Identification in the Presence of Side Information with Application to Watermarking

Yossef Steinberg*and Neri Merhav

Department of Electrical Engineering
Technion - Israel Institute of Technology
Haifa 32000, ISRAEL
[ysteinbe,merhav]@ee.technion.ac.il

Abstract

Watermarking codes are analyzed from an information–theoretic viewpoint as identification codes with side information that is available at the transmitter only or at both ends. While the information hider embeds a secret message (watermark) in a covertext message (typically, text, image, sound, or video stream) within a certain distortion level, the attacker, modeled here as a memoryless channel, processes the resulting watermarked message (within limited additional distortion) in attempt to invalidate the watermark. In most applications of watermarking codes the decoder need not carry out full decoding, as in ordinary coded communication systems, but only to test whether a watermark at all exists and if so, whether it matches a particular hypothesized pattern. This fact motivates us to view the watermarking problem as an identification problem, where the original covertext source serves as side information. In most applications, this side information is available to the encoder only, but sometimes it can be available to the decoder as well. For the case where the side information is available at both encoder and decoder, we derive a formula for the identification capacity and also provide a characterization of achievable error exponents. For the case where side information is available at the encoder only, we derive upper and lower bounds on the identification capacity. All characterizations are obtained as single–letter expressions.

Index Terms: Identification via channels, identification with side information, information hiding, steganography, watermarking.

*Part of this work has been carried out while Y. Steinberg was with the Department of Electrical Engineering, Ben–Gurion University, Beer–Sheva, Israel.

1 Introduction

Digital watermarking (WM) is about embedding a secret message into a data set (covertext), say, an image, with minimum harm to its quality, and in order to protect it (e.g., against potential forgery, copyright protection, etc.). The objective is to prevent an unauthorized party (attacker) from being able to remove or decipher this message without access to a secret key and without significant damage to the image quality. The watermark decoder strives to recover the coded watermark from the (possibly attacked) image, with or without the original covertext, depending on the application. WM techniques must, of course, be robust not only to hostile attacks, but also to standard operations, like printing, lossy compression, and halftoning. The principal tradeoffs in the design of WM codes are among three factors: robustness to attacks, quality, and coding rate.

In contrast to the extensive activity reported in the signal/image processing literature, relatively little attention has been devoted to WM from the information-theoretic perspective. In the few information–theoretic papers that do exist on this topic, capacity and error exponents of WM systems were characterized by viewing them as coded communication systems with side information [4],[5],[8],[9],[10],[11]. The model in these works is essentially as follows: A secret message, encoded by a rate-R block code of length n, is hidden in a memoryless covertext message within small degradation of quality, symbolized by distortion level D_1. An active attacker may introduce additional distortion D_2 in attempt to disrupt the watermark. Finally, the resulting data set is decoded at the receiving end, with or without the original covertext data, in order to extract the watermark.

In the aforementioned papers, WM systems were viewed as ordinary coded communication systems, where the covertext data serves a side information, and where the decoder carries out *full* decoding, i.e., decides on one out of 2^{nR} possible messages. In certain applications, however, full decoding is not really necessary, but only a decision whether or not a *particular* watermark message is present. This observation guides us to view WM codes as *identification codes* rather than as ordinary transmission codes.

Identification (ID) codes were first introduced by Ahlswede and Dueck [2], who have shown that the above–described easier role of the decoder, opens the door for reliable transmission of a double–exponential number N of messages (as a function of n) by using randomized codes. Accordingly, the ID capacity of a channel is defined as the maximal second exponential rate of N, which in the case of the classical discrete memoryless channel (DMC), turns out to coincide with the ordinary Shannon capacity.

One of the fascinating features of the ID problem is the notion of a *common experiment*, a random variable, or an experiment whose outcome is reliably known at both transmitter and receiver. It turns out that whenever the Shannon capacity is positive, the presence of a common experiment, even if totally unrelated to the information to be transmitted (e.g., a TV show), increases the ID capacity of a DMC by an amount that is equal to the entropy rate of this random experiment. In the WM problem, as an application of the ID problem with side information, a common experiment quite clearly exists. When the side information is available at both ends, this is obvious. More interestingly, when it is available at the transmitter only, it can still be learned, at least partially, by the decoder, provided that the encoder is designed cleverly.

Motivated with this background, we derive an ID capacity formula for the DMC with side information at both ends (Section 2), and lower and upper bounds on the ID capacity where side information is available at the transmitter only (Section 3). As mentioned earlier, in both cases, the choice of codes is limited by a distortion constraint that models the required similarity between the watermarked data set and the original one. For the case where the side information is available at both ends, the proof of the direct part also provides a characterization of achievable tradeoffs between the exponents of the error probabilities of the two kinds as a function of the coding rate. The problem of evaluating achievable error exponents for the case where the side information is at the transmitter only, remains open.

While the assumption that the attack channel is memoryless may appear somewhat arbitrary, we adopt it here primarily for reasons of simplicity. It should also be noted that memoryless attack channels are worst–case channels in certain communication situations in the contexts of WM [8], [9], [10], and the jamming game [13],[14]. In addition, as it will turn out, at least for the case of covertext available at both sides, memoryless channels are indeed worst–case channels from the viewpoint of identification capacity.

2 Side Information at Transmitter and Receiver

A Definitions

Throughout the sequel, random variables (RV's) are denoted by capital letters, specific realizations of them are denoted by the corresponding lower–case letters, and their alphabets are symbolized by the respective caligraphic letters (e.g., an RV X may take on the value $x \in \mathcal{X}$). For random vectors of dimension n, the same conventions will apply with all aforementioned notations superscripted by n (e.g., the RV $X^n = (X_1, \ldots, X_n)$ may take on the value $x^n = (x_1, \ldots, x_n) \in \mathcal{X}^n$, the nth Cartesian power of $\mathcal{X}$).

The covertext source will be assumed a discrete memoryless source (DMS) with a finite alphabet $\mathcal{V}$, and with a single-letter marginal denoted by P_V. The probability that $V^n = v^n$ will be denoted by $P_{V^n}(v^n)$.

The attacker will be modeled by a DMC W, henceforth referred to as the *attack channel*, with a finite input alphabet $\mathcal{X}$ and a finite output alphabet $\mathcal{Y}$. Thus, $W(y^n|x^n)$ will denote $\Pr\{Y^n = y^n|X^n = x^n\}$. In general, $W\{E|x^n\}$, or $W(E|x^n)$, for $E \subseteq \mathcal{Y}^n$, will stand for the conditional probability of an event E at the channel output, given that x^n was transmitted.

Let $\rho : \mathcal{V} \times \mathcal{X} \to \mathbb{R}_+$ denote a distortion measure. The distortion between a covertext sequence $v^n = (v_1, \ldots, v_n)$ and its watermarked version $x^n = (x_1, \ldots, x_n)$ is defined as $\rho(v^n, x^n) = \frac{1}{n}\sum_{i=1}^n \rho(v_i, x_i)$. For a set $\mathcal{D} \subseteq \mathcal{V}^n \times \mathcal{Y}^n$ and a covertext vector $v^n \in \mathcal{V}^n$, the section of $\mathcal{D}$ at v^n is defined as $\mathcal{D}(v^n) = \{y^n : \ (v^n, y^n) \in \mathcal{D}\}$.

Definition 1 *An $(n, N, \lambda_1, \lambda_2, D)$ watermark ID code for a covertext source P_V, a distortion measure ρ, a distortion level D, and a channel W, is a collection of pairs of randomized codewords and decoding sets $\{(Q_i(\cdot|\cdot), \mathcal{D}_i)\}_{i=1}^N$ that satisfy the following conditions:*

1. $\mathcal{D}_i \subseteq \mathcal{V}^n \times \mathcal{Y}^n, \ \forall i \in \{1, \ldots, N\}$.

2. $\{Q_i(\cdot|\cdot)\}$ *are conditional probability distributions from $\mathcal{V}^n$ to $\mathcal{X}^n$ such that*

$$\sum_{v^n \in \mathcal{V}^n} \sum_{x^n \in \mathcal{X}^n} P_{V^n}(v^n) Q_i(x^n|v^n) \rho(v^n, x^n) \leq D, \quad \forall i.$$

3.

$$\sum_{v^n \in \mathcal{V}^n} \sum_{x^n \in \mathcal{X}^n} P_{V^n}(v^n) Q_i(x^n|v^n) W(\mathcal{D}_i(v^n)|x^n) \geq 1 - \lambda_1 \quad \forall i.$$

4.

$$\sum_{v^n \in \mathcal{V}^n} \sum_{x^n \in \mathcal{X}^n} P_{V^n}(v^n) Q_j(x^n|v^n) W(\mathcal{D}_i(v^n)|x^n) \leq \lambda_2 \quad \forall i \neq j.$$

Condition (2) in Definition 1 means that the average distortion should not exceed D, condition (3) says that the probability of correct identification (detection) ought to be at least $1-\lambda_1$, and condition (4) requires that the probability of false identification (false alarm) would not be larger than λ_2.

The rate R of the code is defined as $R = \frac{1}{n}\log\log N$, where logarithms and exponents here and throughout the sequel are taken to the base 2 unless specified otherwise. The supremum over all rates R of $(n, N, \lambda_1, \lambda_2, D)$ WM ID codes is denoted by $C_{ID}^{TR}(\lambda_1, \lambda_2, D|V)$, where the superscript TR denotes the availability of the side information at both Transmitter and Receiver. The ID capacity with distortion D, and side information V available at both transmitter and receiver is defined as:

$$C_{ID}^{TR}(D|V) = \lim_{\lambda_1,\lambda_2 \downarrow 0} C_{ID}^{TR}(\lambda_1, \lambda_2, D|V). \tag{1}$$

In this section, we will be interested also in error exponents for WM ID. A triple (R, E_1, E_2) is called *achievable with distortion* D if, for any $\gamma > 0$ and all sufficiently large n, there exists an $(n, N, \lambda_1, \lambda_2, D)$ code with $\frac{1}{n}\log\log N \geq R-\gamma$, and $\lambda_i \leq \exp\{-n(E_i-\gamma)\}$, $i = 1, 2$.

B Watermark ID Codes

We first present a lower bound to the achievable error exponents for WM ID codes, and derive the capacity for the case of covertext known at both ends. Let $E_2 > 0$ be given. Let $\mathcal{Q}$ be a class of conditional probability distributions from $\mathcal{V}$ to $\mathcal{X}$ defined as $\mathcal{Q} = \{Q : \sum_{(x,v)\in\mathcal{X}\times\mathcal{V}} \rho(x,v)P_V(v)Q(x|v) \leq D\}$.

For a given probability distribution P on $\mathcal{V}$, define $E_2' = E_2 - H(P\|P_V)$, where $H(P\|P_V) = \sum_v P(v)\log[P(v)/P_V(v)]$. Next define $h(P) = \min\{H_P(V), R + E_2'\}$, where $H_P(V)$ is the entropy of a random variable V governed by P, and $J(PQW') = I_{PQW'}(X;Y|V) + h(P)$, where

$W' = \{W'(y|x,v)\}$ is a channel from $\mathcal{X} \times \mathcal{V}$ to $\mathcal{Y}$, $I_{PQW'}(X;Y|V)$ is the conditional mutual information between X and Y given V, induced by the joint probability mass function (PMF) $\{P(v)Q(x|v)W'(y|x,v)\}$. For any distribution P on $\mathcal{V}$, and any $Q \in \mathcal{Q}$, we denote by (PQ) the induced marginal on $\mathcal{X}$, i.e., $(PQ)(x) = \sum_{v\in\mathcal{V}} P(v)Q(x|v)$. Finally, for any two channels W', W'' from $\mathcal{X}$ to $\mathcal{Y}$, $H(W''||W'|PQ)$ stands for the conditional relative entropy, i.e.,

$$H(W''\|W'|PQ) = \sum_{x\in\mathcal{X}} (PQ)(x) H(W''(\cdot|x)||W'(\cdot|x)). \tag{2}$$

We have the following achievability result for ID error exponents.

Theorem 1 *Given $R > 0$, $E_2 > 0$, let*

$$E_1(R, E_2) \stackrel{\triangle}{=} \max_{Q\in\mathcal{Q}} \min_{P,W'\in\sigma(R,E_2,Q)} \left[H(P\|P_V) + H(W'\|W|PQ)\right], \tag{3}$$

where $\sigma(R, E_2, Q)$ is the class of all distributions P and channels W' satisfying $H(P\|P_V) \leq E_2$ and $J(PQW') \leq R + 2E_2'$. Then, the triple $(R, E_1(R, E_2), E_2)$ is achievable with distortion D.

The proof of Theorem 1 is given in Section 4. Note that the maximin in the theorem is not over a product region, since the inner minimization may depend on Q.

Before continuing to the capacity formula, recall that the WM transmission capacity is given [10] by $C^{TR}(D|V) = \max I(X;Y|V)$, where the maximum is taken over

$$\left\{Q_{VXY} : \quad Q_{VXY} = P_V Q_{X|V} W_{Y|X}, \sum_x \sum_v P_V(v) Q_{X|V}(x|v)\rho(v,x) \leq D.\right\} \tag{4}$$

The result on WM ID capacity, where the covertext is known at both encoder and decoder, is stated next.

Theorem 2 *For a DMS P_V, a DMC W, and an additive distortion measure ρ with distortion level D, $C_{ID}^{TR}(D|V) = C^{TR}(D|V) + H(V)$.*

Before proving Theorem 2, we introduce our main tool for proving ID capacity upper bounds. Let $(P_V \times (Q_i W))(v^n, y^n) = P_{V^n}(v^n) \cdot \sum_{x^n} Q_i(x^n|v^n) W(y^n|x^n)$. By conditions (3) and (4) of Definition 1, the variational distance between any two output distributions is lower bounded by:

$$
\begin{aligned}
d(P_V \times (Q_i W), P_V \times (Q_j W)) &= 2 \max_{A \subset \mathcal{V}^n \times \mathcal{Y}^n} [(P_V \times (Q_i W))(A) - (P_V \times (Q_j W))(A)] \\
&\geq 2\,[(P_V \times (Q_i W))(\mathcal{D}_i) - (P_V \times (Q_j W))(\mathcal{D}_i)] \\
&\geq 2(1 - \lambda_1 - \lambda_2). \qquad (5)
\end{aligned}
$$

In view of eq. (5), the next lemma from [12] will be used to prove upper bounds on achievable ID rates.

Lemma 1 (Distribution Packing) *[12, Lemma 6] Let W be an arbitrary channel defined by a sequence of conditional probability distributions from $\mathcal{X}^n$ to $\mathcal{Y}^n$, $n = 1, 2, \ldots$. Suppose that for a given $\mu \in (0, 1)$, there is a positive integer $n_0 = n_0(\mu)$ such that for all $n > n_0$ there is a set $\mathcal{Q}_0$ of $N = \lceil 2^{2^{nR}} \rceil$ distributions $\{Q_i\}_{i=1}^N$ on $\mathcal{X}^n$ such that $\min_{i \neq j} d((Q_i W), (Q_j W)) \geq 2(1 - \mu)$. Then, for every $\gamma < \mu \min\{1, R\}/4$, there is a positive integer $n_1 = n_1(\mu, \gamma)$, such that for all $n > n_1$, there exists $\mathcal{Q}_1 \subseteq \mathcal{Q}_0$, with $|\mathcal{Q}_1| \geq \exp\exp(nR) - \exp\exp(n(R - \gamma))$, where for each $Q \in \mathcal{Q}_1$, $I_{QW}(X^n; Y^n) \geq nR(1 - 4\mu)$, $I_{QW}(X^n; Y^n)$ being the mutual information induced by $Q \times W$.*

Proof of Theorem 2. The direct part follows by specializing Theorem 1 to the case $E_1 \to 0$, $E_2 \to 0$. As for the converse part, we define a channel $\hat{W}$ consisting of two parallel channels: the original one W and the identity channel from $\mathcal{V}$ to $\mathcal{V}$. The WM ID code can be viewed as a collection of distributions on the product alphabet $\mathcal{X}^n \times \mathcal{V}^n$, where the $\mathcal{V}^n$ component, that is connected to the identity channel, is kept fixed and equal to the covertext V^n. Note that the variational distance between output distributions on $\mathcal{X}^n \times \mathcal{V}^n$ is lower bounded by the variational distance between the marginals on $\mathcal{X}^n$. Keeping the covertext fixed, and considering only ID codes that satisfy the distortion constraint, Lemma 1, applied to the channel $\hat{W}$, asserts that

$$C_{ID}^{TR}(D|V) \leq \liminf_{n\to\infty} \max \frac{1}{n} I(X^n V^n; Y^n V^n), \tag{6}$$

where Y^n is the output of W due to the input X^n, and the maximum is over of all joint distributions of (X^n, V^n) where V^n is drawn by P_{V^n} and $E\rho(V^n, X^n) \leq D$. Notice that

$$I(X^n V^n; Y^n V^n) = I(X^n; Y^n | V^n) + H(V^n), \tag{7}$$

which proves the converse part. $\square$

Comment: Since $H(V)$ depends neither on the coding strategy, nor on the channel statistics, Theorem 2, together with the known results on WM transmission [9, Proposition 6.4], imply that for WM ID too, memoryless channels are worst–case channels among all block-wise memoryless attack channels that satisfy the distortion constraint.

3 Side Information at the Transmitter Only

A Definitions and Background

Before we actually address the ID capacity for this model, we first give a brief review on some known facts about the transmission capacity, and then provide an informal outline of our plan how to harness these facts for our purposes.

Definition 2 *An (n, M, ϵ, D) WM transmission code with distortion measure ρ, distortion level D, and covertext P_V, available at transmitter only, consists of an encoder $f : \mathcal{V}^n \times \{1, 2, \ldots, M\} \to \mathcal{X}^n$ and a decoder $g : \mathcal{Y}^n \to \{1, 2, \ldots, M\}$ such that*

$$\frac{1}{M} \sum_{m=1}^{M} \sum_{v^n \in \mathcal{V}^n} P_{V^n}(v^n) W\{g(Y^n) = m | X^n = f(v^n, m)\} \geq 1 - \epsilon, \tag{8}$$

and

$$\frac{1}{M}\sum_{m=1}^{M}\sum_{v^n\in\mathcal{V}^n} P_{V^n}(v^n)\rho(v^n, f(v^n, m)) \leq D. \tag{9}$$

The rate is $R = \frac{1}{n}\log M$. The supremum over all rates R for which such codes exist, for every $\epsilon > 0$ and sufficiently large n, is the WM transmission capacity with covertext available at transmitter only, and is denoted by $C^T(D|V)$.

This is a variant of the model studied by Gel'fand and Pinsker [6], who derived the capacity of a channel whose transition probability matrix $\{W(y|s,x)\}$ depends on a random state S, governed by a DMS P_S, where the sequence of states corresponding to each block, $s^n = (s_1, s_2, \ldots, s_n)$, is known prior to transmission (non–causally) and at the encoder only.

In the proof of the direct part in [6], the codebook is divided into 2^{nR} bins, each of which corresponds to a message, and contains a number of codewords. For transmitting a message $m \in \{1, 2, \ldots, 2^{nR}\}$, a particular codeword, depending on s^n, in bin m is chosen. The important point is that it is only the bin number that conveys the information, not the particular codeword within the bin, which in turn depends on the uncontrollable state sequence. Nonetheless, the decoder may strive to estimate the particular codeword transmitted and thereby have at least partial information about s^n. The capacity formula for this channel was shown [6] to be

$$C = \sup_{USXY}\{I(U;Y) - I(U;S)\}, \tag{10}$$

where U is a RV whose alphabet size $|\mathcal{U}|$ need not exceed $|\mathcal{X}| + |\mathcal{S}|$, and $(USXY)$ is such that S is governed by P_S, and $U \to (S,X) \to Y$ is a Markov chain, whose right link is governed by W. For reasons that will become apparent later, we set $|\mathcal{U}| = |\mathcal{X}| \cdot |\mathcal{S}|$.

Before returning to the WM system model, let us see how to harness this result to assess the ID capacity of the Gel'fand–Pinsker channel. While the transmission capacity is always a lower bound to the ID capacity, here we can say somewhat more than that, using the notion of a common experiment [3], which we now briefly explain.

Let Z^n be a RV, whose outcome is known to both encoder and decoder, and let us denote by $|\mathcal{Z}^n|$ its alphabet size. Let $\mathcal{C}_{\theta n}$ be a transmission code of block-length $\theta \cdot n$ ($\theta > 0$, small) with a small error probability.[1] Every mapping from $|Z^n|$ to $\mathcal{C}_{\theta n}$ induces a joint distribution on $Z^n \times \mathcal{C}_{\theta n}$. Since the number of different mappings is $|\mathcal{C}_{\theta n}|^{|Z^n|}$, so is the number of such distributions. Thus, every set of mappings from Z^n to $\mathcal{C}_{\theta n}$ can be used to construct an ID code, whose operation is described as follows. Assume that the receiver is interested in message j. The randomized ID encoder selects a member $(z^n, x^{\theta n}) \in Z^n \times \mathcal{C}_{\theta n}$ according to the joint distribution dictated by the given distribution of Z^n *and* the mapping, and then transmits $x^{\theta n}$. The receiver, that observes $Z^n = z^n$ and reliably estimates $x^{\theta n}$, can also reliably decide whether or not the mapping of the encoder is indeed j, provided that the mappings are, in some sense, "sufficiently different." More precisely, the various distributions on $Z^n \times \mathcal{C}_{\theta n}$ are pairwise separated in terms of the variational distance. We emphasize that the distribution of Z^n does not depend on the message sent, and that no matter how small $\theta > 0$ may be, we obtain a double-exponential behavior, where the exponent depends only on $|\mathcal{Z}^n|$, which for large n, is essentially dictated by the normalized entropy of Z^n, the common experiment rate. Indeed, in Theorem 2, where the side information V^n was available at both ends, and hence could be thought of as a common experiment, the ID capacity exceeds the transmission capacity exactly by $H(V)$. Note that, in general, this method of constructing ID codes holds even if Z^n is not explicitly given to the receiver, but can nevertheless be reliably estimated.

[1]The rate of this code need not be close to capacity.

Next, let us identify what the common experiment of the Gel'fand–Pinsker channel is. Indisputably, the decoder cannot, in general, fully estimate S^n. As we prove in Theorem 3, however, the decoder can reliably estimate, not only the bin number m, but also some more refined information, namely, the index of a subset of bin number m to which the codeword belongs, among $\exp[nI(U;S)]$ many equally–sized subsets that form a partition of this bin. The common experiment rate is then $I(U;S)$. Thus, for any quadruple $(USXY)$ that satisfies the above conditions, one can construct a transmission code of rate $I(U;Y)-I(U;S)$, plus a common experiment of rate $I(U;S)$, thus a total ID rate of $I(U;Y)$. Therefore, the ID capacity of the Gel'fand–Pinsker channel is lower bounded by $\sup I(U;Y)$, where the supremum is over all quadruples $(USXY)$ that satisfy, in addition to the aforementioned conditions, also $I(U;Y)-I(U;S)>0$, which is henceforth referred to as the *positivity constraint*. The roots of this constraint lie in the construction of the ID code: Since the ID code employs an ordinary transmission code of a positive rate, the positivity constraint stems from the Gel'fand–Pinsker capacity formula (10).

The upper bound follows from Lemma 1. To see this, interpret the Gel'fand–Pinsker channel as a channel from $\mathcal{S}\times\mathcal{X}$ to $\mathcal{Y}$, where the marginal of S is constrained to agree with P_S. Thus, by Lemma 1, we obtain $C_{ID}\le\liminf C_{ID}^{(n)}$, where $C_{ID}^{(n)}=\frac{1}{n}\max_{X^n} I(X^nS^n;Y^n)$. Next, by the chain rule, and the fact that the channel is memoryless,

$$I(X^nS^n;Y^n)\le\sum_{i=1}^{n} I(X_iS_i;Y_i), \tag{11}$$

hence

$$C_{ID}\le\sup_X I(XS;Y)\le\sup_{USXY} I(U;Y), \tag{12}$$

where $(USXY)$ is as in (10) and the last inequality holds since we assumed $|\mathcal{U}|=|\mathcal{X}|\cdot|\mathcal{S}|$, which allows for the choice $U=XS$. Note that due to the positivity constraint, there might be a gap between the upper and lower bound.

Returning to the WM problem, the side information V plays the role of the channel state S. The channel transition matrix does not depend on V, but due to the distortion constraint, the codebook does. Thus, two refinements are needed for the construction in [6]: Incorporating the distortion constraint, and reliable estimation the common experiment at the decoder.

We now proceed to formalize these considerations. First observe that when V^n is not available at the receiver, the decoding sets $\mathcal{D}_i$ are subsets of $\mathcal{Y}^n$ only. Accordingly, Definition 1 is now modified so that conditions (3) and (4) include no sectioning of $\mathcal{D}_i$ at v^n. We denote by $C_{ID}^T(D|V)$ the ID capacity, where V^n is available to the transmitter only.

For a given map $f : \mathcal{V}^n \times \{1, 2, \ldots, M\} \to \{1, 2, \ldots, J\} \times \mathcal{X}^n$, $f_{\mathcal{X}}$ will denote its restriction to $\mathcal{X}^n$, and f_J – the restriction to $\{1, 2, \ldots, J\}$. We next define a WM transmission code with a common experiment.

Definition 3 *An $(n, M, J, \delta, \epsilon, D)$ WM transmission code with a common experiment, distortion measure ρ, distortion level D, and covertext P_V available at transmitter only, consists of an encoder $f : \mathcal{V}^n \times \{1, 2, \ldots, M\} \to \{1, 2, \ldots, J\} \times \mathcal{X}^n$ and a decoder $g : \mathcal{Y}^n \to \{1, 2, \ldots, J\} \times \{1, 2, \ldots, M\}$ such that*

$$\frac{1}{M}\sum_{m=1}^{M}\sum_{v^n \in \mathcal{V}^n} P_{V^n}(v^n) W\{g(Y^n) \neq (f_J(v^n, m), m) | X^n = f_{\mathcal{X}}(v^n, m)\} \leq \epsilon, \tag{13a}$$

$$\frac{1}{M}\sum_{m=1}^{M}\sum_{v^n} P_{V^n}(v^n)\rho(v^n, f_{\mathcal{X}}(v^n, m)) \leq D, \tag{13b}$$

and, for every message m there exists a subset $\mathcal{B}_m \subseteq \{1, 2, \ldots, J\}$ such that

$$|\mathcal{B}_m| \geq J \cdot \exp(-n\delta) \tag{14a}$$

$$\frac{\exp(-n\delta)}{J} \leq Pr\{f_J(V^n, m) = j\} \leq \frac{\exp(n\delta)}{J} \quad \forall\, j \in \mathcal{B}_m, \tag{14b}$$

$$\sum_{j \in \mathcal{B}_m} Pr\{f_J(V^n, m) = j\} \quad \geq \quad 1 - \epsilon. \tag{14c}$$

A rate pair (R_1, R_2) is called *achievable with distortion* D if for every $\delta > 0$, $\epsilon > 0$, and sufficiently large n, there exists an $(n, M, J, \delta, \epsilon, D + \epsilon)$ code with $\frac{1}{n} \log M \geq R_1$ and $\frac{1}{n} \log J \geq R_2$.

Discussion: Here, $m \in \{1, 2, \ldots, M\}$ is the message (bin) index, and $j \in \{1, 2, \ldots, J\}$ is the partition index within each bin. Eq. (13a) requires that the error probability in joint estimation of m *and* j would be small. The idea is to use j as a common experiment. For every m, the index j is a RV and $\mathcal{B}_m$ serves as a typical subset of the index set $\{1, \ldots, J\}$. Eqs. (14) guarantee that an AEP holds for j, and that its entropy (the common experiment rate) is close to $\log J$.

We next claim that if (R_1, R_2) is achievable with distortion D, then it is moreover achievable with the averages over m, in eqs. (13), being replaced by maxima over m. To see why this is true, let $P_e(m)$ be the error probability given m, and let $D(m)$ be the conditional expected distortion given m for a given code that satisfies (14). Given $k > 0$ and $\mu > 0$, let $G_k = \{m : \ P_e(m) \leq k\epsilon\}$ and $F_\mu = \{m : \ D(m) \leq D + \mu\}$. By Chebychev's inequality, $|G_k| \geq M(k-1)/k$ and $|F_\mu| \geq M\mu/(D+\mu)$. Given μ, let k be so large that $\mu/(D+\mu) + (k-1)/k > 1$, which means that $|G_k \cap F_\mu| \geq M \cdot (\mu/(D+\mu) + (k-1)/k - 1)$. The transmission code formed by $G_k \cap F_\mu$ satisfies $\max_m D(m) \leq D + \mu$ and $\max_m P_e(m) \leq k\epsilon$, keeping R_1 essentially as large as in the original code since k depends only on μ and D, but not on ϵ or n. This fact will be used in the proof of Theorem 4.

B Results

Let us define

$$\mathcal{P}(D) \quad = \quad \Bigg\{ Q_{UVXY} : \ |\mathcal{U}| = |\mathcal{X}| + |\mathcal{V}|, \ \sum_{v,x} Q_{VX}(v,x)\rho(v,x) \leq D,$$

$$Q_{UVXY}(u,v,x,y) = P_V(v) Q_{XU|V}(x,u|v) W(y|x) \ \forall u,v,x,y \Big\} . \quad (15)$$

We begin with an achievability result for WM transmission with a common experiment.

Theorem 3 *For any $\epsilon > 0$, $0 < \delta < D$, a triple $(UVX) \in \mathcal{P}(D-\delta)$, and large enough n, there exists an $(n, M, J, 2\delta, \epsilon, D)$ WM transmission code with a common experiment, such that $\frac{1}{n}\log M = I(U;Y) - I(U;V) - 4\delta$ and $\frac{1}{n}\log J = I(U;V) + \delta$.*

The proof of Theorem 3 appears in Section 4.

Discussion: Following the earlier discussion on the Gel'fand–Pinsker channel, it is now apparent that $R = \frac{1}{n}\log M + \frac{1}{n}\log J$ is an achievable ID rate for the WM channel, a fact that will be proved in Theorem 4. Note that the term $I(U;V)$ cancels, and we are left with $R = I(U;Y)$. Since we construct the ID code using a transmission code, we must keep $\frac{1}{n}\log M > 0$, or equivalently, $I(U;Y) > I(U;V)$, which is the positivity constraint. This prevents us from using the fact that $U \to X \to Y$ is a Markov chain and setting $\max I(U;Y) = \max I(X;Y)$.

The RV U adds a degree of freedom in the *construction* of ID codes by controlling the tradeoff between the common experiment rate and the transmission rate. This is important in various applications. One of them is *Identification plus Transmission* (I+T) [7], where one transmits information, in the usual sense, to one out of a large number of prospective recipients. The receiver first, identifies the addressee, and then decodes the message itself. A naive approach would be to first, use an ID code, and then an ordinary transmission code. If B information bits should be transmitted and there are N recipients, this would require $(\log\log N + B)/C$ channel uses. It is shown in [7] that, using I+T codes, one can fulfill the same task using only $\max\{\log\log N, B\}/C$ channel uses. Again, the idea is that the transmission code used by the ID code is also used (with proper partitioning) to convey information to the addressee. In our setting, where the transmission code is of size M, for a given ID rate R, we would like to maximize M, or, equivalently, to minimize J.

In view of the aforementioned positivity constraint, define $\mathcal{P}' = \{UVXY : I(U;V) < I(U;Y)\}$. We now provide the main results of this section, namely, upper and lower bounds to $C_{ID}^{T}(D|V)$.

Theorem 4 *For a DMS P_V a DMC W, and an additive distortion measure ρ with distortion level D,*

$$\sup_{\mathcal{P}(D)\cap\mathcal{P}'} I(U;Y) \leq C_{ID}^{T}(D|V) \leq \max_{\mathcal{P}(D)} I(U;Y) = \max_{\mathcal{P}(D)} I(X;Y).$$

The proof is given in Section 4.

Observe that if $\mathcal{X} = \mathcal{V}$ and P_V happens to achieve the capacity of W, then $\max_{\mathcal{P}(D)} I(X;Y) = \max_X I(X;Y)$, and the upper bound of Theorem 4 is independent of D. This indicates that the upper bound may not be tight in general.

4 Proofs

In this section, we prove Theorems 1, 3, and 4.

A Proof of Theorem 1

The proof follows the lines of the proof of Theorem 2 of [2]. The set of all n-sequences of a given type, say P, is denoted by $\mathcal{T}_P^n$. For every realization $v^n \in \mathcal{T}_P^n \subset \mathcal{V}^n$, consider the conditional type $\mathcal{T}_Q^n(v^n) \subset \mathcal{X}^n$ of all x^n-vectors whose joint empirical PMF with v^n is given by $P(v)Q(x|v)$. For every v^n, every message i is encoded by the uniform distribution on a family $\mathcal{X}_i(v^n) \subset \mathcal{T}_Q^n(v^n)$, satisfying $|\mathcal{X}_i(v^n)| = M(v^n)$ for all $i = 1, 2, ..., N$, $N = \lfloor 2^{2^{nR}} \rfloor$, where the function $M(v^n)$ will be specified later.

Let R and E_2 be given. We assign to $\mathcal{X}_i(v^n)$ a decoding set $\mathcal{D}_i(v^n) = \mathcal{D}(\mathcal{X}_i(v^n))$ defined according to the following rules: If v^n belongs to a type class $\mathcal{T}_P^n$ for which $H(P||P_V) \geq E_2$, then $\mathcal{D}_i(v^n) = A^n$ for all i. Else, if $H(P||P_V) < E_2$, let

$$\mathcal{D}(\mathcal{X}_i(v^n)) = \bigcup_{x^n \in \mathcal{X}_i(v^n)} \mathcal{F}_{x^n v^n} \tag{16}$$

where

$$\mathcal{F}_{x^n v^n} = \bigcup_{W':\ J(PQW') > R + 2E_2'} \mathcal{T}_{W'}(x^n, v^n), \tag{17}$$

and $\mathcal{T}_{W'}^n(x^n, v^n)$ is the conditional type of channel output sequences given (x^n, v^n), and associated with the empirical channel $W' = \{W'(y|x,v)\}$.

We first focus on the error probability the first kind. If v^n belongs to a type class $\mathcal{T}_P^n$ for which $H(P||P_V) \geq E_2$, then, by definition, there are never errors of the first kind since $\mathcal{D}_i^c(v^n) = \emptyset$ for all i. If, on the other hand, $H(P||P_V) < E_2$, we have the following:

$$\begin{aligned} W(\mathcal{F}^c_{x^n v^n} | x^n) &\leq \sum_{W':\ J(PQW') \leq R + 2E_2'} W(\mathcal{T}_{W'}(x^n, v^n) | x^n) \\ &\leq \sum_{W':\ J(PQW') \leq R + 2E_2'} \exp\{-nH(W'||W|PQ)\} \\ &\doteq \exp\{-n \min_{W':\ J(PQW') \leq R + 2E_2'} H(W'||W|PQ)\}, \end{aligned} \tag{18}$$

where the notation $a_n \doteq b_n$ means that $\lim_{n\to\infty} n^{-1} \log \frac{a_n}{b_n} = 0$, and thus a_n and b_n are equivalent in the exponential scale (a similar convention holds for $\stackrel{\cdot}{\leq}$, etc.). Thus,

$$\begin{aligned} &\sum_{v^n} P_{V^n}(v^n) \cdot \frac{1}{M(v^n)} \sum_{x^n \in \mathcal{X}_i(v^n)} W(\mathcal{D}_i^c(v^n)|x^n) \\ \leq &\sum_{v^n \in \mathcal{T}_P:\ H(P||P_V) < E_2} P_{V^n}(v^n) \cdot \frac{1}{M(v^n)} \sum_{x^n \in \mathcal{X}_i(v^n)} W(\mathcal{F}^c_{x^n v^n}|x^n) \\ \stackrel{\cdot}{\leq} &\sum_{P:\ H(P||P_V) < E_2} \exp\{-nH(P||P_V)\} \exp\{-n \min_{W':\ J(PQW') \leq R + 2E_2'} H(W'||W|PQ)\} \end{aligned}$$

$$\doteq \exp\left\{-n \min_{P:\ H(P||P_V)\le E_2} \left[H(P||P_V) + \min_{W':\ J(PQW')\le R+2E_2'} H(W'||W|PQ) \right]\right\}, \tag{19}$$

regardless of the choice of $\{\mathcal{X}_i(v^n)\}$. Since it holds for any Q in $\mathcal{Q}$, this proves the achievability of $E_1(R, E_2)$.

We now specify the sets $\{\mathcal{X}_i(v^n)\}$ to achieve (R, E_2). For every $v^n \in \mathcal{T}_P^n$, let

$$M(v^n) = M(P) \stackrel{\triangle}{=} \exp\{n[R + E_2' - h(P)]\} \tag{20}$$

and define the $\{\mathcal{X}_i(v^n)\}$ by random selection as follows. Given $v^n \in \mathcal{T}_P$, let $\boldsymbol{X}_{ij}(v^n)$, $i = 1, ..., N$, $j = 1, ..., M(v^n)$, be independent RV's all uniformly distributed over $\mathcal{T}_Q^n(v^n)$. Define the random sets $\bar{\mathcal{X}}_i(v^n) = \{\boldsymbol{X}_{i1}(v^n), \boldsymbol{X}_{i2}(v^n), ..., \boldsymbol{X}_{iM(v^n)}(v^n)\}$, $i = 1, ..., N$. Every realization of the $\{\boldsymbol{X}_{ij}(v^n)\}$ gives rise to an ID code, where the ith message is encoded with a uniform distribution on the corresponding realization $\mathcal{X}_i(v^n)$ of $\bar{\mathcal{X}}_i(v^n)$.

We want to show that there exists at least one such randomly selected ID code with an error exponent of the second kind at least as large as $E_2 - \delta$, where $\delta > 0$ is arbitrarily small for all sufficiently large n. To this end, consider first the following result for two messages.

For every $v^n \in \mathcal{T}_P^n$, let $\mathcal{X}_1(v^n)$ be any subset of $\mathcal{T}_Q^n(v^n)$ of cardinality $M(v^n)$, and let $\bar{\mathcal{X}}_2(v^n)$ be a random set as described above. We consider the probability P^* that for a given $\gamma > 0$,

$$\sum_{v^n} P_{V^n}(v^n) \cdot \frac{1}{M(v^n)} \sum_{x^n \in \mathcal{X}_1(v^n)} W(\mathcal{D}(\bar{\mathcal{X}}_2(v^n)|x^n) \le \exp\{-n(E_2 - 4\gamma)\} \tag{21}$$

and

$$\sum_{v^n} P_{V^n}(v^n) \cdot \frac{1}{M(v^n)} \sum_{x^n \in \bar{\mathcal{X}}_2(v^n)} W(\mathcal{D}(\mathcal{X}_1(v^n)|x^n) \le \exp\{-n(E_2 - 4\gamma)\}. \tag{22}$$

We then have the following proposition:

Proposition 1 *For every $\gamma > 0$, and every $n \ge n(\gamma)$, $P^* \ge 1 - (n+1)^{|\mathcal{A}|} \exp\{-(n\gamma - 2)\exp(nR)\}$.*

Since P^* converges to unity in a double–exponential rate of at least R, Proposition 1 guarantees the existence of a rate-R ID code with the desired error performance in an analogous way that Proposition 2 of [2] gives rise to the achievability of the indicated error exponents therein (see [2, p. 20]). We next prove Proposition 1.

The following lemma is an extension of [1, Lemma 1].

Lemma 2 *For a given $v^n \in \mathcal{T}_P^n$, let $\boldsymbol{X}$ be uniformly distributed on $\mathcal{T}_Q^n(v^n)$, and let $\mathcal{A} \subset \mathcal{T}_Q^n(v^n)$, $|\mathcal{A}| = \exp(nR^*)$, $R^* \geq 0$. Given $x^n \in \mathcal{X}^n$, and two empirical channels $W' = \{W'(y|x,v)\}$ and $W'' = \{W''(y|x,v)\}$, define*

$$g_{W'W''}(x^n, v^n) = |\mathcal{T}_{W'}^n(x^n, v^n) \bigcap \bigcup_{x^n \in \mathcal{A}} \mathcal{T}_{W''}^n(x^n, v^n)|. \tag{23}$$

Then

(a) $E\{g_{W'W''}(\boldsymbol{X}, v^n)\} \stackrel{\cdot}{\leq} \exp\{n(H_{PQW'}(Y|X,V) - [I_{PQW''}(X;Y|V) - R^*]^+)\}$.

(b) For every $\xi > 0$, $\eta > 0$ and $n \geq n(\eta)$,

$$\begin{aligned} &Pr\{g_{W'W''}(\boldsymbol{X}, v^n) \geq \exp\{n(H_{PQW'}(Y|X,V) - [I_{PQW''}(X;Y|V) - R^*]^+ + \xi + 3\eta)\} \\ &\textit{for some } W', W''\} \leq \exp\{-n(2\eta + \xi)\}. \end{aligned} \tag{24}$$

Proof of Lemma 2. We will prove only part (a) as (b) will follow from (a) by a direct application of Chebychev's inequality. Let QW' denote the conditional PMF defined by $(QW')(y|v) = \sum_x Q(x|v)W'(y|x,v)$. We first observe that $g_{W'W''}$ vanishes (thus, making the lemma trivial) if $QW' \neq QW''$. We shall then henceforth assume that $QW' = QW''$.

$$E\{g_{W'W''}(\boldsymbol{X}, v^n)\} \quad = \quad E|\mathcal{T}_V^n(\boldsymbol{X}, v^n) \cap \bigcup_{x^n \in \mathcal{A}} \mathcal{T}_{W''}^n(x^n, v^n)|$$

$$
\begin{aligned}
&\le \sum_{x^n \in \mathcal{A}} E|\mathcal{T}^n_{W'}(\boldsymbol{X}, v^n) \cap \mathcal{T}^n_{W''}(x^n, v^n)| \\
&= |\mathcal{A}| \cdot E|\mathcal{T}^n_{W'}(\boldsymbol{X}, v^n) \cap \mathcal{T}^n_{W''}(\tilde{x}^n, v^n)| \\
&= |\mathcal{A}| \cdot \sum_{y^n \in \mathcal{T}^n_{W''}(\tilde{x}^n, v^n)} \Pr\{y^n \in \mathcal{T}^n_{W'}(\boldsymbol{X}, v^n)\}, \qquad (25)
\end{aligned}
$$

where $\tilde{x}^n$ is an arbitrary member of $\mathcal{A}$. As $\boldsymbol{X}$ is uniformly distributed over $\mathcal{T}^n_Q(v^n)$, we have

$$
\begin{aligned}
\Pr\{y^n \in \mathcal{T}^n_{W'}(\boldsymbol{X}, v^n)\} &= \frac{|\mathcal{T}^n_Q(v^n) \cap \{x^n :\ y^n \in \mathcal{T}^n_V(x^n, v^n)\}|}{|\mathcal{T}^n_Q(v^n)|} \\
&\stackrel{\cdot}{\le} \frac{\exp\{nH_{PQW'}(X|V,Y)\}}{\exp\{nH_{PQ}(X|V)\}} \\
&= \exp\{-nI_{PQW'}(X;Y|V)\}. \qquad (26)
\end{aligned}
$$

Therefore, since $|\mathcal{A}| = \exp(nR^*)$, we get

$$
\begin{aligned}
E\{g_{W'W''}(\boldsymbol{X}, v^n)\} &\le \exp\{n[R^* + H_{PQW''}(Y|X,V) - I_{PQW'}(X;Y|V)]\} \\
&= \exp\{n[R^* + H_{PQW''}(Y|X,V) - H_{PQW'}(Y|V) + H_{PQW'}(Y|X,V)]\} \\
&= \exp\{n[R^* + H_{PQW''}(Y|X,V) - H_{PQW''}(Y|V) + H_{PQW'}(Y|X,V)]\} \\
&= \exp\{n[R^* + H_{PQW'}(Y|X,V) - I_{PQW''}(X,Y|V)\}. \qquad (27)
\end{aligned}
$$

On the other hand, $E\{g_{W'W''}(\boldsymbol{X}, v^n)\}$ is trivially upper bounded by

$$
E|\mathcal{T}^n_V(\boldsymbol{X}, v^n)| \doteq \exp\{nH_{PQW'}(Y|X,V)\}, \qquad (28)
$$

and so,

$$
\begin{aligned}
E\{g_{W'W''}(\boldsymbol{X}, v^n)\} &\doteq \exp\left\{n[H_{PQW'}(Y|X,V) + \min\{0, R^* - I_{PQW''}(X;Y|V)\}]\right\} \\
&= \exp\left\{n(H_{PQW'}(Y|X,V) - [I_{PQW''}(X;Y|V) - R^*]^+)\right\}, \qquad (29)
\end{aligned}
$$

completing the proof of Lemma 2.

Lemma 3 *For every $v^n \in \mathcal{T}_P^n$, let $\mathcal{A} = \mathcal{X}_1(v^n)$, hence $R^* = R + E_2' - h(P)$, and let $\boldsymbol{x}^* = \boldsymbol{x}^*(v^n)$ satisfy*

$$g_{W'W''}(\boldsymbol{x}^*, v^n) \leq \exp\{n(H_{PQW'}(Y|X,V) - [J(PQW'') - R - E_2']^+ + \xi + 3\eta)\} \tag{30}$$

for all W', W''. Then,

$$W(\mathcal{D}_1(v^n)|\boldsymbol{x}^*) \dot{\leq} \exp\{-n(E_2' - 3\eta - \xi)\} \tag{31}$$

and

$$\sum_{x^n \in \mathcal{X}_1(v^n)} W(\mathcal{F}_{\boldsymbol{x}^* v^n}|x^n) \dot{\leq} \exp\{-n(E_2' - 3\eta - \xi)\}. \tag{32}$$

Proof of Lemma 3. For a given $v^n \in \mathcal{T}_P^n$,

$$\begin{aligned} W(\mathcal{D}_1(v^n)|\boldsymbol{x}^*) &\leq \sum_{W'} \sum_{W'':\ J(PQW'') \geq R + 2E_2'} W(\mathcal{T}_{W'}^n(\boldsymbol{x}^*, v^n) \bigcap \bigcup_{x^n \in \mathcal{X}(v^n)} \mathcal{T}_{V''}^n(x^n, v^n)|\boldsymbol{x}^*) \\ &\doteq \max_{W'} \max_{W'':\ J(PQW'') \geq R + 2E_2'} g_{W'W''}(\boldsymbol{x}^*, v^n) \cdot \\ &\quad \exp\{-n[H_{PQW'}(Y|X) + H(W'||W|PQ)]\}. \end{aligned} \tag{33}$$

Since $H_{PQW'}(Y|X,V) \leq H_{PQW'}(Y|X)$ and $H(W'||W|PQ) \geq 0$, we have

$$\begin{aligned} & g_{W'W''}(\boldsymbol{x}^*, v^n) \cdot \exp\{-n[H_{PQW'}(Y|X) + H(W'||W|PQ)]\} \\ \leq\ & \exp\{-n([J(PQW'') - R - E_2']^+ - 3\eta - \xi\} \\ \dot{\leq}\ & \exp\{-n(E_2' - 3\eta - \xi)\}, \end{aligned} \tag{34}$$

where the last step follows from the fact that $J(PQW'') \geq R + 2E_2'$.

Turning now to the proof of eq. (32), we have

$$\sum_{x^n \in \mathcal{X}_1(v^n)} W(\mathcal{F}_{\boldsymbol{x}^* v^n}|x^n) \leq \sum_{x^n \in \mathcal{X}_1(v^n)} \sum_{W'} \sum_{W'':\ J(PQW'') \geq R + 2E_2'} W(\mathcal{T}_{W''}^n(\boldsymbol{x}^*, v^n) \bigcap \mathcal{T}_{W'}^n(x^n, v^n)|x^n)$$

$$\doteq \sum_{W'} \sum_{W'':\ J(PQW'')\geq R+2E_2'} \exp\{-n[H_{PQW'}(Y|X,V)- H_{PQW''}(Y|X,V)+[J(PQW')-R-E_2']^+ -3\eta-\xi\}$$

$$\doteq \max_{W'} \max_{W'':\ J(PQW'')\geq R+2E_2'} \exp\{-n[H_{PQW'}(Y|X,V)- H_{PQW''}(Y|X,V)+[J(PQW')-R-E_2']^+ -3\eta-\xi\}. \tag{35}$$

It would then be sufficient to show that

$$H_{PQW'}(Y|X,V) - H_{PQW''}(Y|X,V) + [J(PQW') - R - E_2']^+ \geq E_2'. \tag{36}$$

Assume first that $J(PQW') \leq R + E_2'$, in which case the l.h.s. of eq. (36) is just $H_{PQW'}(Y|X,V) - H_{PQW''}(Y|X,V)$. Since $QW' = QW''$, we have

$$H_{PQW'}(Y|X,V) = H_{PQW'}(Y|V) - I_{PQW'}(X;Y|V) \geq H_{PQW''}(Y|V) + h(P) - R - E_2'. \tag{37}$$

Since $J(PQW'') \geq R + 2E_2'$,

$$H_{PQW''}(Y|X,V) = H_{PQW''}(Y|V) - I_{PQW''}(X;Y|V) \leq H_{PQW''}(Y|V) + h(P) - R - 2E_2', \tag{38}$$

and so, $H_{PQW'}(Y|X,V) - H_{PQW''}(Y|X,V) \geq E_2'$.

Assume now that $J(PQW') > R + E_2'$. In this case, the l.h.s. of eq. (36) is

$$\begin{aligned}
& H_{PQW'}(Y|X,V) - H_{PQW''}(Y|X,V) + J(PQW') - R - E_2' \\
= \ & H_{PQW'}(Y|V) - H_{PQW''}(Y|X,V) + h(P) - R - E_2' \\
= \ & H_{PQW''}(Y|V) - H_{PQW''}(Y|X,V) + h(P) - R - E_2' \\
= \ & J(PQW'') - R - E_2' \\
\geq \ & E_2',
\end{aligned} \tag{39}$$

completing the proof of Lemma 3.

Lemma 4 For a given P, let $\boldsymbol{X}_{ij}(v^n), j=1,...,M(v^n), v^n \in \mathcal{T}_P^n$, be $M \triangleq M(P)\cdot|\mathcal{T}_P^n|$ independent random variables, each uniformly distributed in $\mathcal{T}_Q^n(v^n)$. For every $\xi \geq 0$, and $j=1,...,M(v^n)$, let

$$S_{j\xi}(v^n) = \begin{cases} 0 & \text{if } \boldsymbol{X}_{ij}(v^n) \text{ equals a } \boldsymbol{x}^* \text{ that satisfies eq. (30)} \\ 1 & \text{otherwise} \end{cases} \tag{40}$$

Then, for every $\xi \in [0, E_2']$, and sufficiently large n,

$$\Pr\left\{\sum_{v^n \in \mathcal{T}_P^n} \sum_{j=1}^{M(v^n)} S_{j\xi}(v^n) > \exp\{-n\xi\}M\right\} \leq \exp\{-(n\eta-2)\exp nR\}. \tag{41}$$

The proof is identical to the proof of Lemma 3 of [2] as the total number of independent random variables is $M \stackrel{\cdot}{\geq} \exp\{n(R+E_2')\}$.

Finally, we combine Lemmas 1–3 in order to prove Proposition 1. Let $\eta = \gamma$, and define $\xi_k = \frac{k}{n}E_2'$, $k=0,1,...,n$. Then, by the union bound,

$$\Pr\bigcup_P \bigcup_{k=0}^{n}\left\{\sum_{v^n \in \mathcal{T}_P^n} \sum_{j=1}^{M(v^n)} S_{j\xi_k}(v^n) > \exp(-n\xi_k)M\right\} \leq (n+1)^{|A|}\exp\{-(n\gamma-2)\exp(nR)\}. \tag{42}$$

It will be sufficient to show that if

$$\sum_{v^n \in \mathcal{T}_P^n} \sum_{j=1}^{M(v^n)} S_{j\xi_k}(v^n) \leq \exp(-n\xi_k)M \quad \text{for all } P \text{ and all } k=0,1,...,n, \tag{43}$$

then eqs. (21) and (22) hold. Observe that $S_{j\xi}(v^n)$ is monotonically non-increasing in ξ, where the jump from the value 1 to the value 0 occurs at a certain critical value of ξ. For a given v^n and k, the number of $j \in \{1,...,M(v^n)\}$ such that the critical value of ξ falls in $[E_2'k/n, E_2'(k+1)/n]$ is upper bounded by $\sum_{j=1}^{M(v^n)} S_{j\xi_k}(v^n)$. For these j, $S_{j\xi_{k+1}}(v^n) = 0$ and so, eq. (32) holds with ξ_{k+1}. On the other hand, the number of j such that critical ξ is larger than $E_2'(=\xi_n)$ is upper bounded by $\sum_{j=1}^{M(v^n)} S_{j\xi_n}(v^n)$. Hence,

$$\frac{1}{|\mathcal{T}_P^n|}\sum_{v^n \in \mathcal{T}_P^n} \frac{1}{M(v^n)} \sum_{x^n \in \mathcal{X}_1(v^n)} W(\mathcal{D}(\bar{\mathcal{X}}_2(v^n))|x^n)$$

$$
\begin{aligned}
&\le \frac{1}{|\mathcal{T}_P^n|}\sum_{v^n\in\mathcal{T}_P^n}\frac{1}{M(v^n)}\sum_{j=1}^{M(v^n)}\sum_{x^n\in\mathcal{X}_1(v^n)} W(\mathcal{F}_{\boldsymbol{X}_{2j}\boldsymbol{u}}|x^n)\\
&\stackrel{\cdot}{\le} \frac{1}{|\mathcal{T}_P^n|}\sum_{v^n\in\mathcal{T}_P^n}\frac{1}{M(v^n)}\left(\sum_{k=0}^{n}\left[\sum_{j=1}^{M(v^n)} S_{j\xi_k}(v^n)\right]\exp\{-n[E_2'-3\gamma-\xi_{k+1}]\}+\right.\\
&\left.\sum_{j=1}^{M(v^n)} S_{j\xi_n}(v^n)\right). \qquad (44)
\end{aligned}
$$

Since $\xi_{k+1}-\xi_k = E_2'/n$, and $\sum_{j=1}^{M(v^n)} S_{j\xi_k}(v^n)$ is assumed to be less than $\exp(-n\xi_k)M$ for all k, we obtain

$$
\frac{1}{|\mathcal{T}_P^n|}\sum_{v^n\in\mathcal{T}_P^n}\frac{1}{M(v^n)}\sum_{x^n\in\mathcal{X}_1(v^n)} W(\mathcal{D}(\bar{\mathcal{X}}_2(v^n))|x^n) \stackrel{\cdot}{\le} \exp\{-n(E_2'-3\gamma)\}, \qquad (45)
$$

and so,

$$
\begin{aligned}
&\sum_{v^n} P_0(v^n)\frac{1}{M(v^n)}\sum_{x^n\in\mathcal{X}_1(v^n)} W(\mathcal{D}(\bar{\mathcal{X}}_2(v^n))|x^n)\\
\le\ &\sum_{P}\exp\{-nH(P||P_V)\}\exp\{-n(E_2'-3\gamma)\}\\
\doteq\ &\exp\{-n(E_2-4\gamma)\} \qquad (46)
\end{aligned}
$$

In a similar manner,

$$
\begin{aligned}
&\frac{1}{|\mathcal{T}_P^n|}\sum_{v^n\in\mathcal{T}_P^n}\frac{1}{M(v^n)}\sum_{x^n\in\bar{\mathcal{X}}_2(v^n)} W(\mathcal{D}(\mathcal{X}_1(v^n))|x^n)\\
\le\ &\frac{1}{|\mathcal{T}_P^n|}\sum_{v^n\in\mathcal{T}_P^n}\frac{1}{M(v^n)}\sum_{j=1}^{M(v^n)} W(\bigcup_{x^n\in\mathcal{X}_1(v^n)}\mathcal{F}_{x^nv^n}|\boldsymbol{X}_2(v^n)) \qquad (47)
\end{aligned}
$$

is shown to be asymptotically smaller than $\exp\{-n(E_2'-3\gamma)\}$, which again, when averaged over all type classes, gives $\exp\{-n(E_2-4\gamma)\}$. This completes the proof of Proposition 1, and thus also of Theorem 1. □

B Proof of Theorem 3

Throughout the sequel, we use the common notation for δ–typical sets: The generic notation $\mathcal{T}_Z^n(\delta)$ stands for the set of sequences z^n for which the relative frequencies of all letters are within $\delta/|\mathcal{Z}|$ close to the corresponding letter probabilities of a RV Z on $\mathcal{Z}$. Similar conventions will apply to δ–typical sets of pair sequences w.r.t. a given joint PMF of two random variables, and so on. A conditional δ–typical set, which will be denoted by $\mathcal{T}^n_{U|V}(v^n,\delta)$, is defined as $\{u^n:\ (u^n,v^n)\in T^n_{UV}(\delta)\}$.

We will use the following well–known fact: Let (U,V) be a RV drawn by P_{UV}. Then

$$\exp[-n(I(U;V)+3\delta)] \leq \sum_{(u^n,v^n)\in T^n_{UV}(\delta)} P_{U^n}(u^n)P_{V^n}(v^n) \leq \exp[-n(I(U;V)-3\delta)]. \tag{48}$$

Now, choose some quadruple $(UVXY)\in\mathcal{P}(D-\delta)\cap\mathcal{P}'$ and set the values for J and M as in the theorem. Let $\mathcal{U}_c=\{u^n_{jm},\ 1\leq j\leq J,\ 1\leq m\leq M\}$, where $u^n_{jm}\in\mathcal{U}^n$. First, generate $\mathcal{U}_c$ by random and independent selection of nJM letters from $\mathcal{U}$ according to the marginal of U. Then, reveal $\mathcal{U}_c$ to the encoder and decoder.

Encoding:

To encode a message m in the presence of v^n, first, find the smallest j, denoted $j(v^n,m)$, such that $(u^n_{jm},v^n)\in\mathcal{T}^n_{UV}(\delta/6)$. If no such j exists, set $j(v^n,m)=J$. Next, generate $x^n(v^n,m)$ according to the DMC defined by n independent copies of $Q_{X|UV}$, fed by $(u^n_{j(v^n,m),m},v^n)$. With every v^n, we associate $\bar{x}^n(v^n)$ such that $\rho(v^n,\bar{x}^n(v^n))\leq D$. The encoder is defined by

$$f(v^n,m)=\begin{cases}(j(v^n,m),x^n(v^n,m)) & \text{if } \rho(v^n,x^n(v^n,m))\leq D\\ (j(v^n,m),\bar{x}^n(v^n)) & \text{otherwise}\end{cases} \tag{49}$$

Decoding:

Let $E(y^n)=\{u^n_{jm}\in\mathcal{U}_c:\ (u^n_{jm},y^n)\in\mathcal{T}^n_{UY}(\delta/3)\}$. The decoder is defined by

$$g(y^n)=\begin{cases}(jm) & \text{if } E(y^n) \text{ contains only one element}\\ \text{error} & \text{otherwise.}\end{cases} \tag{50}$$

We now evaluate the error probability of this scheme. To this end, define the events

$$A_1(v^n,m) \quad = \quad \{(u^n_{jm},v^n)\notin\mathcal{T}^n_{UV}(\delta/6),\ \forall\ 1\leq j\leq J\} \tag{51}$$

$$A_{11}(v^n,m) \;=\; \{(v^n, x^n(v^n,m)) \notin \mathcal{T}^n_{VX}(\delta/3)\} \tag{52}$$

$$A_2(v^n,m) \;=\; \{(u^n_{j(v^n,m),m}, y^n) \notin \mathcal{T}^n_{UY}(\delta/3)\} \tag{53}$$

$$A_3(v^n,m) \;=\; \{\,\exists\,(j_1 m_1) \neq (j(v^n,m),m) \;\text{ s.t. } (u^n_{j_1,m_1}, y^n) \in \mathcal{T}^n_{UY}(\delta/3)\}. \tag{54}$$

The error probability given m, averaged over the ensemble of randomized codes and the side information V^n, is given by

$$\begin{aligned}\Pr\{\text{error}|m\} \;\leq\; & \sum_{v^n \in \mathcal{V}^n} P_{V^n}(v^n)\,[\Pr\{A_1(v^n,m)\} + \Pr\{A_{11}(v^n,m)|A_1^c(v^n,m)\} \\ & + \Pr\{A_2(v^n,m)|A_1^c(v^n,m) \cap A_{11}^c(v^n,m)\} \\ & + \Pr\{A_3(v^n,m)|A_1^c(v^n,m) \cap A_{11}^c(v^n,m) \cap A_2^c(v^n,m)\}]\,. \end{aligned} \tag{55}$$

For a typical v^n (a-typical ones have small probability anyhow) we have, by the lower bound of (48):

$$\begin{aligned}\Pr\{A_1(v^n,m)\} \;&\leq\; (1-\exp(-n[I(U;V)+\delta/2]))^{\exp(n[I(U;V)+\delta])} \\ &\leq\; \exp_e(-\exp(n\delta/2)). \end{aligned} \tag{56}$$

Next, conditioning on $A_1^c(v^n,m)$ means that the pair $(u^n_{j(v^n,m),m}, v^n)$ is typical. Thus,

$$\Pr\{A_{11}(v^n,m)|A_1^c(v^n,m)\} \longrightarrow 0 \quad \text{as} \;\; n \to \infty, \tag{57}$$

and the probability of Y^n not being jointly typical with the input vanishes. I.e.,

$$\Pr\{A_2(v^n,m)|A_1^c(v^n,m) \cap A_{11}^c(v^n,m)\} \longrightarrow 0 \quad \text{as} \;\; n \to \infty. \tag{58}$$

We turn to evaluate the last term in (55). Conditioning on $A_1^c(v^n,m) \cap A_2^c(v^n,m)$, the output Y^n must be a typical sequence. Using the upper bound of (48), we obtain

$$\begin{aligned}&\Pr\{A_3(v^n,m)|A_1^c(v^n,m) \cap A_2^c(v^n,m) \cap A_{11}^c(v^n,m)\} \\ &= \;\Pr\left\{\bigcup_{j,m}\left\{U^n_{j,m} y^n \in \mathcal{T}^n_{UY}(\delta/3)\right\}\right\}\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{jm}(P_{U^n}P_{Y^n})\{\mathcal{T}^n_{UY}(\delta/3)\} \\
&\leq JM\exp(-nI(U;Y)+n\delta) \\
&= \exp(nI(U;V)+n\delta+nI(U;Y)-nI(U;V)-4n\delta-nI(U;Y)+n\delta) \\
&= \exp(-2n\delta), \qquad (59)
\end{aligned}
$$

which vanishes as $n\to\infty$. On substituting (56), (58), and (59) in (55), the assertion on the error probability of joint estimation of m and j is proved.

As for the estimate of the entropy of the codeword index within its bin, note that we have proved that, with high probability, for each v^n and message m, $(u^n_{j,m}x^n(v^n,m)^n)$ is jointly typical with the covertext. The estimates (14) now follow by noting that

$$
\exp[n(H(V|U)-\delta/2)] \leq |\mathcal{T}^n_{V|U}(u^n_{j,m},\delta/6)| \leq \exp[n(H(V|U)+\delta/2)]. \qquad (60)
$$

Let $G^m_j=\{v^n:\ j(v^n,m)=j\}$. It follows by the encoding procedure that $G^m_j\subseteq\mathcal{T}^n_{V|U}(u^n_{j,m},\delta/6)$, and thus $|G^m_j|\leq\exp(n[H(V|U)+\delta/2)])$. We now derive a lower bound to $|G^m_j|$. Define $\mathcal{B}_m=\{j:\ |G^m_j|>\exp[n(H(V|U)-2\delta)]\}$. Then,

$$
\begin{aligned}
&(1-\epsilon)\exp(nH(V)-n\delta/6) \leq |\mathcal{T}^n_V(\delta/6)| \leq \sum_{j=1}^{J}|G^m_j| \\
&\quad= \sum_{j\in\mathcal{B}_m}|G^m_j| + \sum_{j\in\mathcal{B}^c_m}|G^m_j| \\
&\quad\leq |\mathcal{B}_m|\exp(nH(V|U)+n\delta/2)+(J-|\mathcal{B}_m|)\exp(nH(V|U)-n2\delta) \\
&\quad\leq |\mathcal{B}_m|\exp(nH(V|U)+n\delta/2)+\exp(nH(V)-n\delta), \qquad (61)
\end{aligned}
$$

where in the last inequality we have neglected the $-|\mathcal{B}_m|$ term, and substituted the value of J. From the last equation, we obtain

$$
|\mathcal{B}_m| \geq \exp(nI(U;V)-5n\delta/6) > J\exp(-2n\delta), \qquad (62)
$$

which is essentially (14a). For (14b), we have to bound the probability of the sets G_j^m, for $j \in \mathcal{B}_m$. This is shown by using the upper and lower bounds on the size of G_j^m in conjunction with the standard bound on probability of a typical v^n. Thus, it remains to show that $\mathcal{B}_m$ has a high probability:

$$\begin{aligned}
\Pr\{\mathcal{T}_V^n(\delta/6)\} &\le \sum_{j\in\mathcal{B}_m} \Pr\{G_j^m\} + \sum_{j\in\mathcal{B}_m^c} \Pr\{G_j^m\} \\
&\le \sum_{j\in\mathcal{B}_m} \Pr\{G_j^m\} + (J - |\mathcal{B}_m|)\exp(nH(V|U) - 2n\delta)\exp(-nH(V) + n\delta/6) \\
&\le \sum_{j\in\mathcal{B}_m} \Pr\{G_j^m\} + \exp(nI(U;V) + n\delta + nH(V|U) - n2\delta - nH(V) + n\delta/6) \\
&= \sum_{j\in\mathcal{B}_m} \Pr\{G_j^m\} + \exp(-n5\delta/6) \qquad (63)
\end{aligned}$$

implying that $\sum_{j\in\mathcal{B}_m} \Pr\{G_j^m\} \to 1$ as as $n \to \infty$, from which (14c) follows. $\square$

C Proof of Theorem 4

The proof of the lower bound in Theorem 4 relies on the common experiment result of Theorem 3, and the upper bound relies on the distribution–packing bound of Lemma 1.

We start from the lower bound. Choose $\delta > 0$, $\epsilon > 0$, and $\theta > 0$. According to Theorem 3, construct two WM transmission codes with common experiments: An $(n, M_n, J_n, \delta, \epsilon, D+\epsilon)$ WM transmission code and an $(\theta n, M_{\theta n}, J_{\theta n}, \delta, \epsilon, D+\epsilon)$ WM transmission code, and interpret these codes as satisfying the maximum error probability and maximum distortion criteria (cf. last paragraph in subsection 3A). In the sequel, we use the notation of Theorem 3, i.e., (f_n, g_n) stands for an encoder–decoder pair for the code of length n, $f_{n,J}$ stands for the projection of f_n to the set of codeword indices, and similarly for the short code, of length θn. Let $L = \{1, 2, \ldots, J_n\} \times \{1, 2, \ldots, M_n\}$. Each element of L is associated with bin index $1 \le k \le M_n$, and an index $1 \le j \le J_n$ of a particular element of its partition. As discussed in the description of the Gel'fand–Pinsker model, any family of maps $T_i : L \longrightarrow \{1, 2, \ldots, M_{\theta n}\}$, $1 \leq i \leq N$, can be used to construct ID codes. The number of different maps is $M_{\theta n}^{|L|}$, and this gives the second exponent claimed in the theorem. The goal is to show that out there is a subset whose size is of the same doubly–exponential rate, and whose members all satisfy the distortion and error probability constraints. These maps are constructed by random selection on $\{1, 2, \ldots, M_{\theta n}\}$.

For every $i \in \{1, 2, \ldots, N\}$ and every $(j, k) \in L$, define a RV whose outcomes are triplets of indices $U_{ijk} = (j, k, \bar{m})$, where $\bar{m}$ is uniformly distributed on $\{1, 2, \ldots, M_{\theta n}\}$, independently of i and of (j, k). Define the random sets $\bar{\mathcal{U}}_i = \{U_{ijk},\ (j, k) \in L\}$. For each realization $\mathcal{U}_i$ of $\bar{\mathcal{U}}_i$, define the distributions

$$Q_i(j, \ell, k, m, x^n, \tilde{x}^{\theta n}, v^n, \tilde{v}^{\theta n}) = \begin{cases} \frac{1}{M_n} P_{V^n}(v^n) P_{V^{\theta n}}(\tilde{v}^{\theta n}) & \text{if } U_{ijk} = (j, k, m), \\ & f_n(v^n, k) = (j, x^n) \\ & \text{and } f_{\theta n}(\tilde{v}^{\theta n}, m) = (\ell, \tilde{x}^{\theta n}) \\ 0 & \text{otherwise,} \end{cases} \tag{64}$$

where $1 \leq j \leq J_n$, $1 \leq \ell \leq J_{\theta n}$, $1 \leq k \leq M_n$, $1 \leq m \leq M_{\theta n}$, and the conditional distributions

$$Q_i(j, \ell, k, m, x^n, \tilde{x}^{\theta n} | v^n, \tilde{v}^{\theta n}) = \frac{Q_i(j, \ell, k, m, x^n, \tilde{x}^{\theta n} v^n, \tilde{v}^{\theta n})}{P_{V^n}(v^n) P_{V^{\theta n}}(\tilde{v}^{\theta n})}. \tag{65}$$

Note that by construction, for any choice of $v^n \tilde{v}^{\theta n}$, k, there is at most only one quintet $(j, \ell, m, x^n, \tilde{x}^{\theta n})$ for which $Q_i(j, \ell, k, m, x^n, \tilde{x}^{\theta n}, v^n, \tilde{v}^{\theta n}) > 0$. Moreover, x^n and $\tilde{x}^{\theta n}$ for which $Q_i > 0$ are the transmission codewords for the pairs (v^n, k) and $(\tilde{v}^{\theta n}, \tilde{x}^{\theta n})$, respectively, and the indices j and ℓ are the common experiments. In particular, we make the obvious association between indices and codewords, e.g., given $(v^n, \tilde{v}^{\theta n})$, the quadruple (j, ℓ, k, m) is associated with the pair $(x^n, \tilde{x}^{\theta n})$ for which $Q_i > 0$, where x^n is the transmission codeword corresponding to the transmission of the pair (j, k), and $\tilde{x}^{\theta n}$ corresponding to transmission of the pair (ℓ, m). In the sequel, we will denote these codewords by $x^n_{j,k}(v^n)$ and $\tilde{x}^{\theta n}_{\ell m}(\tilde{v}^{\theta n})$. Thus, for given realizations $\mathcal{U}_i$ of the random sets $\bar{\mathcal{U}}_i$, the conditional distributions $Q_i(j, \ell, k, m, x^n, \tilde{x}^{\theta n} | v^n, \tilde{v}^{\theta n})$ stand for the random ID codes of Definition 1 over block-length of $n(1+\theta)$.

Let $\mathcal{D}_n^T(j,k) = \{y^n : \ g_n(y^n) = (j,k)\}$, and define the decoding sets for ID in the following manner: the (random) decoding set for the i-th message is $\bar{\mathcal{D}}_i = \cup_{j\ell k}\bar{\mathcal{D}}_i(j,\ell,k)$, where $\bar{\mathcal{D}}_i(j,\ell,k) = \mathcal{D}_n^T(j,k) \times \mathcal{D}_{\theta n}^T(\ell,m)$ if $\bar{U}_{i,j,k} = (j,k,m)$. For convenience, we define also $\bar{\mathcal{D}}_{ij} = \cup_{\ell k}\bar{\mathcal{D}}_i(j,\ell,k)$. The decoding sets $\bar{\mathcal{D}}_i$ are random since $\bar{\mathcal{U}}_i$ are random. Specific realizations of the decoding sets are denoted by $\mathcal{D}_i$ and $\mathcal{D}_{ij}$, respectively.

We now turn to assess the performance of this coding scheme. First, observe that the distortion constraint is satisfied by construction since the transmission codes used to construct the distributions Q_i satisfy the maximal distortion criterion. As for the probability of error of first kind, note that for any realization of the random sets $\bar{\mathcal{U}}_i$

$$\begin{aligned}
&\sum_{j,\ell,k,m}\sum_{v^n,\tilde{v}^{\theta n}} P_{V^n}(v^n)P_{V^{\theta n}}(\tilde{v}^{\theta n})Q_i(j,\ell,k,m|v^n,\tilde{v}^{\theta n})W(\mathcal{D}_i|x_{jk}^n(v^n)x_{\ell m}^{\theta n}(\tilde{v}^{\theta n}))\\
&\geq \sum_{j,\ell,k,m}\sum_{v^n,\tilde{v}^{\theta n}} P_{V^n}(v^n)P_{V^{\theta n}}(\tilde{v}^{\theta n})Q_i(j,\ell,k,m|v^n,\tilde{v}^{\theta n}) \times\\
&\qquad W(\mathcal{D}_n^T(j,k)\times\mathcal{D}_{\theta n}^T(\ell,m)|x_{jk}^n(v^n)x_{\ell m}^{\theta n}(\tilde{v}^{\theta n}))\\
&\geq 1-2\epsilon. \qquad (66)
\end{aligned}$$

For simplicity, we henceforth use the notation x_{jk}^n and $x_{\ell m}^{\theta n}$ for $x_{jk}^n(v^n)$ and $x_{\ell m}^{\theta n}(\tilde{v}^{\theta n})$, as the dependence on v^n is clear from the context. It remains to evaluate the probability of error of second kind. Fix a realization $\mathcal{U}_1$, and define the random variables

$$\Psi_{jk}(\bar{\mathcal{U}}_2) = \begin{cases} 1 & \text{if} \quad \bar{U}_{2jk}\in\mathcal{U}_1 \\ 0 & \text{otherwise.}\end{cases}$$

By construction $E\Psi_{jk}(\bar{\mathcal{U}}_2) = 1/M_{\theta n}$, and, if $\bar{\mathcal{U}}_2 = \mathcal{U}_2$ and $(j,k,m)\in\mathcal{U}_1-\mathcal{U}_2$, we have

$$W(\mathcal{D}_{2,j}|x_{jk}^n x_{\ell m}^{\theta n}) \leq 2\epsilon \qquad (67)$$

for all ℓ. For a given realization $\mathcal{U}_1$, we also define the sets

$$F_1(v^n,\tilde{v}^{\theta n}) = \{(j,\ell,k,m):\ (j,k,m)\in\mathcal{U}_1,\ f_{n,J}(v^n,k)=j,\ f_{\theta n,J}(\tilde{v}^{\theta n},m)=\ell\} \tag{68}$$

$$F_2(v^n,\tilde{v}^{\theta n}) = \{(j,\ell,k,m):\ (j,k,m)\in\mathcal{U}_1\cap\bar{\mathcal{U}}_2,\ f_{n,J}(v^n,k)=j,\ f_{\theta n,J}(\tilde{v}^{\theta n},m)=\ell\} \tag{69}$$

$$F_3(v^n,\tilde{v}^{\theta n}) = \{(j,\ell,k,m):\ (j,k,m)\in\mathcal{U}_1-\bar{\mathcal{U}}_2,\ f_{n,J}(v^n,k)=j,\ f_{\theta n,J}(\tilde{v}^{\theta n},m)=\ell\} \tag{70}$$

$$F_4 = \{(j,\ell,k,m):\ (j,k,m)\in\mathcal{U}_1\cap\bar{\mathcal{U}}_2,\} \tag{71}$$

$$F_5(k,m) = \{(j,\ell):\ (j,k,m)\in\mathcal{U}_1\cap\bar{\mathcal{U}}_2,\ j\in\mathcal{B}_n^{(k)},\ \ell\in\mathcal{B}_{\theta n}^{(m)}\} \tag{72}$$

$$G_1(j,\ell,k,m) = \{(v^n\tilde{v}^{\theta n}):\ f_{n,J}(v^n,k)=j,\ f_{\theta n,J}(\tilde{v}^{\theta n},m)=\ell\}. \tag{73}$$

The sets F_2, F_3, F_4, and F_5, are random since $\bar{\mathcal{U}}_2$ is such. The error probability of the second kind can be upper bounded as follows

$$\begin{aligned}
&\sum_{j,\ell,k,m}\sum_{v^n\tilde{v}^{\theta n}} P_{V^{n(1+\theta)}}(v^n\tilde{v}^{\theta n})Q_1(j,\ell,k,m|v^n\tilde{v}^{\theta n})W(\bar{\mathcal{D}}_{2,j}|x^n_{jk}x^{\theta n}_{\ell m})\\
&= \frac{1}{M_n}\sum_{v^n\tilde{v}^{\theta n}}\sum_{F_1(v^n,\tilde{v}^{\theta n})} P_{V^{n(1+\theta)}}(v^n\tilde{v}^{\theta n})W(\bar{\mathcal{D}}_{2,j}|x^n_{jk}x^{\theta n}_{\ell m})\\
&= \frac{1}{M_n}\sum_{v^n\tilde{v}^{\theta n}}\sum_{F_2(v^n,\tilde{v}^{\theta n})} P_{V^{n(1+\theta)}}(v^n\tilde{v}^{\theta n})W(\bar{\mathcal{D}}_{2,j}|x^n_{jk}x^{\theta n}_{\ell m})\\
&\quad + \frac{1}{M_n}\sum_{v^n\tilde{v}^{\theta n}}\sum_{F_3(v^n,\tilde{v}^{\theta n})} P_{V^{n(1+\theta)}}(v^n\tilde{v}^{\theta n})W(\bar{\mathcal{D}}_{2,j}|x^n_{jk}x^{\theta n}_{\ell m})\\
&\triangleq J_1+J_2,
\end{aligned} \tag{74}$$

where J_1 and J_2 are the third and fourth lines of (74), respectively. By eq. (67), $J_2\le 2\epsilon$, and thus it remains to bound J_1. Using (14), we get

$$\begin{aligned}
J_1 &= \frac{1}{M_n}\sum_{F_4}\sum_{G_1(j,\ell,k,m)} P_{V^n}(v^n)P_{V^{\theta n}}(\tilde{v}^{\theta n})W(\bar{\mathcal{D}}_{2,j}|x^n_{jk}x^{\theta n}_{\ell m})\\
&\le \frac{1}{M_n}\sum_{k,m}\sum_{F_5(k,m)}\sum_{G_1(j,\ell,k,m)} P_{V^n}(v^n)P_{V^{\theta n}}(\tilde{v}^{\theta n})W(\bar{\mathcal{D}}_{2,j}|x^n_{jk}x^{\theta n}_{\ell m})+2\epsilon\\
&\le \frac{\exp(2n\delta(1+\theta))}{M_nJ_nJ_{\theta n}}|\mathcal{U}_1\cap\bar{\mathcal{U}}_2|+2\epsilon.
\end{aligned} \tag{75}$$

Now, $|\mathcal{U}_1 \cap \bar{\mathcal{U}_2}| = \sum_{j,k} \Psi_{j,k}(\bar{\mathcal{U}_2})$. Set λ', $0 < \lambda' < 1$. Then

$$\begin{aligned}
&\Pr\left\{\frac{\exp(2n\delta(1+\theta))}{M_n J_n J_{\theta n}} |\mathcal{U}_1 \cap \bar{\mathcal{U}_2}| > \lambda'\right\} \\
&\leq \Pr\left\{\frac{\exp(2n\delta(1+\theta))}{M_n J_n} |\mathcal{U}_1 \cap \bar{\mathcal{U}_2}| > \lambda'\right\} \\
&= \Pr\left\{\sum_{j,k} \Psi_{j,k}(\bar{\mathcal{U}_2}) > M_n J_n \exp(-2n\delta(1+\theta))\lambda'\right\},
\end{aligned} \tag{76}$$

where there are $M_n J_n = \exp(nI(U;Y) - 3n\delta)$ terms in the sum. Define $\lambda \triangleq \exp(-2n\delta(1+\theta))\lambda'$ and

$$\mu \triangleq E\Psi_{jk}(\bar{\mathcal{U}_2}) = \frac{1}{M_{\theta n}} = \exp\left\{-n\theta[I(U;Y) - I(U;V) - 4\delta]\right\}.$$

Now, let θ satisfy

$$\theta > \frac{3\delta}{I(U;Y) - I(U;V) - 6\delta}. \tag{77}$$

It can be verified that with such a choice, we will have $0 < \mu < \lambda < 1$ for any λ' and sufficiently large n. Now, applying Lemma 5 and Lemma 6 (to follow), we get

$$\begin{aligned}
&\Pr\left\{\sum_{j,k} \Psi_{jk}(\bar{\mathcal{U}_2}) > M_n J_n \exp(-2n\delta(1+\theta))\lambda'\right\} \\
&\leq \exp\left(-M_n J_n H(\exp(-2n\delta(1+\theta))\lambda' \| \exp(-n\theta(I(U;Y) - I(U;V) - 4\delta))\right)\right) \\
&\leq \exp(-M_n J_n \lambda' \exp(-2n\delta(1+\theta))).
\end{aligned} \tag{78}$$

Thus, by the union bound, we can construct $\exp(M_n J_n \exp(-3n\delta)) = \exp(\exp(I(U;Y) - 6n\delta))$ mappings $\{T_i\}$ and still get vanishing error probability of second kind. With the upper bound we have for the error probability of the first kind (66), we obtain

$$C_{ID}^T(D|V) \geq \sup_{\mathcal{P}(D)\cap\mathcal{P}'} I(U;Y), \tag{79}$$

where the positivity constraint $\mathcal{P}'$ follows from the construction of the ID code: we used a transmission code with common experiment. This completes the proof of the lower bound.

We now turn to the upper bound. By Lemma 1, we have

$$C_{ID}^{T}(D|V) \leq \lim_{n\to\infty} \frac{1}{n} \max I(V^n X^n; Y^n), \tag{80}$$

where the maximum is taken over all input distributions X^n satisfying the average distortion constraint, i.e., $E\rho(X^n, V^n) \leq D$. Thus, we only have to express the r.h.s. of (80) as a single-letter expression. The average distortion constraint can be written as

$$D \geq \frac{1}{n} \sum_{i=1}^{n} \sum_{v} \sum_{x} P_V(v) P_{X_i|V}(x|v) \rho(v, x). \tag{81}$$

Define a random variable $\tilde{X}$ whose distribution is a uniform mixture of those of the X_i's, $i = 1, 2, \ldots, n$, i.e., $P_{\tilde{X}V}(x, v) = \frac{1}{n} \sum_{i=1}^{n} P_V(v) P_{X_i|V}(x|v)$. Then the V–marginal of $P_{\tilde{X}V}$ remains P_V, and by (81)

$$E_{\tilde{X}V} \rho(\tilde{X}, V) \leq D. \tag{82}$$

Using the concavity of the mutual information w.r.t. the input distribution, we get

$$\frac{1}{n} I(V^n X^n; Y^n) \leq \frac{1}{n} \sum_{i=1}^{n} I(V_i X_i; Y_i) \leq I(V\tilde{X}; \tilde{Y}), \tag{83}$$

where the first inequality is due to the fact that the channel is memoryless, and where $\tilde{Y}$ is the channel output corresponding to $\tilde{X}$. We now set $U = V\tilde{X}$, and the upper bound follows by (80), (82), and (83). $\square$

Lemma 5 (LD) [2]: *If Ψ_j, $j = 1, \ldots, \bar{N}$, are iid random variables with values in $\{0, 1\}$, then*

$$Pr\left\{ \sum_{j=1}^{\bar{N}} \Psi_j > \bar{N}\lambda \right\} \leq \exp(-\bar{N} H(\lambda || \mu)), \tag{84}$$

where $E\Psi_1 \leq \mu < \lambda \leq 1$, and $H(\lambda||\mu)$ is the relative entropy of $(\lambda, 1-\lambda)$ with respect to $(\mu, 1-\mu)$.

Lemma 6 *If $0 < \alpha < \beta$ and $c > 0$, $\lim_{n\to\infty} H(ce^{-\alpha n}||e^{-\beta n})/(ce^{-\alpha n}(\beta - \alpha)n) = 1$, implying that $H(ce^{-\alpha n}||e^{-\beta n}) > ce^{-\alpha n}$ for all sufficiently large n.*

Proof: Use L'hospital's rule. □

5 Conclusion and Future Research

In this paper, we made an attempt to provide an information–theoretic framework to the problem of digital WM by viewing it as an ID system with side information. We provided an exact single–letter capacity formula for memoryless ID channels with side information at both transmitter and receiver, and bounds on the capacity in the case of side information at the transmitter only. In the former case (side information at both ends) our analysis lead also to a characterization of achievable error exponents.

Future research will be aimed at several goals:

First, for completeness, the gap between the upper and lower bounds on ID capacity should be closed, and achievable error exponents should be analyzed, for the case where the side information is available at the transmitter only.

Secondly, it would be interesting to bridge over between the two extremes, where the side information is either fully available or completely unavailable to the receiver. This can be done by adding to our model a secret "key" source $\{K_n\}$ that is available at both ends (but not to the attacker), and which may be statistically related to the covertext $\{V_n\}$, similarly as is done in [9]. Here $K_n = V_n$ corresponds to full side information at the decoder, and on the other extreme, the case where K_n is a deterministic, fixed symbol corresponds to no side information at the decoder. This key model also adds a new dimension to the problem, namely, the optimum design of the conditional probability distribution $P(K_n|V_n)$, which balances the tradeoff among several factors, such as the quality of side information at the decoder, the rate of the common experiment, and the security of the system.

Third, one may argue that capacity may not be the relevant objective function in this problem as it implicitly assumes that the decoder knows enough about the attack channel. It might be interesting, therefore, to characterize achievable identification rates in the absence of such an assumption [13],[14].

Finally, we mentioned in the Introduction that doubly–exponentially many messages may not be really needed in watermarking applications, which typically require relatively low information rates. The natural question to be raised, in this context, is how exactly to exploit this excess power of the system in order to significantly reduce distortion and error probabilities at a given information rate in the practical range.

Acknowledgment: Useful discussions with Shlomo Shamai are gratefully acknowledged.

References

[1] R. Ahlswede and G. Dueck, "Good codes can be produced by a few permutations," *IEEE Trans. Inform. Theory*, vol. IT–28, no. 3, pp. 430–443, May 1982.

[2] —, "Identification via channels," *IEEE Trans. Inform. Theory*, vol. 35, pp. 15-29, January 1989.

[3] —,"Identification in the presence of feedback – a discovery of new capacity formulas," *IEEE Trans. Inform. Theory*, vol. 35, pp. 30-36, January 1989.

[4] M. Barni, F. Bartolini, A. De Rosa, and A. Piva, "Capacity of the watermark channel: how many bits can be hidden within a digital image?" *Proc. SPIE*, vol. 3657, pp. 437-448, San Jose, CA, January 1999.

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

[6] S. I. Gel'fand and M. S. Pinsker, "Coding for channel with random parameters," *Probl. Inform. & Control*, vol. 9, no. 1, pp. 19-31, 1980.

[7] T. S. Han and S. Verdú, "New results in the theory of identification via channels," *IEEE Trans. Inform. Theory*, vol. IT–38, pp. 14–25, Jan. 1992.

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

[9] P. Moulin and J. A. O'Sullivan, "Information-theoretic analysis of information hiding," preprint, 1999.

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

[11] S. D. Servetto, C. I. Podilchuk, and K. Ramchandran, "Capacity issues in digital image watermarking," *Proc. ICIP '98*, 1998.

[12] Y. Steinberg, "New converses in the theory of identification via channels," *IEEE Trans. Inform. Theory*, vol. 44, pp. 984-998, May 1998.

[13] I. G. Stiglitz, "Coding for a class of unknown channels," *IEEE Trans. Inform. Theory*, vol. IT–12, no. 2, pp. 189–195, April 1966. (correction: vol. IT–13, no. 2, p. 195).

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