|}{|T_{\mathbf{v}'}|}\right\} \\
\doteq\ & \min\left\{1, MJe^{-n\hat{I}_{\mathbf{v};\mathbf{y}}}\right\}
\end{aligned} \tag{22}$$

where the last step holds since $\frac{|T_{\mathbf{v}|\mathbf{y}}|}{|T_{\mathbf{v}}|} \doteq e^{-n\hat{I}_{\mathbf{v};\mathbf{y}}}$, and the number of conditional types $T_{\mathbf{v}'|\mathbf{y}}$ grows polynomially with n.

By definition, the joint pmf of the covertext sequence $\mathbf{U}$ and the auxiliary word $\tilde{\mathbf{V}}$, denoted by $P_{\mathbf{U}\tilde{\mathbf{V}}}$ is strongly exchangeable. It is therefore easily verified that under the above assumptions, the average probability of error, as appears in (20), associated with the channel $P_{\mathbf{Y}|\mathbf{X}}$, is equal to that of the channel $P^{\pi}_{\mathbf{Y}|\mathbf{X}}$ defined as $P^{\pi}_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = P_{\mathbf{Y}|\mathbf{X}}(\pi\mathbf{y}|\pi\mathbf{x})$ where π is some permutation of $\{1,...,n\}$. Thus, due to the linearity of P_e (in (20)) in $P_{\mathbf{Y}|\mathbf{X}}$, P_e associated with $P_{\mathbf{Y}|\mathbf{X}}$ is equal to P_e that is associated with $\bar{P}_{\mathbf{Y}|\mathbf{X}} = \frac{1}{n!}\sum_{\pi} P^{\pi}_{\mathbf{Y}|\mathbf{X}}$. This means that, without loss of generality, we can consider attack channels that are strongly exchangeable (and as a consequence, uniform over conditional types).

Hence, we can assume that

$$P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = \frac{\Pr(T_{\mathbf{y}|\mathbf{x}}|\mathbf{X}=\mathbf{x})}{|T_{\mathbf{y}|\mathbf{x}}|}, \tag{23}$$

and since $\Pr(T_{\mathbf{y}|\mathbf{x}}|\mathbf{X}=\mathbf{x}) \le \mathbf{1}\left\{d_2(\mathbf{x},\mathbf{y}) \le nD_2\right\}$, we have the upper bound

$$P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) \le \frac{\mathbf{1}\left\{d_2(\mathbf{x},\mathbf{y}) \le nD_2\right\}}{|T_{\mathbf{y}|\mathbf{x}}|}\ . \tag{24}$$

Define the sets

$$\begin{aligned}
& A_1(\eta_n) \triangleq \left\{\mathbf{u} : ||\hat{P}_{\mathbf{u}} - P_U|| > \eta_n\right\}, \\
& A_2(\delta_n) \triangleq \\
& \left\{(\mathbf{u}, \{\mathbf{v}_{1,j}\}_{j=1}^{J}) : \forall j,\ ||\hat{P}_{\mathbf{u}\mathbf{v}_{1,j}} - P_{U,V}|| > \delta_n\ \right\},
\end{aligned} \tag{25}$$

where it is assumed that the number of auxiliary words per codeword is J (as in (16)), with

$$\epsilon_n = |\mathcal{U}||\mathcal{V}|\frac{\log n}{n} + \frac{\delta_n^2}{2\ln 2} - 3\delta_n \log\frac{\delta_n}{|\mathcal{U}||\mathcal{V}|} + \tilde{\epsilon}_n, \tag{26}$$

and $\tilde{\epsilon}_n = O(n^{-\gamma})$ for some $\gamma < 1$.

Note that,

$$\begin{aligned} P_e &\leq \Pr\{A_1(\eta_n)\} + \Pr\{A_2(\delta_n)|\bar{A}_1(\eta_n)\} \\ &+ \Pr\{error|\bar{A}_1(\eta_n), \bar{A}_2(\delta_n)\}. \end{aligned} \tag{27}$$

The next lemma follows from standard typicality considerations, but for the sake of completeness it is proved in Appendix A. It implies that one can choose $\eta_n \to 0, \delta_n \to 0$ such that $\Pr\{A_1(\eta_n)\} \to 0$ and $\Pr\{A_2(\delta_n)|\bar{A}_1(\eta_n)\} \to 0$.

Lemma 1. *For all $\eta_n > 0$,*

$$Pr\{A_1(\eta_n)\} \leq (n+1)^{|\mathcal{U}|} e^{-n\frac{\eta_n^2}{2\ln 2}} \ , \tag{28}$$

and for all δ_n such that $\eta_n|\mathcal{U}| + \frac{|\mathcal{V}|}{2n} \leq \delta_n < \frac{1}{2}$,

$$Pr\left\{A_2(\delta_n)|\bar{A}_1(\eta_n)\right\} \leq e^{-\exp(n\tilde{\epsilon}_n)}. \tag{29}$$

Thus, it is left to upper bound the rightmost term of (27) which will be denoted by P_e'.

Define the set

$$\mathcal{E}_n \triangleq \left\{ P_{U'V'} \in \mathcal{P}(\mathcal{U}\times\mathcal{V}) : \begin{array}{c} ||P_{U'} - P_U|| \leq \eta_n, \\ ||P_{U'V'} - P_{UV}|| \leq \delta_n \end{array} \right\}. \tag{30}$$

Eqs. (20)-(24) yield

$$\begin{aligned} P_e' &\stackrel{\cdot}{\leq} \max_{\mathbf{uv}:\, \hat{P}_{\mathbf{uv}}\in\mathcal{E}_n} \sum_{\mathbf{x}\in T_{\xi_n}(\mathbf{uv})} \frac{1}{|T_{\xi_n}(\mathbf{uv})|} \\ &\times \sum_{\mathbf{y}:\, d_2(\mathbf{x},\mathbf{y})\leq nD_2} \frac{e^{-n\left|\hat{I}_{\mathbf{v};\mathbf{y}} - \frac{\log J}{n} - R\right|^+}}{|T_{\mathbf{y}|\mathbf{x}}|}. \end{aligned} \tag{31}$$

For a pmf $P' \in \mathbb{P}_n(\mathcal{U}\times\mathcal{V}\times\mathcal{X})$ define the set

$$\mathcal{P}_n(D_2, P') \triangleq \left\{\hat{P}_{Y|UVX} \in \mathbb{P}_n(\mathcal{Y}, P') : Ed_2(X,Y) \leq D_2\right\} \tag{32}$$

where the expectation is w.r.t. the pmf $P' \times \hat{P}_{Y|UVX}$.

As $d_2(\mathbf{x},\mathbf{y}) = nE_{\hat{P}_{\mathbf{xy}}} d_2(X,Y)$, we have

$$\mathcal{P}_n(D_2, P_{\mathbf{uv}}\xi_n(P_{\mathbf{uv}})) = \{\hat{P}_{\mathbf{y}|\mathbf{uvx}} :\ d_2(\mathbf{x},\mathbf{y}) \leq nD_2\}, \tag{33}$$

hence,

$$\begin{aligned}
&\sum_{\mathbf{y}:\, d_2(\mathbf{x},\mathbf{y})\le nD_2} \frac{e^{-n\left|\hat{I}_{\mathbf{v};\mathbf{y}}-\frac{\log J}{n}-R\right|^+}}{|T_{\mathbf{y}|\mathbf{x}}|} \\
&\doteq \max_{\hat{P}_{\mathbf{y}|\mathbf{uvx}}} \frac{|T_{\mathbf{y}|\mathbf{uvx}}|}{|T_{\mathbf{y}|\mathbf{x}}|} e^{-n\left|\hat{I}_{\mathbf{v};\mathbf{y}}-\frac{\log J}{n}-R\right|^+} \\
&\doteq \max_{\hat{P}_{\mathbf{y}|\mathbf{uvx}}} e^{-n\left(\hat{I}_{\mathbf{y};\mathbf{uv}|\mathbf{x}}+\left|\hat{I}_{\mathbf{v};\mathbf{y}}-\frac{\log J}{n}-R\right|^+\right)}
\end{aligned} \tag{34}$$

where the maximizations are over the set $\{\hat{P}_{\mathbf{y}|\mathbf{uvx}} \in \mathcal{P}_n(D_2, P_{\mathbf{uv}}\xi_n(P_{\mathbf{uv}}))\}$ and the last step follows since $\frac{|T_{\mathbf{y}|\mathbf{uvx}}|}{|T_{\mathbf{y}|\mathbf{x}}|} \doteq e^{-n\hat{I}_{\mathbf{y};\mathbf{uv}|\mathbf{x}}}$, and because the number of types $T_{\mathbf{y}|\mathbf{uvx}}$ grows polynomially with n.

Next, for a pmf $P \in \mathcal{P}(\mathcal{U}\times\mathcal{V}\times\mathcal{X}\times\mathcal{Y})$ define

$$\begin{aligned}
G_{R,J,n}(P) &\triangleq I_P(Y;UV|X) + \left|I_P(V;Y) - \frac{\log J}{n} - R\right|^+ \\
G_R(P) &\triangleq I_P(Y;UV|X) + |I_P(V;Y) - I_P(V;U) - R|^+
\end{aligned} \tag{35}$$

The r.h.s. of (31) is therefore upper bounded (in the exponential scale) by

$$\exp\{-n \min_{\hat{P}_{\mathbf{uv}},\hat{P}_{\mathbf{y}|\mathbf{uvx}}} G_{R,J,n}(\hat{P}_{\mathbf{uv}}\xi_n(\hat{P}_{\mathbf{uv}})\hat{P}_{\mathbf{y}|\mathbf{uvx}})\} \tag{36}$$

where the minimization is over the set $\{\hat{P}_{\mathbf{uv}} \in \mathcal{E}_n, \hat{P}_{\mathbf{y}|\mathbf{uvx}} \in \mathcal{P}_n(D_2, P_{\mathbf{uv}}\xi_n(P_{\mathbf{uv}}))\}$.

The rest of the derivation is rather technical and is devoted to replacing the optimizations over sets of empirical measures with optimizations over continuous sets of measures. More specifically, recalling the definition of J in eq. (16), one can establish the following Lemma, proved in Appendix A.

Lemma 2. *There exists a sequence $\{\epsilon_n\}_{n\ge1}$, $\epsilon_n \to 0$ such that the term (36) is upper bounded by*

$$\exp\left\{-n\left(\inf_{P_{Y|UVX}} G_R(P_{UV}\xi_n(P_{UV})P_{Y|UVX}) - \epsilon_n\right)\right\} \tag{37}$$

where the infimum is over the set $\{P_{Y|UVX} :\ Ed_2(X,Y) \le D_2 + \epsilon_n\}$.

As the choice of $P_{V|U}$ and the mapping $\xi_n(P_{UV})$ can be optimized, and $P_{V|U}\times\xi_n(P_{UV}) \in \mathcal{P}(\mathcal{V}\times\mathcal{X}|\mathcal{U})$, Equations (31)-(37) yield

$$P'_e \stackrel{\cdot}{\le} \exp\left\{-n \sup_{P_{VX|U}} \inf_{P_{Y|UVX}} G_R(P_{UVXY})\right\}, \tag{38}$$

where the supremum is over the set $\{P_{VX|U} :\ Ed_1(U,X) \le D_1\}$ and the infimum is over the set $\{P_{Y|UVX} :\ Ed_2(X,Y) \le D_2\}$. Therefore, the smallest rate R which solves the equation

$$\sup_{\{P_{X|UV}:\, Ed_1(U,X)\le D_1\}} \inf_{\{P_{Y|UVX}:\, Ed_2(X,Y)\le D_2\}} \left[I(Y;UV|X) + |I(V;Y) - I(V;U) - R|^+\right] = 0 \tag{39}$$

is a lower bound on the capacity.

The left hand side of (39) vanishes only if both $R = I(V;Y) - I(V;U)$ and $I(Y;UV|X) = 0$, i.e., $UV \leftrightarrow X \leftrightarrow Y$ forms a Markov chain. In other words, the smallest rate satisfying (39) is given by C_ℓ as defined in (14).

6 Proof of Theorem 2

The proof of the converse part consists of the following steps:

- Fano's inequality, stating that if R is an achievable rate then $R \le I(W;\mathbf{Y}|K_n)$, where, as defined earlier, W is the RV designating the hidden watermark message.

- A class of attack channels that are stationary and memoryless *given* the type class of $\mathbf{X}$ is introduced. It is proved that including attack channels belonging to this class in the optimization problem (max min $I(W;\mathbf{Y}|K_n)$ does not help the attacker asymptotically.

- A variation of Gel'fand and Pinsker's result [11, Lemma 4] according to which

$$
\begin{aligned}
& I(W, K_n; \mathbf{Y}) - I(W, K_n; \mathbf{U}) \\
\le\ & \sum_{i=1}^{n} \left[I(Z_i; Y_i) - I(Z_i; U_i) \right],
\end{aligned} \tag{40}
$$

  where $Z_i = (W, K_n, U_{i+1}^n, Y_1^{i-1})$ is used, and Z_i is interpreted as an auxiliary RV whose distribution can be optimized by the information hider.

- Additional technical constraints on the attacker are imposed. Loosely speaking, it is required that if the type class of $\mathbf{X}$ is ϵ-typical w.r.t. the marginal pmf $P_{\tilde{X}}$ induced by $P_U \times \theta(P_U)$, i.e., small fluctuations in the empirical statistics of $\mathbf{X}$ do not change the action of the attacker.

- The auxiliary RV Z is needed in order to preserve the value of $\min_{P_{Y|X}:\, Ed_2(X,Y) \le D_2} [I(Z;Y) - I(Z;U)]$ and to guarantee it, we need to preserve the value of $I(Z;Y) - I(Z;U)$ essentially for *every*[3] pmf $P_{Y|X}$ in the class $\{P_{Y|X} :\ Ed_2(X,Y) \le D_2\}$. By definition, the alphabet of Z_i grows with n. While in [11] it is sufficient to consider an auxiliary RV with alphabet of finite size and still achieve capacity, in our case, Carathéodory's theorem cannot be used straightforwardly because we are facing a problem of communication over an unknown channel. Therefore, we define a "grid of channels", i.e., we approximate every channel $P_{Y|X} \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$ by a quantized one, and then, instead of the infinitely many channels in $\mathcal{P}(\mathcal{Y}|\mathcal{X})$ there are $(\ell+1)^{|\mathcal{X}||\mathcal{Y}|}$ channels (where ℓ is the grid parameter) to which we can apply Carathéodory's Theorem. Using the continuity of $C_\ell = C_\ell(P_U, D_1, D_2)$ w.r.t. D_2, this yields the desired result.

[3] Cf. the discussion after Corollary 1, in the context of the sequence of auxiliary RV's with growing alphabets.

First, note that as there is a finite number of functions from $\mathcal{U}^n \times \mathcal{M}_n$ to $\mathcal{X}^n$, it is sufficient to consider secret code distributions, P_{K_n}, that are defined on a finite alphabet denoted by $\mathcal{K}_n$ (whose cardinality may grow double-exponentially with n).

Let P_{K_n}, $f_n \in \mathcal{F}_n(D_1, \mathcal{U}^n \times \mathcal{M}_n \times \mathcal{K}_n)$ and $P_{\mathbf{Y}|\mathbf{X}}$ be given. Using Fano's inequality, we have

$$nR \leq 1 + P_e nR + I(W; \mathbf{Y}|K_n), \tag{41}$$

and thus, an achievable rate R must satisfy

$$R \leq \liminf_{n\to\infty} \frac{1}{n} \max_{P_{K_n}, f_n} \min_{P_{\mathbf{Y}|\mathbf{X}} \in \mathcal{P}_n^{d_2}(D_2)} I(W; \mathbf{Y}|K_n) \tag{42}$$

where the maximization is over the set $\{P_{K_n} \in \mathcal{P}(\mathcal{K}_n), f_n \in \mathcal{F}_n(D_1, \mathcal{U}^n \times \mathcal{M}_n \times \mathcal{K}_n)\}$, and the joint law of $(\mathbf{U}, K_n, W, \mathbf{X}, \mathbf{Y})$ is given by

$$P_{\mathbf{U},K_n,W,\mathbf{X},\mathbf{Y}} = P_{\mathbf{U}} P_{K_n} P_W \mathbf{1}\{\mathbf{X} = f_n(\mathbf{U}, K_n, W)\} P_{\mathbf{Y}|\mathbf{X}}. \tag{43}$$

Let $\mathcal{N}(D_2)$ be the set of mappings $\Lambda : \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{Y}|\mathcal{X})$ satisfying

$$E_{P\times\Lambda(P)} d_2(X, Y) \leq D_2 \quad \forall P \in \mathcal{P}(\mathcal{X}). \tag{44}$$

Consider the class of attack channels $\{P^{\Lambda}_{\mathbf{Y}|\mathbf{X}}\}$ defined as follows

$$P^{\Lambda}_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{n} \Lambda(\hat{P}_{\mathbf{x}})(y_i|x_i), \tag{45}$$

in other words, given the type class of $\mathbf{X}$, the channel is stationary and memoryless.

Lemma 3. *If $\Lambda \in \mathcal{N}(D_2 - \epsilon)$ then*

$$P^{\Lambda}_{\mathbf{Y}|\mathbf{X}}\{d_2(\mathbf{X}, \mathbf{Y}) > nD_2 | \mathbf{X} = \mathbf{x}\} \leq d_n(\epsilon) \quad \forall \mathbf{x} \in \mathcal{X}^n, \tag{46}$$

where $d_n(\epsilon) = (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp\left\{-n \frac{\epsilon^2}{(2\ln 2) D_{\max}^2 |\mathcal{X}|^2 |\mathcal{Y}|^2}\right\}$ and $D_{\max} = \max_{x,y} d_2(x, y)$.

Lemma 4. *For every channel P^{Λ} such that $\Lambda \in \mathcal{N}(D_2 - \epsilon)$, there exists a channel $P_2 \in \mathcal{P}_n^{d_2}(D_2)$ such that $\sum_{\mathbf{y}} \left|P^{\Lambda}(\mathbf{y}|\mathbf{x}) - P_2(\mathbf{y}|\mathbf{x})\right| \leq 2d_n(\epsilon)$ for all $\mathbf{x}$.*

For convenience denote for $t > 0$ and $A > 1$,

$$\Delta(t, A) = -t \log \frac{t}{A} \tag{47}$$

and note that $\lim_{t\to 0} \Delta(t) = 0$.

Corollary 2. *Assume ϵ is such that $d_n(\epsilon) \leq \frac{1}{4}$. For every channel P^{Λ} such that $\Lambda \in \mathcal{N}(D_2 - \epsilon)$, there exists a channel $P_2 \in \mathcal{P}_n^{d_2}(D_2)$ such that*

$$|I_{P^{\Lambda}}(W; \mathbf{Y}|K_n) - I_{P_2}(W; \mathbf{Y}|K_n)| \leq 2\Delta\left(2d_n(\epsilon), |\mathcal{Y}|^n\right). \tag{48}$$

The corollary follows from [9, Lemma 1.2.7.], according to which if $P_1, P_2 \in \mathcal{P}(\mathcal{A})$ where $|\mathcal{A}| < \infty$ then

$$\|P_1 - P_2\| \leq \alpha \leq \frac{1}{2} \Rightarrow |H(P_1) - H(P_2)| \leq -\alpha \log \frac{\alpha}{|\mathcal{A}|}, \tag{49}$$

$\alpha \log \frac{\alpha}{|\mathcal{A}|} = \Delta(\alpha, |\mathcal{A}|)$ and thus,

$$\begin{aligned} & |I_{P^\Lambda}(W;\mathbf{Y}|K_n) - I_{P_2}(W;\mathbf{Y}|K_n)| \\ \leq\ & |H_{P^\Lambda}(\mathbf{Y}|K_n) - H_{P_2}(\mathbf{Y}|K_n)| \\ & + |H_{P^\Lambda}(\mathbf{Y}|K_n W) - H_{P_2}(\mathbf{Y}|K_n W)| \\ \leq\ & 2\Delta(2d_n(\epsilon), |\mathcal{Y}|^n) \triangleq \zeta_n(\epsilon) \end{aligned} \tag{50}$$

where the second inequality follows from (49) and Lemma 4.

In the sequel, we shall choose $\epsilon = \epsilon_n$ where $\epsilon_n = O(n^{-\gamma})$ with $0 < \gamma < \frac{1}{2}$ so that $d_n(\epsilon) \to 0$ and also $\Delta(2d_n(\epsilon), |\mathcal{Y}|^n) \to 0$.

Corollary 2 implies that including attack channels of the form (45) in the optimization problem ($\max\min I(W;\mathbf{Y}|K_n)$ as in (41)) does not help the attacker much. In other words, for every P_{K_n}, f_n,

$$\begin{aligned} & \min_{P_{\mathbf{Y}|\mathbf{X}} \in \mathcal{P}_n^{d_2}(D_2)} I_{P_{\mathbf{U}} P_{K_n} P_W \mathbf{1}\{\mathbf{X}=f_n(\mathbf{U},K_n,W)\} P_{\mathbf{Y}|\mathbf{X}}}(W;\mathbf{Y}|K_n) \\ \leq\ & \min_{\Lambda \in \mathcal{N}(D_2-\epsilon)} I_{P_{\mathbf{U}} P_{K_n} P_W \mathbf{1}\{\mathbf{X}=f_n(\mathbf{U},K_n,W)\} P^\Lambda_{\mathbf{Y}|\mathbf{X}}}(W;\mathbf{Y}|K_n) + \zeta_n(\epsilon). \end{aligned} \tag{51}$$

Next, observe that

$$\begin{aligned} I(W;\mathbf{Y}|K_n) &= H(\mathbf{Y}|K_n) - H(\mathbf{Y}|K_n, W) \\ &\leq H(\mathbf{Y}) - H(\mathbf{Y}|K_n, W) \\ &= I(K_n, W; \mathbf{Y}) - I(K_n, W; \mathbf{U}), \end{aligned} \tag{52}$$

where the last step is because $(K_n, W, \mathbf{U})$ are independent RVs.

Denote

$$Z_i = (W, K_n, U_{i+1}^n, Y_1^{i-1}). \tag{53}$$

Now, we use a variation of the result[4] of Gel'fand and Pinsker [11, Lemma 4] according to which

$$\begin{aligned} & I(W, K_n; \mathbf{Y}) - I(W, K_n; \mathbf{U}) \\ \leq\ & \sum_{i=1}^n \left[I(Z_i; Y_i) - I(Z_i; U_i)\right]. \end{aligned} \tag{54}$$

[4]To prove the lemma, the same proof as the one of the Lemma 4 in [11] is applied where $((W, K_n), U_i, Z_i)$ in our notation replace $(V, S(i), U(i))$.

This inequality, combined with (42), (51) and (52) yields,

$$R \leq \liminf_{n\to\infty} \frac{1}{n} \max_{P_{K_n}, f_n} \min_{\Lambda} \sum_{i=1}^{n} [I(Z_i;Y_i) - I(Z_i;U_i)] \tag{55}$$

where the minimization is over the set $\{\Lambda \in \mathcal{N}(D_2 - \epsilon)\}$ maximization is over the set $\{P_{K_n} \in \mathcal{P}(\mathcal{K}_n), f_n \in \mathcal{F}_n(D_1, \mathcal{U}^n \times \mathcal{M}_n \times \mathcal{K}_n)\}$, and the probability measure defining $\sum_{i=1}^{n} [I(Z_i;Y_i) - I(Z_i;U_i)]$ is given by

$$\begin{aligned} & P_{\mathbf{U},K_n,W,\mathbf{X},\mathbf{Y}} = P_{\mathbf{U}} P_{K_n} P_W \mathbf{1}\{\mathbf{X} = f_n(\mathbf{U}, K_n, W)\} \\ & \times \prod_{i=1}^{n} \Lambda(\hat{P}_{\mathbf{X}})(Y_i|X_i). \end{aligned} \tag{56}$$

Note that

$$(\mathbf{U}, K_n, W, \mathbf{X}, \{Y_j\}_{j\neq i}) \leftrightarrow (X_i, \hat{P}_{\mathbf{X}}) \leftrightarrow Y_i, \tag{57}$$

which yields

$$Z_i = (W, K_n, U_{i+1}^n, Y_1^{i-1}) \leftrightarrow (X_i, \hat{P}_{\mathbf{X}}) \leftrightarrow Y_i. \tag{58}$$

Now, we introduce a time-sharing RV, T, independent of $(\mathbf{U}, K_n, W, \mathbf{X}, \mathbf{Y})$ and uniformly distributed over $\{1, ..., n\}$, and define the quadruple of RV's $(U, Z, X, Y) = (U_T, Z_T, X_T, Y_T)$.

By definition of T,

$$\begin{aligned} & \sum_{i=1}^{n} I(Z_i;Y_i) - I(Z_i;U_i) \\ = \;& n[I(Z;Y|T) - I(Z;U|T)] \\ = \;& n[I(Z,T;Y) - I(Z,T;U) - I(T;Y) + I(T;U)] \\ \leq \;& n[I(Z,T;Y) - I(Z,T;U)] \end{aligned} \tag{59}$$

where the last step is because $I(T;U) = 0$ since $\mathbf{U}$ is stationary and memoryless.

Define $\tilde{Z} = (Z, T)$ and and note that due to (58), the joint law of $(\mathbf{U}, K_n, W, \mathbf{X}, \mathbf{Y})$ as in (56) induces a joint pmf of the RV's $(U, \tilde{Z}, X, \hat{P}_{\mathbf{X}}, Y)$ that is given by

$$P_{U\tilde{Z}X\hat{P}_{\mathbf{X}}Y} \;=\; P_{U\tilde{Z}X\hat{P}_{\mathbf{X}}} \times \Lambda(\hat{P}_{\mathbf{X}}). \tag{60}$$

Now, for a pmf $P_{U\tilde{Z}Y} \in \mathcal{P}(\mathcal{U} \times \tilde{\mathcal{Z}} \times \mathcal{Y})$ denote

$$J(P_{U\tilde{Z}Y}) = I(\tilde{Z};Y) - I(\tilde{Z};U), \tag{61}$$

and with a little abuse of notation, we shall also denote by $J(P_{U\tilde{Z}X\hat{P}_{\mathbf{X}}Y})$ the quantity $I(\tilde{Z};Y) - I(\tilde{Z};U)$ which is a functional of the marginal distribution of $(U, \tilde{Z}, Y)$ induced by $P_{U\tilde{Z}X\hat{P}_{\mathbf{X}}Y}$.

Recalling (60), we shall be interested in the quantity

$$J(P_{U\tilde{Z}X\hat{P}_{\mathbf{X}}} \times \Lambda(\hat{P}_{\mathbf{X}})). \tag{62}$$

Now we further constrain the attacker as follows. Let $P_{\tilde{X}|U} = \theta(P_U)$ and let $P_{\tilde{X}}$ denote the appropriate marignal pmf induced by $P_U \times \theta(P_U)$. Let $\{b_n\}_{n\geq 1}$ be a sequence such that $b_n \to 0$ and for some $t > 1$

$$\delta^2(b_n) = O(n^{-t}) \tag{63}$$

where the function $\delta(\epsilon)$ is defined in (12). We assume that

$$\Lambda(P) = \Lambda(P_{\tilde{X}}) \ \forall P \text{ s.t. } \|P - P_{\tilde{X}}\| \leq b_n + \delta(b_n) \tag{64}$$

and that $\Lambda(\cdot)$ is such that

$$E_{P_{\tilde{X}} \times \Lambda(P_{\tilde{X}})} d_2(X, Y) \leq D_2 - \alpha_n \tag{65}$$

where

$$\alpha_n = \epsilon_n + (b_n + \delta(b_n))\, |\mathcal{X}||\mathcal{Y}| D_{\max}. \tag{66}$$

From the following two lemmas, it follows that although we introduce these constraints, the attacker is still capable of preventing reliable detection of the watermark in rates higher than the capacity.

Lemma 5. *If $\Lambda(\cdot)$ is such that constraints (64) and (65) are met, then $\Lambda(\cdot) \in \mathcal{N}(D_2 - \epsilon_n)$.*

We shall take a special interest in the event

$$\begin{aligned} &\mathcal{E}_1 = \\ &\{\mathbf{u}: \ \|\hat{P}_{\mathbf{u}} \times \theta(\hat{P}_{\mathbf{u}}) - P_U \times \theta(P_U)\| \leq b_n + \delta(b_n)\}, \end{aligned} \tag{67}$$

which occurs with very high probability (as will be proved in the sequel). Define the RV Z' as follows. The conditional pmf of Z' given U, X is given by

$$P_{Z'|UX}(z|ux) = P_{\mathbf{U}K_n W \mathbf{X}\mathbf{Y}}(\tilde{Z} = z | U = u, X = x, \mathcal{E}_1). \tag{68}$$

Lemma 6. *When the class of constant composition encoders is employed and $\Lambda(\cdot)$ is such that (64) holds, then for sufficiently large n,*

$$\begin{aligned} &J(P_{U\tilde{Z}X\hat{P}_{\mathbf{X}}} \times \Lambda(\hat{P}_{\mathbf{X}})) - J(P_{U\tilde{X}} \times P_{Z'|UX} \times \Lambda(P_{\tilde{X}})) \\ &\leq 2\Delta\left(b_n + \delta(b_n) + 2(n+1)^{|\mathcal{U}|}\exp\{-n\delta^2(b_n)(2\ln 2)\}, |\mathcal{Y}|\right) \\ &\triangleq \gamma_n(b_n). \end{aligned} \tag{69}$$

The a.s. constraint $\Pr\{d_1(\mathbf{U}, \mathbf{X}) \leq nD_1\} = 1$ implies that $Ed_1(\mathbf{U}, \mathbf{X}) \leq nD_1$, and therefore $P_{X|U}$ induced by (f_n, P_{K_n}) must be such that $E_{P_U P_{X|U}} d_1(U, X) \leq D_1$. This fact, together with Lemma 5 and Lemma 6 result in

$$\begin{aligned} &\max_{P_{\tilde{Z}X\hat{P}_{\mathbf{X}}|U}:\, Ed_1(U,X)\leq D_1} \ \min_{\Lambda \in \mathcal{N}(D_2-\epsilon_n)} J(P_{U\tilde{Z}X\hat{P}_{\mathbf{X}}} \times \Lambda(\hat{P}_{\mathbf{X}})), \\ \leq\ &\max_{P_{Z'\tilde{X}|U}:\, Ed_1(U,\tilde{X})\leq D_1} \ \min_{\Lambda(P_{\tilde{X}}) \in \mathcal{N}(D_2-\alpha_n)} J(P_{UZ'\tilde{X}} \times \Lambda(P_{\tilde{X}})) + \gamma_n(b_n) \\ \triangleq\ &\max_{P_{Z',\tilde{X}|U}:\, Ed_1(U,\tilde{X})\leq D_1} \ \min_{P_{Y|\tilde{X}} \in \mathcal{D}(D_2-\alpha_n, P_{\tilde{X}})} J(P_{UZ'\tilde{X}} \times P_{Y|\tilde{X}}) + \gamma_n(b_n), \end{aligned} \tag{70}$$

where $P_{Y|\tilde{X}} = \Lambda(P_{\tilde{X}})$,

$$\begin{aligned}
&\mathcal{D}(D_2, P_{\tilde{X}}) \\
&= \{P_{Y|\tilde{X}} \in \mathcal{P}(\mathcal{Y}|\mathcal{X}): \; E_{P_{\tilde{X}} P_{Y|\tilde{X}}} d_2(\tilde{X}, Y) \leq D_2\}
\end{aligned} \tag{71}$$

and the requirement (64) yields $\gamma_n(b_n) \to 0$.

Recall the definition of Z in (53), the fact that $\tilde{Z} = (Z, T)$, and the definition of Z' (68), hence the alphabet of Z' in (70) has cardinality that is upper bounded by $|\mathcal{M}_n| \times |\mathcal{K}_n| \times |\mathcal{U}|^n \times |\mathcal{Y}|^n \times n$. In the sequel we shall quantify the cost of working in a lower cardinality alphabet (that does not depend on n).

Let $\ell \geq 1$ be a given integer. Define the set $P^{(\ell)}(\mathcal{Y}|\mathcal{X})$ as the set of pmfs $P_{Y|X} \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$ such that for every (x, y), $P_{Y|X}(y|x) = \frac{m}{\ell}$, for some integer m. Neglecting the term $\gamma_n(b_n)$, the r.h.s. of (70) is clearly upper bounded by

$$\begin{aligned}
&\max_{P_{\tilde{X}Z'|U}: \, Ed_1(U,\tilde{X}) \leq D_1} \; \min_{P_{Y|\tilde{X}} \in P^{(\ell)}(\mathcal{Y}|\mathcal{X}) \cap \mathcal{D}(D_2 - \alpha_n, P_{\tilde{X}})} \\
&J(P_{UZ'\tilde{X}} \times P_{Y|\tilde{X}}).
\end{aligned} \tag{72}$$

Lemma 7. *It is sufficient to consider $|\mathcal{Z}'| \leq |\mathcal{X}| \cdot |\mathcal{U}| + 1 + (\ell + 1)^{|\mathcal{X}||\mathcal{Y}|}$ in (72).*

The following lemma concludes the proof of the converse part.

Lemma 8. *Whenever $\ell > \max\{|\mathcal{Y}|^2, \frac{D_{\max}|\mathcal{Y}|}{(D_{\min}/4 - \alpha_n)}\}$ the quantity (72) is upper bounded by*

$$\begin{aligned}
&\max_{P_{\tilde{X}Z'|U}: \, Ed_1(U,\tilde{X}) \leq D_1} \; \min_{P_{Y|\tilde{X}} \in \mathcal{D}(D_2, P_{\tilde{X}})} J(P_{UZ'\tilde{X}} \times P_{Y|\tilde{X}}) \\
&+ 2\Delta\left(\frac{|\mathcal{Y}|^2}{\ell}, |\mathcal{Y}|\right) + 2\Delta(2\beta_{n,\ell}|\mathcal{Y}|, |\mathcal{Y}|),
\end{aligned} \tag{73}$$

where $\beta_{n,\ell} = \frac{\alpha_n + D_{\max}|\mathcal{Y}|/\ell}{|\mathcal{Y}| D_{\min}}$.

Gathering (42)-(73) we obtain that for all $\ell > \max\{|\mathcal{Y}|^2, \frac{D_{\max}|\mathcal{Y}|}{(D_{\min}/4 - \alpha_n)}\}$,

$$\begin{aligned}
R \leq &\liminf_{n \to \infty} \; \max \; \min J(P_{UZ'\tilde{X}} \times P_{Y|\tilde{X}}) \\
&+ 2\Delta\left(\frac{|\mathcal{Y}|^2}{\ell}, |\mathcal{Y}|\right) + 2\Delta(2\beta_{n,\ell}|\mathcal{Y}|, |\mathcal{Y}|),
\end{aligned} \tag{74}$$

where the maximum is over the set $\{P_{Z'\tilde{X}|U} : \; Ed_1(U, \tilde{X}) \leq D_1\}$ and the minimization is over the set $\{P_{Y|\tilde{X}} \in \mathcal{D}(D_2, P_{\tilde{X}})\}$ and it is easily verified that by definition

$$\begin{aligned}
&\lim_{n \to \infty} 2\Delta\left(\frac{|\mathcal{Y}|^2}{\ell}, |\mathcal{Y}|\right) + 2\Delta(2\beta_{n,\ell}|\mathcal{Y}|, |\mathcal{Y}|) \\
&= \; 2 \cdot \frac{|\mathcal{Y}|^2}{\ell} \log \frac{\ell}{|\mathcal{Y}|} + 4 \cdot \frac{D_{\max}|\mathcal{Y}|}{\ell D_{\min}} \log\left(\frac{\ell D_{\min}}{2 D_{\max}}\right)
\end{aligned} \tag{75}$$

and the r.h.s. expression, viewed as a function of ℓ, is obviously $O\left(\frac{\log \ell}{\ell}\right)$, and can be upper bounded by $h \frac{\log \ell}{\ell}$ where h is a constant depending only on $D_{\min}$, $D_{\max}$, and $|\mathcal{Y}|$, as asserted in the Theorem.

7 Acknowledgement

The authors wish to thank the reviewers for their very useful comments.

Appendix A

Proof of Lemma 1

Inequality (28) holds since

$$\begin{aligned}
&\Pr\left\{\mathbf{u}:\ \|\hat{P}_{\mathbf{u}} - P_U\| > \eta_n\right\} \\
&= \Pr\left\{T_{\mathbf{u}} : \|\hat{P}_{\mathbf{u}} - P_U\| > \eta_n\right\} \\
&\leq (n+1)^{|\mathcal{U}|} \max_{T_{\mathbf{u}}:\ \|\hat{P}_{\mathbf{u}} - P_U\| > \eta_n} \Pr\left(T_{\mathbf{u}}\right) && \text{(A. 1)} \\
&\leq (n+1)^{|\mathcal{U}|} \max_{T_{\mathbf{u}}:\ \|\hat{P}_{\mathbf{u}} - P_U\| > \eta_n} e^{-D\left(\hat{P}_{\mathbf{u}}\|P_U\right)} \\
&\leq (n+1)^{|\mathcal{U}|} \max_{T_{\mathbf{u}}:\ \sqrt{D\left(\hat{P}_{\mathbf{u}}\|P_U\right)2\ln 2} > \eta_n} e^{-D\left(\hat{P}_{\mathbf{u}}\|P_U\right)} && \text{(A. 2)} \\
&\leq (n+1)^{|\mathcal{U}|} e^{-n\frac{\eta_n^2}{2\ln 2}}, && \text{(A. 3)}
\end{aligned}$$

where (A. 1) follows since the number of types is polynomial, (A. 2) follows due to Pinsker's inequality ([9]) according to which for any two pmf's P_1, P_2 defined on the same alphabet,

$$\|P_1 - P_2\| \leq \sqrt{D\left(P_1\|P_2\right)2\ln 2}. \tag{A. 4}$$

In order to prove (29), first, note that for any given sequence $\mathbf{u}$ such that $\|\hat{P}_{\mathbf{u}} - P_U\| \leq \eta_n$ and any given sequence $\mathbf{v}$,

$$\begin{aligned}
&\|\hat{P}_{\mathbf{uv}} - P_{UV}\| \\
&= \sum_{uv} \left|\hat{P}_{\mathbf{u}}(u)\hat{P}_{\mathbf{v}|\mathbf{u}}(v|u) - P_U(u)P_{V|U}(v|u)\right| \\
&\leq \sum_{uv} \left(\hat{P}_{\mathbf{u}}(u)\left|\hat{P}_{\mathbf{v}|\mathbf{u}}(v|u) - P_{V|U}(v|u)\right| + \eta_n P_{V|U}(v|u)\right) \\
&= \sum_{uv} \hat{P}_{\mathbf{u}}(u)\left|\hat{P}_{\mathbf{v}|\mathbf{u}}(v|u) - P_{V|U}(v|u)\right| + \eta_n|\mathcal{U}|. && \text{(A. 5)}
\end{aligned}$$

According to Lemma 2.1.2 in [10], there exists an empirical measure $\hat{P}_{\mathbf{v}|\mathbf{u}}$ such that $\left|\hat{P}_{\mathbf{v}|\mathbf{u}}(v|u) - P_{V|U}(v|u)\right| \leq \frac{|\mathcal{V}|}{2n}$, which together with (A. 5), implies that whenever $\delta_n \geq \eta_n|\mathcal{U}| + \frac{|\mathcal{V}|}{2n}$, for any given sequence $\mathbf{u}$ such that $\|\hat{P}_{\mathbf{u}} - P_U\| \leq \eta_n$, the set $\left\{T_{\mathbf{v}|\mathbf{u}}:\ \|\hat{P}_{\mathbf{uv}} - P_{UV}\| \leq \delta_n\right\}$ is non-empty.

Thus, for any given sequence $\mathbf{u}$ satisfying $\|\hat{P}_{\mathbf{u}} - P_U\| \leq \eta_n$, we have,

$$\begin{aligned}
&\Pr\left\{\|\hat{P}_{\mathbf{U}\mathbf{V}_{1,j}} - P_{UV}\| \leq \delta_n | \mathbf{U} = \mathbf{u}\right\} \\
&= \sum_{T_{\mathbf{v}|\mathbf{u}}:\, \|\hat{P}_{\mathbf{u}\mathbf{v}} - P_{UV}\| \leq \delta_n} |T_{\mathbf{v}|\mathbf{u}}| e^{-n\left[D(\hat{P}_{\mathbf{v}}\|P_V) + \hat{H}_{\mathbf{v}}\right]} && \text{(A. 6)} \\
&\geq \max_{T_{\mathbf{v}|\mathbf{u}}:\, \|\hat{P}_{\mathbf{u}\mathbf{v}} - P_{UV}\| \leq \delta_n} |T_{\mathbf{v}|\mathbf{u}}| e^{-n\left[D(\hat{P}_{\mathbf{v}}\|P_V) + \hat{H}_{\mathbf{v}}\right]} \\
&\geq \max_{T_{\mathbf{v}|\mathbf{u}}:\, \|\hat{P}_{\mathbf{u}\mathbf{v}} - P_{UV}\| \leq \delta_n} \frac{e^{-n\left[D(\hat{P}_{\mathbf{v}}\|P_V) + \hat{H}_{\mathbf{v}} - \hat{H}_{\mathbf{v}|\mathbf{u}}\right]}}{(n+1)^{|\mathcal{U}||\mathcal{V}|}} && \text{(A. 7)} \\
&\geq \max_{T_{\mathbf{v}|\mathbf{u}}:\, \|\hat{P}_{\mathbf{u}\mathbf{v}} - P_{UV}\| \leq \delta_n} \frac{1}{(n+1)^{|\mathcal{U}||\mathcal{V}|}} e^{-n\left[\frac{\delta_n^2}{2\ln 2} + \hat{I}_{\mathbf{v};\mathbf{u}}\right]} && \text{(A. 8)} \\
&\geq e^{-n\left[I(U;V) + \frac{\delta_n^2}{2\ln 2} - 3\delta_n \log\left(\frac{\delta_n}{|\mathcal{U}||\mathcal{V}|}\right) + |\mathcal{U}||\mathcal{V}|\frac{\log(n+1)}{n}\right]} && \text{(A. 9)} \\
&= e^{-n[I(U;V) + \epsilon_n - \tilde{\epsilon}_n]}. && \text{(A. 10)}
\end{aligned}$$

where (A. 6) holds since the random vector $\mathbf{V}_{1,j}$ has i.i.d. components with marginal P_V and is independent of $\mathbf{U}$, (A. 7) follows since $|T_{\mathbf{v}|\mathbf{u}}| \geq \frac{1}{(n+1)^{|\mathcal{U}||\mathcal{V}|}} e^{n\hat{H}_{\mathbf{v}|\mathbf{u}}}$, (A. 8) is due to Pinsker's inequality (see (A. 4)), (A. 9) is easily verified as $\|\hat{P}_{\mathbf{u}\mathbf{v}} - P_{UV}\| \leq \delta_n \Rightarrow \left|\hat{I}_{\mathbf{v};\mathbf{u}} - I(V;U)\right| \leq -3\delta_n \log\left(\frac{\delta_n}{|\mathcal{U}||\mathcal{V}|}\right)$, and ϵ_n is defined in (26). Thus,

$$\begin{aligned}
&\Pr\left\{A_2(\delta_n)|\bar{A}_1(\eta_n)\right\} \\
&= \Pr\left\{\forall j,\ \|\hat{P}_{\mathbf{U}\mathbf{V}_{1,j}} - P_{UV}\| > \delta_n | \bar{A}_1(\eta_n)\right\} \\
&= \sum_{\mathbf{u}:\, \|\hat{P}_{\mathbf{u}} - P_U\| \leq \eta_n} P_U^n(\mathbf{u}) \\
&\times \prod_{j=1}^{J} \left[1 - \Pr\left\{\|\hat{P}_{\mathbf{U}\mathbf{V}_{1,j}} - P_{UV}\| \leq \delta_n | \mathbf{U} = \mathbf{u}\right\}\right] \\
&\leq \sum_{\mathbf{u}:\, \|\hat{P}_{\mathbf{u}} - P_U\| \leq \eta_n} P_U^n(\mathbf{u}) \prod_{j=1}^{J} \left[1 - e^{-n[I(U;V) + \epsilon_n - \tilde{\epsilon}_n]}\right] \\
&\leq \left[1 - e^{-n[I(U;V) + \epsilon_n - \tilde{\epsilon}_n]}\right]^J \\
&= \left[1 - e^{-n[I(U;V) + \epsilon_n - \tilde{\epsilon}_n]}\right]^{e^{-n[I(U;V) + \epsilon_n]}} \leq e^{-e^{n\tilde{\epsilon}_n}} && \text{(A. 11)}
\end{aligned}$$

Proof of Lemma 2

First, note that since $\mathbb{P}_n(\mathcal{Y}, \hat{P}_{\mathbf{uv}}\xi_n(\hat{P}_{\mathbf{uv}})) \subseteq \mathcal{P}(\mathcal{Y}|\mathcal{U}\times\mathcal{V}\times\mathcal{X})$, we have for all $\hat{P}_{\mathbf{uv}} \in \mathcal{E}_n$,

$$\begin{aligned}&\max_{\hat{P}_{\mathbf{y}|\mathbf{uvx}}:\, d_2(\mathbf{x},\mathbf{y})\leq nD_2} \exp\{-nG_{R,J,n}(\hat{P}_{\mathbf{uv}}\xi_n(\hat{P}_{\mathbf{uv}})\hat{P}_{\mathbf{y}|\mathbf{uvx}})\} \\ &\leq \exp\left\{-n \inf_{P_{Y|UVX}} G_{R,J,n}(\hat{P}_{\mathbf{uv}}\xi_n(\hat{P}_{\mathbf{uv}})P_{Y|UVX})\right\} \qquad \text{(A. 12)}\end{aligned}$$

where the infimum is over the set $P_{Y|UVX}:\ E_{\hat{P}_{\mathbf{uv}}\xi_n(\hat{P}_{\mathbf{uv}})P_{Y|UVX}} d_2(X,Y) \leq D_2$.

Now, observe that (a) if $\hat{P}_{\mathbf{uv}} \in \mathcal{E}_n$, we have $||\hat{P}_{\mathbf{uv}} - P_{UV}|| \leq \delta_n \to 0$, (b) the mapping ξ_n is required to be continuous around the point P_{UV} (see eq. (18)), (c) the expectation $E_P d_2(X,Y)$ and the operator $G_{R,J,n}(P)$ are uniformly continuous w.r.t. P, and thus, the minimization over $\{\hat{P}_{\mathbf{uv}} \in \mathcal{E}_n\}$ as in (36) can be replaced by P_{UV} with vanishingly small loss. This fact, together with (A. 12) establishes Lemma 2.

Proof of Lemma 3

Proof.

$$\begin{aligned}&\Pr\{d_2(\mathbf{X},\mathbf{Y}) > nD_2|\mathbf{X}=\mathbf{x}\} \\ &\leq \sum_{\mathbf{y}:\, d_2(\mathbf{x},\mathbf{y})>nD_2} e^{-nD(\hat{P}_{\mathbf{y}|\mathbf{x}}||\Lambda(\hat{P}_{\mathbf{x}})|\hat{P}_{\mathbf{x}})} \\ &\leq (n+1)^{|\mathcal{X}||\mathcal{Y}|}\exp\{-n\inf D(P_{Y|X}\|\Lambda(\hat{P}_{\mathbf{x}})|\hat{P}_{\mathbf{x}})\}, \qquad \text{(A. 13)}\end{aligned}$$

where the infimum is over the set $\{P_{Y|X} : E_{\hat{P}_{\mathbf{x}}P_{Y|X}} d_2(X,Y) > D_2\}$. Now, for every $\hat{P}_{\mathbf{x}}$ and $P_{Y|X}$ such that $E_{\hat{P}_{\mathbf{x}}P_{Y|X}} d_2(X,Y) > D_2$ we have

$$\begin{aligned}&\epsilon \leq E_{\hat{P}_{\mathbf{x}}P_{Y|X}} d_2(X,Y) - E_{\hat{P}_{\mathbf{x}}\Lambda(\hat{P}_{\mathbf{x}})} d_2(X,Y) \\ =&\ \sum_{x,y} d_2(x,y)\hat{P}_{\mathbf{x}}(x)\left(P_{Y|X}(y|x) - \Lambda(\hat{P}_{\mathbf{x}})(y|x)\right) \\ \leq&\ |\mathcal{X}||\mathcal{Y}|D_{\max}\max_{x,y}\hat{P}_{\mathbf{x}}(x)\left(P_{Y|X}(y|x) - \Lambda(\hat{P}_{\mathbf{x}})(y|x)\right) \\ \leq&\ |\mathcal{X}||\mathcal{Y}|D_{\max}\|\hat{P}_{\mathbf{x}}\times P_{Y|X} - \hat{P}_{\mathbf{x}}\times\Lambda(\hat{P}_{\mathbf{x}})\|. \qquad \text{(A. 14)}\end{aligned}$$

According to Pinsker's inequality

$$D(P\|Q) \geq \frac{\|P-Q\|^2}{2\ln 2}, \qquad \text{(A. 15)}$$

and similarly, using Jensen's inequality, one obtains,

$$D(P_{Y|X}\|\Lambda(\hat{P}_{\mathbf{x}})|\hat{P}_{\mathbf{x}})\} \geq \frac{\|\hat{P}_{\mathbf{x}}\times P_{Y|X} - \hat{P}_{\mathbf{x}}\times\Lambda(\hat{P}_{\mathbf{x}})\|^2}{2\ln 2} \qquad \text{(A. 16)}$$

we have

$$D(P_{Y|X}\|\Lambda(\hat{P}_{\mathbf{x}})|\hat{P}_{\mathbf{x}}) \geq \tfrac{\epsilon^2}{(2\ln 2)D_{\max}^2|\mathcal{X}|^2|\mathcal{Y}|^2}. \qquad \text{(A. 17)}$$

Eqs. (A. 13) and (A. 17) yield

$$\Pr\{d_2(\mathbf{X},\mathbf{Y}) > nD_2|\mathbf{X}=\mathbf{x}\} \leq d_n(\epsilon). \qquad \text{(A. 18)}$$

□

Proof of Lemma 4

Proof. Let P^Λ be given, and consider the channel $P_2 \in \mathcal{P}_n^{d_2}(D_2)$ defined by

$$\begin{aligned} & P_2(\mathbf{y}|\mathbf{x}) \\ = & \left\{ \begin{array}{cc} 0 & \text{if } d_2(\mathbf{x},\mathbf{y}) > nD_2 \\ \frac{P^\Lambda(\mathbf{y}|\mathbf{x})}{P^\Lambda(\mathbf{y}'\colon d_2(\mathbf{x},\mathbf{y}')\leq nD_2|\mathbf{X}=\mathbf{x})} & \text{otherwise} \end{array} \right. . \end{aligned}$$

We have by Lemma 3,

$$\begin{aligned} & \sum_{\mathbf{y}} \left|P^\Lambda(\mathbf{y}|\mathbf{x}) - P_2(\mathbf{y}|\mathbf{x})\right| \\ & = P^\Lambda(d_2(\mathbf{X},\mathbf{Y}) > nD_2|\mathbf{X}=\mathbf{x}) \\ & \quad + \sum_{\mathbf{y}:d_2(\mathbf{x},\mathbf{y})\leq nD_2} P^\Lambda(\mathbf{y}|\mathbf{x}) \\ & \quad \times \left[\frac{1}{P^\Lambda(\mathbf{y}':\ d_2(\mathbf{x},\mathbf{y}') \leq nD_2|\mathbf{X}=\mathbf{x})} - 1\right] \\ & = P^\Lambda(d_2(\mathbf{X},\mathbf{Y}) > nD_2|\mathbf{X}=\mathbf{x}) \\ & \quad + 1 - P^\Lambda(d_2(\mathbf{X},\mathbf{Y}) \leq nD_2|\mathbf{X}=\mathbf{x}) \\ & = 2P^\Lambda(d_2(\mathbf{X},\mathbf{Y}) > nD_2|\mathbf{X}=\mathbf{x}) \leq 2d_n(\epsilon). \end{aligned} \qquad \text{(A. 19)}$$

□

Proof of Lemma 5

Proof. If constraint (64) is met, then for all P such that $\|P - P_{\tilde{X}}\| \le b_n + \delta(b_n)$,

$$\begin{aligned}
&E_{P\times\Lambda(P)}d_2(X,Y) - E_{P_{\tilde{X}}\times\Lambda(P_{\tilde{X}})}d_2(X,Y)\\
&= \sum_{x,y}\left[P(x)\Lambda(P)(y|x) - P_{\tilde{X}}(x)\Lambda(P_{\tilde{X}})(y|x)\right]d_2(x,y)\\
&\stackrel{(a)}{=} \sum_{x,y}\left[P(x) - P_{\tilde{X}}(x)\right]\Lambda(P_{\tilde{X}})(y|x)d_2(x,y)\\
&\le (b_n + \delta(b_n))\sum_{x,y}\Lambda(P_{\tilde{X}})(y|x)d_2(x,y)\\
&\le (b_n + \delta(b_n))\,|\mathcal{X}||\mathcal{Y}|D_{\max} = \alpha_n - \epsilon_n
\end{aligned} \tag{A. 20}$$

where (a) follows from (64). In other words, for all P such that $\|P - P_{\tilde{X}}\| \le b_n + \delta(b_n)$, if Λ is such that $E_{P_{\tilde{X}}\times\Lambda(P_{\tilde{X}})}d_2(X,Y) \le D_2 - \alpha_n$, then, $E_{P\times\Lambda(P)}d_2(X,Y) \le D_2 - \epsilon_n$, i.e., $\Lambda(\cdot) \in \mathcal{N}(D_2 - \epsilon_n)$. □

Proof of Lemma 6

Proof. Consider the event

$$\mathcal{E}_0 = \left\{\mathbf{u} :\ \|\hat{P}_{\mathbf{u}} - P_U\| \le \sqrt{(2\ln 2)\epsilon}\right\} \tag{A. 21}$$

obviously,

$$\begin{aligned}
&1 - P_U^n\{\mathcal{E}_0\}\\
&\le (n+1)^{|\mathcal{U}|}e^{-n\inf_{P\in\mathcal{P}(\mathcal{U}):\ \|P-P_U\|>\sqrt{(2\ln 2)\epsilon}} D(P\|P_U)}\\
&\le (n+1)^{|\mathcal{U}|}e^{-n\inf_{P\in\mathcal{P}(\mathcal{U}):\ D(P\|P_U)>\epsilon} D(P\|P_U)} && \text{(A. 22)}\\
&\le (n+1)^{|\mathcal{U}|}e^{-n\epsilon} && \text{(A. 23)}
\end{aligned}$$

where (A. 22) is due to Pinsker's inequality (A. 15) which implies that if $\|P - P_U\| > \sqrt{(2\ln 2)\epsilon}$ then $D(P\|P_U) > \epsilon$.

Thus, for all $\mathbf{u}$

$$\begin{aligned}
&1 - (n+1)^{|\mathcal{U}|}e^{-n(\delta^2(b_n)(2\ln 2))}\\
&\stackrel{(a)}{\le} P_U^n\left\{\mathbf{u} : \|\hat{P}_{\mathbf{u}} - P_U\| \le \delta(b_n)\right\}\\
&\stackrel{(b)}{=} P_U^n\left\{\mathbf{u} : \begin{array}{c}\|\hat{P}_{\mathbf{u}} - P_U\| \le \delta(b_n),\\ \|\theta(\hat{P}_{\mathbf{u}})(\cdot|u) - \theta(P_U)(\cdot|u)\| \le b_n \forall u\end{array}\right\}\\
&\stackrel{(c)}{\le} P_U^n\left\{\mathbf{u} : \|\hat{P}_{\mathbf{u}} \times \theta(\hat{P}_{\mathbf{u}}) - P_U \times \theta(P_U)\| \le b_n + \delta(b_n)\right\}
\end{aligned} \tag{A. 24}$$

where (a) follows from (A. 23), (b) follows from (12) and (c) is because

$$
\begin{aligned}
& \|P_1(U,X) - P_2(U,X)\| = \sum_{u,x} |P_1(u,x) - P_2(u,x)| \\
= \ & \sum_{u,x} |P_1(u)\left(P_1(x|u) - P_2(x|u)\right) + \left(P_1(u) - P_2(u)\right) P_2(x|u)| \\
\leq \ & \sum_{u} P_1(u)\|P_1(\cdot|u) - P_2(\cdot|u)\| + |\mathcal{X}| \cdot \|P_1(U) - P_2(U)\|.
\end{aligned} \tag{A. 25}
$$

Inequality (A. 24) implies that with arbitrarily high probability (for sufficiently large n) the empirical statistics of $\mathbf{X}$ is close (in the sense of variational distance) to the marginal of $P_U \times \theta(P_U)$.

Lemma 9. *Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be two events defined on the same probability space, such that $Pr\{\mathcal{E}_1\} \geq 1 - \delta$ where $\delta \leq \frac{1}{2}$, then*

$$
|Pr\{\mathcal{E}_2|\mathcal{E}_1\} - Pr\{\mathcal{E}_2\}| \leq 2\delta. \tag{A. 26}
$$

The proof of Lemma 9 appears in the sequel of the appendix. Now, recall (A. 24) and the definition of $\mathcal{E}_1$ (67). In the sequel, we shall use the notation $a \overset{\delta}{\approx} b$ to abbreviate $|a-b| \leq 2(n+1)^{|\mathcal{U}|}e^{-n\delta^2(b_n)(2\ln 2)}$, we have by lemma 9 applied once to the event $\{U = u, \tilde{Z} = z, X = x, Y = y\}$ in the role of $\mathcal{E}_2$

$$
\begin{aligned}
& \Pr\{U = u, \tilde{Z} = z, X = x, Y = y\} \\
\overset{\delta}{\approx} \ & \Pr\{U = u, \tilde{Z} = z, X = x, Y = y|\mathcal{E}_1\} \\
= \ & \Pr\{U = u, \tilde{Z} = z, X = x|\mathcal{E}_1\} \\
& \cdot \Pr\{Y = y|U = u, \tilde{Z} = z, X = x, \mathcal{E}_1\}.
\end{aligned} \tag{A. 27}
$$

Recall that it is assumed that the attacker satisfies (64) and hence, for sufficiently large n,

$$
\begin{aligned}
& \Pr\{Y = y|U = u, \tilde{Z} = z, X = x, \mathcal{E}_1\} \\
= \ & \Lambda(P_{\tilde{X}})(y|x).
\end{aligned} \tag{A. 28}
$$

Next, we have on the one hand,

$$
\begin{aligned}
& \Pr\{U = u, X = x|\mathcal{E}_1\} \\
\leq \ & \max_{\hat{P}_{\mathbf{u}}:\, \|\hat{P}_{\mathbf{u}}\theta(\hat{P}_{\mathbf{u}}) - P_U\theta(P_U)\| \leq b_n + \delta(b_n)} \Pr\left\{U = u, X = x|\hat{P}_{\mathbf{u}}\right\} \\
\leq \ & P_U(u)\theta(P_U)(x|u) + b_n + \delta(b_n)
\end{aligned} \tag{A. 29}
$$

and on the other hand

$$
\begin{aligned}
& \Pr\{U = u, X = x|\mathcal{E}_1\} \\
\geq \ & \min_{\hat{P}_{\mathbf{u}}:\, \|\hat{P}_{\mathbf{u}}\theta(\hat{P}_{\mathbf{u}}) - P_U\theta(P_U)\| \leq b_n + \delta(b_n)} \Pr\left\{U = u, X = x|\hat{P}_{\mathbf{u}}\right\} \\
\geq \ & P_U(u)\theta(P_U)(x|u) - b_n - \delta(b_n),
\end{aligned} \tag{A. 30}
$$

thus

$$|\Pr\{U=u,X=x|\mathcal{E}_1\}-P_U(u)\theta(P_U)(x|u)|\leq b_n+\delta(b_n). \tag{A. 31}$$

By definition of Z' (see (68)), $\Pr\{\tilde{Z}=z|U=u,X=x,\mathcal{E}_1\}=\Pr\{Z'=z|U=u,X=x\}$, and as a result of (A. 27), (A. 28) and (A. 31) we have for all (u,z,x,y) and sufficiently large n,

$$\begin{aligned}
&\left|\Pr\{U=u,\tilde{Z}=z,X=x,Y=y\}-P_U(u)\theta(P_U)(x|u)P_{Z'|XU}(z|xu)\Lambda(P_{\tilde{X}})(y|x)\right| \\
\leq\ & 2(n+1)^{|\mathcal{U}|}e^{-n\delta^2(b_n)(2\ln 2)}+b_n+\delta(b_n). \qquad\qquad \text{(A. 32)}
\end{aligned}$$

Recall that $P_{\tilde{X}|U}\triangleq\theta(P_U)$, consequently,

$$\begin{aligned}
&\left|J(P_{UX\tilde{Z}\hat{P}_{\mathbf{X}}}\times\Lambda(\hat{P}_{\mathbf{X}}))-J(P_{U\tilde{X}}P_{Z'|UX}\times\Lambda(P_X))\right| \\
=\ &\left|I_{P_{UX\tilde{Z}\hat{P}_{\mathbf{X}}}\times\Lambda(\hat{P}_{\mathbf{X}})}(\tilde{Z};Y)-I_{P_{U\tilde{X}}P_{Z'|UX}\times\Lambda(P_X)}(\tilde{Z};Y)\right| \\
\leq\ &\left|H_{P_{UX\tilde{Z}\hat{P}_{\mathbf{X}}}\times\Lambda(\hat{P}_{\mathbf{X}})}(Y)-H_{P_{U\tilde{X}}P_{Z'|UX}\times\Lambda(P_X)}(Y)\right| \\
&+\left|H_{P_{UX\tilde{Z}\hat{P}_{\mathbf{X}}}\times\Lambda(\hat{P}_{\mathbf{X}})}(Y|\tilde{Z})-H_{P_{U\tilde{X}}P_{Z'|UX}\times\Lambda(P_{\tilde{X}})}(Y|\tilde{Z})\right| \\
\leq\ & 2\Delta\left(b_n+\delta(b_n)+2(n+1)^{|\mathcal{U}|}\exp\{-n\delta^2(b_n)(2\ln 2)\},|\mathcal{Y}|\right) \qquad\qquad \text{(A. 33)}
\end{aligned}$$

where the last step follows from (A. 32) and (49). □

Proof of Lemma 7

Proof. To prove the lemma we will use the support lemma [9, p. 310] based on Carathéodory's theorem, according to which, given L real valued continuous functionals $f_l, l=1,...,L$ on the set $\mathcal{P}(\mathcal{X})$, and given any probability measure μ on the Borel σ-algebra of $\mathcal{P}(\mathcal{X})$, there exist L elements $Q_1,..,Q_L$ of $\mathcal{P}(\mathcal{X})$ and L non-negative real numbers $\alpha_1,...,\alpha_L$ such that $\sum_{l=1}^L\alpha_l=1$ and for every $l=1,...,L$

$$\int_{\mathcal{P}(\mathcal{X})}f_l(Q)\mu(dQ)=\sum_{i=1}^L\alpha_i f_l(Q_i). \tag{A. 34}$$

Let $P_{U\tilde{X}Z'}$ be a given, we shall show that there exists another RV $\tilde{Z}'$ with alphabet cardinality $|\tilde{\mathcal{Z}}'|\leq|\mathcal{X}|\cdot|\mathcal{U}|+1+(\ell+1)^{|\mathcal{X}||\mathcal{Y}|}$ such that

$$\begin{aligned}
&\sum_{z\in\mathcal{Z}}P_{Z'}(z)P_{U\tilde{X}|Z'}(u,x|z) \\
=\ &\sum_{z'\in\mathcal{Z}'}P_{\tilde{Z}'}(z')P_{U\tilde{X}|Z'}(u,x|z') \quad \forall u\in\mathcal{U},x\in\mathcal{X} \qquad\qquad \text{(A. 35)} \\
&J(P_{UZ'\tilde{X}}\times P_{Y|\tilde{X}}) \\
=\ &J(P_{U\tilde{Z}'\tilde{X}}\times P_{Y|\tilde{X}}) \quad \forall P_{Y|\tilde{X}}\in P^{(\ell)}(\mathcal{Y}|\mathcal{X}). \qquad\qquad \text{(A. 36)}
\end{aligned}$$

Note that

$$\begin{aligned} & J(P_{UZ'\tilde{X}} \times P_{Y|\tilde{X}}) \\ = \;& H(Y) - H(Y|Z') + H(U|Z') - H(U). \end{aligned} \tag{A. 37}$$

The entropy $H(Y)$ can be viewed as a function of $P_{\tilde{X}Y}$ and $H(U)$ is a function of P_U, both are unaffected by Z' (provided that (A. 35) holds). It is therefore sufficient to prove that

$$H(U|Z') - H(Y|Z') \;=\; H(U|\tilde{Z}') - H(Y|\tilde{Z}'), \tag{A. 38}$$

or, more explicitly,

$$\begin{aligned} & \sum_{z\in\mathcal{Z}} P_{Z'}(z)\,[H(U|Z'=z) - H(Y|Z'=z)] \\ = \;& \sum_{z\in\mathcal{Z}'} P_{\tilde{Z}'}(z')\left[H(U|\tilde{Z}'=z') - H(Y|\tilde{Z}'=z')\right], \end{aligned} \tag{A. 39}$$

for every $P_{Y|\tilde{X}} \in P^{(\ell)}(\mathcal{Y}|\mathcal{X})$.

In view of equations (A. 35) and (A. 39) we now invoke the support lemma. Let $l_1(u,x)$ be a function that indexes all the elements $(u,x) \in \mathcal{U}\times\mathcal{X}$ (but one) and let $l_2(P_{Y|\tilde{X}})$ be a function that indexes all the elements of $P^{(\ell)}(\mathcal{Y}|\mathcal{X})$. Define

$$\begin{aligned} & f_{l_1(u,x)}(P_{U\tilde{X}|Z'=z}) = P_{U\tilde{X}|Z'=z}(u,x|z) \\ & f_{|\mathcal{U}||\mathcal{X}|-1+l_2(P_{Y|\tilde{X}})}(P_{U\tilde{X}|Z'=z}) \\ & = H(U|Z'=z) - H(Y|Z'=z). \end{aligned} \tag{A. 40}$$

Since the number of elements of $P^{(\ell)}(\mathcal{Y}|\mathcal{X})$ is upper bounded by $(\ell+1)^{|\mathcal{X}||\mathcal{Y}|}$ the lemma follows from the support lemma. □

Proof of Lemma 8

Proof. The proof of the lemma is divided to the following two parts denoted (a) and (b):

$$\begin{aligned} & \max_{P_{\tilde{X}Z'|U}:\, Ed_1(U,\tilde{X})\le D_1}\; \min_{P_{Y|\tilde{X}}\in P^{(\ell)}(\mathcal{Y}|\mathcal{X})\cap\mathcal{D}(D_2-\alpha_n,P_{\tilde{X}})} J(P_{UZ'\tilde{X}} \times P_{Y|\tilde{X}}) \\ \overset{(a)}{\le} \;& \max_{P_{\tilde{X}Z'|U}:\, Ed_1(U,\tilde{X})\le D_1}\; \min_{P_{Y|\tilde{X}}\in\mathcal{D}\left(D_2-\alpha_n-\frac{D_{\max}|\mathcal{Y}|}{\ell},P_{\tilde{X}}\right)} J(P_{UZ'\tilde{X}} \times P_{Y|\tilde{X}}) + 2|\mathcal{Y}|\frac{\log\ell}{\ell} \\ \overset{(b)}{\le} \;& \max_{P_{\tilde{X}Z'|U}:\, Ed_1(U,\tilde{X})\le D_1}\; \min_{P_{Y|\tilde{X}}\in\mathcal{D}(D_2,P_{\tilde{X}})} J(P_{UZ'\tilde{X}} \times P_{Y|\tilde{X}}) + 2|\mathcal{Y}|\frac{\log\ell}{\ell} \\ & +2\Delta(2\beta_{n,\ell}|\mathcal{Y}|,|\mathcal{Y}|). \end{aligned} \tag{A. 41}$$

Proof of part (a)

Let $P_{Y|\tilde{X}} \in \mathcal{D}\left(D_2 - \alpha_n - \frac{D_{\max}|\mathcal{Y}|}{\ell}, P_{\tilde{X}}\right)$ be given. Let $P^{(\ell)}_{Y|\tilde{X}}$ be defined as a quantized version of $P_{Y|\tilde{X}}$, i.e.,

$$|P^{(\ell)}_{Y|\tilde{X}}(y|x) - P_{Y|\tilde{X}}(y|x)| \leq \frac{|\mathcal{Y}|}{\ell} \quad \forall (x, y). \tag{A. 42}$$

This yields,

$$\begin{aligned}
&\sum_y |P^{(\ell)}_{Y|\tilde{X}}(y|x) - P_{Y|\tilde{X}}(y|x)| \leq \frac{|\mathcal{Y}|^2}{\ell} \quad \forall x \in \mathcal{X} \\
&\sum_x P_{\tilde{X}}(x) \sum_y |P^{(\ell)}_{Y|\tilde{X}}(y|x) - P_{Y|\tilde{X}}(y|x)| \leq \frac{|\mathcal{Y}|^2}{\ell} \\
&\sum_x P_{\tilde{X}}(x|z) \sum_y |P^{(\ell)}_{Y|\tilde{X}}(y|x) - P_{Y|\tilde{X}}(y|x)| \leq \frac{|\mathcal{Y}|^2}{\ell} \quad \forall z \in \mathcal{Z}
\end{aligned} \tag{A. 43}$$

and hence, applying (49), we have for all $P_{U\tilde{X}Z'}$,

$$\begin{aligned}
&\left| J(P_{UZ'\tilde{X}} \times P_{Y|\tilde{X}}) - J(P_{UZ'\tilde{X}} \times P^{(\ell)}_{Y|\tilde{X}}) \right| \\
= &\left| I_{P_{UZ'\tilde{X}} \times P_{Y|\tilde{X}}}(\tilde{Z}; Y) - I_{P_{UZ'\tilde{X}} \times P^{(\ell)}_{Y|\tilde{X}}}(\tilde{Z}; Y) \right| \\
\leq & \; 2\Delta\left(\frac{|\mathcal{Y}|^2}{\ell}, |\mathcal{Y}|\right).
\end{aligned} \tag{A. 44}$$

Moreover, eq. (A. 42) implies that for all $P_{\tilde{X}}$

$$\left| E_{P_{\tilde{X}} P_{Y|\tilde{X}}} d_2(\tilde{X}, Y) - E_{P_{\tilde{X}} P^{(\ell)}_{Y|\tilde{X}}} d_2(\tilde{X}, Y) \right| \leq \frac{|\mathcal{Y}| D_{\max}}{\ell}. \tag{A. 45}$$

Proof of part (b)

Let $P_{Y|\tilde{X}} \in \mathcal{D}(D_2, P_{\tilde{X}})$ be given. We shall show that there exists $P^*_{Y|\tilde{X}} \in \mathcal{D}\left(D_2 - \alpha_n - \frac{D_{\max}|\mathcal{Y}|}{\ell}, P_{\tilde{X}}\right)$ such that $\left| J(P_{UZ'\tilde{X}} \times P_{Y|\tilde{X}}) - J(P_{UZ'\tilde{X}} \times P^*_{Y|\tilde{X}}) \right| \leq 2\Delta(2\beta_{n,\ell}|\mathcal{Y}|, |\mathcal{Y}|)$.

Denote $c(x) = \sum_{y'} \mathbf{1}\{P(y'|x) \geq \beta_{n,\ell}\}$ and define $P^*_{Y|\tilde{X}}$ as follows

$$P^*_{Y|\tilde{X}}(y|x) = \begin{cases} P_{Y|\tilde{X}}(x|x) + \beta_{n,\ell} \cdot c(x) & y = x \\ P_{Y|\tilde{X}}(y|x) - \beta_{n,\ell} & y \neq x, P(y|x) \geq \beta_{n,\ell} \\ P_{Y|\tilde{X}}(y|x) & \text{otherwise} \end{cases}. \tag{A. 46}$$

Since we assume $d_2(x, x) = 0$, for all $P_{\tilde{X}}$

$$\begin{aligned}
&E_{P_{\tilde{X}} P^*_{Y|\tilde{X}}} d_2(\tilde{X}, Y) - E_{P_{\tilde{X}} P_{Y|\tilde{X}}} d_2(\tilde{X}, Y) \\
= & \; -\sum_x P_{\tilde{X}}(x) \sum_{y \neq x:\, P_{Y|\tilde{X}}(y|x) \geq \beta_{n,\ell}} \beta_{n,\ell} d_2(\tilde{X}, Y) \\
\leq & \; -\beta_{n,\ell} \cdot D_{\min}|\mathcal{Y}| = -\left(\alpha_n + \frac{D_{\max}|\mathcal{Y}|}{\ell}\right),
\end{aligned} \tag{A. 47}$$

thus, $P^*_{Y|\tilde{X}} \in \mathcal{D}\left(D_2 - \alpha_n - \frac{D_{\max}|\mathcal{Y}|}{\ell}, P_{\tilde{X}}\right)$ and by definition,

$$\sum_y \left|P_{Y|\tilde{X}}(y|x) - P^*_{Y|\tilde{X}}(y|x)\right| \leq 2|\mathcal{Y}|\beta_{n,\ell}, \tag{A. 48}$$

and hence, applying (49), we have for all $P_{U\tilde{X}Z'}$,

$$\begin{aligned} & \left|J(P_{UZ'\tilde{X}} \times P_{Y|\tilde{X}}) - J(P_{UZ'\tilde{X}} \times P^*_{Y|\tilde{X}})\right| \\ = & \left|I_{P_{UZ'\tilde{X}} \times P_{Y|\tilde{X}}}(\tilde{Z};Y) - I_{P_{UZ'\tilde{X}} \times P^*_{Y|\tilde{X}}}(\tilde{Z};Y)\right| \\ \leq & \; 2\Delta(2\beta_{n,\ell}|\mathcal{Y}|, |\mathcal{Y}|). \end{aligned} \tag{A. 49}$$

□

Proof of Lemma 9

Proof. On the one hand we have

$$\begin{aligned} \Pr\{\mathcal{E}_2|\mathcal{E}_1\} &= \frac{\Pr\{\mathcal{E}_2, \mathcal{E}_1\}}{\Pr\{\mathcal{E}_1\}} \\ &\geq \Pr\{\mathcal{E}_2, \mathcal{E}_1\} \\ &\geq \Pr\{\mathcal{E}_2\} + \Pr\{\mathcal{E}_1\} - 1 \\ &\geq \Pr\{\mathcal{E}_2\} - \delta. \end{aligned} \tag{A. 50}$$

On the other hand we have,

$$\begin{aligned} \Pr\{\mathcal{E}_2|\mathcal{E}_1\} &= \frac{\Pr\{\mathcal{E}_2, \mathcal{E}_1\}}{\Pr\{\mathcal{E}_1\}} \\ &\leq \frac{\Pr\{\mathcal{E}_2\}}{1-\delta} \\ &= \Pr\{\mathcal{E}_2\} + \Pr\{\mathcal{E}_2\} \cdot \frac{\delta}{1-\delta} \\ &\leq \Pr\{\mathcal{E}_2\} + 2\delta. \end{aligned} \tag{A. 51}$$

□

References

[1] R. Anderson, "Stretching the limits of steganography," in *Information Hiding*, Springer Lecture Notes in Computer Science vol. 1174, pp. 39–48, 1996. Also, available at `http://www.bluespike.com/research/research.html`.

[2] R. J. Anderson and A. Petitcolas, "On the limits of steganography," *IEEE J. Select. Areas in Communications*, vol. 16, no. 4, pp. 463-473, May 1998.

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

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

[5] M. H. M. Costa, “Writing on dirty paper',” *IEEE Trans. Inform. Theory*, vol. 29, no. 4, pp. 439-441, May 1983.

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

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

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

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

[10] A. Dembo and O. Zeitouni, *Large Deviations Tecniques and Applications,* Springer-Verlag, New York, 1998.

[11] S. Gel'fand and M. Pinsker, “Coding for channels with random parameters,” *Probl. Contr. Inform. Theory,* vol. 9, no. 1 pp. 19-31, 1980.

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

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

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

[16] 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.

[17] 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.

[18] 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`

Neri Merhav (S'86–M'87–SM'93–F'99) was born in Haifa, Israel, on March 16, 1957. He received the B.Sc., M.Sc., and D.Sc. degrees from the Technion, Israel Institute of Technology, in 1982, 1985, and 1988, respectively, all in electrical engineering.

From 1988 to 1990 he was with AT&T Bell Laboratories, Murray Hill, NJ, USA. Since 1990 he has been with the Electrical Engineering Department of the Technion, where he is now the Irving Shepard Professor. During 1994–2000 he was also serving as a consultant to the Hewlett–Packard Laboratories – Israel (HPL-I). His research interests include information theory, statistical communications, and statistical signal processing. He is especially interested in the areas of lossless/lossy source coding and prediction/filtering, relationships between information theory and statistics, detection, estimation, and Shannon Theory, including topics in joint source–channel coding, source/channel simulation, and coding with side information with applications to information hiding and watermarking systems.

Dr. Merhav was a co-recipient of the 1993 Paper Award of the IEEE Information Theory Society and he is a Fellow of the IEEE since 1999. He also received the 1994 American Technion Society Award for Academic Excellence and the 2002 Technion Henry Taub Prize for Excellence in Research. From 1996 until 1999 he served as an Associate Editor for Source Coding of the IEEE TRANSACTIONS ON INFORMATION THEORY. He also served as a co–chairman of the Program Committee of the 2001 IEEE International Symposium on Information Theory. He is currently an editor of FOUNDATIONS AND TRENDS IN COMMUNICATIONS AND INFORMATION THEORY.

Anelia Somekh-Baruch (S'01) received the B.Sc. degree from Tel-Aviv University, Israel, and the M.Sc. and Ph.D. degrees from the Technion-Israel Institute of Technology in 1996, 1999 and 2003 respectively, all in electrical engineering. She is currently a post-doc in the Department of Electrical Engineering at the Technion. Her research interests include topics in information theory and watermarking codes.